Computational Statistics (2005) 20:177-196

Implementing Random Scan Gibbs Samplers[1]

Richard A. Levine[2], Zhaoxia Yu[3], William G. Hanley[4], and John J. Nitao[4]

[2] Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA
[3] Department of Statistics, Rice University, PO Box 1892, MS-138, Houston, TX 77251 USA
[4] Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94551, USA

Summary

The Gibbs sampler, being a popular routine amongst Markov chain Monte Carlo sampling methodologies, has revolutionized the application of Monte Carlo methods in statistical computing practice. The performance of the Gibbs sampler relies heavily on the choice of sweep strategy, that is, the means by which the components or blocks of the random vector $\mathbf{X}$ of interest are visited and updated. We develop an automated, adaptive algorithm for implementing the optimal sweep strategy as the Gibbs sampler traverses the sample space. The decision rules through which this strategy is chosen are based on convergence properties of the induced chain and precision of statistical inferences drawn from the generated Monte Carlo samples. As part of the development, we analytically derive closed form expressions for the decision criteria of interest and present computationally feasible implementations of the adaptive random scan Gibbs sampler via a Gaussian approximation to the target distribution. We illustrate the results and algorithms presented by using the adaptive random scan Gibbs sampler developed to sample multivariate Gaussian target distributions, and screening test and image data.

Keywords: Markov chain Monte Carlo, Gaussian approximation, adaptive algorithms, optimal sweep strategies, convergence rate, asymptotic risk

1 Introduction

Markov chain Monte Carlo methods (MCMC) have had an enormous impact on the development and application of statistical methodologies. The Gibbs sampler, as introduced to the statistics literature by Gelfand and Smith (1990), is one of the most popular implementations within this class of Monte Carlo methods. The Gibbs sampler provides the practitioner with samples of a multivariate random vector $\mathbf{X} = \{X(1), \ldots, X(d)\}'$ distributed according to some distribution $\pi(\mathbf{X})$, difficult if not impossible to sample directly, by generating random variates from each of the full conditional distributions. In particular, the algorithm iteratively visits each component X_i (or block of components of $\mathbf{X}$) and updates that component (or block) with a sample from the distribution conditioned on the most recently simulated values of all the other coordinates (blocks) excluding X_i (or excluding the given block). Under general regularity conditions, in the limit over the number of iterations, the random variates approach samples from the distribution $\pi(\mathbf{X})$.

The original MCMC method introduced by Metropolis et al. (1953) as well as the groundbreaking Gibbs sampler work of Geman and Geman (1984) uses a *random sweep strategy* whereby the selection of coordinates or blocks of $\mathbf{X}$ to update at each iteration is made at random. Convergence properties of this MCMC implementation are dependent on the frequency at which the coordinates or blocks of $\mathbf{X}$ are visited (Levine and Casella, 2003; Liu et al., 1995). In this paper we develop automated routines for optimal implementations of the random scan strategy in the Gibbs sampler.

We approach the problem of optimal implementation of the random scan Gibbs sampler from two angles: convergence of the induced chain to the stationary distribution (convergence rate) and precision in inferences drawn from the Monte Carlo sample (variance). We develop these decision criteria in closed form, however implementation of these criteria for optimizing the random scan Gibbs sampler in general is hindered by computational complexity. We thus introduce an automated algorithm for choosing the optimal random sweep strategy through the application of Gaussian approximations to the chain of interest. This algorithm is distinguished by an adaptive rule whereby the visitation strategy is constructed as the chain progresses through the sample space. The result is a routine feasible in practice that does not require much more coding hardship nor computational expense than a random scan Gibbs sampler with predetermined visitation rule, though providing at least a near-optimal sweep strategy with respect to the criteria of interest.

[1]Research by RL and ZY supported in part by a US National Science Foundation FRG grant 0139948 and a grant from Lawrence Livermore National Laboratory, Livermore, California, USA.

Mira (2001) and Besag et al. (1995) observe that implementations of MCMC samplers that optimize convergence rate are different than those that optimize with respect to a variance decision criterion due to the role played by the eigenvalues of the transition kernels in each case. Our implementation strategies lead us to the same conclusions and, consequently, a recommendation of applying different random sweep strategies during the burn-in and post-processing phases of the Gibbs sampler. We discuss and illustrate this option as part of the automated random scan Gibbs sampler proposed.

The paper unfolds as follows. In Section 2, we formalize the random scan Gibbs sampler and develop the convergence rate and variance decision criteria with respect to which we propose to choose the optimal visitation strategy for the sampler. In Section 3, we present the decision criteria for the associated chain with approximate Gaussian target distribution. In Section 4, we present the pseudocode for various implementations of the adaptive random scan Gibbs sampler. In Section 5, we illustrate the algorithms developed when the target distribution is truly Gaussian, in a screening test problem, and in a Bayesian image analysis problem. In each case, we compare our optimal random scan Gibbs sampler to the typical implementation of the random scan Gibbs sampler, namely a sampler with equal probability of visiting a given component or block of $\mathbf{X}$. We also study the effect of decision criteria, be it convergence rate or variance, on the choice of visitation strategy.

2 Random scan Gibbs sampler

Assume we wish to simulate $\mathbf{X} = \{X(1), \ldots, X(d)\}'$, a continuous random d-vector with distribution $\pi(\mathbf{X})$. Further suppose all full conditional distributions $\pi(X(i) \mid \mathbf{X}_{-i})$, $i = 1, \ldots, d$, where $\mathbf{X}_{-i} = \{X(j) : j \neq i\}$, are *available* in the sense that a sample can readily be drawn from the distributions. The Gibbs sampler is an iterative scheme which constructs a Markov chain through these easy to simulate full conditionals with $\pi(\mathbf{X})$ as the equilibrium distribution.

Formally, the Gibbs sampler updates components $X(i)$ of $\mathbf{X}$ with a sample from the distribution $\pi(X(i) \mid \mathbf{X}_{-i})$ conditioned on the current states of the other components. The order in which the components are visited may vary. The most common sweep strategy is the random scan, where the visiting order is random. In the next subsections, we will briefly detail the random scan Gibbs sampler, the convergence theory for this sampler focusing on the convergence rate, the variance of estimators derived from the random variates generated by the sampler, and the adaptive random scan Gibbs sampler.

2.1 Random scan

The random scan Gibbs sampler randomly chooses at each step the sequence in which states or the d components of $\mathbf{X}$ are visited. Let $\{\alpha_1, \ldots, \alpha_d\}$ be the

set of selection probabilities of visiting a component. Assume $0 < \alpha_i < 1$ for all $i \in \{1, \ldots, d\}$ and $\sum \alpha_i = 1$. The random scan Gibbs sampling algorithm can be stated as follows.

Algorithm 2.1 Random scan

1. `Select` an initial point $\mathbf{X}^{(0)}$ and selection probabilities $\boldsymbol{\alpha}$.
2. On the tth iteration
   a. `Randomly choose` $i \in \{1, \ldots, d\}$ with probability α_i;
   b. `Generate` $X^{(t)}(i) \sim \pi(X(i) \mid \mathbf{X}_{-i}^{(t-1)})$.
3. `Repeat` step two until reaching equilibrium.

Note that on each iteration, j, a single component $X(i)$ is updated so $\mathbf{X}_{-i}^{(t)} = \mathbf{X}_{-i}^{(t-1)}$.

