


DISCUSSION ARTICLE

The Art of Data Augmentation

David A. VAN DYK and Xiao-Li MENG

The term *data augmentation* refers to methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables. For deterministic algorithms, the method was popularized in the general statistical community by the seminal article by Dempster, Laird, and Rubin on the EM algorithm for maximizing a likelihood function or, more generally, a posterior density. For stochastic algorithms, the method was popularized in the statistical literature by Tanner and Wong's Data Augmentation algorithm for posterior sampling and in the physics literature by Swendsen and Wang's algorithm for sampling from the Ising and Potts models and their generalizations; in the physics literature, the method of data augmentation is referred to as the method of *auxiliary variables*. Data augmentation schemes were used by Tanner and Wong to make simulation feasible and simple, while auxiliary variables were adopted by Swendsen and Wang to improve the speed of iterative simulation. In general, however, constructing data augmentation schemes that result in both simple and fast algorithms is a matter of art in that successful strategies vary greatly with the (observed-data) models being considered. After an overview of data augmentation/auxiliary variables and some recent developments in methods for constructing such efficient data augmentation schemes, we introduce an effective search strategy that combines the ideas of *marginal augmentation* and *conditional augmentation*, together with a *deterministic approximation* method for selecting good augmentation schemes. We then apply this strategy to three common classes of models (specifically, multivariate t, probit regression, and mixed-effects models) to obtain efficient Markov chain Monte Carlo algorithms for posterior sampling. We provide theoretical and empirical evidence that the resulting algorithms, while requiring similar programming effort, can show dramatic improvement over the Gibbs samplers commonly used for these models in practice. A key feature of all these new algorithms is that they are positive recurrent subchains of nonpositive recurrent Markov chains constructed in larger spaces.

Key Words: Auxiliary variables; Conditional augmentation; EM algorithm; Gibbs sampler; Haar measure; Hierarchical models; Marginal augmentation; Markov chain Monte Carlo; Mixed-effects models; Nonpositive recurrent Markov chain; Posterior distributions; Probit regression; Rate of convergence.

David A. van Dyk is Associate Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: vandyk@stat.harvard.edu). Xiao-Li Meng is Professor, Department of Statistics, The University of Chicago, Chicago, IL 60637 (E-mail: meng@galton.uchicago.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 1–50

1. DATA AUGMENTATION AND AUXILIARY VARIABLES

Suppose Y_{obs} is our observed data and we want to sample from the posterior $p(\theta|Y_{\text{obs}}) \propto p(Y_{\text{obs}}|\theta)p(\theta)$, where $p(Y_{\text{obs}}|\theta)$ is a probability density with respect to a measure $\mu(\cdot)$ and $p(\theta)$ is our prior density on $\Theta \subset R^d$. It is well known that even with common models, such as those discussed in this article, the posterior sampling required for Monte Carlo integrations may not be trivial. Indeed, until recently, this burden was a major block in the routine use of Bayesian techniques in practice. The situation has changed considerably in the last ten years or so, thanks to powerful Markov chain Monte Carlo (MCMC) sampling methods. The relevant literature on MCMC is simply too extensive to list, but the book edited by Gilks, Richardson, and Spiegelhalter (1996) is worthy of being singled out, because it provides a fairly general picture of MCMC techniques and illustrates them in a variety of real-data applications. It also contains accessible theoretical background as well as a fairly extensive list of references up to 1996. Another useful resource is Neal (1993), especially because it contains an insightful overview of many MCMC methods developed outside of statistics. For the most recent developments in MCMC methodologies in statistics, the MCMC preprint service at http://www.statslab.cam.ac.uk/~mcmc is an excellent resource. For some of the most advanced recent developments in physics, Ceperley's (1995) long review article, in the context of simulating boson superfluid, is essential reading. For detailed illustrations and discussions of MCMC in Bayesian and likelihood computation, the books by Gelman, Stern, and Rubin (1995), Carlin and Louis (1996), and Tanner (1996) cover many models that are routinely encountered in practice.

One very effective tool in the MCMC toolkit is the so-called *data augmentation* technique. The technique was popularized in general for constructing deterministic mode-finding algorithms by Dempster, Laird, and Rubin (1977) in their seminal article on the EM algorithm, but the term *data augmentation* originated with Tanner and Wong's (1987) Data Augmentation (DA) algorithm, which provides a perfect illustration of this technique in a simulation setting. The DA algorithm starts with the construction of the so-called *augmented data*, Y_{aug}, which are linked to the observed data via a many-to-one mapping $\mathcal{M} : Y_{\text{aug}} \to Y_{\text{obs}}$. A data augmentation scheme is a model for Y_{aug}, $p(Y_{\text{aug}}|\theta)$, that satisfies the following constraint

$$\int_{\mathcal{M}(Y_{\text{aug}})=Y_{\text{obs}}} p(Y_{\text{aug}}|\theta)\mu(dY_{\text{aug}}) = p(Y_{\text{obs}}|\theta). \tag{1.1}$$

That is, to be qualified as an augmentation scheme, the marginal distribution of Y_{obs} implied by $p(Y_{\text{aug}}|\theta)$ must be the original model $p(Y_{\text{obs}}|\theta)$. The necessity of this requirement is obvious because $p(Y_{\text{aug}}|\theta)$ is introduced purely for computational purposes and thus should not alter our posited analysis model. (Throughout this article, whenever appropriate, all equalities and inequalities, such as (1.1), are understood to hold almost surely with respect to an appropriate dominating measure.)

The utility of the DA algorithm stems from the fact that with an appropriate choice of $p(Y_{\text{aug}}|\theta)$, sampling from both $p(\theta|Y_{\text{aug}})$ and $p(Y_{\text{aug}}|Y_{\text{obs}}, \theta)$ is much easier than sampling directly from $p(\theta|Y_{\text{obs}})$. Consequently, starting with an initial value, $\theta^{(0)} \in \Theta$, we can form a Markov chain $\{(\theta^{(t)}, Y_{\text{aug}}^{(t)}), t \geq 1\}$ by iteratively drawing $Y_{\text{aug}}^{(t+1)}$ and $\theta^{(t+1)}$ from

$p(Y_{\text{aug}}|\theta^{(t)}, Y_{\text{obs}})$ and $p(\theta|Y_{\text{aug}}^{(t+1)})$, respectively. This is simply a two-step version of the more general Gibbs sampler (Geman and Geman 1984), and thus under the standard regularity conditions for the Gibbs sampler [see Roberts (1996) or Tierney (1994, 1996)], the limiting distribution of $(\theta^{(t)}, Y_{\text{aug}}^{(t)})$ is given by $p(\theta, Y_{\text{aug}}|Y_{\text{obs}})$.

Besides simple implementation, another desirable requirement for the data augmentation scheme is that the resulting Markov chains mix quickly, thus reducing the required computation time. In fact, in some cases a data augmentation scheme has been introduced mainly to improve mixing. This is the case with the well known Swendsen and Wang (1987) algorithm for simulating from the Ising (1925) and Potts (1952) models. Their algorithm is a special case of what Neal (1997) termed *slice sampling*, a general class of which can be formulated as follows. Suppose we want to simulate from a density $f(x)$, which can be written as $f(x) \propto \pi(x)\prod_{k=1}^{K} l_k(x)$. We then can introduce an *auxiliary variable* $u = (u_1, \ldots, u_K) \in (0, \infty)^K$ such that the joint density of x and u (with respect to Lebesgue measure) is given by

$$f(x, u) \propto \pi(x) \prod_{k=1}^{K} I\{u_k \leq l_k(x)\}, \tag{1.2}$$

where $I\{\cdot\}$ is the indicator function. It is clear that the marginal density of x implied by (1.2) is $f(x)$. The Gibbs sampler can then be implemented by (a) simulating u from $f(u|x)$, which amounts to independently simulating u_k from Uniform$(0, l_k(x))$, $k = 1, \ldots, K$, and (b) simulating x from $f(x|u)$, which is $\pi(x)$ truncated to the region $\bigcap_{k=1}^{K}\{x : l_k(x) \geq u_k\}$; when x is multidimensional, further Gibbs steps may be needed to sample from $f(x|u)$. In some applications, such as the Ising model where x is a lattice, $\pi(x)$ is a simple distribution with independence structures among the components of x, and therefore is easy to sample from. The factor $\prod_{k=1}^{K} l_k(x)$, however, reflects the dependence structure among the components of x (e.g., the neighborhood interaction structure in the Ising and Potts models, where k indexes adjacent pixels). This dependence is responsible for the slow mixing when one implements the Gibbs sampler or the Metropolis–Hastings algorithm directly on $f(x) \propto \pi(x)\prod_{k=1}^{K} l_k(x)$. The use of the auxiliary variable u effectively eliminates such interactions and thus reduces the strong autocorrelation in the MCMC draws, as discussed by Besag and Green (1993), Green (1997), and Higdon (1998).

The success of the Swendsen–Wang algorithm has stimulated much interest in the general use of the method of auxiliary variables in the physics literature, most importantly in Edwards and Sokal (1989). In the statistical literature, there also has been growing general interest in this method, apparently starting from the overview article of Besag and Green (1993); important methodological and/or theoretical papers include Damien, Wakefield and Walker (1999), Higdon (1998), Mira and Tierney (1997), Neal (1997), and Roberts and Rosenthal (1997). It is worthwhile to note that the statistical literature on auxiliary variables has grown largely independently of the literature on data augmentation, despite the fact that the two methods are identical in their general form. The general form of the former, of which (1.2) is a special case, is to embed our target distribution (or density) $f(x)$ into $f(x, u)$, where u is an auxiliary variable of arbitrary dimension. This is the same as in

(1.1) if we express (1.1) in the equivalent form:

$$\int_{\mathcal{M}(Y_{\text{aug}})=Y_{\text{obs}}} p(\theta, Y_{\text{aug}}|Y_{\text{obs}})\mu(dY_{\text{aug}}) = p(\theta|Y_{\text{obs}}), \tag{1.3}$$

and identify θ with x, Y_{aug} with u, and $p(\cdot|Y_{\text{obs}})$ with $f(\cdot)$, where "$\cdot$" can be either x or $\{x, u\}$. In other words, we can either view the method of auxiliary variables as data augmentation without any observed data (or equivalently by fixing the observed data to be constant), or view data augmentation as introducing an auxiliary variable into $p(\theta|Y_{\text{obs}})$.

The lack of communication between these two literatures could be due to the "backwards" nature of (1.1) compared to (1.3), and/or due to the initial difference in emphasis between the two methods, namely, the easy implementation from data augmentation versus the improved speed from auxiliary variables. Indeed, until recently, common experience and belief have held that there is a general conflict between simplicity and speed. This is evident, for example, in the literature on the EM algorithm, the predecessor and the deterministic counterpart of the DA algorithm, where it is well known that the theoretical rate of convergence of EM is determined by the so-called "fraction of missing information" (see Section 2). Thus, in terms of the augmented-data Fisher information, the less we augment, the faster the algorithm will be as measured by its theoretical rate of convergence. On the other hand, the less we augment, the more difficult the implementation is expected to be. For example, in the extreme case of no augmentation, $Y_{\text{aug}} = Y_{\text{obs}}$, we are faced with sampling from $p(\theta|Y_{\text{obs}})$ directly.

Although the conflict between speed and simplicity is a common phenomenon with many standard augmentation schemes, we demonstrated recently (Meng and van Dyk 1997, 1998) that with more creative augmentation schemes it is entirely possible to construct EM-type algorithms that are both fast and simple. Finding such an efficient augmentation scheme, however, is largely a matter of art in the sense that it needs to be worked out on a case-by-case basis, sometimes with substantial effort (for those of us who create algorithms, *not for the users*). For example, while the "slicing" technique in (1.2) is a general strategy, it can be difficult to implement when $p(x|u)$ is not easy to sample from and can result in extremely slow algorithms when certain asymmetries arise in the target density (e.g., Gray 1993; Green 1997). Much recent work has been devoted to the development of general strategies for constructing MCMC algorithms that are both fast and simple; see, for example, the work by Damien et al. (1999), Higdon (1998), and Neal (1997) on auxiliary variables and in particular on slice sampling.

Likewise, this article introduces a constructive search strategy for improving standard augmentation schemes and then applies this strategy to construct efficient MCMC algorithms for three common classes of models. This constructive strategy combines the *conditional augmentation* and *marginal augmentation* approaches developed by Meng and van Dyk (1999), which were inspired, respectively, by Meng and van Dyk's (1997) *working parameter* approach and Liu, Rubin, and Wu's (1998) *parameter expansion* approach, both designed to speed up EM-type algorithms. The marginal augmentation approach was developed independently by Liu and Wu (1999) under the name parameter-expanded DA algorithm; see also C. Liu (1999) for a related method called the "covariance-adjusted DA algorithm." Our strategy includes a method we call the *deterministic approximation* for choosing optimal or nearly optimal data augmentation schemes. This method circumvents

the difficulties of directly comparing the theoretical rates of convergence of stochastic DA algorithms by comparing the rates of their deterministic counterparts; that is, EM-type algorithms. An interesting phenomenon in all three applications presented in this article is that the resulting algorithms use positive recurrent subchains of nonpositive recurrent Markov chains.

The remainder of this article is divided into eight sections. Sections 2 and 3 review the basic ideas underlying conditional and marginal augmentation respectively. Section 4 discusses the use and theory of improper priors for marginal augmentation, which leads to nonpositive recurrent Markov chains containing properly converging subchains with the desired limiting distributions. Section 5 provides some comparisons of conditional augmentation and marginal augmentation, and introduces our general search strategy which uses the two approaches in tandem. Sections 6–8 apply this general strategy, respectively, to three common models: multivariate t, probit regression, and mixed-effects models. Section 9 concludes with discussion of limitations and generalizations of our search strategies and calls for more theoretical research on nonpositive recurrent Markov chains.

2. CONDITIONAL AUGMENTATION AND THE EM CRITERION

The key to the methods discussed in this article is the introduction of a "working parameter" that is identifiable under the augmented model but not under the observed-data model. Specifically, we introduce a working parameter α into (1.1),

$$\int_{\mathcal{M}(Y_{\text{aug}})=Y_{\text{obs}}} p(Y_{\text{aug}}|\theta,\alpha)\mu(dY_{\text{aug}}) = p(Y_{\text{obs}}|\theta). \tag{2.1}$$

That is, we create a class of augmentation schemes, $p(Y_{\text{aug}}|\theta,\alpha)$ or, equivalently, a class of auxiliary variables indexed by $\alpha \in \mathcal{A}$. In real applications, such as those in Sections 6–8, the working parameter is chosen so that a common augmentation scheme corresponds to a specific value of α (e.g., $\alpha = 1$) and thus direct comparisons can be made with the common augmentation scheme when we search for better schemes.

Once such a class of augmentation schemes is constructed, we can search for the best value of α according to some sensible criterion. This strategy was referred to as the *conditional augmentation* approach by Meng and van Dyk (1999) because, once a desirable value of α is found, it is conditioned upon throughout the algorithm. Meng and van Dyk (1999) discussed three criteria for choosing α, in the order of decreasing theoretical appeal but of increasing practicality. The first is to minimize the geometric rate of convergence of the DA algorithm (see Amit 1991 and Liu, Wong, and Kong 1994)

$$\lambda^{\text{DA}}(\alpha) = 1 - \inf_{h:\text{var}[h(\theta)|Y_{\text{obs}}]=1} \text{E}[\text{var}(h(\theta)|Y_{\text{aug}},\alpha)|Y_{\text{obs}},\alpha], \tag{2.2}$$

where the expectation is with respect to the stationary density $p(\theta, Y_{\text{aug}}|Y_{\text{obs}},\alpha)$. The second is to minimize the maximum autocorrelation over linear combinations (Liu 1994)

$$\sup_{x\neq 0} \text{corr}(x^\top\theta^{(t)}, x^\top\theta^{(t+1)}) = \sup_{x\neq 0} \frac{x^\top \text{var}[\text{E}(\theta|Y_{\text{aug}},\alpha)|Y_{\text{obs}},\alpha]x}{x^\top \text{var}(\theta|Y_{\text{obs}})x} = \rho(\mathcal{F}_B(\alpha)), \tag{2.3}$$

where $\mathcal{F}_B(\alpha)$ is the so-called Bayesian fraction of missing information

$$\mathcal{F}_B(\alpha) = I - [\text{var}(\theta|Y_{\text{obs}})]^{-1}\text{E}[\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha], \tag{2.4}$$

and $\rho(A)$ is the spectral radius of A. Thus, if we have two augmentation schemes indexed by α_1 and α_2, the second criterion will prefer the scheme with larger expected conditional variance, $\text{E}[\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ (using a positive semidefinite ordering when θ is a vector). This autocorrelation criterion is more general than the geometric-rate criterion as it can be applied to Markov chains that do not converge at a geometric rate.

The third criterion is based on the intrinsic connection between the DA algorithm and its deterministic counterpart and predecessor, the EM algorithm. Specifically, given a conditional augmentation scheme $p(Y_{\text{aug}}|\theta, \alpha)$, the corresponding EM algorithm for computing the posterior mode(s) of $p(\theta|Y_{\text{obs}})$, denoted by $\theta^\star$, has a theoretical rate of convergence given by (Dempster, Laird, and Rubin 1977)

$$\mathcal{F}_{\text{EM}}(\alpha) = I - I_{\text{obs}} I_{\text{aug}}^{-1}(\alpha),$$

where

$$I_{\text{aug}}(\alpha) = \text{E}\left[-\frac{\partial^2 \log p(\theta|Y_{\text{aug}}, \alpha)}{\partial\theta \cdot \partial\theta}\Big|Y_{\text{obs}}, \theta, \alpha\right]\Bigg|_{\theta=\theta^\star} \tag{2.5}$$

is the expected augmented Fisher information matrix, and

$$I_{\text{obs}} = -\frac{\partial^2 \log p(\theta|Y_{\text{obs}})}{\partial\theta \cdot \partial\theta}\Big|_{\theta=\theta^\star}$$

is the observed Fisher information matrix. Here we adopt the traditional terms (e.g., Fisher information) of the EM literature, which primarily focuses on the likelihood computation, even though we are dealing with the more general posterior computation. In particular, $\mathcal{F}_{\text{EM}}(\alpha)$, which is called the matrix fraction of missing information, can be viewed as the likelihood analogue of $\mathcal{F}_B(\alpha)$. Indeed, when $p(\theta, Y_{\text{aug}}|Y_{\text{obs}}, \alpha)$ is normal, $\mathcal{F}_{\text{EM}}(\alpha) = \mathcal{F}_B(\alpha)$ (e.g., Sahu and Roberts 1999). We propose the EM criterion for choosing α. Namely, we suggest minimizing $I_{\text{aug}}(\alpha)$ via a positive semidefinite ordering and thus minimizing $\rho(\mathcal{F}_{\text{EM}}(\alpha))$. Strictly speaking, we should call this the *matrix-rate* EM criterion in contrast to the *global-rate* EM criterion which directly minimizes $\rho(\mathcal{F}_{\text{EM}}(\alpha))$. The latter is more general since $I_{\text{aug}}(\alpha)$ may not exhibit the positive semidefinite ordering (see Section 8), but is often much more difficult to implement. See Meng (1994) and Meng and Rubin (1994) for discussion on the relationship between the matrix rate and the global rate.

In general, it is a weaker requirement for the minimizer of $\rho(\mathcal{F}_{\text{EM}}(\alpha))$ to approximate that of $\rho(\mathcal{F}_B(\alpha))$ well than for $\mathcal{F}_{\text{EM}}(\alpha)$ to approximate $\mathcal{F}_B(\alpha)$ well as functions of α; empirical evidence is provided by Meng and van Dyk (1999) as well as in this article (e.g., in Sections 6 and 7 the *deterministic approximation* method finds the exact optimal algorithms as defined by Liu and Wu's group theoretic formulation). The essence of this method is that whenever it is too difficult to compare two stochastic algorithms (e.g., DA) directly, we compare their deterministic counterparts (e.g., EM) to decide which stochastic algorithm to use. This does not necessarily lead to the best stochastic algorithm even if we find the optimal deterministic algorithm, but it often leads to good stochastic algorithms

with reasonable analytical effort. The utility of the EM criterion is that it is much easier to handle analytically and typically does not require knowing the value of $\theta^\star$, as demonstrated in Sections 6–8.

We emphasize that whereas the EM criterion is a useful strategy for finding good choices of α, it is not always applicable. And, obviously, there are other ways of finding suitable choices of α, especially when aided by considerations for specific applications. For example, Higdon (1993, 1998) proposed the method of *partial decoupling* to combat the slow mixing of both the direct Gibbs sampler and the Swendsen–Wang algorithm for Ising-type models with multiple modes. His method introduces a working parameter $\alpha = (\alpha_1, \ldots, \alpha_K) \in [0,1]^K$ into (1.2):

$$f(x,u|\alpha) \propto \pi(x) \prod_{k=1}^{K} l_k^{1-\alpha_k}(x) I\{u_k \le l_k^{\alpha_k}(x)\}. \tag{2.6}$$

He discussed many methods for choosing α so that the resulting algorithm is faster than either the direct Gibbs sampler (with $\alpha = (0, \ldots, 0)$) or the Swendsen–Wang algorithm (with $\alpha = (1, \ldots, 1)$). In particular, Higdon (1998) demonstrated empirically a dramatic improvement by setting $\alpha_{\{i,j\}} = 1/(1 + |y_i - y_j|)$ where $\{y_i, y_j\}$ are recorded data from a pair of adjacent pixels indexed by $k \equiv \{i,j\}$, which are used to formulate $\pi(x)$. This choice is not based on the EM criterion, which in fact is not applicable here because $\log f(x,u|\alpha)$ is undefined, but rather it is based on heuristic arguments and empirical evidence including more rapid jumps between modes. It is conceivable, however, to work directly with $\lambda^{\mathrm{DA}}(\alpha)$ to determine optimal or near optimal choices of α, though it is unclear whether such effort will pay off as the computation needed for finding a (nearly) optimal choice of α may completely wipe out the savings offered by this choice.

3. MARGINAL AUGMENTATION AND A MARGINALIZATION STRATEGY

A second method for using the working parameter α is to integrate both sides of (2.1) with respect to a proper *working prior* $p(\alpha)$; that is,

$$\int_{\mathcal{M}(Y_{\mathrm{aug}})=Y_{\mathrm{obs}}} \left[\int p(Y_{\mathrm{aug}}|\theta,\alpha) p(d\alpha) \right] \mu(dY_{\mathrm{aug}}) = p(Y_{\mathrm{obs}}|\theta). \tag{3.1}$$

Meng and van Dyk (1999) referred to this as *marginal augmentation* because it creates a new data augmentation scheme by marginalizing out the working parameter α:

$$p(Y_{\mathrm{aug}}|\theta) = \int p(Y_{\mathrm{aug}}|\theta,\alpha) p(d\alpha). \tag{3.2}$$

Note that in (3.1) we have implicitly assumed that θ and α are a priori independent. Although this assumption simplifies certain theoretical results and appears to be adequate for practical purposes (see Meng and van Dyk 1999 and Liu and Wu 1999), it is not necessary. That is, (3.1) still holds if $p(\alpha)$ is replaced by $p(\alpha|\theta)$.

Initially, it may appear that we have accomplished nothing, since (3.1) is symbolically identical to (1.1) via the notation in (3.2). As discussed by Meng and van Dyk (1999), this symbolic equivalence is both correct and deceptive. It is correct because $p(Y_{\text{aug}}|\theta)$ in (3.2) is a legitimate data augmentation scheme (when $p(\alpha)$ is proper) and thus should satisfy the general definition given by (1.1). It is deceptive because in the context of (3.1) and (3.2), the dependency of the conditional augmentation scheme on the working parameter is suppressed in (1.1).

