J. R. Statist. Soc. B (1997)
59, *No.* 2, *pp.* 291–317

Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler

By G. O. ROBERTS† and S. K. SAHU

University of Cambridge, UK

[Received November 1995. Revised June 1996]

SUMMARY

In this paper many convergence issues concerning the implementation of the Gibbs sampler are investigated. Exact computable rates of convergence for Gaussian target distributions are obtained. Different random and non-random updating strategies and blocking combinations are compared using the rates. The effect of dimensionality and correlation structure on the convergence rates are studied. Some examples are considered to demonstrate the results. For a Gaussian image analysis problem several updating strategies are described and compared. For problems in Bayesian linear models several possible parameterizations are analysed in terms of their convergence rates characterizing the optimal choice.

Keywords: BAYESIAN INFERENCE; BLOCKING; CORRELATION STRUCTURE; GAUSSIAN DISTRIBUTION; GIBBS SAMPLER; MARKOV CHAIN MONTE CARLO METHOD; MARKOV RANDOM FIELD; PARAMETERIZATION; RANDOM SCAN; RATES OF CONVERGENCE; STOCHASTIC RELAXATION; UPDATING SCHEMES

1. INTRODUCTION

The Gibbs sampler has enjoyed wide popularity, in particular for the implementation of the Bayesian paradigm. Although rival techniques based on the Hastings–Metropolis algorithm often have at least as good theoretical properties, the Gibbs sampler is often preferred because of its easy programmability and tremendous simplicity in implementation. Thus, the Gibbs sampler is considered as a default option in a wide range of problems.

The Gibbs sampler iterates by sampling from the conditional distributions of some set of co-ordinates given the values on the complement. In implementing the Gibbs sampler, there are many practical issues which need to be considered. They range from choosing a good starting point to finding an iteration number to stop sampling.

In this paper we address some of the most pertinent practical issues relating to convergence of the Gibbs sampler. Broadly we investigate the following issues:

(a) the rate of convergence of the induced Markov chain;
(b) the choice of sampling (updating, scanning) strategy and the use of random strategies;
(c) the use of blocking of some components to hasten convergence;
(d) the extent of influence of correlation structure and dimensionality of the target distribution on the rates;

†*Address for correspondence*: Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge, CB2 1SB, UK.
E-mail: g.o.roberts@statslab.cam.ac.uk

© 1997 Royal Statistical Society 0035–9246/97/59291

(e) applications of the above, e.g. to Bayesian inference in image analysis and parameterization issues for a class of generalized linear modelling problems.

We cannot hope to answer all these in total generality; instead our aims are more modest. We attempt to provide guidelines regarding these issues for a broad range of problems in Bayesian statistics. First, we review and characterize the exact computable rates of convergence of various Gibbs samplers for any multivariate Gaussian target distribution. The convergence rates turn out to be dominant eigenvalues of certain associated matrices. Using these rates we compare various updating strategies and blocking combinations for fairly general structured problems. We also obtain results on the effect of correlation on convergence, and in examples we study the effect of dimensionality and parameterization on the speed of convergence.

We consider two major examples (and many more minor ones) to illustrate our results. Our first example comes from the Bayesian analysis of a corrupted Gaussian image, where we can directly apply our results on updating schemes and blocking. Also we analyse the hierarchically recentred parameterizations (see Gelfand *et al.* (1995, 1996)) and rival parameterization schemes for the investigation of Bayesian linear models.

1.1. *Rates of Convergence of Markov Chains*

The emergence and success of Markov chain Monte Carlo (MCMC) techniques have inspired a renewed interest in Markov chains. Attention has largely focused on obtaining theoretical bounds on rates of convergence for the appropriate Markov chain. See Amit and Grenander (1991), Mengersen and Tweedie (1996), Polson (1996), Roberts and Tweedie (1996), Tierney (1994) and the references therein for a broad overview of the subject. Here we give a brief description of some of the concepts that we require.

Suppose that $\{\boldsymbol{\theta}^{(t)}, t = 0, 1, \ldots\}$ is a Markov chain with stationary density function $h(\boldsymbol{\theta})$. The rate of convergence of the Markov chain can be characterized by studying how quickly the expectations of arbitrary square h-integrable functions approach their stationary values. Let f be a square h-integrable function of $\boldsymbol{\theta}$ and $h(f)$ denote the expectation of f under the target density h.

We shall consider the rate at which $P^t f(\boldsymbol{\theta}^{(0)}) \equiv E_h[f(\boldsymbol{\theta}^{(t)})|\boldsymbol{\theta}^{(0)}]$ approaches $h(f)$ in L^2. Specifically, define ρ to be the minimum number such that for all square h-integrable functions f, and for all $r > \rho$,

$$\lim_{t\to\infty}(E_h[\{P^t f(\boldsymbol{\theta}^{(0)}) - h(f)\}^2]r^{-t}) = 0. \tag{1}$$

For one of the Gibbs sampler variants that we consider (the random permutation Gibbs sampler), the computation of ρ is not possible. However, a related *linear* convergence rate, measuring the convergence rate of only linear functions, is readily available in this case. We write ρ_L to be the largest value such that for all linear functions f, and for all $r > \rho$, equation (1) holds. ρ_L is also a natural quantity to consider when interest is restricted to linear functions only. Clearly, it is always true that $\rho_L \leqslant \rho$. However, in practice the two rates are usually equal, particularly for the Gaussian problems that we consider.

1.2. *Designing a Gibbs Sampler: Practical Convergence Issues*

Although Markov chains are well understood theoretically, bounds on convergence rates are often too conservative to be of any practical value. As a result, an assortment of convergence diagnostics based on the output of single or multiple chains are used in practice to detect convergence. For example, see Brooks and Roberts (1995) and Cowles and Carlin (1996) and the references therein. Though the diagnostics convey important information about the behaviour of the Markov chain, their interpretation must be treated with extreme care, and they cannot be used emphatically to *prove* convergence.

The choice of updating strategy for the sampler can often dramatically affect its speed of convergence. Recently, Fishman (1996) has studied a collection of random and non-random updating strategies for finite state space Markov random field target distributions. All the strategies considered there induce reversible Markov chains. Consequently a complete analysis of the various strategies was performed using eigenvalues of appropriate transition matrices.

Comparing upper bounds on rates of convergence of the Gibbs samplers on Gaussian target distributions, Amit and Grenander (1991) recommended the use of random updating strategies for arbitrary correlation structure. Barone and Frigessi (1990) also considered this problem and showed examples where random schemes are faster. In this paper we show that deterministic schemes are better in two classes of problem: for a class of hierarchically structured problem and for the class of densities with non-negative partial correlations. However, we also demonstrate that deterministic updating strategies can be considerably slower to converge outside both of our classes. We give guidelines on which updating strategy should be adopted in practical problems and illustrate these with some practical examples.

Gaussian Markov random field priors with Gaussian data are used routinely in image analysis. See Winkler (1995) and the references therein. For such problems we show that deterministic updating strategies are faster to converge than random strategies. We also consider a checker-board type of updating scheme which is more convenient to use in such situations.

It is generally believed that blocking of the components leads to faster convergence rates, e.g. 'the larger the blocks that are updated simultaneously — the faster the convergence' (Amit and Grenander, 1991). Updating in a block or group is often more computationally demanding than the corresponding componentwise updating scheme. However, blocking is still worth consideration, because it 'moves any high correlation . . . from the Gibbs sampler over to the random vector generator' (Seewald, 1992). Liu *et al.* (1994) compared blocking and collapsing for the three-component Gibbs samplers and the related data augmentation schemes by using norms of the forward and backward operators of the induced Markov chains. However, as they mentioned, comparisons using spectral radii (which describe the true rate of convergence) are less clear from their analysis. Besag *et al.* (1995) noted that blocking or grouping some components of a multivariate normal target distribution which has all partial correlations non-negative reduces the variance of ergodic averages.

Here we obtain a result that is similar in spirit to that cited by Besag *et al.* (1995), giving sufficient conditions (but certainly not necessary) for blocking to improve the convergence of a sampler. However, it is important to emphasize that *blocking can also make an algorithm converge more slowly*. We give two examples to illustrate this

in Section 2.4. Furthermore, any gain in reducing the rate of convergence may be offset by other computational concerns, since block updating requires more computational effort. However, in this paper we do not consider such issues and restrict ourselves to examining cases where blocking reduces the rate of convergence.

It is well known that high correlations between the co-ordinates diminish the speed of convergence of the Gibbs sampler; see, for example, Hills and Smith (1992). The correlations among the co-ordinates are determined by the particular parameterization of the problem. Gelfand *et al.* (1995, 1996) argue that a *hierarchically centred* parameterization leads to faster convergence and mixing because it generally leads to smaller intercomponent correlations among the co-ordinates in practical problems in Bayesian linear models. Using the results of this paper, we give a complete analysis of the hierarchically centred parameterization and some of its rival parameterizations, demonstrating that hierarchical recentring gives faster mixing Gibbs samplers than others for such problems.

1.3. *Plan of the Paper*

The remainder of the paper is organized as follows. Section 2 analyses the Gibbs sampler for a Gaussian target distribution. In Section 2.1 we introduce all the updating strategies. We develop general methodology for calculating the exact rate of convergence of the Gibbs sampler in Section 2.2. Using these exact rates we compare different updating strategies in Section 2.3. Theoretical results on blocking are discussed in Section 2.4. In Section 3 we calculate rates for two practical examples. The analysis of Gaussian Markov random field target distributions is given in Section 3.1. Section 3.2 considers a Gaussian target distribution with exchangeable correlation structure, and in this simple class of examples the effect of correlation structure on the Gibbs sampler updating scheme is investigated in detail. We also study the effect of dimensionality on the Gibbs sampler for this problem. Section 4 analyses the hierarchically centred parameterization, and other schemes for Gaussian linear models, and a summary is given in Section 5. To enhance clarity and continuity, we have placed many of the proofs of our results in Appendix A.

2. RATES OF CONVERGENCE

2.1. *Gibbs Sampler and Updating Strategies*

Suppose that our m-dimensional target vector $\boldsymbol{\theta}$ has a density $h(\boldsymbol{\theta})$. To sample from the distribution $h(\boldsymbol{\theta})$, the Gibbs sampler creates a transition from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ as follows. We partition $\boldsymbol{\theta}$ into s blocks, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_s)$, where the ith block contains r_i components with $\Sigma r_i = m$. We obtain $\boldsymbol{\theta}_1^{(t+1)}$ as a draw from the conditional density,

$$h(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_s^{(t)}).$$

We then obtain $\boldsymbol{\theta}_2^{(t+1)}$ as a draw from

$$h(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_s^{(t)}),$$

and so on, until finally $\boldsymbol{\theta}_s^{(t+1)}$ is drawn from

$$h(\boldsymbol{\theta}_s|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t+1)}),$$

whence $\boldsymbol{\theta}$ has been fully updated. This is the most widely used version of the Gibbs sampler. Also this is the deterministic sweep strategy considered by Amit and Grenander (1991); henceforth we call it DUGS. Note that this sampler updates the components in the natural ordering. In what follows, we shall also consider and compare updating orders other than the natural order.