The chain $\{\mathbf{X}^{(i)}\}_{i=1}^{n}$ induced by the random scan Gibbs sampler is Markov with invariant distribution π. The transition function is a weighted sum of the full conditional probabilities $P(\mathbf{Y} \mid \mathbf{X}) = \sum_{i=1}^{d} \alpha_i \pi(\mathbf{Y} \mid \mathbf{X}_{-i})$ since, as seen in step two of Algorithm 2.1, any component may be updated during a given iteration. The weights, α_i, are the probabilities a component i is visited during an iteration. The forward operator is $F_{RS} = E(\cdot \mid \mathbf{X}) = E[E(\cdot \mid \mathbf{i}, \mathbf{X}) \mid \mathbf{X}] = \sum_{i=1}^{d} \alpha_i E(\cdot \mid \mathbf{X}_{-i})$. If $A_i = E(\cdot \mid \mathbf{X}_{-i})$ represents the operation of mapping the current state $\mathbf{X}$ to its expected value upon updating component i during a given iteration, then F_{RS} is a weighted sum of these operators.

2.2 Convergence rate and maximal correlation

The Markov chain induced by the random scan Gibbs sampler visits every state infinitely often. Liu et al. (1995) show that under mild regularity conditions, the chain is reversible and ergodic with geometric convergence rate the spectral radius of the forward operator, $\|F_{RS}\|$.

These convergence issues are closely related to the ρ-mixing condition and the notion of maximal correlation in the Markov chain literature. A Markov chain $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ is *ρ-mixing* if $\rho_n = \sup_{t,s \in L_0^2(\pi)} \operatorname{corr}\{t(\mathbf{X}^{(0)}), s(\mathbf{X}^{(n)})\}$ converges to zero as n approaches infinity. Note that the ρ-mixing coefficient ρ_n is defined as the *maximal correlation* between the nth step of the chain, $\mathbf{X}^{(n)}$, and the initial state, $\mathbf{X}^{(0)}$, denoted $\gamma(\mathbf{X}^{(0)}, \mathbf{X}^{(n)})$. Therefore, elements further apart in a ρ-mixing chain have smaller maximal correlation with the value decreasing as a function of distance. See Bradley (1986) for more about ρ-mixing.

An interesting connection can be drawn between the random scan Gibbs sampler convergence rate and the ρ-mixing or maximal correlation term.

Theorem 2.1 *A Markov chain* $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ *induced by the random scan Gibbs sampling strategy has convergence rate* $\lambda_{RS} = \rho_1 = \gamma(\mathbf{X}^{(0)}, \mathbf{X}^{(1)})$. *Furthermore, for a positive integer* k, $\|F_{RS}\|^{2k} = \sup_{t \in L_0^2(\pi)} \operatorname{cov}\{t(\mathbf{X}^{(0)}), t(\mathbf{X}^{(2k)})\}$.

The proof is a consequence of the definition of the operator norm and reversibility of the random scan Gibbs chain.

Theorem 2.1 provides expressions for the convergence rate for the random scan through the maximal correlation and k-lag covariances. We note too that the theorem may be stated in terms of functions on $L^2(\pi)$ without restricting the mean to zero. The same consequences hold under this more general restriction.

2.3 Asymptotic variance

An important characterization of the Gibbs sampler, in addition to the convergence rate, is the variance of estimators derived from the generated random variates. In particular, we will need the variance as a function of the selection probabilities $\boldsymbol{\alpha}$. In this subsection, we will present the asymptotic variance of an estimator, as the number of samples generated approaches infinity.

Assume a random $d \times 1$ vector $\mathbf{X}$ with distribution $\pi(\mathbf{X})$ is generated by a random scan Gibbs sampler. The sampler generates a (Markov) chain $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ with stationary distribution π. Suppose interest lies in estimating $\mu = E_\pi\{h(\mathbf{X})\}$ where $h \in L^2(\pi)$. The natural estimator for μ is the sample mean $\hat{\mu} = \sum_{t=1}^{t} h(\mathbf{X}^{(t)})/t$ based on the samples generated by the random scan Gibbs sampler. The asymptotic variance is then

$$\lim_{t\to\infty} t \operatorname{VAR}_\pi(\hat{\mu}) = \operatorname{VAR}_\pi\{h(\mathbf{X})\} + 2 \lim_{t\to\infty} \sum_{i=1}^{t-1} \left(\frac{t-i}{t}\right) \operatorname{cov}\{h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(i)})\}. \tag{1}$$

Assume $h \in L_0^2(\pi)$ from here onward. All the preceding definitions hold for this subspace of $L^2(\pi)$. The restriction of zero mean is the same as that assumed in earlier sections and is required for future results. Levine and Casella (2003) prove the following result for the asymptotic variance (1). Define $\eta(h(\mathbf{X}), i_j) = E\left(\cdots E(E(h(\mathbf{X}) \mid \mathbf{X}_{-i_1}) \mid \mathbf{X}_{-i_2}) \mid \cdots \mid \mathbf{X}_{-i_j}\right)$ and $Q_t(\boldsymbol{\alpha}) = \operatorname{cov}\{h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(t)})\}$ where $\{i_1, \ldots i_j\}$ is a set of integers $i_q \in \{1, \ldots, d\}$; $q = 1, \ldots, j$. These t-lag covariances can be expressed as follows.

Theorem 2.2 (Levine and Casella, 2003) *For the Markov chain* $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ *induced by the random scan Gibbs sampler, the* t*-lag autocovariances with* $r = t/2$ *and* $s = (t+1)/2$ *can be represented by*

$$Q_t(\boldsymbol{\alpha}) = \sum_{i_1,\ldots,i_r=1}^{d} (\alpha_{i_1} \cdots \alpha_{i_r})^2 \operatorname{VAR}_\pi\{\eta(\mathrm{h}(\mathbf{X}), \mathrm{i_r})\}$$

$$+ \sum_{i_1,\ldots,i_r=1}^{d} \sum_{m_1,\ldots,m_r=1;\, i_j \neq m_j, \forall j}^{d} \alpha_{i_1} \cdots \alpha_{i_r} \alpha_{m_1} \cdots \alpha_{m_r}$$
$$\times \operatorname{cov}\{\eta(\mathrm{h}(\mathbf{X}), \mathrm{i_r}), \eta(\mathrm{h}(\mathbf{X}), \mathrm{m_r})\}$$

if t is even and

$$Q_t(\boldsymbol{\alpha}) = \sum_{i_1,\ldots,i_s=1}^{d} (\alpha_{i_1} \cdots \alpha_{i_{s-1}})^2 \alpha_{i_s} \mathrm{VAR}_\pi\{\eta(h(\mathbf{X}), i_s)\}$$
$$+ \sum_{i_s=1}^{d} \alpha_{i_s} \sum_{i_1,\ldots,i_{s-1}=1}^{d} \sum_{m_1,\ldots,m_{s-1}=1;\, i_j \neq m_j, \forall j}^{d} \alpha_{i_1} \cdots \alpha_{i_{s-1}} \alpha_{m_1} \cdots \alpha_{m_{s-1}}$$
$$\times \operatorname{cov}\{\eta(\mathrm{h}(\mathbf{X}), \mathrm{i_s}), \eta(\mathrm{h}(\mathbf{X}), \mathrm{m_s})\}$$

if t is odd.

