METROPOLIS METHODS, GAUSSIAN PROPOSALS AND ANTITHETIC VARIABLES

Peter J. Green and Xiao-liang Han

University of Bristol,
Department of Mathematics,
Bristol, BS8 1TW, UK.

ABSTRACT

We investigate various aspects of a class of dynamic Monte Carlo methods, that generalises the Metropolis algorithm and includes the Gibbs sampler as a special case. These can be used to estimate expectations of marginal distributions in stochastic systems. A distinction is drawn between speed of weak convergence and precision of estimation. For continuously distributed processes, a particular gaussian proposal distribution is suggested: this incorporates a parameter that may be varied to improve the performance of the sampling method, by adjusting the magnitude of an "antithetic" element introduced into the sampling. The suggestion is examined in detail in some experiments based on an image analysis problem.

Keywords: *autocorrelation time, convergence rate, dynamic Monte Carlo, Gibbs sampler, statistical image analysis.*

1. Introduction

Complex stochastic systems, large collections of random variables with non-trivial dependence structure, arise in probability modelling in many contexts. Examples include statistical mechanics, geographical epidemiology, pedigrees in genetics, statistical image analysis, and general multi-parameter Bayesian inference. In practice, distributions of variables in such systems are usually not amenable either to exact numerical calculation or to direct simulation. The dynamic Monte Carlo approach to the computation of probabilities and expectations in these systems is receiving much attention. The basic idea in such methods is to turn the given static problem into a dynamic one by constructing what is usually an artificial temporal stochastic process, that is known to converge weakly to the distribution of the original system, yet is easy to simulate. This temporal process is generally taken to be a time-homogeneous Markov chain, so that apart from the matter of checking the aperiodicity and irreducibility of the constructed process (and some additional regularity conditions when the state space is uncountable), analysis of the original problem is reduced to the construction of a Markov chain with a specified equilibrium distribution.

There will always be very many such chains, and the art of dynamic Monte Carlo is to choose one that uses computational resources effectively by striking the right balance between simplicity and speed. This paper is a contribution to the discussion of how to strike this balance. It is stimulated by our interest in a class of problems arising in a Bayesian approach to low-level image analysis, but in fact our observations and conclusions will be quite generally applicable.

After a brief survey of the statistical literature on dynamic Monte Carlo, the remainder of the paper is divided into three sections. In Section 2, we discuss how to define speed of convergence appropriately when estimating properties of the equilibrium distribution, and how to monitor and quantify convergence from a sample realisation. Section 3 contains a new proposal for a class of Markov chains with a given equilibrium, and an analysis of some properties of this class in certain simple situations. In Section 4, we present the results of some experiments making use of this class of chains in an idealised Bayesian image analysis context.

Dynamic Monte Carlo methods originated in computational physics research; the paper by Metropolis *et al.* (1953) is usually taken as starting the subject. An enormous literature has developed: the current state of the subject, with particular reference to applications in statistical mechanics and quantum field theory, is summarised in the excellent lecture notes of Sokal (1989). The monograph of Hammersley and Handscomb (1964) introduced the ideas to the statistical community, but only with the development of stochastic models for image analysis, and Geman and Geman's proposal (1984) to use the Gibbs sampler in image reconstruction, did interest really develop. The Metropolis algorithm and Gibbs sampler are described in Ripley (1987, p.113), and will be further discussed in Section 3. Many other applications have now been addressed by such approaches, including Monte Carlo testing, for example in the Rasch model, (Besag and Clifford, 1989), marginal distributions in Bayesian inference (Gelfand and Smith, 1990), and geographical epidemiology (Besag, York and Mollié, 1991).

There are connections with the development of simulated annealing as an approach to combinatorial optimisation, proposed by Kirkpatrick *et al.* (1983) as an analogue of annealing in physical systems, and also utilised by Geman and Geman (1984) in image reconstruction. Here a time-inhomogeneous Markov chain is constructed; instantaneously the transition mechanism has an equilibrium distribution that is a renormalised power of the distribution of interest. As the process evolves, this power is gradually increased: in the limit the distribution is concentrated on the value(s) of maximum probability, and if the power is increased sufficiently slowly, the process can be shown to converge to this maximum. We do not consider annealing in this paper.

2. Speed of convergence

We are concerned with a random vector $\mathbf{x}$ with components x_i indexed by $i \in S$, a finite set of sites or pixels. The set of possible values for $\mathbf{x}$ will be denoted by Ω which usually has the form C^S where C might be finite, countable, or an interval in $\mathbf{R}$ or $\mathbf{R}^d$ depending on context. For the most part, in this paper we will use notation appropriate to the countable case. In addition, in most applications there is a vector of observables $\mathbf{y}$. The object of interest is the distribution of $\mathbf{x}$ given $\mathbf{y}$: we denote its density with respect to an appropriate measure as $p(\mathbf{x}|\mathbf{y})$. In the context of image analysis, $\mathbf{y}$ represents an observed pixellated degraded digital image, and $\mathbf{x}$ an unobservable true image representing the "state of nature": study of the posterior distribution $p(\mathbf{x}|\mathbf{y})$ is our route to drawing inference about this truth. Note that in this and other examples, $\mathbf{x}$ and $\mathbf{y}$ are vectors of high dimensionality.

Our discussion is also applicable to the case where there are no observables: just regard $\mathbf{y}$ as null. We will use the symbol $\pi(\mathbf{x})$ to denote whichever of $p(\mathbf{x})$ or $p(\mathbf{x}|\mathbf{y})$ is of interest. The same computational ideas apply to both cases: thus Grenander's perceptive observation (1983, p.83) that "pattern analysis equals pattern synthesis".

Attention usually focusses on one or more functionals of the distribution $\pi(\mathbf{x})$: suppose that we wish to estimate the expectation $E_\pi(f) = \sum f(\mathbf{x})\pi(\mathbf{x})$. This is very general, for example $f(\mathbf{x})$ might be $\sum_{i \in A} x_i$ (total truth in a region $A \subset S$) or $I_{[x_k=0]}$ (to estimate the probability of a zero; $I_{[]}$ is the indicator function). If $V(\mathbf{x})$ is a sufficient statistic for a parameter β in the model $\pi(\mathbf{x})$, then $f(\mathbf{x})$ might be $V(\mathbf{x})$ or $I_{[V(\mathbf{x}) \leq t]}$, in order to construct procedures for inference about β.

Let P be a Markov transition function on the state space Ω, that is irreducible and aperiodic, and has π as equilibrium distribution, so that

$$\sum_{\mathbf{x} \in \Omega} \pi(\mathbf{x}) P(\mathbf{x}, \mathbf{x}') = \pi(\mathbf{x}')$$

for all $\mathbf{x}' \in \Omega$. Suppose we have a partial realisation $\{\mathbf{x}^{(t)} : t = 0,1,2,\ldots,N\}$ from this Markov chain. Then our estimator of $E_\pi(f)$ will be the empirical average

$$\bar{f}_N = \frac{1}{N} \sum_{t=1}^{N} f(\mathbf{x}^{(t)}).$$

Motivated by the observation in Gelfand and Smith (1990) that Rao-Blackwellisation can be used to reduce mean squared error, in the case of replicated independent Monte Carlo runs, we are investigating the performance of modified estimators exploiting conditioning. For example, if $f(\mathbf{x})$ is actually $f_1(x_1)$, a function of a single component of $\mathbf{x}$, then $g(\mathbf{x}) = E_\pi(f_1(x_1)|\mathbf{x}_{S\setminus 1})$ may sometimes be cheaply computed. It has the same expectation $E_\pi(f) = E_\pi(g)$, but $\bar{g}_N$ will have smaller mean squared error than $\bar{f}_N$. Since these have the same general form, the ensuing treatment continues to apply; we will not discuss this modification specifically here, but further details will be reported elsewhere.

How good are such estimators? This seems to depend on how fast the Markov chain $\{\mathbf{x}^{(t)}\}$ converges weakly to π. Under the irreducibility condition, P has only the single eigenvalue 1 on the unit circle, with the constant vectors as the corresponding right eigenvectors, so that the rate of convergence is given by R, the spectral radius of P acting on the orthogonal complement of the constant vectors, and we find

$$\sup_{\mathbf{x}, \mathbf{x}' \in \Omega} |P^t(\mathbf{x}, \mathbf{x}') - \pi(\mathbf{x}')| \sim cR^t. \tag{1}$$

Then for any bounded, continuous function f,

$$|E(f(\mathbf{x}^{(t)})) - E_\pi(f)| \leq c(f)R^t. \tag{2}$$

Using R, we can define the exponential autocorrelation time $(-1/\log R)$: the number of steps of the Markov chain needed to reduce the "errors" $|P^t(\mathbf{x}, \mathbf{x}') - \pi(\mathbf{x}')|$ by a factor of e asymptotically. Small R (small autocorrelation time) indicates rapid convergence, but will actually be a pessimistic measure for any particular f, for which the chain may achieve faster convergence of $E(f(\mathbf{x}^{(t)}))$.

But this discussion does not address performance of the estimator $\bar{f}_N$, obtained from the sample path of the process by integration over time, not over realisation. This estimator has bias and variance, whose asymptotic forms are:

$$\begin{aligned} E(\bar{f}_N)-E_\pi(f) &= \frac{1}{N}\sum_{t=1}^{N}\{E(f(\mathbf{x}^{(t)}))-E_\pi(f)\} \\ &\sim \frac{1}{N}\sum_{t=1}^{\infty}\{E(f(\mathbf{x}^{(t)}))-E_\pi(f)\}; \qquad (3) \\ var(\bar{f}_N) &= \frac{1}{N^2}\sum_{s=1}^{N}\sum_{t=1}^{N} cov(f(\mathbf{x}^{(s)}),f(\mathbf{x}^{(t)})) \\ &\sim \frac{\sigma^2}{N}\sum_{t=-(N-1)}^{N-1}(1-\frac{|t|}{N})\rho_t(f) \\ &\sim \frac{\sigma^2}{N}\sum_{t=-\infty}^{\infty}\rho_t(f) \qquad (4) \end{aligned}$$

where $\rho_t(f)$ is the autocorrelation function of the process $\{f(\mathbf{x}^{(t)})\}$, calculated under the equilibrium distribution π, and σ^2 is the equilibrium variance of $f(\mathbf{x})$. The asymptotic variance is a factor

$$\tau(f) = \sum_{t=-\infty}^{\infty}\rho_t(f)$$

times what would be obtained if independent random sampling of $\mathbf{x}$ from $\pi(\mathbf{x})$ could be achieved: we call $\tau(f)$ the integrated autocorrelation time (differing from Sokal's definition (1989) by a factor of 2). From (2), (3) and (4), it is evident that the asymptotic mean squared error of $\bar{f}_N$ as an estimator of $E_\pi(f)$ is determined by the variance, which is of order N^{-1} while that of the squared bias is N^{-2}.

For clarification of the distinction between rapid convergence (small R in (1) and (2)) and good estimation performance (small $\tau(f)$), it is helpful to study the finite reversible case, where explicit expressions can be given.

