The Annals of Statistics
2000, Vol. 28, No. 4, 1128–1149

ANTITHETIC COUPLING OF TWO GIBBS SAMPLER CHAINS[1]

BY ARNOLDO FRIGESSI, JØRUND GÅSEMYR AND HÅVARD RUE

Norwegian Computing Center, University of Oslo and Norwegian University for Science and Technology

Two coupled Gibbs sampler chains, both with invariant probability density π, are run in parallel so that the chains are negatively correlated. We define an asymptotically unbiased estimator of the π-expectation $\mathrm{E}(f(\mathbf{X}))$ which achieves significant variance reduction with respect to the usual Gibbs sampler at comparable computational cost. The variance of the estimator based on the new algorithm is always smaller than the variance of a single Gibbs sampler chain, if π is attractive and f is monotone nondecreasing in all components of $\mathbf{X}$. For nonattractive targets π, our results are not complete: The new antithetic algorithm outperforms the standard Gibbs sampler when π is a multivariate normal density or the Ising model. More generally, nonrigorous arguments and numerical experiments support the usefulness of the antithetically coupled Gibbs samplers also for other nonattractive models. In our experiments the variance is reduced to at least a third and the efficiency also improves significantly.

1. Introduction. Markov chain Monte Carlo (MCMC) algorithms allow the approximate calculation of expectations with respect to multivariate probability density functions $\pi(x), x \in \Omega$ defined up to a normalizing constant. We refer the reader to Gilks, Richardson and Spiegelhalter (1996) as a starting point for a vast literature about MCMC methodology. The underlying idea is to construct an ergodic discrete time Markov chain with invariant density function π, whose trajectory is easy to simulate without normalizing π. In order to approximate the expectation $\mathrm{E}(f(\mathbf{X})) < \infty$ of a function $f(x)$ with respect to π, one just needs to compute the empirical average of f along the generated trajectory $\mathbf{X}^1, \ldots, \mathbf{X}^T$. To avoid strong dependence on the initial conditions, an initial part is dropped. The sample mean with *burn-in* of length T_0,

$$\hat{f} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} f(\mathbf{X}^t),$$

is used. In this paper we propose a new algorithm for the estimation of $\mathrm{E}(f(\mathbf{X}))$. The idea is to simulate two MCMC trajectories in parallel, both invariant with respect to π, which are coupled in such way that variance

Received March 1999; revised April 2000.

[1]Supported by the Universitá di Roma Tre the EU-TMR project on Spatial Statistics (ERB-FMRX-CT960095), the ESF program on Highly Structured Stochastic Systems and the Norwegian Research Council project 11441420.

AMS 1991 *subject classifications*. Primary 62M05, 65C05; secondary 62M10.

Key words and phrases. Antithetic Monte Carlo, associated random variables, attractive models, decay of cross-autocorrelations, Markov chain Monte Carlo, variance reduction.

reduction can be achieved. We use the Gibbs sampler, a particular MCMC scheme where samples from a one-dimensional conditional density computed from π are drawn.

After the burn-in we continue the simulation by running two parallel Gibbs sampler chains, both ergodic with respect to π. Let us denote the two chains $\mathbf{X}^t$ and $\mathbf{Y}^t$ for $t = T_0+1, T_0+2, \ldots$. Marginally, the two chains are ordinary Gibbs samplers, but their joint probability measure is constructed in such a way that $f(\mathbf{X}^t)$ and $f(\mathbf{Y}^t)$ have negative covariance. The coupling is simple, based on using a common sequence of random numbers. Specifically, if $\mathbf{X}^t$ uses a uniform $[0, 1)$ random number U^t to proceed to $\mathbf{X}^{t+1}$, then $\mathbf{Y}^t$ uses $1-U^t$ to proceed to $\mathbf{Y}^{t+1}$. This coupling is well known to reduce the variance of empirical averages of i.i.d. samples. A pleasant fact of the antithetically coupled Gibbs sampler is that, starting from the code of the usual Gibbs sampler, the modifications required in order to implement the new algorithm are simple if Ω is discrete. If the needed conditional univariate distribution functions cannot be inverted analytically, then the implementation requires either numerical inversion or some more advanced techniques.

We combine the output of the two coupled chains into the asymptotically unbiased estimator

$$\hat{\hat{f}} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \frac{f(\mathbf{X}^t) + f(\mathbf{Y}^t)}{2}. \tag{1}$$

To make a fair comparison between this algorithm and the usual, single trajectory Gibbs sampler, we have to take into consideration that each iteration of the new algorithm takes twice the computing time of a single Gibbs sampler iteration. Hence we allow the single Gibbs sampler to turn for twice as many iterations as the new algorithm. This means that $\hat{\hat{f}}$ in (1) has to be compared with

$$\hat{f} = \frac{1}{2T} \sum_{t=T_0+1}^{T_0+2T} f(\mathbf{X}^t). \tag{2}$$

We define precisely the new algorithm in Section 2. In Section 3 we assume that $\mathbf{X}^{T_0}$ and $\mathbf{Y}^{T_0}$ are independent and π-distributed. Hence (1) and (2) are unbiased and we prove that $\mathrm{Var}(\hat{\hat{f}}) \le \mathrm{Var}(\hat{f})$, for component-wise monotone functions f, attractive π, and for all T. Not surprisingly, the key point is the sign of the cross-covariances between the two coupled chains. Under the given conditions, we prove that the cross-covariances are all negative. Section 4 is devoted to a study of the multivariate normal density and the Ising model. These distributions are not necessarily attractive but have a certain local symmetry property. If f is linear, then as $T \to \infty$, we have $\mathrm{Var}(\hat{\hat{f}}) \le \mathrm{Var}(\hat{f})$ even when π is not attractive. In Section 5 we discuss the joint asymptotic properties of the coupled chains and the existence of a unique joint stationary measure. In Section 6 we present some heuristic arguments supporting the claim that $\mathrm{Var}(\hat{\hat{f}}) \le \mathrm{Var}(\hat{f})$ for other nonattractive targets π and give some

precise results for a nonattractive example that mimics the behavior of the new algorithm. The variance reduction, defined as $\mathrm{Var}(\hat{\hat{f}})/\mathrm{Var}(\hat{f})$, seems to depend only mildly on the mixing property of the single Gibbs sampler chain; if the single Gibbs sampler chain is slowly mixing, then the joint Gibbs sampler will also be slow; however, the variance reduction remains roughly the same. In Section 7 we discuss some practical implementation issues and we test our new algorithm on two data sets: the hierarchical Poisson model [Gelfand and Smith (1990)] and the ordered normal means example [Gelfand, Hills, Racine-Poon and Smith (1990)]. The experiments show that the antithetically coupled Gibbs sampler is significantly better than the standard one. The variance reduction is often larger than five and always larger than three. In practice $\mathbf{X}^{T_0}$ and $\mathbf{Y}^{T_0}$ are not π-distributed. Hence, we include bias in our comparison and find that the ratio $(\mathrm{bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}))/(\mathrm{bias}(\hat{\hat{f}}) + \mathrm{Var}(\hat{\hat{f}}))$ is larger than 10 in our experiments. Looking beyond the Gibbs sampler, we apply the antithetic coupling to Metropolis–Hastings and show empirically that improvement can still be achieved, although with a variance reduction of about 2. Comments are in Section 8.

2. The new algorithm. Let $\Omega = S \times S \times \cdots \times S = S^n$ be the n-fold product space of a set S, which may be either discrete or continuous. For simplicity we consider two cases: $\Omega = \mathbb{R}^n$ and π is a probability density function that is absolutely continuous with respect to, say, Lebesgue measure, or S is discrete and π is a discrete probability. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. The random scan Gibbs sampler for sampling from π is a Markov chain $\mathbf{X}^0, \mathbf{X}^1, \ldots$ constructed as follows. Given $\mathbf{X}^t = x^t$, one component in $\{1, 2, \ldots, n\}$ is chosen uniformly at random. Denote this component by I^t. Only $X^t_{I^t}$ will be updated by sampling the new value $X^{t+1}_{I^t}$ from the conditional density

$$\pi(x_{I^t} \mid \mathbf{X}_{-I^t} = \mathbf{x}_{-I^t}), \tag{3}$$

where $\mathbf{x}_{-A} = \{x_i : i \notin A\}$, for $A \subset \{1, \ldots, n\}$. The remaining components are left unchanged, $\mathbf{X}^{t+1}_{-I^t} = \mathbf{x}^t_{-I^t}$. We assume (3) to be strictly positive, so that the resulting Markov chain is ergodic and π-variant. The transition of a random scan Gibbs sampler can be written as

$$\mathbf{X}^{t+1} = \mathbf{\Phi}(\mathbf{X}^t, I^t, U^t), \tag{4}$$

where $U^0, U^1, \ldots$ is a sequence of i.i.d. random numbers, uniformly distributed in $[0, 1)$, and $I^0, I^1, \ldots$ are i.i.d. random numbers uniform in $\{1, 2, \ldots, n\}$, that identify the component to be updated at step $t + 1$. The I^tth component of the vector function $\mathbf{\Phi}$ is the inverse distribution function corresponding to the local conditional density (3),

$$\Phi_{I^t}(\mathbf{X}^t, I^t, U^t) = \Phi_{I^t}(\mathbf{X}^t_{-I^t}, U^t) = \inf\{x \in \mathbb{R} : \pi(X_{I^t} \le x \mid \mathbf{X}^t_{-I^t}) = U^t\},$$

where the inf is needed only if S is discrete. The other components of $\mathbf{\Phi}$ are identity functions $\Phi_j(\mathbf{X}^t, I^t, U^t) = X^t_j$, for $j \neq I^t$. We will also give results for another visitation schedule, where each component is updated in a raster

scan. Then we shall adopt a similar notation using lowercase letters i^t for the site to be updated at time t, $i^t = (t-1)(\mathrm{mod}\, n) + 1$.

