DISCUSSION ARTICLE

Comment

James P. HOBERT and Jorge Carlos ROMÁN

We begin by congratulating Professors Yu and Meng on an outstanding article, and thanking Professor Levine for giving us the opportunity to discuss their work. Our discussion focuses mainly on the GIS and ASIS algorithms. Section 1 concerns the relationship between the GIS and sandwich algorithms. In Section 2, we consider a family of toy GIS algorithms based on the bivariate normal distribution, and show how this family is related to the toy example in Section 2 of the article by Yu and Meng (2001) (hereafter Y&M). Finally, in Section 3, we provide a simple example of a non-reversible GIS algorithm.

1. {DA ALGORITHMS} ⊂ {SANDWICH ALGORITHMS} ⊂ {GIS ALGORITHMS}

Let $f_X : \mathsf{X} \to [0, \infty)$ be an intractable target density, and suppose that $f : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$ is a joint density whose x-marginal is the target; that is, $\int_{\mathsf{Y}} f(x, y)\, dy = f_X(x)$. If straightforward sampling from the associated conditional densities is possible, then we can use the data augmentation (DA) algorithm to explore f_X. Of course, running the algorithm entails alternating between draws from $f_{Y|X}$ and $f_{X|Y}$, which simulates the Markov chain whose Markov transition density (Mtd) is

$$k_{\mathrm{DA}}(x'|x) = \int_{\mathsf{Y}} f_{X|Y}(x'|y) f_{Y|X}(y|x)\, dy.$$

If we denote the DA Markov chain by $\{X_n\}_{n=0}^{\infty}$, then $k_{\mathrm{DA}}(\cdot|x)$ is simply the conditional density of X_{n+1} given that $X_n = x$. It is easy to see that $k_{\mathrm{DA}}(x'|x) f_X(x)$ is symmetric in (x, x'), so the DA Markov chain is reversible. We assume throughout that all Markov chains on the target space, X, satisfy the usual regularity conditions: Harris recurrence, irreducibility, and aperiodicity.

James P. Hobert is Professor (E-mail: *jhobert@stat.ufl.edu*) and Jorge Carlos Román is Graduate Student (E-mail: *jcroman7@stat.ufl.edu*), Department of Statistics, University of Florida, 221 Griffin–Floyd Hall, P.O. Box 118545, Gainesville, FL 32611-8545.

Journal of Computational and Graphical Statistics, Volume 20, Number 3, Pages 571–580
DOI: 10.1198/jcgs.2011.203a

Following Liu and Wu (1999), Meng and van Dyk (1999), and van Dyk and Meng (2001), Hobert and Marchev (2008) introduced an alternative to DA that employs an extra move on the Y space that is "sandwiched" between the two conditional draws. Define $f_Y(y) = \int_X f(x, y)\,dx$ and suppose that $R(y, dy')$ is *any* Markov transition function (Mtf) on Y that is reversible with respect to f_Y; that is, $R(y, dy')f_Y(y)\,dy = R(y', dy)f_Y(y')\,dy'$. The *sandwich algorithm* simulates the Markov chain whose Mtd is

$$k_S(x'|x) = \int_Y \int_Y f_{X|Y}(x'|y')R(y, dy')f_{Y|X}(y|x)\,dy.$$

Again, it is easy to see that $k_S(x'|x)f_X(x)$ is symmetric in (x, x'). To run the sandwich algorithm, we simply run the DA algorithm as usual, except that after each y is drawn, we perform the extra step $y' \sim R(y, \cdot)$ before drawing the new x. In practice, R is usually chosen so that, for fixed y, the chain driven by $R(y, \cdot)$ lives in a one-dimensional subspace of Y. Consequently, drawing from R is much less expensive (computationally) than drawing from the conditional densities. Note that the sandwich algorithm reduces to DA if we take R to be the trivial Mtf whose chain is absorbed at the starting point.

Here is our interpretation of Y&M's GIS algorithm. Suppose that, in addition to $f(x, y)$, we have another joint density $\tilde{f}: X \times Y \to [0, \infty)$ for which $\int_Y \tilde{f}(x, y)\,dy = f_X(x)$. Now let $Q(y, dy')$ be *any* Mtf on the space Y that satisfies

$$\int_Y Q(y, dy')f_Y(y)\,dy = \tilde{f}_Y(y')\,dy', \tag{1.1}$$

where $\tilde{f}_Y(y) = \int_X \tilde{f}(x, y)\,dx$. Note that $\tilde{f}_Y(y)$ need not be the same as $f_Y(y)$. Y&M's GIS algorithm simulates the Markov chain whose Mtd is

$$k_{YM}(x'|x) = \int_Y \int_Y \tilde{f}_{X|Y}(x'|y')Q(y, dy')f_{Y|X}(y|x)\,dy. \tag{1.2}$$

This chain is not necessarily reversible (see Section 3), but $f_X(x)$ is the invariant density. As with the sandwich algorithm, we allow the Markov chain defined by Q to be reducible, as long as the GIS chain itself is well behaved. We can see that the sandwich chain is a special case of the GIS chain by taking $\tilde{f}(x, y) = f(x, y)$ and $Q(y, dy') = R(y, dy')$.

Here is a simple example of a GIS algorithm. Given $f(x, y)$ and $\tilde{f}(x, y)$, let $Q_A(y, dy') = q_A(y'|y)\,dy'$, where the Mtd q_A is defined as

$$q_A(y'|y) = \int_X \tilde{f}_{Y|X}(y'|x)f_{X|Y}(x|y)\,dx. \tag{1.3}$$

Clearly, $\int_Y q_A(y'|y)f_Y(y)\,dy = \tilde{f}_Y(y')$, so condition (1.1) is satisfied. Now, plugging into (1.2), we see that the resulting GIS algorithm has Mtd given by

$$k_A(x'|x) = \int_Y \int_X \int_Y \tilde{f}_{X|Y}(x'|y')\tilde{f}_{Y|X}(y'|x'')f_{X|Y}(x''|y)f_{Y|X}(y|x)\,dy\,dx''\,dy',$$

which is exactly the Mtd associated with the *alternating scheme* defined by Y&M's Equation (2.12). Hence, the alternating scheme is a special case of GIS.

It seems that our definition of the GIS algorithm is slightly more general than that of Y&M. Suppose that $f(x, y)$ and $\tilde{f}(x, y)$ are fixed. In contrast with our version of GIS, in

which any Q satisfying (1.1) will do, Y&M restrict attention to those Q's that stem from joint distributions on $\mathsf{X} \times \mathsf{Y} \times \mathsf{Y}$ that are consistent with $f(x, y)$ and $\tilde{f}(x, y)$. For example, suppose that $g : \mathsf{X} \times \mathsf{Y} \times \mathsf{Y} \to [0, \infty)$ is a joint density such that

$$\int_{\mathsf{Y}} g(x, y, y')\, dy' = f(x, y) \qquad \text{and} \qquad \int_{\mathsf{Y}} g(x, y', y)\, dy' = \tilde{f}(x, y). \tag{1.4}$$

Then take $Q(y, dy') = q(y'|y)\, dy'$ where

$$q(y'|y) = \frac{\int_{\mathsf{X}} g(x, y, y')\, dx}{f_Y(y)}. \tag{1.5}$$

It is easy to see that this Q satisfies condition (1.1). An example of a joint density that satisfies (1.4) is $g(x, y, y') = f(x, y)\tilde{f}(x, y')/f_X(x)$. Interestingly, applying (1.5) with this particular g yields q_{A}.

On the other hand, given $f(x, y)$, $\tilde{f}(x, y)$ and an arbitrary Q that satisfies (1.1), there does not necessarily exist a joint distribution that is consistent with all three of these. Hence, the class of Q's considered by Y&M is a strict subset of the Q's satisfying (1.1). However, some of Y&M's theoretical results concerning the GIS algorithm still hold when the set of allowable Q's is expanded to include all those satisfying (1.1). An example is given later.

We now discuss relationships among the convergence rates of the Markov chains that we have defined. As in Y&M's Section 5.1, let $L_0^2(f_X)$ denote the set of real-valued functions with domain X that are square integrable and have mean zero with respect to f_X. Suppose that $k(x'|x)$ is a (generic) Mtd satisfying $\int_{\mathsf{X}} k(x'|x) f_X(x)\, dx = f_X(x')$. This Mtd defines an operator, $K : L_0^2(f_X) \to L_0^2(f_X)$, that maps $g \in L_0^2(f_X)$ to $Kg \in L_0^2(f_X)$ where

$$(Kg)(x) = \int_{\mathsf{X}} g(x')k(x'|x)\, dx'.$$

Let $\|K\|$ and $r(K)$ denote the norm and spectral radius of K, respectively, which both lie in $[0, 1]$. For definitions, see the work of Retherford (1993). As explained by Rosenthal (2003), $r(K)$ is equal to the (asymptotic) convergence rate of the Markov chain defined by k. Since $r(K) \leq \|K\|$ and small values of $r(K)$ are associated with fast convergence, the norm provides an upper bound on the "slowness" of the chain. In the special case where the chain is reversible, $r(K) = \|K\|$ and the chain is geometrically ergodic if and only if $\|K\| < 1$ (Roberts and Rosenthal 1997; Roberts and Tweedie 2001).

Let K_{DA} and K_{YM} denote the Markov operators defined by k_{DA} and k_{YM}, respectively. Also, let $\tilde{K}_{\mathrm{DA}}$ denote the Markov operator associated with the DA chain based on $\tilde{f}(x, y)$. Y&M's Theorem 1 states that

$$r(K_{\mathrm{YM}}) \leq \|K_{\mathrm{YM}}\| \leq \mathcal{R}_{12}\sqrt{\|K_{\mathrm{DA}}\|\|\tilde{K}_{\mathrm{DA}}\|}, \tag{1.6}$$

where $\mathcal{R}_{12}$ is the maximal correlation of the pair (Y_0, Y_1), whose joint probability distribution is $Q(y, dy') f_Y(y)\, dy$. We note that, in the formal statement of Y&M's Theorem 1, there is an explicit assumption concerning the existence of a joint distribution that is consistent with $f(x, y)$ and $\tilde{f}(x, y)$, but the proof (in Appendix B) does not seem to rely on this assumption.

The inequality (1.6) can be used to prove a new result concerning the sandwich algorithm. As mentioned earlier, we can recover the sandwich chain from the GIS chain by taking $\tilde{f}(x, y) = f(x, y)$ and $Q(y, dy') = R(y, dy')$. In this case, (1.6) becomes

$$r(K_S) = \|K_S\| \leq \mathcal{R}_{12}\sqrt{\|K_{DA}\|\|K_{DA}\|} = \mathcal{R}_{12}\|K_{DA}\|, \tag{1.7}$$

which was noted by Y&M (at the bottom of page 539–top of page 540). Now, $R(y, dy') \times f_Y(y)\,dy$ is actually the joint distribution of the first two steps of the stationary version of the Markov chain driven by R. Therefore, by Liu, Wong, and Kong's (1994) Lemma 2.3, $\mathcal{R}_{12} = \|K_R\|$, where $K_R : L_0^2(f_Y) \to L_0^2(f_Y)$ is the Markov operator defined by $R(y, dy')$. Hence, (1.7) becomes

$$r(K_S) = \|K_S\| \leq \|K_R\|\|K_{DA}\|. \tag{1.8}$$

The inequality (1.8) strengthens and generalizes a result of Hobert and Rosenthal (2007) who showed that, if K_S is a positive operator, then $\|K_S\| \leq \|K_{DA}\|$. The additional factor of $\|K_R\|$ on the right side of (1.8) is important because it shows that a sub-geometric DA chain can be transformed into a geometrically ergodic sandwich chain by adding an extra move according to a geometrically ergodic chain on the Y space. In other words, the left side of (1.8) will be less than 1 as long as $\|K_R\| < 1$, even if $\|K_{DA}\| = 1$.