We also consider a reversible version of DUGS. For this version we perform one forward update, i.e. update in the order 1, 2, . . ., s immediately followed by one backward update, i.e. update in the order s, $s-1$, . . ., 1. We label this strategy as REGS. In practice the behaviour of REGS will be very similar to that of DUGS. However, the reason that it has been considered in the literature is that, unlike DUGS, it induces a reversible Markov chain and in general such Markov chains are easier to analyse.

Next we describe the random sweep strategy Gibbs sampler. At any update, we generate a uniformly distributed random variable i over $\{1, 2, \ldots, s\}$ and decide to update the block $\boldsymbol{\theta}_i$. We repeat this s times and treat the entire exercise as one iteration for meaningful comparison with other strategies. We name this strategy RSGS for future reference. Amit and Grenander (1991) and also Fishman (1996) considered a variant of this strategy in which successive updated components are required to be distinct. But we do not consider that here.

For the final updating strategy, suppose that at the tth iteration we generate a random permutation $\mathcal{Z} = (z_1, z_2, \ldots, z_s)$ of $\{1, 2, \ldots, s\}$ and decide to update the components in that order. The resulting Gibbs sampler, denoted by RPGS, is called a random permutation Gibbs sampler.

To examine blocking issues we investigate how to group univariate components of $\boldsymbol{\theta}$. We compare strategies which update each blocked component of $\boldsymbol{\theta}$ simultaneously against the strategy which updates each univariate component of at least one block, say $\boldsymbol{\theta}_i$ with $r_i > 1$, sequentially. Without loss of generality, we only consider strategies which group only the adjacent components of $\boldsymbol{\theta}$. If we wish to consider the effect of blocking non-adjacent components, we can permute the individual (univariate) components of $\boldsymbol{\theta}$ so that they become adjacent and proceed from there.

2.2. *Exact Rate of Convergence of Gibbs Sampler for Gaussian Target Distributions*

We assume that $h(\boldsymbol{\theta})$ is the density of an m-dimensional Gaussian target vector $\boldsymbol{\theta}$ with mean $\boldsymbol{\mu}$ and dispersion Σ. Let $\boldsymbol{\mu}_i$, $r_i \times 1$, denote the mean for the ith component block of $\boldsymbol{\theta}$. We partition Σ and $Q = \Sigma^{-1}$ according to the Gibbs sampler blocking scheme being considered, i.e.

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \ldots & \Sigma_{1s} \\ \Sigma_{21} & \Sigma_{22} & \ldots & \Sigma_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{s1} & \Sigma_{s2} & \ldots & \Sigma_{ss} \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} Q_{11} & Q_{12} & \ldots & Q_{1s} \\ Q_{21} & Q_{22} & \ldots & Q_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s1} & Q_{s2} & \ldots & Q_{ss} \end{pmatrix}, \tag{2}$$

where Σ_{ij} and Q_{ij} are matrices of order $r_i \times r_j$. Let

$$A = I - \text{diag}(Q_{11}^{-1}, \ldots, Q_{ss}^{-1})Q \tag{3}$$

where I is the identity matrix of order m. We partition A also in this manner and let A_{ij} of order $r_i \times r_j$ denote the (i, j)th block of A. Observe that all the diagonal blocks of A are null matrices. In what follows, the matrix A is the starting point in many computations. Henceforth we shall refer to this as the A-matrix. Straightforward calculation yields

$$E(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j, j \neq i) = \sum_{j=1}^{s} A_{ij}\boldsymbol{\theta}_j + \mathbf{a}_i,$$

$$\text{disp}(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j, j \neq i) = Q_{ii}^{-1}$$

where

$$\mathbf{a}_i = Q_{ii}^{-1} \sum_{j=1}^{s} Q_{ij}\boldsymbol{\mu}_j.$$

The first of the above identities simply expresses the means of the complete conditionals as linear combinations of the other components. Let L be the block lower triangular matrix with blocks in the lower triangle being identical with those of A and $U = A - L$. Immediately, we can calculate the transition kernel for the Markov chain induced by DUGS. Define

$$B = (I - L)^{-1}U, \tag{4}$$

and $\mathbf{b} = (I - B)\boldsymbol{\mu}$. Note that if Σ is positive definite it is guaranteed that the inverse of $I - L$ exists.

Lemma 1. The Markov chain induced by DUGS has a normal transition density with mean

$$E(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = B\boldsymbol{\theta}^{(t)} + \mathbf{b} \tag{5}$$

and dispersion $\Sigma - B\Sigma B'$. Thus $\{\boldsymbol{\theta}^{(t)}\}$ induced by DUGS is a multivariate AR(1) process.

The matrix B, given in equation (4), describes the rotational component of the affine map studied in Amit and Grenander (1991). Since the Gibbs sampler is positive recurrent the roots of B should lie inside the unit disc but might not be all real. It turns out that the maximum modulus eigenvalue of B, i.e. the *spectral radius* of B, denoted by $\rho(B)$, is the exact rate of convergence of DUGS.

Theorem 1. Suppose that the Markov chain $\boldsymbol{\theta}^{(t)}$ follows the transition law $N(B\boldsymbol{\theta}^{(t-1)} + \mathbf{b},\ \Sigma - B\Sigma B')$. The rate of convergence of the Markov chain to its stationary distribution $N(\boldsymbol{\mu}, \Sigma)$ is given by $\rho = \rho(B)$.

We label this convergence rate ρ_{DUGS} for future reference. Very similar results have also been obtained by other researchers. Goodman and Sokal (1989) showed that autocorrelation functions of $\boldsymbol{\theta}^{(t)}$ decay by this rate. Barone and Frigessi (1990) obtained this to be the rate for variational norm convergence.

Observe that B defined in equation (4) corresponds to the update ordering (1, 2, . . ., s). However, if we use a different updating order, say $\mathcal{Z} = (z_1, z_2, \ldots, z_s)$, we

shall end up with a different coefficient matrix, say $B_{\mathcal{Z}}$, satisfying equation (5) and there is no guarantee that the maximum modulus eigenvalues of B and $B_{\mathcal{Z}}$ are equal. For each different $\mathcal{Z}$, we need to evaluate the matrix $B_{\mathcal{Z}}$ and the rate thereof.

Henceforth, we use the following notation:

(a) $B_{\mathcal{Z}}$ to denote the matrix B in equation (5) for any general updating order $\mathcal{Z}$;
(b) B_+ to denote the matrix B in equation (4) to emphasize the forward updating order;
(c) B_- to denote the matrix B in equation (5) for the backward updating order $(s, s-1, \ldots, 2, 1)$;
(d) we drop the subscript when the order is irrelevant or is understood from the context without confusion.

We can obtain $B_{\mathcal{Z}}$ in three steps. First, we obtain a block matrix A^* with $A^*_{ij} = A_{z_i z_j}$. In the second step, we split A^* into two matrices L^* and U^* as above, i.e. L^* is strictly block lower triangular and U^* is strictly block upper triangular. Then we set $B^* = (I - L^*)^{-1}U^*$. B^* is the coefficient matrix in equation (5) when $\boldsymbol{\theta}$ is ordered according to $\mathcal{Z}$. At the final step, we rearrange the blocks of the B^*-matrix to obtain $B_{\mathcal{Z}}$. More precisely, we set the $(z_i z_j)$th block of $B_{\mathcal{Z}}$ to be B^*_{ij}. We remark that in general the first two steps do not commute, i.e. B^*_{ij} is not the same as $B_{z_i z_j}$.

The optimal DUGS updating order is a permutation $\mathcal{Z}$ such that $B_{\mathcal{Z}}$ has the least spectral radius. Consider the indices $\{z_1, z_2, \ldots, z_s\}$ placed on the circumference of a circle. DUGS corresponding to the updating order $\mathcal{Z}$ can be thought of as visiting each index in the clockwise direction. It is clear by simple argument that DUGS for two updating orders have equal convergence rates if the orders give rise to the same circular permutation of the indices $\{1, 2, \ldots, s\}$. Note that there are $(s-1)!$ distinct circular permutations. Though we cannot provide a general method of choosing the best order we can halve the number of permutations to be searched for because of the result given below.

Lemma 2. The matrices B_+ and B_- have the same eigenvalues.

Therefore the forward updating strategy has the same convergence rate as the backward updating strategy, i.e. the rate of convergence of the Gibbs sampler is not affected by the direction (clockwise or anticlockwise) in which it updates. Hence, we can restrict attention to $(s-1)!/2$ of the possible $s!$ DUGS.

Recall that REGS updates one in the order $1, 2, \ldots, s$ immediately followed by one in the order $s, s-1, \ldots, 1$, i.e. REGS operates on $\boldsymbol{\theta}$ twice, once with B_+ and then with B_-. Hence, we take the positive square root of the maximum eigenvalue of the product B_+B_- as the convergence rate to provide a meaningful comparison with other strategies. We call this convergence rate ρ_{REGS}.

The RSGS convergence rate can be found by using an expansion of functions by the Hermite polynomials considered by Amit (1995). Let $\lambda(A)$ denote the maximum eigenvalue of the matrix A defined in equation (3). We point out that in general $\lambda(A) \neq \rho(A)$, the spectral radius of A.

Theorem 2. The RSGS convergence rate for target distribution $N(\boldsymbol{\mu}, \Sigma)$ is

$$\rho_{\mathrm{RSGS}} = [s^{-1}\{s - 1 + \lambda(A)\}]^s. \tag{6}$$

In the absence of any blocking, i.e. $s = m$, this rate is exactly equal to the convergence rate given by Amit. For RPGS we can find ρ_L analytically.

Theorem 3. Consider the following random scan sampler. Let C_i, $1 \leqslant i \leqslant n$, be $m \times m$ matrices and Ψ_i, $1 \leqslant i \leqslant n$, be $m \times m$ non-negative definite matrices. Given $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\theta}^{(t+1)}$ is chosen from $N(C_i\boldsymbol{\theta}^{(t)} + \mathbf{c}_i, \Psi_i)$, with probability w_i, $0 \leqslant w_i \leqslant 1$ for each i and $\Sigma w_i = 1$. Thus at each transition the chain chooses from a mixture of autoregressive alternatives. For this chain, ρ_L is the maximum modulus eigenvalue of the matrix $C = \Sigma_{i=1}^{n} w_i C_i$.