We now define the companion chain. It is marginally a π-stationary Gibbs sampler with the same type of scan and transition rule as (4),

$$\mathbf{Y}^{t+1} = \mathbf{\Phi}(\mathbf{Y}^t, I^t, 1 - U^t), \tag{5}$$

but the common random numbers U^t and I^t couple the two chains and make $\mathbf{X}^{t+1}$ and $\mathbf{Y}^{t+1}$ dependent. We call the coupling antithetic because we use $1 - U^t$ in (5). The same component I^t is updated in both chains. Looking to the coupled chains jointly, notice that $X_{I^t}^{t+1}$ is conditionally independent of $\mathbf{Y}_{-I^t}^t$ given $\mathbf{X}_{-i^t}^t$, because of (4) and (5) and since U^t is independent of $\mathbf{Y}_{-I^t}^t$. The two coupled Gibbs sampler chains allow us to construct the estimator $\hat{\hat{f}}$ in (1) which we shall compare to $\hat{f}$ given in (2) in the rest of this paper.

When Ω is discrete, or when Ω is continuous and the conditional distributions $\pi(X_i \le x \mid \mathbf{X}_{-i})$ can be inverted analytically, the implementation of the new algorithm is straightforward. If this is not the case, even if formally $\mathbf{X}^t$ can be represented as inverse transform of uniform variables as in (4), there might be no closed form for $\mathbf{\Phi}$. Then, one has to numerically invert $\pi(X_i \le x \mid \mathbf{X}_{-i})$, which is generally easy and can be efficiently performed off-line. Alternatively, accept–reject mechanisms can be used. We shall come back to these practical issues in Section 7.

3. Comparing variances for attractive target densities. We assume that the two chains are started at time $T_0 = 0$ in the marginal stationary distribution $\mathbf{X}^0 \sim \pi$, $\mathbf{Y}^0 \sim \pi$, independently, and then coupled. We shall return to this assumption later in this section and again in Section 7. Then both $\hat{\hat{f}}$ and $\hat{f}$ are unbiased. Hence, to evaluate the performance of the antithetically coupled Gibbs sampler, we compare the variance of $\hat{f}$ with the variance of $\hat{\hat{f}}$ (both assumed to be finite). In comparing variances we shall need both autocovariances for the marginal chains, and cross covariances for the two chains jointly. Let $\gamma_k = \mathrm{Cov}(f(\mathbf{X}^0), f(\mathbf{X}^k))$, $k = 0, 1, \ldots$ be the marginal stationary autocovariance at lag k of one of the two components. We do not assume stationarity of the joint (bivariate) Markov chain, hence the cross covariances $\beta(t, s) = \mathrm{Cov}(f(\mathbf{X}^t), f(\mathbf{Y}^s))$ depend on time.

We consider a special class of target distributions π and functions f. The target π is assumed to be attractive; see Møller (1999) for several examples. A model π is attractive if $\pi(X_i \le x_i \mid \mathbf{x}_{-i}) \le \pi(X_i \le x_i \mid \mathbf{x}'_{-i})$, for $\mathbf{x}_{-i} \ge \mathbf{x}'_{-i}$, $\forall \mathbf{x}, \mathbf{x}' \in \Omega$, assuming the partial ordering of Ω given by $\mathbf{x}_A \ge \mathbf{x}'_A$ if $x_i \ge x'_i$ for all $i \in A$. We assume from now on and without loss of generality that the expected value of $f(\mathbf{X})$ is zero. To be able to study the two estimators $\hat{\hat{f}}$ and $\hat{f}$, we need to restrict the space of functions f, too. Our algorithm induces antithetic dependency between $\mathbf{X}^t$ and $\mathbf{Y}^t$; we want this structure to transfer to $f(\mathbf{X}^t)$ and $f(\mathbf{Y}^t)$ as well. For this we require $f \in \mathscr{F}$, where $\mathscr{F}$ is the class of nonconstant functions $f\colon \Omega \to \mathbb{R}$ which are monotone nondecreasing

in all components. In practice, often $f(\mathbf{x}) = \sum_i g_i(x_i)$ where the $g_i(\cdot)$'s are monotonic increasing functions. If the function of interest is decreasing in, say, component i, we can replace X_i with $-X_i$ to obtain a function in $\mathscr{F}$ and change π accordingly.

THEOREM 1. *Suppose $f \in \mathscr{F}$ and π is attractive. Consider the coupled Gibbs sampler chains given in* (4) *and* (5) *using a random scan or a raster scan. If $\mathbf{X}^0$ and $\mathbf{Y}^0$ are independent and distributed according to π, then*

$$\operatorname{Var}(\hat{f}) - \operatorname{Var}(\hat{\hat{f}}) \geq 0 \tag{6}$$

for every $T > 0$.

Proofs are collected in the Appendix. The theorem is based on Lemma 1, which is interesting in itself. It states that under the same assumptions of Theorem 1, $\beta(t, s) \leq 0$ for all t and s. For the raster scan we prove (6) also under the different assumption that all components of $\mathbf{X}^0$ and $\mathbf{Y}^0$ are independent, but $\mathbf{X}^0$ and $\mathbf{Y}^0$ are not required to be distributed according to π. This condition is more appealing in practice. See the Appendix for details. In the next section we move to nonattractive models to see if the variance of $\hat{\hat{f}}$ is still smaller than the variance of $\hat{f}$.

4. Comparing variances for some nonattractive target densities. We first consider a multivariate normal target distribution: π is normal with mean vector zero and inverse covariance matrix $\mathbf{Q} = (q_{ij})$. We assume without loss of generality that the diagonal of $\mathbf{Q}$ consists of ones. When updating component i, the Gibbs sampler samples from a univariate normal density with mean $-\sum_{j \neq i} q_{ij} x_j^t$ and variance 1. Note that the off-diagonal terms in $\mathbf{Q}$ can be both negative and positive, allowing for nonattractive π.

In a Bayesian setting, the posterior density of $\mathbf{X}$ given m data points often tends to a normal with mean equal to the maximum likelihood estimator of $\mathbf{X}$ and variance of order $1/m$, as $m \to \infty$. Then we can Taylor expand $f(\mathbf{X})$ around the maximum likelihood estimate to first order, so that a linear approximation of f is enough.

THEOREM 2. *Let π be the multivariate normal density. Let f be a linear function with zero π-mean. Assume a deterministic scan for the Gibbs sampler. For T large enough,* $\operatorname{Var}(\hat{f}) \geq \operatorname{Var}(\hat{\hat{f}})$. *Moreover,* $\operatorname{Var}(\hat{\hat{f}}) = \mathscr{O}(T^{-2})$ *as* $T \to \infty$.

The last part of the statement of Theorem 2 is curious because it shows that in this special case, coupling two Gibbs sampler chains reduces the variance by a full order of magnitude, since for the single stationary chain $\operatorname{Var}(\hat{f}) = \mathscr{O}(T^{-1})$. The reason is the following. As shown in the proof of Theorem 2, we have that

$$X_{i_t}^{t+1} + Y_{i_t}^{t+1} = -\sum_{j \neq i_t} q_{ij}(X_j^t + Y_j^t). \tag{7}$$

This means that in the limit as $t \to \infty$, the process $(\mathbf{X}^t, \mathbf{Y}^t)$ is attracted and trapped in the set $\{(\mathbf{x}, \mathbf{y}) \in \Omega \times \Omega | \mathbf{x} + \mathbf{y} = 0\}$. Once $(\mathbf{X}^t, \mathbf{Y}^t)$ lies in this set, then $\beta(t, t+k) = -\gamma_k$. Because of the deterministic nature of (7), the process with probability 1 will never get to the set, unless started there. (If $\mathbf{X}^0 + \mathbf{Y}^0 = 0$, then the variance is zero.) The setting is indeed special. Theorem 2 holds also for other target densities than the multivariate normal, if they satisfy the following symmetry condition: $\pi(x_i \mid \mathbf{x}_{-i}) = \psi_i(x_i - \tilde{x}_i)$, $\forall i$, where $\psi_i(\cdot)$ is symmetric around zero, and $\tilde{x}_i$ is the median in $\pi(x_i \mid \mathbf{x}_{-i})$ which can be written as $\tilde{x}_i = \mathbf{a}_i^T \mathbf{x}_{-i}$ for some vector $\mathbf{a}_i$. Not all models that satisfy this symmetry condition are attractive. The multivariate normal satisfies this condition because the conditional median equals the conditional mean which is linear in x_{-i}, and the conditional variance does not depend on x_{-i}. Another π, sometimes used for smoothing, that satisfies the symmetry conditions is $\pi(\mathbf{x}) \propto \exp(-\sum_{i,j} b_{ij}(x_i - x_j)^k)$, where $b_{ii} \geq 0$, the off-diagonal coefficients b_{ij} must be chosen appropriately, k is even (say 4), and x_1 is fixed. Though a different proof would be needed, we think that Theorem 2 remains valid for a random scan.