2.4 Adaptive random scan Gibbs sampler

The random scan Gibbs sampler is heavily dependent on the choice of the selection probabilities $\boldsymbol{\alpha}$. As seen in previous subsections, the convergence rate of the induced Markov chain and variances of estimators based on random variates from this chain depend on $\boldsymbol{\alpha}$. How may we construct a general, automated routine to choose the optimal random scan Gibbs sampler in terms of the selection probabilities and the decision criterion, be it convergence rate, precision, or some other measure, over which we select these probabilities? Levine and Casella (2003) make some headway in answering this question through their adaptive random scan Gibbs sampler. For generality of the approach, let $R(\boldsymbol{\alpha}, h)$ denote the criteria over which we will optimize to find the selection probabilities where the letter R signifies "risk function" from a decision theoretic perspective. The adaptive random scan Gibbs sampler may be written as follows.

Algorithm 2.2 Adaptive random scan Gibbs sampler

1. `Select` an initial point $\mathbf{X}^{(0)}$ and selection probabilities $\boldsymbol{\alpha}^{(0)}$.
2. On the tth iteration
   a. `Randomly choose` $i \in \{1, \ldots, d\}$ with probability $\alpha_i^{(t-1)}$.
   b. `Choose` $\boldsymbol{\alpha}^{(t)}$ to minimize $R^{(t)}(\boldsymbol{\alpha}, h|\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})$.
   c. `Generate` $X^{(t)}(i) \sim \pi(X_i|\mathbf{X}_{-i}^{(t-1)})$ and `set` $\mathbf{X}_{-i}^{(t)} = \mathbf{X}_{-i}^{(t-1)}$.
3. `Repeat` step two until reaching equilibrium.

In words, the adaptive random scan Gibbs sampler in Algorithm 2.2 estimates the risk function $R(\boldsymbol{\alpha}, h)$ and chooses $\boldsymbol{\alpha}$ each iteration using previous Gibbs samples generated. In this sense, the random scan learns from and adapts to the random variates as the Markov chain traverses the sample space.

We propose to study two alternative objective functions or optimization criteria $R(\boldsymbol{\alpha}, h)$ when implementing adaptive random scan Gibbs samplers: convergence rate and asymptotic variance. The convergence rate and asymptotic variance describe speed of convergence to the stationary distribution and precision of estimates from the Monte Carlo samples respectively. Mira (2001) and Besag et al. (1995) argue that the convergence rate is of importance for determining an appropriate burn-in of the chain. Variance considerations are relevant for posterior inferences following the burn-in period. Thus we may use both of these criteria during different parts, pre and post burn-in, of our Gibbs sampling, for choosing $\boldsymbol{\alpha}$.

3 Gaussian target distribution

In this section, we will characterize the convergence rate and asymptotic variance of Sections 2.2 and 2.3 in the case of Gaussian target distributions. The motivation is that in many situations, the Gibbs sampler of interest may be well approximated by a Gibbs chain with an appropriate Gaussian target distribution. We may thus study the convergence rate and asymptotic variance of the induced Markov chain of interest through the approximate chain with Gaussian stationary distribution. The first subsection discusses the Gaussian approximation theory. The remaining subsections derive the convergence rates and asymptotic variance for a random scan Gibbs sampler for a Gaussian target distribution.

3.1 Gaussian approximation

Normal approximations to posterior distributions has a long history dating back to the work of Laplace in the 18th and early 19th centuries (see Chapter 5.3 of Bernardo and Smith, 1994, for relevant historical references from the time of Laplace to the present). The literature focuses on regularity conditions under which such an approximation holds. We require conditions under which not only the posterior distribution of interest is approximately normal, but the properties of the Markov chain induced by our Gibbs sampler is well approximated by the convergence properties of the Markov chain Monte Carlo for sampling from the Gaussian approximation. We first introduce some notation in order to formally clarify this idea.

Assume interest lies in a d-dimensional parameter $\mathbf{X}$ characterizing a probability model $p(\mathbf{Y}|\mathbf{X})$ from which we observe the n data points $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. We consider the case of implementing a Gibbs sampler to generate random variates from the posterior distribution of interest $p_n(\mathbf{X}|\mathbf{Y})$. The Gibbs sam-

pler induces a Markov chain with stationary distribution $p(\mathbf{X}|\mathbf{Y})$ by successively sampling the full conditional distributions $p_n(X(i)|\mathbf{X}_{-i}, \mathbf{Y})$. We subscript the distributions with an n to signify the dependence on a data set of size n.

Roberts and Sahu (2001) provide the conditions under which we may apply a Gaussian approximation in our setting. In particular, under regularity conditions as presented in Bernardo and Smith (1994; page 289), if for all i the full conditional distributions $p_n(X(i)|\mathbf{X}_{-i}, \mathbf{Y})$ converge pointwise and in L^1 to normal distributions, then the posterior distribution $p_n(\boldsymbol{\theta}|\mathbf{x})$ is approximately normal. Furthermore, the convergence properties of the Gibbs sampler with stationary distribution $p_n(\mathbf{X}|\mathbf{Y})$ may be approximated by the Gibbs sampler with the approximating normal distribution as the stationary distribution.

The regularity conditions of Bernardo and Smith (1994) and Roberts and Sahu (2001) place restrictions on the "steepness," "smoothness," and "concentration" of the posterior distribution. That is, in a neighborhood of the posterior mean, the function $p_n(\mathbf{X}|\mathbf{Y})$ becomes highly peaked and behaves like a normal density. Outside this neighborhood, the probability mass must be negligible. We note too that many other regularity conditions exist for application of a asymptotic Gaussian distribution. See for example Yee et al. (2002) for references and alternative sufficient conditions.

3.2 Gaussian case: Random scan Gibbs sampler

Let $\pi(\mathbf{X})$ be a d-dimensional normal density $N(\mathbf{0}, \boldsymbol{\Sigma})$ with dispersion $\boldsymbol{\Sigma}$ and, without loss of generality, mean $\mathbf{0}$. The conditional distributions required by the Gibbs samplers are easily shown (Amit and Grenander, 1991) to be

$$X(i) \mid \mathbf{X}_{-i} \sim N(-\sum_{j \neq i} \frac{r_{ij}}{r_{ii}} X(j),\ 1/r_{ii}) \tag{2}$$

where r_{ij} is the (i,j)th element of $\mathbf{R} = \boldsymbol{\Sigma}^{-1}$.

The Gibbs sampler discussed in Algorithm 2.1 updates each component of $\mathbf{X}$ by generating a random variate from the conditional distributions $X(i) \mid \mathbf{X}_{-i}$, $i = 1, \ldots, d$. Let $\mathbf{D}_i = \text{diag}(0, \ldots, 0, 1/r_{ii}, 0, \ldots, 0)$, a $d \times d$ diagonal matrix with all entries zero except for the ith diagonal element which is $1/r_{ii}$. The transition kernel and convergence of the random scan Gibbs Markov chain is described by the following theorem, a result easily proved using (2) and the theoretical construct of Liu et al. (1995). Denote the spectral radius of the transition operator F by $\rho(F)$.

Theorem 3.1 *The Markov chains $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ induced by the random scan Gibbs sampler has normal transition distribution $P(\mathbf{X}^{(k)} \mid \mathbf{X}^{(k-1)})$ with mean*

$$\mu_{RS} = \sum_{i_1, \ldots, i_d = 1}^{d} \prod_{j=1}^{d} \alpha_{i_j}\ (I - \mathbf{D}_{i_1}\mathbf{R}) \cdots (\mathbf{I} - \mathbf{D}_{i_d}\mathbf{R})\ \mathbf{X}^{(k-1)} \tag{3}$$

and dispersion $\Sigma - \mu_{RS}\Sigma\mu'_{RS}$. *Under these transition laws, the Markov chain converges geometrically to the stationary distribution* $N(\mathbf{0}, \Sigma)$ *at rate* $\rho(F_{RS})$ *where* F_{RS} *corresponds to the transition operator for the random scan scheme.*

3.3 Gaussian case: Convergence rate via maximal correlation

Theorem 3.1 proves geometric convergence of the Markov chain induced by the random scan Gibbs sampler, but does not provide analytic expressions for the convergence rates. The mixing properties and maximal correlation discussed in Section 2.2 establish analytic representations for these rates.