Suppose the Markov chain P is finite, reversible, irreducible and aperiodic, and that B is the diagonal matrix with entries $(\pi(\mathbf{x}),\mathbf{x}\in\Omega)$, the equilibrium probabilities for P. Reversibility means that $\pi(\mathbf{x})P(\mathbf{x},\mathbf{x}')=\pi(\mathbf{x}')P(\mathbf{x}',\mathbf{x})$ for all $\mathbf{x},\mathbf{x}'\in\Omega$, so that BP is a symmetric matrix. We then have the spectral representation

$$P = E\Lambda E^{\mathrm{T}}B$$

where $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\ldots)$ is a diagonal matrix of eigenvalues of P, which are real, and E is the matrix whose columns are the corresponding right eigenvectors, normalised to be orthogonal with respect to B, so that $E^{\mathrm{T}}BE=I$. We take λ_1 to be the unique unit eigenvalue, so that $E_{\mathbf{x}1}=1$ for all $\mathbf{x}\in\Omega$. Then $P^n=E\Lambda^nE^{\mathrm{T}}B\to E\,\mathrm{diag}(1,0,\ldots,0)E^{\mathrm{T}}B=\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$ as expected.

If $\mathbf{f}$ is the vector with components $(f(\mathbf{x}),\mathbf{x}\in\Omega)$ and $\mathbf{a}$ the vector of initial probabilities for the chain, then

$$\begin{aligned} E(f(\mathbf{x}^{(t)})) &= \mathbf{a}^{\mathrm{T}}P^t\mathbf{f} \\ &= \sum_k \lambda_k^t(E^{\mathrm{T}}\mathbf{a})_k(E^{\mathrm{T}}B\mathbf{f})_k. \end{aligned}$$

So

$$E(f(\mathbf{x}^{(t)}))-E_\pi(f) = E(f(\mathbf{x}^{(t)}))-\sum_{\mathbf{x}}\pi(\mathbf{x})f(\mathbf{x})$$
$$= \sum_{k\geq 2}\lambda_k^t(E^{\mathrm{T}}\mathbf{a})_k(E^{\mathrm{T}}B\mathbf{f})_k.$$

If λ_2 is an eigenvalue second largest in absolute value, then $|\lambda_2|=R$ and

$$|E(f(\mathbf{x}^{(t)}))-E_\pi(f)| = O(R^t), \tag{5}$$

where the multiplier depends on the initial distribution, the particular functional of interest, and on the transition matrix P.

Turning to the empirical average $\bar{f}_N$, we have

$$E(\bar{f}_N) = \frac{1}{N}\sum_{t=1}^{N}\sum_k \lambda_k^t(E^{\mathrm{T}}\mathbf{a})_k(E^{\mathrm{T}}B\mathbf{f})_k,$$

whence the bias is

$$E(\bar{f}_N)-E_\pi(f) = \sum_{k\geq 2}\{\frac{1}{N}\sum_{t=1}^{N}\lambda_k^t\}(E^{\mathrm{T}}\mathbf{a})_k(E^{\mathrm{T}}B\mathbf{f})_k$$
$$\sim \frac{1}{N}\sum_{k\geq 2}\frac{\lambda_k}{1-\lambda_k}(E^{\mathrm{T}}\mathbf{a})_k(E^{\mathrm{T}}B\mathbf{f})_k. \tag{6}$$

As for the variance, the equilibrium autocovariance is

$$\sigma^2\rho_t(f) = \sum_{\mathbf{x}}\sum_{\mathbf{x}'}f(\mathbf{x})f(\mathbf{x}')\pi(\mathbf{x})(P^t(\mathbf{x},\mathbf{x}')-\pi(\mathbf{x}'))$$
$$= \mathbf{f}^{\mathrm{T}}BP^t\mathbf{f}-\mathbf{f}^{\mathrm{T}}\boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{f}$$
$$= \mathbf{f}B(P^t-\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})\mathbf{f}.$$

Now $(P^t-\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}) = (P-\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})^t$ for $t\geq 1$ so

$$\sum_{t=0}^{\infty}(P^t-\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}) = (I-P+\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})^{-1}-\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}.$$

But from (4)

$$N\mathrm{var}(\bar{f}_N) \sim \sigma^2\tau(f)$$
$$= \sigma^2\sum_{t=-\infty}^{\infty}\rho_t(f) = 2\sigma^2\sum_{t=0}^{\infty}\rho_t(f)-\sigma^2$$
$$= \mathbf{f}^{\mathrm{T}}\{2B(I-P+\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})^{-1}-2B\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}-B(P^0-\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})\}\mathbf{f}$$
$$= \mathbf{f}^{\mathrm{T}}BE\,\mathrm{diag}\{2(1-\lambda_k+\delta_{k1})^{-1}-2\delta_{k1}-1+\delta_{k1}\}E^{\mathrm{T}}B\mathbf{f}$$
$$= \sum_{k\geq 2}\frac{1+\lambda_k}{1-\lambda_k}(E^{\mathrm{T}}B\mathbf{f})_k^2. \tag{7}$$

The matrix expression is noted by Peskun (1973), and the spectral expansion by Sokal (1989) and Frigessi, Hwang and Younes (1990).

Contrasting (5) with (7), we see that rapid weak convergence to equilibrium is obtained by having all eigenvalues λ_k other than $\lambda_1 = 1$ small in absolute value, whilst

good asymptotic mean squared error of estimation is suggested by having $(1+\lambda_k)/(1-\lambda_k)$ small: "negative eigenvalues help". The rôle played by the eigenvectors in (7) should not be neglected, however, as two alternative transition matrices P will in general differ not only in their eigenvalues.

In practice, with a finite Monte Carlo sample size N, both of these aspects of convergence are relevant. The very complexity of the distribution π which suggested consideration of Monte Carlo simulation in the first place inhibits explicit calculation of eigen-decompositions, of course, and we need diagnostics for studying the rate of weak convergence and methods for estimating the integrated autocorrelation time. Such tools will be used both in studies aimed at making general recommendations, such as the present one, and routinely in the actual use of dynamic Monte Carlo methods (not least in order to attach standard errors to estimates of $E_\pi(f)$). Blind application of Gibbs or Metropolis samplers, with no examination of these issues, can produce completely meaningless results.

The conflicting demands of small $\sup_{k\geq 2}|\lambda_k|$ and small $(1+\lambda_k)/(1-\lambda_k)$ suggest a revised strategy of switching between different transition mechanisms as the simulation proceeds, producing a time-inhomogeneous Markov chain. In its simplest form, which seems to be commonly used in the physics literature, the idea would be to use an initial process P_0 for the first N_0 iterations, then to switch to P for another N updates. P_0 would be chosen to give rapid convergence to equilibrium and P for a small $\tau(f)$. The switch would take place when diagnostics suggested that the process was effectively in equilibrium, and the first N_0 iterations discarded for estimation purposes, so that the estimator is

$$\frac{1}{N}\sum_{t=N_0+1}^{N_0+N} f(\mathbf{x}^{(t)}).$$

More complicated variants of this, perhaps involving continuous alteration of the transition mechanism, and/or weighted averages of $\{f(\mathbf{x}^{(t)})\}$ may be worth exploring; another factor that can influence this discussion arises when the cost of computing f is high relative to that of the Markov transition, which will support sub-sampling the chain at equally spaced times at which $f(\mathbf{x}^{(t)})$ is computed, with a corresponding modification to the definition of autocorrelation time.

Rather than sample repeatedly from a single run of the process, some authors, for example Gelfand and Smith (1990), propose evaluating $f(\mathbf{x}^{(t)})$ only once (so that $N=1$, although N_0 is large), but then repeatedly restarting the whole process, so as to be able to average completely independent values of $f(\mathbf{x})$. But this seems to us to be inefficient, at least in the situations of our experience, where $\tau(f)$ is much less than $(-1/\log R)$ and $var_\pi(f)$ is sufficiently large that say 100 or 1000 effectively independent observations will be needed to estimate $E_\pi(f)$ to adequate precision. In this situation, more computing effort would be used to achieve the same precision if the chain were restarted.

We have only tentative recommendations to make regarding the diagnostic monitoring of convergence. Practical considerations limit attention to a few scalar-valued functionals $f_1(\mathbf{x}^{(t)}), f_2(\mathbf{x}^{(t)}), \ldots$, although there may be merit in also measuring aspects of several $\mathbf{x}^{(t)}$ jointly, such as a summary of the magnitude of the difference between

$\mathbf{x}^{(t-1)}$ and $\mathbf{x}^{(t)}$. Each functional f_j will have a characteristic spectral radius R_j governing the rate of convergence of its distribution to equilibrium: none of these can exceed R (see equation(2)). The hope in selecting a range of such functionals for study is that at least one R_j is close to R, so that we are not misled into an over-optimistic impression of the rate of convergence of the process as a whole. Our approach is then simply to plot the values of these functionals against iteration number: the clarity of the visual impression given about convergence depends on the equilibrium variance of the functional. See the examples in Section 4.

Estimating the integrated autocorrelation time $\tau(f)$, for any particular functional f, is a standard problem from the analysis of stationary time series. (We suppose that we only address this question after discarding the initial N_0 iterations according to the criteria just described, so that we can regard the process as in equilibrium). The integrated autocorrelation time is simply 2π times the normalised power spectral density function of the process evaluated at frequency 0 (Priestley, 1981, p.225), so we are dealing with a special case of spectral density estimation. The difficulties are well-known, the naive estimator $\sum_{t=-\infty}^{+\infty}\hat{\rho}_t(f)$ using the sample autocorrelations of the observed process being inconsistent as the length of the observed series increases (Priestley, p.429). The conventional solution is to apply a spectral window, that is to use a weighted estimator $\sum w_t\hat{\rho}_t(f)$, where the lag window function w_t decreases to 0 as $t\to\pm\infty$. In particular, Sokal (1989) recommends the truncated periodogram estimator $\sum_{|t|\le M}\hat{\rho}_t(f)$, with the window width M chosen adaptively as the minimum integer with $M\ge 3\hat{\tau}(f)$. In our implementation of this we estimate the autocorrelations from the Fourier transform of the process, thus wrapping the time axis onto a circle; this approximation is acceptable if $\tau(f)$ is small relative to the length of the series.

An alternative non-parametric estimator of $\tau(f)$ which is also appealing turns out to be related to the spectral density estimator using the Bartlett window (Priestley, p.439), and is recommended by Hastings (1970). If $N=bk$ and the series is broken into b non-overlapping blocks of k consecutive observations, then the between-blocks mean square

$$\frac{k}{b-1}\sum_{j=1}^{b}\left[\{\frac{1}{k}\sum_{t=(j-1)k+1}^{jk} f(\mathbf{x}^{(t)})\}-\bar{f}_N\right]^2 \tag{8}$$

is an approximately unbiased estimator of $\sigma^2\tau(f)$ as b and $k\to\infty$.