We conclude this section with a second (discrete) example, the two-dimensional Ising model, for which we obtain a result similar to the multivariate normal case. When π is the Ising model the n variables x_i are positioned on the sites of a finite squared grid, and $\pi(\mathbf{x}) = \exp(\delta \sum_{i \sim j} x_i x_j)/Z$, where $x_i \in \{-1, +1\}$, the sum is taken over the four nearest neighboring pairs and Z is the normalizing constant. The so-called inverse temperature δ can be either positive, in which case the model is attractive, or negative, which gives a repulsive interaction model. We shall consider our algorithm with a deterministic scan. Define the set $C = \{(\mathbf{x}, \mathbf{y}) \in \Omega \times \Omega | \mathbf{x} + \mathbf{y} = 0\}$. We observe that C is an absorbing set for the joint chain: if $(\mathbf{X}^t, \mathbf{Y}^t) \in C$ then also $(\mathbf{X}^s, \mathbf{Y}^s) \in C$ for $s > t$, because of the antithetic coupling and the form of the conditional distribution $\pi(x_i \mid \mathbf{x}_{-i})$. Furthermore, C is reachable from any initial state within one full sweep with a probability larger than $p = [\exp(-8|\delta|)/(1 + \exp(-8|\delta|))]^n > 0$. Hence the random time τ at which C is entered is stochastically dominated by a geometric random variable τ' with mean $1/p$ and finite variance. For any linear f with zero mean, we have, as $t \to \infty$,

$$\mathrm{Var}(\hat{\bar{f}}) = \mathrm{Var}\left(\frac{1}{2T} \sum_{t=1}^{\min\{T, \tau\}} (f(\mathbf{X}^t) + f(\mathbf{Y}^t))\right) \leq \frac{1}{T^2} c\mathrm{E}((\tau')^2) = \mathcal{O}(T^{-2}),$$

where c is a finite constant.

5. Joint properties of the coupled Gibbs sampler chains. The coupled Gibbs samplers $(\mathbf{X}^t, \mathbf{Y}^t)$ form a Markov chain evolving on $\Omega \times \Omega$ that updates components blockwise, the block being $B_i = (X_i, Y_i)$. Although each marginal component is a Gibbs sampler chain, $(\mathbf{X}^t, \mathbf{Y}^t)$ does not need to be. An algorithm that updates a component B_i using a conditional probability that does not depend on the current value in B_i is not necessarily a Gibbs sampler. The full conditionals have to satisfy complicated consistency conditions.

It is interesting to know if the joint chain $(\mathbf{X}^t, \mathbf{Y}^t)$ is ergodic and if so what the properties of the stationary measure are, which of course has π as marginals. The difficulty is well illustrated by the multivariate Gaussian case. As explained in Section 4, if there is a limit distribution $\mu(\mathbf{x}, \mathbf{y})$ of $(\mathbf{X}^t, \mathbf{Y}^t)$, as $t \to \infty$, then its support must be

$$\operatorname{supp}(\mu) = \{(\mathbf{x}, \mathbf{y}) \in \Omega \times \Omega | \mathbf{x} + \mathbf{y} = 0\}. \tag{8}$$

In this case, when $t \to \infty$, the density of $(\mathbf{X}^t, \mathbf{Y}^t)$ is attracted towards the subspace $\mathbf{x} = -\mathbf{y}$. Hence $\Omega \times \Omega$ can be decomposed into a transient class and an ergodic one and μ is singular with respect to $\pi \times \pi$. For general state space $\Omega \times \Omega$, the picture could be more complicated: it could be that the marginal components converge (to π) while jointly they do not converge, or that there is more than one ergodic class. We are not able to exclude such situations. However the asymptotic behavior of the joint chains does not influence the efficacy of the new algorithm.

The theory in the Appendix of Arjas and Gasbarra (1996) can be used to prove that if the joint chain $(\mathbf{X}^t, \mathbf{Y}^t)$ is started in the ergodic class then there exists a unique stationary distribution μ on this class. In the multivariate normal case, this means that if $(\mathbf{X}^0, \mathbf{Y}^0)$ is such that $\mathbf{X}^0 = -\mathbf{Y}^0$, then there exists a unique stationary distribution μ on the set $\mathbf{x} = -\mathbf{y}$. We give the precise statement; see Arjas and Gasbarra (1996) for more information on the assumptions.

THEOREM 3. *Let $\mathbf{X}^t$ and $\mathbf{Y}^t$ be positive recurrent Markov chains on a complete separable metric space Ω. Let $\mathbf{Z}^t = (\mathbf{X}^t, \mathbf{Y}^t)$ be a φ-irreducible Markovian coupling of $\mathbf{X}^t$ and $\mathbf{Y}^t$. Consider the closure (with respect to the product topology of $\Omega \times \Omega$) $\overline{\operatorname{supp}\{\varphi\}}$ with the relative topology inherited from the product topology. Let $(\mathbf{X}^0, \mathbf{Y}^0) \in \overline{\operatorname{supp}\{\varphi\}}$. If, as a Markov chain on $\overline{\operatorname{supp}\{\varphi\}}$, $\mathbf{Z}^t$ is weakly Feller with respect to the relative topology, and if $\operatorname{supp}\{\varphi\}$ contains an open set (with respect to the relative topology), then $\mathbf{Z}^t$ is positive recurrent.*

It can be seen that the coupled Gibbs samplers $(\mathbf{X}^t, \mathbf{Y}^t)$ realize a Markovian coupling. In the multivariate normal case φ can be chosen to be the Lebesgue measure. We have not experienced more than one ergodic class in our numerical experiments. What can we say about the form of the support of μ? In the Gaussian and Ising models it is the symmetry of the conditional density with respect to the median, which is linear in the conditioning components, that makes the limiting support of the joint chain of the type $\{(\mathbf{x}, \mathbf{y})\colon \mathbf{y} = H(\mathbf{x})\}$. It is interesting to note that if there is such a function H and if $\pi(\mathbf{x}) > 0$ for all $\mathbf{x} \in \Omega$, then this function must act componentwise, that is, $y_i = h_i(x_i)$ for all i, as happens in the Gaussian case; see Theorem 4. Note that if π is an n-fold product measure on $\Omega = S^n$, that is $\pi = \pi_1 \times \cdots \times \pi_n$, then $(\mathbf{X}^t, \mathbf{Y}^t)$ has a stationary distribution reached after one single sweep with support $y_i = h_i(x_i)$, $i = 1, \ldots, n$, where the functions h_i are nonlinear. Hence constant $\mathbf{x} + \mathbf{y}$ is not the only possible form for a degenerate $\operatorname{supp}(\mu)$.

6. Nonrigorous variance comparison for general nonattractive target densities. Because we are not able to extend rigorous theory beyond attractive π's, we present some rough arguments and conjectures. We assume that the coupled chains have been started in an ergodic class, on which there exists a joint stationary measure μ. Denote by $\beta_k = \beta(t, t+k)$ the stationary cross covariances. Let $f \in \mathscr{F}$ and assume a random scan. We argue that all β_k, $k > 0$, have the same sign as β_0. The heuristic argument is based on the approximations

$$\mathrm{E}(f(\mathbf{Y}^{t+k}) \mid \mathbf{Y}^t) \approx \frac{\mathrm{Cov}(f(\mathbf{Y}^t), f(\mathbf{Y}^{t+k}))}{\mathrm{Var}(f(\mathbf{Y}^t))} f(\mathbf{Y}^t) = \frac{\gamma_k}{\gamma_0} f(\mathbf{Y}^t), \tag{9}$$

$$\mathrm{E}(f(\mathbf{X}^t) \mid \mathbf{Y}^t) \approx \frac{\mathrm{Cov}(f(\mathbf{Y}^t), f(\mathbf{X}^t))}{\mathrm{Var}(f(\mathbf{Y}^t))} f(\mathbf{Y}^t) = \frac{\beta_0}{\gamma_0} f(\mathbf{Y}^t). \tag{10}$$

Approximation (9) is explained as follows: among all quantities $cf(\mathbf{Y}^t)$, linear in $f(\mathbf{Y}^t)$, the one given in (9) minimizes the mean squared error, $\mathrm{E}_{\mathbf{Y}^t}\mathrm{E}((cf(\mathbf{Y}^t) - f(\mathbf{Y}^{t+k}))^2 \mid \mathbf{Y}^t)$. The same argument applies to (10). The k-step conditional expectation is the best predictor for $f(\mathbf{Y}^{t+k})$ in terms of mean squared error, but is not generally linear in $f(\mathbf{Y}^t)$. If f is linear, $f(\mathbf{Y}^t) = \mathbf{a}^T\mathbf{Y}^t$, and if π is multivariate normal, then the k-step conditional expectation is linear in $\mathbf{Y}^t$ and approximately linear in $\mathbf{a}^T\mathbf{Y}^t$ unless the dependency among the Y_i's is very strong.

We obtain the following expression for β_k:

$$\begin{aligned}
\beta_k &= \mathrm{E}(f(\mathbf{X}^t)f(\mathbf{Y}^{t+k})) = \mathrm{E}_{\mathbf{Y}^t}\mathrm{E}(f(\mathbf{X}^t)f(\mathbf{Y}^{t+k}) \mid \mathbf{Y}^t) \\
&= \mathrm{E}_{\mathbf{Y}^t}\Big[\mathrm{E}(f(\mathbf{X}^t) \mid \mathbf{Y}^t)\mathrm{E}(f(\mathbf{Y}^{t+k}) \mid \mathbf{Y}^t)\Big] \\
&\approx \mathrm{E}_{\mathbf{Y}^t}\left[\frac{\beta_0}{\gamma_0} f(\mathbf{Y}^t)\frac{\gamma_k}{\gamma_0} f(\mathbf{Y}^t)\right] = \beta_0\frac{\gamma_k}{\gamma_0},
\end{aligned} \tag{11}$$

using (9), (10) in the last line and conditional independence. If (9) and (10) are good, so will (11) be. For a random scan Gibbs sampler Liu, Wong and Kong (1995) show that $\gamma_k \geq 0$ for all k, regardless of the attractivity. If (9) and (10) were correct, β_k would have the same sign as β_0 for all $k > 0$. Figure 3 shows a plot of the estimated values of β_k/γ_0 and the approximation $\beta_0\gamma_k/\gamma_0^2$ for the pump example described in Section 7.1. The fit is very good.