The maximal correlation

$$\gamma(\mathbf{X}^{(0)}, \mathbf{X}^{(n)}) = \sup_{t,s \in L_0^2(\pi)} \operatorname{corr}\{t(\mathbf{X}^{(0)}), s(\mathbf{X}^{(n)})\}$$

is easily determined under normality for the random scan Gibbs sampler after noting the following. First, Sarmanov and Zaharov (1960) show that the supremum over t and s is attained by linear functions when the distribution of $\mathbf{X}^{(0)}$ and $\mathbf{X}^{(n)}$ is normal. Second, the covariance of two linear functions is concisely determined by the following lemma.

Lemma 3.1 *If* $t(\mathbf{X}) = \mathbf{a}'\mathbf{X}$ *and* $s(\mathbf{X}) = \mathbf{b}'\mathbf{X}$ *for some* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, *then for a Markov chain* $\{\mathbf{X}^{(i)}\}_{i=0}^n$ *with transition operator* F *and stationary distribution* $\pi = N(\mathbf{0}, \Sigma)$, $cov_\pi\{t(\mathbf{X}^{(0)}), s(\mathbf{X}^{(n)})\} = \mathbf{a}'\Sigma(F^n)'\mathbf{b}$.

Proof: Note that $E_\pi\{t(\mathbf{X})\} = E_\pi\{\mathbf{a}'\mathbf{X}\} = 0$ and $E_\pi\{s(\mathbf{X})\} = E_\pi\{\mathbf{b}'\mathbf{X}\} = 0$. Hence $\operatorname{cov}_\pi\{t(\mathbf{X}^{(0)}), s(\mathbf{X}^{(n)})\} = E_\pi\{t(\mathbf{X}^{(0)}) \cdot s(\mathbf{X}^{(n)})\} = E_\pi\{t(\mathbf{X}^{(0)}) \cdot F^n s(\mathbf{X}^{(0)})\} = E_\pi\{\mathbf{a}'\mathbf{X}^{(0)} \cdot (\mathbf{X}^{(0)})'(F^n)'\mathbf{b}\} = \mathbf{a}'\Sigma(F^n)'\mathbf{b}$ where the second equality follows from the definitions of the operator F and the iterative expected value (see Liu et al., 1995).

□

The one-lag maximal correlation for the random scan Gibbs sampler, and thus the convergence rate, is as follows. Let $\lambda(A)$ denote the largest eigenvalue of the matrix A.

Theorem 3.2 *The convergence rate for the random scan Gibbs sampler with transition operator* F_{RS} *is* $\lambda_{RS}{}^2 = \lambda(\Sigma^{-1/2}F_{RS}\Sigma F_{RS}{}^T\Sigma^{-1/2})$.

This result follows directly from a constrained maximization over the one-lag covariance term in Theorem 2.1 using Lemma 3.1. Note that the expression for λ_{RS} in Theorem 3.2 is independent of the iteration count. This fact is a consequence of reversibility of the Markov chain induced under the random sampling scheme.

Recall from Theorem 2.1, for linear functions $t(\mathbf{X}) = \mathbf{a}'\mathbf{X}$ where $\mathbf{a} \in \mathbb{R}^d$, $\lambda_{RS}{}^2 = \sup_{\mathbf{a}'\Sigma\mathbf{a}=1} \operatorname{corr}(\mathbf{a}'\mathbf{X}^{(0)}, \mathbf{a}'\mathbf{X}^{(2)})$, a quantity based on one coefficient $\mathbf{a}$.

This expression may provide an alternative specification for λ_{RS} than the one derived in Theorem 3.2 where the maximum covariance with respect to two coefficients $\mathbf{a}$ and $\mathbf{b}$ is utilized. However, the following theorem, proven analogously to Theorem 3.2, shows no additional information is attained from this formulation.

Theorem 3.3 *If* $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ *is a Markov chain induced by the random scan Gibbs sampler with transition operator* F_{RS}*, then for some* $\mathbf{a} \in I\!R^d$ $\sup_{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}=1} \mathrm{corr}(\mathbf{a}'\mathbf{X}^{(0)}, \mathbf{a}'\mathbf{X}^{(2)})^2 = \lambda(\boldsymbol{\Sigma}^{-1/2} F_{RS} \boldsymbol{\Sigma} F_{RS}{}^T \boldsymbol{\Sigma}^{-1/2})$.

3.4 Gaussian case: Convergence rate via spectral radii

The convergence rate expression obtained in Theorem 3.2 is still difficult to compute and study due to the reliance on the currently unspecified operator F_{RS}. However, the maximal correlations are attained by linear functions corresponding to the eigenspace of the largest eigenvalue of the operator F_{RS}. Therefore, computations utilizing these operators can be restricted to the space of linear functions. For example, the random scan operator updates $\mathbf{X}^{(0)}$ to obtain a new state $\mathbf{X}^{(1)}$, say, through $\mathbf{X}^{(1)} = F_{RS}\, t(\mathbf{X}^{(0)}) = \mathbf{a}' \sum_{i_1,\ldots,i_d=1}^{d} \prod_{j=1}^{d} \alpha_{i_j}\ (\mathbf{I} - \mathbf{D}_{i_1}\mathbf{R}) \cdots (\mathbf{I} - \mathbf{D}_{i_d}\mathbf{R})\ \mathbf{X}^{(0)} = \mathbf{a}' \mu_{RS} \mathbf{X}^{(0)}$. Since the maximum eigenvalue of F_{RS} occurs in the space of linear functions, the convergence rate may then be written $\lambda_{RS} = \rho(\sum_{i_1,\ldots,i_d=1}^{d} \prod_{j=1}^{d} \alpha_{i_j}\ (\mathbf{I} - \mathbf{D}_{i_1}\mathbf{R}) \cdots (\mathbf{I} - \mathbf{D}_{i_d}\mathbf{R}))$. See Amit and Grenander (1991) and Roberts and Sahu (1997) for more formal proofs of this convergence rate. This expression may be simplified through matrix manipulations. Amit (1996) and Roberts and Sahu (1997) find the convergence rate for the random scan under equal selection probabilities, $\alpha_i = 1/d$ for all i, is

$$\lambda_{RS} = \rho\{(I - \frac{1}{d}\mathbf{SQ})^d\} \tag{4}$$

where $\mathbf{S} = \mathrm{diag}(1/r_{11}, \ldots, 1/r_{dd})$ and $\mathbf{I}$ is a $d \times d$ identity matrix. This result may be generalized to random scans with arbitrary selection probabilities.