The following identity given by Meng and van Dyk (1999) and Liu and Wu (1999) is a key to understanding the marginal augmentation approach. Under the joint distribution of $(\theta, \alpha, Y_{\text{aug}})$ given by

$$p(\theta, \alpha, Y_{\text{aug}}) = p(Y_{\text{aug}}|\theta, \alpha)p(\theta)p(\alpha), \tag{3.3}$$

we have

$$\begin{aligned}\text{E}\left[\text{var}(h(\theta)|Y_{\text{aug}})|Y_{\text{obs}}\right] = \text{E}\left\{\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]|Y_{\text{obs}}\right\} \\ +\text{E}\left\{\text{var}[\text{E}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{aug}}]|Y_{\text{obs}}\right\},\end{aligned} \tag{3.4}$$

for any square-integrable $h(\theta)$. Consequently, if $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ does not depend on α, the expected conditional variance of $h(\theta)$ under the marginal augmentation scheme (i.e., $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}})|Y_{\text{obs}}]$) cannot be smaller than the expected conditional variance under any conditional augmentation scheme (i.e., $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$). It follows then from (2.2) that the rate of convergence of the DA algorithm under marginal augmentation cannot exceed its rate under the conditional augmentation scheme. Note that when $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ does not depend on α, all conditional augmentation schemes are equivalent in terms of the rate of convergence of the resulting DA algorithms (see (2.2)). We emphasize that when $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ does depend on α, maximizing this quantity can be beneficial, and thus in general the marginal augmentation approach does not dominate the conditional augmentation approach (see Section 5 of this article and Liu and Wu 1999).

Meng and van Dyk (1999) proved that, starting from any augmentation scheme of the form $\tilde{Y}_{\text{aug}} = \{Y_{\text{obs}}, \tilde{Y}_{\text{mis}}\}$, the following strategy ensures that $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ is free of α.

A Marginalization Strategy

Step 1: For α in a selected set $\mathcal{A}$, construct a one-to-one mapping, $\mathcal{D}_\alpha$, on the $\tilde{Y}_{\text{mis}}$ space and then define $Y_{\text{aug}} = \{Y_{\text{obs}}, \mathcal{D}_\alpha(\tilde{Y}_{\text{mis}})\}$. The set $\mathcal{A}$ should include some α_0 such that the corresponding $\mathcal{D}_{\alpha_0}$ is an identity mapping. The distribution of Y_{aug} induced by the distribution of $\tilde{Y}_{\text{mis}}$ and $\mathcal{D}_\alpha$ gives a class of conditional augmentation schemes indexed by α.

Step 2: Choose a proper prior distribution $p(\alpha)$ (*independent* of θ) to define a marginal augmentation scheme as in (3.2).

Promising choices of $\mathcal{D}_\alpha$ include *rescaling* (e.g., $\mathcal{D}_\alpha(\tilde{Y}_{\text{mis}}) = \alpha\tilde{Y}_{\text{mis}}$), *recentering* (e.g., $\mathcal{D}_\alpha(\tilde{Y}_{\text{mis}}) = \alpha + \tilde{Y}_{\text{mis}}$), and more generally the affine transformation, $\mathcal{D}_\alpha(\tilde{Y}_{\text{mis}}) = \alpha_1\tilde{Y}_{\text{mis}} +$

α_2, as discussed in the rejoinder of Meng and van Dyk (1997) and illustrated in Sections 6–8.

In practice, the integration in (3.2) is avoided by first drawing α from the prior distribution $p(\alpha)$ and then drawing Y_{aug} from $p(Y_{\text{aug}}|\theta, \alpha)$. When using marginal augmentation, there are (at least) three ways to implement a Gibbs sampler, corresponding to the three schemes of Liu, Wong, and Kong (1994). The three schemes iteratively sample from the following distributions

Scheme 1: $p(Y_{\text{aug}}|\theta, Y_{\text{obs}})$ and $p(\theta|Y_{\text{aug}})$
(inducing a Markov chain for θ);

Scheme 2: $p(Y_{\text{aug}}|\theta, \alpha, Y_{\text{obs}})$ and $p(\theta, \alpha|Y_{\text{aug}})$
(inducing a Markov chain for (θ, α));

Scheme 3: $p(Y_{\text{aug}}|\theta, \alpha, Y_{\text{obs}})$, $p(\theta|\alpha, Y_{\text{aug}})$, and $p(\alpha|Y_{\text{aug}}, \theta)$
(inducing a Markov chain for (θ, α)).

As discussed by Meng and van Dyk (1999), Scheme 1 is preferable to Scheme 2 when using a proper working prior, but Scheme 2 is useful when using improper priors for α (see Section 4). Scheme 3, which no longer is a DA algorithm but rather is a three-step Gibbs sampler, typically has a slower rate of convergence than either Scheme 1 or Scheme 2. In fact, Scheme 3 can completely wipe out the benefit of marginal augmentation; see Section 9.1. But this scheme can be useful in some applications as a trade-off between easy implementation and a fast mixing rate, when it is easier to draw from $p(\alpha|Y_{\text{aug}}, \theta)$ and $p(\theta|Y_{\text{aug}}, \alpha)$ than from $p(\theta, \alpha|Y_{\text{aug}})$ or $p(\theta|Y_{\text{aug}})$. More generally, for simpler implementation, any of Y_{aug}, θ, and α can be further split into their respective subcomponents to be sampled via Gibbs sampling steps or Metropolis–Hastings steps.

4. MARGINAL AUGMENTATION WITH AN IMPROPER PRIOR

Because our goal is to increase the expected conditional variance of $h(\theta)$ (see (2.2)), one may expect that with certain choices of $\mathcal{D}_\alpha$ the maximum of this variance is achieved when the prior density $p(\alpha)$ becomes very diffuse, or even improper. An example was given by Meng and van Dyk (1999), and further examples appear in Sections 6–8. When using an improper prior $p(\alpha)$, however, any induced Markov chain involving α cannot be positive recurrent because $p(\alpha|Y_{\text{obs}})$ is the same as $p(\alpha)$ and thus it is improper. This is not necessarily a problem, however, since our interest is in the marginal posterior distribution $p(\theta|Y_{\text{obs}})$, not the improper joint posterior distribution $p(\theta, \alpha|Y_{\text{obs}})$.

Currently, there are two types of theoretical results that guide the choices of improper working prior. The more general type of results involves a limiting argument obtained independently by Meng and van Dyk (1999) and Liu and Wu (1999). Briefly, if an improper prior $p(\alpha)$ results in a transition kernel for θ that is the limit of a sequence of transition kernels each resulting from a proper prior, then the stationary distribution of the subchain

$\{\theta^{(t)}, t \geq 0\}$ is our desired posterior $p(\theta|Y_{\text{obs}})$. Often this limiting condition can be verified by explicitly deriving the stochastic mapping under Scheme 1, $\theta^{(t+1)} = M_\omega(\theta^{(t)})$, where ω indexes a class of proper working priors, $\{p(\alpha|\omega),\ \omega \in \Omega\}$, and then showing that the stochastic mapping under Scheme 2 using an improper working prior is the limit of $M_\omega(\theta^{(t)})$ as, say, $\omega \to \infty$. This is the approach taken by Meng and van Dyk (1999) and Liu and Wu (1999), and is further applied in Sections 6–7. When it is not convenient to explicitly express M_ω, as in the application in Section 8, the following lemma offers an alternative method of establishing the limiting condition [i.e., the two conditions of the lemma are sufficient for the conditions of Lemma 1 of Liu and Wu (1999) and of Theorem 2 of Meng and van Dyk (1999)].

Lemma 1. *Consider implementing Scheme 2 under the Marginalization Strategy with an improper working prior, $p_0(\alpha)$. Let $p_0(\theta, \alpha|Y_{\text{aug}})$ be the corresponding (proper) joint posterior of (θ, α) given the augmented data, $Y_{\text{aug}} = \mathcal{D}_\alpha(\tilde{Y}_{\text{aug}}) \equiv \{Y_{\text{obs}}, \mathcal{D}_\alpha(\tilde{Y}_{\text{mis}})\}$. Suppose*

1. *there exists a sequence of proper working priors indexed by ω, $p(\alpha|\omega)$, and an ω_0 such that the corresponding $p(\theta, \alpha|Y_{\text{aug}}, \omega)$ converges to $p_0(\theta, \alpha|Y_{\text{aug}})$ as $\omega \to \omega_0$; and*
2. *$p_0(\theta|\mathcal{D}_\alpha(\tilde{Y}_{\text{aug}}))$ is invariant to α.*

Then the subchain $\{\theta^{(t)}, t \geq 0\}$ induced by Scheme 2 under $p_0(\alpha)$ is Markovian and its transition kernel is the limit, as $\omega \to \omega_0$, of the transition kernel for θ from Scheme 1 with the working prior $p(\alpha|\omega)$.

Proof: At the $(t+1)$st iteration, the transition kernel from Scheme 1 under a proper prior $p(\alpha|\omega)$, $p^{(1)}(\theta^{(t+1)}|\theta^{(t)}, \omega)$, is given by the following two steps:

1.1 draw $\tilde{Y}_{\text{aug}}^{(t+1)}$ from $p(\tilde{Y}_{\text{aug}}|\theta^{(t)}, Y_{\text{obs}})$ and $\alpha_\omega^{(t+1)}$ from $p(\alpha|\omega)$; and

1.2 draw $\theta^{(t+1)}$ from $p(\theta|Y_{\text{aug}} = \mathcal{D}_{\alpha_\omega^{(t+1)}}(\tilde{Y}_{\text{aug}}^{(t+1)}), \omega)$.

Similarly, for Scheme 2 under $p_0(\alpha)$, the transition kernel $p^{(2)}(\theta^{(t+1)}, \alpha^{(t+1)}|\theta^{(t)}, \alpha^{(t)})$ is given by:

2.1 draw $\tilde{Y}_{\text{aug}}^{(t+1)}$ from $p(\tilde{Y}_{\text{aug}}|\theta^{(t)}, Y_{\text{obs}})$; and

2.2 draw $(\theta^{(t+1)}, \alpha^{(t+1)})$ from $p_0(\theta, \alpha|Y_{\text{aug}} = \mathcal{D}_{\alpha^{(t)}}(\tilde{Y}_{\text{aug}}^{(t+1)}))$.

Under Condition 1 and by the Fatou Lemma, $p_0(\theta|Y_{\text{aug}}) = \lim_{\omega\to\omega_0} p(\theta|Y_{\text{aug}}, \omega)$. Therefore given the same value of Y_{aug}, the transition kernel under Step 2.2 for $\theta^{(t+1)}$ is the limit of that of Step 1.2 when $\omega \to \omega_0$. However, because of Condition 2, the transition kernel for $\theta^{(t+1)}$ in Step 2.2 is unchanged if we replace $\alpha^{(t)}$ with $\alpha_\omega^{(t+1)}$ from Step 1.1 for any ω. Consequently $p^{(2)}(\theta^{(t+1)}|\theta^{(t)}, \alpha^{(t)}) = \lim_{\omega\to\omega_0} p^{(1)}(\theta^{(t+1)}|\theta^{(t)}, \omega)$, and hence we have both conclusions. □

The simplicity of applying Lemma 1 is that Condition 1 is typically automatic when we obtain the improper working prior as the limit of a sequence of proper working priors, as is the case in all of the applications in this article, and that Condition 2 deals only with the *limiting* case. It is also clear that Scheme 2 differs from Scheme 1 when using the same proper prior $p(\alpha|\omega)$, since it sets $\alpha_\omega^{(t+1)} = \alpha^{(t)}$ instead of drawing $\alpha_\omega^{(t+1)}$ from $p(\alpha|\omega)$. Note also that in practice it is often easier to implement Step 1.2 in the manner of Step 2.2 and then discard $\alpha^{(t+1)}$.

The second class of theoretical results justifying the use of improper working priors are due to Liu and Wu (1999) and involve the use of an invariant measure (i.e., Haar measure)

on $\{\alpha^{-1}, \alpha \in \mathcal{A}\}$, where $\mathcal{A}$ is a unimodular group (Nachbin 1965). Here α^{-1} is defined through $\mathcal{D}_{\alpha^{-1}} = \mathcal{D}_\alpha^{-1}$; note that Liu and Wu's (1999) "data transformation" is defined through $\tilde{Y}_{\text{aug}} = t_\alpha(Y_{\text{aug}})$ and thus their t_α is our $\mathcal{D}_\alpha^{-1}$. The beauty of the group formulation of Liu and Wu (1999) is that it not only guarantees the validity of the choice but also a type of optimality—within a *single* data augmentation scheme no proper working prior can produce a faster rate of convergence than the Haar measure. A restriction is that this result does not cover applications like the one given in Section 8 because the affine transformation does not form a unimodular group. More recent work by Liu and Sabatti (2000) has proved the validity, but not the optimality, of using *right Haar* measure of α [corresponding to the left Haar measure in Liu and Wu's (1999) notation]. Establishing the optimality of the right Haar measure remains an open problem, in particular when compared to other improper priors (see the examples in Sections 6–8). Furthermore, no general theoretical results that compare the performance of different data augmentation schemes are currently available (see Section 9).

5. COMPARING AND COMBINING CONDITIONAL AND MARGINAL AUGMENTATION

The previous discussion on the use of improper prior distributions for α hinted that the marginal augmentation approach is also conditional in the sense that it conditions on a particular choice of $p(\alpha)$ (or more generally, $p(\alpha|\theta)$). Thus, mathematically speaking, for a given working parameter α, we can consider optimizing over the choice of $p(\alpha)$. However, optimizing over all possible priors is not always practical nor desirable in real applications—recall that our goal is to find algorithms that are both easy to implement and fast to converge. A more fruitful approach, in general, is to optimize over a class of conveniently parameterized prior distributions, say, $p(\alpha|\omega)$ for $\omega \in \Omega$. That is, we move the conditioning to a higher level in the augmented-data model when we condition on the optimal value of ω rather than the optimal value of α. In other words, we can extend the Marginalization Strategy to

A Combined Strategy:

Step 1: Same as Step 1 of the Marginalization Strategy (p. 8).

Step 2: Same as Step 2 of the Marginalization Strategy except that $p(\alpha)$ is now $p(\alpha|\omega)$, $\omega \in \Omega$. A convenient and useful choice of $p(\alpha|\omega)$ is the (conditional) conjugate prior for the augmented model $p(Y_{\text{aug}}|\theta, \alpha)$.

Step 3: Use a conditional augmentation criterion to select a desirable value of $\omega \in \Omega$, by treating

$$p(Y_{\text{aug}}|\theta, \omega) = \int p(Y_{\text{aug}}|\theta, \alpha)p(d\alpha|\omega) \tag{5.1}$$

as the class of conditional augmentation schemes.

Although any sensible criterion can be used in Step 3, in practice it is often convenient to use the EM criterion. Further simplification is needed, however, when implementing

the EM criterion at the *level-two conditional augmentation* (i.e., Step 3) in order to avoid the integration in (5.1). We found the following normal approximation quite useful in our applications (see Sections 6–8).

Typically, since we start with $p(\tilde{Y}_{\text{aug}}|\theta)$, a standard data augmentation scheme, it is natural to consider the reduction in the augmented Fisher information from the Marginalization Strategy compared to that resulting from $p(\tilde{Y}_{\text{aug}}|\theta)$. The augmented information from the original augmentation is given by (2.5) with $\alpha = \alpha_0$ (recall that Step 1 requires that $\tilde{Y}_{\text{aug}}$ correspond to the conditional augmentation scheme when $\alpha = \alpha_0$) and the augmented information resulting from the augmentation scheme given by (5.1) is defined the same way but with $p(Y_{\text{aug}}|\theta, \alpha)$ replaced by $p(Y_{\text{aug}}|\theta, \omega)$. (That is, the level-one working parameter α is replaced by the level-two working parameter ω, which indicates the change of augmentation schemes as well.) To distinguish between these two different levels of conditioning more explicitly, we use $I^{(1)}_{\text{aug}}(\alpha)$ for the level-one augmented information and $I^{(2)}_{\text{aug}}(\omega)$ for level-two augmented information. We also denote by

$$\Delta^{(2)}_{\text{EM}}(\omega) = I^{(1)}_{\text{aug}}(\alpha_0) - I^{(2)}_{\text{aug}}(\omega) \tag{5.2}$$

the absolute reduction achieved by (5.1). Clearly, minimizing $I^{(2)}_{\text{aug}}(\omega)$ as suggested by the EM criterion for the level-two working parameter is equivalent to maximizing $\Delta^{(2)}_{\text{EM}}(\omega)$. Likewise, we have:

Criterion 5.1. *The EM criterion for selecting α, the level-one working parameter (exactly as described in Section 2), is equivalent to maximizing*

$$\Delta^{(1)}_{\text{EM}}(\alpha) = I^{(1)}_{\text{aug}}(\alpha_0) - I^{(1)}_{\text{aug}}(\alpha). \tag{5.3}$$

We suggest a normal approximation to compute $\Delta^{(2)}_{\text{EM}}(\omega)$ and thus to avoid the integration in (5.1). Under the assumption that $p(\theta, \alpha|Y_{\text{aug}}, \omega)$ is normal, it is easy to verify that

$$\Delta^{(2)}_{\text{EM}}(\omega) = \mathcal{I}_{\theta\alpha}(\omega)\mathcal{I}^{-1}_{\alpha\alpha}(\omega)\mathcal{I}^{\top}_{\theta\alpha}(\omega), \tag{5.4}$$

where $\mathcal{I}_{\theta\alpha}(\omega)$ and $\mathcal{I}_{\alpha\alpha}(\omega)$ are submatricies of the augmented Fisher information for the joint parameter $\tilde{\theta} = \{\theta, \alpha\}$ given by

$$\tilde{I}_{\text{aug}}(\omega) = \mathrm{E}\left[-\frac{\partial^2 \log p(\tilde{\theta}|Y_{\text{aug}}, \omega)}{\partial\tilde{\theta} \cdot \partial\tilde{\theta}}\Big|Y_{\text{obs}}, \tilde{\theta}, \omega\right]\Bigg|_{\tilde{\theta}=\tilde{\theta}^\star} \equiv \begin{pmatrix} \mathcal{I}_{\theta\theta}(\omega) & \mathcal{I}_{\theta\alpha}(\omega) \\ \mathcal{I}^{\top}_{\theta\alpha}(\omega) & \mathcal{I}_{\alpha\alpha}(\omega) \end{pmatrix} \tag{5.5}$$

using the standard submatrix notation, where $\tilde{\theta}^\star = \{\theta^\star, \hat{\alpha}(\omega)\}$ with $\hat{\alpha}(\omega)$ being the mode of $p(\alpha|\omega)$.

Criterion 5.2 *We select ω by maximizing the right side of (5.4) even when the normal assumption is not true (which is typically the case in practice).*

In other words, although we arrived at (5.4) under a normal assumption, in practice we will treat maximizing (5.4) as a criterion in its own right; the effectiveness of this criterion

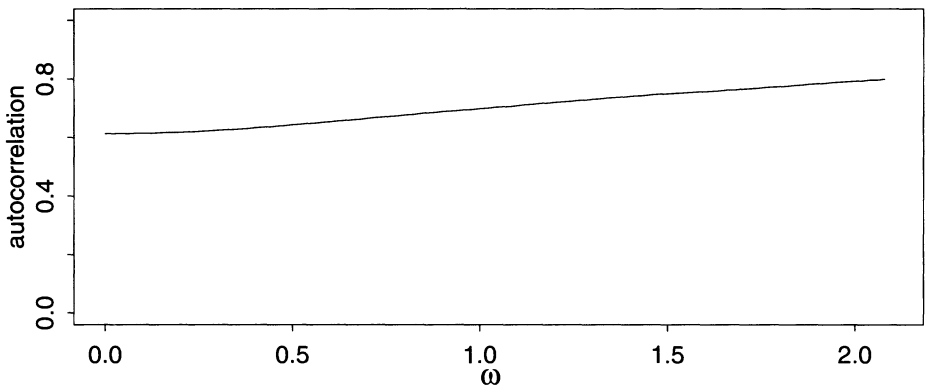

Figure 1. Marginalizing out a Conditional Augmentation Working Parameter. Shown is the (approximate) lag-one autocorrelation for σ^2 as a function of the width of the uniform prior for the working parameter α in model (5.6). Note that the autocorrelation increases *with ω, the level-two working parameter, and thus the optimal value of ω is zero.*

is demonstrated in Sections 6–8. Note that we do not need to compute $\mathcal{I}_{\theta\theta}(\omega)$ in order to use this criterion, which can be a real boon in practice (e.g., Section 6).

Similar to the situation discussed in Section 2 (p. 6), under the further assumption that $p(Y_{\text{aug}}, \theta, \alpha | Y_{\text{obs}}, \omega)$ is normal, $\Delta^{(2)}_{\text{EM}}(\omega)$ in (5.2) is the same as

$$\{\text{E}[\text{var}(\theta|Y_{\text{aug}}, \alpha_0)|Y_{\text{obs}}, \alpha_0]\}^{-1} - \{\text{E}[\text{var}(\theta|Y_{\text{aug}}, \omega)|Y_{\text{obs}}, \omega]\}^{-1}.$$

Consequently, in general, we can view maximizing $\Delta^{(2)}_{\text{EM}}(\omega)$ over ω as an attempt to approximately maximize $\text{E}[\text{var}(\theta|Y_{\text{aug}}, \omega)|Y_{\text{obs}}, \omega]$ and thus approximately minimize the maximum lag-one autocorrelation, as discussed in Section 2.

Logically, one may wonder about putting a hyper-working prior on the hyper-parameter ω, instead of optimizing over ω—indeed, conditional augmentation is a special case of marginal augmentation with a point-mass prior. Although it is clear that one has to stop at some point, another reason for not marginalizing at level two (i.e., averaging over ω) is that it is not guaranteed to be beneficial since $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \omega)|Y_{\text{obs}}, \omega]$ will generally depend on ω. This is in contrast to level-one, where $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ is invariant to α when we follow the marginalization strategy. The importance of this invariance is seen from (3.4), where $\max_\alpha \text{E}[\text{var}(h(\theta)|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha]$ can be larger than $\text{E}[\text{var}(h(\theta)|Y_{\text{aug}})|Y_{\text{obs}}]$ when the invariance fails, in which case conditional augmentation may outperform marginal augmentation.

An illustration of this possibility is displayed in Figure 1 using the common t model. Algorithms using data augmentation to fit t models typically use the well known decomposition, $t = \mu + \sigma Z/\sqrt{q}$, where $Z \sim \text{N}(0, 1)$ and $q \sim \chi^2_\nu/\nu$ with Z and q independent. The data augmentation scheme employed to produce Figure 1, with $\nu = 1$, introduces a working parameter into this decomposition,

$$y_i|q_i \sim \text{N}\left(\mu, \frac{\sigma^{2(1-\alpha)}}{q_i}\right) \quad \text{and} \quad q_i \sim \frac{\sigma^{-2\alpha}\chi^2_\nu}{\nu} \quad \text{for} \quad i = 1, \ldots, 100. \tag{5.6}$$

This working parameter was introduced by Meng and van Dyk (1997) to implement the EM

algorithm for the t model, and they showed that $\alpha = 1/(1+\nu)$ is optimal and leads to a faster EM implementation than the standard implementation which corresponds to $\alpha = 0$.

Since no simple conjugate prior exists for α (with respect to the augmented-data log-likelihood), we used $\alpha \sim \text{Uniform}\left((1-\omega)/2, (1+\omega)/2\right)$. Here ω is the length of the interval and $\alpha = 1/2$ is chosen as the prior mean since it satisfies the EM criterion (when $\nu = 1$) for conditional augmentation and (approximately) minimizes the geometric rate of the corresponding DA algorithm (see Meng and van Dyk 1999). Figure 1, which was obtained via simulation, indicates that the optimal value of ω is zero, and thus there is no gain in averaging over α, at least within the class of uniform priors we considered.