3. Gaussian proposals in the Metropolis method

Motivated by our interest in image analysis problems, we consider here a new class of samplers appropriate to the continuous case, where $\Omega=\mathbf{R}^S$, with particular emphasis on designing Markov chain methods with small integrated autocorrelation time.

The best known dynamic Monte Carlo method is the Metropolis algorithm (Metropolis, *et al.*, 1953). Here we describe it in the interesting variant due to Hastings (1970). Recall that we wish to construct a Markov chain with a prescribed equilibrium distribution $\pi(\mathbf{x})$. Let $q(\mathbf{x},\mathbf{x}')$ be an arbitrary irreducible aperiodic transition function on $\Omega\times\Omega$: how can this be modified to achieve the required equilibrium? Given $\mathbf{x}^{(t)}=\mathbf{x}$, a proposal $\mathbf{x}'$ is drawn from $q(\mathbf{x},\mathbf{x}')$, but not immediately taken as the

new state of the chain. Rather, it is only accepted, and $\mathbf{x}^{(t+1)}$ set equal to $\mathbf{x}'$, with probability $\alpha(\mathbf{x},\mathbf{x}')$; otherwise it is rejected, and no move is made, so that $\mathbf{x}^{(t+1)} = \mathbf{x}$. The acceptance probability can always be chosen so that detailed balance is obtained:

$$\pi(\mathbf{x})P(\mathbf{x},\mathbf{x}') = \pi(\mathbf{x}')P(\mathbf{x}',\mathbf{x}) \tag{9}$$

for all $\mathbf{x},\mathbf{x}' \in \Omega$. One possibility for α is

$$\alpha(\mathbf{x},\mathbf{x}') = \min\left\{1, \frac{\pi(\mathbf{x}')q(\mathbf{x}',\mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x},\mathbf{x}')}\right\}, \tag{10}$$

for which (9) is easily verified, the corresponding transition function being

$$\begin{aligned} P(\mathbf{x},\mathbf{x}') &= q(\mathbf{x},\mathbf{x}')\alpha(\mathbf{x},\mathbf{x}') \quad \mathbf{x}' \neq \mathbf{x} \\ &= 1 - \sum_{\mathbf{x}' \neq \mathbf{x}} q(\mathbf{x},\mathbf{x}')\alpha(\mathbf{x},\mathbf{x}') \quad \mathbf{x}' = \mathbf{x} \end{aligned}$$

Among all possible α achieving detailed balance for a given q, the particular choice in (10) is shown by Peskun (1973) to give minimum integrated autocorrelation time.

This prescription is very general, and can be used to generate a wide variety of Markov chain simulation methods for different problems. The process $\mathbf{x}^{(t)}$ is usually highly multivariate, and in practice we usually concentrate on algorithms which only change one component of $\mathbf{x}$ at a time. (There are notable exceptions in special cases, for example the algorithm of Swendsen and Wang (1987)). This does not affect the validity of (10), but merely facilitates its computation. There are various valid ways to choose which component, i, of $\mathbf{x}^{(t)}$ is to be updated in the transition to $\mathbf{x}^{(t+1)}$: the common ones being a systematic choice, cycling through $i \in S$ in some fixed order, or a random choice, drawing i at random each time. The choice is reflected in q.

In the original application of this idea, Metropolis *et al.* (1953) considered a finite set of "colours": $\Omega = \{0,1,\ldots,L-1\}^S$, and the proposal that takes a uniformly distributed choice from among the $L-1$ colours different from the current one. In this case, and whenever there is symmetry of the proposal distribution, $q(\mathbf{x},\mathbf{x}') = q(\mathbf{x}',\mathbf{x})$, the expression for $\alpha(\mathbf{x},\mathbf{x}')$ simplifies to $\min\{1,\pi(\mathbf{x}')/\pi(\mathbf{x})\}$. But we shall see there is something to be gained by the slightly greater generality.

Two other points might be made about this prescription. One is the "distribution-free" nature of the simulation step: the transition function q is quite arbitrary (provided that $q(\mathbf{x},\mathbf{x}')$ and $q(\mathbf{x}',\mathbf{x})$ are either both zero or both positive). So use of the method is not restricted to those $\pi(\mathbf{x})$ which are convenient for simulation: the model π only enters the algorithm through the calculation of $\pi(\mathbf{x}')/\pi(\mathbf{x})$ in the definition of α. The second point is that although the whole procedure has the flavour of the conventional rejection methods for static Monte Carlo simulation, there is no requirement, as there, for the density that is used for simulation to envelope (a submultiple of) the density of interest.

One particular algorithm in this class has received a good deal of attention in the recent statistical literature: the Gibbs sampler. The proposal distribution q is defined as follows: a pixel i is chosen from S uniformly at random, the current value $x_i^{(t)}$ deleted, and the proposed new value drawn from the conditional distribution, under π, of x_i given the values of all other pixels: thus

$$q(\mathbf{x},\mathbf{x}') = \frac{1}{|S|}\sum_{i\in S}\pi(x'_i|x_{S\setminus i})$$

(Systematic scanning over the pixels is also commonly used). It is trivial to see that for $\mathbf{x}\neq\mathbf{x}'\in\Omega$, $q(\mathbf{x},\mathbf{x}')$ and $q(\mathbf{x}',\mathbf{x})$ are each positive if and only if $\mathbf{x}$ and $\mathbf{x}'$ differ in at most one coordinate, and that in that case

$$\frac{q(\mathbf{x},\mathbf{x}')}{q(\mathbf{x}',\mathbf{x})} = \frac{\pi(\mathbf{x}')}{\pi(\mathbf{x})}$$

so that the corresponding acceptance probability $\alpha(\mathbf{x},\mathbf{x}')$ is identically 1. From this point of view, the Gibbs sampler, or heat-bath method as it is known in the physical literature, is but an extreme form of the Metropolis method, with a highly model-dependent choice of proposal and zero probability of rejection.

Although informal heuristics suggest that eliminating rejection should reduce the integrated autocorrelation time, the computational price paid may be high. For the Gibbs sampler requires simulation from $\pi(x'_i|x_{S\setminus i})$, which may be quite unwieldy. Except when C consists of a small number of discrete colours, or when $\pi(\mathbf{x})$ is Gaussian, even normalisation of $\pi(x'_i|x_{S\setminus i})$ may be expensive. On the other hand, if some model-independent choice of proposal is made, we only need to be able to compute the ratio $\pi(x'_i|x_{S\setminus i})/\pi(x_i|x_{S\setminus i})$ of the posterior probabilities of the proposed and current values.

Thus if we know that the Gibbs sampler does yield good convergence properties, it may nevertheless be preferable in terms of computational cost to choose a proposal distribution that is merely reasonably close to $\pi(x'_i|x_{S\setminus i})$ from which it is easy to simulate, and to tolerate the consequent small probability of rejection.

The Gibbs sampler is not the only Metropolis method that gives zero rejection probability. In their study of stochastic relaxation in gaussian processes, Barone and Frigessi (1989) derived a class of samplers that include the Gibbs sampler as a special case. Suppose that μ_i and σ_i^2 are the expectation and variance of the conditional distribution $\pi(x'_i|x_{S\setminus i})$. The Gibbs sampler proceeds by drawing the new value x'_i from $N(\mu_i,\sigma_i^2)$. Barone and Frigessi's ω-stochastic relaxation (ω-*SR*) approach draws instead from $N((1+\theta)\mu_i-\theta x_i,(1-\theta^2)\sigma_i^2)$. (We use θ in place of their $\omega-1$). Validity of this method is most easily checked in the present context by noting that $q(\mathbf{x},\mathbf{x}')/q(\mathbf{x}',\mathbf{x})$ does not depend on θ. Barone and Frigessi prove that in the case of entirely positive association between the variables (all non-diagonal entries in the inverse of the variance matrix non-positive), the spectral radius R of the corresponding Markov chain is a decreasing function of θ at $\theta=0$. An intuitive explanation for this advantage of using $\theta>0$ in the case of positive association comes from noting that then the current value x_i is positively correlated with the values of its neighbours. If x_i is, say, in the lower tail of its marginal distribution under π, then the whole local conditional distribution $\pi(x'_i|x_{S\setminus i})$ will be biased towards this lower tail: hence the advantage in modifying the Gibbs sampler to improve convergence by "over-correcting" this bias.

A simpler yet stronger result holds for the asymptotic variance: for any linear function of $\mathbf{x}$, the asymptotic variance when using Barone and Frigessi's modified sampler, with systematic scanning of pixels, is proportional to $(1-\theta)/(1+\theta)$. Without

loss of generality, we assume that the process has zero expectation and is in equilibrium.

Theorem.
Suppose that $\pi(\mathbf{x})$ is the gaussian distribution $N(0,V)$ where V is non-singular, and that the pixels are indexed by $i \in S=\{1,2,\ldots,n\}$. Let a stationary process $\{\mathbf{x}^{(t)}, t \in \mathbf{Z}\}$ with marginal distribution $\pi(\mathbf{x})$ be defined by updating x_i cyclically for $i=1,2,\ldots,n,1,2,\ldots$ by resampling x_i from

$$N((1+\theta)\mu_i - \theta x_i, (1-\theta^2)\sigma_i^2) \tag{11}$$

where μ_i and σ_i^2 are the mean and variance of the distribution $\pi(x_i|x_{S\setminus i})$. Then for any vector of constants $\mathbf{c}$,

$$N\,var(\mathbf{c}^{\mathrm{T}} \frac{1}{N} \sum_{t=1}^{N} \mathbf{x}^{(t)}) \to \frac{1-\theta}{1+\theta} \mathbf{c}^{\mathrm{T}} V \mathrm{diag}(V^{-1}) V \mathbf{c}.$$

Proof. We first consider the stationary first-order matrix autoregression defined by $\mathbf{x}^{(t+1)} = A\mathbf{x}^{(t)} + \mathbf{z}^{(t+1)}$, where $\{\mathbf{z}^{(t)} : t \in \mathbf{Z}\}$ are independent and identically distributed gaussian random vectors with zero mean. (Since the process is stationary, and is to have marginal distribution π, it follows that we must have $var(\mathbf{z}^{(t)}) = V - AVA^{\mathrm{T}}$). Now for any $t \geq 0$, $E(\mathbf{x}^{(t)}\mathbf{x}^{(0)\mathrm{T}}) = E((A^t\mathbf{x}^{(0)} + \sum_{r=1}^{t} A^{t-r}\mathbf{z}^{(r)})\mathbf{x}^{(0)\mathrm{T}}) = A^t E(\mathbf{x}^{(0)}\mathbf{x}^{(0)\mathrm{T}}) + 0 = A^t V$. Thus

$$\begin{aligned} \sum_{t=-\infty}^{\infty} E(\mathbf{x}^{(t)}\mathbf{x}^{(0)\mathrm{T}}) &= \sum_{t=0}^{\infty} A^t V + \sum_{t=0}^{\infty} (A^t V)^{\mathrm{T}} - V \\ &= (I-A)^{-1} V + V(I-A)^{-1} - V \end{aligned}$$