Following theorem 3, for RPGS we average over all possible $B_{\mathcal{Z}}$-matrices corresponding to all possible permutations $\mathcal{Z}$. The ρ_L for RPGS is the spectral radius of the matrix $(1/s!)\,\Sigma_{\mathcal{Z}}\, B_{\mathcal{Z}}$. We call the convergence rate (ρ_L) for this strategy ρ_{RPGS}. Since ρ_{RPGS} is the convergence rate corresponding to the linear functions only, it is not directly comparable with ρ_{DUGS}. However, ρ_{RPGS} provides a lower bound to the true rate of convergence of RPGS and we use this lower bound to compare its performance with the other updating strategies considered here.

We summarize the above presentation in the following theorem.

Theorem 4.

(a) $\rho_{DUGS} = \rho(B_+)$.
(b) $\rho_{REGS} = \rho(B_+B_-)^{1/2}$.
(c) $\rho_{RSGS} = [s^{-1}\{s - 1 + \lambda(A)\}]^s$.
(d) ρ_L for RPGS is $\rho\{(1/s!)\,\Sigma_{\mathcal{Z}}\, B_{\mathcal{Z}}\}$ where the summation is over all possible permutations $\mathcal{Z}$ of $(1, 2, \ldots, s)$.

Using theorem 4, we can calculate the exact convergence rate for any updating strategy and any blocking structure when the target dispersion matrix Σ is given. If any other form of random updating strategy is considered then we can use theorem 3 to find the convergence rate for the linear functionals. For example, consider the following random updating strategy. At each iteration we make a forward or a backward update with equal probability. Following theorem 3 we find the exact convergence rate for the linear functionals to be the maximum modulus eigenvalue of $\frac{1}{2}(B_+ + B_-)$.

2.3. *Comparing Deterministic Updating with Random Sweep Gibbs Sampling*

Insight into the behaviour of the samplers for the Gaussian case comes from the study of the iterative solution of linear equations $Q\mathbf{x} = \mathbf{c}$, where $\mathbf{c}$ is a vector of constants. It is well known that DUGS and RSGS have non-stochastic counterparts in the numerical analysis literature; see, for example Barone and Frigessi (1990), Goodman and Sokal (1989) and Neal (1995). Non-stochastic versions of DUGS and RSGS are known as the *Gauss–Siedel* and *Jacobi* relaxation techniques respectively. Varga (1962) and Young (1971) are among the key references on these topics. However, care must be taken in the possible reformulation and applicability of the numerical analysis results in our stochastic cases. Sometimes the non-stochastic versions may not converge at all, whereas their random counterparts will; for example, consider a Q-matrix for which $\rho(A) > 1$, where A is given in equation (3). In such a situation the Jacobi method will fail to converge but RSGS will converge

without any difficulty. However, we can still reformulate some of the matrix theory results to study the stochastic versions. In the remainder of this subsection we examine DUGS and RSGS from the theoretical perspective. Applications to particular cases are considered in Sections 3 and 4.

Stochastic overrelaxation and underrelaxation techniques are suggested to speed up convergence and/or to improve efficiency in estimation; see, for example, Barone and Frigessi (1990), Green and Han (1992) and Neal (1995). In this paper we do not consider these since they have limited applicability in the non-Gaussian case. However, see, for example, Neal (1995), who reviews generalizations of these methods and gives many references including Green and Han (1992), who showed how these methods can be used for non-Gaussian image analysis problems by incorporating the relaxation parameter in a Metropolis proposal distribution.

We first consider a structured situation that is often encountered in practice. Suppose that after suitable permutation of the rows and the corresponding columns, if necessary, the inverse dispersion matrix Q has the block tridiagonal form, i.e. Q_{ij} is a null matrix for $|i-j|>1$. Therefore Q can be written in the form (7). We consider the Gibbs samplers which use this blocking scheme. Note, however, that this blocking strategy includes sequential updating of all the components of any block, say the ith, for which Q_{ii} is diagonal:

$$Q=\begin{pmatrix} Q_{11} & Q_{12} & 0 & \dots & 0 & 0 \\ Q_{21} & Q_{22} & Q_{23} & \dots & 0 & 0 \\ 0 & Q_{32} & Q_{33} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & Q_{s-1,s-1} & Q_{s-1,s} \\ 0 & 0 & 0 & \dots & Q_{s,s-1} & Q_{ss} \end{pmatrix}. \tag{7}$$

Theorem 5. If a Gaussian target density has the inverse dispersion matrix Q of the form (7), then the rate of convergence of DUGS (with the above blocking scheme) can be found as follows.

$$\rho_{\mathrm{DUGS}}=\lambda(A)^2. \tag{8}$$

Theorem 5 has many important implications. We do not need to compute the matrix B to find the rate of convergence. It turns out that many practical examples can be analysed using the above theorem; for example see the image model (11) and the hierarchical version of the model (14). Also in many general hierarchical models we deal with such covariance structure; see, for example, Smith and Roberts (1993), section 5.2.

Corollary 1. For Gaussian target distributions having inverse dispersion matrix Q in the block tridiagonal form (7), $\rho_{\mathrm{RSGS}} \geqslant \rho_{\mathrm{DUGS}}$.

Corollary 2. For Gaussian target distributions having inverse dispersion matrix Q in the block tridiagonal form (7),

$$\lim_{\rho_{\mathrm{DUGS}}\to 1}\left(\frac{1-\rho_{\mathrm{DUGS}}}{1-\rho_{\mathrm{RSGS}}}\right)=2.$$

Therefore, RSGS will take approximately twice as many iterations as DUGS will

take to achieve the same level of accuracy (at least for fairly slowly converging problems). This is parallel to the result from the numerical analysis literature that the Gauss–Siedel method is twice as fast as the Jacobi method in this type of situation; see Goodman and Sokal (1989) for more in this regard.

Sometimes a checker-board type of updating order is suggested for the correlation structure Q in equation (7). In this scheme an update of all the odd-numbered blocked components is followed by an update of all the even-numbered components. Hence the basic Gibbs sampler here is a two-component sampler and parallel updating can be performed within each block. See the image analysis example in Section 3.1 for a practical application of this. However, the rates of convergence for the two Gibbs samplers are the same as the following corollary demonstrates.

Corollary 3. The rate of convergence of the blocked deterministic checker-board updating order and the rate for the blocked lexicographic updating order are the same for the Gaussian target distribution with inverse dispersion matrix Q of the form (7).

Though theorem 5 gives many nice results, it should be used with caution on two counts. We emphasize that the rate relationship (8) is only true for the updating orders which put Q in the form (7). It may not be true for DUGS corresponding to other updating orders for which Q is not of the form (7). Also the role of blocking is important. When components of one block, say $\boldsymbol{\theta}_i$, are updated by using a sequential updating scheme, but $r_i > 1$ and Q_{ii} is not diagonal, the above result, i.e. equation (8), may not hold. However, it is valid if we update each component of $\boldsymbol{\theta}_i$ sequentially with $r_i > 1$ but the corresponding Q_{ii} is diagonal, i.e. components of $\boldsymbol{\theta}_i$ are conditionally independent.

The results that we describe below are for the case when all the off-diagonal elements of Q are non-positive, i.e. all partial correlations are non-negative. This type of correlation structure for the Gibbs sampler has been studied previously. The physical field for the Gaussian density is 'attractive' (see, for example, Barone and Frigessi (1990)) and overrelaxation leads to faster convergence. (Recall that partial correlation between the components i and j is $-q_{ij}/\sqrt{(q_{ii}q_{jj})}$; see, for example, Whittaker (1990), p. 143.) For such a Q-matrix $\Sigma\,(= Q^{-1}) \geqslant 0$, elementwise, i.e. all ordinary correlations are non-negative. For a proof of this fact, see Young (1971), pages 43–44. But note that in general $\Sigma \geqslant 0$ elementwise does not imply that Q has all non-positive off-diagonal elements.

Theorem 6. If all partial correlations of a Gaussian target density are non-negative, i.e. all off-diagonal elements of $Q\,(= \Sigma^{-1})$ are non-positive, then $\rho_{\mathrm{RSGS}} > \rho_{\mathrm{DUGS}}$.

Theorem 6 and corollary 1 of theorem 5 demonstrate that the random strategies can be slower than the deterministic strategies. However, the conclusions of theorem 6 are often false when the above assumptions are relaxed. In Section 3.2 we consider cases where the random strategies converge substantially faster. Many practical examples can be studied by using theorem 6, e.g. the imaging model (11) in Section 3.1 and the hierarchical version of the Bayesian linear model (14) in Section 4.2.

We conclude this discussion by considering one more structured situation. Sometimes it is necessary to compare the strategies for two different target inverse dispersion matrices. We obtain the following result which has important practical applications, e.g. the imaging model (11) in Section 3.1.

Theorem 7. Suppose that we have two target inverse dispersion matrices Q and V. Also suppose that $s = m$ (i.e. no blocking) and

$$q_{ii} \leqslant v_{ii} \qquad \text{for each } i \text{ and } q_{ij} = v_{ij} \leqslant 0 \text{ for } i \neq j. \tag{9}$$

Then we have $\rho_{\mathrm{DUGS}}(V) \leqslant \rho_{\mathrm{DUGS}}(Q)$ and also $\rho_{\mathrm{RSGS}}(V) \leqslant \rho_{\mathrm{RSGS}}(Q)$, where $\rho_{\mathrm{DUGS}}(H)$ and $\rho_{\mathrm{RSGS}}(H)$ denote respectively DUGS and RSGS rates of convergence for the Gaussian target distribution with inverse dispersion matrix H.

2.4. *Blocking Strategies*

In the description of the Gibbs sampler in Section 2.1, there is scope for updating one or all of the complete conditionals in blocks as each r_i can be bigger than 1. We aim to compare the blocked scheme with a scheme which updates all the components of at least one blocked component sequentially.

First, suppose that all the components of $\boldsymbol{\theta}_i$ with $r_i > 1$ are conditionally independent. Such a block is termed a 'coding set' by Besag *et al.* (1995), section 2.4.4. Then, it is clear that the blocked strategy (blocking all components in the coding set) and the unblocked strategy which updates the components in the ith block using the sequential univariate updating procedure lead to the same A- and B-matrices as given in equations (3) and (4). Hence this blocking does not reduce ρ_{DUGS} and $\lambda(A)$ remains unaltered for the two schemes.

However, note that ρ_{RSGS} as given in equation (6) is an increasing function of s, the number of blocks, when $\lambda(A)$ is held fixed. Therefore, this convergence rate improves when we block the conditionally independent components in a single group (because this reduces s without altering the matrix A). Section 3.1 gives an example of this blocking structure. In the remainder of this section we only compare DUGS for various blocking strategies.