The foregoing comparisons assume a fixed augmentation scheme with the same working parameter. A more difficult comparison is between different augmented-data schemes with working parameters of different forms. For example, in the t model, an alternative working parameter formulation of the augmented-data model is (Liu, Rubin, and Wu 1998)

$$y_i|q_i \sim \text{N}\left(\mu, \frac{\alpha\sigma^2}{q_i}\right) \quad \text{and} \quad q_i \sim \frac{\alpha\chi^2_\nu}{\nu} \quad \text{for} \quad i = 1, \ldots, n. \tag{5.7}$$

Empirical comparisons suggest that conditional augmentation with $\alpha = 1/(\nu+1)$ in model (5.6) has the same rate of convergence as the marginal augmentation scheme from (5.7) using an improper prior on α, $p(\alpha) \propto 1/\alpha$ (see Meng and van Dyk 1999 for details). Although currently we do not have a theoretical proof (or disproof) of this equivalence, the Δ quantities defined previously (i.e., (5.2) and (5.3)) are useful, because they allow comparisons between the improvements resulting from different working parameter formulations that share a common special case. For example, the common augmentation scheme for the t model (Rubin 1983) corresponds to $\alpha = 0$ in (5.6) and $\alpha = 1$ in (5.7). In Section 9.1, we show that for slice sampling the very successful marginalization strategy via affine transformations used in the next three sections turns out to be useless, whereas the conditional augmentation approach using the power transformation given in (2.6) [e.g., Higdon's (1998) partial decoupling method] is quite fruitful. We also emphasize that the Δ quantities can be very useful for providing insight into when (e.g., as a function of the fitted model parameters) the improvement over the standard algorithms is likely to be substantial, as demonstrated in Section 7.

6. APPLICATION: THE MULTIVARIATE *t* DISTRIBUTION

6.1 DATA AUGMENTATION AND ALGORITHMS

As our first example, we consider the multivariate version of the t model introduced in Section 5. As a generalization of the marginal augmentation scheme (5.7) we write

$$Y = \mu + \frac{\sqrt{\alpha}\Sigma^{\frac{1}{2}}Z}{\sqrt{q}}, \quad Z \sim \text{N}_d(0, I), \quad q \sim \alpha\chi^2_\nu/\nu, \quad Z \perp q, \tag{6.1}$$

and would like to draw from the posterior $p(\theta|Y_{\text{obs}})$ where $\theta = (\mu, \Sigma)$, $Y_{\text{obs}} = \{Y_i, i = 1, \ldots, n\}$, $Y_{\text{aug}} = \{(Y_i, q_i), i = 1, \ldots, n\}$, and the degrees of freedom, ν, is assumed

known. Since $q_i = \alpha\tilde{q}_i$, where $\tilde{q}_i$ corresponds to q_i when $\alpha = 1$, Step 1 of the Marginalization Strategy (p. 8) is accomplished. Consequently, we expect marginal augmentation with a working prior independent of θ to improve the rate of convergence over the corresponding standard augmentation scheme, $\tilde{Y}_{\text{aug}} = \{(Y_i, \tilde{q}_i), i = 1, \ldots, n\}$.

As suggested by the Combined Strategy (p. 11), we choose $p(\alpha)$ to be the conditional conjugate prior for $p(Y_{\text{aug}}|\theta, \alpha)$, namely, $\beta\chi_\gamma^{-2}$, where $\beta > 0, \gamma > 0$ are level-two working parameters and χ_γ^{-2} is an inverse chi square random variable with γ degrees of freedom. Under this proper prior for α and the standard improper prior $p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$, the joint posterior density of θ, α, and $q \equiv \{q_1, \ldots, q_n\}$ is given by

$$p(\theta, \alpha, q|Y_{\text{obs}}, \gamma, \beta) \propto \alpha^{-[\frac{\gamma+n(d+\nu)+2}{2}]} \prod_{i=1}^{n} q_i^{\frac{d+\nu}{2}-1} |\Sigma|^{-\frac{n+d+1}{2}} \times \exp\left\{-\frac{\sum_{i=1}^n q_i[(Y_i-\mu)^\top\Sigma^{-1}(Y_i-\mu)+\nu]+\beta}{2\alpha}\right\}. \quad (6.2)$$

It follows that

$$q_i\Big|\mu, \Sigma, Y_{\text{obs}}, \alpha \sim \frac{\alpha\chi^2_{\nu+d}}{(Y_i-\mu)^\top\Sigma^{-1}(Y_i-\mu)+\nu}, \quad (6.3)$$

independently for $i = 1, \ldots, n$,

$$\mu\Big|\Sigma, Y_{\text{aug}}, \alpha \sim \text{N}_d\left(\hat{\mu}, \frac{\alpha\Sigma}{\sum_{i=1}^n q_i}\right), \quad \text{where} \quad \hat{\mu} = \frac{\sum_{i=1}^n q_i Y_i}{\sum_{i=1}^n q_i}, \quad (6.4)$$

$$\Sigma^{-1}\Big|Y_{\text{aug}}, \alpha \sim \alpha\text{Wishart}_{n-1}\left[\left(\sum_{i=1}^n q_i(Y_i-\hat{\mu})(Y_i-\hat{\mu})^\top\right)^{-1}\right], \quad (6.5)$$

and

$$\alpha\Big|Y_{\text{aug}} \sim \frac{\beta+\nu\sum_{i=1}^n q_i}{\chi^2_{\gamma+n\nu}}, \quad (6.6)$$

where $\text{Wishart}_k(A)$ denotes the Wishart distribution with scale matrix A and k degrees of freedom.

To implement Criterion 5.2 for selecting $\omega = \{\gamma, \beta\}$, we first note that the terms of $\log p(\theta, \alpha|Y_{\text{aug}}, \omega)$ involving $\{\theta, \alpha, \omega\}$ are linear in the missing data $q = (q_1, \ldots, q_n)$; see (6.2). Thus, $\tilde{I}_{\text{aug}}(\omega)$ of (5.5) can be computed by first calculating second derivatives of $\log p(\theta, \alpha|Y_{\text{aug}}, \omega)$ as a function of $\{\theta, \alpha\}$, replacing q_i with

$$q_i^*(\theta) \equiv \text{E}(q_i|Y_i, \theta, \alpha) = \alpha\text{E}(\tilde{q}_i|Y_i, \theta) = \frac{\alpha(d+\nu)}{\nu+(Y_i-\mu)^\top\Sigma^{-1}(Y_i-\mu)}, \quad (6.7)$$

and evaluating the resulting expression at $\theta = \theta^\star$, the observed-data posterior mode of θ, and at $\alpha = \hat{\alpha} = \beta/(\gamma+2)$, the mode of the prior $p(\alpha|\omega)$. (This is actually the general scenario when the augmented-data model is from an exponential family, and q corresponds to appropriate augmented-data sufficient statistics.) In fact, we do not even need to compute

any derivatives with respect to θ, nor do we need the value of $\theta^\star$. This is because, from (6.2),

$$\frac{\partial \log p(\theta, \alpha | Y_{\text{aug}}, \omega)}{\partial \alpha} = -\frac{\gamma + n(d+\nu) + 2}{2\alpha} + \frac{\sum_{i=1}^n q_i[(Y_i - \mu)^\top \Sigma^{-1}(Y_i - \mu) + \nu] + \beta}{2\alpha^2}. \tag{6.8}$$

This implies, together with (6.7), that the $\mathcal{I}_{\theta\alpha}(\omega)$ of (5.5) must be of the form $V/\hat{\alpha}$, where V is a nonzero vector of length $d(d+3)/2$ that is free of ω, and

$$\begin{aligned}\mathcal{I}_{\alpha\alpha}(\omega) &= -\frac{\gamma + n(d+\nu) + 2}{2\hat{\alpha}^2} \\ &\quad + \frac{\sum_{i=1}^n q_i^*(\theta^\star)[(Y_i - \mu^\star)^\top \Sigma^{\star -1}(Y_i - \mu^\star) + \nu] + \beta}{\hat{\alpha}^3} \\ &= \frac{\gamma + n(d+\nu) + 2}{2\hat{\alpha}^2}.\end{aligned}$$

Consequently, $\Delta_{\text{EM}}^{(2)}(\omega)$ of (5.4) is $2VV^\top[\gamma + n(d+\nu) + 2]^{-1}$, which achieves its maximum as $\gamma \downarrow 0$; note that $\Delta_{\text{EM}}^{(2)}(\omega)$ is free of β, and thus Criterion 5.2 suggests that the optimal rate does not depend on β. This result covers the $d = 1$ case treated in Meng and van Dyk (1999), where the optimal algorithm was found by numerically inspecting an autocorrelation as a function of γ.

When $\gamma \downarrow 0$, the prior distribution for α becomes improper: $p(\alpha|\gamma = 0, \beta) \propto \alpha^{-1} \exp(-\frac{\beta}{2\alpha}), \beta \geq 0$. As in Meng and van Dyk (1999), to prove that the choice $\{\gamma = 0, \beta = 0\}$ yields a valid algorithm, we first provide the explicit stochastic mappings under Scheme 1, $(\mu^{(t)}, \Sigma^{(t)}) \to (\mu^{(t+1)}, \Sigma^{(t+1)})$, and under Scheme 2, $(\mu^{(t)}, \Sigma^{(t)}, \alpha^{(t)}) \to (\mu^{(t+1)}, \Sigma^{(t+1)}, \alpha^{(t+1)})$. The mappings are given by the following steps:

Step 1: Make n independent draws of $\chi^2_{d+\nu}$ and denote them by $\{\chi^2_{d+\nu,i}, i = 1, \ldots, n\}$. And independently, draw $\chi^2_{n\nu}$, χ^2_γ, $Z \sim \text{N}_d(0, I)$, and $W \sim \text{Wishart}_{n-1}(I)$. For Scheme 1, also independently draw another χ^2_γ, denoted by $\tilde{\chi}^2_\gamma$.

Step 2: Set

$$\tilde{q}_i = \frac{\chi^2_{d+\nu,i}}{\nu + (Y_i - \mu^{(t)})^\top [\Sigma^{(t)}]^{-1}(Y_i - \mu^{(t)})}, \quad \text{for} \quad i = 1, \ldots, n,$$

$$B = \text{Chol}\left(\sum_{i=1}^n \tilde{q}_i (Y_i - \hat{\mu}^{(t+1)})(Y_i - \hat{\mu}^{(t+1)})^\top\right) \quad \text{with} \quad \hat{\mu}^{(t+1)} = \frac{\sum_{i=1}^n \tilde{q}_i Y_i}{\sum_{i=1}^n \tilde{q}_i},$$

and

$$\mu^{(t+1)} = \hat{\mu}^{(t+1)} + \frac{1}{\sqrt{\sum_{i=1}^n \tilde{q}_i}} \text{Chol}\left(BW^{-1}B^\top\right) Z,$$

where $\text{Chol}(A)$ represents the lower triangular matrix in the Choleski decomposition of A (one can also use any other appropriate decomposition).

Step 3: For Scheme 1, compute

$$\Sigma^{(t+1)} = \frac{\chi^2_{n\nu} + \chi^2_\gamma}{\tilde{\chi}^2_\gamma + \nu \sum_{i=1}^n \tilde{q}_i} BW^{-1}B^\top.$$

For Scheme 2, compute

$$\alpha^{(t+1)} = \frac{\beta + \nu\alpha^{(t)} \sum_{i=1}^n \tilde{q}_i}{\chi^2_{n\nu} + \chi^2_\gamma} \quad \text{and} \quad \Sigma^{(t+1)} = \frac{\chi^2_{n\nu} + \chi^2_\gamma}{\beta/\alpha^{(t)} + \nu \sum_{i=1}^n \tilde{q}_i} BW^{-1}B^\top.$$

Since χ^2_γ becomes a point mass at zero when $\gamma \to 0$, it is clear from Step 3 that the transition kernel under Scheme 2 with the choice $\beta = \gamma = 0$ is the limit of the corresponding kernels under Scheme 1 with $\gamma \to 0$ (and with any fixed $\beta > 0$), and the limiting mapping is given by

$$\Sigma^{(t+1)} = \frac{\chi^2_{n\nu}}{\nu \sum_{i=1}^n \tilde{q}_i} BW^{-1}B^\top.$$

Since the mapping in Step 2 for μ is invariant to the choice of either β or γ, we have verified the limiting condition of Theorem 2 of Meng and van Dyk (1999) and thus we know that the subchain $\{\mu^{(t)}, \Sigma^{(t)}; t \geq 1\}$ induced by Scheme 2 with the choice $\beta = \gamma = 0$ will converge in distribution to the desired posterior distribution.

The validity and optimality of the choice $\gamma = \beta = 0$ is also confirmed by Liu and Wu's (1999) group-theoretic results because $p(\alpha) \propto \alpha^{-1}$ is the Haar measure for the scale group. This agreement illustrates the effectiveness of Criterion 5.2. It is also interesting to note that Criterion 5.2 suggests more than the group theoretic optimality results of Liu and Wu (1999), which do not cover the class of priors $p(\alpha|\beta) \propto \alpha^{-1} \exp(-\beta/2\alpha)$ with $\beta > 0$ because they are neither proper nor invariant. In fact, under this class of priors, the subchain $\{\theta^{(t)}, t \geq 0\}$ is not even Markovian because $\Sigma^{(t+1)}$ depends on $\alpha^{(t)}$, as indicated in Step 3. Nevertheless, $\{\theta^{(t)}, t \geq 0\}$ has the correct limiting distribution and has the same optimal convergence rate as the chain generated with the Haar measure (see Meng and van Dyk 1999).

6.2 COMPUTATIONAL PERFORMANCE

The standard and the optimal (marginal augmentation) algorithms (i.e., with $\beta = \gamma = 0$) were applied to a simulated dataset with $n = 100$ observations from a ten dimensional t distribution, $t_{10}(0, I_{10}, \nu = 1)$. With each algorithm three chains were run, each with one of three starting values: $(\mu^{(0)}, \Sigma^{(0)}) = (0, I_{10}), (10, 100I_{10})$, and $(-10, I_{10}/1000)$. Figure 2 compares, for all 65 model parameters, the lag one autocorrelation, lag two autocorrelation, and the minimum k to obtain a lag-k autocorrelation less than .05. The computations are based on 2,000 draws (from one chain) after discarding the first 1,000 draws for both algorithms. The symbols in Figure 2 distinguish between mean, standard deviation, and correlation parameters, and it is evident that the optimal algorithm substantially reduces the autocorrelations for the standard deviation parameters while maintaining them for the other two groups of parameters. This effect is not too surprising given that the working parameter is a rescaling parameter, though it is not clear how general this phenomenon

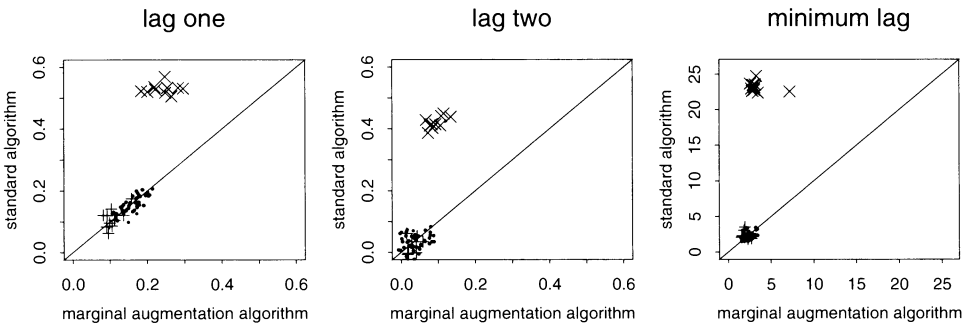

Figure 2. Comparing the Improvement for the 65 Model Parameters in the Multivariate t Example. The three plots compare the lag one autocorrelation, lag two autocorrelation, and minimum k such that the lag-k autocorrelation is less than .05 for each of the 65 model parameters. The symbols '+', '×', and '•' represent mean, standard deviation, and correlation parameters, respectively. In the last plot, the symbols are plotted with Unif (−.5, .5) jittering. The figure emphasizes the reduction in autocorrelation for the ten standard deviation parameters.

is; namely, that rescaling working parameters have a direct effect only on the scale model parameters. However, this does not imply that our optimal algorithm only improves the convergence of the standard deviation (or variance) parameters, because these three groups of parameters are not a posteriori independent, and the overall algorithm converges only if each component does so.

To illustrate the improvement at a more detailed level, Figure 3 shows several graphical summaries of the draws of the first diagonal element of Σ. The columns in the figure correspond to the two algorithms and the rows, from the top to bottom, contain an autocorrelation plot, a time series plot, a lag-one scatterplot, and Gelman and Rubin's (1992) $\sqrt{\hat{R}}$ statistic as a function of iteration (computed on the log of the first diagonal element of Σ). The $\sqrt{\hat{R}}$ statistic is a measure of the between-chain variance relative to the within-chain variance and values close to one indicate acceptable mixing (when the starting values are over dispersed relative to the target distribution). Judging from the various plots, the optimal algorithm is a substantial improvement over the standard algorithm. In particular, we see that $\sqrt{\hat{R}}$ stabilizes near one and the autocorrelation function dies out much more rapidly.

The convergence results described here and in the remaining examples are in terms of the number of iterations required for convergence [although the global rate is sometimes a poor predictor of the actual number of iterations required for convergence; see van Dyk and Meng (1997)]. For a fair comparison, we need to consider the computational burden required by each iteration and it is clear that the marginal DA algorithm is somewhat more costly per iteration simply because it samples more variables at each iteration. For the current problem, the additional cost is essentially zero (e.g., an additional χ^2 variable needed by (6.6)). In general, with sensible choices of the working parameter, the additional computational load required by marginal augmentation algorithms is a very small premium for the substantial improvement in reliability and efficiency of the resulting chains. Such improvements, as seen in Figures 2–3, are even more pronounced in the next two applications.

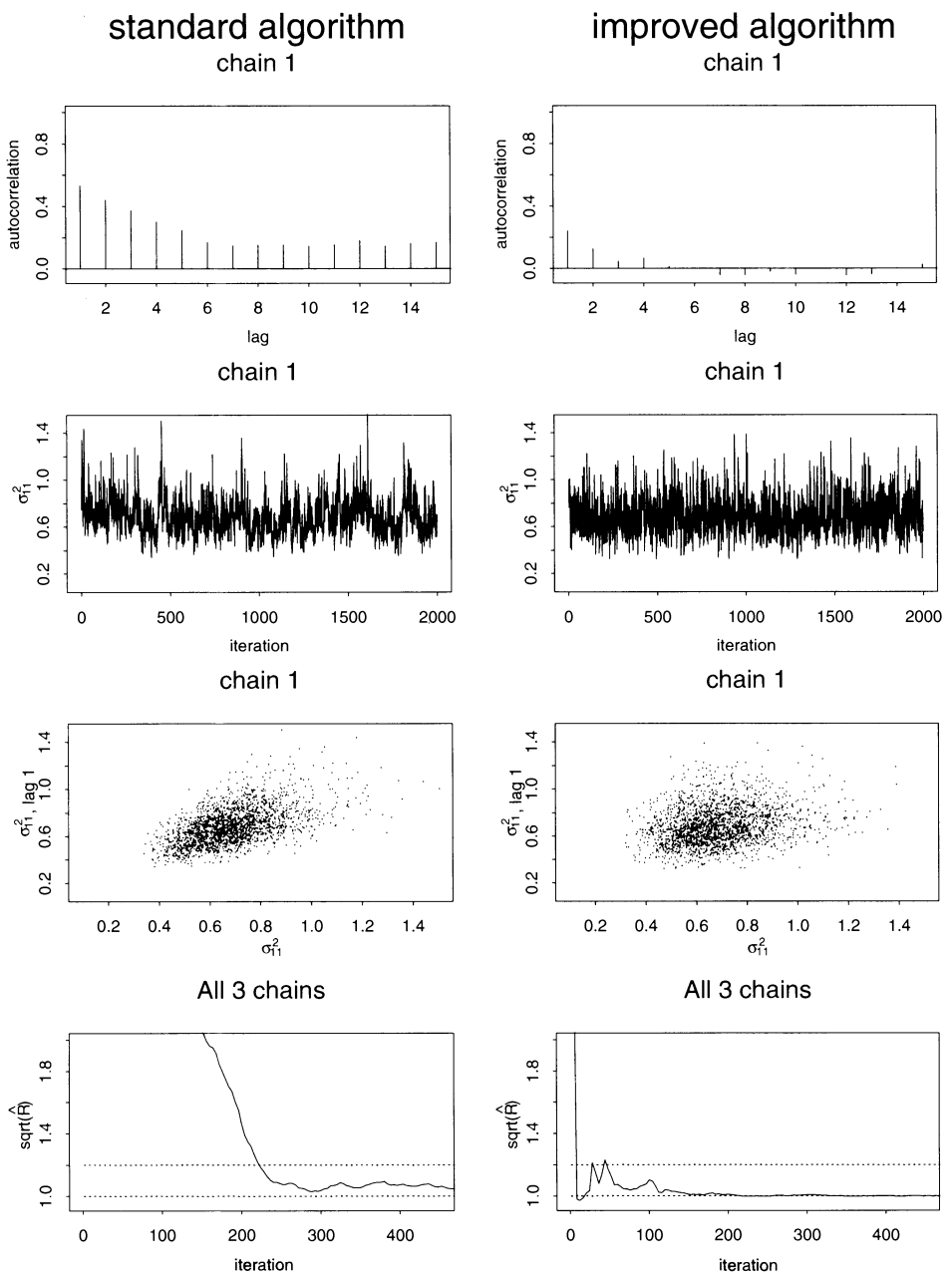

Figure 3. Convergence of Posterior Sampling Algorithms for Fitting a Multivariate t Model. The columns of the figure correspond to the standard algorithm and the optimal algorithm, respectively. The rows of the figure illustrate the autocorrelation plot, a time-series plot, a lag-one scatterplot, and Gelman and Rubin's (1992) $\sqrt{\hat{R}}$ statistic as a function of iteration, all computed using the first diagonal element of Σ. (A log transformation was used to compute $\sqrt{\hat{R}}$.) The dashed lines in the final row correspond to the convergence value of 1 and a threshold of 1.2. Note that the autocorrelation dies out and the chains mix more quickly with the improved algorithm.

7. APPLICATION: PROBIT REGRESSION

7.1 DATA AUGMENTATION AND ALGORITHMS

As a second application, we consider a probit regression model that we formalize by assuming we observe n independent Bernoulli random variables, $Y_i \sim \text{Ber}(\Phi(x_i^\top \beta))$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function, x_i a $p \times 1$ vector of observed covariates, and β a $p \times 1$ parameter. Here we code successes of the Bernoulli random variable as 1 and failures as -1. We assume the standard non-informative prior $p(\beta) \propto 1$ and denote the common data augmentation scheme as $\tilde{Y}_{\text{aug}} = \{(Y_i, \psi_i), i = 1, \ldots, n\}$, where $\psi_i \sim \text{N}(x_i^\top \beta, 1)$ is a latent variable of which we only observe its sign Y_i (see, e.g., Albert 1992; Albert and Chib 1993; McCulloch and Rossi 1994; and Meng and Schilling 1996). The complete conditional distributions for the corresponding Gibbs sampler are given by

$$\beta|\tilde{Y}_{\text{aug}} \sim \text{N}(\tilde{\beta}, (X^\top X)^{-1}) \quad \text{with} \quad \tilde{\beta} = (X^\top X)^{-1} X^\top \psi,$$

where X is an $n \times p$ matrix with ith row equal to $x_i^\top$ and $\psi = (\psi_1, \ldots, \psi_n)^\top$, and by

$$\psi_i|\beta, Y_i \overset{\text{indep}}{\sim} \text{TN}(x_i^\top \beta, 1, Y_i), \quad \text{for} \quad i = 1, \ldots, n,$$

where $\text{TN}(\mu, \sigma^2, Y_i)$ specifies a normal distribution with mean μ and variance σ^2 truncated to be positive if $Y_i = 1$ and negative if $Y_i = -1$. This algorithm was studied in detail in Albert and Chib (1993), and we will label it Albert and Chib's algorithm in our comparison, as suggested by reviewers.