It should be noted that, in most practical applications of the sandwich algorithm, R is *idempotent*; that is,

$$\int_Y R(y, dy'')R(y'', dy') = R(y, dy'). \tag{1.9}$$

Loosely speaking, if R is idempotent, then K_R is a projection and $\|K_R\| = 1$, so (1.8) becomes $\|K_S\| \leq \|K_{DA}\|$, which is exactly Hobert and Rosenthal's (2007) result (which is applicable here because K_S is a positive operator when R is idempotent). See the work of Khare and Hobert (2011) for more on this.

2. A FAMILY OF GIS ALGORITHMS FOR A NORMAL TARGET

Suppose that the target density, $f_X(x)$, is $N(\mu, \sigma^2)$, and that $f(x, y)$ is bivariate normal, denoted by $BN(\mu, \sigma^2, \theta, \tau^2, \rho)$. Obviously, the x-marginal of $f(x, y)$ is the target. We now specify an $\tilde{f}$ so that we can construct a family of GIS algorithms. Fix $c \in \mathbb{R}$. If $(U_1, U_2) \sim BN(\mu, \sigma^2, \theta, \tau^2, \rho)$, then $(W_1, W_2) = (U_1, U_2 + cU_1)$ is $BN(\mu, \sigma^2, \tilde{\theta}, \tilde{\tau}^2, \tilde{\rho})$, where $\tilde{\theta} = c\mu + \theta$, $\tilde{\tau}^2 = c^2\sigma^2 + \tau^2 + 2c\rho\sigma\tau$, and

$$\tilde{\rho} = \frac{c\sigma + \rho\tau}{\sqrt{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2}}.$$

Let $\tilde{f}(x, y)$ denote this second bivariate normal density, and note that the x-marginal of $\tilde{f}$ is still the target. Obviously, $f_Y(y)$ is $N(\theta, \tau^2)$ and $\tilde{f}_Y(y)$ is $N(\tilde{\theta}, \tilde{\tau}^2)$. Now, let $Q(y, dy') = q(y'|y)\,dy'$ where $q(\cdot|y)$ is a univariate normal density given by

$$N(y + c\mu + c\rho(\sigma/\tau)(y - \theta), c^2\sigma^2(1 - \rho^2)).$$

A simple calculation shows that $\int_{\mathbb{R}} q(y'|y)f_Y(y)\,dy = \tilde{f}_Y(y')$, so this Q satisfies condition (1.1). We now have all the ingredients that we need to construct a family of GIS

algorithms, indexed by (θ, τ^2, ρ) and c. A direct calculation shows that the resulting Mtd, $k_{\mathrm{YM}}(\cdot|x)$, is normal with mean

$$\mu + \rho\left[\frac{(c\sigma + \rho\tau)(\rho c\sigma + \tau)}{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2}\right](x - \mu)$$

and variance

$$\sigma^2 - \rho^2\sigma^2\left[\frac{(c\sigma + \rho\tau)(\rho c\sigma + \tau)}{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2}\right]^2.$$

Interestingly, $k_{\mathrm{YM}}(x'|x) f_X(x)$ is symmetric in (x, x'), so every member of this family of GIS chains is reversible. Consequently, the convergence rate (spectral radius) is exactly equal to the norm, $\|K_{\mathrm{YM}}\|$.

By Liu, Wong, and Kong's (1994) Lemma 2.3, we know that $\|K_{\mathrm{YM}}\|$ is equal to the maximal correlation of a random pair with joint density $k_{\mathrm{YM}}(x'|x) f_X(x)$. It is easy to show that this joint density is bivariate normal, so the maximal correlation is equal to the absolute value of the correlation in the bivariate normal, and we have

$$\|K_{\mathrm{YM}}\| = \frac{|\rho||c\sigma + \rho\tau||\rho c\sigma + \tau|}{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2}, \tag{2.1}$$

which is less than 1 if and only if $|\rho| < 1$. It is interesting to compare the exact value in (2.1) with the upper bound from Y&M's Theorem 1, that is, with the right side of (1.6). Liu, Wong, and Kong's (1994) Theorem 3.2 implies that $\|K_{\mathrm{DA}}\|$ is equal to the square of the maximal correlation of a random pair with joint density $f(x, y)$, which is bivariate normal. Thus, $\|K_{\mathrm{DA}}\| = \rho^2$. Similarly, $\|\tilde{K}_{\mathrm{DA}}\| = \tilde{\rho}^2$. Moreover, the joint distribution $Q(y, dy') f_Y(y)\, dy$ is also bivariate normal, and it follows that

$$\mathcal{R}_{12} = \frac{|\rho c\sigma + \tau|}{\sqrt{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2}}.$$

Therefore,

$$\begin{aligned}
\mathcal{R}_{12}\sqrt{\|K_{\mathrm{DA}}\|}\sqrt{\|\tilde{K}_{\mathrm{DA}}\|} &= \frac{|\rho c\sigma + \tau|}{\sqrt{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2}} \\
&\quad \times \sqrt{\rho^2}\sqrt{\frac{(c\sigma + \rho\tau)^2}{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2}} \\
&= \frac{|\rho||c\sigma + \rho\tau||\rho c\sigma + \tau|}{(c\sigma + \rho\tau)^2 + (1-\rho^2)\tau^2} \\
&= \|K_{\mathrm{YM}}\|.
\end{aligned}$$

So, for every member of this family of GIS algorithms, the upper bound on $\|K_{\mathrm{YM}}\|$ from Y&M's Theorem 1 is exactly equal to $\|K_{\mathrm{YM}}\|$.

When the norm of a Markov operator is zero, the corresponding algorithm is "perfect" in the sense that it produces an i.i.d. sample from the target. For example, it is clear that the DA algorithm based on $f(x, y)$ is perfect when $\rho = 0$, since in that case, $f(x, y) = f_X(x) f_Y(y)$. Equation (1.6) shows that the GIS algorithm must be perfect whenever one of the underlying DA algorithms is perfect.

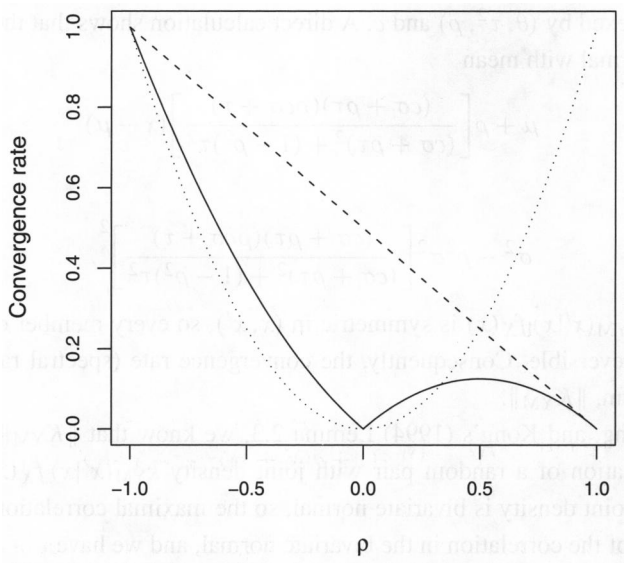

Figure 1. This plot shows how convergence rate varies with ρ for the DA chain based on f (dotted line), the DA chain based on $\tilde{f}$ (dashed line), and the GIS chain (solid line). In this example $\mu = \theta = 0$, $\sigma^2 = \tau^2 = 1$, and $c = -1$.

Now consider a specific example in which $\mu = \theta = 0$, $\sigma^2 = \tau^2 = 1$, and $c = -1$. In this case, $\|K_{\mathrm{DA}}\| = \rho^2$, $\|\tilde{K}_{\mathrm{DA}}\| = (1-\rho)/2$, and $\|K_{\mathrm{YM}}\| = |\rho|(1-\rho)/2$. Figure 1 shows a plot of the convergence rates of the three different algorithms as ρ ranges between -1 and 1. It is interesting to note that the first DA algorithm actually converges strictly faster than the GIS algorithm for all ρ in $(-1, 1/3) \setminus \{0\}$. This example serves as a warning that it is possible for a GIS algorithm to converge more slowly than the faster of the two underlying DA algorithms.

2.1 ANALYSIS OF Y&M'S NORMAL HIERARCHY

Consider the following simple hierarchy:

$$\begin{aligned} Z|X, Y &\sim \mathrm{N}(Y, 1), \\ Y|X &\sim \mathrm{N}(X, V), \\ X &\sim \mathrm{N}(0, A), \end{aligned} \tag{2.2}$$

where V and A are fixed positive numbers. This is the same as Y&M's normal hierarchy (from their Section 2.1), except that we have replaced the flat (Haar) prior by a proper normal prior. (Also, in order to keep our notation consistent, we are using (X, Y, Z) in place of Y&M's $(\theta, Y_{mis}, Y_{obs})$.) As in Y&M, we take the target to be the posterior density of X given the data, z. This posterior, which is denoted by $f_X(x)$, is

$$\mathrm{N}\left(\frac{Az}{V+A+1}, \frac{A(V+1)}{V+A+1}\right).$$

The joint density of (X, Y) given z, which we denote by $f(x, y)$, is bivariate normal with $\mu = Az/(V + A + 1)$, $\sigma^2 = A(V + 1)/(V + A + 1)$,

$$\theta = \frac{(A + V)z}{A + V + 1}, \qquad \tau^2 = \frac{A + V}{A + V + 1},$$

and

$$\rho = \frac{\sqrt{A}}{\sqrt{(A + V)(V + 1)}}.$$

Hence, we can employ the GIS algorithm developed earlier in this section. For general c, the convergence rate is given by

$$\|K_{\mathrm{YM}}\| = \frac{A^2|1 + V/A + c||1 + c(V + 1)|}{(V + A)[A(1 + c(V + 1))^2 + V(V + A + 1)]}. \tag{2.3}$$

Despite the fact that we are not using a Haar prior, we can still get perfect GIS algorithms by setting $c = -(V + 1)^{-1}$ or $c = -V/A - 1$. Figure 2 shows a plot of the convergence rate of the GIS algorithm as c ranges between -5 and 5 when $A = 3$ and $V = 2$. Note that the ASIS algorithm, which corresponds to the value $c = -1$, is not perfect. In fact, its convergence rate is 0.1. Perhaps this example could be used to settle the open problem described in Y&M's Section 6 concerning the optimality of AA–SA pairs.

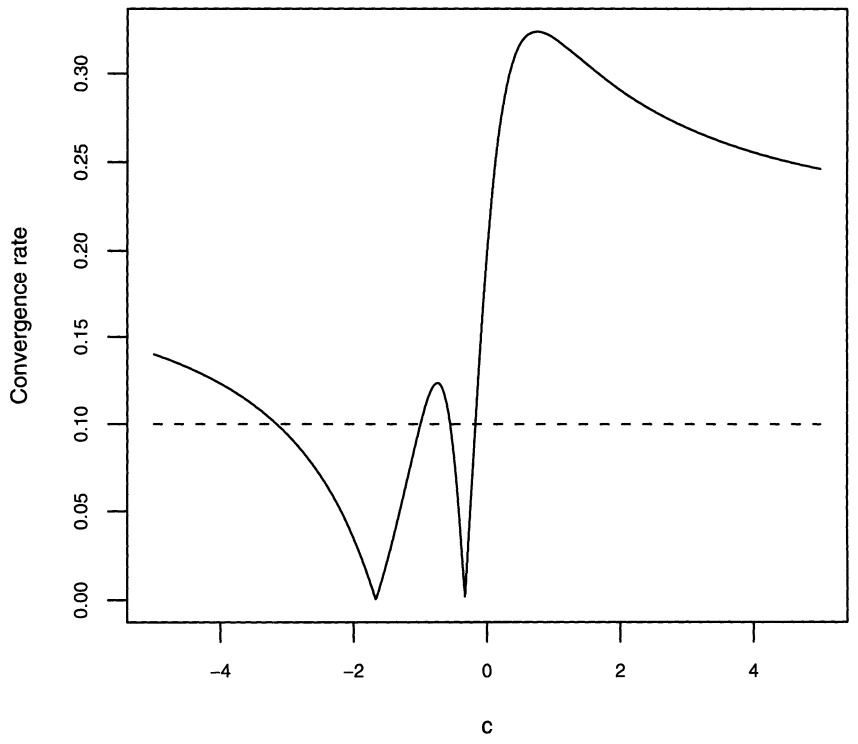

Figure 2. The convergence rate of the GIS algorithm versus c. The dashed horizontal line at 0.1 is the convergence rate of the ASIS algorithm.