We next consider the case when all partial correlations of a Gaussian target density are non-negative, i.e. $q_{ij} \leqslant 0$ for all $i \neq j$. Suppose that, written as a block matrix, the inverse dispersion matrix Q is as in equation (2). We want to compare the corresponding blocked DUGS with the scheme which updates at least one block which has more than one component, say the ith such that $r_i > 1$, componentwise and also at least one of the off-diagonal elements in the matrix Q_{ii} is non-zero, i.e. Q_{ii} is not diagonal.

Theorem 8. If all partial correlations of a Gaussian target density are non-negative, i.e. all the off-diagonal elements of Q $(= \Sigma^{-1})$ are non-positive, then the blocked DUGS has a faster rate of convergence than the DUGS which updates at least one block, say the ith, componentwise for which $r_i > 1$ and Q_{ii} is not a diagonal matrix.

What happens if the non-negativity assumptions of all partial correlations do not hold? For general correlation structure, it is difficult to formulate a blocking strategy which will *guarantee* a decrease in the DUGS rate of convergence. We investigate a three-dimensional Gaussian target distribution with the following inverse dispersion matrix:

$$\Sigma^{-1} = Q = \begin{pmatrix} 1 & q_{12} & q_{13} \\ q_{12} & 1 & q_{23} \\ q_{13} & q_{23} & 1 \end{pmatrix}. \tag{10}$$

(Without loss of generality we have assumed that all diagonal elements of Q are 1.) Let $\Delta = |Q| = 1 - q_{12}^2 - q_{13}^2 - q_{23}^2 + 2q_{12}q_{13}q_{23}$.

Consider the DUGS which blocks the first two components together. By direct calculations, it is easy to see that the rate of convergence of this DUGS is $1 - \Delta/(1 - q_{12}^2)$. Observe that this is a non-increasing function of $|q_{12}|$. Hence, we have the following theorem.

Theorem 9. For a three-dimensional Gaussian target distribution, the DUGS which blocks the two components having the maximum absolute partial correlation is faster than any other two-component DUGS for the same target distribution.

The more important question regarding the comparison of the best two-component DUGS with the usual three-component scheme is more involved. The convergence rate for the three-component DUGS is the maximal root of the quadratic $\lambda^2 - b\lambda + c = 0$, where $b = q_{12}^2 + q_{13}^2 + q_{23}^2 - q_{12}q_{13}q_{23}$ and $c = q_{12}q_{13}q_{23}$. Several cases depending on the roots being real or imaginary arise.

Theorem 10. For a three-dimensional Gaussian target distribution, with inverse covariance matrix Q given in equation (10), the DUGS which blocks components 1 and 2 is faster than the three-component DUGS if any one of the following conditions holds:

(a) $b^2 = 4c$ and $b/2 > 1 - \Delta/(1 - q_{12}^2)$;
(b) $b^2 > 4c$ and $q_{12}^4 - q_{12}^2 b + c < 0$;
(c) $b^2 > 4c$ and $q_{12}^4 - q_{12}^2 b + c > 0$ and $b/2 > 1 - \Delta/(1 - q_{12}^2)$;
(d) $b^2 < 4c$ and $\sqrt{c} > 1 - \Delta/(1 - q_{12}^2)$.

Otherwise the univariate componentwise DUGS is faster than the blocked DUGS.

We illustrate these results by considering two examples. Both the examples reveal that blocking may worsen the rate of convergence.

Example 1 is taken from Liu *et al.* (1994). Suppose that we have the following target dispersion matrix:

$$\Sigma = \begin{pmatrix} 1 & & \\ a & 1 & \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

For $a > \frac{1}{4}$, the dispersion matrix satisfies the conditions of theorem 8; therefore blocking any two components will lead to a faster rate of convergence. When $a < \frac{1}{4}$ the DUGS which blocks components 1 and 2 will be slower than the full three-component scheme by theorem 10.

Example 2, taken from Whittaker (1990), p. 319, demonstrates a real situation where blocking worsens the rate of convergence. Here the inverse dispersion matrix is

$$Q = \begin{pmatrix} 1 & & & & & & \\ 0.611 & 1 & & & & & \\ -0.108 & -0.152 & 1 & & & & \\ 0.25 & 0.277 & -0.1 & 1 & & & \\ 0.248 & 0.294 & -0.105 & 0.572 & 1 & & \\ 0.410 & 0.446 & -0.213 & 0.489 & 0.597 & 1 & \\ 0.331 & 0.303 & -0.153 & 0.335 & 0.478 & 0.651 & 1 \end{pmatrix}.$$

The unblocked DUGS rate is 0.4843, whereas if components 1 and 2 are blocked the rate is 0.4928.

3. EXAMPLES

3.1. *Exact Rates of Convergence for Bayesian Image Analysis*

In image analysis problems (see, for example, Winkler (1995)), often the true image $\boldsymbol{\theta}$ and the observed corrupted image $\mathbf{y}$ are modelled as follows. First, an *Ising model* type of prior distribution on the $p \times p$ lattice with the inverse temperature $\beta > 0$ is assumed, e.g.

$$\pi(\boldsymbol{\theta}) \propto \exp\left\{ -\beta \sum_{[i,j]} (\theta_i - \theta_j)^2 \right\}$$

where the sum is over all neighbours. In the second step, y_i for each pixel i is assumed to follow an independent Gaussian distribution with mean θ_i and variance σ^2. Hence the posterior distribution is of the form

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i} (y_i - \theta_i)^2 \right\} \tag{11}$$

where the sum is over all pixels i. We assume orthogonal neighbourhood structure. Each pixel has four neighbours except for those in the boundary of the lattice which have three or two. Four corner pixels have two neighbours each and all other boundary pixels have three neighbours.

There are many possible updating orders and blocking schemes for this problem. First, the most popular lexicographic updating order can be used. Also we can consider a checker-board type of updating order as below. Suppose that we partition the pixels as in Fig. 1 (for $p = 5$) into two types: black and white. The checker-board type of DUGS updates each pixel of one colour and then updates each pixel of the other colour. This scheme may be preferred because a suitable parallel updating schedule may be implemented here. It is guaranteed by lemma 2 that the rate of convergence is unaffected by the order in which the colours are updated. The inverse dispersion matrix for the checker-board type of updating scheme is in the form (7)

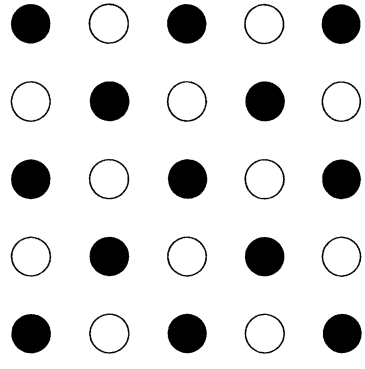

Fig. 1. Checker-board type of updating in a Markov random field

with $s = 2$. Here Q_{11} and Q_{22} are diagonal matrices. Hence we can use theorem 5 to find its rate of convergence. However, we cannot use corollary 3 to compare the above two deterministic schemes because we update the pixels sequentially in the lexicographic case. However, see the numerical illustration below. There we see numerically that both of the schemes have an identical rate of convergence.

We also consider a blocked lexicographic updating order where pixels in each row are put into separate blocks. Here block updating can be done easily by using Cholesky decomposition for moderate dimensional problems. However, this will be impractical for high dimensional problems, e.g. $p = 64$. The inverse dispersion matrix for this scheme is of the form (7). Also consider the following blocked checker-board type of updating scheme. Here we update all the odd-numbered blocks first and then we update all the even-numbered blocks. By corollary 3, the two updating orders will converge at the same rate. Again, parallel updating can be implemented for the later scheme.

DUGS for any updating order is faster than RSGS for this example. However, it is easy to observe that the two types of pixel are conditionally independent given the other type. Hence, recalling the discussion in Section 2.4, we shall succeed in reducing the RSGS rate of convergence by blocking the pixels of each colour together. This grouping structure is also advocated by Besag *et al.* (1995) and Fishman (1996) for the discrete case. This is also known in the numerical analysis literature as the red–black ordering; see, for example, Goodman and Sokal (1989). Observe that the deterministic updating scheme corresponding to this random scheme is the checker-board type of updating protocol described above.

The role of σ^2, the inverse precision, can be studied by using theorem 7. The posterior inverse covariance matrix differs from the prior inverse covariance matrix only in the diagonal elements. The data only increase the diagonal elements of the inverse covariance matrix by σ^{-2}, leaving all off-diagonal elements unaltered. Theorem 7 applies directly to this case. (Observe that all the off-diagonal elements of the prior inverse covariance matrix are non-positive.) Hence, the precision of the data directly improves the rate of convergence of the Gibbs sampler.

We give a few illustrative examples of rates of convergence for both the unblocked and the blocked schemes. We take σ to be 5 as in Green and Han (1992). In each case we have considered both DUGS and RSGS. The exact rates for three different inverse temperatures, $\beta = 0.001, 0.01, 0.1$, and for two different lattice sizes $p = 16$ and $p = 25$ are given in Table 1. The rates of convergence for the pixelwise lexicographic and checker-board type of updating schemes were found to be the same in each case. It may be observed from Table 1 that DUGS in each case has a faster convergence rate than RSGS agreeing with corollary 2. The parameter β has a large influence on the convergence of the algorithm; however, we do not investigate it in detail here.

3.2. *Exchangeable Correlation Structure*

Here we assume that the target dispersion matrix is of the form

$$\Sigma_{m\times m} = aI + bJ, \qquad a > 0, \quad a + mb > 0, \tag{12}$$

where J is an $m \times m$ matrix of all 1s.

Correlation between any two components is $b/(a + b)$ whereas partial correlation

TABLE 1
Convergence rates for the imaging model (11)†

β	*Convergence rates for the following values of p:*							
	$p = 16$				$p = 25$			
	Pixelwise updating		*Blocked*		*Pixelwise updating*		*Blocked*	
	DUGS	*RSGS*	*DUGS*	*RSGS*	*DUGS*	*RSGS*	*DUGS*	*RSGS*
0.001	0.02688	0.33870	0.00799	0.29670	0.02739	0.33959	0.00815	0.29716
0.010	0.43425	0.68805	0.24315	0.55734	0.43953	0.69137	0.24685	0.56014
0.100	0.90191	0.95032	0.81839	0.90692	0.90403	0.95141	0.82194	0.90879

†DUGS rates correspond to both the checker-board and the lexicographic updating orders. RSGS rates for the pixelwise updating case were computed by blocking the pixels of each colour together. The blocked scheme refers to blocking of the pixels row-wise.

is $q = b/\{a + (m-1)b\}$. Therefore, the matrix A in equation (3) has all elements q except the diagonals, which are 0, i.e. $A = q(J - I)$. It is easy to see that by symmetry ρ_{DUGS} is not dependent on the updating order. We can compute the B-matrix as defined in equation (4) exactly as follows.