Liu, Rubin, and Wu (1998) identified the variance in the prior (and posterior) distribution for ψ_i as a candidate working parameter in this model. To formalize this, we define a family of data augmentation schemes to be $Y_{\text{aug}} = \{(Y_i, \xi_i), i = 1, \ldots, n\}$, where $\xi \equiv (\xi_1, \ldots, \xi_n)^\top = \mathcal{D}_\sigma(\psi) = (\sigma\psi_1, \ldots, \sigma\psi_n)^\top$. A class of conditional conjugate priors for σ under the augmented-data model is $\sigma^2 \sim \nu_0 s_0^2 / \chi^2_{\nu_0}$, where $\{\nu_0, s_0^2\}$ are the level-two working parameters to be determined. The resulting complete conditional distributions are given by

$$\beta|\sigma^2, Y_{\text{aug}} \sim \text{N}\left(\frac{\hat{\beta}}{\sigma}, (X^\top X)^{-1}\right) \quad \text{with} \quad \hat{\beta} = (X^\top X)^{-1} X^\top \xi, \tag{7.1}$$

$$\sigma^2|Y_{\text{aug}} \sim \frac{(n-p)s^2 + \nu_0 s_0^2}{\chi^2_{n+\nu_0}} \quad \text{with} \quad s^2 = \frac{1}{n-p} \sum_{i=1}^{n} (\xi_i - x_i^\top \hat{\beta})^2, \tag{7.2}$$

and

$$\xi_i|\beta, \sigma^2, Y_i \overset{\text{indep}}{\sim} \text{TN}(x_i^\top(\sigma\beta), \sigma^2, Y_i), \quad \text{for} \quad i = 1, \ldots, n. \tag{7.3}$$

Criterion 5.2 suggests that the optimal algorithm is obtained when we set $\nu_0 = 0$, because $\tilde{I}_{\text{aug}}(\omega)$ of (5.5) with $\theta = \beta$, $\alpha = \sigma^2$, and $\omega = (\nu_0, s_0^2)$ is

$$\tilde{I}_{\text{aug}}(\omega) = \begin{pmatrix} X^\top X & \frac{\nu_0+2}{2s_0^2\nu_0} X^\top X \beta^\star \\ \frac{\nu_0+2}{2s_0^2\nu_0} (X^\top X \beta^\star)^\top & \frac{(\nu_0+2)^2}{4s_0^4\nu_0^2} [2(2+n+\nu_0) + (X\beta^\star)^\top X\beta^\star] \end{pmatrix},$$

where $\beta^\star$ is the observed-data posterior mode of β. Thus,

$$\Delta^{(2)}_{\mathrm{EM}}(\omega) = \mathcal{I}_{\theta\alpha}(\omega)\mathcal{I}^{-1}_{\alpha\alpha}(\omega)\mathcal{I}^{\top}_{\theta\alpha}(\omega) = \frac{X^\top X\beta^\star(X^\top X\beta^\star)^\top}{2(2+n+\nu_0)+(X\beta^\star)^\top X\beta^\star}, \tag{7.4}$$

which is free of s_0^2 and maximized on the boundary of the parameter space as $\nu_0 \to 0$. This leads to an improper prior given by $p(\sigma^2) \propto \sigma^{-2}$.

As with the t application in Section 6, to verify this improper working prior will yield a properly converging subchain for β, we explicitly express the stochastic mappings for Scheme 1, $\beta^{(t)} \to \beta^{(t+1)}$, and Scheme 2, $([\sigma^2]^{(t)}, \beta^{(t)}) \to ([\sigma^2]^{(t+1)}, \beta^{(t+1)})$, under the working prior $\sigma^2 \sim \nu_0/\chi^2_{\nu_0}$ (i.e., we have set $s_0^2 = 1$). The mappings are given by the following steps:

Step 1: Draw independently $\psi_i^{(t+1)} \sim \mathrm{TN}(x_i^\top\beta^{(t)}, 1, Y_i)$ and denote $\psi^{(t+1)} = (\psi_1^{(t+1)}, \ldots, \psi_n^{(t+1)})^\top$; also draw independently χ^2_n, $\chi^2_{\nu_0}$, and $Z \sim \mathrm{N}_p(0, I)$; let $[\tilde\sigma^2]^{(t+1)} \equiv \nu_0\chi^{-2}_{\nu_0}$.

Step 2: For Scheme 1, set

$$[\hat\sigma^2]^{(t+1)} = \frac{\nu_0 + [\tilde\sigma^2]^{(t+1)}R^{(t+1)}}{(\chi^2_n + \chi^2_{\nu_0})}, \tag{7.5}$$

where $R^{(t+1)} = \sum_{i=1}^n(\psi_i^{(t+1)} - x_i^\top\tilde\beta^{(t+1)})^2$ with $\tilde\beta^{(t+1)} = (X^\top X)^{-1}X^\top\psi^{(t+1)}$.

For Scheme 2, set

$$[\sigma^2]^{(t+1)} = \frac{\nu_0 + [\sigma^2]^{(t)}R^{(t+1)}}{(\chi^2_n + \chi^2_{\nu_0})}. \tag{7.6}$$

Step 3: For Scheme 1, set

$$\beta^{(t+1)} = \frac{\tilde\sigma^{(t+1)}}{\hat\sigma^{(t+1)}}\tilde\beta^{(t+1)} + \mathrm{Chol}[(X^\top X)^{-1}]Z. \tag{7.7}$$

For Scheme 2, set

$$\beta^{(t+1)} = \frac{\sigma^{(t)}}{\sigma^{(t+1)}}\tilde\beta^{(t+1)} + \mathrm{Chol}[(X^\top X)^{-1}]Z. \tag{7.8}$$

Note that the quantities $[\tilde\sigma^2]^{(t+1)}$ and $[\hat\sigma^2]^{(t+1)}$ are not part of the Markov chain under Scheme 1 since Scheme 1 induces a marginal chain for β. These intermediate quantities are introduced to facilitate sampling under marginal augmentation.

Noting that when $\nu_0 \to 0$, $[\tilde\sigma^2]^{(t+1)} \equiv \nu_0\chi^{-2}_{\nu_0}$ becomes a point mass at 1, we see from (7.5)–(7.8) that the transition kernel under Scheme 2 with $\nu_0 = 0$ is the limit of the corresponding kernels under Scheme 1 as $\nu_0 \to 0$. This limiting mapping is given by

$$\beta^{(t+1)} = \sqrt{\frac{\chi^2_n}{R^{(t+1)}}}\tilde\beta^{(t+1)} + \mathrm{Chol}[(X^\top X)^{-1}]Z, \tag{7.9}$$

Table 1. The Latent Membranous Lupus Nephritis Dataset. The table records the number of latent membranous lupus nephritis cases (the numerator) and the total number of cases (the denominator) for each combination of the values of the two covariates.

	IgA				
IgG3 − IgG4	0	.5	1	1.5	2
−3.0	0/1	—	—	—	—
−2.5	0/3	—	—	—	—
−2.0	0/7	—	—	—	0/1
−1.5	0/6	0/1	—	—	—
−1.0	0/6	0/1	0/1	—	0/1
− .5	0/4	—	—	1/1	—
.0	0/3	—	0/1	1/1	—
.5	3/4	—	1/1	1/1	1/1
1.0	1/1	—	1/1	1/1	4/4
1.5	1/1	—	—	2/2	—

where $R^{(t+1)}$ and $\tilde{\beta}^{(t+1)}$, as defined earlier, are stochastically determined by $\beta^{(t)}$. Consequently, the Markov chain defined by (7.9) will converge properly with the target posterior $p(\beta|Y_{\text{obs}})$ as its stationary distribution. Section 7.2 provides empirical evidence of the advantage of this chain over Albert and Chib's chain. Additional empirical evidence can be found in Liu and Wu (1999), whose theoretical results again confirm the validity and optimality of the algorithm given by (7.9) because $p(\sigma^2) \propto \sigma^{-2}$ is the Haar measure for the scale group. Analogous to the finding in Section 6, theoretically more general improper priors of the form $p(\sigma^2|s_0^2) \propto \sigma^{-2}\exp(-s_0^2/2\sigma^2)$ also produce the optimal algorithm (7.9) in the limit for any $s_0^2 > 0$. This can be verified by replacing ν_0 with $\nu_0 s_0^2$ as the scale term in the working prior for σ^2.

7.2 Computational Performance

To compare empirically Albert and Chib's algorithm and the optimal algorithm (7.9), we implemented both using a dataset supplied by M. Haas, who was a client of the Statistical Consulting Program at the University of Chicago. Table 1 displays the data with two clinical measurements (i.e., covariates), which are used to predict the occurrence of latent membranous lupus nephritis. The first covariate is the difference between IgG3 and IgG4, and the second covariate is IgA, where IgG and IgA stand for immunoglobulin G and immunoglobulin A, two classes of antibody molecules. The dataset consists of measurements on 55 patients of which 18 have been diagnosed with latent membranous lupus. Haas was interested in regressing the disease indicator on the first covariate alone, as well as on additional covariates [the original dataset contains additional covariates that are not used here; see Haas (1994, 1998) for scientific background].

We consider two models, the first with an intercept and the first covariate, and the second with an intercept and both covariates. Under each model we run both algorithms, each with three different starting values. Figures 4 and 5 give for each model the autocorrelation plots, time series plots, and lag-one scatterplots of β_1 (the coefficient for the covariate common to both models) from one chain, as well as the $\sqrt{\hat{R}}$ statistic for β_1 using all three chains. Looking across the plots, it is clear that the performance of both algorithms degrades when the second covariate is added. The improvement offered by the optimal algorithm, however, is particularly striking under the second model. This

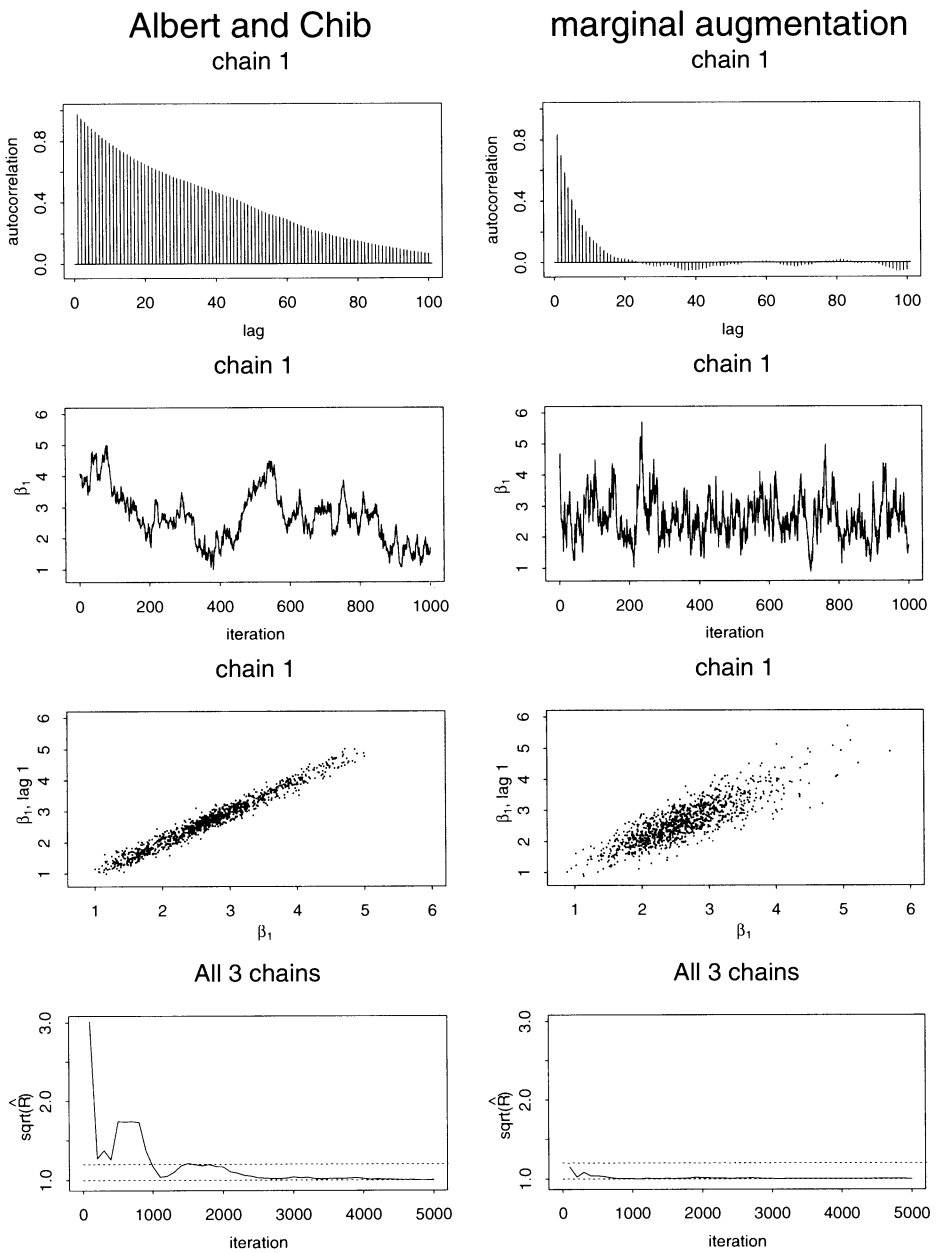

Figure 4. Convergence of Posterior Sampling Algorithms for Fitting a Probit Regression Model with One Covariate. The columns of the figure correspond to the standard algorithm and the optimal algorithm, respectively. The rows are as in Figure 3 with all summaries computed for β_1. The improved chain significantly reduces the autocorrelation and improves the mixing of β_1.

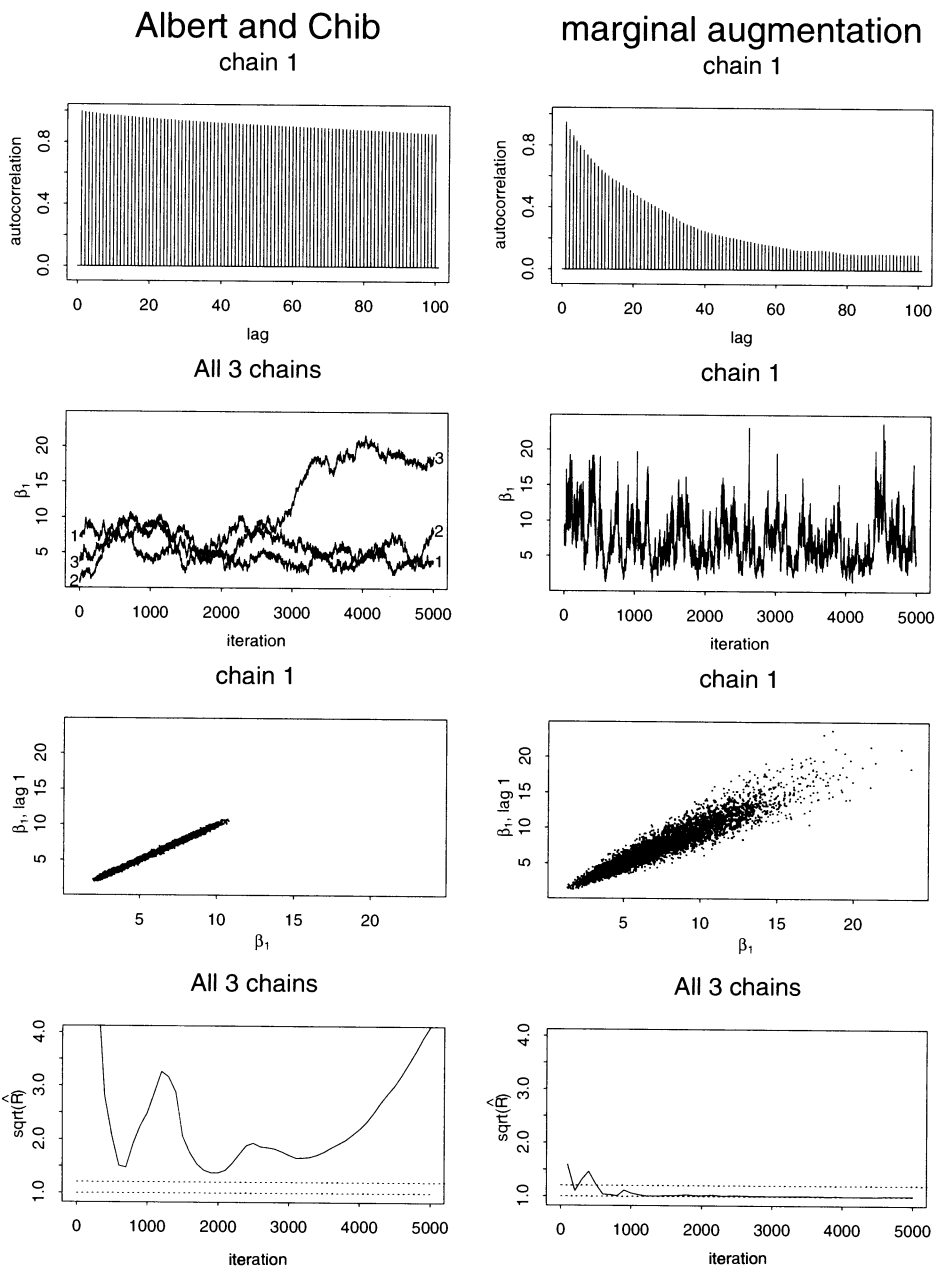

Figure 5. Convergence of Posterior Sampling Algorithms for Fitting a Probit Regression Models with Two Covariates. The figure contains the same summaries as Figure 4. The substantial improvement due to marginalizing out σ^2 is evident in the reduction in autocorrelation in the draws of β_1 and the faster mixing (e.g., as measured by $\sqrt{\hat{R}}$). We plot all three of the chains generated with Albert and Chib's algorithm to explain the increase in $\sqrt{\hat{R}}$ after iteration 3,000.

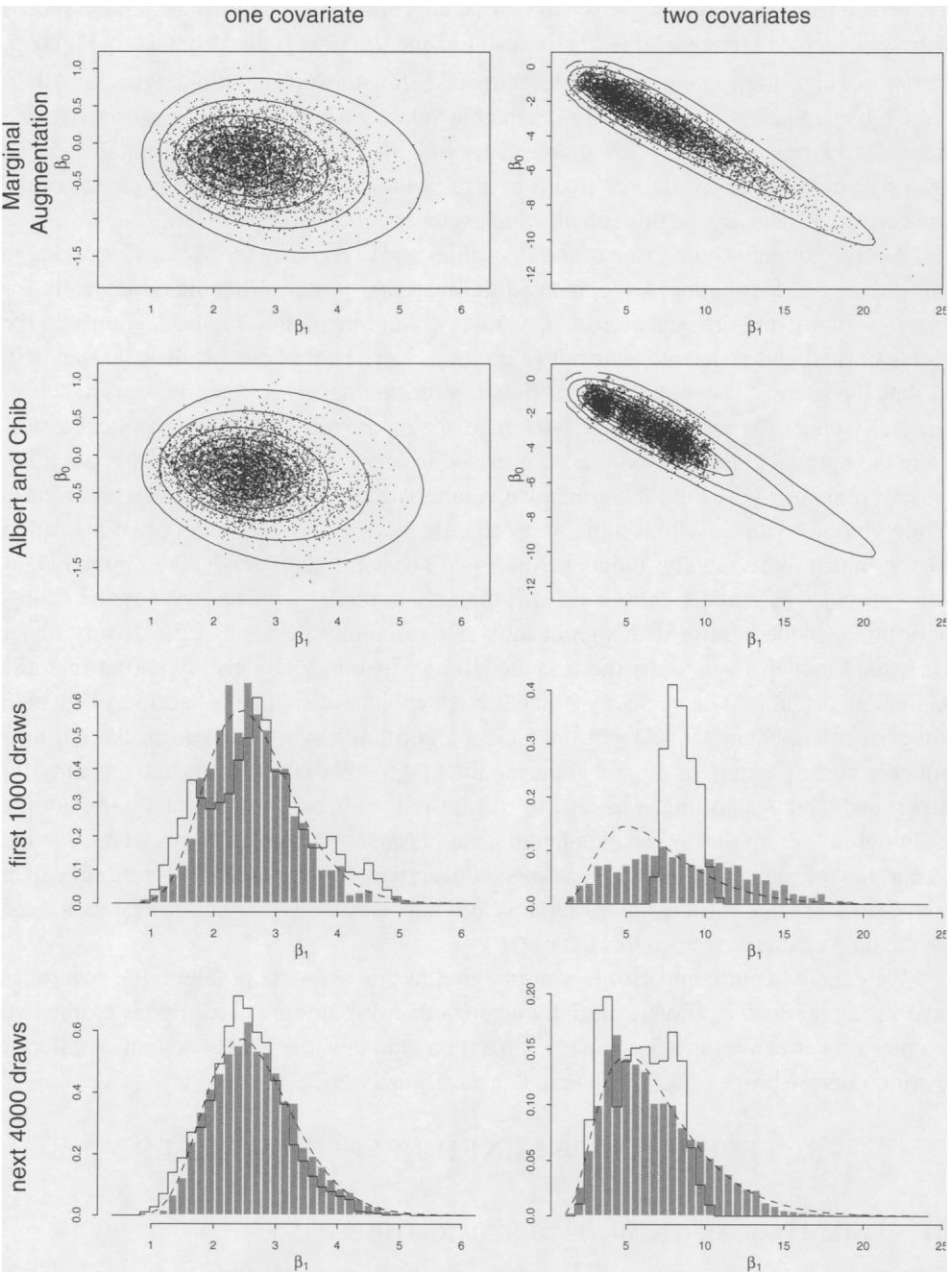

Figure 6. The Reliability of Albert and Chib's Sampler and the Optimal Marginal Augmentation Sampler. This plot compares the Monte Carlo generated draws with the posterior (target) distribution. The columns correspond to the model fit with one and two covariates respectively. The first two rows illustrate the draws from the marginal augmentation and Albert and Chib's sampler, respectively. (The contours represent the target posterior distribution.) In the final two rows, the shaded histograms represent the draws from the marginal augmentation sampler, the solid lines outline the corresponding histograms using the draws from Albert and Chib's sampler, and the dotted lines are the actual marginal posterior densities. The first two rows illustrate 5,000 draws, the third row the first 1,000 draws, and the final row the next 4,000 draws. The benefit of marginal augmentation is most pronounced in the early draws in the first column and is pronounced throughout the second column.

is expected from the Δ quantity given in (7.4), which indicates that the improvement should increase with the magnitude of $X\beta^\star$; see also Liu and Wu (1999). In particular, in Figure 5, the optimal algorithm attains acceptable values of $\sqrt{\hat{R}}$ within 600 iterations, whereas Albert and Chib's algorithm is not close to acceptable values even after 5,000 iterations. To see more clearly the exceedingly slow convergence of Albert and Chib's algorithm, the second row of Figure 5 includes all three chains to display the clearly separated trajectories of the three chains, especially the third chain which wanders off.