Now

$$\begin{aligned} var(\mathbf{c}^{\mathrm{T}} \frac{1}{N} \sum_{t=1}^{N} \mathbf{x}^{(t)}) &= \frac{1}{N^2} \mathbf{c}^{\mathrm{T}} \left[\sum_{s=1}^{N} \sum_{t=1}^{N} E(\mathbf{x}^{(s)}\mathbf{x}^{(t)\mathrm{T}}) \right] \mathbf{c} \\ &\sim \frac{1}{N} \mathbf{c}^{\mathrm{T}} [(I-A)^{-1} V + V(I-A)^{-1} - V] \mathbf{c} \\ &= N^{-1} \mathbf{c}^{\mathrm{T}} (I-A)^{-1} (I+A) V \mathbf{c}. \end{aligned}$$

We now have to write Barone and Frigessi's sampler in the matrix autoregressive form. But

$$\begin{aligned} \mu_i &= E(x_i|x_{S\setminus i}) = -g_{ii}^{-1} \sum_{j \neq i} g_{ij} x_j, \\ \sigma_i^2 &= var(x_i|x_{S\setminus i}) = g_{ii}^{-1}, \end{aligned}$$

where $G = (g_{ij}) = V^{-1}$. Thus the sampler (11) can be written

$$x_i^{(t+1)} = \sum_{j=1}^{i-1} b_{ij} x_j^{(t+1)} + \sum_{j=i}^{n} b_{ij} x_j^{(t)} + z_i^{(t+1)},$$

where $B = (b_{ij}) = I - (1+\theta)\Gamma G$ and $var(z_i^{(t+1)}) = (1-\theta^2)\Gamma_{ii}$ where $\Gamma = (\mathrm{diag}(G))^{-1}$. Let L denote the lower triangle of B. Then in matrix form we have

$$\mathbf{x}^{(t+1)} = L\mathbf{x}^{(t+1)} + (B-L)\mathbf{x}^{(t)} + \mathbf{z}^{(t+1)}$$

or

$$\mathbf{x}^{(t+1)} = (I-L)^{-1}(B-L)\mathbf{x}^{(t)} + (I-L)^{-1}\mathbf{z}^{(t+1)}.$$

This is a matrix autoregression with $A = (I-L)^{-1}(B-L)$. Thus

$$\begin{aligned}(I-A)^{-1}(I+A)V &= (I-B)^{-1}(I+B-2L)V \\ &= (1+\theta)^{-1}G^{-1}\Gamma^{-1}\{2I-(1+\theta)\Gamma G+2(1+\theta)\Gamma H\}V\end{aligned}$$

where H is the lower triangle of G

$$\begin{aligned}&= (1+\theta)^{-1}V\{2\Gamma^{-1}-(1+\theta)G+2(1+\theta)H\}V \\ &= (1+\theta)^{-1}V\{(1-\theta)\Gamma^{-1}+(1+\theta)(H-H^{\mathrm{T}})\}V\end{aligned}$$

since G is symmetric. On pre- and post-multiplying by the same vector $\mathbf{c}$, the anti-symmetric term vanishes, and we obtain the required result.

The implication of this result is that, for linear functionals in the gaussian case, and considering only the asymptotic variance, best performance in this class of procedures is obtained by letting $\theta \to +1$. This is a dynamic analogue of the conventional idea of using antithetic variables to reduce Monte Carlo variance. It is interesting to note that this effect is anticipated, without explanation, in a simple example in Hastings (1970, p.101).

All of this applies only to gaussian distributions $\pi(\mathbf{x})$, and our real interest is in other cases with continuously distributed $\mathbf{x}$. Only in rather special cases could we expect to find a family of samplers analogous to that of Barone and Frigessi, indexed by an "antithetic parameter" θ and including the Gibbs sampler, yet convenient for simulation. As a general procedure, however, we suggest using a gaussian proposal of the form

$$x_i' \sim N((1+\theta)\mu-\theta x_i, (1-\theta^2)\sigma^2) \tag{12}$$

in the Metropolis/Hastings algorithm, with appropriately chosen μ, σ^2 and θ (these can depend on all variables in the model except x_i). The acceptance probability is

$$\alpha(\mathbf{x},\mathbf{x}') = \min\{1, \frac{\pi(\mathbf{x}')q(\mathbf{x}',\mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x},\mathbf{x}')}\}$$

which simplifies to

$$= \exp[\min\{0, g(x_i')-g(x_i)\}],$$

where

$$g(x_i') = \log \pi(x_i'|x_{S\setminus i}) + \frac{1}{2\sigma^2}(x_i'-\mu)^2. \tag{13}$$

Note that α does not depend on the antithetic parameter θ. We can now choose μ and σ, depending on $x_{S\setminus i}$, to ensure that $g(x_i')$ is approximately constant over a range of arguments including x_i and the most probable values from the proposal distribution, so that $\alpha(\mathbf{x},\mathbf{x}')$ is close to 1 with high probability. Such a choice of μ and σ can be made by expanding $g(x_i')$ to second order about an appropriate approximate centre: for example, we have used the mean of neighbouring x_j when simulating from Gibbs distributions.

The nub of the idea is to use a gaussian approximation to the Gibbs sampler, but

- this need only be a good approximation in the centre of the conditional distribution,
- exact detailed balance is restored by the acceptance/rejection decisions, and
- we still have the parameter θ free to help improve asymptotic variance.

A full analysis of the spectrum of such a Markov chain seems to be a challenging problem, but intuitively one might be concerned that as θ increases towards 1, the spectral radius may approach or even attain the value 1. There may therefore be less freedom of choice in general than in Barone and Frigessi's gaussian case. This underlines the need to monitor convergence carefully as the simulation proceeds.

We are not aware of any classes of distributions for use in generating proposals, other than the gaussian family (12), into which it is possible to introduce an antithetic parameter θ that cancels on forming the ratio $q(\mathbf{x},\mathbf{x}')/q(\mathbf{x}',\mathbf{x})$. Thus if our procedure were modified to use a non-gaussian proposal distribution, the details would be a little more complicated. The simplest way to extend the idea to state spaces other than $\Omega = \mathbf{R}^S$ would be by transformation, for example, by replacing x throughout by $\log(x)$ if x takes only non-negative values.

Further insight into these sampling methods can be gained by considering a toy example. Suppose we have just two sites, and three possible values $\{1,2,3\}$ at each. There are then only nine possible states of the system, $\{11,12,\ldots,33\}$, and spectral decompositions are easily computed numerically. Updating is by uniform random choice of site, and when visiting site 1, the probabilities of some possible transitions are

$12 \to 12$	α
$12 \to 22$	$1-\alpha-\gamma$
$12 \to 32$	γ
$11 \to 21$	δ
$11 \to 11$	$1-2\delta$

All other probabilities are determined from these by symmetry over permutations of sites and values. It is easily verified that the unique equilibrium distribution of the chain has $\pi(11) = \kappa/(6+3\kappa) = \pi(22) = \pi(33)$, and $\pi(12) = 1/(6+3\kappa) = \pi(13) = \ldots$etc., where $\kappa = (1-\alpha-\gamma)/\delta$. This is a symmetric Potts model on 3 colours (Potts (1952)). For any fixed value of $\kappa > 1$, the simple Metropolis method, in which the proposal is an equally likely choice among the colours different from the current one, is the case $\alpha = 0, \gamma = 0.5$. The Gibbs sampler is the case $\alpha = \gamma = \delta$, where the new value is (conditionally) independent of the old. A crude analogue of the sampler (12) for $\theta > 0$ is obtained by reducing α and γ, and increasing δ accordingly to preserve the equilibrium, thus increasing the (equilibrium) probability of change at a transition, which is proportional to $(1-\alpha-\gamma/2)$.

Table 1 displays some values of $\tau(f)$ and R for selected parameter values, including the simple Metropolis method, the Gibbs sampler, those achieving minimum $\tau(f)$ or R, and for contrast, an extreme case with very poor performance. The values for $\tau(f)$ apply to any function f of x_1 alone: the invariance follows from the symmetry in this example. These figures confirm that minimum $\tau(f)$ and R are not the same thing,

that Gibbs sampling achieves neither, and that Metropolis methods *can* be very poor.

Table 1. $\tau(f)$ and R in the toy example.

$\kappa = 1.5$

	α	γ	$\tau(f)$	R
min $\tau(f)$	0	0.25	1.7701	0.7500
min R	0	0.4	1.9762	0.4000
simple	0	0.5	2.2063	0.5000
Gibbs	0.2857	0.2857	3.1667	0.5714
poor	0.99	0	265.67	0.9943

$\kappa = 3$

	α	γ	$\tau(f)$	R
min $\tau(f)$	0	0	3.0667	0.6667
min R	0	0.0659	3.1111	0.6052
simple	0	0.5	4.5111	0.7101
Gibbs	0.2	0.2	4.5238	0.7000
poor	0.99	0	405.67	0.9961

There is an analysis of the spectral radius R for various samplers in the finite state space case in Frigessi, Hwang, Sheu, and di Stefano (1990), including some numerical comparisons for the Ising model.

4. Experiments with the new sampler

In this last section, we present a few of the results from some fairly extensive experimentation with the Metropolis algorithm with the gaussian proposal distribution suggested in the previous section. Kirkland (1989) performed a thorough study of a number of samplers for the case of binary Markov random fields; here, of course, we are considering only the continuous case.

The context is of an idealised image analysis problem based on artificial data. In all of the experiments to be described, both the true and observed images, $\mathbf{x}$ and $\mathbf{y}$ respectively, consist of 64×64 pixels. The model to be assumed in the analysis for $\mathbf{x}$ is gaussian:

$$p(\mathbf{x}) \propto \exp\{-\beta \sum_{[i,j]} (x_i - x_j)^2\} \tag{14}$$

where the sum is over orthogonal neighbours only. Each pixel has four neighbours, except for those on the boundary of the array which have three or two. The true $\mathbf{x}$ images from which our artificial data are generated are drawn from the same model except that (a) a possibly different parameter value β_0 is used, and (b) the overall average x value is adjusted to the level 25 by adding a constant to all x_i (under (14), the average has an improper distribution).

Two different models for $p(\mathbf{y}|\mathbf{x})$ will be used, in each case both for simulating and analysing the data. Under each of the models, the $\{y_i\}$ are conditionally

independent, given $\mathbf{x}$, and we have respectively:

$$y_i \sim N(x_i, 25) \qquad \text{(gaussian)}$$

$$y_i \sim Poisson(25\exp(x_i/25-1)) \qquad \text{(Poisson)}$$

Note that these models have been devised so that they are comparable in terms of mean and variance; the second allows us to study Monte Carlo methods in the presence of Poisson variation with mean of similar order to that found in much of the medical imagery we see.