Lemma 3. The B-matrix for a normal target distribution with dispersion matrix (12) has elements b_{ij} given by

$$b_{ij} = \begin{cases} q\{(q+1)^{i-1} - (q+1)^{i-j}\}, & j < i, \\ q\{(q+1)^{i-1} - 1\}, & j = i, \\ q\{(q+1)^{i-1}\}, & j > i. \end{cases}$$

Consider first the trivial case $b = 0$, in which case $b_{ij} = 0$ for all i and j. All the variants of the Gibbs sampler, except RSGS, have rate exactly 0, i.e. they will converge immediately. The exact RSGS rate is $(1 - 1/m)^m$ which approaches $\exp(-1)$ as $m \to \infty$. RSGS is less effective because it does not guarantee to update all the components in any fixed number of iterations.

The target dispersion matrix of the form (12) has been used extensively in the literature as a test case for proving effectiveness of convergence diagnostics (see, for example, Raftery and Lewis (1992)) and new MCMC algorithms (see, for example, Polson (1996)). The tractability of its convergence rate makes it appealing in this respect. To illustrate this, we take $a = 0.1$ and $b = 0.9$. For $m = 10$, Raftery and Lewis (1992) recommended a DUGS burn-in period of 36 iterations, but for accurate estimation their method diagnoses over 20000 iterations, i.e. they need more than 20000 samples from the chain to estimate any feature of the distribution accurately. Theoretical bounds due to Amit (1991) suggest that a running time of 180 million iterations suffices for adequate convergence. Polson (1996) stipulates a running time of 13.8 million iterations to achieve an accuracy of 0.001 for a local jump Metropolis algorithm named the 'FKP Metropolis algorithm'. We find that the exact convergence rate of DUGS for this target distribution is 0.9758. Consequently, to achieve an accuracy of 0.001 we need to run DUGS for a burn-in period of 282 ($= \log 0.001/\log 0.9758$) iterations only!

The rates of convergence for the remaining schemes (described in Section 2.1) can be found easily by using theorem 4. The REGS convergence rate is

$$\rho_{\mathrm{REGS}} = \sqrt{\rho(BP'BP)},$$

with P having 1 in the reverse diagonal, i.e. $p_{im-i+1} = 1$, $i = 1, \ldots, m$, and the remaining elements all 0. Again, by theorem 4, ρ_{L} for RPGS is the maximum eigenvalue of the matrix with all diagonals being $\mathrm{tr}(B)/m$ and the off-diagonals being $\mathbf{1}'B\mathbf{1} - \mathrm{tr}(B)/m(m-1)$.

Theorem 11. Suppose that we have the normal target distribution with dispersion matrix (12). The exact RSGS convergence rate is

$$\rho_{\mathrm{RSGS}} = \begin{cases} \left\{(q+1)\,\dfrac{m-1}{m}\right\}^m & \text{if } q > 0, \\ \left(\dfrac{m-1-q}{m}\right)^m & \text{if } q < 0, \end{cases}$$

whereas ρ_{L} for RPGS is

$$\rho_{\mathrm{RPGS}} = \begin{cases} \dfrac{1}{m}\left\{(q+1)^m\left(m-1-\dfrac{1}{q}\right)+1+\dfrac{1}{q}\right\} & \text{if } q > 0, \\ \dfrac{1}{m(m-1)}\left\{\dfrac{1}{q}\,(q+1)^m - qm^2 - m - 1 - \dfrac{1}{q}\right\} & \text{if } q < 0. \end{cases}$$

We can compare RSGS and RPGS exactly by using theorem 11. RSGS rates are higher than RPGS rates if $q > 0$. If $q < 0$ is independent of m, RSGS has an asymptotic rate of $\exp(-1-q)$. However, we cannot compare DUGS with other schemes exactly analytically. Instead, we consider a few interesting special cases numerically. The interested reader is referred to Amit and Grenander (1991) and Barone and Frigessi (1990): both compared the updating schemes asymptotically (and also numerically) when the inverse dispersion matrix is circulant.

First, we investigate the case when all the components are negatively correlated. We take $a = 1$ and $b = -1/(m+\delta)$ in the region $0 < \delta \leqslant 1$. For $\delta = 0$ the distribution is singular. RSGS has an asymptotic rate of $\exp\{-\delta/(1+\delta)\}$ whereas RPGS has a rate of $1/(1+\delta)$. To compare these with DUGS and REGS rates we plot the exact convergence rates for all the four updating strategies in Fig. 2(a) when $\dim(m)$ is 100. DUGS and REGS are almost indistinguishable from one another. Both of them behave very badly. Surprisingly, the randomized strategies are far better. We shall consider one common practical example having this type of correlation structure in Section 4.1.

We consider another class of special cases with $a = 1$ and $b = k/(m+1)$ where $k > -1$. As k increases to 0, correlation between any two components decreases to 0 but remains negative. The rates for $m = 100$ are plotted in Fig. 2(b) for $-1 < k < 0$. It may be a little surprising that RPGS has an asymptotic rate of 0 whereas the asymptotic RSGS rate is $\exp(-1)$.

Continuing with the previous example we examine the situation for $k > 0$. As k increases, correlation between the components also increases. Using theorem 6 we

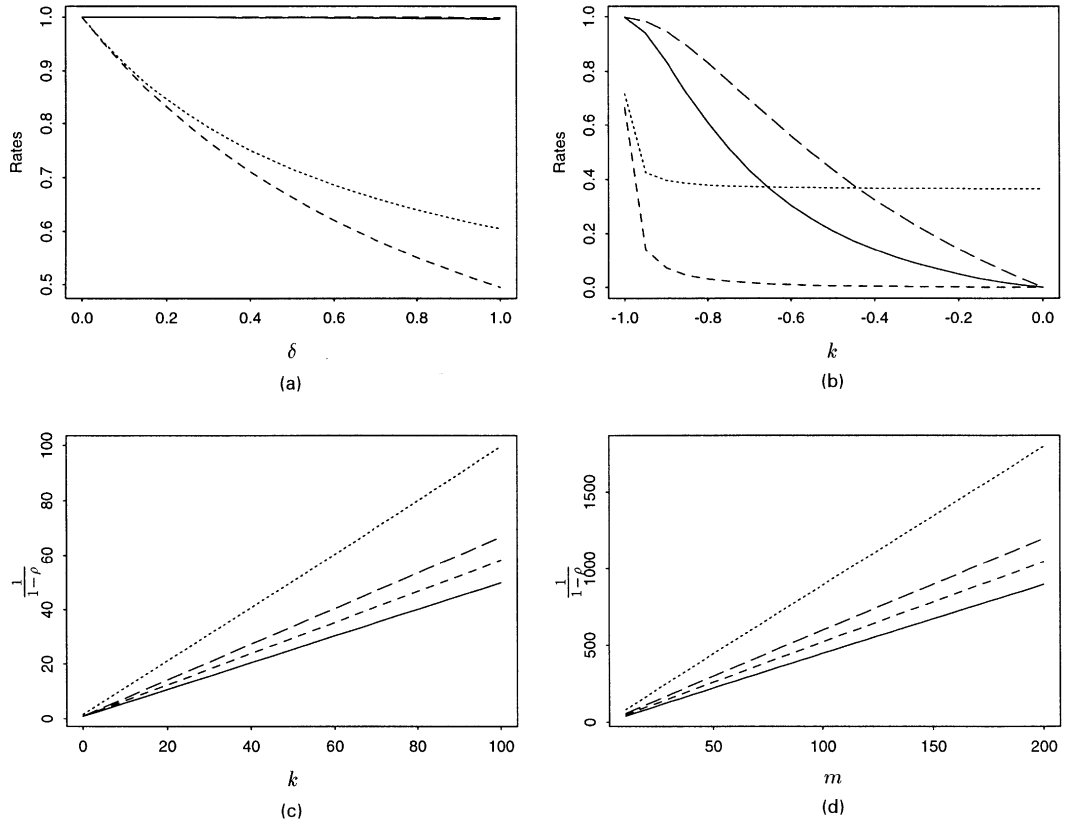

Fig. 2. Convergence rates and running times for various target Σ-matrices: (a) $\Sigma_{100} = I - J/(100+\delta)$ $(0 < \delta \leqslant 1)$; (b) $\Sigma_{100} = I + kJ/(100+1)$ $(-1 < k < 0)$; (c) $\Sigma_{100} = I + kJ/(100+1)$ $(1/(1-\rho)$ for $k > 0)$; (d) $\Sigma_m = 0.1I + 0.9J$ $(1/(1-\rho)$ for $10 < m < 200)$ (———, DUGS; ·····, RSGS; - - - -, RPGS; — — —, REGS)

conclude that the deterministic strategies have faster rates of convergence. RPGS has an asymptotic rate of

$$1 + \frac{1}{k} - \frac{1}{k} \exp\left(\frac{k}{1+k}\right)$$

whereas the asymptotic RSGS rate is $\exp\{-1/(1+k)\}$. We consider the quantity $T = 1/(1-\rho)$ to see how increasing correlation affects the running times of the algorithms. Fig. 2(c) provides a plot of T for all the strategies when m is 100. It is clear from Fig. 2(c) that increasing positive correlation only increases the running time approximately linearly.

We investigate the effect of increasing dimension on the running times of various Gibbs sampling schemes for the target correlation structure with $a = 0.1$ and $b = 0.9$. We plot $T = 1/(1-\rho)$ in Fig. 2(d) as the dimension increases from 10 to 200. We see that DUGS is uniformly better than any other strategy. It is equivalent to the

random permutation strategy asymptotically. Also the running times increase approximately linearly as the dimension increases.

As in the previous example when $a > 0$ and $b > 0$ are fixed, q (the partial correlation) goes to 0 as the reciprocal of dimension goes to 0, i.e. $q = O(m^{-1})$. From the exact calculation of ρ_{RSGS} in theorem 11, we see that the convergence rate goes to 1 at the same rate as $m \to \infty$. Thus in this simple example, if partial correlations are $O(m^{-1})$, the RSGS algorithm has polynomial (in fact linear) complexity in dimension. However, if partial correlations are $O(1)$, then the DUGS algorithm has exponential complexity, i.e. the algorithm exhibits *critical slowing down*. Thus the interplay between correlation and dimension is complicated, even in this simple example. For highly structured hierarchical models, the inverse covariance matrix typically contains $O(m)$ non-zero elements. The complexity of such a class of algorithms is clearly difficult to analyse without imposing further (and possibly artificial) restrictions. However, it is reasonable to conjecture that a suitably well-defined class of such algorithms might be polynomial in dimension.