As an additional comparison of the algorithms and to compare the Monte Carlo sample with the target distribution, we computed the marginal posterior/likelihood counters for $\{\beta_0, \beta_1\}$ jointly and the posterior of β_1 via numerical integration. Figure 6 compares the Monte Carlo draws generated from both algorithms using chain 1 against the gold standard, the actual posterior. The improvement offered by the optimal algorithm is now crystal clear, especially under the second model. (All three chains provide essentially identical results using the optimal algorithm, but the comparison is even more dramatic under chain 3 because it remains in the tail longer than we would like with Albert and Chib's algorithm.) While both algorithms will eventually provide the correct answer, the optimal algorithm does so much more rapidly. Judging from Figure 6, with 5,000 draws, the optimal algorithm provided essentially correct answers under both models, whereas Albert and Chib's algorithm provided draws with significantly less variability than the target density under the second model. Even under the first model, its performance is visibly poorer than the optimal algorithm. As seen clearly in the second column of Figure 6, because of its high autocorrelation, chain 1 of Albert and Chib's algorithm has not traveled to the tail area (roughly corresponding to $\beta_1 \geq 11$) during the first 5,000 draws. Incidentally, chain 3 of Albert and Chib's algorithm was able to travel to the tail area by iteration 3,000, but the chain did not return during the simulation, again because of high autocorrelation; see the second row of Figure 5. Neither phenomenon occurred with the optimal algorithm, which used the same three starting values (and we did not selectively present our results—these are the only chains we ran on this dataset).

We conclude our comparison by noting that in this real-data problem, the covariates have strong predictive power, especially when both covariates are used. However, the two samples do not exhibit quasi-complete separation, and thus the MLEs exist and posterior distributions are proper (see Speckman, Lee, and Sun 1999).

8. APPLICATION: MIXED-EFFECTS MODELS

8.1 TWO MARGINAL AUGMENTATION ALGORITHMS

In this section we use the marginal augmentation approach to derive two new algorithms for posterior sampling under the popular mixed-effect models (with the standard diffuse prior), which belong to the class of Gaussian hierarchical models. In the next two sections, we provide both theoretical evidence (Section 8.2) and empirical evidence (Section 8.3) to demonstrate that the new algorithms can provide substantial improvement over several existing algorithms, all of which can be derived from a common conditional augmentation scheme.

We consider a mixed-effects model of the following general form

$$y_i = X_i\beta + Z_i\xi + Z_i b_i + e_i, \quad b_i \sim \mathrm{N}_q(0, T), \quad e_i \sim \mathrm{N}(0, \sigma^2 I_{n_i}), \quad b_i \perp e_i, \tag{8.1}$$

for $i = 1, \ldots, m$, where the observed response y_i is $n_i \times 1$, X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ are known covariates (throughout we assume that the Z_i are such that T is identifiable), b_i $(q \times 1)$ is the random effect, β $(p \times 1)$ and ξ $(q \times 1)$ are fixed effects, and $\theta = (\beta, \xi, \sigma^2, T)$. Model (8.1) can be modified to accommodate a more general variance-covariance matrix known up to a scalar, say $\sigma^2\Xi_i$, for the e_i, by premultiplying by the Choleski factor of Ξ_i.

We can fit model (8.1) in the ML or empirical Bayesian paradigm with a myriad of EM-type algorithms (e.g., Dempster, Laird, and Rubin 1977; Laird and Ware 1982; Laird, Lange, and Stram 1987; Liu and Rubin 1994; Meng and van Dyk 1998; van Dyk 2000), all of which are based on the idea of data augmentation; some of these EM-type algorithms have been translated (and extended) into Gibbs samplers for full posterior inference (e.g., Lange, Carlin, and Gelfand 1992; Gelfand, Sahu, and Carlin 1995). The standard and most obvious data augmentation is simply to treat the random effect as missing data; that is, $\tilde{Y}_{\text{aug}} = \{(y_i, X_i, Z_i, b_i), i = 1, \ldots, m\}$. To improve on this scheme, we adopt the Combined Strategy of Section 5 using the following augmentation scheme with working parameter $\alpha = (\gamma, \Upsilon)$:

$$Y_{\text{aug}} = \{(y_i, X_i, Z_i, d_i = \Upsilon^{-1}b_i + \gamma), \quad i = 1, \ldots, m\}, \tag{8.2}$$

where γ is $q \times 1$ and Υ is $q \times q$. This is a generalization of the augmentation scheme introduced by Liu, Rubin, and Wu (1998), which fixed $\gamma = 0$ and assumes y_i is a scalar.

We consider two choices of Υ, which lead to two different algorithms, each with its own advantages. The first one is a full $q \times q$ nonsingular Υ, and the second is a lower triangular nonsingular Υ. For each choice, suggested by the conditional-conjugacy and simple implementation, we choose the working prior to be $\mathrm{vec}(\Upsilon) \sim \mathrm{N}(\mathrm{vec}(I_q), \omega I_{v(\Upsilon)})$ and $\gamma|\Upsilon \sim \mathrm{N}(0, \omega[\Upsilon\Upsilon^\top]^q)$, where ω is a positive scalar parameter serving as the level-two working parameter. Here $\mathrm{vec}(B)$ is a $v(B) \times 1$ vector containing the elements of B arranged in dictionary order with the row index listed first, but it skips elements above the diagonal if B is lower triangular. Thus, $v(B) = q^2$ when $B_{q\times q}$ is a full matrix and $v(B) = q(q+1)/2$ when $B_{q\times q}$ is lower triangular. Note that the definition of $\mathrm{vec}(I_q)$ varies accordingly with the definition of $\mathrm{vec}(\Upsilon)$.

Under (8.2), (8.1) can be re-expressed as (e_i is unchanged)

$$y_i = X_i\beta + Z_i\tilde{\xi} + \tilde{Z}_i(d_i)\mathrm{vec}(\Upsilon) + e_i, \quad d_i \sim \mathrm{N}_q(\gamma, \tilde{T}), \quad d_i \perp e_i, \tag{8.3}$$

where $\tilde{\xi} = \xi - \Upsilon\gamma$, $\tilde{T} = \Upsilon^{-1}T[\Upsilon^{-1}]^\top$, and

$$\tilde{Z}_i(d_i) = \begin{cases} \left(Z_{i1}(E_1 d_i)^\top, \ldots, Z_{iq}(E_q d_i)^\top\right)_{n_i \times [q(q+1)/2]}, & \text{if } \Upsilon \text{ is lower trangular;} \\ \left(Z_{i1}d_i^\top, \ldots, Z_{iq}d_i^\top\right)_{n_i \times q^2}, & \text{if } \Upsilon \text{ is a full matrix;} \end{cases}$$

with $\left(Z_{i1}, \ldots, Z_{iq}\right) = Z_i$ and E_j a $j \times q$ matrix formed by taking the first j rows of the identity matrix I_q, $j = 1, \ldots, q$. The model prior we use is the standard diffuse prior, $p(\beta, \log(\sigma), \xi, T) \propto |T|^{-1/2}$. Van Dyk (2000) verified that this prior results in a proper

posterior if the covariates are full rank, $n = \sum_{i=1}^{m} n_i > p + q^2$, and $m > 2q - 1$. [See Hobert and Casella (1996, 1998) and Sun, Tsutakawa, and He (1997) for related conditions.] As in the previous two applications, Criterion 5.2 suggests a diffuse working prior for α, namely, we let $\omega \to \infty$; see the Appendix for theoretical derivations as well as verification of the conditions of Lemma 1 for justifying the resulting algorithms. Note that the algorithms can also be justified by Liu and Sabatti's (2000) validity result because the optimal prior $p(\gamma, \Upsilon|\omega = \infty) \propto |\Upsilon|^{-q}$ found by the Combined Strategy is also the right Haar measure for the affine-transformation group (i.e., $b \to \Upsilon(b - \gamma)$), with Υ either a full matrix or a lower triangular matrix. However, the optimality result of Liu and Wu (1999) is not applicable here because the affine-transformation group is not a unimodular group (Nachbin 1965, pp. 71–73).

With either choice of Υ, the resulting algorithm for sampling from the desired posterior has the following steps, where t indexes iteration and $\tilde{\theta} = (\beta, \tilde{\xi}, \sigma^2, \tilde{T})$.

Step 1: For $i = 1, \ldots, m$, draw independently

$$b_i \sim \mathrm{N}\left(\hat{b}_i(\theta^{(t)}), \quad \left(T^{(t)} - T^{(t)} Z_i^\top W_i(\theta^{(t)}) Z_i T^{(t)}\right)\right), \tag{8.4}$$

where

$$\hat{b}_i(\theta) = T Z_i^\top W_i(\theta)(y_i - X_i\beta - Z_i\xi) \quad \text{and} \quad W_i(\theta) = \left[\sigma^2 I_{n_i} + Z_i T Z_i^\top\right]^{-1}. \tag{8.5}$$

Step 2: Given the output from Step 1, draw $(\tilde{\theta}, \alpha)$ as

$$[\sigma^2]^{(t+1)} \sim \frac{Y^\top (I_n - \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top) Y}{\chi^2_{n-p-q-v(\Upsilon)}}, \tag{8.6}$$

$$\begin{pmatrix} \beta^{(t+1)} \\ \tilde{\xi} \\ \mathrm{vec}(\Upsilon) \end{pmatrix} \sim \mathrm{N}\left((\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y, \quad [\sigma^2]^{(t+1)} (\tilde{X}^\top \tilde{X})^{-1}\right), \tag{8.7}$$

$$\tilde{T}^{-1} \sim \mathrm{Wishart}_{m-q-1}\left[\left(\sum_{i=1}^{m} (b_i - \mu_\gamma)(b_i - \mu_\gamma)^\top\right)^{-1}\right], \tag{8.8}$$

and

$$\gamma \sim \mathrm{N}\left(\mu_\gamma, \quad \frac{1}{m}\tilde{T}\right), \tag{8.9}$$

where $\mu_\gamma = \sum_{i=1}^{m} b_i/m$,

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \quad \text{and} \quad \tilde{X} = \begin{pmatrix} X_1 & Z_1 & \tilde{Z}_1(b_1) \\ \vdots & \vdots & \vdots \\ X_m & Z_m & \tilde{Z}_m(b_m) \end{pmatrix}.$$

Finally, we set $\xi^{(t+1)} = \tilde{\xi} + \Upsilon\gamma$ and $T^{(t+1)} = \Upsilon\tilde{T}\Upsilon^\top$.

Table 2. Some Conditional Data Augmentation Schemes for Fitting Model (8.1). The table indicates what parameters will be updated using regression (observation level) and the marginal distribution of c_i (system level). Here $L = \Delta U$ (and thus $T = LL^\top$).

Working parameter	*Parameters updated or sampled at*	
α	*observation level*	*system level*
$\alpha_{\text{reg}} = (1_q, 0_q, 1)$	$\beta \quad \xi \quad L \quad \sigma^2$	–
$(0_q, 0_q, 1)$	$\beta \quad \xi \quad \Delta \quad \sigma^2$	U^2
$(0_q, 1_q, 1)$	$\beta \quad \Delta \quad \sigma^2$	$[\Delta^{-1}\xi] \quad U^2$
$\alpha_{\text{std}} = (0_q, 0_q, 0)$	$\beta \quad \xi \quad \sigma^2$	T
$\alpha_{\text{gsc}} = (0_q, 1_q, 0)$	$\beta \quad \sigma^2$	$\xi \quad T$

We note that when Υ is a full matrix, the denominator of (8.6) requires $n > p+q(q+1)$ to implement the algorithm. This requires n to be larger than is necessary for ensuring the propriety of the targeted posterior distribution. In contrast, when Υ is lower triangular, the degrees-of-freedom requirement of (8.6) is that n not be smaller than the number of parameters (i.e., $p + q(q + 3)/2 + 1$), which is a necessary condition for the posterior to be proper, so there is no additional requirement on the data size. However, the advantage of choosing a full matrix working parameter is the improved stability and efficiency of the algorithm, as demonstrated in Section 8.3. For the EM algorithm, the advantages of using full Υ over lower triangular Υ are illustrated in van Dyk (2000) and Foulley and van Dyk (2000).

8.2 COMPARISONS WITH CONDITIONAL AUGMENTATION ALGORITHMS

Several existing and alternative algorithms for posterior sampling under mixed-effect models can all be derived from the following conditional augmentation scheme indexed by a working parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, where $\alpha_1, \alpha_2 \in R^q$ and $\alpha_3 \in \{0, 1\}$,

$$Y_{\text{aug}} = \{(y_i, X_i, Z_i, c_i = \tilde{U}(-\alpha_1)\Delta^{-\alpha_3}(b_i + \alpha_2 * \xi)), i = 1, \ldots, m\}, \tag{8.10}$$

where "$*$" indicates component-wise multiplication, $T = \Delta U^2 \Delta^\top$ with Δ lower triangular with ones on the diagonal and $U \equiv \text{diag}\{u_1, \ldots, u_q\}$, and $\tilde{U}(\alpha_1) \equiv \text{diag}\{u_1^{\alpha_{11}}, \ldots, u_q^{\alpha_{1q}}\}$.

It is clear that the standard augmentation, $\tilde{Y}_{\text{aug}}$ corresponds to conditioning on $\alpha = \alpha_{\text{std}} = (0_q, 0_q, 0)$ in (8.10), where 0_q is the zero vector of length q. The "re-centered" Gibbs sampler presented by Gelfand, Sahu, and Carlin (1995) corresponds to fixing α at $\alpha_{\text{gsc}} = (0_q, 1_q, 0)$. Meng and van Dyk (1998) derived an EM-type algorithm which uses (8.10) with $\alpha = \alpha_{\text{reg}} = (1_q, 0_q, 1)$, and suggested that such a scheme can be useful for a Gibbs sampler implementation as well. Here the subscript "reg" stands for "full regression," because under this choice of α, all the system level parameters are reparameterized into regression parameters, as seen in Table 2. Table 2 also lists some other choices of α which result in complete conditional distributions that are relatively easy to sample from.

In Table 2, the *observation level* (level 1) of the model refers to the sampling distribution of y_i conditional on the unobserved c_i; that is,

$$y_i|\theta, \alpha, c_i \sim \text{N}\left(X_i\beta + Z_i((1_q - \alpha_2) * \xi) + \tilde{Z}_i(c_i)\text{vec}(\Delta^{\alpha_3}\tilde{U}(\alpha_1)), \quad \sigma^2 I_{n_i}\right). \tag{8.11}$$

The distribution of c_i is refereed to as the *system level* (level 2) of the model, and is given by

$$c_i|\theta,\alpha \sim \mathrm{N}_q(\tilde{U}(-\alpha_1)\Delta^{-\alpha_3}(\alpha_2 * \xi),\quad \tilde{U}(-\alpha_1)\Delta^{1-\alpha_3}U^2[\Delta^{1-\alpha_3}]^\top\tilde{U}(-\alpha_1)). \quad (8.12)$$

Since for any value of α in Table 2, each of the parameters falls into either the observation or the system level of the model, drawing from $p(\theta|Y_{\text{aug}},\alpha)$ is typically straightforward; see the Appendix for the actual steps with $\alpha = \alpha_{\text{std}}, \alpha_{\text{gsc}}$, and α_{reg}. Drawing c_i from $p(c_i|Y_{\text{obs}},\theta,\alpha)$ for $i = 1,\ldots,m$ is simple for any value of α, namely, by first drawing b_i according to (8.4)–(8.5), and then adding and premultiplying by the appropriate function of θ (depending on α) to obtain c_i.

The comparisons among the three conditional augmentation algorithms (i.e., with $\alpha = \alpha_{\text{std}}, \alpha_{\text{gsc}}$, and α_{reg}) and with the marginal augmentation algorithm of Section 8.1 using Υ lower triangular can be summarized as follows. Note that we choose Υ to be lower triangular for the theoretical comparisons. The results are even stronger with a full matrix Υ since the larger working parameter necessarily dominates the smaller one in terms of the EM rate of convergence (Liu, Rubin, and Wu 1998), which is our deterministic approximation criterion underlying the following comparisons.

Comparison 1: The marginal augmentation algorithm dominates the conditional augmentation algorithm with any of the values of α given in Table 2.

Comparison 2: The conditional augmentation algorithm with $\alpha = \alpha_{\text{std}}$ (i.e., the standard algorithm) is dominated by either the algorithm using $\alpha = \alpha_{\text{gsc}}$ or the algorithm using $\alpha = \alpha_{\text{reg}}$.

Comparison 3: The conditional augmentation algorithm with $\alpha = \alpha_{\text{gsc}}$ is better than the one with $\alpha = \alpha_{\text{reg}}$ when the coefficients of determination (as defined in the Appendix, (A.8)) are large. When the coefficients of determination are small, the algorithm using $\alpha = \alpha_{\text{reg}}$ is better.

The precise conditions and statements for these comparisons are given in Results A.1–A.3 in the Appendix; empirical demonstrations are given in the next section. These comparisons suggest that the marginal augmentation algorithms, especially the one with full Υ, should be preferred in general practice, while the standard algorithm should be avoided whenever possible. The slight disadvantage of the marginal augmentation algorithms of Section 8.1, especially when q is large, is its requirement of drawing γ and Υ. This is particularly a problem when using a full matrix working parameter. Thus, it may occasionally be more efficient to start with one of the simpler algorithms, perhaps using the conditional augmentation algorithm with $\alpha = \alpha_{\text{reg}}$ or α_{gsc}. Based on the initial iterates we can then determine if the extra programming effort required by the more efficient algorithm is necessary. The results in the Appendix help us to see when the extra effort will be fruitful (e.g., when the posterior mode $T^\star$ is singular or nearly singular, or when the posterior mode $[\sigma^2]^\star$ is very small). The simpler algorithms could also be used to debug the more sophisticated one, since they require a subset of the computations but still sample from $p(\theta|Y_{\text{obs}})$.

8.3 COMPUTATIONAL PERFORMANCE

Our empirical investigation uses one real and two artificial datasets. We begin with a dataset described by Pierson and Ginther (1987) which consists of the number of ovarian follicles greater than 10mm in diameter recorded daily for each of 11 mares. For this dataset, $m = 11, n_i$ varies from 25 to 31, and $n = 308$. As suggested by Lindstrom and Bates (1988), we fit the model

$$y_{ij} = (\xi_1 + b_{i1}) + \sin(2\pi z_{ij})(\xi_2 + b_{i2}) + \cos(2\pi z_{ij})(\xi_3 + b_{i3}) + e_{ij},$$

where y_{ij} is the jth measurement on the ith mare, z_{ij} records the time of this measurement scaled so that 0 and 1 correspond to ovulation, $b_i = (b_{i1}, b_{i2}, b_{i3})^\top \sim \mathrm{N}(0, T)$ and $e_i = (e_{i1}, \ldots, e_{in_i})^\top \sim \mathrm{N}(0, \sigma^2 R_i)$ with the (u, v) element of R_i given by $e^{-\rho|u-v|}$. For our purposes, we fit ρ via a REML estimate and conditioned on this value in the Gibbs samplers. Figure 7 illustrates time series plots of the first 1,000 draws of ξ_1 and $\log(T_{11})$ (i.e., $\log(\mathrm{var}(b_{i1}|T))$) using each of the five samplers being compared; that is, condition augmentation using α_{std}, α_{reg}, and α_{gsc} and marginal augmentation with both a lower triangular and a full matrix working parameter Υ. The first column shows the draws of ξ_1 and illustrates the gross autocorrelation in the standard conditional augmentation and full regression samplers. It is clear that the sampler using α_{gsc} and the marginal augmentation samplers behave much better than the other two (for this dataset, the coefficients of determination are large). The second column of Figure 7 represents $\log(T_{11})$; although the behavior is much better for $\log(T_{11})$ than for ξ_1 when $\alpha = \alpha_{\mathrm{std}}$, the sampler using α_{reg} is unacceptable for either parameter.

The last two rows of Figure 7 illustrate the low autocorrelation in the draws from the two marginal augmentation samplers. To compare these two algorithms more carefully, Figure 8 shows the lag-one autocorrelation of the ten scalar parameters in level 1 variance σ^2 (one parameter), level 2 variance T (six parameters), and level 1 means ξ (three parameters), some with transformations (e.g., log of variance). For comparison, we include conditional augmentation with $\alpha = \alpha_{\mathrm{std}}$ (open triangles), $\alpha = \alpha_{\mathrm{reg}}$ (open squares), and $\alpha = \alpha_{\mathrm{gsc}}$ (open diamonds) along with marginal augmentation with Υ lower triangular (filled squares) and a full matrix (filled triangles). The marginal augmentation algorithm with a lower triangular Υ is dominated by the algorithm with full matrix Υ, which results in nearly "perfect simulation," as the lag-one autocorrelations of all the parameters are essentially zero.

To further investigate the performance of these samplers, two artificial datasets were generated with 100 observations from the model

$$y_i = (\xi_1 + b_{i1}) + z_i(\xi_2 + b_{i2}) + e_i, \tag{8.13}$$

where y_i is $n_i \times 1$, $b_i = (b_{i1}, b_{i2})^\top \sim \mathrm{N}(0, T)$, $e_i \sim \mathrm{N}(0, \sigma^2 I_{n_i})$, and b_i and e_i

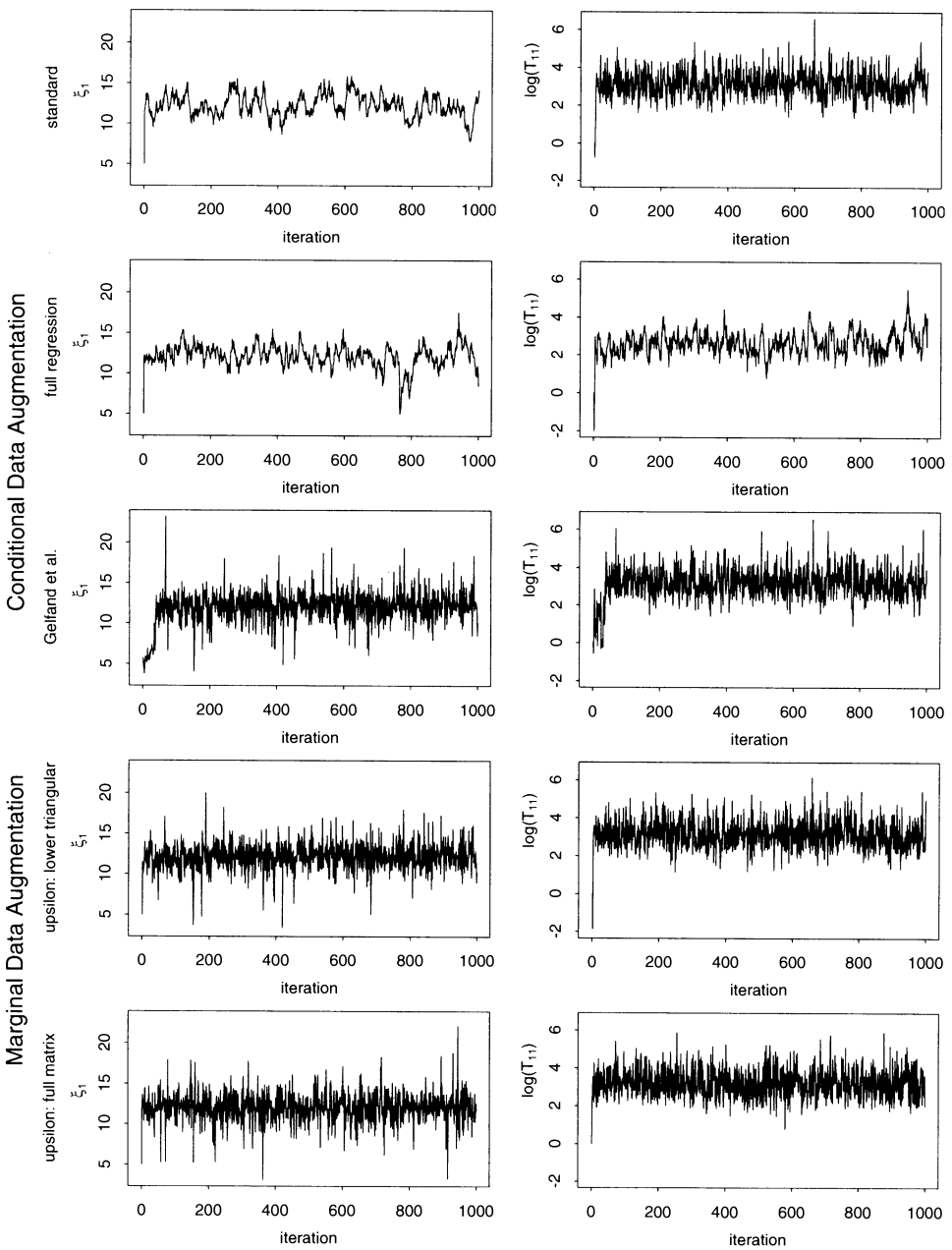

Figure 7. Time Series Plots of ξ_1 and $\log(T_{11}) = \log(var(b_1|T))$ for the Mare Dataset. The plots were generated from the Gibbs samplers using conditional augmentation with α_{std}, α_{reg}, and α_{gsc} and marginal augmentation with $\omega = \infty$ using a lower triangular and a full matrix working parameter (in the rows in this order). As the theoretical criterion of Result A.1 predicts when σ^2 is relatively small (and thus the coefficients of determination are large) choosing $(\alpha_{1j}, \alpha_{2j}) = (1,0)$ as with α_{gsc} is preferable among the conditional DA algorithms.