For the gaussian model, we use Barone and Frigessi's sampler; this is straightforward. A little work is needed to set up the corresponding Metropolis algorithm for the Poisson case, however. The function g defined in (13) is given by

$$\begin{aligned}
g(x_i') &= \log p(y_i|x_i') + \log p(x_i'|x_{S\setminus i}) + \frac{1}{2\sigma^2}(x_i'-\mu)^2 + \text{constant} \\
&= y_i \log(me^{x_i'/m-1}) - me^{x_i'/m-1} - \beta\nu_i(x_i'-\bar{x}_i)^2 \\
&\quad + \frac{1}{2\sigma^2}(x_i'-\mu)^2 + \text{constant} \\
&= \frac{y_i x_i'}{m} - me^{x_i'/m-1} - \beta\nu_i(x_i'-\bar{x}_i)^2 \qquad (15) \\
&\quad + \frac{1}{2\sigma^2}(x_i'-\mu)^2 + \text{constant} \\
&= x_i'\{\frac{y_i}{m} - me^{\bar{x}_i/m-1}(\frac{1}{m} - \frac{\bar{x}_i}{m^2}) + 2\beta\nu_i\bar{x}_i - \frac{\mu}{\sigma^2}\} \\
&\quad + x_i'^2\{-\frac{me^{\bar{x}_i/m-1}}{2m^2} - \beta\nu_i + \frac{1}{2\sigma^2}\} \\
&\quad + O((x_i'-\bar{x}_i)^3) + \text{constant}
\end{aligned}$$

where $m = 25$ is the overall level assumed in the model, and ν_i and $\bar{x}_i$ are the number of neighbouring x values and their mean respectively. Thus if we choose

$$\mu = \bar{x}_i + \frac{\sigma^2}{m}(y_i - me^{\bar{x}_i/m-1})$$

and

$$\sigma^2 = \{2\beta\nu_i + \frac{e^{\bar{x}_i/m-1}}{m}\}^{-1}$$

the first and second order terms vanish, and there is a prospect that g will be nearly constant in the range of interest. These are the values used in the experiments we report. There may be merit in examining alternative quadratic approximations to the exponential function in (15) in the hope of obtaining values for μ and σ^2 that give higher average acceptance probability by making g closer to constant over a wider range, but we do not pursue that here.

Our experiments consider three different functionals f, chosen to reflect different aspects of the distribution π, but in no sense claimed to be thoroughly exploring the eigenspace of P. The functionals are

- Mean: the overall mean
- 8-Co: the lag-(8,0) spatial autocorrelation
- PL: the statistic that would give the maximum pseudo-likelihood estimate of β for directly observed $\mathbf{x}$ from the model (14), namely $N/(2\sum v_i(x_i-\bar{x}_i)^2)$.

We first present, in Table 2, estimates of the integrated autocorrelation time, for all three functionals, for three values of β, for both the gaussian and Poisson cases, and for three independent replicates of each. Four different samplers are compared: three of these are the $\theta=0.5, \theta=0$ and $\theta=-0.5$ versions of our proposed method. The other is a simple Metropolis method using a proposal drawn from a gaussian distribution centred at the current value, and with standard deviation 3: thus $x_i' \sim N(x_i, 3^2)$. This corresponds formally to (12), in the limit as $\theta \to -1$ and $\sigma^2 \to \infty$ while $(1-\theta^2)\sigma^2 \to 9$. We use Sokal's estimator $\hat{\tau}(f)$ (see section 2). Each estimate is based on the last 4096 sweeps of a run of 5000, starting from $\mathbf{x}=\mathbf{y}$. Our experience has been that Sokal's estimator is somewhat more stable than the between-blocks mean square (8) with $N=bk=4000$ and $k=50$ or 100; most of the exceptions to this pattern being with the simple Metropolis sampler.

Table 2. Estimates of integrated autocorrelation time.

(a) Gaussian

		Mean			8-Co			PL		
$\beta = 0.001$	θ=0.5	0.09	0.15	0.10	0.14	0.15	0.16	0.14	0.16	0.15
	θ=0.0	1.15	1.21	1.12	1.12	1.16	1.18	0.92	0.98	0.98
	θ=−0.5	3.42	3.31	3.25	4.29	3.28	3.24	2.50	2.36	2.62
	simple	14.96	17.14	14.87	22.47	11.34	14.39	12.14	11.09	10.66
$\beta = 0.01$	θ=0.5	0.97	0.97	0.96	0.94	0.90	0.88	1.48	1.38	1.47
	θ=0.0	2.72	3.25	2.89	2.39	2.51	3.03	0.98	0.98	1.04
	θ=−0.5	9.39	6.53	6.63	7.41	6.66	6.85	1.79	1.73	1.65
	simple	19.75	17.60	20.05	19.88	16.61	19.64	5.33	6.08	6.15
$\beta = 0.1$	θ=0.5	6.55	6.65	8.52	5.39	4.58	6.13	2.21	1.94	2.09
	θ=0.0	21.24	14.91	18.53	14.11	10.47	16.03	1.02	1.08	1.20
	θ=−0.5	55.59	56.84	108.8	28.03	26.65	27.51	1.59	1.37	1.59
	simple	93.77	104.0	68.84	58.62	62.36	70.64	4.41	3.85	3.77

(b) Poisson

		Mean			8-Co			PL		
$\beta = 0.001$	θ=0.5	1.29	1.22	1.12	1.10	1.03	1.11	3.52	3.31	3.16
($\beta_0 = 0.1$)	θ=0.0	2.16	1.84	1.84	1.75	1.49	1.40	2.58	3.61	2.64
	θ=−0.5	5.48	5.25	4.73	3.48	3.52	3.36	5.29	5.65	5.55
	simple	22.00	20.37	13.17	12.66	14.51	11.43	9.83	13.80	12.62
$\beta = 0.01$	θ=0.5	1.18	1.25	1.34	1.14	1.17	1.18	1.44	1.45	1.53
	θ=0.0	3.41	3.50	3.34	3.23	2.98	3.05	1.13	1.13	1.13
	θ=−0.5	9.14	10.87	10.23	8.58	10.15	8.10	2.08	2.27	2.26
	simple	18.28	27.69	21.28	11.72	22.75	19.88	3.73	5.63	4.86
$\beta = 0.1$	θ=0.5	8.33	6.55	7.85	4.41	5.75	5.75	2.45	2.13	2.15
	θ=0.0	19.75	21.61	19.64	12.28	12.61	14.98	1.24	1.15	1.27
	θ=−0.5	43.63	81.21	80.97	49.27	34.53	70.60	1.58	1.52	1.62
	simple	50.99	56.13	40.62	46.57	46.55	81.15	3.95	4.43	3.63

In earlier experiments, we found that in the case of low interaction parameter, $\beta = 0.001$, the consequent wide range in values in the generated true $\mathbf{x}$ led to very unstable results. Such wide variation in $\mathbf{x}$ does not occur in most real image analysis problems, and so our studies in this case have used simulations using $\beta_0 = 0.1$ instead.

It is clear from Table 2 that the different samplers have very different behaviour as measured by autocorrelation time. For two of the functionals, the $\theta = 0.5$ sampler is always the best, often giving asymptotic variance as small as would arise from independent random sampling. This is remarkably good performance, and very encouraging. It confirms the heuristic interpretation given earlier of the antithetic properties of the sampler when $\theta > 0$. In contrast, the simple Metropolis method performs very badly, suggesting in some cases that a run 100 times as long as for independent sampling is needed to give the same asymptotic variance. The pattern for the third functional, namely the pseudo-likelihood statistic, is somewhat different: in most of the cases considered, the best performance is obtained with $\theta = 0$, corresponding exactly or approximately to the Gibbs sampler. Of the three, this functional depends most directly on local conditional distributions, so it is intuitively reasonable that resampling directly from these distributions should be close to optimal.

Other features of the Table are that there is apparently little difference in performance in the gaussian and Poisson cases, and that, as would be expected, the autocorrelation time increases with β.

It is of interest to compare these numerical estimates with the conclusions of the Theorem in the previous section. The only case to which the Theorem applies exactly is that of the mean functional in the gaussian case, for which it is apparent from Table 2 that the estimated autocorrelation times are indeed approximately proportional to $(1 - \theta)/(1 + \theta)$. The Table also suggests that the conclusions of the Theorem hold more widely, to a rough approximation.

In Table 3, we present another property of the same four samplers: the empirical acceptance rates, expressed as percentages, and computed only after equilibrium is reached. For the Barone-Frigessi samplers in the gaussian case, of course there is

100% acceptance, but we see that the rate is about 90% or better even in the Poisson case with $\beta = .001$ (the situation among those considered where the quadratic approximation to the exponential function is (15) is least adequate).

Table 3. Empirical Metropolis acceptance rates in equilibrium.

(a) Gaussian

	$\beta = 0.001$	$\beta = 0.01$	$\beta = 0.1$
θ=0.5	100%	100%	100%
θ=0.0	100%	100%	100%
θ=−0.5	100%	100%	100%
simple	79.82%	69.64%	40.38%

(b) Poisson

	$\beta = 0.001$ ($\beta_0 = 0.1$)	$\beta = 0.01$	β =0.1
θ=0.5	89.73%	98.18%	99.93%
θ=0.0	91.27%	98.45%	99.94%
θ=−0.5	93.58%	98.83%	99.95%
simple	79.89%	69.50%	40.35%

Before coming to a general conclusion that these Metropolis methods all perform well in the circumstances of this example, we should seek some reassurance that the Markov chains we are simulating do actually converge in a reasonable number of steps.

In Figure 1, we display the values of our three functionals, for a single realisation of the chain, plotted against "time" measured in units of complete sweeps through the image; this Figure is for the gaussian model, with $\beta = 0.01$. In order to make the initial transient more visible, we have deliberately chosen a poor starting value for the run, namely $x_i \sim U(0,10)$. Each panel of the Figure displays four trajectories, one for each of the four sampling procedures represented in Tables 2 and 3. It is evident that in this case, the three Barone-Frigessi methods all converge quickly (as judged by these functionals): equilibrium is effectively reached by time 25. The simple Metropolis method takes somewhat longer, until approximately time 70. Of course, there are visible differences in character between the three sets of trajectories, reflecting the differing equilibrium variances of the three functionals.

Figure 2 reveals a dramatically different picture for the Poisson model: all other details are the same as for Figure 1. With a poor starting value, the $\theta = +0.5$ sampler converges extremely slowly, and has not reached equilibrium even by time 1000. For the PL functional, this is also true for $\theta = 0.0$. However, the remaining two samplers apparently converge by time 80. The unacceptably slow convergence when $\theta \geq 0.0$ is apparently due to a very low average acceptance probability when the process is far from equilibrium: recall that the values presented in Table 3 applied to equilibrium only.

Recommendation of choice of sampler using the criteria of convergence speed is therefore in stark contrast to that suggested by our discussion of the autocorrelation

times in Table 2. We feel these observations strongly support the strategy, mentioned in Section 2, of switching from one type of sampler to another, as convergence is achieved. Indeed, one might envisage working entirely with our gaussian proposals, and adaptively varying the value of θ according to the behaviour of the sample trajectories of functionals such as those in Figure 2. We have not investigated this suggestion, but it seems a promising line for future enquiry.