Using (11), and the expressions for $\mathrm{Var}(\hat{f})$ and $\mathrm{Var}(\hat{\hat{f}})$ given in the proof of Theorem 1, we calculate the variance reduction factor of $\hat{\hat{f}}$ with respect to $\hat{f}$ when $T \to \infty$, as

$$\frac{\mathrm{Var}(\hat{f})}{\mathrm{Var}(\hat{\hat{f}})} \sim \frac{1}{1 + \beta_0/\gamma_0}. \tag{12}$$

The antithetic algorithm is always better if $\beta_0 \leq 0$ and (9) and (10) (approximately) hold. We conjecture that this is true in many cases. For example, suppose the cross-autocorrelation at lag zero β_0/γ_0 is equal to, say, $-2/3$.

Then the variance reduction factor is circa 3. In Section 7 we observe estimated variance reduction factors larger than 3. Note further that the ratio (12) does not depend on γ_k, $k > 0$, which may indicate that the efficiency of the new algorithm does not depend on the mixing properties of the marginal chains.

Next we present a nonattractive example that mimics the behavior of the new algorithm but that allows for rigorous analysis. The two coupled chains are stationary non-Gaussian autoregressive processes. Since the approximations (9) and (10) are exact, the sign of β_k follows the sign of β_0. The variance reduction (12) is valid and is not influenced by the mixing properties of the marginal chains. Besides being nonattractive, the process does not satisfy the symmetry condition used in Section 4. Let X^t be the real valued process.

$$X^t = \phi X^{t-1} + \varepsilon_x^t, \qquad t > 0, \tag{13}$$

started in equilibrium at time zero, with $|\phi| < 1$. Let ε_x^t be i.i.d. zero–one binary variables with $\mathrm{p}(\varepsilon_x^t = 1) = p \geq 1/2$. This is not a Gibbs sampler, but it has the same flavor; see (20). We choose $f(x) = x$ with the aim of estimating the mean $\mathrm{E}(\mathbf{X}) = p/(1-\phi)$. The variance of $\hat{f}$ is

$$\mathrm{Var}(\hat{f}) = \mathrm{Var}\left(\frac{1}{2T}\sum_{t=1}^{2T} X^t\right) \sim \tau_x/(2T) \quad \text{where } \tau_x = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \tag{14}$$

is the integrated autocovariance time. Also, $\tau_x = \gamma_0(1+\phi)/(1-\phi)$ where $\gamma_0 = p(1-p)/(1-\phi^2)$. We compare the variance in (14) with that obtained using two realizations of (13), X^t and Y^t, where X^t is sampled using the uniform random variable U^t and Y^t is sampled using $1 - U^t$. We get

$$\mathrm{Var}(\hat{\hat{f}}) = \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{X^t + Y^t}{2}\right) \equiv \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} Z^t\right) \sim \tau_z/T.$$

Z^t is an autoregressive process of the same form as (13), with ε_z^t equal to either 1, with probability $2p-1$, or to 1/2 otherwise. We get $\tau_z = (p-1/2)(1-p)/(1-\phi^2)$. Hence, we obtain the factor of variance reduction of $(\hat{\hat{f}})$ w.r.t. $\hat{f}$ as

$$\frac{\mathrm{Var}(\hat{f})}{\mathrm{Var}(\hat{\hat{f}})} \sim \frac{\tau_x}{2\tau_z} = \frac{1}{2 - 1/p}, \qquad p > 1/2, \tag{15}$$

where we make use of the exponentially decaying autocovariances of X^t and Z^t. This shows that the antithetic estimate is always better, and that the variance reduction factor tends to ∞ as the symmetry increases; that is, $p \to 1/2$. It tends to 1 as the symmetry decreases, $p \to 1$. For $p = 1/2$ (perfect symmetry), the variance of $\hat{\hat{f}}$ is again $\mathcal{O}(T^{-2})$. Notice that the joint and marginal chains require a burn-in of similar length, as both are autoregressive processes of the same form (13). For the cross covariances we get

$$\beta_k = -\frac{(1-p)^2}{1-\phi^2}\phi^{|k|}$$

and (11) and (12) hold exactly. Furthermore, β_0 is minimal for $p = 1/2$, it takes value $\beta_0 = 0$ if $p = 1$, and the efficiency in (15) increases as β_0 becomes more negative.

7. Practical implementation and numerical experiments. According to Theorem 1, we should start the two marginal chains independently in π. Only if a raster scan is used can all components of $\mathbf{X}^0$ and $\mathbf{Y}^0$ be sampled independently. The latter method is easy, while sampling $\mathbf{X}^0$ and $\mathbf{Y}^0$ independently from π requires the independent running to convergence of two ordinary Gibbs Sampler chains. The bias of the two estimators is influenced by the initialization. In general, the asymptotic mean squared error of either estimator is determined by the variance, which is of order T^{-1}, and by the squared bias, which is of order T^{-2}. We discuss more about the bias in our first example. In practice, we run a single Gibbs sampler for T_0 steps. We keep $\mathbf{X}^{T_0}$ and discard the rest. Let $\mathbf{Y}^{T_0} = \mathbf{X}^{T_0}$, start two (dependent) trajectories, one using (4) and the other (5) and terminate the coupled chains after a further T transitions. In this way we fail to fulfill precisely the requirement of $\mathbf{X}^{T_0}$ and $\mathbf{Y}^{T_0}$ in Theorem 1 (independence and π-distributed). Nevertheless, we will compare this algorithm, which gives an estimator $\hat{\hat{f}}$ based on a total of $2T$ Gibbs sampler updates, with a single Gibbs sampler chain of length $2T$, started in $\mathbf{X}^{T_0}$. If the burn-in is long enough, the two estimators will be approximately unbiased. We shall apply our algorithm to two well-studied data sets, the hierarchical Poisson model [Gelfand and Smith (1990)] and the ordered normal means example [Gelfand, Hills, Racine-Poon and Smith (1990)]. The main purpose is to evaluate the performance of the new algorithm and to quantify its variance reduction and the efficiency w.r.t. the usual Gibbs sampler. We also introduce antithetically coupled Metropolis–Hastings chains and discuss their performance.

7.1. *Hierarchical Poisson model.* Gelfand and Smith (1990) present counts $s = (s_1, \ldots, s_n)$ of failures in $n = 10$ pump systems at a nuclear power plant, where the times of operation $t = (t_1, \ldots, t_n)$ for each system are known. The hierarchical model assumes $s_k \sim \text{Poisson}(\lambda_k t_k)$, and a common Gamma prior for the failure rate λ_k of each pump, $\lambda_k \sim \Gamma(\alpha, \beta)$. The problem is to make inferences about α and the inverse scale β computing the posterior means. Here α has exponential prior distribution with mean 1, and β a $\Gamma(0.1, 1.0)$ distribution.

The conjugate priors ensure that λ_1 is Γ-distributed conditional on the remaining variables, as are $\lambda_2, \ldots, \lambda_n$ and β. It is therefore easy to update each of these variables using a Gibbs sampler. The conditional density for α is, however, nonstandard since

$$\pi(\alpha \mid \lambda_1, \ldots, \lambda_{10}, \beta) \propto \exp(\alpha a - n \log \Gamma(\alpha))$$
$$\text{where } a = n \log \beta + \sum_{k=1}^{n} \log \lambda_k - 1. \tag{16}$$

In this case it is most natural to perform a Metropolis–Hastings step for the α-parameter update. This means that, using a proposal density, a new value for α is proposed and then accepted or rejected. We suggest three different updating strategies for α:

1. Gibbs sampler update. To implement the full Gibbs sampler, we compute numerically $F^{-1}(u; a_x)$ and $F^{-1}(1-u; a_y)$, where F is the cumulative conditional distribution function for α.
2. Hastings update. We approximate the conditional density (16) with a normal ($\widetilde{F}$) with the mean and variance matching the mode and the curvature in the mode. We update α using a Hastings step, proposing to move the current values of α to $\widetilde{F}_x^{-1}(u)$ and $\widetilde{F}_y^{-1}(1-u)$, respectively. We accept–reject the proposals using a common uniform variate and get an 90% average acceptance rate for α.
3. Metropolis update. We update α using a random walk Metropolis step and propose a new state from a uniform density centered at the old state using u and $1-u$. We accept the proposals using the same uniform variate. The width of the proposal density is determined to obtain an average acceptance rate for α close to 50%.

To verify the robustness with respect to various scanning schedules, we apply each of these three updating rules for α with three different visiting schedules: random scan (RS), where we look to 12 variable updates as one step; random permutation scan (RPS), where at each iteration we update our 12 variables in a random permutation and deterministic scan (DET), where at each iteration we update $\lambda_1, \ldots, \lambda_{10}, \alpha, \beta$ and then $\alpha, \lambda_{10}, \ldots, \lambda_1$.

We run a single Markov chain using $T_0 = 1000$ iterations as burn-in, and then we split the chain into two components and run it according to (4) and (5) for $\lambda_1, \ldots, \lambda_{10}, \beta$. For α, we run it according to one of the three above methods. The algorithm is set to perform a further $T = 100{,}000$ iterations. Figure 1 shows small parts of the sample paths for the β variables in the two chains, denoted by $\beta_x^t$and β_y^t, respectively, where we use the Gibbs sampler also for α and RPS. The paths show a clear negative correlation. In Figure 2(a) we plot the sampled points (α_x^t, α_y^t) of the two coupled chains to show the shape of empirical joint density, using 5000 subsequent samples. The second panel of Figure 2 illustrates the empirical joint density of (β_x^t, β_y^t). The negative cross-correlation structure is clearly visible. Figure 3 shows how good the approximation of the cross covariances given in (11) is, for α and β (using Gibbs sampling to update also α).