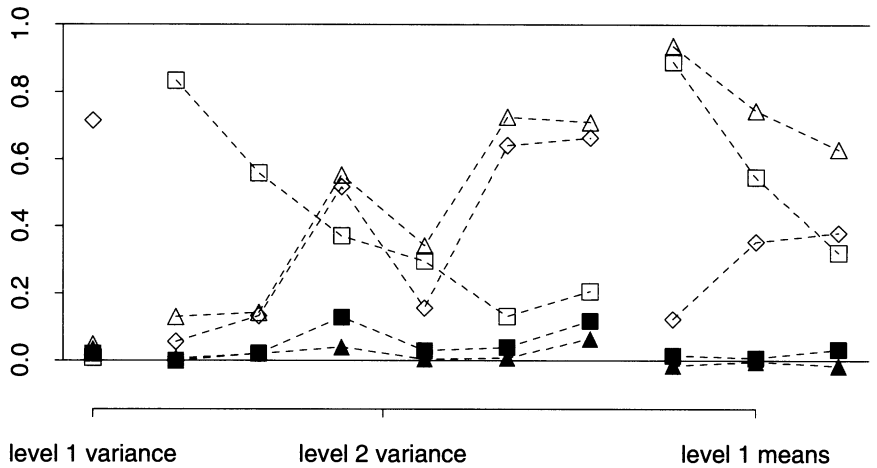

Figure 8. The Overall Improvement of Marginal Augmentation Using a Full Matrix Working Parameter. The plot shows the lag-one autocorrelation for ten scalar parameters using conditional augmentation ($\alpha = \alpha_{\text{std}}$, open triangles; $\alpha = \alpha_{\text{gsc}}$, open diamonds, and $\alpha = \alpha_{\text{reg}}$, open squares) and marginal augmentation (Υ a lower triangular matrix, filled squares and a full matrix, filled triangles). The dramatic overall improvement of marginal augmentation is evident and results in nearly "perfect simulation" with full matrix Υ.

independent, with $\xi_1 = \xi_2 = 1$, $T = I_2$, and the z_i's generated independently from $N(0, 1)$. The subgroup sizes, n_i, varied from three to seven with $\sum_i n_i = 100$. Because the coefficients of determination play a key role in comparing the conditional augmentation samplers, we generated one dataset with $\sigma^2 = .04$ and one dataset with $\sigma^2 = 36$, which correspond respectively to large and small values of the coefficients of determination.

For the dataset generated with $\sigma^2 = .04$, each of the five samplers was run with three different starting values, the ML estimate and two values far from the ML estimate. Figure 9, which displays the Gelman-Rubin $\sqrt{\hat{R}}$ statistic as a function of the iteration number for the six model parameters in (8.13) (some with transformations), illustrates the magnitude of the computational gain resulting from marginal augmentation. The two solid lines, which correspond to two marginal augmentation samplers and are hardly distinguishable, are immediately close to one. In contrast, the $\sqrt{\hat{R}}$ statistics for ξ_1 and ξ_2 with the standard sampler (dotted line) never fall below 10 and thus are out of the range of the two plots in the first row of Figure 9. It is evident that, in terms of $\sqrt{\hat{R}}$, the two marginal augmentation samplers dominate all three conditional augmentation samplers in this example.

To further demonstrate the extremely slow convergence of the standard algorithm (i.e., $\alpha = \alpha_{\text{std}}$), Figure 10 displays the $\sqrt{\hat{R}}$ statistics of ξ_2 and of $\log(T_{11})$ for an additional 19,000 iterations. The figure also provides the corresponding time-series plots for all three chains, which identify the slow convergence with chain 3. It is noteworthy that initially chain 3 appears to have converged for $\log(T_{11})$ at a value far from the posterior mode (see first 10,000 iterations in chain 3), demonstrating once again the danger of using a single chain (Gelman and Rubin 1992).

As in Section 7, we again performed (time consuming) numerical integration to evaluate some low-dimensional marginal posterior densities. Figure 11 displays the marginal posterior densities of $\log(T_{11})$ and $\log((1 + r)/(1 - r))$ and the joint posterior contours of

these two quantities, where r is the (prior) correlation of b_1 and b_2. The histograms and scatterplots display the first 5,000 draws of chain 3 for each algorithm. The first 10 draws were removed to help unify plotting ranges. The standard algorithm completely missed the target density, as expected from Figure 10; note that the centers of the two target marginal densities are near zero, as expected, because the true values of $\log(T_{11})$ and $\log[(1+r)/(1-r)]$ are zero. The slow mixing in the full regression sampler is also evident. In contrast, Gelfand, Sahu, and Carlin's (1995) and the marginal augmentation samplers perform exceptionally well.

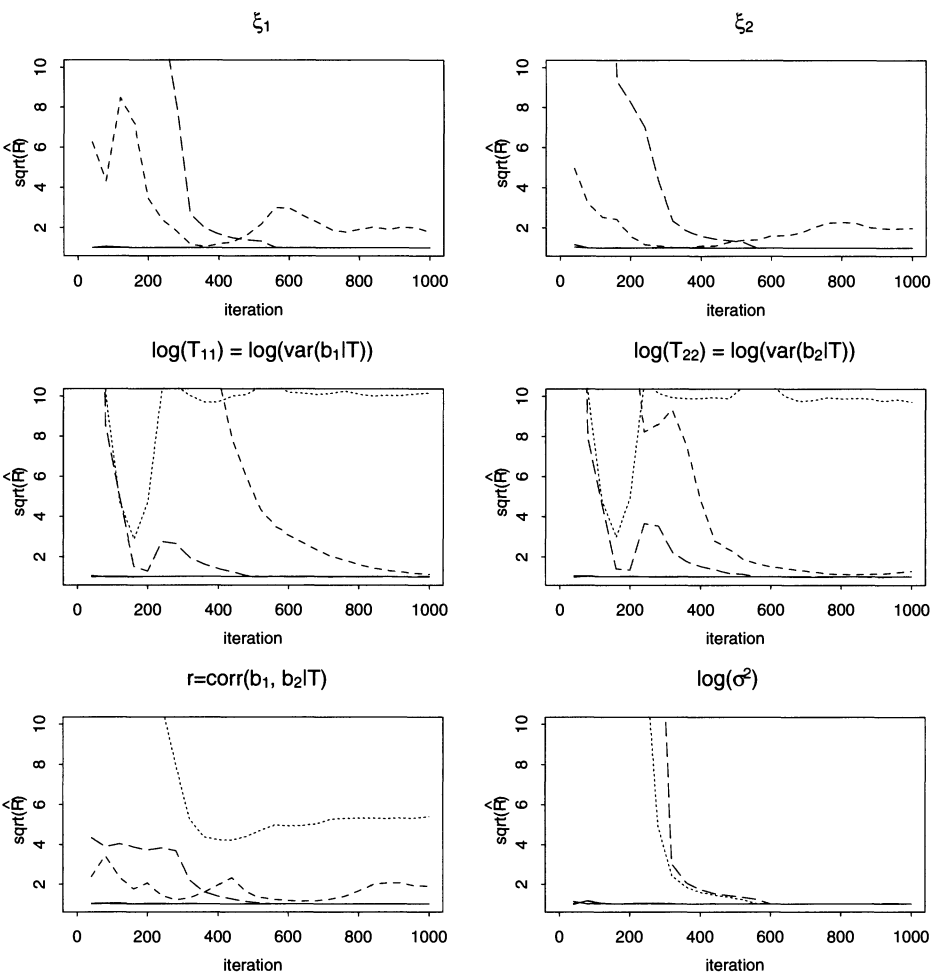

Figure 9. The Gelman-Rubin $\sqrt{\hat{R}}$ Statistic for Five Algorithms for the Mixed-Effects Model. The figure shows $\sqrt{\hat{R}}$ as a function of the iteration number for each of the six parameters in (8.13) using conditional augmentation with α_{std} (the dotted line), α_{reg} (the long dashed line), and α_{gsc} (the short dashed line), and the two marginal augmentations with $\omega = \infty$ (the two nearly indistinguishable solid lines, which are also virtually indistinguishable from the unprinted line $\sqrt{\hat{R}} \equiv 1$). (The data were generated with $\sigma^2 = .04$.) Marginalizing out the working parameter dramatically improves the rate of convergence in terms of $\sqrt{\hat{R}}$ of the samplers. Note that the dotted line is not visible from the plot for ξ_1 and ξ_2—see the first row of Figure 10 for explanation and comparison.

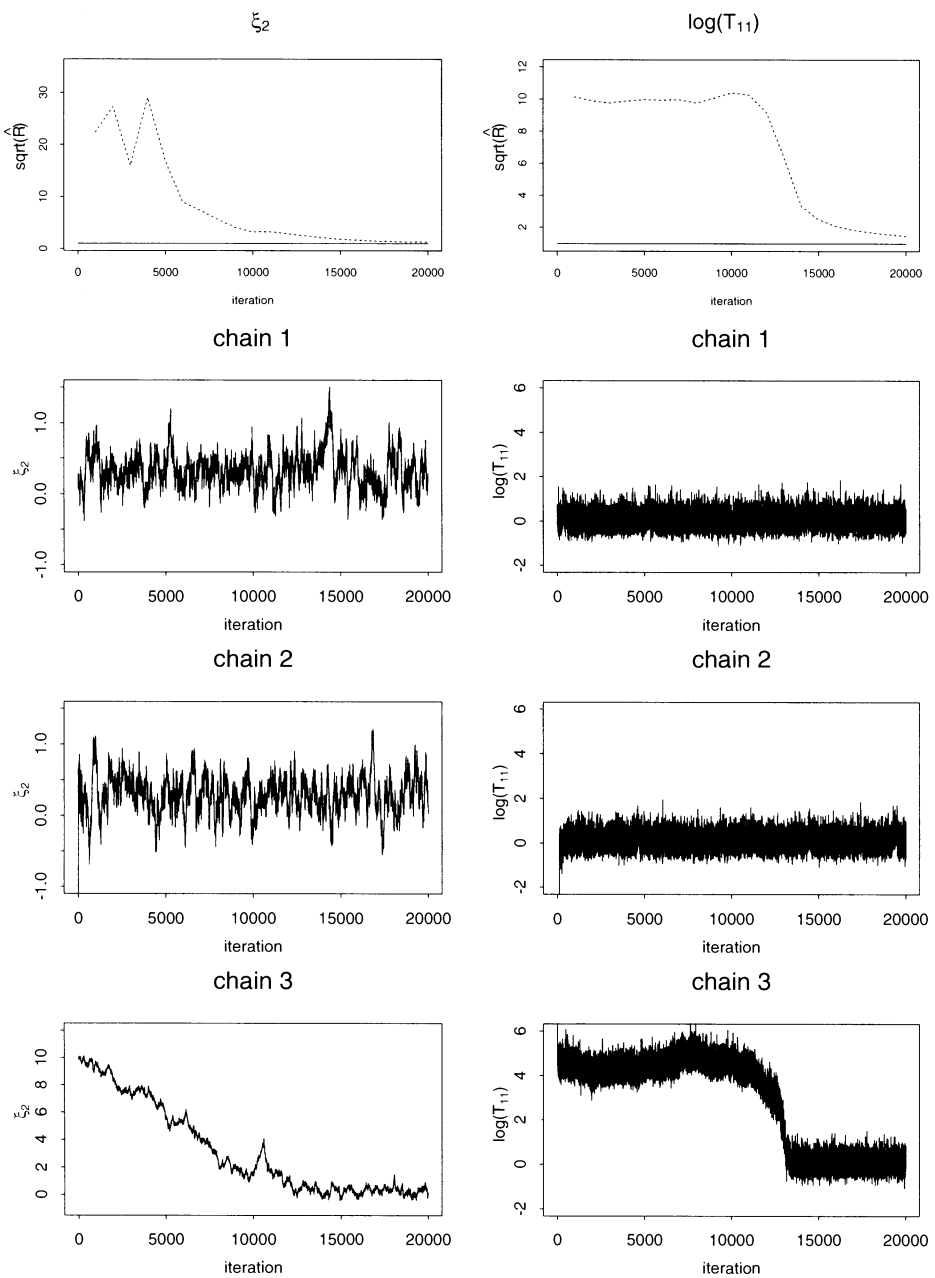

Figure 10. The Gelman-Rubin $\sqrt{\hat{R}}$ Statistic for the Standard Algorithm for the Mixed-Effects Model. The plots represent $\sqrt{\hat{R}}$ as a function of iteration number and time series plots of three chains for ξ_2 and $\log(T_{11})$ (again with $\sigma^2 = .04$ in data generation). Notice the extremely slow convergence of ξ_2 in chain 3 and the illusion of convergence of $\log(T_{11})$ in chain 3 in the first 10,000 iterations. Also note that the disappearance of the dip around iteration 180 in the $\sqrt{\hat{R}}$ plot for $\log(T_{11})$, which was seen in the corresponding plot in Figure 9 (first column, second row), was due to the coarser choice of grid used for the current plot.

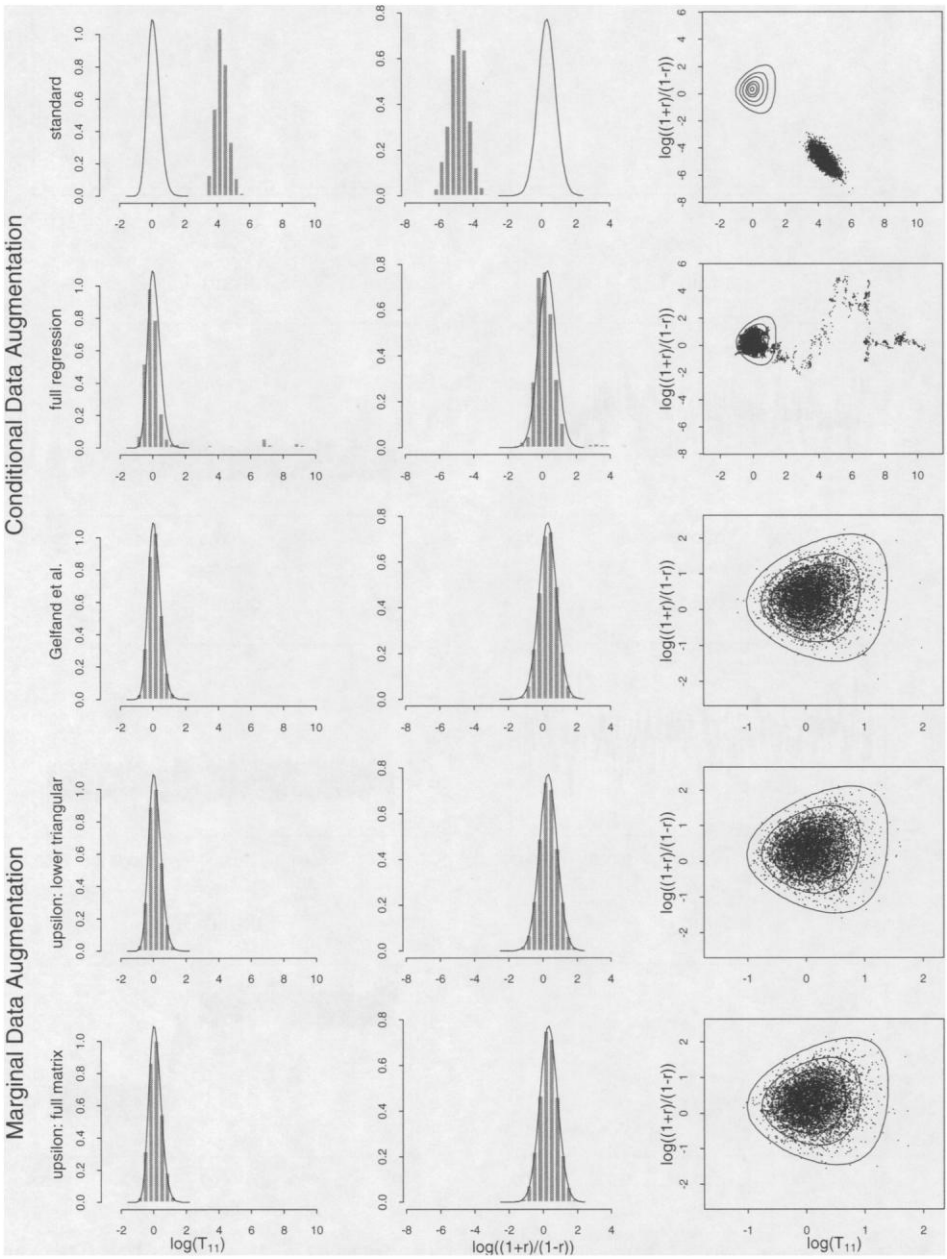

Figure 11. Comparing the MCMC Draws with Low Dimensional Marginal Posteriors. The plots compare histograms and scatter plots of 5,000 draws (omitting the first 10 draws) with marginal posteriors obtained by numerical integration. The slow convergence of the standard and full regression samplers is evident. (The data was generated with $\sigma^2 = .04$.)

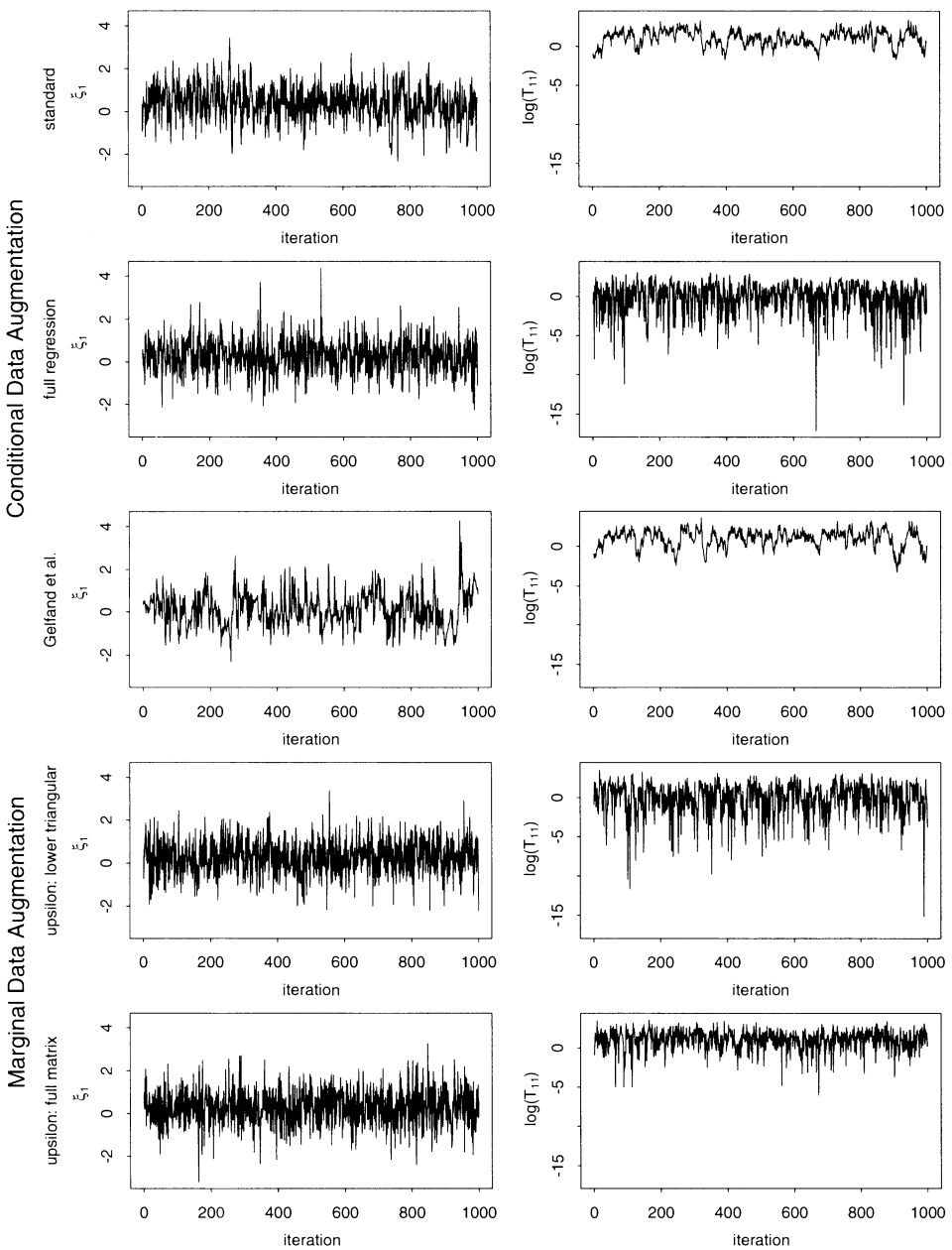

Figure 12. Time Series Plots of ξ_1 *and* $\log(T_{11}) = \log(\text{var}(b_1|T))$ *for the Dataset Generated with* $\sigma^2 = 36$. *Notice the very large autocorrelation in* $\log(T_{11})$ *for the standard and Gelfand, Sahu, and Carlin (1995) samplers. As the theoretical criterion of Result A.1 predicts when* σ^2 *is relatively large choosing* $(\alpha_{1j}, \alpha_{2j}) = (0,1)$ *as in* α_{reg} *is preferable among the conditional DA algorithms.*

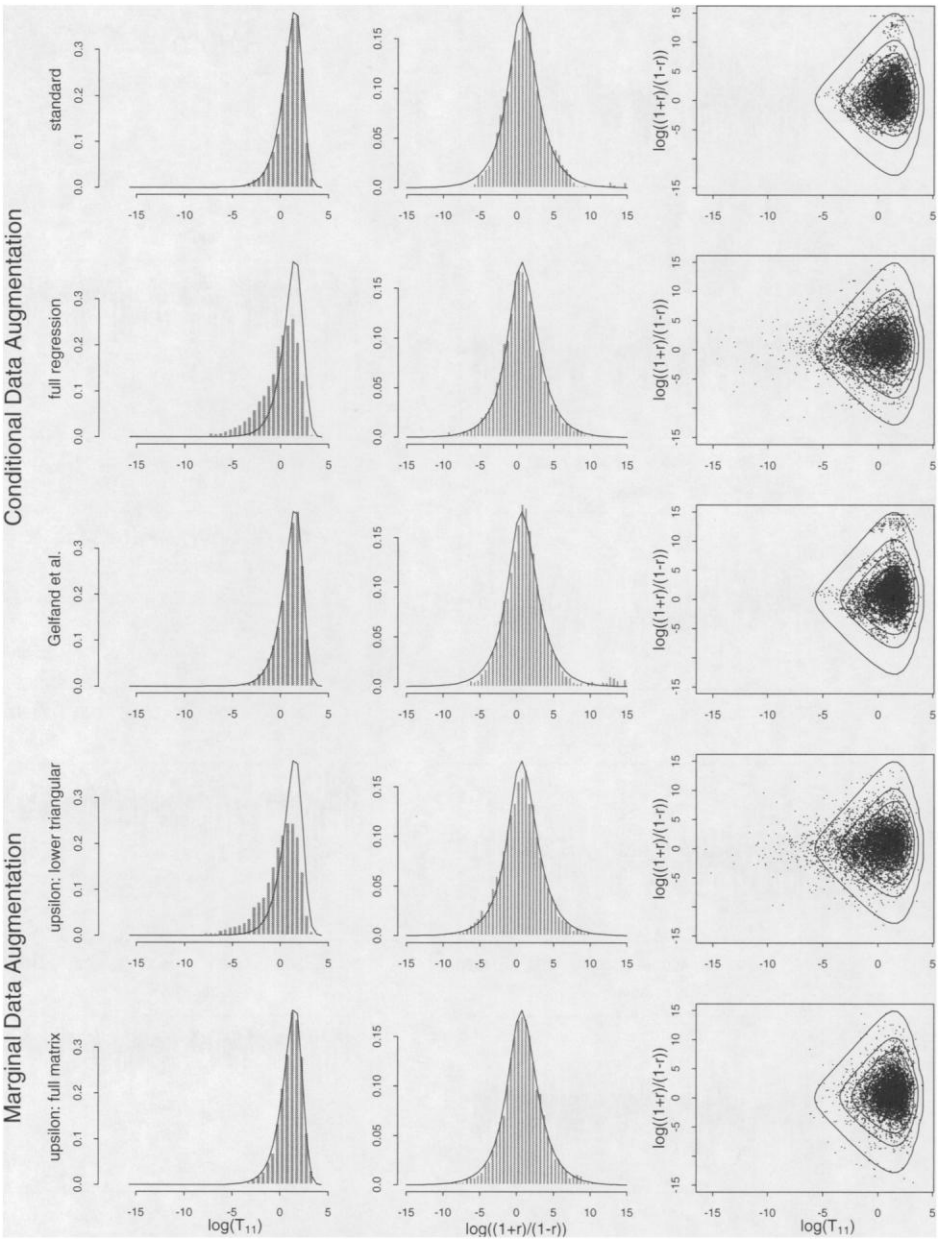

Figure 13. Comparing the MCMC Draws with Low Dimensional Marginal Posteriors. The plots are as in Figure 11, but based on the data generated with $\sigma^2 = 36$. The difficulty, which we believe is numerical, with the full regression sampler and the marginal augmentation sampler with lower triangular working parameter is evident as is the slow mixing of the standard and Gelfand, Sahu, and Carlin (1995) samplers. The marginal augmentation sampler with full matrix working parameter, however, performs well even with such an ill-behaved posterior.