In Figure 3, we display similar information for the Poisson case, with the prior interaction parameter β increased to 0.1. As we observed earlier, increased β means increased $\tau(f)$, but we see that the performance in terms of converge speed is improved, presumably because the quadratic approximation to (15) is better for larger β.

Finally, we turn in Figure 4 to the case of low prior interaction, $\beta = 0.001$, in the Poisson case. As with Tables 2 and 3, we simulated the true $\mathbf{x}$ using $\beta_0 = 0.1$. The results are consistent with the pattern observed in the other examples, but perhaps convey a suggestion that the poor performance in Figure 2 is partly due to the wide intensity range in $\mathbf{x}$ in consequence of simulating from (14) with $\beta = 0.01$.

In conclusion, we underline the points made in Section 2. Successful use of dynamic Monte Carlo methods demands attention to two different aspects of the constructed Markov chains: speed of weak convergence and integrated autocorrelation time. Both are determined by the spectral structure of the transition mechanism of the chain, but are computationally inaccessible except in trivial examples. However, careful monitoring of sample paths of the process can give useful information. Without such monitoring, conclusions of Monte Carlo calculations on stochastic systems can be very misleading.

The performance of Metropolis methods using the gaussian proposal (12) can be very good, providing that care is taken in selecting the antithetic parameter θ.

Acknowledgements

We are grateful for the insightful comments of Julian Besag, Arnoldo Frigessi and Brian Ripley in discussing this work, and acknowledge the financial support of the Complex Stochastic Systems initiative of the Science and Engineering Research Council.

References

Barone, P. and Frigessi, A. (1989) Improving stochastic relaxation for gaussian random fields. Probability in the Engineering and Informational sciences, 4, 369-389.

Besag, J. and Clifford, P. (1989) Generalized Monte Carlo significance tests. Biometrika, 76, 633-642.

Besag, J., York, J. C., and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math., 43, 1-59.

Frigessi, A., Hwang, C-R., Sheu, S-J. and di Stefano, P. (1990) Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. IAC quaderno 6/90, Rome.

Frigessi, A., Hwang, C-R. and Younes, L. (1990) Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. Submitted to Ann. Appl. Probab.

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

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell., 12, 609-628.

Grenander, U. (1983) Tutorial in pattern theory, Brown University, Division of Applied Mathematics, Providence, RI.

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

Kirkland, M. (1989) Simulating Markov random fields. Ph. D. thesis, University of Strathclyde.

Kirkpatrick, S., Gellatt, C. D. and Vecchi, M. P. (1983) Optimization by simulated annealing. Science, 220, 671-680.

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

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

Potts, R. B. (1952) Some generalised order-disorder transformations. Proc. Camb. Phil. Soc., 48, 106-109.

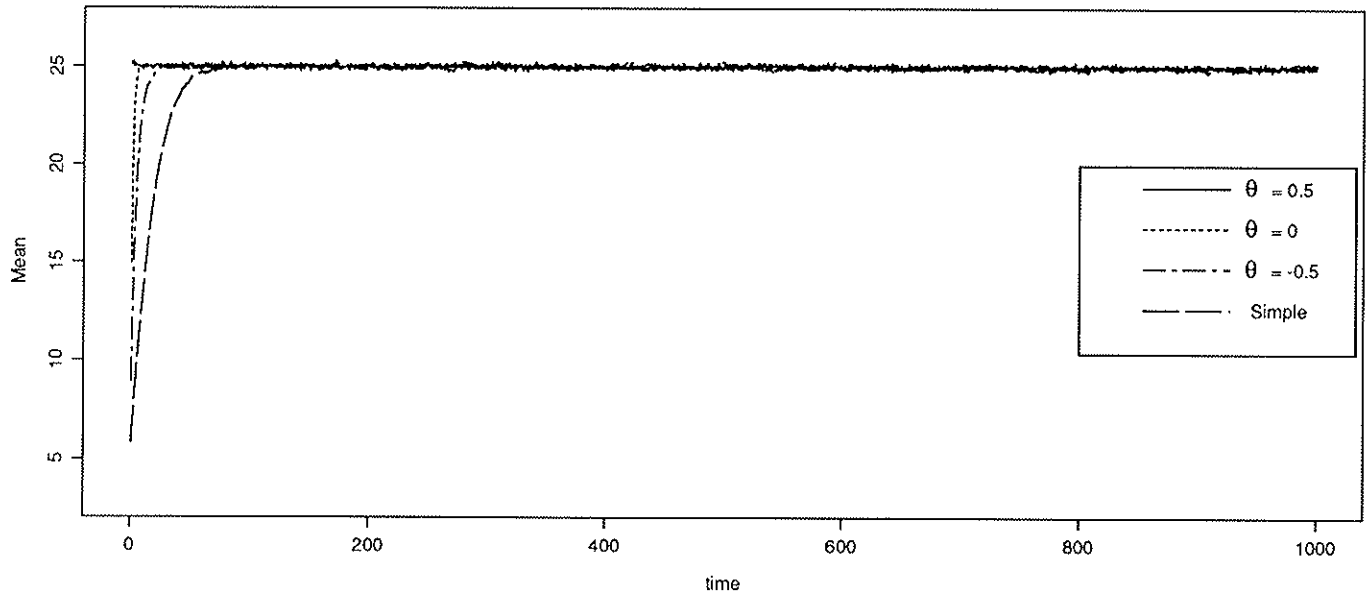
Priestley, M. (1981) Spectral analysis and time series. Academic Press, London.

Ripley, B. D. (1987) Stochastic simulation. Wiley, New York.

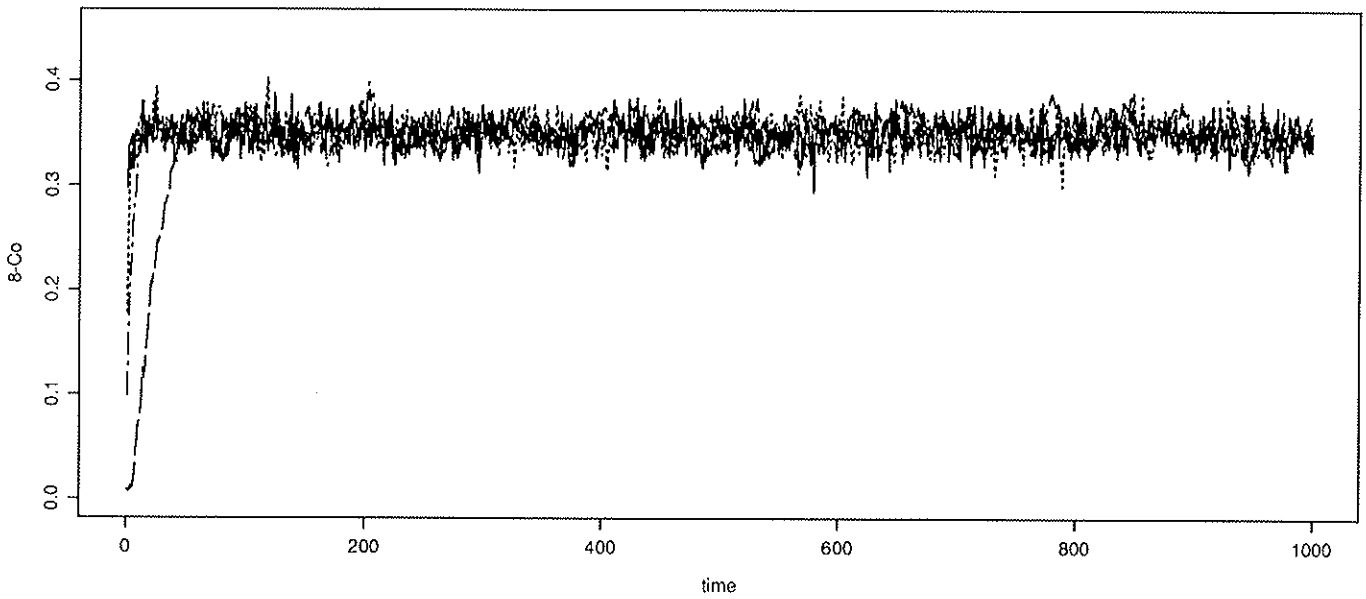
Sokal, A. D. (1989) Monte Carlo methods in statistical mechanics: foundations and new algorithms. Troisième cycle de la Physique en Suisse Romande lecture notes.