To give a quantitative measure of the variance reduction using the antithetic chains, we estimate the integrated autocovariance time using all 100,000 iterates; see Geyer (1992). The ratios $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$, for f projecting on the single components α and β, are listed in Table 1 for the three different updating rules and the three visitation schemes. These ratios do not seem to depend significantly on the visitation schedules. The variance reduction factors for the Gibbs samplers are around 9 and 6 for α and β, respectively. The variance

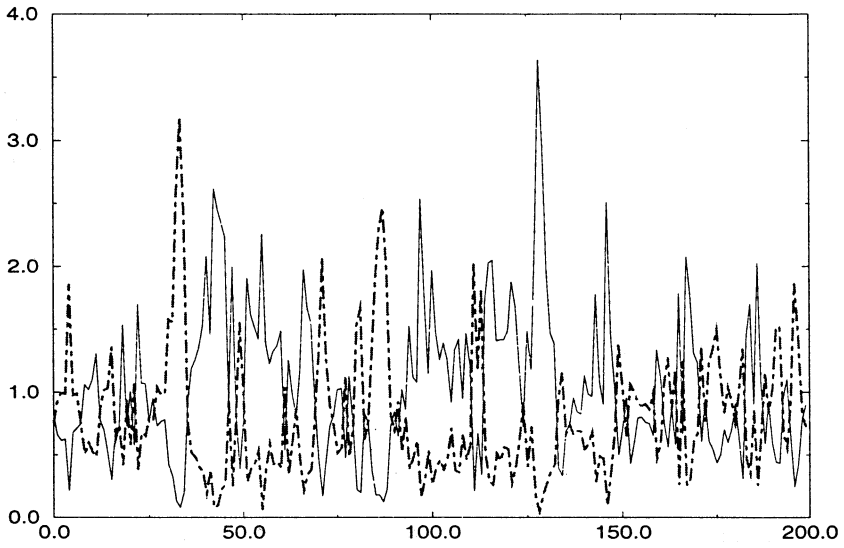

FIG. 1. *The sample-paths of the β component of the two antithetically coupled Gibbs sampler chains show a clear negative correlation.* 200 *consecutive iterations.*

reduction of the two other algorithms (Hastings update and Metropolis update) drops to around 2–2.5 for α. This occurs despite the fact that the acceptance rate was 90% for the Hastings update. A further experiment using a random walk Metropolis update, for α with a uniform proposal with larger width and an acceptance rate of 25%, still gave a variance reduction around 2. The reason for this is that the two antithetic chains get out of phase when an antithetic proposal is rejected by one chain but not by the other; the antithetic coupling between the two chains weakens. We do not adjust for this in later iterations, since only shared random numbers are used to introduce antithetic dependency between the two chains and the current states of the two chains are not considered in the proposal. The antithetic Gibbs sampler is also better than a single hybrid $2T$ long chain, using Gibbs sampling for $\lambda_1, \ldots, \lambda_{10}, \beta$ and a Hastings update for α. The asymptotic variance of such a hybrid sampler is larger than that of a single pure Gibbs sampler.

We now include the biases in our analysis. Let the efficiency be defined as the squared bias plus the variance and consider the ratio

$$\frac{\text{bias}(\hat{f})^2 + \text{Var}(\hat{f})}{\text{bias}(\hat{\hat{f}})^2 + \text{Var}(\hat{\hat{f}})}. \tag{17}$$

The estimated ratios (17) regarding the estimation of α and β, for the pure Gibbs sampler and the deterministic visitation scheme are 30.9 and 12.1, respectively. The true posterior mean values of α and β, needed in the bias calculation, were estimated with a very long (antithetic) run (4,000,000 full sweeps). Exact sampling or numerical integration could also have been used

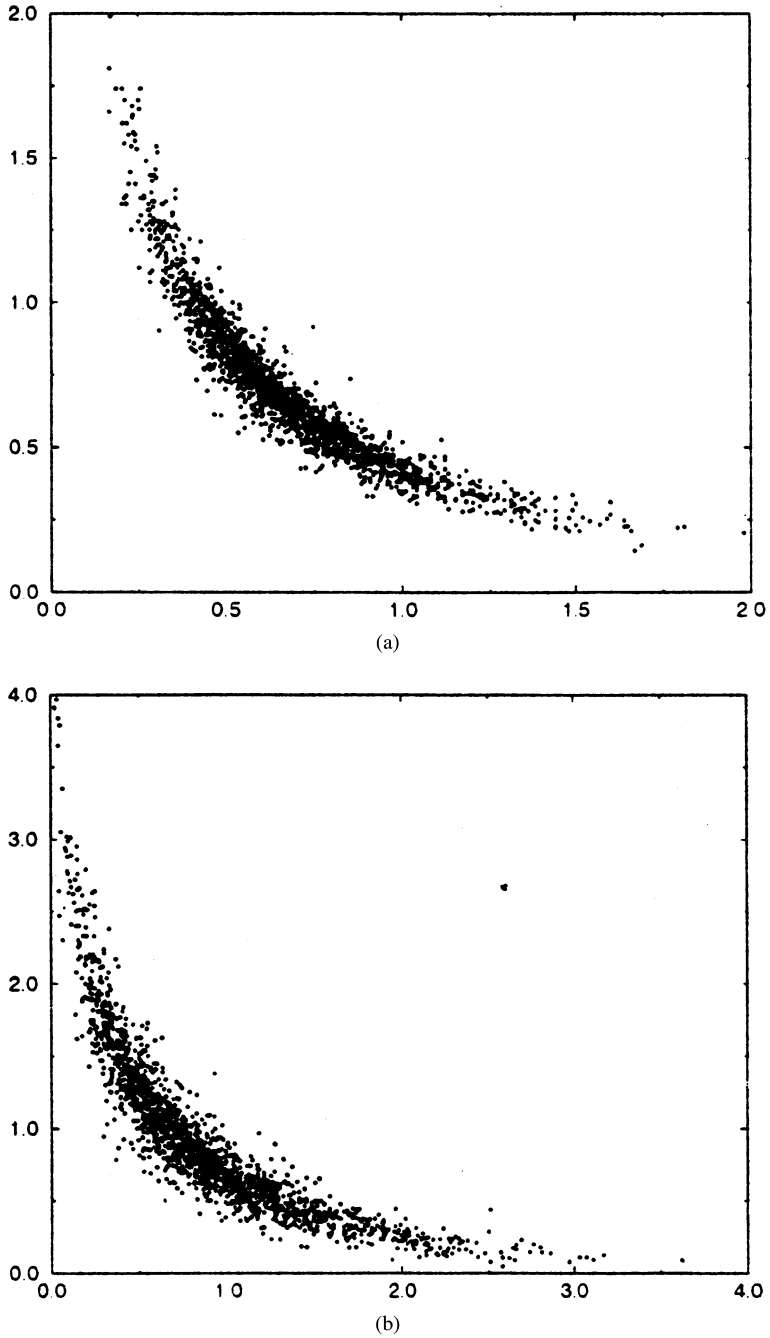

FIG. 2. *Point plots of* 5,000 *samples from the two antithetic Gibbs sampler chains for* (*a*) α *and* (*b*) β. *The support of the joint density and a clear negative correlation are illustrated.*

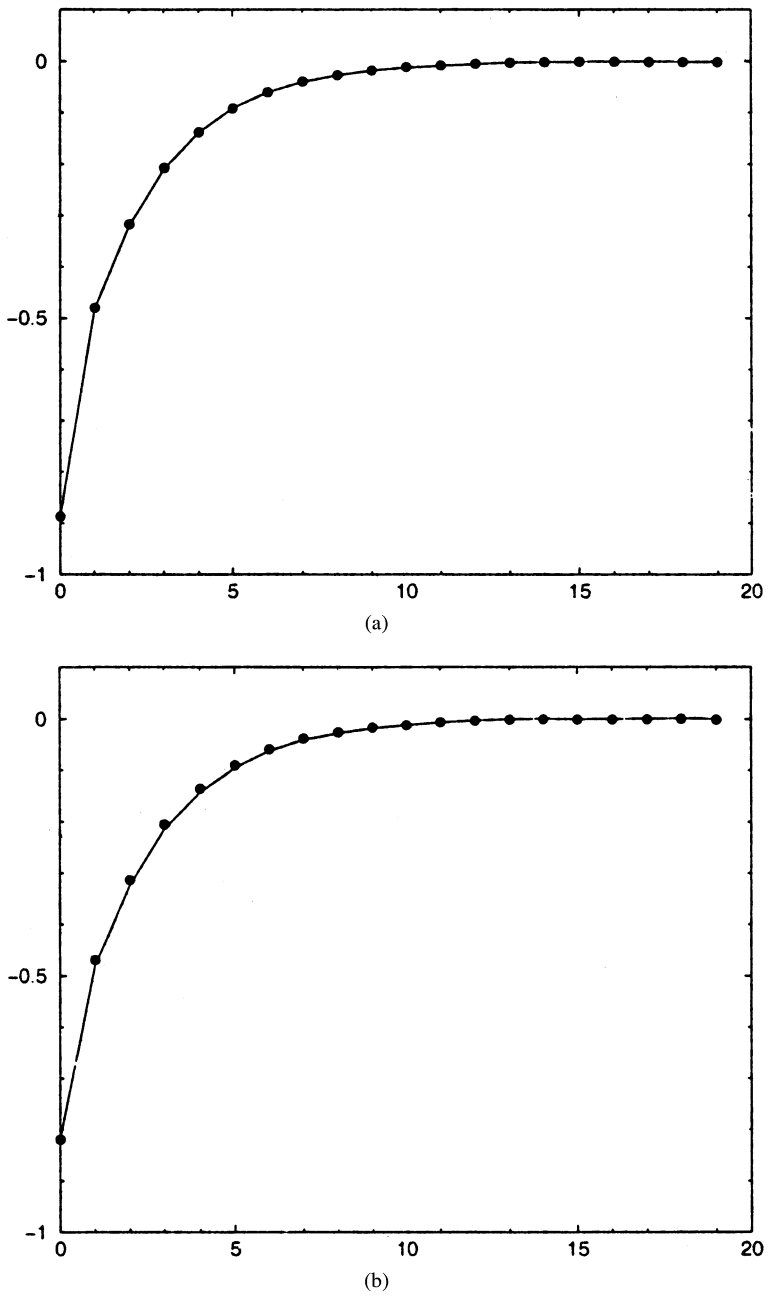

FIG. 3. *The estimated cross-autocorrelation (solid line) for* α *[in (a)] and* β *[in (b)] together with the approximated cross-correlation (dots) based on approximation* (11)*, for the pump example using the Gibbs sampler and the visitation schedule DET. The approximation is very good.*