Theorem 3.4 *If* $\boldsymbol{\Psi} = diag(\alpha_1, \ldots, \alpha_d)$, $\sum_{i=1}^{d} \alpha_i = 1$, *contains the selection probabilities for the random scan Gibbs sampler, then*

$$\lambda_{RS} = \rho\{(I - \boldsymbol{\Psi}\mathbf{SR})^d\}. \tag{5}$$

Proof: The transition operator for the random scan for a single update is, according to (2), $\sum_{i=1}^{d} \alpha_i\ (\mathbf{I} - \mathbf{D}_i\mathbf{R}) = \mathbf{I} - \sum_{i=1}^{d} \alpha_i \mathbf{D}_i\mathbf{R} = I - \boldsymbol{\Psi} \sum_{i=1}^{d} \mathbf{D}_i\mathbf{R} = I - \boldsymbol{\Psi}\mathbf{SR}$. Since there are d updates each iteration, the transition operator for the random Gibbs scan restricted to linear functions is $(\mathbf{I} - \boldsymbol{\Psi}\mathbf{SR})^d$. □

The expression for λ_{RS} in Theorem 3.4 may be more easily manipulated than the formulation in Theorem 3.2 as dependence on α is isolated to the single diagonal matrix $\boldsymbol{\Psi}$. Its application will be seen later.

3.5 Gaussian case: Asymptotic variance

Consider a d-dimensional Gaussian random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and dispersion $\boldsymbol{\Sigma}$ with $\mathbf{R} = \boldsymbol{\Sigma}^{-1}$. Let μ_i denote the mean for the ith component. We denote the $(i,j)th$ element of $\boldsymbol{\Sigma}$ and $\mathbf{R}$ by s_{ij} and r_{ij} respectively. Let $\mathbf{A} = \mathbf{I} - diag(r_{11}^{-1}, \cdots, r_{dd}^{-1})\mathbf{R}$ with $(i,j)th$ element A_{ij}. Under this Gaussian distribution, we may derive the covariance lags $Q_k(\boldsymbol{\alpha})$. In this section, we will derive the first four covariance lags, $k = 1, 2, 3, 4$, under the assumption that interest lies in a linear function $h(\mathbf{X}) = \mathbf{l}'\mathbf{X}$ with d-vector of coefficients $\mathbf{l}$. As we will discuss in Section 5, an order four expansion of $Q(\boldsymbol{\alpha})$ reasonably balances the tradeoff of approximation accuracy and computational cost in calculating the asymptotic variances of Section 2.3. The choice of linear functions $h(\mathbf{X})$ is made partially for simplicity and partially due to the optimality results of Kagan et al. (1973) for Gaussian distributions, which we will exploit in later sections.

The first four covariance lags follow directly from Theorem 2.2. Using the notation of Section 2.3, we have that

$$\begin{aligned} Q_1(\boldsymbol{\alpha}) &= \sum_{i=1}^{d} \alpha_i \mathrm{VAR}[\eta(\mathrm{h(x)}, \mathrm{i})] \\ &= \sum_{i=1}^{d} \alpha_i \mathbf{l}'\{\mathbf{I} - \mathbf{e}_i\mathbf{e}_i'(\mathbf{I} - \mathbf{A})\}'\boldsymbol{\Sigma}\{\mathbf{I} - \mathbf{e}_i\mathbf{e}_i'(\mathbf{I} - \mathbf{A})\}\mathbf{l} \end{aligned}$$

where $\mathbf{e}_i = (0, \cdots 0, 1, 0, \cdots, 0)'$ with the ith element equal to one. Additionally,

$$\begin{aligned} Q_2(\boldsymbol{\alpha}) &= \mathrm{VAR_r}\left\{ \sum_{\mathrm{i_1,i_2,\cdots,i_r=1}}^{\mathrm{d}} \eta(\mathrm{h(X)}, \mathrm{i_r}) \right\} \\ &= \mathrm{VAR}\left\{ \sum_{\mathrm{i}}^{\mathrm{d}} \alpha_{\mathrm{i}} \eta(\mathrm{h(X)}, \mathrm{i}) \right\} \\ &= \{\boldsymbol{\alpha}'\mathbf{1}\mathbf{1}' - \boldsymbol{\alpha}' diag(\mathbf{l})(\mathbf{I} - \mathbf{A})\}\boldsymbol{\Sigma}\{\boldsymbol{\alpha}'\mathbf{1}\mathbf{1}' - \boldsymbol{\alpha}' diag(\mathbf{l})(\mathbf{I} - \mathbf{A})\}' \end{aligned}$$

where $\mathbf{1}$ is a d-vector of ones.

The odd and even covariance lags are now easily computed by expanding out the first two covariance lags. In particular, we have that

$$\begin{aligned} Q_3(\boldsymbol{\alpha}) &= \sum_{i_2=1}^{d} \mathrm{VAR}\left\{ \sum_{\mathrm{i_1}=1}^{\mathrm{d}} \alpha_{\mathrm{i_1}} \eta(\mathrm{h(x)}, \mathrm{i_2}) \right\} \\ &= \sum_{j=1}^{d} \alpha_j [l'\{I + (\alpha_j \mathbf{e}_j\mathbf{e}_j' - diag(\boldsymbol{\alpha}) - \mathbf{e}_j\mathbf{e}_j')(\mathbf{I} - \mathbf{A})\}] \\ &\quad \boldsymbol{\Sigma}[\mathbf{l}'\{I + (\alpha_j \mathbf{e}_j\mathbf{e}_j' - diag(\boldsymbol{\alpha}) - \mathbf{e}_j\mathbf{e}_j')(\mathbf{I} - \mathbf{A})\}]' \end{aligned}$$

and

$$\begin{aligned} Q_4(\boldsymbol{\alpha}) &= \mathrm{VAR}\left\{\sum_{i_1=1}^{d}\sum_{i_2=1}^{d}\alpha_{i_1}\alpha_{i_2}\eta(h(x),i_2)\right\} \\ &= [l'\{\mathbf{I}+diag(\boldsymbol{\alpha}^2-2\boldsymbol{\alpha})(\mathbf{I}-\mathbf{A})-diag(\boldsymbol{\alpha})\mathbf{A}diag(\boldsymbol{\alpha})(\mathbf{I}-\mathbf{A})\}]\boldsymbol{\Sigma} \\ &\quad [l'\{\mathbf{I}+diag(\boldsymbol{\alpha}^2-2\boldsymbol{\alpha})(\mathbf{I}-\mathbf{A})-diag(\boldsymbol{\alpha})\mathbf{A}diag(\boldsymbol{\alpha})(\mathbf{I}-\mathbf{A})\}]' \end{aligned}$$

where $\boldsymbol{\alpha}^2$ represents the squaring of each element of $\boldsymbol{\alpha}$.

In the following sections, we will approximate $Q(\boldsymbol{\alpha}) \approx \sum_{k=1}^{4} Q_k(\boldsymbol{\alpha})$ and use this function as a decision criterion to choose optimal selection probabilities in the adaptive scan.

4 Implementing the adaptive random scan Gibbs sampler

Sections 3.4 and 3.5 show that under Gaussian target distributions, the convergence rate of the induced Markov chain is available in closed form and precision measures such as the asymptotic variance (1) are computationally feasible. If the Gibbs sampler of interest may be reasonably approximated by a Markov chain with a Gaussian stationary distribution, we may use these measures to, in some sense, optimally choose selection probabilities for implementing the random scan Gibbs sampler as well as study convergence properties of the induced Markov chain. In this section, we address implementation of the random scan Gibbs sampler through development of an adaptive random scan Gibbs sampler which chooses selection probabilities $\boldsymbol{\alpha}$ "on-the-fly." Such an algorithm automates the task of implementing the random scan Gibbs sampler and, as far as the asymptotic results of Section 3.1 takes us, will find the optimal random scan Gibbs sampler with respect to the decision criteria of choice.