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

Figure 1. Gaussian Case (β =0.01) (a) Mean



(b) 8-Co



(c) PL

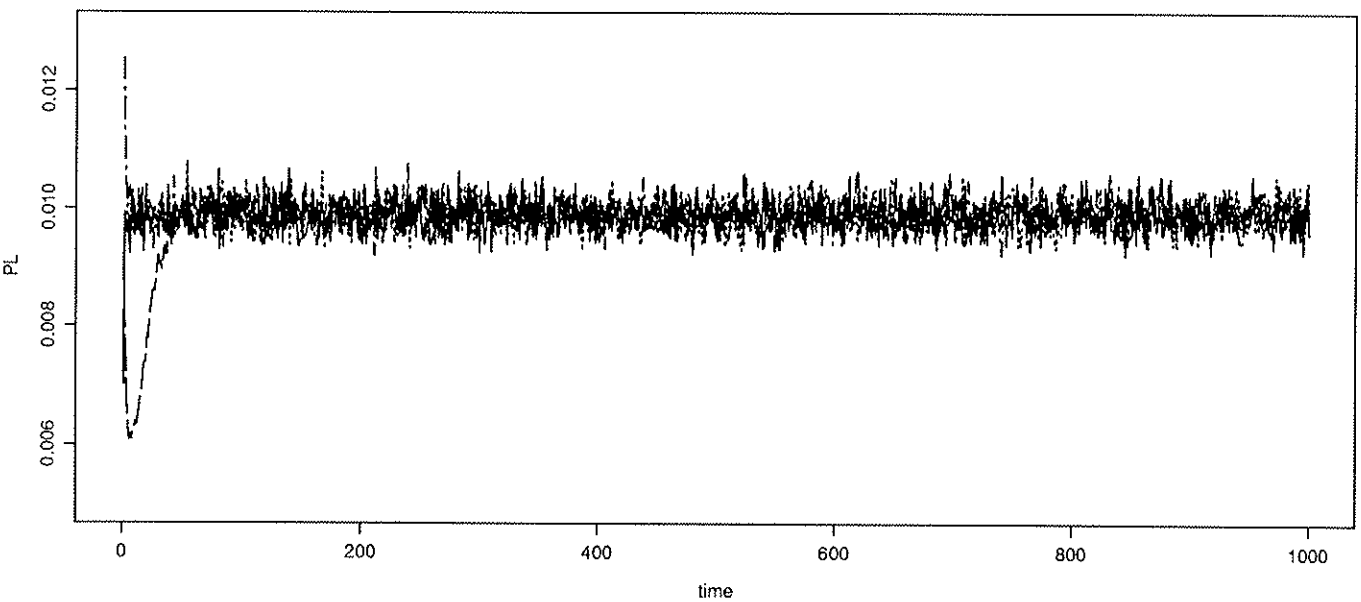
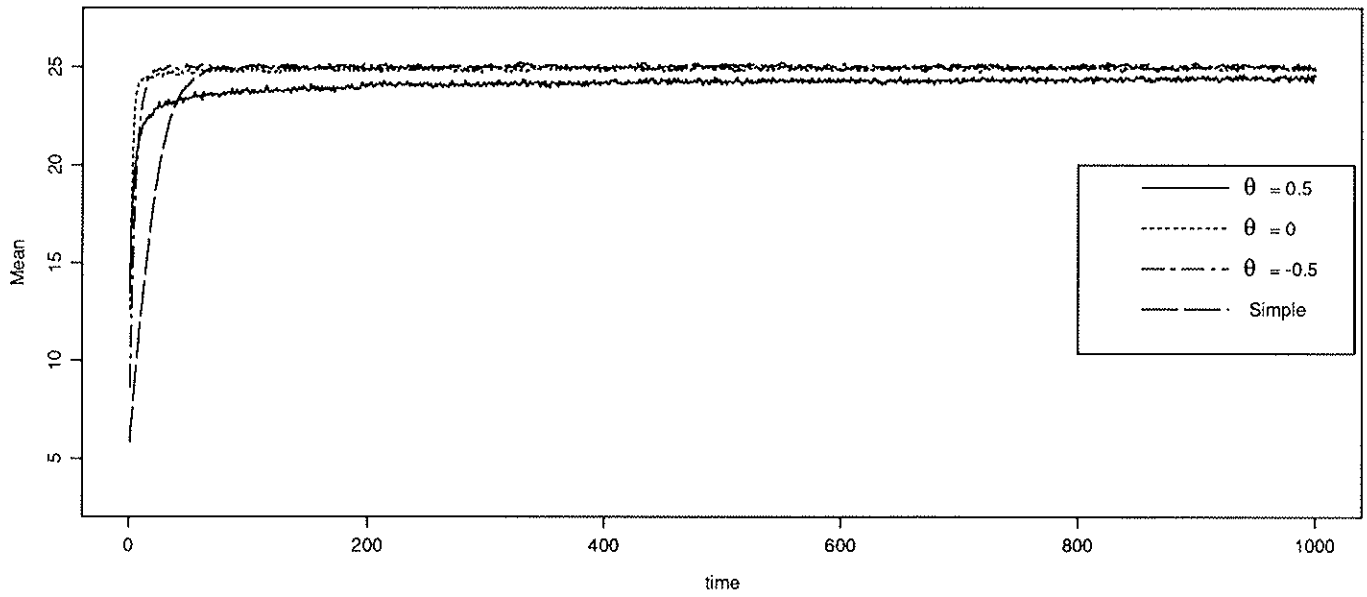
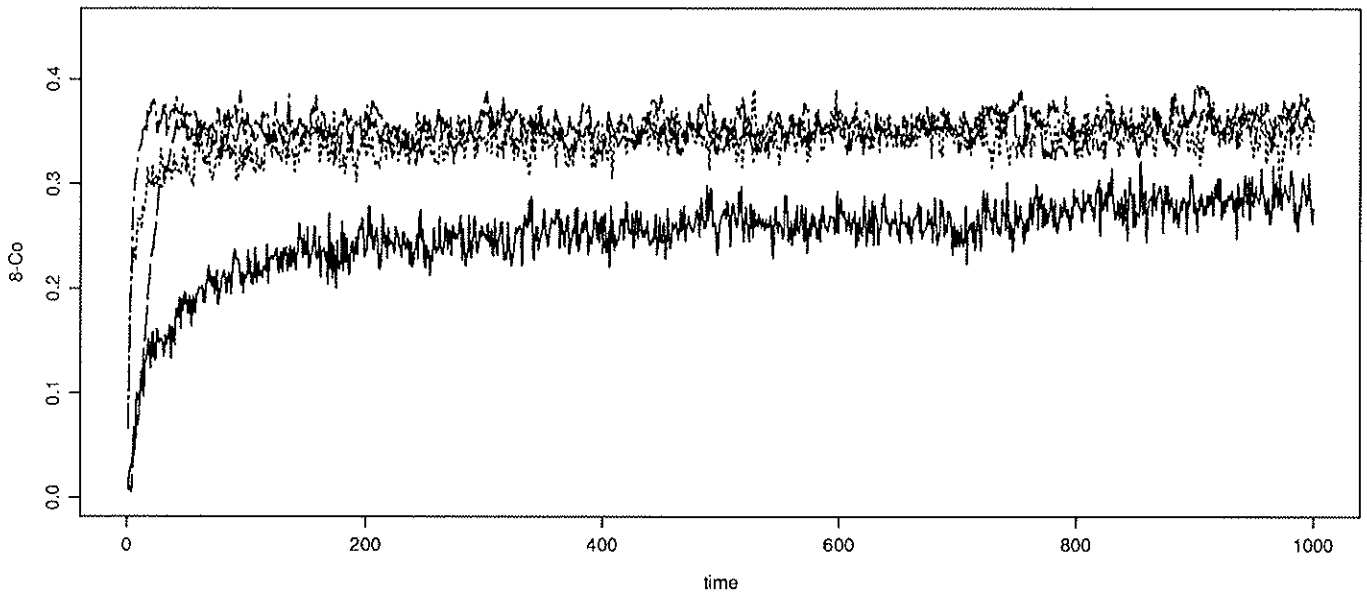


Figure 2. Poisson Case (β=0.01) (a) Mean



(b) 8-Co



(c) PL

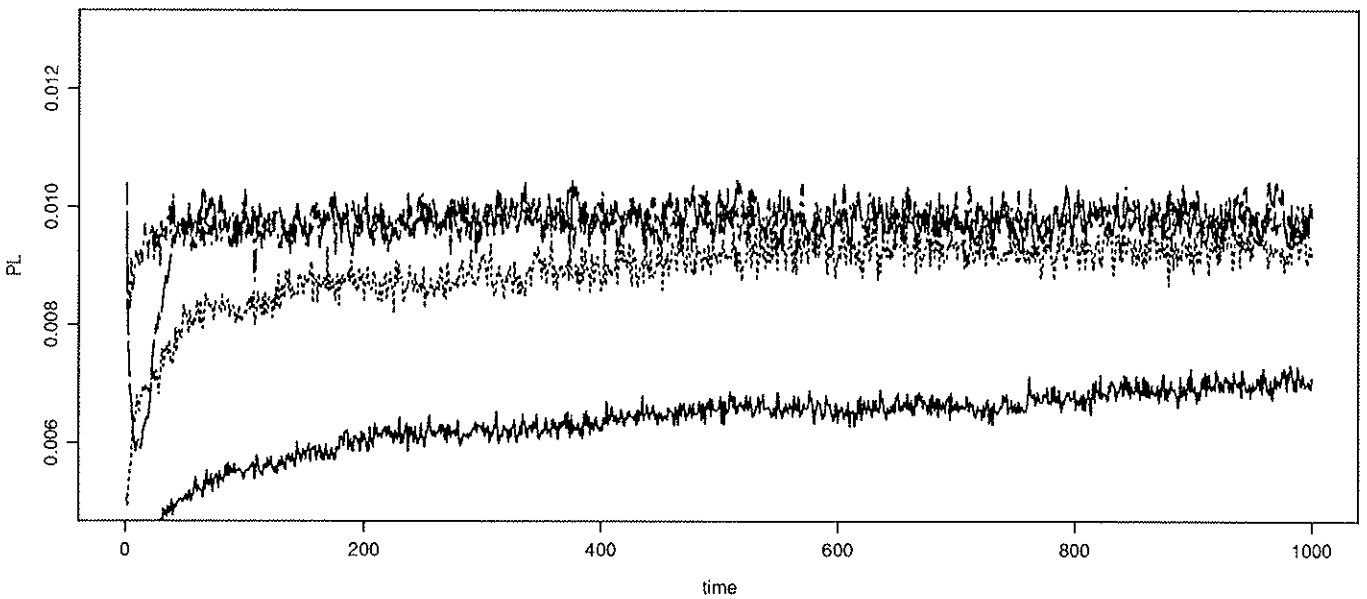
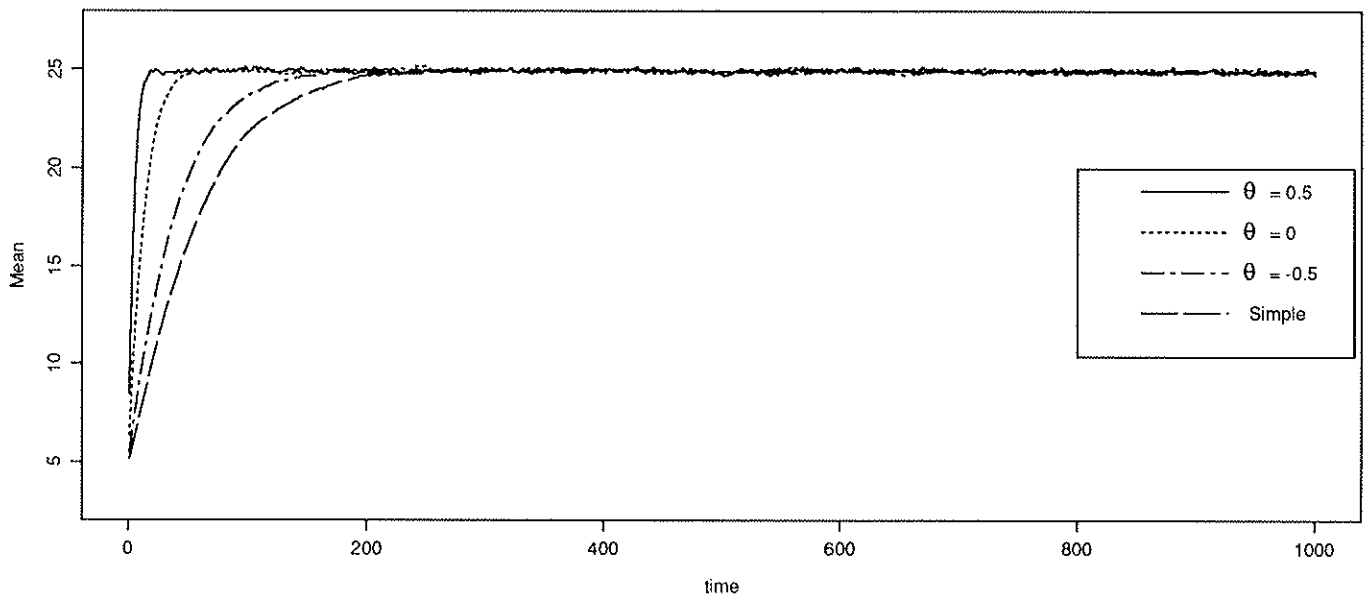
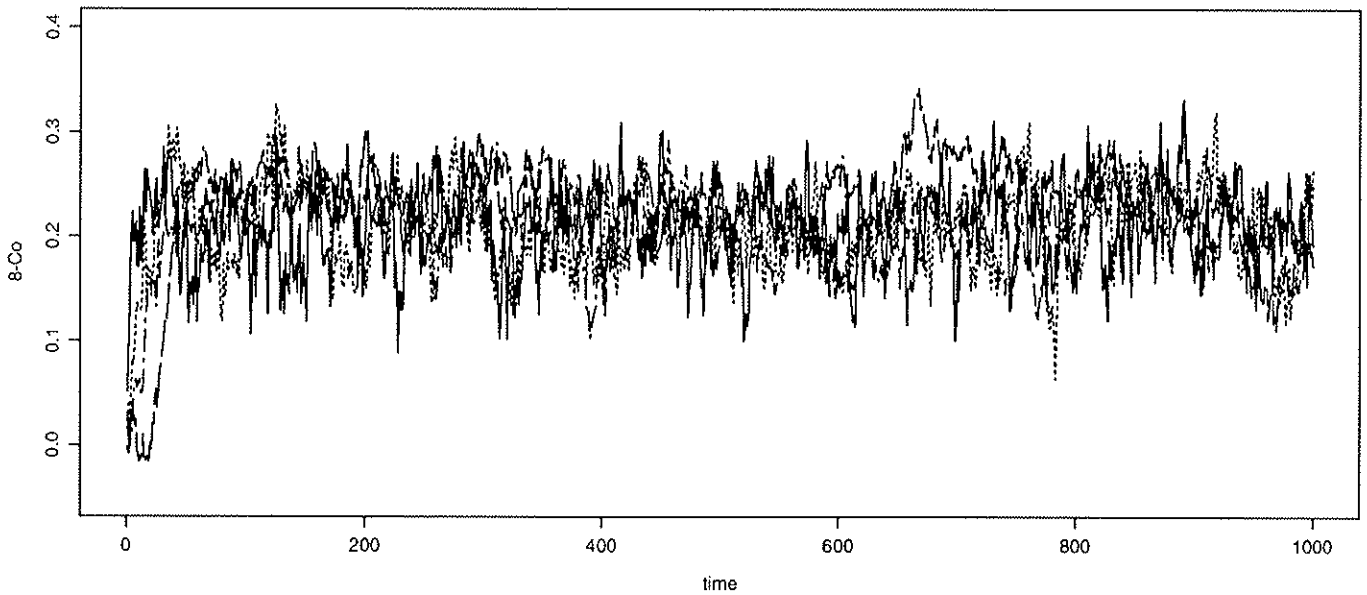