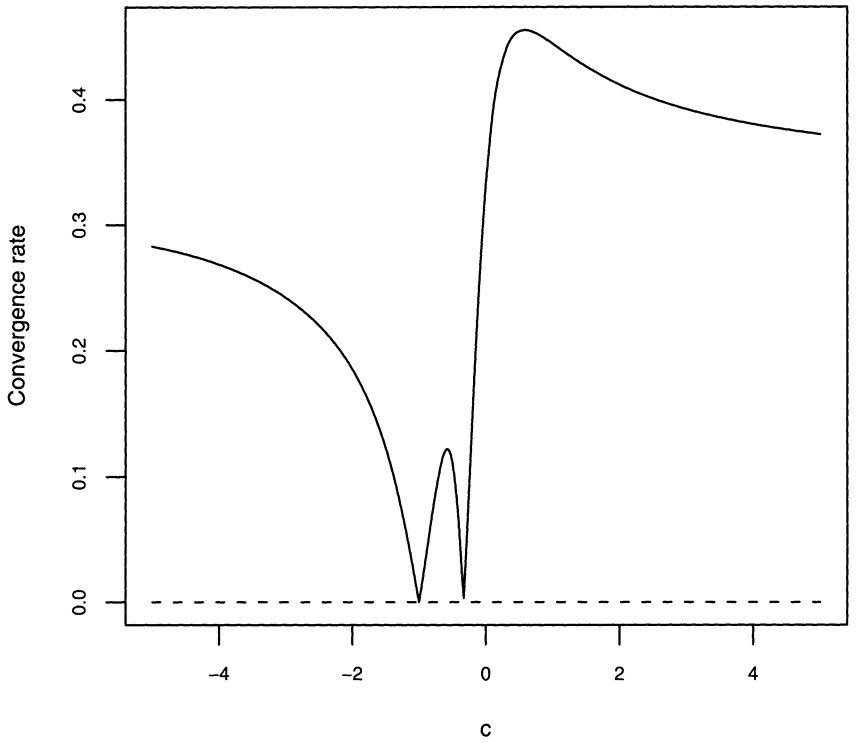

Figure 3. The convergence rate of the GIS algorithm versus c for the model with a flat (Haar) prior. The dashed horizontal line is at zero, which is the convergence rate of both the ASIS algorithm and the GIS algorithm with $c = -(V+1)^{-1}$.

If we replace the proper normal prior on X in the hierarchical model (2.2) with a flat (Haar) prior, then the model becomes exactly the one studied in Section 2.1 of Y&M. In that case, the convergence rate becomes

$$\|K_{\mathrm{YM}}\| = \frac{|1+c||1+c(V+1)|}{(1+c(V+1))^2+V}. \tag{2.4}$$

(Not surprisingly, (2.4) is simply the limit of (2.3) as $A \to \infty$.) As Y&M noted, the ASIS algorithm ($c = -1$) is perfect in this case. Indeed, their Theorem 4 is applicable here because the Haar assumptions are satisfied. Note, however, that one of the GIS algorithms ($c = -(V+1)^{-1}$) is also perfect. This shows that, even when a Haar prior is used (and the conditions of Y&M's Theorem 4 are satisfied), there may still exist a GIS algorithm (that is not ASIS) that has the same convergence rate as the optimal ASIS algorithm. Figure 3 shows a plot of the convergence rate of the GIS algorithm as c ranges between -5 and 5 when $A = 3$ and $V = 2$.

3. A NON-REVERSIBLE GIS ALGORITHM

Y&M did not provide an example of a non-reversible GIS chain, so we present one here. It was shown in Section 1 of this discussion that the alternating scheme defined by Y&M's

Equation (2.12) is a GIS algorithm. While this seems like an obvious place to look for an example of a non-reversible GIS algorithm, we decided to go in a different direction. Let $\mathsf{X} = \{1, 2, 3\}$ and take the target mass function to be $f_X = (0.4 \;\; 0.3 \;\; 0.3)^T$. Take $\mathsf{Y} = \mathsf{X}$ and consider the following joint mass function:

		Y		
		1	2	3
X	1	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
	2	$\frac{2}{10}$	$\frac{1}{10}$	0
	3	$\frac{2}{10}$	0	$\frac{1}{10}$

The x-marginal is clearly f_X, and the y-marginal is $f_Y = (0.6 \;\; 0.2 \;\; 0.2)^T$. The Markov transition matrix (Mtm) of the corresponding DA chain is

$$K = \begin{pmatrix} \frac{10}{24} & \frac{7}{24} & \frac{7}{24} \\ \frac{7}{18} & \frac{7}{18} & \frac{4}{18} \\ \frac{7}{18} & \frac{4}{18} & \frac{7}{18} \end{pmatrix}.$$

The (i, j)th entry is the probability that the DA Markov chain moves from i to j in one step. Note that K is reversible with respect to f_X. Now define a second joint mass function:

		Y		
		1	2	3
X	1	$\frac{3}{10}$	$\frac{1}{10}$	0
	2	0	$\frac{1}{10}$	$\frac{2}{10}$
	3	$\frac{2}{10}$	$\frac{1}{10}$	0

The x-marginal is again f_X, but the y-marginal in this case is $\tilde{f}_Y = (0.5 \;\; 0.3 \;\; 0.2)^T$. The Mtm of the new DA algorithm is

$$\tilde{K} = \begin{pmatrix} \frac{32}{60} & \frac{5}{60} & \frac{23}{60} \\ \frac{1}{9} & \frac{7}{9} & \frac{1}{9} \\ \frac{23}{45} & \frac{5}{45} & \frac{17}{45} \end{pmatrix}.$$

Again, $\tilde{K}$ is reversible with respect to f_X.

Consider a Mtm on Y given by

$$Q = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix},$$

and note that $f_Y^T Q = \tilde{f}_Y^T$. Thus, (the discrete analogue of) (1.1) holds, and we can use Q to construct a GIS algorithm. A simple calculation reveals that the Mtm of this GIS algorithm is

$$K_{\mathrm{YM}} = \begin{pmatrix} \frac{49}{120} & \frac{35}{120} & \frac{36}{120} \\ \frac{31}{90} & \frac{35}{90} & \frac{24}{90} \\ \frac{4}{9} & \frac{2}{9} & \frac{3}{9} \end{pmatrix}.$$

Clearly, K_{YM} is not reversible with respect to f_X, since, for example, $\frac{35}{120} \times \frac{4}{10} \neq \frac{31}{90} \times \frac{3}{10}$. However, it is true that $f_X^T K_{\mathrm{YM}} = f_X^T$, so f_X is indeed invariant for the GIS chain.

The eigenvalues of K_{YM} are $\{1, 0.1230, 0.0075\}$, so its spectral radius is 0.123. The eigenvalues of K and $\tilde{K}$ are $\{1, 0.1667, 0.0278\}$ and $\{1, 0.6835, 0.0054\}$, respectively. Thus, in this case, the GIS algorithm converges faster than either of the two underlying DA algorithms. Again, it is interesting to compare the exact answer, 0.123, with the upper bound from Y&M's Theorem 1. A straightforward, but somewhat tedious calculation shows that $\mathcal{R}_{12} = 0.5477$. Thus, Y&M's Theorem 1 says that the spectral radius of K_{YM} is bounded above by $0.5477\sqrt{0.1667 \times 0.6835} = 0.1849$.

ACKNOWLEDGMENTS

The authors thank Hani Doss and Vivekananda Roy for helpful discussions. The first author's work was supported by NSF grant DMS-08-05860.

ADDITIONAL REFERENCES

Hobert, J. P., and Rosenthal, J. S. (2007), "Norm Comparisons for Data Augmentation," *Advances and Applications in Statistics*, 7, 291–302. [574]

Khare, K., and Hobert, J. P. (2011), "A Spectral Analytic Comparison of Trace-Class Data Augmentation Algorithms and Their Sandwich Variants," *The Annals of Statistics*, to appear. [574]

Retherford, J. R. (1993), *Hilbert Space: Compact Operators and the Trace Theorem*, Cambridge: Cambridge University Press. [573]

Roberts, G. O., and Rosenthal, J. S. (1997), "Geometric Ergodicity and Hybrid Markov Chains," *Electronic Communications in Probability*, 2, 13–25. [573]

Rosenthal, J. S. (2003), "Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms," *Journal of the American Statistical Association*, 98, 169–177. [573]

Yu, Y., and Meng, X.-L. (2011), "To Center or Not to Center: That Is Not the Question—An Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency" (with discussion), *Journal of Computational and Graphical Statistics*, 20, 531–570 (this issue). [571]


DISCUSSION ARTICLE

Data Augmentation, Internal Representation, and Unsupervised Learning

Ying Nian WU

I am grateful to the editor for inviting me to contribute this discussion. I have learned a great deal from this exceedingly clever article by Yu and Meng. Ever since the groundbreaking work of Meng and van Dyk (1997), there have been many interesting developments in the art of data augmentation for both EM and MCMC. This article is yet another significant contribution to this line of research. While I feel I can contribute little to the discussion of the proposed method, I would like to mention a different perspective of data augmentation, in the hope of broadening the scope of the discussion. Data augmentation is not only a useful tool for MCMC, but it is also an essential ingredient in the so-called unsupervised learning, which involves augmenting latent variables or hidden units to explain the observed or visible data. In the context of neural science, the observed data are collected by the sensors in the form of images or sounds, and the latent variables or hidden units form the internal representations of the sensory data. The learning of such internal representations often does not require class labels or detailed annotations of the training examples, thus the learning is said to be unsupervised.

Latent variable models abound in statistical literature, such as factor analysis, mixture model, t-model, random effects model, probit regression, hidden Markov model, just to name a few. In what follows, I shall briefly review two popular latent variable models in neural science and unsupervised learning, as well as their hierarchical extensions.

The first model is the sparse coding model of Olshausen and Field (1996). Let $Y = (y_1, \ldots, y_M)$ be the M-dimensional vector, such as an image (where M is the number of pixels). Let $Z = (z_1, \ldots, z_K)$ be the K-dimensional vector of hidden units for representing Y. The model is of the following form:

$$z_k \sim p(z) \quad \text{independently}, \tag{0.1}$$

$$Y = \sum_{k=1}^{K} z_k B_k + \epsilon, \tag{0.2}$$

Ying Nian Wu is Professor, Department of Statistics, UCLA, Los Angeles, CA 90095-1554 (E-mail: *ywu@stat.ucla.edu*).

Journal of Computational and Graphical Statistics, Volume 20, Number 3, Pages 581–583
DOI: 10.1198/jcgs.2011.203b

where B_k's are unknown M-dimensional basis vectors, and ϵ is the residual. The model appears to be very similar to factor analysis, except that K is often assumed to be greater than M, so that the representation is said to be "overcomplete." Moreover, $p(z)$ is assumed to be a heavy tailed distribution, such as Laplacian distribution, t-distribution, or a mixture of a point mass at 0 and a normal distribution with a large variance. Such $p(z)$ captures the sparsity of Z in the sense that most of the K components of Z are small or 0. ϵ is often assumed to be white noise although this is quite unrealistic. The goal is to learn the dictionary of the basis elements $\mathbf{B} = (B_k, k = 1, \ldots, K)$ from training data $\{Y_i, i = 1, \ldots, n\}$, such as n image patches randomly cropped from some images of natural scenes.