TABLE 1

Hierarchical Poisson model: *estimated* $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$ *for* f *equal to* α *or* β*

Estimated $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$	Gibbs sampler			Gibbs–Hastings			Gibbs–Metropolis		
	RS	**RPS**	**DET**	**RS**	**RPS**	**DET**	**RS**	**RPS**	**DET**
α	9.53	9.00	9.64	2.33	2.23	2.46	2.31	2.13	2.05
β	6.56	6.40	6.05	3.05	2.97	2.60	2.72	2.50	2.39

*Three different ways of updating the parameter α and three different scan strategies are compared using 100,000 iterates. The antithetic coupling is very convenient for the pure Gibbs sampler, but the variance reduction decreases using a Hastings-update or a Metropolis-update for α.

for this purpose. We can conclude that the new antithetic algorithm is still better than a single long chain. This is surprising, but it seems that the bias of the average based on the $\mathbf{X}^t$ chain has the opposite sign to the bias of the estimator based on the antithetic $\mathbf{Y}^t$ chain, so that these contributions to the bias of $\hat{\hat{f}}$ cancel. This seems to be a further advantage for the new method. In Figure 4 we plot the bias of these two chains. Observe the antithetic sign. In the same figures (one for α and one for β) we have also plotted the total bias of the antithetic Gibbs sampler, which oscillates around zero. To avoid a further figure we have shrunk the time of this total bias, so that it can be compared to the bias of the estimate based on $\mathbf{X}^t$ (or on $\mathbf{Y}^t$). This now corresponds to the bias of a single, twice as long, run. The bias is smaller.

7.2. *The ordered normal means problem.* Gelfand, Hills, Racine-Poon and Smith (1990) use the Gibbs sampler to estimate the mean and precision in normal populations, when the ordering of the means is known in advance. We have repeated their example using the antithetic Gibbs sampler to investigate its variance reduction and efficiency in estimating the posterior mean of the parameters of interest.

Let Y_{ij} be the jth observation ($j = 1, \ldots, n_i$) from the ith group ($i = 1, \ldots, n_g$). Assuming conditional independence throughout, let $Y_{ij} \sim N(\theta_i, 1/\tau_i)$, $\theta_i \sim N(\mu, 1/\tau_g)$, $\tau_i \sim \Gamma(a_1, b_1)$, $\tau_g \sim \Gamma(a_2, b_2)$, and $\mu \sim N(\mu_0, 1/\tau_0)$. Here τ_i, τ_g, τ_0 denote the precision or inverse variance. A priori it is known that the means θ_i satisfy the constraint $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{n_g}$. Gelfand, Hills, Racine-Poon and Smith (1990) demonstrate that the Gibbs sampler is easy to implement even in this case. We refer to Gelfand, Hills, Racine-Poon and Smith (1990) for details about the Gibbs sampler and for the specific choices of the (flat) priors of the hyperparameters.

We simulated data using $n_g = 5$ and sampled from the ith population $n_i = 2i+4$ observations from $\mathrm{N}(i, i^2)$. Table 2 lists the empirical mean and variance within each group. Note that the observed ordering of the means is not in agreement with the a priori constraint. We used the deterministic schedule DET with a burn-in of 1000 cycles. The variance reduction factor $\mathrm{Var}(\hat{f})/(\hat{\hat{f}})$ was estimated using the following 50,000 iterates of the coupled chains as

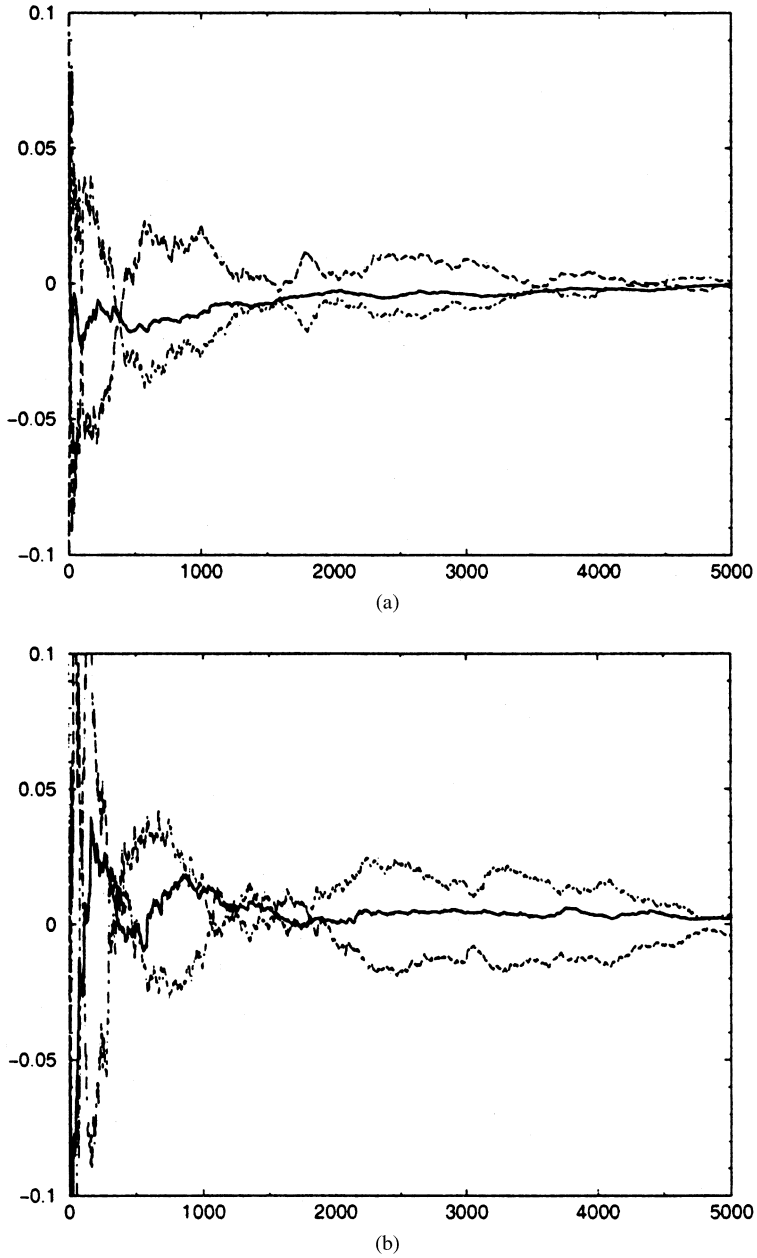

FIG. 4. *Bias in estimation of the posterior mean for* α [*in* (*a*)] *and* β [*in* (*b*)] *for each of the two Gibbs sampler chains* $\mathbf{X}^t$ *and* $\mathbf{Y}^t$ (*dashed and dot–dashed lines*) *and the antithetic Gibbs sampler* (*solid line*) *as a function of number of iterations* (*counting two for the antithetic Gibbs sampler*). *The bias for the antithetic Gibbs sampler is scaled so the amount of computational work is comparable.*

TABLE 2
*Ordered normal means problem: characteristics of the simulated data**

Sample values	**1**	**2**	**3**	**4**	**5**
n_i	6	8	10	12	14
$\overline{Y}_i$	0.645	2.212	3.576	2.401	4.195
S_i^2	1.473	2.279	3.452	20.186	11.330

*Note the exchange in the empirical ordering of the means.

in Section 7.1 Table 3 displays the estimated ratios for (θ_i, τ_i), $i = 1, \ldots, n_g$. The new antithetic Gibbs sampler gives a significant speedup with variance reduction between 2.97 and 6.69 with an average of 4.7. Similar results were obtained for the other visiting schedules.

8. Conclusions. We have suggested a simple way to couple two Gibbs sampler chains in order to reduce the variance of the empirical average as an estimator of an expectation. The coupling induces negative cross covariances. The new estimator is also asymptotically unbiased and the reduction of the variance can be remarkable with respect to the simple Gibbs sampler run for the same time. The coding of the proposed algorithm is easy, given a standard Gibbs sampler implementation.

Other authors have introduced antithetic behaviors into a single MCMC chain. If the density π is symmetric around zero, Geweke (1988) proposes using the estimator $(1/T)\sum_{t=1}^{T/2}(f(\mathbf{X}^t) + f(-\mathbf{X}^t))$ and proves that its asymptotic variance is smaller than Var $(\hat{f})$. Barone and Frigessi (1989) propose a variation of the Gibbs sampler where each step moves antithetically to the current state and show a faster weak convergence rate in some cases. Neal (1998) improves the single updating step further. Green and Han (1992) show that in such a way the asymptotic variance could also be reduced in certain special cases. We show that with two chains a more authentic antithetic behavior can be established.