4. OPTIMUM PARAMETERIZATIONS FOR GAUSSIAN LINEAR MODELS

4.1. *One-way Analysis of Variance*

We consider the customary one-way analysis of variance with random effects. We assume that the error variance σ_e^2 is known and after reducing by sufficiency we have a single observation y_i for each population, i.e.

$$y_i = \mu + \alpha_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{13}$$

where ϵ_i are independent and identically distributed (IID) $N(0, \sigma_e^2)$ and α_i are IID $N(0, \sigma_\alpha^2)$. We assume that μ has a flat prior, i.e. $\pi(\mu) = 1$. Let $\mathbf{y}$ be $(y_1, \ldots, y_n)'$ and $\bar{y}$ be $\Sigma\, y_i/n$.

For this model we consider three possible parameterizations. We call the above μ–$\boldsymbol{\alpha}$ representation of the model the standard parameterization. It is immediate that the exact rate for DUGS is $1 - \kappa$ by using theorem 1, where $\kappa = \sigma_e^2/(\sigma_e^2 + \sigma_\alpha^2)$, and for RSGS the asymptotic rate (as $n \to \infty$) is $\exp\{\sqrt{(1-\kappa)} - 1\}$. Hence DUGS is better than RSGS for this parameterization asymptotically for large data sets.

Following Gelfand *et al.* (1996), model (13) can also be written in a hierarchical form. Defining $\gamma_i = \mu + \alpha_i$, we have $y_i|\gamma_i \sim N(\gamma_i, \sigma_e^2)$, $\gamma_i|\mu \sim N(0, \sigma_\alpha^2)$ and a flat prior for μ. We call the μ–$\boldsymbol{\gamma}$ representation the *hierarchically centred* parameterization. The exact rate for DUGS is κ and the asymptotic RSGS rate is $\exp(\sqrt{\kappa} - 1)$. As claimed by Gelfand *et al.* (1996), the hierarchically centred parameterization will be a better choice when $\sigma_e^2 < \sigma_\alpha^2$ for both DUGS and RSGS.

Next we consider the *parameterization by sweeping* as analysed by Gilks and Roberts (1996). They imposed the restriction that $\bar{\alpha} = 0$. With the above restriction, μ and $\boldsymbol{\alpha}$ will be independent *a posteriori* and the posterior for α will be $n-1$ dimensional. With α_n deleted, their method tries to sample from the $(n-1)$-dimensional normal distribution with mean $\tau(\mathbf{y}_{-n} - \bar{y}\mathbf{1}_{n-1})$ and dispersion $\tau(I - J/n)$ where $\tau = \sigma_e^2\sigma_\alpha^2/(\sigma_e^2 + \sigma_\alpha^2)$ and $\mathbf{y}_{(-n)} = (y_1, \ldots, y_{n-1})'$. After scaling etc. the Gibbs sampler for this parameterization tries to sample from an $(n-1)$-dimensional normal distribution with dispersion matrix $I - J/n$. This target dispersion matrix has been studied as the first example in the previous section with $\delta = 1$ in Fig. 2(a).

Hence, a random permutation Gibbs sampler is the best choice for this parameterization. Also RPGS has an asymptotic rate of 0.5 whereas RSGS converges at the rate of $\exp(-0.5)$. The systematic schemes, e.g. DUGS and REGS, have asymptotic rates 1.

When the variance components are unknown (as in most practical situations), the posterior distribution will cease to be Gaussian. The variance components will be included in the model with their respective prior specifications. The Gibbs sampler needs to sample from the joint posterior distribution of the μs, αs and the variance components. However, the conditional distribution of μ and α given the variance components will still be the above normal distributions. Consequently, the behaviour of the Gibbs sampler for various parameterizations should still be guided by the above investigation.

4.2. *Multilevel Mixed Linear Models*

The findings of Section 4.1 can be generalized for the *multilevel* or *nested* mixed linear models. These models have been considered by Gelfand *et al.* (1995) to show that a *hierarchically centred* model specification leads to a more efficient Gibbs sampling scheme. With our exact rate calculation we can reveal more about the striking features of the posterior distributions arising out of these model specifications. Let, after reducing by sufficiency,

$$y_{ij} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ij}, \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, I, \tag{14}$$

where ϵ_{ij} are IID normal random variables with mean 0 and variance σ_e^2. The standard model specification assumes that the β_{ij} are IID $N(0, \sigma_\beta^2)$ and α_i are IID $N(0, \sigma_\alpha^2)$. Again all the variance components are assumed known. The centring specification assumes that $\eta_{ij} = \mu + \alpha_i + \beta_{ij}$, whence $y_{ij}|\eta_{ij} \sim N(\eta_{ij}, \sigma_e^2)$. Further, we define $\gamma_i = \mu + \alpha_i$ so that $\eta_{ij}|\gamma_i \sim N(\gamma_i, \sigma_\beta^2)$ and $\gamma_i|\mu \sim N(\mu, \sigma_\alpha^2)$. The Bayesian model specification in each case is completed by assuming a flat prior for μ.

It is straightforward to write the conditional distributions needed for the Gibbs sampling in both the cases. Let us define $n = \Sigma_i n_i$, $y_{..} = \Sigma\, y_{ij}/n$, $y_{i.} = \Sigma_j\, y_{ij}/n_i$. The complete conditionals for the uncentred specifications (μ–$\boldsymbol{\alpha}$–$\boldsymbol{\beta}$) are

$$\mu|\alpha_i, \beta_{ij} \sim N\left(y_{..} - \sum_i \frac{\alpha_i n_i}{n} - \sum_{ij} \frac{\beta_{ij}}{n},\ \frac{\sigma_e^2}{n}\right),$$

$$\alpha_i|\mu, \beta_{ij} \sim N\left\{\frac{n_i\sigma_e^{-2}}{n_i\sigma_e^{-2} + \sigma_\alpha^{-2}}\left(y_{i.} - \mu - \sum_j \frac{\beta_{ij}}{n_i}\right),\ \frac{1}{n_i\sigma_e^{-2} + \sigma_\alpha^{-2}}\right\},$$

$$\beta_{ij}|\mu, \alpha_i \sim N\left\{\frac{\sigma_e^{-2}}{\sigma_e^{-2} + \sigma_\beta^{-2}}(y_{ij} - \mu - \alpha_i),\ \frac{1}{\sigma_e^{-2} + \sigma_\beta^{-2}}\right\},$$

whereas the centred parameterization leads to the conditional distributions

$$\mu|\gamma_i, \eta_{ij} \sim N\left(\frac{\sum_i \gamma_i}{I}, \frac{\sigma_\alpha^2}{I}\right),$$

$$\gamma_i|\mu, \eta_{ij} \sim N\left(\frac{\sigma_\beta^{-2}\sum_j \eta_{ij} + \mu\sigma_\alpha^{-2}}{n_i\sigma_\beta^{-2} + \sigma_\alpha^{-2}}, \frac{1}{n_i\sigma_\beta^{-2} + \sigma_\alpha^{-2}}\right),$$

$$\eta_{ij}|\mu, \alpha_i \sim N\left(\frac{\sigma_e^{-2}y_{ij} + \gamma_i\sigma_\beta^{-2}}{\sigma_e^{-2} + \sigma_\beta^{-2}}, \frac{1}{\sigma_e^{-2} + \sigma_\beta^{-2}}\right).$$

In the μ–$\boldsymbol{\alpha}$–$\boldsymbol{\beta}$-parameterization the α_is are conditionally independent given the others, i.e. μ and β_{ij}. It is also true for the β_{ij}s, γ_is and η_{ij}s. Therefore, blocking them together does not alter the performance of the Gibbs sampler. Henceforth, we consider the three-block Gibbs samplers on μ, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the standard parameterization case and μ, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ in the hierarchical situation.

In the μ–$\boldsymbol{\alpha}$–$\boldsymbol{\beta}$ parameterization the partial correlation between any component of one block and any component of any other block is negative whereas for the μ–$\boldsymbol{\gamma}$–$\boldsymbol{\eta}$ case it is non-negative. Moreover, for the μ–$\boldsymbol{\gamma}$–$\boldsymbol{\eta}$-case all partial correlations between μ and η_{ij} are exactly 0.

Examples in Section 3.2 demonstrate that DUGS can be slower than RSGS or RPGS when partial correlations are all negative. Therefore, we can infer that DUGS for the μ–$\boldsymbol{\alpha}$–$\boldsymbol{\beta}$-parameterization is likely to be slower than RSGS or RPGS. Hence, we recommend random updating schemes for this parameterization.

In contrast, for the μ–$\boldsymbol{\gamma}$–$\boldsymbol{\eta}$-parameterization the appropriate inverse dispersion matrix Q can be written in the form (7). Also it satisfies the conditions of theorem 6, i.e. all partial correlations are non-negative. Therefore, using theorem 6 we see that random updating methods will be slower than DUGS. Hence, we shall choose DUGS in this case.

5. DISCUSSION

In this paper we have dealt only with the Gaussian models for reasons of analytical tractability. However, we may hope that the results described here can hold in more realistic non-Gaussian cases, particularly for the models which have a linear mean part. Also we can use a very crude Gaussian approximation of the target density to find an approximate rate of convergence of the Gibbs sampler. We can then use the approximate rates to decide on many practical issues discussed in Section 1. We investigate these ideas in Roberts and Sahu (1996).

ACKNOWLEDGEMENTS

This work has been supported by the Engineering and Physical Sciences Research Council of the UK. The authors thank a referee for many helpful suggestions and Yali Amit for providing a preprint of his paper.

APPENDIX A

To prove theorem 1 we need the following lemma from Horn and Johnson (1990), p. 299.

Lemma 4. Let P be an $m \times m$ given complex matrix, and let $\epsilon > 0$ be given. There is a constant $C = C(P, \epsilon)$ such that

$$|(P^k)_{ij}| \leqslant C\{\rho(P) + \epsilon\}^k$$

for all $k = 1, 2, 3, \ldots$ and all $i, j = 1, 2, 3, \ldots m$, where $\rho(P)$ is the spectral radius of P.