The second model is the restricted Boltzmann machine (Hinton, Osindero, and Teh 2006):

$$p(Y, Z|W) \propto \exp\left\{\sum_{k,m} w_{km} z_k y_m\right\}, \tag{0.3}$$

where both y_m and z_k are assumed to be binary, and $W = (w_{km}, k = 1, \ldots, K, m = 1, \ldots, M)$ are the unknown parameters or the connection weights between hidden units z_k and the visible units y_m. This model looks rather unusual to statisticians, in the sense that it is not in the form of $p(Z|W)$ and $p(Y|Z, W)$. In fact, the prior distribution $p(Z|W)$ is implicit, and only the joint distribution $p(Y, Z|W)$ is specified. However, this model has the advantage that both $p(Y|Z, W)$ and $p(Z|Y, W)$ are simple. Given Z and W, y_m are independent, and given Y and W, z_k are independent. The model can be extended to the situation where y_k are continuous, so that $p(Y|Z, W)$ is in a comparable form as in Equation (0.2).

Both the sparse coding model (0.1) and (0.2) and the restricted Boltzmann machine (0.3) can be extended by introducing a higher layer of hidden variables on top of the layer of Z. The extension of (0.3) leads to the so-called deep belief network (Hinton, Osindero, and Teh 2006). The key observation in this endeavor is that the undirected graphical model $p(Y, Z|W)$ is equivalent to an infinite layer directed graphical model where each layer is a step of Gibbs sampler with $p(Y, Z|W)$ being the target distribution. The extension of (0.1) and (0.2) is quite different because of the sparsity of Z. Recently we proposed an active basis model (Wu et al. 2010), where we assumed that on top of Z is a layer of templates or partial templates, each being a composition of a small number of B_k's selected from the dictionary $\mathbf{B} = (B_k, k = 1, \ldots, K)$. Each selected B_k can be considered a "stroke" for sketching the template. See Figure 1 for three templates learned from natural images,

Figure 1. Templates formed by different sets of selected B_k's, where each B_k is depicted by a small line segment. Numbers of the selected B_k's are, respectively, 80, 30, 50.

where each B_k is illustrated by a small line segment, and the compositions of different sets of selected B_k's form different templates.

While statisticians may not be at home with discriminative supervised learning such as max-margin classification, the hierarchical latent variable models and the associated likelihood or Bayesian learning should be very familiar and natural to statisticians, who are well equipped to make useful contributions.

ACKNOWLEDGMENT

I acknowledge the support of NSF DMS 1007889.

ADDITIONAL REFERENCES

Hinton, G. E., Osindero, S., and Teh, Y. (2006), "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, 18, 1527–1554. [582]

Olshausen, B. A., and Field, D. J. (1996), "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, 381, 607–609. [581]

Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2010), "Learning Active Basis Model for Object Detection and Recognition," *International Journal of Computer Vision*, 90, 198–235. [582]

DISCUSSION ARTICLE

Comment

Brandon C. KELLY

1. INTRODUCTION

I would like to congratulate Professors Yu and Meng on a valuable contribution to the MCMC toolkit. The ASIS framework is conceptually simple and straightforward to implement, while still providing a significant boost in efficiency, at least for some problems. These qualities of ASIS are attractive to me as an astrophysicist, because I spend most of my time doing astrophysics research, and not designing MCMC samplers.

The astrophysical application I discuss involves joint fitting of multiple astronomical sources that are subject to both instrumental noise and a common calibration error. The particular object I applied an ASIS to is a cloud of astronomical dust located in a region that is forming stars within our galaxy. Astronomical dust is thought to be important for a number of reasons, including playing a role in the formation of stars (e.g., Hirashita and Ferrara 2002) and planets (Watson et al. 2007). Moreover, dust acts as a screen, attenuating the light from astronomical objects behind it, making it difficult to study these objects. Unfortunately, there is much that we do not understand about astronomical dust. Research into the properties of astronomical dust is currently of significant interest, both because it will improve our understanding of the physics and formation of dust, and because it will better enable us to correct for its attenuating effects.

2. ACCOUNTING FOR NOISE AND CALIBRATION ERROR: THE STATISTICAL MODEL

Suppose we have a sample of $i = 1, \ldots, n$ data points, each of which has flux measurements at $j = 1, \ldots, J$ observational frequencies; the flux of an object is a measure of the amount of energy we detect from that object per time interval per area as a function of the frequency of light at which we observe the object. Denote the set of observational

Brandon C. Kelly is Hubble Fellow, Harvard-Smithsonian Center for Astrophysics, 60 Garden St, Cambridge, MA 02138 (E-mail: *bckelly@cfa.harvard.edu*).

Journal of Computational and Graphical Statistics, Volume 20, Number 3, Pages 584–591
DOI: 10.1198/jcgs.2011.203c

frequencies as ν; the values of $\nu_1, \ldots, \nu_J$ are known and fixed by the instrument. Values of n can be as small as $n \sim 100$ and as large as $n \sim 10^5$, while typical values of J range from $J = 4$ to $J = 15$. The scientific goal is to use the measured fluxes to estimate a set of physical parameters (e.g., mass, temperature, etc.) for the set of data points, as well as their joint distribution. Denote the set of physical parameters as $\theta_1, \ldots, \theta_n$, where each θ_i is a vector containing the values of the p physical parameters for the ith data point. We can relate the parameters θ_i to the J-element vector of measured flux values, $\mathbf{y}_i$, as

$$\mathbf{y}_i = X(\nu)\gamma_i + \mathbf{f}(\nu, \phi_i) + \epsilon_i + \delta, \qquad \theta_i = (\gamma_i, \phi_i). \tag{2.1}$$

Here, I have divided the parameters θ_i into those parameterizing a linear component to the flux model, and those parameterizing a nonlinear component. The linear component is $X(\nu)\gamma_i$, where $X(\nu)$ is a matrix whose elements are a function of the observational frequencies, while the nonlinear component is $\mathbf{f}(\nu, \phi_i)$. The quantity ϵ_i is a zero-mean normally distributed J-element vector representing the contribution of measurement noise to the observations. Because the noise is independent at different observational frequencies, the covariance matrix of ϵ_i is diagonal. The last term, δ, is a zero-mean normally distributed J-element vector representing the calibration uncertainties, that is, our uncertainty in converting the detector output to a physical flux measurement. This quantity is the same for every data point that is output by a detector, and therefore is the same for all i. The term δ may be considered to be an unknown bias at each observational frequency that affects each data point in the same manner. As is typical in astrophysics, the covariance matrices of ϵ_i and δ, denoted as V_i and V_δ, respectively, are assumed known.

I model the distribution of θ to be a normal distribution with mean μ and covariance Σ, leading to the following hierarchical model:

$$\mathbf{y}_i | \theta_i, \delta \overset{\text{i.i.d.}}{\sim} N(X(\nu)\gamma_i + \mathbf{f}(\nu, \phi_i) + \delta, V_i), \tag{2.2}$$

$$\delta \sim N(0, V_\delta), \tag{2.3}$$

$$\theta_i | \mu, \Sigma \overset{\text{i.i.d.}}{\sim} N(\mu, \Sigma), \tag{2.4}$$

$$\mu, \Sigma \sim p(\mu, \Sigma). \tag{2.5}$$

Here, $p(\mu, \Sigma)$ is the prior distribution on μ and Σ, which I set to be uniform subject to $|\Sigma| > 0$. The calibration errors are considered nuisance parameters and are of no scientific interest. I could calculate the posterior distribution for this problem by analytically integrating over the unknown δ. Unfortunately, doing so implies that I have to invert an $n \times n$ non-diagonal covariance matrix for each evaluation of the likelihood, which is prohibitively slow for typical values of n, for example, $n > 1000$. Instead, I model δ as missing data and update it in an MCMC sampler. Under this formulation, Equations (2.2)–(2.5) form an *ancillary augmentation*, in the language of Professors Yu and Meng, as δ and θ are independent a priori.

2.1 MCMC Sampler Under the Ancillary Augmentation

The ancillary augmentation suggests the following MCMC sampler, which employs a Metropolis-within-Gibbs strategy:

Step 1. Update the calibration errors by drawing δ from $p(\delta|\mathbf{y}_1, \ldots, \mathbf{y}_n, \theta)$:

$$\delta|\mathbf{y}_1, \ldots, \mathbf{y}_n, \theta \sim N(\hat{\delta}, (V_\delta^{-1} + S^{-1})^{-1}), \tag{2.6}$$

$$\hat{\delta} = (V_\delta^{-1} + S^{-1})^{-1} \sum_{i=1}^{n} V_i^{-1}(\mathbf{y}_i - X(\nu)\gamma_i - \mathbf{f}(\nu, \phi_i)), \tag{2.7}$$

$$S^{-1} = \left(\sum_{i=1}^{n} V_i^{-1}\right)^{-1}. \tag{2.8}$$

Step 2_A. Update $\gamma_1, \ldots, \gamma_n$. This can be done independently for each γ_i by noting that $p(\gamma_i|\mathbf{y}_i, \delta, \phi_i, \mu, \Sigma)$ has the form of a Bayesian linear regression of $\mathbf{y}_i - \mathbf{f}(\nu, \phi_i) - \delta$ onto $X(\nu)$ with normal prior distribution:

$$\gamma_i|\mathbf{y}_i, \phi_i, \delta, \mu, \Sigma \overset{\text{i.i.d.}}{\sim} N(\hat{\gamma}_i, V_{\hat{\gamma}_i}), \tag{2.9}$$

$$\hat{\gamma}_i = V_{\hat{\gamma}_i}^{-1}\left[X^T(\nu)V_i^{-1}(\mathbf{y}_i - \mathbf{f}(\nu, \phi_i) - \delta) + \Sigma^{-1}(\phi_i)\mu(\phi_i)\right], \tag{2.10}$$

$$V_{\hat{\gamma}_i} = (X^T(\nu)V_i^{-1}X(\nu) + \Sigma^{-1}(\phi_i))^{-1}. \tag{2.11}$$

Here, $\mu(\phi_i)$ and $\Sigma(\phi_i)$ are the conditional prior mean and covariance of $\gamma_i|\phi_i$, respectively.

Step 3_A. Update $\phi_i|\mathbf{y}_i, \gamma_i, \delta, \mu, \Sigma$ for $i = 1, \ldots, n$. For most cases there is no closed form for drawing directly from the conditional density, so I employ a Metropolis–Hastings move.

Step 4. Update $\mu, \Sigma|\theta_1, \ldots, \theta_n$. This step is just the usual update for the mean and covariance of a normal density.

Unfortunately, the above algorithm is very inefficient for the typical case of $n \gg J$. The reason for this is because for large n, δ is very precisely determined conditional on θ, so Step 1 produces a value of δ that is very close to that expected from the current values of θ. But Steps 2_A and 3_A produce values of θ which are on average close to the current value of δ. This is especially true when there are at least some data points with very high signal-to-noise, that is, when the diagonal elements of V_i^{-1} are very large for some i. For these data points, θ_i is updated very close to the value determined by the current value of δ. The end result is that the above sampler moves very slowly through the parameter space, and convergence to the target density is extremely slow. The performance of the above algorithm becomes worse when the sample size increases, or when the measurement noise is decreased ($V_i \to 0$). The slow convergence is particularly a concern when the calibration uncertainties are large, as the regions containing significant posterior probability for θ become larger, and the above sampler takes even longer to explore them.

2.2 Implementing an ASIS Step Into the MCMC Sampler

I have had success using ASIS on the above problem. I reparameterized the calibration uncertainties to find a *sufficient augmentation* for γ:

$$\tilde{\delta}_i = \delta + X(\nu)\gamma_i. \tag{2.12}$$

Under the SA parameterization, each source has its own set of "calibration uncertainties." Ideally, one would like to find an SA for $\theta = (\gamma, \phi)$, but the nonlinearity of $\mathbf{f}(\nu, \phi)$ makes it difficult to implement this in practice. Under this reparameterization, the hierarchical model defined by Equations (2.2)–(2.5) becomes

$$\mathbf{y}_i|\phi_i, \tilde{\delta}_i \overset{\text{i.i.d.}}{\sim} N(\mathbf{f}(\nu, \phi_i) + \tilde{\delta}_i, V_i), \tag{2.13}$$

$$\tilde{\delta}_i|\tilde{\delta}_k, \gamma_k, \gamma_i = \tilde{\delta}_k - X(\nu)(\gamma_k - \gamma_i), \qquad i \neq k, \tag{2.14}$$

$$\tilde{\delta}_k|\gamma_k \sim N(X(\nu)\gamma_k, V_\delta), \tag{2.15}$$

$$\theta_i|\mu, \Sigma \overset{\text{i.i.d.}}{\sim} N(\mu, \Sigma), \tag{2.16}$$

$$\mu, \Sigma \sim p(\mu, \Sigma). \tag{2.17}$$

Equation (2.14) arises because it must be true that $\tilde{\delta}_k - X(\nu)\gamma_k = \delta = \tilde{\delta}_i - X(\nu)\gamma_i$ for all i and k. When implementing my MCMC sampler, I set $k = 1$ for Equations (2.14) and (2.15).