As the example showed, it is not trivial to extend equally successfully the antithetic idea to Metropolis–Hastings type algorithms. It is more difficult to induce antithetic correlation when an accept–reject step may well reject a proposed antithetic move. More research is needed in order to understand how to couple such chains properly.

TABLE 3
The estimated variance reduction of the estimates based on antithetic coupled Gibbs sampler w.r.t. *the estimates based on a simple Gibbs sampler in the ordered normal means problem using* 50,000 *iterates*

Estimated Var($\hat{f}$)/ Var($\hat{\hat{f}}$)	**1**	**2**	**3**	**4**	**5**
θ_i	5.44	4.02	2.97	3.09	4.31
τ_i	4.20	5.08	4.71	6.69	6.53

The Gibbs sampler is rarely the fastest MCMC algorithm. In fact, other Metropolis–Hastings schemes have often a smaller asymptotic variance. However, the new antithetically coupled Gibbs sampler may compete with such algorithms. A further interesting idea, suggested to us by an Associate Editor, is to apply the antithetic chains to slice sampling, which relies on uniform distributions.

It is also possible to couple more than two chains. Four chains could share the random variates $U, 1-U, (U+0.5) \bmod(1), (1-U-0.5) \bmod(1)$ and so on. This is likely to lead to a further gain, though we expect that the advantage with respect to only two chains is not large.

APPENDIX

A.1. Proof of Theorem 1.

LEMMA 1. *Suppose $f \in \mathscr{F}$ and let π be attractive. Consider the coupled Gibbs sampler chains given in* (4) *and* (5). *If the components $X_1^0, \ldots, X_n^0$, $Y_1^0, \ldots, Y_n^0$ are generated independently and if a deterministic raster scan is used, then $\beta(t, t+k) \le 0$ and $\operatorname{Cov}(f(\mathbf{X}^t), f(\mathbf{X}^{t+k})) \ge 0$ for all $t \ge 0$ and $k \ge 0$. The same assertions hold true if $\mathbf{X}^0$ and $\mathbf{Y}^0$ are drawn independently from π and either a deterministic raster or a random scan is used.*

PROOF OF LEMMA 1. The proof relies on the construction of sets of associated random variables. We shall use properties of associated random variables referenced as P1 to P4 in Esary, Proschan and Walkup (1967).

Deterministic scan. First assume that $X_1^0, \ldots, X_n^0$, $Y_1^0, \ldots, Y_n^0$ are independent. The component i^t is updated in the transition from $(\mathbf{X}^t, \mathbf{Y}^t)$ to $(\mathbf{X}^{t+1}, \mathbf{Y}^{t+1})$, which happens according to $X_{i^t}^{t+1} = \Phi_{i^t}(\mathbf{X}_{-i^t}^t, U^t)$ and $Y_{i^t}^{t+1} = \Phi_{i^t}(\mathbf{Y}_{-i^t}^t, 1-U^t)$. Then $\Phi_{i^t}(\cdot, \cdot)$ is nondecreasing in each variable, by attractivity and because of the monotonicity of inverse conditional distribution functions. Now suppose

$$\{X_1^t, \ldots, X_n^t, -Y_1^t, \ldots, -Y_n^t\} \tag{18}$$

is a set of associated random variables. Then also $S^t = \{X_1^t, \ldots, X_n^t, -Y_1^t, \ldots, -Y_n^t, U^t\}$ are associated, since U^t is independent of the other variables (P2). By the monotonicity of Φ_{i^t} it follows that $X_{i^t}^{t+1}$ and $-Y_{i^t}^{t+1}$ are nondecreasing functions of the variables in S^t. For $j \ne i^t$, X_j^{t+1} and $-Y_j^{t+1}$ are trivially nondecreasing functions of the variables in S^t. Hence $\{X_1^{t+1}, \ldots, X_n^{t+1}, -Y_1^{t+1}, \ldots, -Y_n^{t+1}\}$ is also a set of associated random variables (P4). Now if $X_1^0, \ldots, X_n^0$, $Y_1^0, \ldots, Y_n^0$ are independent, then in particular S^0 is associated (P2), and by induction it follows that (18) is a set of associated random variables for all t. For fixed t, it follows in the same way that $\{X_1^t, \ldots, X_n^t, -Y_1^{t+k}, \ldots, Y_n^{t+k}\}$ is associated for each $k \ge 0$. We now use induction on k, changing only $-Y_{i^{t+k-1}}$ in the kth step.

Define two functions, $g(x, -y) = f(x)$ and $h(x, -y) = -f(y) = -f(-(-y))$. Because f is nondecreasing, then g and h are nondecreasing functions of $\{x_1, \ldots, x_n, -y_1, \ldots, -y_n\}$, and it follows by associativity [Esary, Proschan and Walkup (1967), Definition 1.1] that

$$\operatorname{Cov}(f(\mathbf{X}^t), -f(\mathbf{Y}^{t+k})) = \operatorname{Cov}(g(\mathbf{X}^t, -\mathbf{Y}^{t+k}), h(\mathbf{X}^t, -\mathbf{Y}^{t+k})) \geq 0.$$

Changing the sign gives the asserted nonpositivity of the cross covariances for each t.

An induction argument similar to the one above shows that the sets $\{X_1^t, \ldots, X_n^t, X_1^{t+k}, \ldots, X_n^{t+k}\}$, are associated for each $k \geq 0$ and each $t \geq 0$. Replacing the function h in the preceding argument by $h(x, y) = f(y)$, we obtain $\operatorname{Cov}(f(\mathbf{X}^t), f(\mathbf{X}^{t+k})) = \operatorname{Cov}(g(\mathbf{X}^t, \mathbf{X}^{t+k}), h(\mathbf{X}^t, \mathbf{X}^{t+k})) \geq 0$. Taking the limit as $t \to \infty$ we also have that $\gamma_k \geq 0$.

We move now to the actual assumption of Theorem 1, that $\{X_1^0, \ldots, X_n^0\}$ and $\{Y_1^0, \ldots, Y_n^0\}$ are π-distributed and independent. Since π is attractive, for $i = 1, \ldots, n$ and arbitrary x_i, we have that $P(X_i \geq x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1})$ is nondecreasing in $x_1, \ldots, x_{i-1}$ if $\mathbf{X}$ is distributed according to π. This means that the variables $X_1, \ldots, X_n$ are conditionally nondecreasing in sequence. By Barlow and Proschan (1975) they are associated. Since association is preserved by multiplying all variables by -1, the same holds for $-X_1, \ldots, -X_n$. Therefore, if the initial state $(\mathbf{X}^0, \mathbf{Y}^0)$ is obtained by drawing $\mathbf{X}^0$ and $\mathbf{Y}^0$ independently from π, the set $\{X_1^0, \ldots, X_n^0, -Y_1^0, \ldots, -Y_n^0\}$ is associated. Therefore, the same argument used for the case of independent components in the initial state can be followed.

Random Scan. Let the initial states $\mathbf{X}^0$ and $\mathbf{Y}^0$ be π-distributed and independent. Let $\mathbf{I}^t = (I^0, \ldots, I^{t-1})$ be the site updating sequence. Then it holds that

$$\begin{aligned}\operatorname{Cov}(f(\mathbf{X}^t), f(\mathbf{Y}^{t+k})) &= \mathrm{E}\big(\operatorname{Cov}\big(f(\mathbf{X}^t), f(\mathbf{Y}^{t+k})\big) \mid \mathbf{I}^{t+k}\big) \\ &\quad + \operatorname{Cov}\big(\mathrm{E}\big(f(\mathbf{X}^t) \mid \mathbf{I}^{t+k}\big), \mathrm{E}\big(f(\mathbf{Y}^{t+k}) | \mathbf{I}^{t+k}\big)\big).\end{aligned} \tag{19}$$

If $\mathbf{X}^0$ is distributed according to π, then $\mathbf{X}^t$ given I^{t+k} is also π-distributed for all t. Hence, $\mathrm{E}(f(\mathbf{X}^t) \mid \mathbf{I}^{t+k}) = \mathrm{E}(f(\mathbf{Y}^{t+k}) \mid \mathbf{I}^{t+k}) = 0$ and the second term in (19) is zero. The proof for the deterministic scan shows that the first term in (19) is nonpositive. By the same argument, $\operatorname{Cov}(f(\mathbf{X}^t), f(\mathbf{X}^{t+k})) \geq 0$ for all $k \geq 0$, $t \geq 0$ if $\mathbf{X}^0$ is drawn from π and a random scan is used. Again, letting $t \to \infty$ we get $\gamma_k \geq 0$ for all $k \geq 0$. □

PROOF OF THEOREM 1. We compute first $\operatorname{Var}(\hat{f})$ and $\operatorname{Var}(\hat{\hat{f}})$ as functions of $\beta(t, s)$and γ_k. Then Theorem 1 follows using Lemma 1. The variances are as follows:

$$\operatorname{Var}(\hat{f}) = \operatorname{Var}\left(\frac{1}{2T}\sum_{t=1}^{2T} f(\mathbf{X}^t)\right) = \frac{1}{2T}\gamma_0 + \frac{1}{T}\sum_{k=1}^{2T-1}\gamma_k\left(1 - \frac{k}{2T}\right),$$

$$\operatorname{Var}(\hat{\hat{f}}) = \operatorname{Var}\left(\frac{1}{2T}\sum_{t=1}^{T}(f(\mathbf{X}^t) + f(\mathbf{Y}^t))\right)$$
$$= \frac{1}{2T}\gamma_0 + \frac{1}{T}\sum_{k=1}^{T-1}\gamma_k\left(1 - \frac{k}{T}\right) + \frac{1}{2T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\beta(s,t).$$

Thus, $T(\operatorname{Var}(\hat{f}) - \operatorname{Var}(\hat{\hat{f}})) = S - D$, where

$$S = \sum_{k=T}^{2T-1}\gamma_k\left(1 - \frac{k}{2T}\right) + \frac{1}{2T}\sum_{k=1}^{T-1}k\gamma_k, \quad D = \frac{1}{2T}\sum_{t=1}^{T}\sum_{s=1}^{T}\beta(s,t).$$

We can now study the sign of $S - D$ when π is attractive and $f \in \mathscr{F}$, in the case of random or raster scan. Using Lemma 1 we have that $\gamma_k \geq 0$ for all k and $\beta(s,t) \leq 0$ for all s and t, hence $S \geq 0$ and $D \leq 0$ for all T. This concludes the proof. □

Notice that the proof of Lemma 1 does not require that the components to be updated in the two chains are the same. Hence variance reduction is achieved also if the two chains update different components at each step. However, we expect strongest variance reduction when the components are the same.