The mean and dispersion of the approximating normal distribution are the mode and Hessian, evaluated at the mode, of the target distribution $\pi(\mathbf{X})$ respectively. Of course, in practice, the posterior mode is typically not available to us, thus the need for MCMC estimation routines in the first place. The EM algorithm, which has been shown to have an intimate connection with the Gibbs sampler (Sahu and Roberts, 1999), presents quick methods for computing these quantities however. We thus propose the following strategy for finding the optimal random scan Gibbs sampler. In all subsequent algorithms, we will suppress dependence of conditional distributions on the observed data $\mathbf{Y}$ for clarity of presentation.

Algorithm 4.1 Adaptive random scan Gibbs sampler using Gaussian approximations

1. `Select` an initial point $\mathbf{X}^{(0)}$ and selection probabilities $\boldsymbol{\alpha}^{(0)}$.

2. On the tth iteration

   a. Choose $\boldsymbol{\alpha}^{(t)}$ to minimize $R^{(t)}(\boldsymbol{\alpha}, h|\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}, N(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}))$.

   b. Randomly choose $i \in \{1, \ldots, d\}$ with probability $\alpha_i^{(t)}$.

   c. Generate $X^{(t)}(i) \sim \pi(X(i)|\mathbf{X}_{-i}^{(t-1)})$ and set $\mathbf{X}_{-i}^{(t)} = \mathbf{X}_{-i}^{(t-1)}$.

3. Repeat step two until reaching equilibrium.

The conditioning arguments in the risk function in step 2a implies that the risk function is constructed under the *approximating Gaussian* distribution whose parameters, the mean vector $\boldsymbol{\mu}^{(t)}$ and the dispersion $\boldsymbol{\Sigma}^{(t)}$, are computed at the tth iteration using perhaps all random variates $\{\mathbf{X}^{(j)}\}_j$ generated to that point.

Two simple modifications of Algorithm 4.1 lead to practical implementations of the routine. For the first implementation, note that, following Roberts and Sahu (2001), we can estimate the asymptotic Gaussian distribution independent of the random variates generated by the Gibbs sampler. In such a scenario, we may modify the random scan Gibbs sampler as follows.

Algorithm 4.2 Adaptive random scan Gibbs sampler via Roberts and Sahu (2001)

1. Obtain the posterior mode, $\mathbf{X}^*$ using the EM algorithm.

2. Obtain the Hessian matrix $H = -\partial^2 \log\{\pi(\mathbf{X})\}/\partial\mathbf{X} \cdot \partial\mathbf{X}|_{\mathbf{X}=\mathbf{X}^*}$.

3. Select an initial point $\mathbf{X}^{(0)}$ and selection probabilities $\boldsymbol{\alpha}^{(0)}$.

4. On the tth iteration

   a. Choose $\boldsymbol{\alpha}^{(t)}$ to minimize $R^{(t)}(\boldsymbol{\alpha}, h|\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}, N(\mathbf{X}^*, H^{-1}))$.

   b. Randomly choose $i \in \{1, \ldots, d\}$ with probability $\alpha_i^{(t)}$.

   c. Generate $X^{(t)}(i) \sim \pi(X(i)|\mathbf{X}_{-i}^{(t-1)})$ and set $\mathbf{X}_{-i}^{(t)} = \mathbf{X}_{-i}^{(t-1)}$.

5. Repeat step two until reaching equilibrium.

For the second implementation, note that we do not need to perform an adaptive random scan if the objective function or decision criteria $R^{(t)}(\boldsymbol{\alpha}, h|\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}, N(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}))$ is computable independent of the random variates $\{\mathbf{X}^{(j)}\}_j$. The idea is that we use the Gibbs sampler with approximate normal stationary distribution, for which we can compute the decision criteria, to choose the selection probabilities and then implement the Gibbs sampler of interest using these pre-selected selection probabilities. In such a scenario, we may modify the random scan Gibbs sampler as follows.

Algorithm 4.3 Random scan via a Gaussian approximation

1. `Obtain` the posterior mode, $\mathbf{X}^*$ using the EM algorithm.
2. `Obtain` the Hessian matrix $H = -\partial^2 \log\{\pi(\mathbf{X})\}/\partial\mathbf{X} \cdot \partial\boldsymbol{\theta}|_{\mathbf{X}=\mathbf{X}^*}$.
3. `Choose` $\boldsymbol{\alpha}^{(t)}$ to minimize $R(\boldsymbol{\alpha}, h|N(\mathbf{X}^*, H^{-1}))$.
4. `Select` an initial point $\mathbf{X}^{(0)}$.
5. On the tth iteration
   a. `Randomly choose` $i \in \{1, \ldots, d\}$ with probability α_i;
   b. `Generate` $X^{(t)}(i) \sim \pi(X(i)|\mathbf{X}_{-i}^{(t-1)})$ and `set` $\mathbf{X}_{-i}^{(t)} = \mathbf{X}_{-i}^{(t-1)}$.
6. `Repeat` step two until reaching equilibrium.

5 Examples

In this section, we illustrate our results and adaptive random scan Gibbs sampler on three examples: sampling from Gaussian target distributions, studying a screening test problem, and performing Bayesian image analysis. In each case we implement a random scan with equal selection probabilities (ERS), a random scan with selection probabilities chosen to optimize the convergence rate (RRS), and a random scan with selection probabilities chosen to optimize the asymptotic variance (VRS). ERS is the sweep strategy typically used in applications of the random scan Gibbs sampler. The implementations of VRS assume a linear function of interest as in Section 3.5. This assumption is reasonable from an optimality viewpoint as linear functions are optimal among unbiased estimators of linear estimands under Gaussian target distributions (Kagan et al., 1973, Section 7.6). However, we may lift this assumption to consider alternative functions of interest. See Levine et al. (2003).

In each example, we compare the three scan strategies, ERS, RRS, and VRS, with respect to convergence rate and asymptotic variance. We also put forth the idea of implementing RRS during the burn-in phase of the Gibbs sampler and VRS during the post-processing (statistical inference) phase of the sampler.

5.1 Gaussian target distributions

5.1.1 Bivariate Gaussian variates

Consider sampling from a bivariate normal distribution with mean vector zero, variances σ_1^2 and σ_2^2, and covariance τ. By Theorem 3.4, the random scan Gibbs sampler has convergence rate $\lambda_{RS} = \alpha_1^2/2 + \alpha_2^2/2 + \alpha_1\alpha_2\tau^2/(\sigma_1^2\sigma_2^2) +$

$0.5\sigma_1^{-2}\sigma_2^{-2}\cdot(\alpha_1+\alpha_2-2)\,\sigma_1^{3/2}\sigma_2^{3/2}\cdot\{4\alpha_1\alpha_2\tau^2+\sigma_1^2\sigma_2^2(\alpha_1+\alpha_2)\}^{1/2}$. The best random scan can be calculated by minimizing λ_{RS} over α_1 and α_2 under the constraint $\alpha_1+\alpha_2=1$. The resulting optimal random scan is the strategy with equal selection probabilities, $\alpha_1=\alpha_2=0.5$. When the selection probabilities are equal, $\lambda_{RS}=0.25\{1+\tau/(\sigma_X\sigma_Y)\}^2=0.25(1+\lambda_{SS}+2\sqrt{\lambda_{SS}})$.