Proof of theorem 1. The χ^2-distance between $h^{(t)}(\boldsymbol{\theta}^{(0)}, \cdot)$, the density at the tth iteration started from $\boldsymbol{\theta}^{(0)}$, and the target density h is defined as

$$\chi^2\{h^{(t)}(\boldsymbol{\theta}^{(0)}, \cdot), h\} = \int \frac{\{h^{(t)}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}) - h(\boldsymbol{\theta})\}^2}{h(\boldsymbol{\theta})}\, \mathrm{d}\boldsymbol{\theta}.$$

For $h = N_m(\boldsymbol{\mu}, \Sigma)$ the DUGS induces a normal density $h^{(t)}$ with mean $\boldsymbol{\mu}^{(t)}$ and variance $\Sigma^{(t)}$ where

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu} + B^t(\boldsymbol{\theta}^{(0)} - \boldsymbol{\mu}),$$

$$\Sigma^{(t)} = \Sigma + B^t(\Sigma^{(0)} - \Sigma)(B')^t$$

with $\boldsymbol{\theta}^{(0)}$ is the starting point having dispersion $\Sigma^{(0)}$. Here, the χ^2-distance between $h^{(t)}$ and h is

$$|W||2W - I|^{-1/2} \exp(S/2) - 1 \tag{15}$$

where

$$W = W(\Sigma^{(0)}) = \Sigma^{(t)}\Sigma^{-1} = I + B^t(\Sigma^{(0)} - \Sigma)(B')^t\Sigma^{-1}$$

and

$$S = (\boldsymbol{\theta}^{(0)} - \boldsymbol{\mu})'(B')^t\Sigma^{(t)-1}\{I + (2W - I)^{-1}\}B^t(\boldsymbol{\theta}^{(0)} - \boldsymbol{\mu}).$$

Using lemma 4 we see that, for sufficiently large t, $W(\Sigma^{(0)})$ approaches the identity matrix. This already shows more than we require (i.e. pointwise convergence from all starting configurations with finite dispersion matrix, according to the χ^2-distance). For theorem 1 itself, we firstly note that by integrating expression (15) with respect to h

$$E_h[\chi^2\{h^t(\boldsymbol{\theta}^{(0)}, \cdot), h\}] = |W(0)||2\,W(0) - I|^{-1/2}|D\Sigma - I|^{-1/2} - 1, \tag{16}$$

where

$$D = (B')^t\{\Sigma - B^t\Sigma(B')^t\}^{-1}[I + \{2\,W(0) - I\}^{-1}]B^t.$$

Hence, for large t, $E_h[\chi^2\{h^t(\boldsymbol{\theta}^{(0)}, \cdot), h\}]$ is of the order $\rho(B)^2$. Therefore, we have $\rho_\chi \leqslant \rho(B)$, where ρ_χ^2 is the rate by which equation (16) goes to 0.

It remains to show that $\rho(B) = \rho$, the rate corresponding to the definition (1). It is easy to observe that $\rho_\chi \leqslant \rho(B) = \rho_{\mathrm{L}}$ which is less than or equal to ρ. As for the other inequality, i.e.

$$\rho_\chi \geqslant \rho, \tag{17}$$

we proceed as follows.

$$E_h[\{P^t f(\boldsymbol{\theta}^{(0)}) - h(f)\}^2] = \int \left\{ \int \frac{h^t(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}) - h(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} f(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \, \mathrm{d}(\boldsymbol{\theta}) \right\}^2 h(\boldsymbol{\theta}^{(0)}) \, \mathrm{d}\boldsymbol{\theta}^{(0)}$$

$$\leqslant ||f||_2^2 \, E_h[\chi^2\{h^t(\boldsymbol{\theta}^{(0)}, \cdot), h\}].$$

Here $||\cdot||$ denotes the L^2-norm with respect to h, and the inequality follows by the Cauchy–Schwartz inequality. □

Proof of lemma 2. We give the proof when all $r_i = 1$. The extension to the general case is immediate. Recall that the matrix A for the natural updating order is $A = I - \mathrm{diag}(q_{11}^{-1}, \ldots, q_{ss}^{-1})Q = L + U$ where L is strictly lower triangular. Also we have $B_+ = (I - L)^{-1}U$. Let A_- be the starting matrix for the deterministic updating in the reverse order. We write $A_- = L_- + U_-$ where L_- is a strictly lower triangular matrix and U_- is a strictly upper triangular matrix. By definition, we have $B_- = (I - L_-)^{-1}U_-$. Let P be the permutation matrix which has elements $p_{is-i+1} = 1$ and all other elements 0. Here we have $A_- = PAP = P(L + U)P$. But note that $PUP = L_-$ and $PLP = U_-$ where L_- is a strictly lower triangular matrix and U_- is a strictly upper triangular matrix. Hence it is easy to see that B_- is similar to B_+. □

Proof of theorem 2. The basic proof is due to Amit (1995). Here we generalize the proof given by Amit to the block updating case. The proof involves the generating function of Hermite polynomials. Before we start the proof we define the following. Let D_i be an $m \times m$ matrix blocked in the same pattern as Q and with all blocks null except the (i, i)th block which is an identity matrix of order r_i, $i = 1, \ldots, s$. Let V be a symmetric square-root matrix of Q. Again we partition V according to Q. Let

$$C_i = \begin{pmatrix} I & 0 & \ldots & 0 \\ 0 & I & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{i1} & A_{i2} & \ldots & A_{is} \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & I \end{pmatrix}.$$

Note that $C_i = I - D_i + D_iA$. Let $T^{(i)}$ be the projection in $\mathbb{R}^m$ onto the vector space spanned by the rows of C_i. Also note that $C_i = I - \mathrm{diag}(0, \ldots, Q_{ii}^{-1}, \ldots, 0)Q$. Let $F_i^{(j)}$ denote the projection in R^m onto the vector space spanned by the jth row of the matrix $Q_{ii}^{-1}(Q_{i1}, \ldots, Q_{is})$. Observe that $T^{(i)}$ can be written as a direct sum of the matrices $W_i^{(j)} = I - F_i^{(j)}$, $j = 1, \ldots, r_i$, and the rows of $W_i^{(j)}$ can be taken orthogonal to the rows of the other $W_i^{(j)}$s.

Following Amit we define a multi-index $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_s)$, and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s)$ where $\boldsymbol{\alpha}_i = (\alpha_{i_1}, \ldots, \alpha_{i_{r_i}}) \in \mathbb{Z}_+^{r_i}$, and $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_{r_i}}) \in \mathbb{R}^{r_i}$. We also define $\mathbf{x}_i^{\boldsymbol{\alpha}_i} = x_{i_1}^{\alpha_{i_1}}, \ldots, x_{i_{r_i}}^{\alpha_{i_{r_i}}}$ and $\boldsymbol{\alpha}_i! = \alpha_{i_1}!, \ldots, \alpha_{i_{r_i}}!$ and $\mathbf{x}^{\boldsymbol{\alpha}} = \Pi_{i=1}^s \mathbf{x}_i^{\boldsymbol{\alpha}_i}$ and $\boldsymbol{\alpha}! = \Pi_{i=1}^s \boldsymbol{\alpha}_i!$. Let

$$\mathcal{H}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \prod_{i=1}^{s} H_{\boldsymbol{\alpha}_i}(\boldsymbol{\theta}_i)$$

where

$$H_{\alpha_i}(\theta_i) = \prod_{j=1}^{r_i} h_{\alpha_{ij}}(\theta_{ij}),$$

with h_k being the kth-order Hermite polynomial (not to be confused with the target density h).

Let $f_{\mathbf{x}}(\boldsymbol{\theta}) = \exp(\mathbf{x}'V\boldsymbol{\theta} - \mathbf{x}'\mathbf{x}/2)$ denote the generating function parameterized by $\mathbf{x}$ for the family of Hermite polynomials, i.e. $f_{\mathbf{x}}(\boldsymbol{\theta})$ can also be written as $\Sigma_{\alpha}\, \mathbf{x}^{\alpha}\, \mathcal{H}_{\alpha}(V\boldsymbol{\theta})/\boldsymbol{\alpha}!$. Let

$$P_i\{f_{\mathbf{x}}(\boldsymbol{\theta})\} = \int f_{\mathbf{x}}(\boldsymbol{\theta})\, h(\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_s)\, \mathrm{d}\boldsymbol{\theta}_i.$$

Simple computation yields

$$P_i(f_{\mathbf{x}}) = f_{T^{(i)}\mathbf{x}}. \tag{18}$$

Also it is easy to see that

$$f_{T^{(i)}\mathbf{x}} = f_{\Sigma_{j=1}^{r_i} W_i^{(j)}\mathbf{x}}.$$

Now we use the following lemma which can be proved by using Amit's result.

Lemma 5. If equation (18) holds, the rate of convergence of the randomly updated Gibbs sampler (which is RSGS but without the repetition) is the largest eigenvalue of the matrix

$$\frac{1}{s}\sum_{i=1}^{s} T^{(i)}.$$

Since C_i corresponds to the projection matrix $T^{(i)}$, and $\Sigma\, C_i = (s-1)I + A$, the RSGS convergence rate is the maximum modulus eigenvalue

$$\left(s^{-1}\sum_{i=1}^{s} C_i\right)^s = [s^{-1}\{(s-1)I + A\}]^s.$$

Observe that A has all eigenvalues real and less than 1. Moreover, since the trace of A is 0, it cannot have all negative eigenvalues. Therefore, we have $\rho_{\mathrm{RSGS}} = [s^{-1}\{s - 1 + \lambda(A)\}]^s$, where $\lambda(A)$ is the maximum non-negative eigenvalue of A. In general $\lambda(A) \neq \rho(A)$. □

A.1. *Splitting of Matrices*

We can study the various Gibbs sampling schemes by introducing the notion of splitting. Given an inverse dispersion matrix Q, the pair (M, N) such that $Q = M - N$ is called a *splitting* of Q. It is also called a *regular splitting* when M is non-singular and M^{-1}, $N \geqslant 0$, elementwise. DUGS and RSGS convergence rates depend on the eigenvalues of the matrix $M^{-1}N$ for different choices of M and N. Before we see that, we need to introduce further notation. We write $Q = D - E - F$ where D is $\mathrm{diag}(Q_{11}, \ldots, Q_{ss})$, E is the lower triangular matrix with the blocks in the lower triangle being identical with those of Q but with an opposite sign and diagonal blocks are all null and F is obtained by subtraction, i.e.

$$D = \begin{pmatrix} Q_{11} & & \dots & 0 \\ & Q_{22} & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \dots & Q_{ss} \end{pmatrix}, \qquad E = -\begin{pmatrix} 0 & 0 & \dots & 0 \\ Q_{21} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s1} & Q_{s2} & \dots & 0 \end{pmatrix},$$

$$F = -\begin{pmatrix} 0 & Q_{12} & \dots & Q_{1s} \\ 0 & 0 & \dots & Q_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Observe that the matrices L and U defined in Section 2.2 are $D^{-1}E$ and $D^{-1}F$ respectively. Taking $M = D$ and $N = E + F$ we obtain $M^{-1}N = A$ ($\equiv L + U$), and choosing $M = D - E$ and $N = F$ we obtain $M^{-1}N = B$.

To prove theorem 5 we need the following lemma.

Lemma 6. We can verify by direct calculation that

$$|\alpha E + \alpha^{-1}F - \nu D|$$

is independent of $\alpha \neq 0$ and for all ν where $Q = D - E - F$ is in the form (7).

Proof of theorem 5. The basic proof can be found in Young (1971). We have adapted it here to suit our purpose. Observe that Q is of block tridiagonal form (7). We write the matrix $Q = D - E - F$ where D, E and F are as above.