PROOF OF THEOREM 2. When the ith component is updated, we can write the usual Gibbs sampler in matrix notation as

$$\mathbf{X}^{t+1} = (\mathbf{I} - \mathbf{D}_i\mathbf{Q})\mathbf{X}^t + \boldsymbol{\varepsilon}^t, \tag{20}$$

where $\mathbf{Q}$ is the inverse covariance matrix, $\mathbf{D}_i$ is a matrix of zeros, except for a single 1 in ith position along the diagonal and $\boldsymbol{\varepsilon}^t$ is a vector of zeros except for ε_i^t which is a normal variate with zero mean and variance 1. Similarly, $\mathbf{Y}^{t+1} = (\mathbf{I} - \mathbf{D}_i\mathbf{Q})\mathbf{Y}^t + \eta^t$. Due to the antithetic coupling and the normal assumption, $\boldsymbol{\eta}^t = -\boldsymbol{\varepsilon}^t$. Hence

$$\mathbf{X}^{t+1} + \mathbf{Y}^{t+1} = (\mathbf{I} - \mathbf{D}_i\mathbf{Q})(\mathbf{X}^t + \mathbf{Y}^t). \tag{21}$$

It is shown in Barone and Frigessi (1989) that the spectral radius q of the matrix

$$\mathbf{A}_n = (\mathbf{I} - \mathbf{D}_n\mathbf{Q}) \times \cdots \times (\mathbf{I} - \mathbf{D}_1\mathbf{Q}) \tag{22}$$

that governs a full deterministic raster scan is strictly smaller than one. The spectral radius of the single components $\mathbf{I} - \mathbf{D}_i\mathbf{Q}$ is smaller or equal to 1. Consider a linear function f with zero π-mean. We can write (for T a multiple of n)

$$\hat{\hat{f}} = \frac{1}{2T}\sum_{k=0}^{T/n-1}\sum_{s=0}^{n-1}\left(f(\mathbf{X}^{nk+s}) + f(\mathbf{Y}^{nk+s})\right),$$

so that

$$|\hat{\hat{f}}| \leq \frac{n}{2T}\sum_{k=0}^{T/n-1}q^k|f(\mathbf{X}^0) + f(\mathbf{Y}^0)|$$

and

$$\mathrm{Var}(\hat{\hat{f}}) \le \frac{cn^2}{4T^2}\left(\frac{1-q^{T/n}}{1-q}\right)^2$$

for some positive constant c, which gives $\mathrm{Var}(\hat{\hat{f}}) = \mathcal{O}(T^{-2})$. Furthermore, $\mathrm{Var}(\hat{f}) \sim c'/T$ for some constant c'. Hence for T large enough, $\mathrm{Var}(\hat{f}) \ge \mathrm{Var}(\hat{\hat{f}})$. □

PROOF OF THEOREM 3. We follow Arjas and Gasbarra (1996). A φ-irreducible Markov chain has to be either transient or recurrent. If $\mathbf{Z}^t$ is recurrent, it has to be positive recurrent since the marginals are positive recurrent. We show that it cannot be transient. Fix a starting point $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \overline{\mathrm{supp}\{\varphi\}}$. For each k, the joint k-step transition kernel $K_Z^n(\mathbf{z}, \cdot)$ is a coupling of the marginal k-step transition kernels $K_X^k(\mathbf{x}, \cdot)$ and $K_Y^k(\mathbf{y}, \cdot)$. Since the marginal chains are positive recurrent, it follows that the sequences of probability measures $\{K_X^k(\mathbf{x}, \cdot),\ k \in N\}$ and $\{K_Y^{k'}(\mathbf{y}, \cdot),\ k \in N\}$ are tight [see, e.g., Meyn and Tweedie (1993)]. This implies tightness of the sequence of couplings $\{K_Z^k(\mathbf{z}, \cdot),\ k \in N\}$ in the product space, as well as the tightness in $\overline{\mathrm{supp}\{\varphi\}}$ with the relative topology. This is because if A is a topological space, B is a closed subset and K is compact in A, then $B \cap K$ is closed w.r.t. the relative topology. In particular, there is a compact set C (compact w.r.t. the relative topology) such that $K_Z^k(\mathbf{z}, C) > 1/2$, $\forall\, k$. It is not a restriction to take $\mathbf{z} \in C$ (add $\mathbf{z}$ to C if necessary). Since $\sum_{k=1}^{\infty} K_Z^k(\mathbf{z}, C) = \infty$, C is a compact set which is not uniformly transient. By the lemma in the Appendix of Arjas and Gasbarra (1996), it follows that the Markov chain $\mathbf{Z}^t$ is not transient. □

THEOREM 4. *Let $P(\cdot \mid (\mathbf{x}, \mathbf{y}))$ be a transition kernel on $\Omega \times \Omega$ which updates only one block component (x_i, y_i) at each transition, and which is ergodic with joint stationary measure μ. Suppose that $\pi(\mathbf{x}) > 0$ for all $\mathbf{x} \in \Omega = S \times \cdots \times S$ and* $\mathrm{supp}(\mu) = \{(\mathbf{x}, \mathbf{y}) \in \Omega \times \Omega | \mathbf{Y} = H(\mathbf{X})\}$. *Then this support can be defined componentwise: for each i, y_i is function of x_i only.*

PROOF. We proceed by induction. At step i, consider $(y_1, \ldots, y_i) = h_i(x_1, \ldots, x_i)$ for some function h_i. This is true for $i = n$ with $h_n = H$. Suppose that the chain is stationary at time t, and $(\mathbf{X}^t, \mathbf{Y}^t) = (\mathbf{x}, \mathbf{y})$. Assume that $I^t = i$. Then $(\mathbf{X}^{t+1}, \mathbf{Y}^{t+1}) = ((x_i', \mathbf{x}_{-i}), (y_i', \mathbf{y}^{-i}))$ belongs to the support of μ, by stationarity. Hence, the induction assumption says that $(y_1, \ldots, y_i') = h_i(x_1, \ldots, x_i')$. In particular $(y_1, \ldots, y_{i-1}) = h_i(x_1, \ldots, x_i')_{-i}$ is a function of $(x_1, \ldots, x_i')$. But we also have $(y_1, \ldots, y_{i-1}) = h_i(x_1, \ldots, x_i)_{-i}$. By the assumption on π, x_i' could be any point in S, and it follows that $h_i(x_i, \ldots, x_{i-1})_{-i}$ is constant as a function of x_i'. Hence $(y_1, \ldots, y_{i-1})$ is determined by $(x_1, \ldots, x_{i-1})$ only. Proceeding by induction back from $i = n$ to $i = 1$, it follows that y_1 is a function of x_1. The ordering of the components is arbitrary, so that y_k is a function of x_k only, and the theorem follows. □

Acknowledgment. We thank Dario Gasbarra who provided us with the proof of Theorem 3.

REFERENCES

ARJAS, E. and GASBARRA, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *J. Amer. Statist. Assoc.* **91** 1101–1109.

BARLOW, R. E. and PROSCHAN, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston, New York.

BARONE, P. and FRIGESSI, A. (1989). Improving stochastic relaxation for Gaussian random fields. *Probab. Engng. Inform. Sci.* **3** 369–389.

ESARY, J. D., PROSCHAN, J. D. and WALKUP, D. W. (1967). Association of random variables, with applications. *Ann. Math. Statist*. **38** 651–655.

GELFAND, A. E., HILLS, S. E., RACINE-POON, A. and SMITH A. F. M. (1990). Illustration of Bayesian inference in normal data models using the Gibbs sampler. *J. Amer. Statist. Assoc.* **85** 972–985.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–509.

GEWEKE, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayseian inference. *J. Econometrics* **38** 73–90.

GEYER, C. (1992). Practical Markov chain Monte Carlo (with discussion). *Statist. Sci*. **7** 473–511.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

GREEN, P. J. and HAN, X. L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables. *Stochastic Models, Statistical Methods and Algorithms in Image Analysis. Lecture Notes in Statist.* **74** 142–164. Springer, Berlin.

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

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

MØLLER, J. (1999). Perfect simulation of conditionally specified models. *J. Roy. Statist. Soc. Ser. B* **61** 251–264.

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

A. FRIGESSI
NORWEGIAN COMPUTING CENTER
P.O. BOX 114 BLINDERN
N-0314 OSLO
NORWAY
E-MAIL: Arnoldo.Frigessi@nr.no

J. GÅSEMYR
DEPARTMENT OF MATHEMATICS
P.O. BOX 1053 BLINDERN
N-0316 OSLO
NORWAY
E-MAIL: gassemyr@math.uio.no

H. RUE
DEPARTMENT OF MATHEMATICAL SCIENCES
NORWEGIAN UNIVERSITY FOR SCIENCE
AND TECHNOLOGY
N-7491 TRONDHEIM
E-MAIL: Havard.Rue@math.ntnu.no