5.1.2 Independent Gaussian variates

Consider sampling from a d-variate normal distribution with independent components, mean vector zero, and variances $\sigma_1^2,\sigma_2^2,\ldots,\sigma_d^2$. Since the components are independent, the Gibbs sampler obtains samples directly from the marginal distribution. The random scan, the way we are implementing it, is penalized for not guaranteeing an update of every component each iteration. This statement can be shown as follows. Note that $\lambda_{RS}=\rho\{(\mathbf{I}-\mathbf{\Psi SR})^d\}=\rho\{(\mathbf{I}-\mathbf{\Psi})^d\}$ since $\mathbf{R}=\mathbf{\Sigma}^{-1}=\mathbf{S}^{-1}$ in this situation. Thus, the best random scheme has rate $\lambda_{RS}=\max_{\boldsymbol{\alpha}}\{(1-\alpha_i)^d:\alpha_i\in(0,1),\sum\alpha_i=1\}=d^{-d}>0$.

5.1.3 Equicorrelated Gaussian variates

Consider a trivarate Gaussian variable $\mathbf{X}=\{X(1),X(2),X(3)\}'$ with zero mean vector, variances $(100,10,1)$, and covariance -0.125. We are interested in drawing inferences about $h(\mathbf{X})=\sum_{i=1}^3 X(i)$. In this section, we find the set of selection probabilities that induces a random scan Gibbs sampler with the smallest convergence rate as well as the set that induces a random scan Gibbs sampler with the smallest asymptotic variance. Unlike the derivations in last two subsections, closed form manageable expressions of the spectral radius in equation (4) and the asymptotic variance (1) are unavailable for analytic optimization. Nonetheless, we may optimize the convergence rate numerically over $\boldsymbol{\alpha}$.

The selection probabilities that minimize the convergence rate is $\alpha_{RRS}=(0.32,0.33,0.34)$ for the RRS strategy. The selection probability that minimizes the asymptotic risk is $\alpha_{VRS}=(0.80,0.10,0.10)$. We take the asymptotic risk as put forth in Section 3.5 being a four term expansion of $Q(\boldsymbol{\alpha})$. Our simulations indicate that the remainder term beyond this fourth term is negligible towards optimizing the random scan Gibbs sampler. We thus find that the convergence rate under VRS, RRS, and ERS is 0.74, 0.31, and 0.32 respectively. The asymptotic risk under VRS, RRS, and ERS is 226.37, 471.57, and 463.45 respectively.

Note that the RRS and ERS algorithms produce similar scans. We find this result true in all simulations and applications we have performed using the optimal random scan with respect to a convergence rate criterion. In fact, we find and postulate that $(1+\lambda_{ERS})(1-\lambda_{RRS})/\{(1-\lambda_{ERS})(1+\lambda_{RRS})\}<1.5$. While this result is based on empirical evidence, it suggests that the ERS is sufficient for optimizing the convergence rate in that the gain in convergence speed does not outweigh the cost in finding RRS. However,

the asymptotic variance can be substantially reduced, even in this simple example, by applying VRS. In this case, the risk reduction between VRS and ERS is about 52%.

Our proposal in practice is then to burn-in the Gibbs sampler using ERS, in which the goal is convergence to the stationary distribution in as few iterations as possible. After burn-in, when precise inferences are of interest, apply VRS to minimize the variance of estimators of interest.

We next consider a Gaussian target distribution with equicorrelated variates studied by Roberts and Sahu (1997), in our case illustrating VRS under four dispersion structures. We assume that the mean vector in each case is zero and define $\mathbf{B} = diag\{10, 5, ones(1,8)\}$, $\mathbf{C} = diag\{100, 10, ones(1,8)\}$, $\delta = \{0, 0.05, 0.10, \ldots, 0.95\}$, $k = \{-0.95, 0.90, \ldots, 0\}$, and J as a matrix of all ones. The four dispersion matrices we consider are $\mathbf{\Sigma}_1 = \mathbf{B} - \mathbf{J}/(10+\delta)$, $\mathbf{\Sigma}_2 = \mathbf{B} + k\mathbf{J}/(10+1)$, $\mathbf{\Sigma}_3 = \mathbf{C} - \mathbf{J}/(10+\delta)$, and $\mathbf{\Sigma}_4 = \mathbf{C} + k\mathbf{J}/(10+1)$. We consider the risk for estimating the linear function $h(\mathbf{X}) = \sum_{i=1}^{d} X(i)$.

Figure 1 displays the asymptotic variance (risk) as a function of δ or k for each of these four scenarios. Note that VRS significantly improves upon ERS and RRS. Though not displayed here, the convergence rates for ERS and RRS differ by less than a factor of 1.5. Thus the recommendation of using random scans with equal selection probabilities put forth by Roberts and Sahu (1997) is recommended for the burn-in process as convergence to the stationary distribution is not significantly improved by alternative scan strategies. However, for drawing statistical inferences, the random scan with equal selection probabilities is not preferred in terms of estimator precision.

We finally illustrate the performance of the three scan strategies for a larger dimensional Gaussian target distribution with mean vector zero and dispersion matrix $\Sigma_5 = \mathbf{B} - \mathbf{J}/(100+\delta)$ where $\mathbf{B} = diag\{100, ones(99,1)\}$. We find that the convergence rate under VRS, RRS, and ERS is 1.00, 0.61, and 0.61 respectively. VRS is definitely not desirable in the burn-in phase and again, RRS and ERS are not different enough to warrant use of RRS. The asymptotic risk under VRS, RRS, and ERS is 5881.46, 21994.27, 25204.25 respectively. VRS shows a 73% improvement in risk over RRS.

5.2 Screening Test Problem

Joseph et al. (1995) considers the evaluation of two diagnostic tests for the determination of Strongyloides infection status among patients in which no gold standard is available. Out of 162 subjects, we observe $y_{11} = 38$ positive cases under both tests, $y_{12} = 87$ positive cases only under diagnostic test one, $y_{21} = 2$ positive cases only under diagnostic test two, and $y_{22} = 35$ negative cases under both tests. The model assumes the number of cases under each test of these four scenarios is multinomially distributed with parameters $p_{ij} = \beta\eta_i\eta_j + (1-\beta)(1-\theta_i)(1-\theta_j)$, $i,j = 1,2$. Following Joseph et al. (1995), the five unknown parameters are assumed to have independent beta prior distributions with parameters $(a_{1\eta}, b_{1\eta}) = (21.96, 5.49)$,

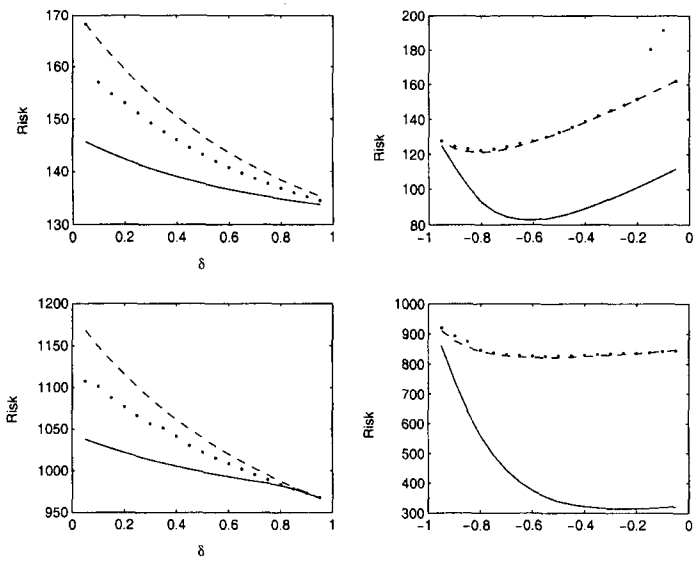

Figure 1: Risk comparison between the three random scan strategies ERS (dotted), RRS (dashed), and VRS (solid). The subplots from upper left to lower right apply these scans to ten-dimensional Gaussian target distributions with zero mean vectors and dispersions, being functions of either δ or k, $\mathbf{\Sigma}_1 = \mathbf{B} - \mathbf{J}/(10+\delta)$, $\mathbf{\Sigma}_2 = \mathbf{B} + k\mathbf{J}/(10+1)$, $\mathbf{\Sigma}_3 = \mathbf{C} - \mathbf{J}/(10+\delta)$, and $\mathbf{\Sigma}_4 = \mathbf{C} + k\mathbf{J}/(10+1)$ where $\mathbf{B} = diag\{10, 5, ones(1,8)\}$ and $\mathbf{C} = diag\{100, 10, ones(1,8)\}$.