Choosing $\alpha = \pm 1$ we have $|E + F + \nu D| = |-E - F + \nu D|$. Hence, if ν is an eigenvalue of $A = D^{-1}(E + F)$ of any multiplicity then $-\nu$ is also an eigenvalue of A with the same multiplicity. Therefore it follows that $\rho(A) = \lambda(A)$. Suppose that $\delta \neq 0$ is an eigenvalue of $B = (D - E)^{-1}F$ and $\nu^2 = \delta$. Therefore, we have

$$|(D - E)^{-1}F - \delta I| = |F - \nu^2 D + \nu^2 E| = \nu^m |\nu E + \nu^{-1}F - \nu D| = 0$$

and applying lemma 6 with $\alpha = \nu$ we see that $|E + F - \nu D| = 0$. Hence ν is an eigenvalue of A.

Conversely, if ν is an eigenvalue of A and δ satisfies $\delta = \nu^2$ then δ is an eigenvalue of B by using lemma 6 similarly. Therefore, we have $\lambda(A)^2 = \rho(B)$. □

Proof of corollary 1. Using equation (6) and theorem 5 we have

$$\rho_{\text{RSGS}} = \{s^{-1}(s - 1 + \rho_{\text{DUGS}}^{1/2})\}^s.$$

For notational convenience, we write $y = \rho_{\text{RSGS}}$ and $x = \rho_{\text{DUGS}}$. At $x = 1$, $y = 1$, and, at $x = 0$, $y > 0$. We claim that, for any $0 < x < 1$, $y \neq x$. For $s = 2$, it is straightforward that $y = x$ only at $x = 1$. Therefore, assume that $s > 2$. To have $y = x$, we must have

$$sx^{1/s} - x^{1/2} = s - 1. \tag{19}$$

Note that the left-hand side of equation (19) is an increasing function which increases to the right-hand side at $x = 1$. □

Proof of corollary 3. The Q-matrix for the checker-board type of updating order is also of

the form (7). Hence by theorem 5 the DUGS rate of convergence is $\rho(A_c)^2$ where A_c is the A-matrix for this updating order. However, note that $\rho(A_c) = \rho(A)$ since A_c is obtained by permuting the rows and columns of A suitably. □

Before proving theorem 6 we state the following result from Varga (1962), p. 90.

Lemma 7. Let $Q = \Sigma^{-1}$ have all off-diagonal elements non-positive and (M_1, N_1) and (M_2, N_2) be two regular splittings of Q. If $N_2 \geqslant N_1 \geqslant 0$, elementwise, equality excluded (i.e. neither N_1 nor $N_2 - N_1$ is a null matrix), then

$$1 > \rho(M_2^{-1}N_2) > \rho(M_1^{-1}N_1) \geqslant 0.$$

Proof of theorem 6. Choose $M_1 = D - E$ and $N_1 = F$, and $M_2 = D$ and $N_2 = E + F$. Using lemma 7 we can show that $0 < \rho(B_+) < \rho(A)$. However, we note that $\rho(A) = \lambda(A) < [s^{-1}\{s - 1 + \lambda(A)\}]^s$. The equality in the previous statement is due to the non-negativity of A. □

Proof of theorem 7. RSGS convergence rates are dependent on the matrices A_Q and A_V where

$$A_Q = \begin{pmatrix} 0 & -q_{12}/q_{11} & \dots & -q_{1s}/q_{11} \\ -q_{21}/q_{22} & 0 & \dots & -q_{2s}/q_{22} \\ \vdots & \vdots & \ddots & \vdots \\ -q_{s1}/q_{ss} & -q_{s2}/q_{ss} & \dots & 0 \end{pmatrix},$$

$$A_V = \begin{pmatrix} 0 & -v_{12}/v_{11} & \dots & -v_{1s}/v_{11} \\ -v_{21}/v_{22} & 0 & \dots & -v_{2s}/v_{22} \\ \vdots & \vdots & \ddots & \vdots \\ -v_{s1}/v_{ss} & -v_{s2}/v_{ss} & \dots & 0 \end{pmatrix}.$$

Using inequality (9) we have $A_Q \geqslant A_V \geqslant 0$. Hence by the Perron–Frobenius theorem (see, for example, Seneta (1981), p. 22) we obtain $\rho_{\mathrm{RSGS}}(V) \leqslant \rho_{\mathrm{RSGS}}(Q)$. For DUGS we use the splitting idea. We first write

$$M_Q = \begin{pmatrix} q_{11} & 0 & \dots & 0 \\ q_{21} & q_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{s1} & q_{s2} & \dots & q_{ss} \end{pmatrix}, \qquad M_V = \begin{pmatrix} v_{11} & 0 & \dots & 0 \\ v_{21} & v_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_{s1} & v_{s2} & \dots & v_{ss} \end{pmatrix}.$$

Let $Q = M_Q - N_Q$ and $V = M_V - N_V$. Again by inequality (9) we have $N_Q = N_V \geqslant 0$. DUGS rates for Q and V are maximum modulus eigenvalues of $B_Q = M_Q^{-1}N_Q$ and $B_V = M_V^{-1}N_V$ respectively. The theorem will follow if we establish the claim that $M_Q^{-1} \geqslant M_V^{-1} \geqslant 0$, elementwise. The proof of the claim is by induction on the dimension of the matrix Q and V. We write $M_Q(k)$ and $M_V(k)$ instead of M_Q and M_V to emphasize their orders. It is easy to verify the claim if $s = 1$ or $s = 2$ where s is the dimension of the matrix Q. Let the claim be true for $s = k$. We have

$$M_Q(k+1) = \begin{pmatrix} M_Q(k) & 0 \\ \mathbf{c}' & q_{k+1,k+1} \end{pmatrix}, \qquad M_V(k+1) = \begin{pmatrix} M_V(k) & 0 \\ \mathbf{c}' & v_{k+1,k+1} \end{pmatrix}$$

where $\mathbf{c}$ contains the off-diagonal elements $(q_{k+1,1}, \ldots, q_{k+1,k})'$. By a standard result on the inverse of partitioned matrices, we have

$$M_Q^{-1}(k+1) = \begin{pmatrix} M_Q^{-1}(k) & 0 \\ -q_{k+1,k+1}^{-1}\mathbf{c}' M_Q^{-1}(k) & q_{k+1,k+1}^{-1} \end{pmatrix}$$

and

$$M_V^{-1}(k+1) = \begin{pmatrix} M_V^{-1}(k) & 0 \\ -v_{k+1,k+1}^{-1}\mathbf{c}' M_V^{-1}(k) & v_{k+1,k+1}^{-1} \end{pmatrix}.$$

By assumptions and by the inductive hypothesis, we have

$$-q_{k+1,k+1}^{-1}\mathbf{c}' \geqslant -v_{k+1,k+1}^{-1}\mathbf{c}',$$
$$M_Q^{-1}(k) \geqslant M_V^{-1}(k).$$

Multiplying these two together we obtain that $M_Q^{-1}(k+1) \geqslant M_V^{-1}(k+1)$. Therefore the claim holds. □

Proof of theorem 8. Let (M_1, N_1) be the regular splitting of Q with some $r_i > 1$. Recall that r_i is the size of the ith block, $1 \leqslant i \leqslant s$. Let Q_{ii} have some non-zero elements above the main diagonal. Notice that, since N_1 corresponds to the block updating, all the elements of $(N_1)_{ii}$ are 0. Let (M_2, N_2) be the splitting of Q corresponding to a componentwise update of all the components in the ith block, $1 \leqslant i \leqslant s$. Since Q_{ii} has at least one negative off-diagonal element, we have the corresponding element in $(N_2)_{ii} > 0$. Hence, we have $N_2 \geqslant N_1 \geqslant 0$, elementwise. Using lemma 7, we immediately see that the first strategy will converge faster than the second, i.e. blocking leads to faster convergence. □

REFERENCES

Amit, Y. (1995) Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Ann. Statist.*, **24**, 122–140.

———(1991) On the rates of convergence of stochastic relaxation for Gaussian and Non-Gaussian distributions. *J. Multiv. Anal.*, **38**, 82–99.

Amit, Y. and Grenander, U. (1991) Comparing sweep strategies for stochastic relaxation. *J. Multiv. Anal.*, **37**, 197–222.

Barone, P. and Frigessi, A. (1990) Improving stochastic relaxation for Gaussian random fields. *Probab. Engng Inform. Sci.*, **4**, 369–389.

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

Brooks, S. P. and Roberts, G. O. (1995) Diagnosing convergence of Markov chain Monte Carlo. *Research Report 95-12*. Statistical Laboratory, University of Cambridge, Cambridge.

Cowles, M. K. and Carlin, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Statist. Ass.*, **91**, 883–904.

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

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrization for normal linear mixed models. *Biometrika*, **82**, 479–488.

———(1996) Efficient parametrizations for generalised linear mixed models (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 165–180. Oxford: Oxford University Press.

Gilks, W. R. and Roberts, G. O. (1996) Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 89–114. London: Chapman and Hall.

Goodman, J. and Sokal, A. D. (1989) Multigrid Monte Carlo method: conceptual foundations. *Phys. Rev.* D, **40**, 2035–2071.

Green, P. J. and Han, X.-L. (1992) Metropolis methods, Gaussian proposals, and antithetic variables. *Lect. Notes Statist.*, **74**, 142–164.

Hills, S. E. and Smith, A. F. M. (1992) Parameterization issues in Bayesian inference (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 641–649. Oxford: Oxford University Press.

Horn, R. A. and Johnson, C. R. (1990) *Matrix Analysis*. Cambridge: Cambridge University Press.

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

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

Neal, R. M. (1995) Suppressing random walks in Markov chain Monte Carlo using ordered over-relaxation. *Technical Report*. Department of Statistics, University of Toronto, Toronto.

Polson, N. G. (1996) Convergence of Markov chain Monte Carlo algorithms (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 297–321. Oxford: Oxford University Press.

Raftery, A. and Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 765–776. Oxford: Oxford University Press.

Roberts, G. O. and Sahu, S. K. (1996) Rate of convergence of the Gibbs sampler by Gaussian approximation. *Research Report 96-21*. Statistical Laboratory, University of Cambridge, Cambridge.

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

Seewald, W. (1992) Discussion on Parameterization issues in Bayesian inference (by S. E. Hills and A. F. M. Smith). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 241–243. Oxford: Oxford University Press.

Seneta, E. (1981) *Non-negative Matrices and Markov Chains*. New York: Springer.

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

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

Varga, R. S. (1962) *Matrix Iterative Analysis*. Englewood Cliffs: Prentice Hall.

Whittaker, J. (1990) *Graphical Models in Applied Mathematical Multivariate Analysis*. New York: Wiley.

Winkler, G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Berlin: Springer.

Young, D. M. (1971) *Iterative Solution of Large Linear Systems*. New York: Academic Press.