The SA defined by Equations (2.12)–(2.17) results in an ASIS MCMC sampler which inserts the following step between Steps 2_A and 3_A:

Step 2_S. Update δ and $\gamma_1, \ldots, \gamma_n$ under the SA. This is done by calculating $\tilde{\delta}_i$ according to Equation (2.12), and choosing a value of k, say $k = 1$. Then update the value of $\gamma_k|\tilde{\delta}_k$ by drawing from $p(\gamma_k|\tilde{\delta}_k)$. This is just a Bayesian linear regression of $\tilde{\delta}_k$ on $X(\nu)$:

$$\gamma_k|\tilde{\delta}_k \sim N\big(\hat{\Gamma}_k, V_{\hat{\Gamma}_k}\big), \tag{2.18}$$

$$\hat{\Gamma}_k = V_{\hat{\Gamma}_k} X^T(\nu) V_\delta^{-1} \tilde{\delta}_k, \tag{2.19}$$

$$V_{\hat{\Gamma}_k} = [X^T(\nu) V_\delta^{-1} X(\nu)]^{-1}. \tag{2.20}$$

Given the updated value of γ_k, for $i \neq k$ set

$$\gamma_i = \gamma_k + [X^T(\nu)X(\nu)]^{-1} X^T(\nu)(\tilde{\delta}_i - \tilde{\delta}_k). \tag{2.21}$$

Next, invert Equation (2.12) using the updated value of γ_k to update the calibration errors:

$$\delta = \tilde{\delta}_k - X(\nu)\gamma_k. \tag{2.22}$$

Finally, because we updated $\gamma_k|\tilde{\delta}_k$ after marginalizing over μ, assuming a uniform prior on μ, we need to update the components of μ corresponding to γ. Denote these components as μ_γ. Update μ_γ as

$$\mu_\gamma|\phi, \mu_\phi, \gamma \sim N(\hat{\mu}_\gamma, n^{-1} \operatorname{var}(\gamma|\phi)), \tag{2.23}$$

$$\hat{\mu}_\gamma = \bar{\gamma} - \Sigma_{\gamma\phi}\Sigma_\phi^{-1}(\bar{\phi} - \mu_\phi), \tag{2.24}$$

where $\bar{\gamma}$ and $\bar{\phi}$ are the sample mean of γ and ϕ, respectively, $\operatorname{var}(\gamma|\phi)$ is the conditional variance of γ given ϕ, and $\Sigma_{\gamma\phi}$, Σ_ϕ, and μ_ϕ denote the appropriate components of Σ and μ.

Step 2_S results in a significant improvement in efficiency because it breaks the dependence between δ and γ. By reparameterizing δ, we are able to propose values of δ and γ

which leave the quantity $\tilde{\delta}$ unchanged. This improves MCMC efficiency because the primary culprit for the slow convergence under the AA is Equation (2.2). Under Step 2_S, we obtain values of δ and γ sampled from $p(\delta)$ which leave Equation (2.2) unchanged, that is, along the directions defined by $\delta(\Delta\gamma) = -X(\nu)(\gamma + \Delta\gamma)$.

3. APPLICATION OF AN ASIS TO A SOURCE OF ASTRONOMICAL DUST

I applied the above ASIS to observations from the *Herschel* Space Observatory, presented by Stutz et al. (2010). The data are a set of $n = 5503$ flux measurements at different spatial locations for a source of astronomical dust observed at $J = 5$ frequencies. A common model for the spectra of dust is that of a blackbody modified by a power-law (Hildebrand 1983):

$$S_{ij} = C_i \left(\frac{\nu_j}{\nu_0}\right)^{\beta_i} B_{\nu_j}(T_i). \tag{3.1}$$

Here, S_{ij} is the measured flux at ν_j for the ith location, ν_0 is a fixed reference frequency, C_i is a scale factor that is related to the density of dust at the ith location, β_i is a parameter describing the dust opacity, and $B_{\nu_j}(T_i)$ is the Planck function evaluated at the frequency ν_j as a function of temperature, T_i. After taking a logarithmic transformation, Equation (3.1) is equivalent to Equation (2.1) for $y_{ij} = \log S_{ij}$, $\gamma_i = (\log C_i, \beta_i)$, $\phi_i = \log T_i$, $f_j(\nu, \phi_i) = \log B_{\nu_j}(T_i)$, and $X(\nu)$ is a $J \times 2$ matrix with the jth row $\mathbf{x}_j^T(\nu) = [1, \log(\nu_j/\nu_0)]$. The scientific goal is to estimate the values of C_i, β_i, and T_i, as well as their distribution; of particular interest are the conditional distributions, as these will provide insight into how environment and other physical quantities affect the dust properties. While it would be best to also include a component to the statistical model which incorporates the spatial correlations in θ, for simplicity I did not do this (although presumably an ASIS could be developed for this as well).

I applied the above MCMC sampler, both with and without Step 2_S, to the dataset from Stutz et al. (2010). I ran the chains for 2×10^5 iterations, after performing 2×10^4 iterations of burn-in. Figures 1 and 2 show the results for the chains corresponding to the values of β for two different data points. One data point is typical of the dataset, with the dominant component of uncertainty being the noise, that is, $V_i > V_\delta$ along the diagonal. The other data point represents the data point with the most precise measurements, and thus the dominant component of uncertainty for this data point are the calibration errors, that is, $V_i \ll V_\delta$ along the diagonal.

For the data point with lower signal-to-noise, the ASIS represents a modest improvement over the sampler based purely on the AA. However, the improvement for the data point with very high signal-to-noise is more significant. The inclusion of Step 2_S enables much larger moves at fixed T_i, which help to more efficiently explore the parameter space corresponding to C_i, β_i, and δ at fixed T_i. Moreover, this improvement in efficiency results in faster convergence to the marginal distribution of β_i, as suggested by the much

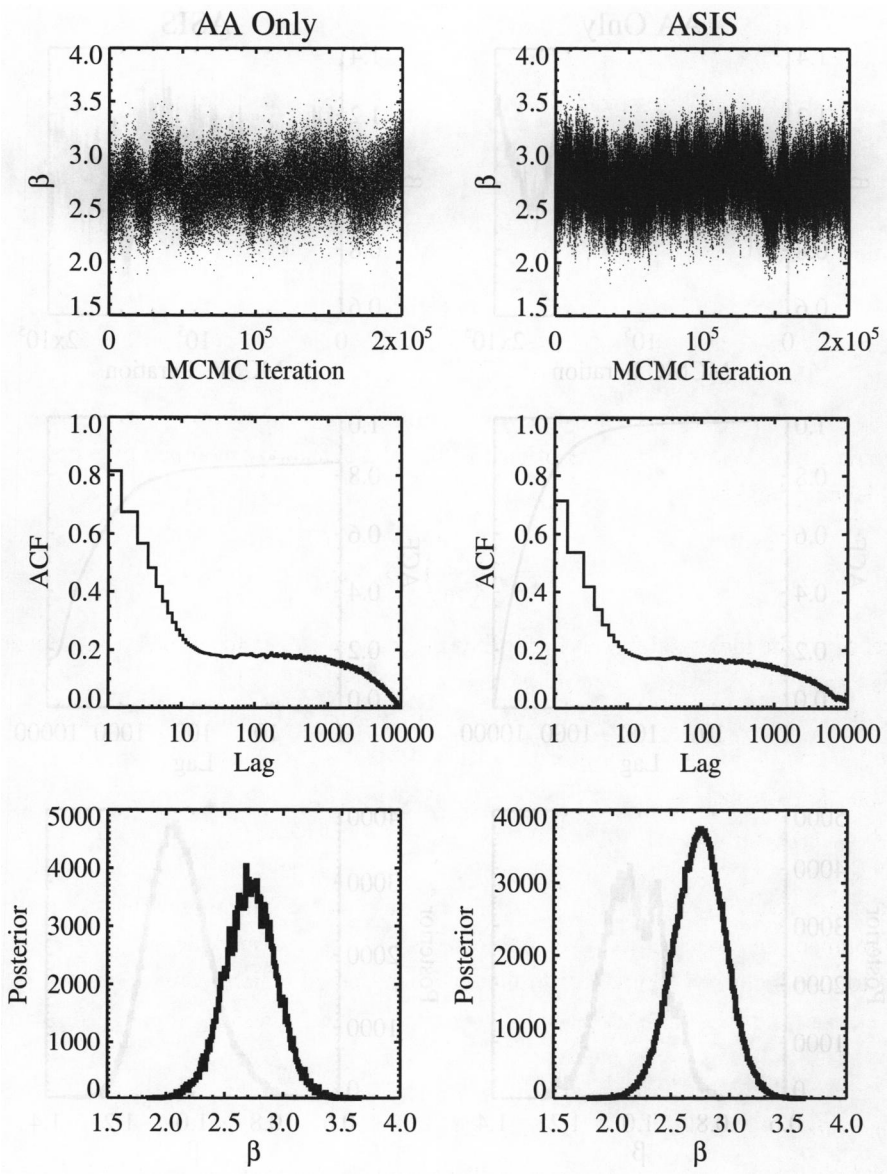

Figure 1. Results for the MCMC sampler under the original AA (left) and implementing an ASIS (right). Shown are the evolution of the chain for β for a data point with typical signal-to-noise (top), the autocorrelation function for this chain (middle), and the marginal posterior estimated from this chain (bottom). The inclusion of Step 2_S under the ASIS increases the variations in β at fixed temperature, but the lack of an ASIS for temperature enables long-time scale correlations to persist.

smoother estimated posterior when including Step 2_S. It is worth noting that for this problem the ASIS I implemented appears to reduce the autocorrelation function by a constant factor, although the shape remains unchanged.