$(a_{2\eta}, b_{2\eta}) = (4.44, 13.31)$, $(a_{1\theta}, b_{1\theta}) = (4.1, 1.76)$, $(a_{2\theta}, b_{2\theta}) = (71.25, 3.75)$, $(a_\beta, b_\beta) = (1, 1)$. A Gibbs sampler for drawing Monte Carlo posterior inferences is easily obtained by augmenting the data with the actual disease status, categorized similarly to the observed data as $\mathbf{z} = (z_{11}, z_{12}, z_{21}, z_{22})$.

Yee et al. (2002) shows that the Gaussian distribution serves as an approximate target distribution for this Gibbs sampler. We apply the routine of Section 4 to this data. For drawing statistical inferences, we use VRS in which we obtain a 55% reduction in risk over ERS. Interestingly, the VRS strategy draws more precise inferences than the Gibbs sampling routine of Yee et al. (2002), though we do not present the results here due to space limitations.

5.3 Bayesian image analysis

In this example, we consider dimension reduction as a means of implementing the VRS for analyzing image data. We consider modeling a temperature image over the globe from Jones et al. (1999). Our goal is to illustrate our algorithm, though the implications towards analysis of images, and high dimensional data, is apparent.

Let θ be the true image and y be the observed image. An Ising model type of prior distribution on the $p \times p$ lattice with the inverse temperature $\beta > 0$ is assumed where $\pi(\theta) \propto \exp\left\{-\beta \sum_{[i,j]}(\theta_i - \theta_j)^2\right\}$, $[i,j]$ specifies i and j as neighbors in the image. Here $\left\{-\beta \sum_{[i,j]}(\theta_i - \theta_j)^2\right\}$ can be expressed as the quadratic form $\{-\beta' W \beta\}$, where if $i = j$, $W(i,j)$ is the number of neighbors of i; if $|i-j| = 1$, $W(i,j) = -1$; otherwise $W(i,j) = 0$.

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 $\pi(\theta|y) \propto \pi(\theta) exp\{-0.5 \sum_i (\theta_i - \theta_j)^2/\sigma^2\} \propto exp\{-0.5(\theta - \mu)' \Sigma^{-1} (\theta - \mu)\}$ where $\Sigma = (2\beta W + I/\sigma^2)^{-1}$.

We may draw posterior inferences through an application of the random scan Gibbs sampler. Determining optimal selection probabilities through the algorithms of Section 4 may be computationally intensive for high-dimensional images. However, our experience shows that even a slight deviation from random scans with equal selection probabilities may lead to significant improvement in Monte Carlo precision, even if the scan strategy is not optimal. We thus use the structure of the image to suggest dimension reduction strategies for which implementation of VRS is computationally inexpensive.

The pairwise difference prior of Besag et al. (1995) forces symmetry in the parameters depending on their location in the image. In particular, the model assumes the covariance structure for points with the same number of neighbors are equal. We thus may group pixels into four groups depending on whether the pixel is on a corner, caddy-corner, edge, or center (see Figure 2).

We thus define four selection probabilities according to the four groups within which a pixel may lie. Assuming $\beta = 0.0002$ and δ estimated from the standard deviation of our climate data, the optimal scan probabilities for the four groups based on a random scan Gibbs sampler fit to the Ising model are $\boldsymbol{\alpha} = (0.08, 0.03, 0.02, 0.001)$. The corresponding asymptotic risk is 17032.98 improving upon ERS by 20%. This problem is small enough that we can optimize the risk over the 64 dimensional space of selection probabilities. The resulting gain in risk is minimal (less than 1%) with a substantial increase in cost (1 second versus 10 minutes). The user must be aware then of the tradeoff between computational expense and efficiency in larger dimensional problems. For a given level of coding skill, the efficiency gain may not be worth the loss in computational expense without reducing the dimension of the space of selection probabilities.

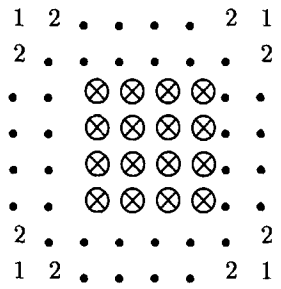

Figure 2: Grouping of pixels according to location in an 8×8 image.

References

Amit, Y. (1996). Convergence Properties of the Gibbs Sampler for Perturbations of Gaussians. *Annals of Statistics* 24, 122-140.

Amit, Y. and Grenander, U. (1991). Comparing Sweep Strategies for Stochastic Relaxation. *Journal of Multivariate Analysis* 37, 197-222.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Wiley, New York.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems. *Statistical Science* 10, 3-41.

Bradley, R. C. (1986). Basic Properties of Strong Mixing Conditions. In *Dependence in Probability and Statistics*, E. Eberlein and M. S. Taqqu (Eds.). Birkhaüser. Boston.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85, 398-409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721-741.

Jones, P. D., New, M., Parker, D. E., Martin, S., and Rigor, I. G. (1999). Surface air temperature and its changes over the past 150 years. *Review of Geophysics* 37, 173-199.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995). Bayesian Estimation of Disease Prevalence and Parameters for Diagnostics Tests in the Absence of a Gold Standard. *American Journal of Epidemiology* 141, 263-272.

Kagan, A. M., Linnik, Yu. V., Rao, C. (1973). *Characterization problems in mathematical statistics*, translated from the Russian by B. Ramachandran. Wiley, New York.

Levine, R. A. and Casella, G. (2003). Optimizing Random Scan Gibbs Samplers. Lawrence Livermore National Laboratory, Livermore, CA, UCRL-JC-145679.

Levine, R. A., Yu, Z., Hanley, W. G., and Nitao, J. J. (2003). Implementing Componentwise Hastings Algorithms. Lawrence Livermore National Laboratory, Livermore, CA, UCRL-JC-145678.

Liu, J., Wong, W. H., and Kong, A. (1995). Correlation Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *Journal of the Royal Statistical Society, B* 57, 157-169.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21, 1087-1092.

Mira, A. (2001). Ordering and Improving Performance of Monte Carlo Markov Chains. *Statistical Science* 16, 340-350.

Roberts, G. O. and Sahu, S. K. (2001). Approximate Predetermined Convergence Properties of the Gibbs Sampler. *Journal of Computational and Graphical Statistics* 10, 216-229.

Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society, B* **59** 291–317.

Sahu, S. K. and Roberts, G. O. (1999). On Convergence of the EM Algorithm and the Gibbs Sampler. *Statistics and Computing* 9, 55-64.

Sarmanov, O. V. and Zaharov, V. K. (1960). Maximum Coefficients of Multiple Correlation. *Doklady Akademii Nauk SSSR* 29, 269-271.

Yee, J., Johnson, W. O., and Samaniego, F. J. (2002). Asymptotic Approximations to Posterior Distributions via Conditional Moment Equations. *Biometrika* 89, 755-767.