Figure 12 contains plots corresponding to those in Figure 7 for the dataset generated with $\sigma^2 = 36$. The patterns in Figures 7 and 12 are markedly different. Looking at the plots representing $\log(T_{11})$ (the second column) it is clear that both α_{gsc} and α_{std} exhibit high autocorrelation. This is supported by the plots for ξ_1 which reveal that α_{gsc} performs relatively poorly when σ^2 is large (i.e., when coefficients of determination are small). In contrast, α_{reg} as well as the two marginal augmentation samplers display more acceptable autocorrelations for both parameters.

Figure 12 also illustrates a curious behavior of the draws of $\log(T_{11})$ from the full regression sampler and the marginal augmentation sampler with lower triangular Υ. In particular, these samplers travel to very small values of T_{11} much more often than the other three samplers. To investigate this behavior, we again computed several low dimensional marginal posterior densities using numerical integration. The plots in Figure 13 correspond to those in Figure 11 and illustrate the poor behavior of the two algorithms which travel too often to very small values of T_{11}. The third column of Figure 13 also illustrates the interesting "triangular shape" of the posterior, which stretches out in three directions towards singularity for T; the marginal posterior of r is bimodal with modes at 1 and -1. We suspect the difficulty with the full regression sampler and the marginal augmentation sampler with lower triangular Υ is numerical, as computations with nearly singular matrices are known to be unstable. This is partially substantiated by the fact that the computer code we used for the two marginal samplers are very similar, differing only by a change in the degrees of freedom for σ^2 and the set of covariates in the level 1 regression. It appears that by summing over more terms with the full matrix Υ has helped to circumvent the numerical instabilities associated with the near singularity of T, though this is speculative.

The "triangular shape" of the posterior appears to have caused the poor mixing of the standard and Gelfand, Sahu, and Carlin (1995) samplers, even with 5,000 draws. For example, it is seen that both samplers missed the area close to $r = -1$, and stayed a bit too long in the area close to $r = 1$. This example was chosen to be somewhat extreme (i.e., the residual variance is much larger than the variances of the random effects) to test the robustness of our marginal augmentation algorithms. But it is not pathological as such triangular shape posteriors/likelihoods do occur with real-data modeling, even for normal one-way random effects analysis [e.g., see Pawitan (2000) for a triangular shape likelihood contour plot for a dataset on estrogen level from postmenopausal women]. Whereas the marginal augmentation with lower triangular Υ suffers from an apparent numerical problem, the one with full Υ performs exceedingly well for all aspects we have examined for such an apparently difficult problem.

9. WARNINGS AND FUTURE WORK

9.1 LIMITATIONS

Like any statistical or computational method, including the auxiliary variable methods (see Liu 1994), there are limitations to the methods we propose. In addition to demonstrating the nonapplicability of the EM criterion as discussed at the end of Section 2, slice sampling also provides a good example of the limitations of optimal marginalization via

affine transformations, a key to the successful strategy underlying all three applications presented in Sections 6–8. Specifically, consider the simple case of (1.2) with $K = 1$, that is, our "standard" augmentation model is

$$f(x, \tilde{u}) \propto \pi(x) I\{0 < \tilde{u} \leq l(x)\}. \tag{9.1}$$

To implement the optimal marginalization using an affine transformation we introduce a scale working parameter α, i.e., $u = \alpha\tilde{u}$ with the Haar measure $p(\alpha) \propto \alpha^{-1}$. (Since u and $\tilde{u}$ need to have the same support $(0, \infty)$, location shift is not allowed.) This implies a joint measure (see (3.3))

$$f(x, u, \alpha) \propto \pi(x) I\{u \leq \alpha l(x)\}\alpha^{-2}. \tag{9.2}$$

It follows then that in order to implement Scheme 2 of Section 3, at the $(t + 1)$st iteration, we need to (I) draw $u^{(t+1)}$ from $f(u|x^{(t)}, \alpha^{(t)})$ and (II) draw $(x^{(t+1)}, \alpha^{(t+1)})$ from $f(x, \alpha|u^{(t+1)})$ (and discard $\alpha^{(t+1)}$). Step (I) is trivial since $f(u|x, \alpha)$ is uniform on $(0, \alpha l(x))$, but if we could implement Step (II), then we would not need slice sampling because under (9.2), $f(x|u^{(t+1)}) \propto \pi(x)l(x)$, our target density. In other words, the optimal marginalization via scale transformation does provide an "optimal" sampler—it provides independent draws from our target density, but only if we know how to make such draws in the first place!

This is a good example of the "artistic" aspect of the search for efficient data augmentation schemes because it appears impossible to quantify the notion of "implementability" in general. Therefore, the search for balance between the speed and complexity of the resulting algorithms is largely a matter of art, at least at present. The need for balance is further highlighted by another aspect of this example. If implementability/simplicity is the only consideration, then we would adopt Scheme 3 (i.e., the three-step Gibbs sampler) which replaces Step (II) by (IIa) drawing $\alpha^{(t+1)}$ from $f(\alpha|x^{(t)}, u^{(t+1)}) \propto I\{\alpha > u^{(t+1)}/l(x^{(t)})\}\alpha^{-2}$ and (IIb) drawing $x^{(t+1)}$ from $f(x|\alpha^{(t+1)}, u^{(t+1)}) \propto \pi(x)I\{l(x) > u^{(t+1)}/\alpha^{(t+1)}\}$. We can implement this algorithm, but it is (stochastically) identical to the original slice sampler and thus the working parameter offers no improvement! This can be seen by comparing the stochastic mapping $x^{(t)} \to x^{(t+1)}$ for both algorithms. Let $X_\pi(c)$ represent a random draw from the truncated density $\pi(x)I\{l(x) \geq c\}$. Noting that the $\alpha^{(t+1)}$ from (IIa) can be written as $\alpha^{(t+1)} = \alpha_0 u^{(t+1)}/l(x^{(t)})$, where α_0 follows the Pareto density, $\alpha^{-2}I\{\alpha > 1\}$, the mapping under Scheme 3 can be written as $x^{(t+1)} = X_\pi(u^{(t+1)}/\alpha^{(t+1)}) = X_\pi(l(x^{(t)})/\alpha_0)$. But this is the same as the mapping under the original slice sampler, $x^{(t+1)} = X_\pi(l(x^{(t)})u_0)$, where u_0 is a uniform variate on (0,1), because α_0^{-1} is uniform on (0,1) as well.

9.2 Possibilities and Open Problems

The Combined Strategy (p. 11) illustrates how the marginal and conditional augmentation strategies compliment each other and can lead to very efficient algorithms when used in conjunction. This reflects a general principle: many MCMC strategies become more effective when combined. As an example, model reduction techniques, as defined by Meng and van Dyk (1997), can lead to algorithms that are easier to implement by replacing one

draw with several conditional draws (e.g., the Gibbs sampler). In some cases it is possible to use different data augmentation schemes for different conditional draws, which can lead to algorithms that converge faster. In the context of EM-type algorithms, this possibility is demonstrated by the ECME algorithm (Liu and Rubin 1994), the SAGE algorithm (Fessler and Hero 1994), and more generally the AECM algorithm (Meng and van Dyk 1997). We expect that the Combined Strategy is also effective for searching for efficient stochastic versions of these algorithms. In this context, a needed investigation is that into the possibility and the effectiveness of using the more complex formulas of the rate of convergence of the AECM algorithm (see Meng and van Dyk 1997, Theorem 4) in place of that of the EM algorithm in implementing the deterministic approximation component of our search procedures.

Finally, the use of an improper working prior, which leads to a nonpositive recurrent joint Markov chain with positive recurrent subchains, may provide additional motivation for theoretical investigation of Markov chains with improper invariant distributions, a difficult area with a number of open problems (e.g., Tierney 1996). It also adds an interesting component to the recent debate over the use of nonpositive recurrent Markov chains in the MCMC setting; see, for example, the paper by Casella (1996) and the discussions by J. Berger and by E. George. The use of marginal augmentation with improper priors demonstrates that sometimes it is beneficial to purposely construct a nonpositive recurrent Markov chain on a larger space in order to have fast mixing positive recurrent subchains that simulate the target densities. In fact, Hobert (in press) showed that these chains have to be *sub-subchains* when the chain in the larger space is a two-step Gibbs sampler (e.g., in Scheme 2 of Section 3, the resulting chain on θ is a subchain of the chain on (θ, α), which itself is a subchain of the parent Gibbs sampler). This is because the direct subchains (e.g., the chain on (θ, α)) are necessarily nonpositive recurrent as well. The construction of such embedded subchains or sub-subchains may well be a matter of art, but we hope the Combined Strategy proposed in this article can help more people be artistic in the development of many practically important MCMC algorithms that exhibit simplicity, stability, and speed.

APPENDIX: TECHNICAL DETAILS FOR SECTION 8

I. VERIFYING THE LIMITING CONDITION FOR THE ALGORITHMS IN SECTION 8.1

Following the notation and assumptions in Section 8.1, and noting that $|\frac{\partial\theta}{\partial\tilde{\theta}}| = |\Upsilon|^{q+1}$, we have

$$\begin{aligned}
&\log p(\tilde{\theta}, \alpha, Y_{\text{aug}} | Y_{\text{obs}}, \omega) \\
&\quad \propto \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{m}[y_i - X_i\beta - Z_i\tilde{\xi} - \tilde{Z}_i(d_i)\text{vec}(\Upsilon)]^\top[y_i - X_i\beta - Z_i\tilde{\xi} - \tilde{Z}_i(d_i)\text{vec}(\Upsilon)] \\
&\qquad -(1+\frac{n}{2})\log\sigma^2 - \frac{m+1}{2}\log|\tilde{T}| \\
&\qquad -\frac{1}{2}\sum_{i=1}^{m}(d_i-\gamma)^\top\tilde{T}^{-1}(d_i-\gamma) + h_\omega(\alpha), \qquad \text{(A.1)}
\end{aligned}$$

where

$$h_\omega(\alpha) = -\frac{1}{2\omega}\gamma^\top(\Upsilon\Upsilon^\top)^{-q}\gamma - \frac{1}{2\omega}\left[\text{vec}(\Upsilon) - \text{vec}(I)\right]^\top\left[\text{vec}(\Upsilon) - \text{vec}(I)\right].$$

Since $\lim_{\omega\to\infty} h_\omega(\alpha) = 0$, and $\exp\{h_\omega(\alpha)\} \le 1$ for all α, by the Dominated Convergence Theorem, it is easy to verify Condition 1 of Lemma 1. Namely, the full conditional distribution for (θ, α) in Step 2 is consistent with the corresponding full conditional derived from (A.1) given $d_i = b_i, i = 1, \ldots, m$ and $\omega = \infty$. To verify Condition 2, we see from (8.3) that if we replace d_i by $Ab_i + a$, then, conditioning on $b = \{b_1, \ldots, b_m\}$, we have

$$y_i = X_i\beta + Z_i(\tilde{\xi} + \Upsilon a) + \tilde{Z}_i(b_i)\text{vec}(\Upsilon A) + e_i, \quad i = 1, \ldots, m. \tag{A.2}$$

Comparing (A.2) with (8.3), it is clear that (assuming the same random numbers) $\beta_b^{(t+1)} = \beta_d^{(t+1)}$, $[\sigma^2]_b^{(t+1)} = [\sigma^2]_d^{(t+1)}$, and

$$\Upsilon_b = \Upsilon_d A, \qquad \tilde{\xi}_b = \tilde{\xi}_d + \Upsilon_d a, \tag{A.3}$$

where the subscript indicates which missing data (i.e., b or d) are conditioned upon in performing Step 2. From (8.8)–(8.9), we have

$$\tilde{T}_b = A^{-1}\tilde{T}_d(A^{-1})^\top \quad \text{and} \quad \gamma_b = A^{-1}(\gamma_d - a). \tag{A.4}$$

Consequently,

$$\xi_b^{(t+1)} = \tilde{\xi}_b + \Upsilon_b\gamma_b = \tilde{\xi}_d + \Upsilon_d a + \Upsilon_d AA^{-1}(\gamma_d - a) = \xi_d^{(t+1)},$$

and

$$T_b^{(t+1)} = \Upsilon_b\tilde{T}_b\Upsilon_b^\top = \Upsilon_d AA^{-1}\tilde{T}_d(A^{-1})^\top A^\top\Upsilon_d^\top = T_d^{(t+1)}.$$

Thus, Condition 2 of Lemma 1 holds, and we can use b instead of d (i.e., taking $\Upsilon = I_q$ and $\gamma = 0$) in Step 1.

II. Conditional Augmentation Algorithms for Mixed-Effects Models

We now give the details of the Gibbs samplers for sampling from the posterior $p(\theta, Y_{\text{aug}}|Y_{\text{obs}}, \alpha)$ using α_{std}, α_{gsc}, and α_{reg}. After sampling from $p(Y_{\text{aug}}|Y_{\text{obs}}, \theta, \alpha)$ as described in Section 8.2, all three algorithms sample from $p(\theta|Y_{\text{aug}}, \alpha)$, the implementation of which varies with α.

For $\alpha = \alpha_{\text{std}}$, we will use the noninformative prior distribution $p_1(\beta, \log(\sigma), \xi, T) \propto |T|^{-1/2}$. Given $\{c_i, i = 1, \ldots, m\}$, we draw θ as follows. Using the system level of the model we draw

$$T^{-1}|\alpha, Y_{\text{aug}} \sim \text{Wishart}_{m-q}\left[\left(\sum_{i=1}^m c_i c_i^\top\right)^{-1}\right] \tag{A.5}$$

and using the observation level of the model, we draw

$$\sigma^2|\alpha, Y_{\text{aug}} \sim \frac{1}{\chi^2_{n-\tilde{p}}} \sum_{i=1}^{m} (y_i - \tilde{X}_i \mu_{\tilde{\beta}} - Z_i c_i)^\top (y_i - \tilde{X}_i \mu_{\tilde{\beta}} - Z_i c_i),$$

and

$$\tilde{\beta}|\sigma^2, \alpha, Y_{\text{aug}} \sim N_{\tilde{p}}\left(\mu_{\tilde{\beta}}, \quad \sigma^2 \tilde{B}^{-1}\right) \quad \text{with} \quad \mu_{\tilde{\beta}} = \tilde{B}^{-1} \sum_{i=1}^{m} \tilde{X}_i^\top (y_i - Z_i c_i),$$

where $\tilde{p} = p + q$, $\tilde{X}_i = (X_i\ , Z_i)$, $\tilde{\beta}^\top = (\beta^\top, \xi^\top)$, and $\tilde{B} = \sum_{i=1}^{m} \tilde{X}_i^\top \tilde{X}_i$.

With $\alpha = \alpha_{\text{gsc}}$, we use the same noninformative prior distribution and draw θ, using the system level of the model

$$T^{-1}|\alpha, Y_{\text{aug}} \sim \text{Wishart}_{m-q-1}\left[\left(\sum_{i=1}^{m} (c_i - \mu_\xi)(c_i - \mu_\xi)^\top\right)^{-1}\right], \tag{A.6}$$

and

$$\xi|T, \alpha, Y_{\text{aug}} \sim \text{N}_q(\mu_\xi, \frac{1}{m} T) \quad \text{with} \quad \mu_\xi = \frac{1}{m} \sum_{i=1}^{m} c_i.$$

Using the observation level of the model, we then draw

$$\sigma^2|\alpha, Y_{\text{aug}} \sim \frac{1}{\chi^2_{n-p}} \sum_{i=1}^{m} (y_i - X_i \mu_\beta - Z_i c_i)^\top (y_i - X_i \mu_\beta - Z_i c_i),$$

and, defining $B = \sum_{i=1}^{m} X_i^\top X_i$,

$$\beta|\sigma^2, \alpha, Y_{\text{aug}} \sim \text{N}_p\left(\mu_\beta, \quad \sigma^2 B^{-1}\right), \quad \text{with} \quad \mu_\beta = B^{-1} \sum_{i=1}^{m} X_i^\top (y_i - Z_i c_i).$$

Finally when using $\alpha = \alpha_{\text{reg}}$, we change the noninformative prior distribution to $p_2(\beta, \log(\sigma), \xi, L) \propto 1$. We use this prior for computational ease with the understanding that the effect will be small for comparative purposes especially because the two priors are equivalent if we constrain T to be diagonal. (Indeed, the effect of this different prior is not visible in the marginal density of $\log[(1 + r)/(1 - r)]$ for $\alpha = \alpha_{\text{reg}}$ in Figure 13.) For this implementation we draw all the parameters in θ from the observation level. In particular, we draw

$$\sigma^2|\alpha, Y_{\text{aug}} \sim \frac{1}{\chi^2_{n-\tilde{p}}} \sum_{i=1}^{m} (y_i - \tilde{X}_i \mu_{\tilde{\beta}})^\top (y_i - \tilde{X}_i \mu_{\tilde{\beta}})$$

and $\tilde{\beta}|\sigma^2, \alpha, Y_{\text{aug}} \sim \text{N}_{\tilde{p}}\left(\mu_{\tilde{\beta}},\ \sigma^2 \tilde{B}^{-1}\right)$, where $\tilde{p} = p + q(q + 3)/2$, $\tilde{X}_i = (X_i, Z_i, \tilde{Z}_i(c_i))$, $\tilde{\beta} = (\beta, \xi, \text{vec}(\Delta U))$, $\tilde{B} = \sum_{i=1}^{m} \tilde{X}_i^\top\ \tilde{X}_i$, and $\mu_{\tilde{\beta}} = \tilde{B}^{-1} \sum_{i=1}^{m} \tilde{X}_i^\top y_i$.

III. THEORETICAL COMPARISONS OF CONDITIONAL AUGMENTATION ALGORITHMS

Result A.1 When (I) the mean and variance structure of the model (8.1) are correctly specified, (II) T is diagonal in the fitted model, and (III) $S_{XZ} \equiv \sum_{i=1}^m X_i^\top Z_i = 0$ and $S_Z \equiv \sum_{i=1}^m Z_i^\top Z_i$ is diagonal, the asymptotically (as $m \to \infty$) optimal value of $(\alpha_{1j}, \alpha_{2j}) \in \{(0,0),(0,1),(1,0)\} \equiv \mathcal{A}_0$ according to Criterion 5.1 (with $\alpha_k = (\alpha_{k1}, \ldots, \alpha_{kq})$ for $k = 1, 2$) has the following properties:

(a) The optimal value is $(\alpha_{1j}, \alpha_{2j}) = (1,0)$ if

$$\frac{u_j^{2\star}}{m} \sum_{i=1}^m Z_{ij}^\top Z_{ij} \le \sigma^{2\star}, \quad \text{i.e.,} \quad D_j < 1/2, \tag{A.7}$$

where $U^\star = \text{diag}(u_1^\star, \ldots, u_q^\star)$ and $\sigma^{2\star}$ are posterior modes, and

$$D_j = \frac{u_j^{2\star} \sum_{i=1}^m Z_{ij}^\top Z_{ij}}{m\sigma^{2\star} + u_j^{2\star} \sum_{i=1}^m Z_{ij}^\top Z_{ij}} \tag{A.8}$$

can be viewed as a measure of the overall coefficient of determination from the jth component of the random effect.

(b) If

$$\frac{u_j^{2\star}}{m} \sum_{i=1}^m Z_{ij}^\top Z_{ij} \ge 2\sigma^{2\star}, \quad \text{i.e.,} \quad D_j > 2/3, \tag{A.9}$$

the optimal value is $(\alpha_{1j}, \alpha_{2j}) = (0,1)$.

(c) If neither (A.7) nor (A.9) hold, both $(\alpha_{1j}, \alpha_{2j}) = (0,1)$ and $(1,0)$ are preferable to $(0,0)$. This implies that at least one of $(0,1)$ and $(1,0)$ is always superior to $(0,0)$.