The persistent dependency of the chain on long time scales is most likely the result of the strong correlations among δ, β_i, C_i, and T_i combined with the slow convergence of T_i,

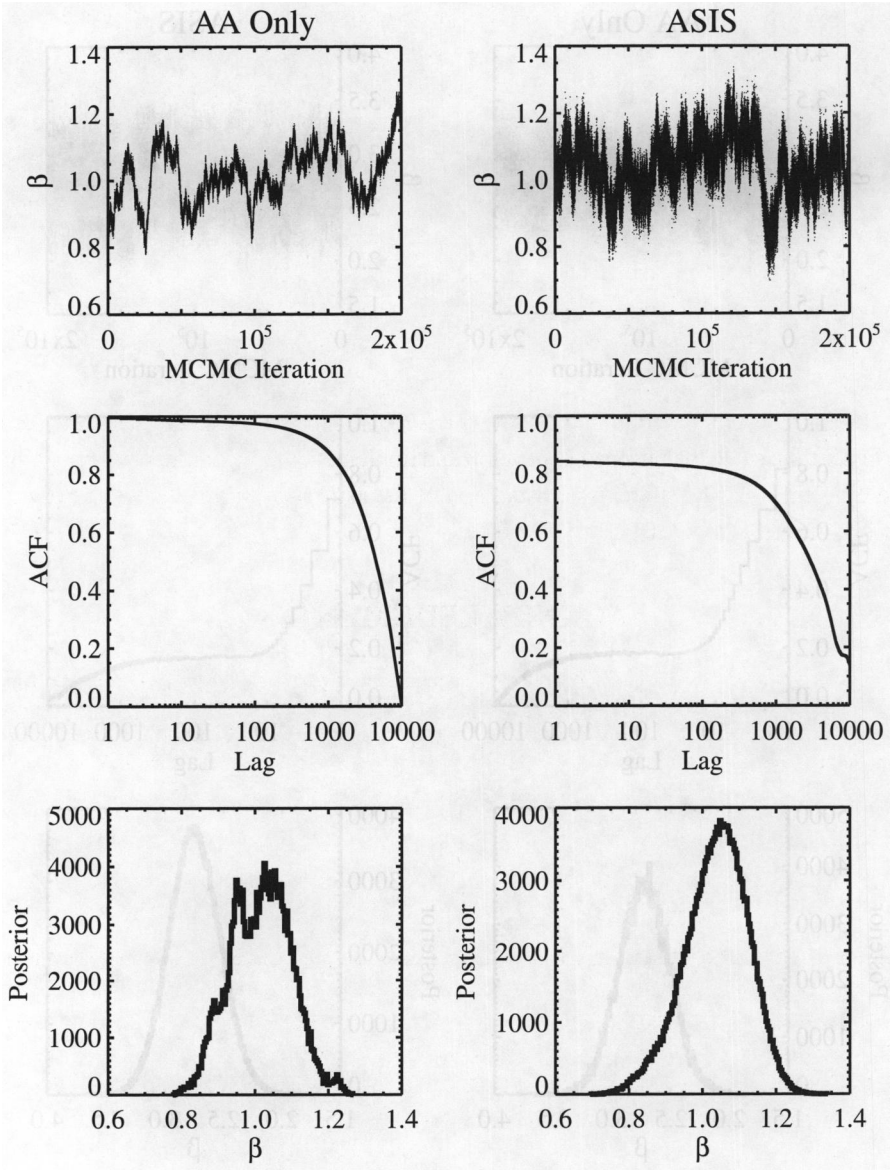

Figure 2. Same as Figure 1, but for the data point with the highest signal-to-noise. As with Figure 1, the inclusion of Step 2_S improves the efficiency of the sampler, but the improvement is more dramatic for this data point, for which the uncertainties are dominated by the calibration errors.

as Step 2_S has only partially mitigated the strong correlation between δ and θ. Ideally, one would like to have an ASIS for all of these parameters, instead of an ASIS for only δ, C_i, and β_i. In general, such an ASIS could include an additional step analogous to Step 2_S and set $\tilde{\delta}_i = \delta + \mathbf{f}(\nu, \phi_i)$. However, updating ϕ_k under this representation would prove difficult. Moreover, it is unclear whether it is always possible to update ϕ and δ under this SA such that the quantity $\delta = \tilde{\delta}_i - \mathbf{f}(\nu, \phi_i)$ is invariant for all i. If such a strategy is possible, it would likely go a long way toward further improving the efficiency of this ASIS sampler.

The implementation of ASIS I studied here leads to a noticeable improvement of MCMC efficiency, albeit not as dramatic as some of the applications studied by Professors Yu and Meng. That being said, the improvement is still significant, as inspection of the estimated posterior for β in Figure 2 suggests that the MCMC chain without ASIS has not even converged to the target distribution after 2×10^5 iterations. Further improvement to MCMC efficiency for this problem could likely be developed by also including various other MCMC techniques; however, the gain in efficiency provided by ASIS for this problem is sufficient for my scientific purposes. For one, ASIS provided a gain in efficiency with only a negligible increase in computational speed. Second, because the ASIS is simple and flexible, it was relatively straightforward for me to develop and insert Step 2_S into my MCMC sampler. Therefore, it was not necessary for me to devote significant numbers of man-hours to implementing the ASIS, which enables me to obtain the scientific results on a faster time scale than if I had to develop and implement more complicated, but perhaps more efficient, MCMC strategies. Discussion of the astrophysical results will be given in the work of Kelly et al. (2011).

ACKNOWLEDGMENTS

I thank the editor, Professor Richard A. Levine, for inviting me to participate in this discussion, and Dr. Amelia Stutz for providing me with her data. I acknowledge support from NASA through Hubble Fellowship grants HF-01220.01 and HF-51243.01 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS 5-26555.

ADDITIONAL REFERENCES

Hildebrand, R. H. (1983), "The Determination of Cloud Masses and Dust Characteristics From Submillimetre Thermal Emission," *Quarterly Journal of the Royal Astronomical Society*, 24, 267–282. [588]

Hirashita, H., and Ferrara, A. (2002), "Effects of Dust Grains on Early Galaxy Evolution," *Monthly Notices of the Royal Astronomical Society*, 337, 921–937. [584]

Kelly, B. C., Shetty, R., Stutz, A. M., Kauffmann, J., Goodman, A. A., and Launhardt, R. (2011), "Dust SEDs in the Era of Herschel and Planck: A Bayesian Fitting Technique," to appear. [591]

Stutz, A., Launhardt, R., Linz, H., Krause, O., Henning, T., Kainulainen, J., Nielbock, M., Steinacker, J., and André, P. (2010), "Dust-Temperature of an Isolated Star-Forming Cloud: Herschel Observations of the Bok Globule CB244," *Astronomy & Astrophysics*, 518, L87. [588]

Watson, A. M., Stapelfeldt, K. R., Wood, K., and Ménard, F. (2007), "Multiwavelength Imaging of Young Stellar Object Disks: Toward an Understanding of Disk Structure and Dust Evolution," in *Protostars and Planets V*, eds. B. Reipurth, D. Jewitt, and K. Keil, Tucson: University of Arizona Press, pp. 523–538. [584]


Whether 'tis Nobler in the Mind to Suffer the Slings and Arrows of Outrageous Mixing Problems, or to Take Arms Against a Sea of Troubles, and by Opposing End Them?

Author(s): O. Papaspiliopoulos, G. O. Roberts and G. Sermaidis

Source: *Journal of Computational and Graphical Statistics*, Vol. 20, No. 3 (September 2011), pp. 592-602
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: https://www.jstor.org/stable/23248841
Accessed: 11-08-2020 20:21 UTC

DISCUSSION ARTICLE

Whether 'tis Nobler in the Mind to Suffer the Slings and Arrows of Outrageous Mixing Problems, or to Take Arms Against a Sea of Troubles, and by Opposing End Them?

O. PAPASPILIOPOULOS, G. O. ROBERTS, and G. SERMAIDIS

1. INTRODUCTION

Parameterization of Markov chain Monte Carlo algorithms, particularly involving data augmentation, has been a topical and important area of computational statistics for over 20 years now. The article by Meng and Yu offers us an exciting new idea building on rather established ideas of centering and non-centering of algorithms.

This discussion will focus on providing some further insights into the apparently magical properties of the interweaved algorithm. We will also describe the construction of non-centered algorithms in latent Poisson process models, and describe our experience of using the interweaved approach in MCMC for discretely observed diffusion models using an approach with no discretization error.

2. BAYESIAN FRACTION OF MISSING INFORMATION AND NON-CENTERING

As the article hints in Section 2, the most powerful connection between statistics and the properties of MCMC algorithms comes through the Bayesian fraction of missing information (BFMI). Expanding on the authors' definition and just denoting by Y^*_{mis} an arbitrary

O. Papaspiliopoulos is Assistant Professor, Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, Barcelona 08005, Spain (E-mail: *omiros.papaspiliopoulos@upf.edu*). G. O. Roberts is Professor, Department of Statistics, Warwick University, Coventry CV4 7AL, U.K. G. Sermaidis is Research Associate, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster LA1 4YF, U.K.

Journal of Computational and Graphical Statistics, Volume 20, Number 3, Pages 592–602
DOI: 10.1198/jcgs.2011.203d

augmentation, we set the BFMI as

$$\kappa = \sup_{f \in L^2(\pi)} \frac{\text{var}(f(\theta)|Y_{obs}, Y^*_{mis})}{\text{var}(f(\theta)|Y_{obs})}. \tag{2.1}$$

Here we take π to be the marginal posterior density of θ given Y. The rate of convergence of the "pure" Gibbs sampler which augments using Y^*_{mis} is given by

$$\rho_* = 1 - \kappa. \tag{2.2}$$

The power of this result comes from its explicit and general characterization of the algorithm's convergence properties, but also from its statistical interpretability, which in turn can guide choice of augmentation; see, for instance, the work of Papaspiliopoulos and Roberts (2008) and Papaspiliopoulos, Roberts, and Sköld (2007) and many of the references therein.

Consider again the two-level normal hierarchical model of (2.1) and (2.2) in the article. Simple application of (2.2) shows the rate of convergence of the centered and non-centered Gibbs samplers, ρ_c and ρ_{nc}, respectively, as

$$\rho_c = \frac{1}{1+V}, \qquad \rho_{nc} = \frac{V}{1+V}.$$

Consider the following parameterization as introduced by Papaspiliopoulos, Roberts, and Sköld (2003). For the same model,

$$\begin{aligned} Y_{obs} &\sim N(w\theta + Y^w_{mis}, 1), \\ Y^w_{mis} &\sim N((1-w)\theta, V), \end{aligned} \tag{2.3}$$

where w is a fixed number in $[0, 1]$. This defines a continuum of *partially non-centered* parameterizations with $w = 0$ corresponding to the centered case, while $w = 1$ is the non-centered parameterization. Again applying (2.2), it can easily be seen that the rate of convergence for the Gibbs sampler on (θ, Y^w_{mis}) has rate of convergence ρ_w given by

$$\rho^w_{pnc} = \frac{(w - (1-\kappa))^2}{w^2\kappa + (1-w)^2(1-\kappa)}, \tag{2.4}$$

where κ was defined in (2.2). In Figure 1 we plot this against w for $V = 1/3$. From (2.4) we can easily derive that

$$\begin{aligned} &\rho^w_{pnc} \le \max(\rho_c, \rho_{nc}) \quad \forall w \in (0,1), \\ &\rho^w_{pnc} = 0 \quad \text{for } w = 1 - \kappa, \end{aligned} \tag{2.5}$$

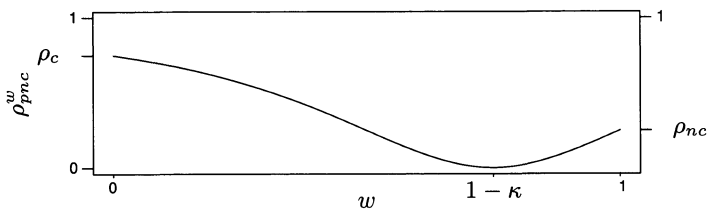

Figure 1. The partial non-centered algorithm convergence *smile*. For the Gaussian case the rate reaches 0 for the optimal sampler. However, knowing this optimal strategy a priori is usually not possible.

which suggests not only that the PNCP algorithm (2.3) can outperform both CP and NCP, but also it can be tuned appropriately to produce i.i.d. samples, by setting $w = 1 - \kappa$.

Partially non-centered parameterizations in considerably more complex situations can be used to speed up MCMC in many situations (see, e.g., Neal and Roberts 2005 for applications in infectious disease epidemiology).

However, although the partially non-centered method is in principle just as flexible and powerful as the interweaved approach, its implementation is complicated by the need to know (or estimate) the "optimal" value of w—something which can be nontrivial in more complex problems than the normal hierarchical model. In contrast, one great strength of the interweaved approach is that it is completely automatic.

3. MORE IS NOT NECESSARILY BETTER!

We can gain further insight into the power of the interweaved approach as follows. We shall consider exclusively the Gaussian linear model for this section.

One might be inclined to dismiss the apparent improvement of the interweaved algorithm over the centered and non-centered methods. After all, the interweaved algorithm involves more simulation:

1. Update $Y_{mis}|Y_{obs}, \theta$.
2. Update $\theta|Y_{mis}$.
3. Update $\theta|Y^*_{mis}, Y_{obs}$.

However, adding a further step, essentially concatenating the centered and non-centered updates produces:

1. Update $Y_{mis}|\theta, Y_{obs}$.
2. Update $\theta|Y_{mis}$.
3. Update $Y^*_{mis}|\theta, Y_{obs}$.
4. Update $\theta|Y^*_{mis}, Y_{obs}$.

This is readily shown to have convergence rate $V/(1+V)^2$. Therefore, carrying out more simulation actually slows down the convergence of the algorithm!