Figure 3. Poisson Case (β =0.1) (a) Mean



(b) 8-Co



(c) PL

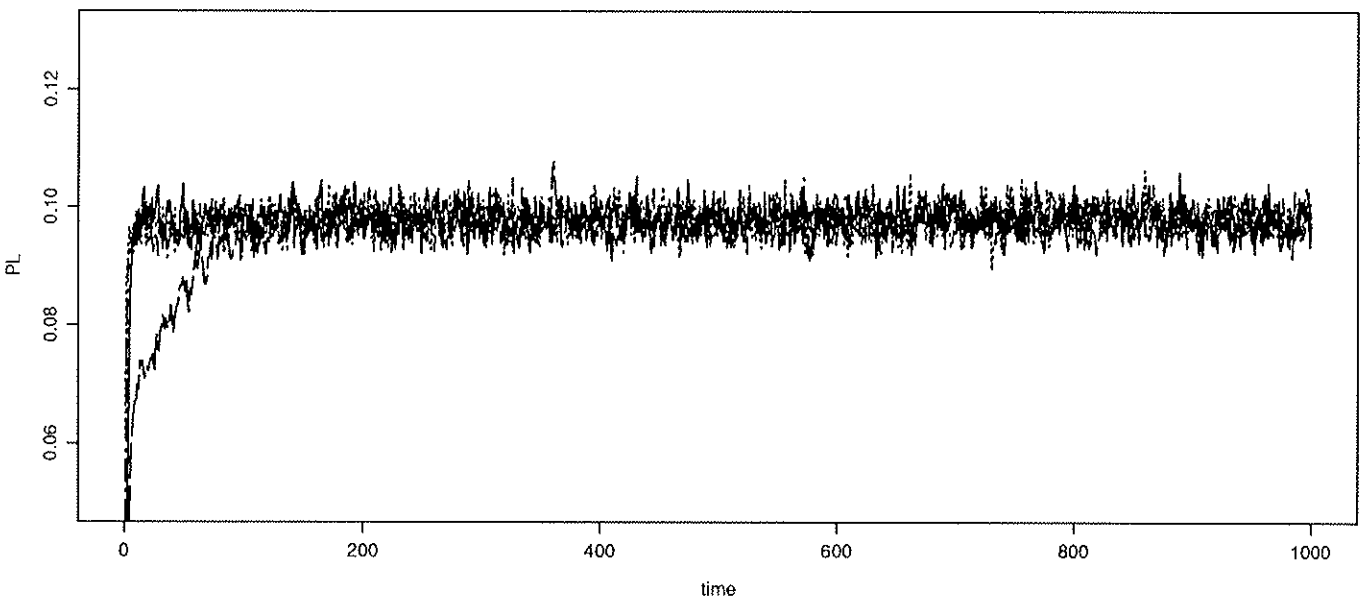
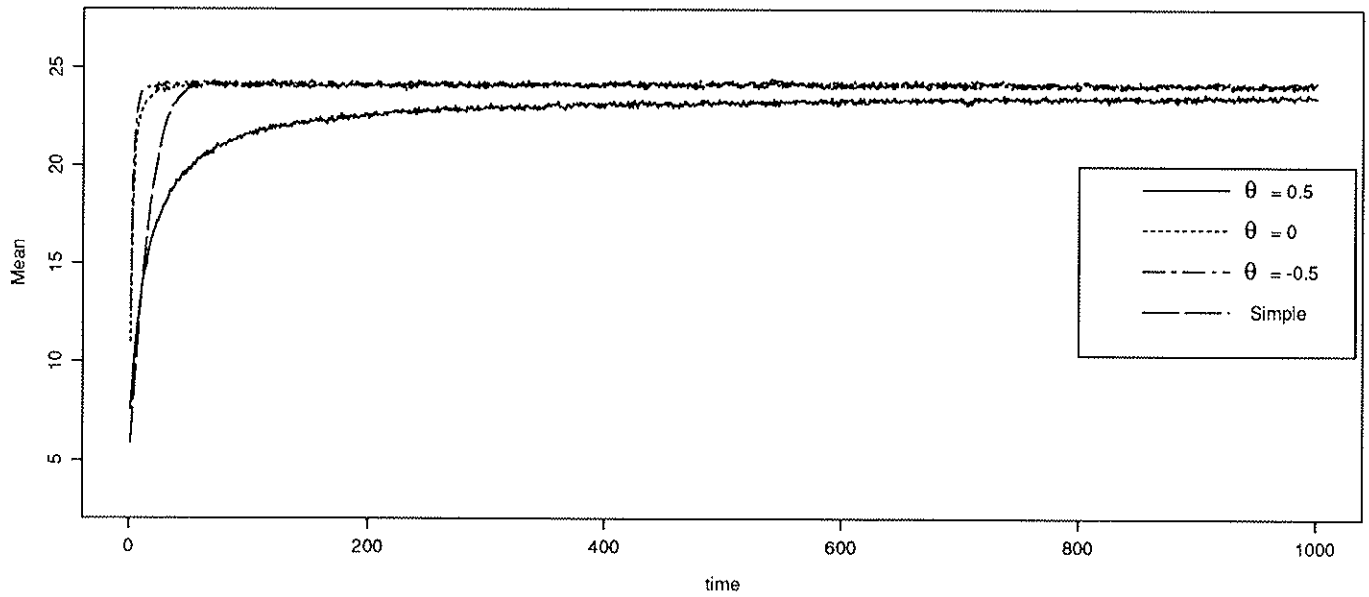
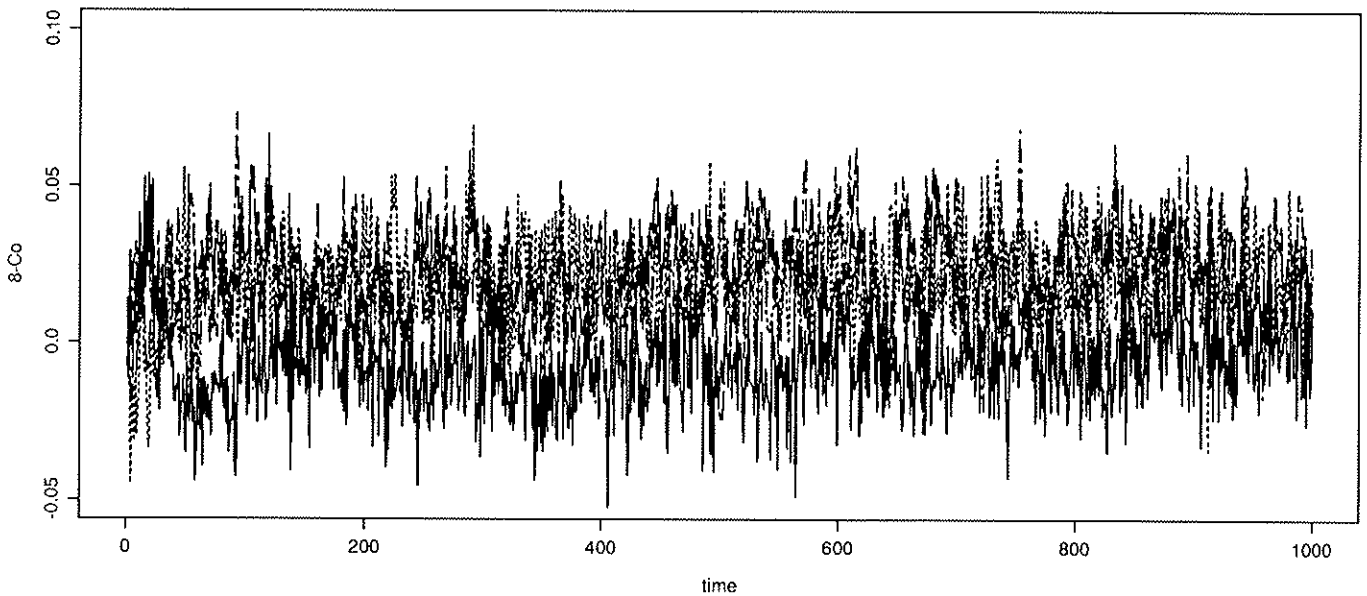


Figure 4. Poisson Case (β =0.001 [βo=0.1]) (a) Mean



(b) 8-Co



(c) PL