Proof: We need only consider the submatrix

$$\tilde{I}_{\text{aug}}^{(1)}(\alpha) = \begin{pmatrix} I_{\xi\xi}(\alpha) & I_{\xi U}(\alpha) \\ I_{\xi U}^\top(\alpha) & I_{UU}(\alpha) \end{pmatrix} \tag{A.10}$$

of $I_{\text{aug}}^{(1)}(\alpha)$ since the rest of the matrix is (asymptotically) free of α. To derive the elements of $I_{\xi\xi}(\alpha)$, we differentiate $Q_\alpha(\theta|\theta^\star) = \text{E}[\log p(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \theta^\star, \alpha]$ twice with respect to ξ:

$$-\left.\frac{\partial^2 Q_\alpha(\theta|\theta^\star)}{\partial\xi \cdot \partial\xi}\right|_{\theta=\theta^\star} = \frac{1}{\sigma^{2\star}}(I - A_2)S_Z(I - A_2) + mA_2[U^\star]^{-2}A_2, \tag{A.11}$$

where $A_2 = \text{diag}\{\alpha_{21}, \ldots, \alpha_{2q}\}$. Likewise we can derive the elements of $I_{\xi U}(\alpha)$ as

$$-\left.\frac{\partial^2 Q_\alpha(\theta|\theta^\star)}{\partial\xi_k \partial u_l^2}\right|_{\theta=\theta^\star} = \frac{(1-\alpha_{2k})\alpha_{1l}}{2\sigma^{2\star}u_k^{2\star}} \sum_{i=1}^m Z_{ik}^\top Z_{il}(\hat{b}_{il}^\star + \alpha_{2l}\xi_k^\star) + \zeta_{kl}, \tag{A.12}$$

where $\hat{b}_{il}^\star$ is the lth component of the vector $\hat{b}_i^\star = \text{E}(b_i|Y_{\text{obs}}, \theta^\star)$ and $\zeta_{kl} = 0$ if $\alpha_{1k}\alpha_{2k} = \alpha_{1l}\alpha_{2l} = 0$. Since $\text{E}(\hat{b}_i^\star|\theta) = 0$, $\sum_{i=1}^m Z_{ik}^\top Z_{il} = 0$ for $k \ne l$ and $(1-\alpha_{2k})\alpha_{2l} = 0$ for $k = l$, we find $\lim_{m\to\infty} I_{\xi U}(\alpha) = 0$. Finally, we derive the matrix $I_{UU}(\alpha)$ as

$$
\begin{aligned}
&-\frac{\partial^2 Q_\alpha(\theta|\theta^\star)}{\partial U^2 \cdot \partial U^2}\bigg|_{\theta=\theta^\star} \\
&= \frac{1}{4\sigma^{2\star}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \operatorname{diag}\left\{\frac{\alpha_{11} z_{ij1}}{u_1^{2\star}}, \ldots, \frac{\alpha_{1q} z_{ijq}}{u_q^{2\star}}\right\} \hat{S}_i^\star(\alpha_2) \operatorname{diag}\left\{\frac{\alpha_{11} z_{ij1}}{u_1^{2\star}}, \ldots, \frac{\alpha_{1q} z_{ijq}}{u_q^{2\star}}\right\}, \\
&\quad + \operatorname{diag}\left\{(1-\alpha_{11})^2 \frac{m}{2u_1^{4\star}} + (\alpha_{11}\alpha_{21})^2 \frac{m\xi_1^{2\star}}{4u_1^{6\star}}, \ldots, \right. \\
&\qquad \left. (1-\alpha_{1q})^2 \frac{m}{2u_q^{4\star}} + (\alpha_{1q}\alpha_{2q})^2 \frac{m\xi_q^{2\star}}{4u_q^{6\star}}\right\} - \frac{1}{2} U^{-4\star}, \qquad \text{(A.13)}
\end{aligned}
$$

where $\hat{S}_i^\star(\alpha_2) = \mathrm{E}[(b_i + \alpha_2 * \xi)(b_i + \alpha_2 * \xi)^\top | Y_{\text{obs}}, \theta^\star, \alpha]$, which can be expressed, using (8.4) evaluated at $\theta^{(t)} = \theta^\star$, as

$$
\begin{aligned}
\hat{S}_i^\star(\alpha_2) &= T^\star + T^\star Z_i^\top W_i(\theta^\star) \\
&\quad \times \left[(y_i - X_i\beta^\star - Z_i\xi^\star)(y_i - X_i\beta^\star - Z_i\xi^\star)^\top W_i(\theta^\star) - I\right] Z_i T^\star \\
&\quad + \hat{b}_i(\theta^\star)(\alpha_2 * \xi^\star)^\top + (\alpha_2 * \xi^\star)\hat{b}_i(\theta^\star)^\top + (\alpha_2 * \xi^\star)(\alpha_2 * \xi^\star)^\top. \qquad \text{(A.14)}
\end{aligned}
$$

Since $\mathrm{E}[\hat{b}_i(\theta)|\theta] = 0$ and $\mathrm{E}[(y_i - X_i\beta - Z_i\xi)(y_i - X_i\beta - Z_i\xi)^\top|\theta] = W_i^{-1}(\theta)$, asymptotically (as $m \to \infty$) $I_{UU}(a)$ is diagonal with the kth diagonal term equal to (noting that $T = U$ when $\Delta = I$),

$$
(1-\alpha_{1k})^2 \frac{m}{2u_k^{4\star}} + \alpha_{1k}^2 \frac{\sum_{i=1}^m Z_{ik}^\top Z_{ik}}{4\sigma^{2\star} u_k^{2\star}} + (\alpha_{1k}\alpha_{2k})^2 \left[\frac{m\xi_k^{2\star}}{4u_k^{6\star}} + \frac{\xi_k^{2\star} \sum_{i=1}^m Z_{ik}^\top Z_{ik}}{4\sigma^{2\star} u_k^{4\star}}\right] - \frac{1}{2u_k^{4\star}}. \qquad \text{(A.15)}
$$

Thus, $\tilde{I}_{\text{aug}}^{(1)}(\alpha)$ is asymptotically diagonal with diagonal elements given in (A.11) and (A.15). In order to minimize $\tilde{I}_{\text{aug}}^{(1)}(\alpha)$, we need only minimize each of its diagonal terms. We are now in a position to prove each of the statements in the result.

(a) If (A.7) holds, the kth diagonal term of $I_{\xi\xi}(\alpha)$ is minimized by $\alpha_{2k} = 0$. Noting that if (A.7) holds then (A.9) does not hold, we see that (A.15) is minimized by $\alpha_{1k} = 1$.

(b) If (A.9) holds, (A.15) is minimized by $\alpha_{1k} = 0$ and (A.7) does not hold. These two facts yield the result.

(c) The first statement follows immediately from (A.11) and (A.15). The second follows since the suppositions of statements (a), (b), and (c) are the only possibilities. □

IV. THEORETICAL RESULTS ON MARGINAL AUGMENTATION ALGORITHMS

The search for an optimal ω in the mixed-effects model is somewhat more complicated than that of Sections 6–7, partially because $\Delta_{\text{EM}}^{(2)}(\omega)$ does not admit the positive semidefinite ordering. In such cases Criterion 5.2 can still be effective and suggestive as long as a major submatrix of $\Delta_{\text{EM}}^{(2)}(\omega)$ can be ordered. Recall that maximizing $\Delta_{\text{EM}}^{(2)}(\omega)$ is only an approximately sufficient condition for finding optimal ω, and it is by no means necessary.

If desired, one can carry out a more involved analytic search for an optimal ω by directly working with, say, $\rho(\mathcal{F}_{EM}(\omega))$. In our experience, ignoring a relatively unimportant part of $\Delta^{(2)}_{\text{EM}}(\omega)$ can be very effective, as shown in the following. Note that for the rest of this Appendix, the marginal augmentation uses a lower triangular Υ.

Result A.2 Under the conditions of Result A.1, the submatrix $\Delta^{(2)}_{\text{EM}}(\omega)$ corresponding to (β, ξ, T), that is, all the parameters except σ^2, is an increasing function of ω, in terms of a positive semidefinite ordering. This suggests using the improper working prior corresponding to $\lim_{\omega\to\infty} p(\gamma, \Upsilon|\omega) \propto |\Upsilon|^{-q}$.

Proof: Under (8.2), I_{aug} is (asymptotically) block diagonal (similar to the previous calculations), and we need only consider the submatrix,

$$\begin{pmatrix} I_{\sigma^2\sigma^2} & I_{\sigma^2\xi} & I_{\sigma^2 U^2} & I_{\sigma^2\gamma} & I_{\sigma^2\Upsilon} \\ I_{\sigma^2\xi}^\top & I_{\xi\xi} & I_{\xi U^2} & I_{\xi\gamma} & I_{\xi\Upsilon} \\ I_{\sigma^2 U^2}^\top & I_{\xi U^2}^\top & I_{U^2U^2} & I_{U^2\gamma} & I_{U^2\Upsilon} \\ I_{\sigma^2\gamma}^\top & I_{\xi\gamma}^\top & I_{U^2\gamma}^\top & I_{\gamma\gamma} & I_{\gamma\Upsilon} \\ I_{\sigma^2\Upsilon}^\top & I_{\xi\Upsilon}^\top & I_{U^2\Upsilon}^\top & I_{\gamma\Upsilon}^\top & I_{\Upsilon\Upsilon} \end{pmatrix}. \tag{A.16}$$

By differentiating $\text{E}[\log p(\tilde{\theta}|Y_{\text{aug}})|\theta^\star, \alpha^\star, Y_{\text{obs}}]$ twice with respect to $(\sigma^2, \xi, U, \alpha)$, and noting that $\gamma^\star = 0$ and $\Upsilon^\star = I_q$, it can be shown that (A.16) is of the form

$$\begin{pmatrix} \frac{N+2}{2\sigma^{4\star}} & 0 & 0 & 0 & \frac{1}{2\sigma^{2\star}\omega}1_q^\top \\ 0 & \frac{1}{\sigma^{2\star}}S_Z & 0 & -\frac{1}{\sigma^{2\star}}S_Z & 0 \\ 0 & 0 & \frac{m+1}{2}(T^\star)^{-2} & 0 & -(m+1)(T^\star)^{-1} \\ 0 & -\frac{1}{\sigma^{2\star}}S_Z & 0 & \frac{1}{\sigma^{2\star}}S_Z + m(T^\star)^{-1} + \frac{1}{\omega}I_q & 0 \\ \frac{1}{2\sigma^{2\star}\omega}1_q & 0 & -(m+1)(T^\star)^{-1} & 0 & I_{\Upsilon\Upsilon} \end{pmatrix},$$

and asymptotically (replacing $\text{E}[d_i d_i^\top|Y_{\text{obs}}, \theta]$ by its expectation over Y_{obs} evaluated at $\theta = \theta^\star$, that is, as in (A.15))

$$I_{\Upsilon\Upsilon} \approx 2m + \frac{1}{4}\left[\frac{1}{\sigma^{2\star}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{Z}_{ij}T^\star\tilde{Z}_{ij} + \left(\frac{1}{\omega} + 2m - q\right)I_q\right],$$

with $\tilde{Z}_{ij} = \text{diag}\{z_{ij1}, \ldots, z_{ijq}\}$. We can now compute

$$\begin{aligned} \Delta^{(2)}_{\text{EM}}(\omega) &= \begin{pmatrix} I_{\sigma^2\gamma} & I_{\sigma^2\Upsilon} \\ I_{\xi\gamma} & I_{\xi\Upsilon} \\ I_{U\gamma} & I_{U\Upsilon} \end{pmatrix}\begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\Upsilon} \\ I_{\gamma\Upsilon}^\top & I_{\Upsilon\Upsilon} \end{pmatrix}^{-1}\begin{pmatrix} I_{\sigma^2\gamma} & I_{\sigma^2\Upsilon} \\ I_{\xi\gamma} & I_{\xi\Upsilon} \\ I_{U\gamma} & I_{U\Upsilon} \end{pmatrix}^\top \\ &= \begin{pmatrix} \Delta_{11} & 0 & \Delta_{13}^\top \\ 0 & \Delta_{22} & 0 \\ \Delta_{13} & 0 & \Delta_{33} \end{pmatrix}, \end{aligned}$$

where $\Delta_{11} = T^{2\star}1_q^\top M 1_q/(\sigma^{2\star}\omega^2)$, $\Delta_{13} = -(m+1)T^\star M 1_q/\omega$, $\Delta_{33} = (m+1)^2\sigma^{2\star}M$, and $\Delta_{22} = S_Z\left(\sigma^{2\star}S_Z + \sigma^{4\star}m(T^\star)^{-1} + \frac{\sigma^{4\star}}{\omega}I_q\right)^{-1}S_Z$ with $M = \left(T^{3\star}S_Z + (2m - q + \frac{1}{\omega})\sigma^{2\star}T^{2\star}\right)^{-1}$. It is clear that $\Delta^{(2)}_{\text{EM}}(\omega)$ is singular and that both Δ_{22} and Δ_{33} are

increasing in ω, but Δ_{11} is decreasing in ω (and thus, as a whole, $\Delta^{(2)}_{\text{EM}}(\omega)$ does not admit the positive semidefinite ordering). □

The optimality of using $\omega = \infty$ is further supported by the following result, which suggests that the algorithms in Section 8.1 are superior, in terms of the Δ quantities, to all the algorithms described in Table 2.

Result A.3 Under the assumptions of Result A.1, asymptotically

$$\lim_{\omega \to \infty} \Delta^{(2)}_{\text{EM}}(\omega) - \Delta^{(1)}_{\text{EM}}(\alpha) \geq 0, \tag{A.17}$$

for any α such that $(\alpha_{1j}, \alpha_{2j}) \in \mathcal{A}_0$ for each j. The diagonal terms of (A.17) are positive for any $(\alpha_1, \alpha_2) \in R^{2q}$.

Proof: Since Δ_{11} and Δ_{13} are both zero in the limit as $\omega \to \infty$, we need only compare the diagonal matrices, Δ_{22} and Δ_{33}, with the corresponding terms of $\Delta^{(1)}_{\text{EM}}(\alpha)$, which are also asymptotically diagonal if $(\alpha_{1j}, \alpha_{2j}) \in \mathcal{A}_0$ for each j. Thus, we need only show the diagonal terms of (A.17) are positive for all $(\alpha_1, \alpha_2) \in R^{2q}$. We first compare Δ_{22} with (A.11). By minimizing (A.11) over (α_1, α_2) we find that the optimal value of $A_2 = \text{diag}(\alpha_2)$ is $(S_Z/\sigma^{2\star})(S_Z/\sigma^{2\star} + m(T^\star)^{-2})$ for any $\alpha_1 \in R^q$ and that the resulting minimally augmented information for ξ is

$$\frac{1}{\sigma^{2\star}} S_Z - S_Z[\sigma^{2\star} S_Z + \sigma^{4\star} m(T^\star)^{-1}]^{-1} S_Z, \tag{A.18}$$

where the first term is the augmented information for ξ when using the standard algorithm (i.e., using (8.10) with $\alpha = \alpha_{\text{std}}$). Comparing the second term in (A.18) to Δ_{22} (with $\omega = \infty$) we see that Δ_{22} equals the reduction in augmented information for ξ resulting from using the optimal conditional augmentation scheme determined by the optimal value of A_2. (In fact, it can be shown that Δ_{22} corresponds to incorporating the working parameter γ alone. In particular, the introduction of γ results in an algorithm that is more efficient than using $\alpha = \alpha_{\text{gsc}}$ which fixes $\alpha_1 = 0_q$.)

Similar arguments can be made to show that Δ_{33} is larger than the reduction in (A.15) resulting from a similarly defined optimal value of α_1 when α_2 if fixed at 0_q. (If α_2 is fixed at some other value the reduction in (A.15) will be still smaller.) Thus, the result follows since $(\alpha_{1j}, \alpha_{2j}) \in \mathcal{A}_0$ for each j, $\Delta^{(1)}_{\text{EM}}(\alpha)$ is asymptotically diagonal (see (A.11)–(A.13)), and the reduction in the diagonal terms of $I^{(1)}_{\text{aug}}(\alpha_{\text{std}})$ is dominated by Δ_{22} and Δ_{33}. □

ACKNOWLEDGMENTS

Van Dyk's research was supported in part by NSF grant DMS 97-05156 and in part by the U.S. Census Bureau. Meng's research was supported in part by NSF grant DMS 96-26691 and in part by NSA grant MDA 904-9610007. We thank Andreas Buja for the prompt handling of our article, A. Gelman, C. Liu, J. Liu, and Y. Wu for helpful exchanges; J. Servidea for proofreading; and a total of ten editors and anonymous AEs and referees for stimulating comments. We also thank M. Hass for permission of using his data. Some of the computer programs and data are available at the URL www.fas.harvard.edu/~vandyk and others are available upon request (E-mail: vandyk@stat.harvard.edu).

[Received December 1999. Revised April 2000.]

REFERENCES

Albert, J. H. (1992), "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling," *Journal of Educational Statistics*, 17, 251–269.

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Amit, Y. (1991), "On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions," *Journal of Multivariate Analysis*, 38, 82–89.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society*, Ser. B, 55, 25–37.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, New York: Chapman and Hall.

Casella, G. (1996), "Statistical Inference and Monte Carlo Algorithms" (with discussion), *Test*, 5, 249–344.

Ceperley, D. M. (1995), "Path Integrals in the Theory of Condensed Helium," *Reviews of Modern Physics*, 67, 279–355.

Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904.

Damien, P., Wakefield, J., and Walker, S.(1999), "Gibbs Sampling for Bayesian Nonconjugate and Hierarchical Models Using Auxiliary Variables," *Journal of the Royal Statistical Society*, Ser. B, 61, 331–344.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Edwards, R. G., and Sokal, A. D. (1988), "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang Representation and Monte Carlo Algorithm," *Physical Review Letters*, 38, 2009–2012.

Fessler, J. A., and Hero, A. O. (1994), "Space-Alternating Generalized Expectation-Maximization Algorithm," *IEEE Transactions on Signal Processing*, 42, 2664–2677.

Foulley, J. L., and van Dyk, D. A. (2000), "The PX-EM Algorithm for Fast Stable Fitting of Henderson's Mixed Model," *Genetics Selection Evolution*, 32, 143–163.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parameterization for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–472.

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

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds) (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Gray, A. J. (1993), "Discussion of the Gibbs Sampler and Other Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 58–61.

Green, P. (1997), Discussion of "The EM Algorithm—An Old Folk Song Sung to a Fast New Tune," by Meng and van Dyk, *Journal of the Royal Statistical Society*, Ser. B, 59, 554–555.

Haas, M. (1994), "IgG Subclass Deposits in Glomeruli of Lupus and Nonlupus Membranous Nephropathies," *American Journal of Kidney Disease*, 23, 358–364.

——— (1998), "Value of IgG Subclasses and Ultrastructural Markers in Predicting Latent Membranous Lupus Nephritis," *Modern Pathology*, 11, 147A.

Higdon, D. M. (1993), "Discussion of the Gibbs Sampler and Other Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 78.

——— (1998), "Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications," *Journal of the American Statistical Association*, 93, 585–595.

Hobert, J. P. (in press), "Stability Relationships Among the Gibbs Sampler and its Subchains," *Journal of Computational and Graphical Statistics*.

Hobert, J. P., and Casella, G. (1996), "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association*, 91, 1461–1473.

——— (1998), "Functional Compatibility, Markov Chains, and Gibbs Sampling With Improper Posteriors," *Journal of Computational and Graphical Statistics*, 7, 42–60.

Ising, E. (1925), "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift fur Physik*, 31, 253–258.

Laird, N., Lange, N., and Stram, D. (1987), "Maximizing Likelihood Computations With Repeated Measures: Application of the EM Algorithm," *Journal of the American Statistical Association*, 82, 97–105.

Laird, N. M., and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 967–974.

Lange, N., Carlin, B. P., and Gelfand, A. E. (1992), "Hierarchical Bayes Models for the Progression of HIV Infection Using Longitudinal CD4 T-cells Numbers," *Journal of the American Statistical Association*, 87, 615–632.

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measure Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Liu, C. (1999), "Covariance Adjustment for Markov Chain Monte Carlo—A General Framework and the Covariance-Adjusted Data Augmentation Algorithm," Technical Report, Bell Labs, Lucent Technologies.

Liu, C., and Rubin, D. B. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM With Fast Monotone Convergence," *Biometrika*, 81, 633–648.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM—The PX-EM Algorithm," *Biometrika*, 85, 755–770.

Liu, J. S. (1994), "Fraction of Missing Information and Convergence Rate of Data Augmentation," in *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface*. eds J. Sall and A. Lehman, Fairfax Station, VA: Interface Foundation, pp. 490–497.

Liu, J. S., and Sabatti, C. (2000), "Generalized Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation," *Biometrika*, 87, 353–369.

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

Liu. J. S., and Wu, Y. N. (1999), "Parameter Expansion Scheme for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

McCulloch, R., and Rossi, P. (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207–240.

Meng, X.-L. (1994), "On the Rate of Convergence of The ECM Algorithm," *The Annals of Statistics*, 22, 326–339.

Meng, X.-L., and Rubin, D. B. (1994), "On the Global and Component-Wise Rates of Convergence of the EM Algorithm," *Linear Algebra and its Applications*, 199, 413–425.

Meng, X.-L., and Schilling, S. (1996), "Fitting Full-Information Factor Models and an Empirical Investigation of Bridge Sampling," *Journal of the American Statistical Association*, 91, 1254–1267.

Meng, X.-L., and van Dyk, D. A. (1997), "The EM Algorithm—An Old Folk Song Sung to a Fast New Tune" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 511–567.

——— (1998), "Fast EM Implementations for Mixed-Effects Models," *Journal of the Royal Statistical Society*, Ser. B., 60, 559–578.

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320.

Mira, A., and Tierney, L. (1997), "On the Use of Auxiliary Variables in Markov Chain Monte Carlo Sampling," Technical Report, University of Minnesota, School of Statistics.

Nachbin, L. (1965), *The Haar Integral*, Princeton: Van Nostrand.

Neal, R. M. (1993), *Probabilistic Inference Using Markov chain monte Carlo Methods*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 140 pages.

——— (1997), "Markov Chain Monte Carlo Methods Based on 'Slicing' the Density Function," Techincal Report No. 9722, Department of Statistics, University of Toronto.

Pawitan, Y. (2000), "A Reminder of the Fallibility of the Ward Statistic: Likelihood Explanation," *The American Statistician*, 54, 54–56.

Pierson, R. A., and Ginther, O. J. (1987), "Follicular Population Dynamics During the Estrus Cycle of the Mare," *Animal Reproduction Science*, 14, 219–231.

Potts, R. B. (1952), "Some Generalized Order-Disorder Transformations," *Proceedings of the Cambridge Philosophic Society*, 48, 106–109.

Roberts G. O. (1996), "Markov Chain Concepts Related to Sampling Algorithms," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 45–58.

Roberts, G. O., and Rosenthal, J. S. (1997), "Convergence of Slice Sampler Markov Chains," Technical Report, Cambridge University, Statistical Laboratory.

Rubin, D. B. (1983), "Iteratively Reweighted Least Squares," in *Encyclopedia of Statistical Sciences*, 4, eds. S. Kotz, N. L. Johnson, and C. B. Reed, New York: Wiley, pp. 272–275.

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

Speckman, P. L., Lee, J., and Sun, D. (1999), "Existence of the MLE and Propriety of Posteriors for a General Multinomial Choice Model," Technical Report, Department of Statistics, The University of Missouri-Columbia.

Sun, D., Tsutakawa, R. K., and He, Z. (1997), "Propriety of Posteriors With Improper Priors in Hierarchical Linear Mixed Models," Technical Report, Department of Statistics, The University of Missouri-Columbia.

Swendsen, R. H., and Wang, J. S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters*, 58, 86–88.

Tanner, M. A. (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

——— (1996), "Introduction to General State-Space Markov Chain Theory," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 59–74.

van Dyk, D. A. (2000), "Fitting Mixed-Effects Models Using Efficient EM-Type Algorithms," *Journal of Computational and Graphical Statistics*, 9, 78–98.

van Dyk, D. A., and Meng, X.-L. (1997), "On the Orderings and Groupings of Conditional Maximizations Within ECM-Type Algorithms," *Journal of Computational and Graphical Statistics*, 6, 202–223.