To understand this phenomenon more thoroughly, we shall adopt the geometric interpretation of the Gibbs sampler as introduced by Amit (1991). Consider the space of square integrable functions under π: $L^2(\pi)$. Let P_θ denote the transition kernel of the step of the Gibbs sampler which updates Y_{mis} (or Y^*_{mis}) for fixed θ. Correspondingly define $P_{Y_{mis}}$ and $P_{Y^*_{mis}}$, respectively, for the moves which fix Y_{mis} and Y^*_{mis}, respectively. The various samplers can be described by the action of these transition kernels on $L^2(\pi)$, for instance,

$$P_\theta f(\theta, y_{mis}) = \mathbf{E}(f(\theta, Y_{mis})),$$

where the expectation here is taken with respect to the conditional $Y_{mis}|\theta$. Note that P_θ maps general functions in $L^2(\pi)$ to the subspace of functions which can be written as

functions of θ alone. In fact P_θ (and of course $P_{Y_{mis}}$ and $P_{Y^*_{mis}}$) can be viewed as orthogonal projections on $L^2(\pi)$ under the inner product

$$\langle f, g \rangle = \mathbf{E}_\pi(f(\theta, Y_{mis}), g(\theta, Y_{mis})).$$

P_θ, $P_{Y_{mis}}$, and $P_{Y^*_{mis}}$ are projections onto V_θ, $V_{Y_{mis}}$, and $V_{Y^*_{mis}}$, subsets of $L(\pi)$ which can be expressed respectively as functions of θ, Y_{mis}, and Y^*_{mis}.

In this framework, the interweaved sampler can be written

$$P_{IW} = P_\theta P_{Y_{mis}} P_{Y^*_{mis}}$$

while the concatenated sampler is just

$$P_{con} = P_\theta P_{Y_{mis}} P_\theta P_{Y^*_{mis}}.$$

Since $V_{Y_{mis}}$ and $V_{Y^*_{mis}}$ are orthogonal subspaces of $L^2(\pi)$, $P_{Y_{mis}} P_{Y^*_{mis}}$ just maps functions in $L^2(\pi)$ to constants (the mean of the function under π). Therefore the interweaved approach would yield i.i.d. draws from θ even if the θ update step 1 is omitted!

Conversely, the concatenated approach does not apply $P_{Y_{mis}}$ and $P_{Y^*_{mis}}$ consecutively and so instant convergence fails. This is illustrated in Figure 2.

4. GEOMETRIC ERGODICITY ROBUSTNESS

Turning from quantitative rate of convergence results which are generally only available in the Gaussian case, positive results can be obtained concerning the robustness of geometric and uniform ergodicity of the interweaved approach to the tail behavior of the target distribution.

We recall that a Markov chain is geometrically ergodic if we can find a constant $\rho < 1$ and a finite function $V(x)$ such that for all starting points x

$$\| P^n(x, \cdot) - \pi \| \leq V(x)\rho^n$$

(where $\| \cdot \|$ denotes total variation distance) and that this convergence is deemed to be *uniform* if V can be taken to be bounded. While not guaranteeing rapid convergence, geometric ergodicity is a very desirable robustness property, particularly as it guarantees the existence of CLTs when the chain is reversible and with minimal extra conditions for nonreversible chains (see Roberts and Rosenthal 2008).

As an example, and following Papaspiliopoulos and Roberts (2008), we consider the following model:

$$\begin{aligned} Y_{obs} &\sim D_{obs}(Y_{mis}), \\ Y_{mis} &\sim D_{hidden}(\theta), \end{aligned} \tag{4.1}$$

where D_{obs} and D_{hidden} denote symmetric distributions centered around their respective arguments. Here we shall focus on the cases where these distributional forms can be either Gaussian (denoted N) or Cauchy (denoted C).

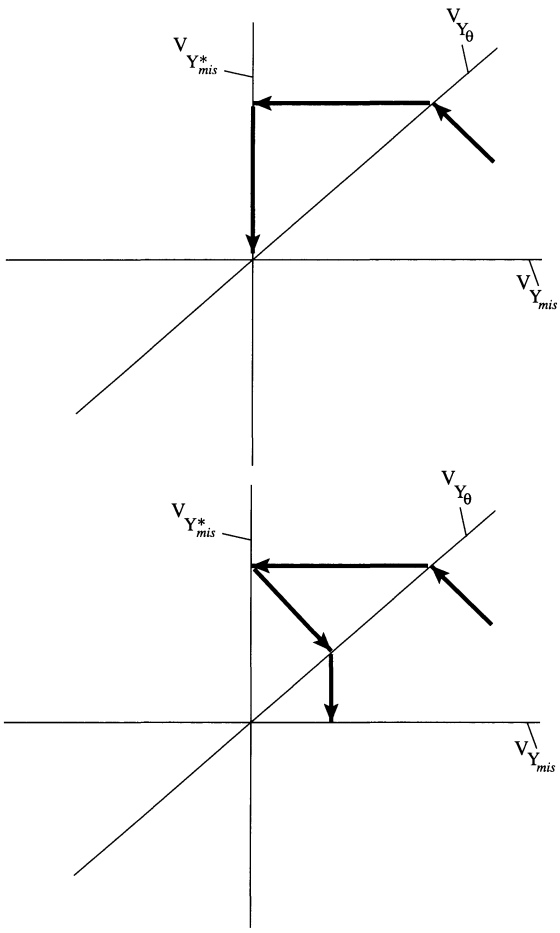

Figure 2. These figures depict convergence in $L^2(\pi)$. Stationarity is represented by the origin. The upper figure shows the interweaved approach and its immediate convergence. The lower figure shows the concatenated sampler which does not achieve immediate convergence, and instead shrinks to 0 geometrically.

Where $D_{obs} = D_{hidden} = N$, both the centered and non-centered methods are geometrically ergodic, though not uniformly in the algorithm starting value. Of course as we have seen, in this case the interweaved approach is uniformly ergodic, even producing i.i.d. output.

If $D_{obs} = D_{hidden} = C$, it turns out that all three algorithms are uniformly ergodic. This can be seen by a simple extension of the results of Papaspiliopoulos and Roberts (2008).

The most interesting case is where the distributions are different. If $D_{obs} = C$ and $D_{hidden} = N$, the centered algorithm is not geometrically ergodic while the non-centered method is in fact uniformly ergodic. Conversely, if $D_{obs} = N$ and $D_{hidden} = C$, then it is the centered algorithm which is uniformly ergodic while the non-centered algorithm is non-geometrically ergodic. However, in both cases, the interweaved approach is provably uniformly ergodic, again by minor extensions on the techniques of Papaspiliopoulos and Roberts (2008).

It is worth noting that these robustness properties are actually shared by (for instance) any random scan mixture of the centered and non-centered algorithms.

5. THE VERSATILITY OF NON-CENTERING

In general, the essential contrast between centering and non-centering lies within the choice between conditional independence and a priori independence. Of course employing the interweaving strategy requires being able to adopt both in separate updates of θ. In many relatively simple models such as those in the article, it is straightforward to construct a one-to-one map between Y_{mis} and Y^*_{mis}. However, other simple relationships between θ and Y_{mis} are substantially more difficult to decouple. A collection of techniques based around auxiliary variable constructions can be used in many of these situations; see, for instance, the work of Papaspiliopoulos, Roberts, and Sköld (2007).

Here we shall concentrate on a very common scenario where the a priori relationship is Poisson. In fact we shall consider the case where the Y_{mis} represents the sample path of a Poisson process which we shall denote by the more usual notation in this context, X, while Y^*_{mis} will be denoted by $\tilde{X}$. Options for non-centering this link are described in the work of Papaspiliopoulos (2003) which we follow closely.

We shall illustrate the techniques through a specific example. Here we let X be a Poisson process on $S = [0, T] \times (0, \infty)$ with intensity function

$$\lambda(c, \epsilon; \Theta) = r\phi \exp\{-\phi\epsilon\}, \qquad (c, \epsilon) \in S, \Theta = (r, \phi). \tag{5.1}$$

Processes such as these, and their non-centered parameterizations, have been used for latent time-series modeling, for instance, for the parameterization of a shot-noise process in the work of Roberts, Papaspiliopoulos, and Dellaportas (2004). We shall also see in Section 6 that such models have an important part to play in exact simulation of diffusions and their related exact MCMC methods Sermaidis et al. (2011) through the use of the *Wiener–Poisson* factorization of diffusion measure (Beskos, Papaspiliopoulos, and Roberts 2008).

By introducing auxiliary variables into non-centered algorithms, many more alternative approaches are available. In this context, for instance, there are (at least) three interesting options for non-centering which can be constructed for this process. We have options for non-centering by Poisson thinning, by using the inverse CDF simulation method, and in addition we have the option to ignore (or not) the product structure which is present in the process in (5.1).

The first non-centering is termed MPP-THIN-NCP (marked Poisson process thinned non-centered parameterization) and takes $\tilde{X}$ to be a Poisson process on $[0, T] \times (0, \infty) \times (0, \infty)$ with mean measure

$$e^{-\tilde{\epsilon}}\, dc\, dm\, d\tilde{\epsilon}, \tag{5.2}$$

and X is retrieved from $\tilde{X} = \{(C_i, M_i, \tilde{E}_i), i = 1, 2, \ldots\}$ and Θ as follows (see also Figure 3):

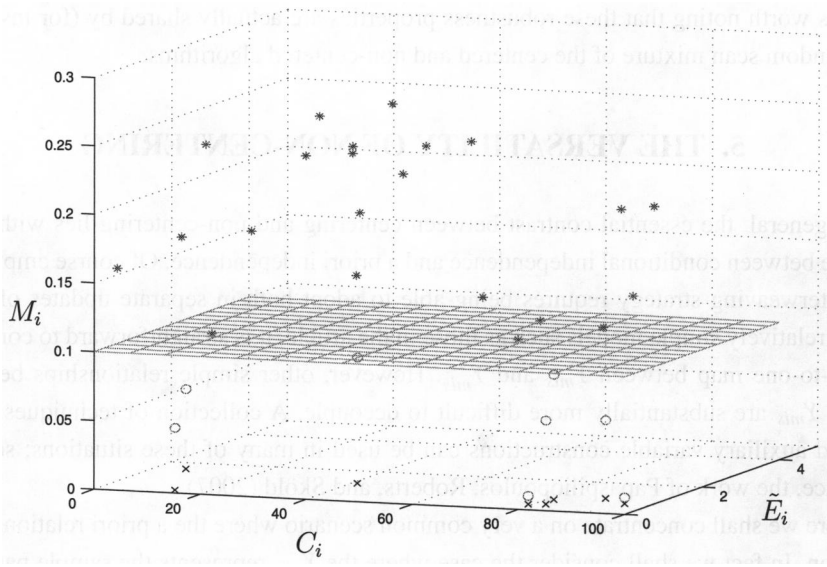

Figure 3. The MPP-THIN-NCP of (Θ, X) for the Poisson process with intensity (5.1). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$, and $T = 100$. $\tilde{X}$ is a Poisson process on $[0, T] \times (0, \infty) \times (0, \infty)$ with mean measure $e^{-\tilde{\epsilon}}\, dc\, dm\, d\tilde{\epsilon}$; choose all $(C_i, M_i, \tilde{E}_i) \in \tilde{X}$ with $M_i \leq r$ (denoted by circles as opposed to the points with $M_i > r$ denoted by asterisks); project them to S; set $E_i = \tilde{E}_i/\phi$. The online version of this figure is in color.

MPP-THIN-NCP:

Let $\tilde{X} = \{(C_i, M_i, \tilde{E}_i), i = 1, 2, \ldots\}$.
Select all points from $\tilde{X}$ for which $M_i < r$.
Project these points to $[0, T] \times (0, \infty)$.
Transform $\{(C_i, \tilde{E}_i)\}$ to $\{(C_i, E_i)\}$ where $E_i = \tilde{E}_i/\phi$.
X consists of the transformed points.

We can also apply the MPP-CDF-NCP as follows. We take $\tilde{X} = \{(C_i, \tilde{E}_i), i = 1, 2, \ldots\}$ to be a unit rate Poisson process on S and transform $\tilde{X} \to X$ as follows (see also Figure 4).

MPP-CDF-NCP:

Let $\tilde{X} = \{(C_i, \tilde{E}_i), i = 1, 2, \ldots\}$.
Select all points $(C_i, \tilde{E}_i) \in \tilde{X}$ for which $\tilde{E}_i < r$.
Set $E_i = -\log\{\tilde{E}_i/r\}/\phi$.
$X = \{(C_i, E_i), i = 1, 2, \ldots\}$.

The third possible NCP ignores the product space structure of S entirely. Thus $\tilde{X} = \{(C_i, E_i, M_i), i = 1, 2, \ldots\}$ is a unit rate Poisson process on $S \times (0, \infty)$ and obtain X from $\tilde{X}$ as described below (see also Figure 5).

THIN-NCP:

Let $\tilde{X} = \{(C_i, E_i, M_i), i = 1, 2, \ldots\}$.
Select all points from $\tilde{X}$ for which $M_i < r\phi \exp\{-\phi E_i\}$.

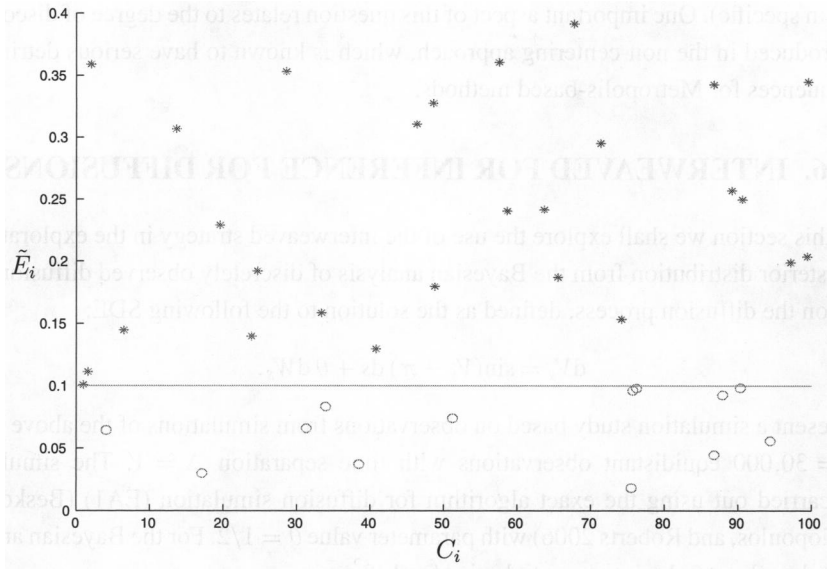

Figure 4. The MPP-CDF-NCP of (Θ, X) for the Poisson process with intensity (5.1). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$, and $T = 100$. $\tilde{X}$ is a unit rate Poisson process on S; choose all $(C_i, \tilde{E}_i) \in \tilde{X}$ with $\tilde{E}_i < r$; set $E_i = -\log(\tilde{E}_i/r)/\phi$. The online version of this figure is in color.

```
Project these points to [0, T] × (0, ∞).
X consists of the projected points.
```

While all these three non-centering mechanisms are plausible, it is an open question as to which should be preferred in MCMC applications (and the answer is almost certainly

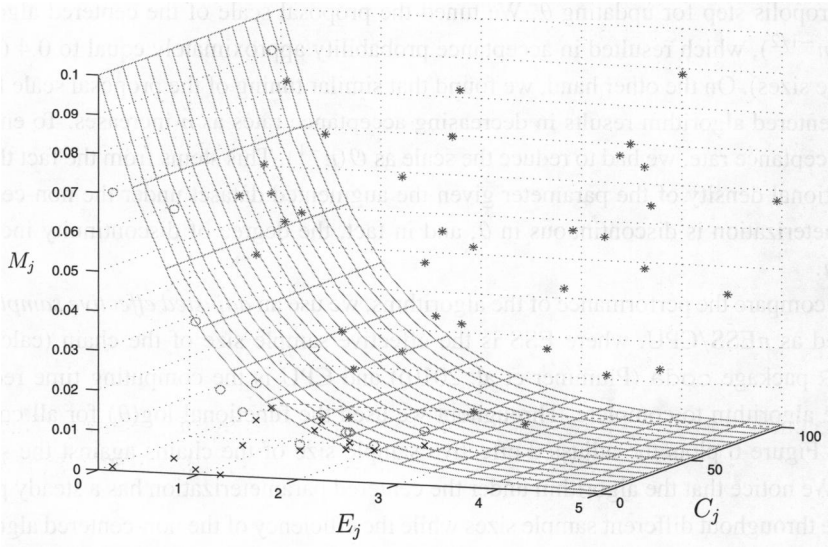

Figure 5. The THIN-NCP of (Θ, X) for the Poisson process with intensity (5.1). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$, and $T = 100$. $\tilde{X}$ is a unit rate Poisson process on $S \times (0, \infty)$ and X consists of all (C_i, E_i) such that $(C_i, E_i, M_i) \in \tilde{X}$ and $M_i < \lambda(C_i, E_i)$. The online version of this figure is in color.

problem specific). One important aspect of this question relates to the degree of discontinuity introduced in the non-centering approach, which is known to have serious detrimental consequences for Metropolis-based methods.

6. INTERWEAVED FOR INFERENCE FOR DIFFUSIONS

In this section we shall explore the use of the interweaved strategy in the exploration of the posterior distribution from the Bayesian analysis of discretely observed diffusions. We focus on the diffusion process, defined as the solution to the following SDE:

$$\mathrm{d}V_s = \sin(V_s - \pi)\,\mathrm{d}s + \theta\,\mathrm{d}W_s.$$

We present a simulation study based on observations from simulations of the above model at $n = 30{,}000$ equidistant observations with time separation $\Delta = 1$. The simulations were carried out using the exact algorithm for diffusion simulation (EA1) (Beskos, Papaspiliopoulos, and Roberts 2006) with parameter value $\theta = 1/2$. For the Bayesian analysis we employed a standard exponential prior for $1/\theta^2$.

Introducing a latent Poisson process through the so-called *Wiener–Poisson* factorization of diffusion measure (Beskos, Papaspiliopoulos, and Roberts 2008), the posterior distribution can be written as an explicit if complex form (see theorem 1 in Sermaidis et al. 2011). Non-centering by thinning can be employed as in Section 5 to provide a non-centered alternative, and thus an interweaved approach can readily be applied also.

We studied the convergence properties of the three competing algorithms on posterior distributions using different subsets of the simulated dataset (all taken to be equally spaced with time separation $\Delta = 1$).

In this problem we are unable to do pure Gibbs updates and we have to incorporate a Metropolis step for updating θ. We tuned the proposal scale of the centered algorithm as $\mathcal{O}(n^{-1/2})$, which resulted in acceptance probability approximately equal to 0.4 (for all sample sizes). On the other hand, we found that similar tuning of the proposal scale for the non-centered algorithm results in decreasing acceptance rates as n increases. To ensure a 0.4 acceptance rate, we had to reduce the scale as $\mathcal{O}(n^{-1})$. This stems from the fact that the conditional density of the parameter given the augmented dataset under the non-centered parameterization is discontinuous in θ, and in fact, the degree of discontinuity increases with n.

To compare the performance of the algorithms, we use an *adjusted effective sample size*, defined as nESS/CPU, where ESS is the effective sample size of the chain (calculated with `R` package `coda` (Plummer et al. 2010)) and CPU is the computing time required by the algorithm to complete. Throughout we used the functional $\log(\theta)$ for all comparisons. Figure 6 plots the adjusted effective sample size of the chains against the sample size. We notice that the algorithm under the centered parameterization has a steady performance throughout different sample sizes while the efficiency of the non-centered algorithm decreases as the sample sizes increases. On the other hand, the interweaved strategy performs better than both until $n = 5000$, after which the increase in the effective sample size is counterbalanced by the increase in the CPU time.

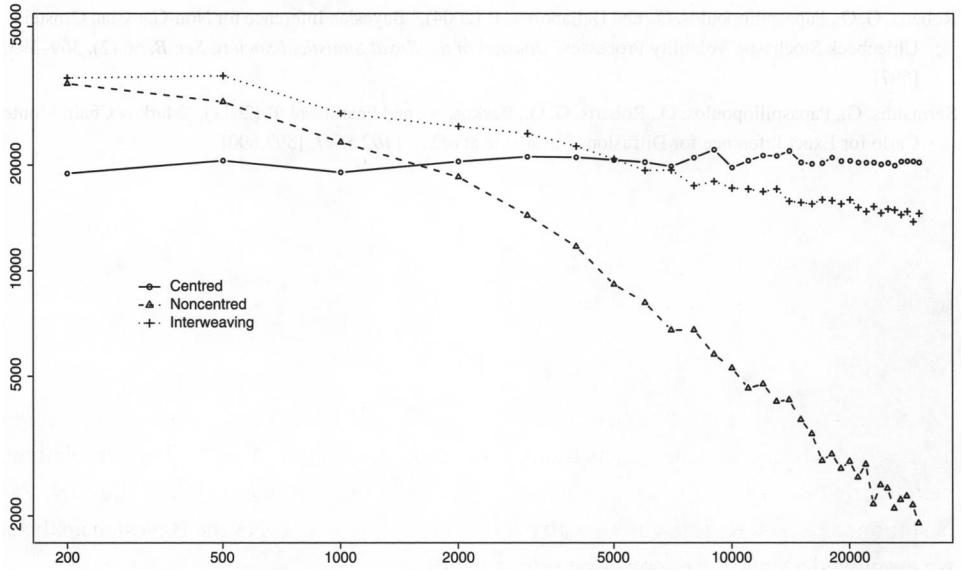

Figure 6. "Exact" MCMC for Bayesian inference for diffusions with increasing number of simulated data points (outfill asymptotics). Adjusted effective sample size (log-scale) of the chains using the centered, non-centered, and interweaved strategies versus the sample size (log-scale).

7. DISCUSSION

The interweaved approach is an important innovation which is extremely easy to implement, at least once a non-centering algorithm is available. Our empirical experience suggests that, at least in some cases, it can lead to significant computational savings. However, there is a good deal to be done, in terms of both the theoretical understanding of the method and empirical investigations.

The authors are to be congratulated on this thought-provoking initial study into the interweaved approach, and we are certain this will stimulate considerable further research developments in the years to come.

ADDITIONAL REFERENCES

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2006), "Retrospective Exact Simulation of Diffusion Sample Paths With Applications," *Bernoulli*, 12 (6), 1077–1098. [600]

——— (2008), "A Factorisation of Diffusion Measure and Finite Sample Path Constructions," *Methodology and Computing in Applied Probability*, 10 (1), 85–104. [597,600]

Neal, P., and Roberts, G. O. (2005), "A Case Study in Non-Centering for Data Augmentation: Stochastic Epidemics," *Statistics and Computing*, 15, 315–327. [594]

Papaspiliopoulos, O. (2003), "Non-Centered Parametrisations for Hierarchical Models and Data Augmentation," Ph.D. thesis, Dept. of Mathematics and Statistics, Lancaster University, Lancaster. [597]

Plummer, M., Best, N., Cowles, K., and Vines, K. (2010), "coda: Output Analysis and Diagnostics for MCMC," R package version 0.13-5. [600]

Roberts, G. O., and Rosenthal, J. S. (2008), "Variance Bounding Markov Chains," *The Annals of Applied Probability*, 18, 1201–1214. [595]

Roberts, G. O., Papaspiliopoulos, O., and Dellaportas, P. (2004), "Bayesian Inference for Non-Gaussian Ornstein–Uhlenbeck Stochastic Volatility Processes," *Journal of the Royal Statistical Society, Ser. B*, 66 (2), 369–393. [597]

Sermaidis, G., Papaspiliopoulos, O., Roberts, G. O., Beskos, A., and Fearnhead, P. (2011), "Markov Chain Monte Carlo for Exact Inference for Diffusions," available at *arXiv:1102.5541*. [597,600]