Statistica Sinica **20** (2010), 1423-1454

MARGINAL MARKOV CHAIN MONTE CARLO METHODS

David A. van Dyk

University of California, Irvine

Abstract: Marginal Data Augmentation and Parameter-Expanded Data Augmentation are related methods for improving the convergence properties of the two-step Gibbs sampler known as the Data Augmentation sampler. These methods expand the parameter space with a so-called working parameter that is unidentifiable given the observed data but is identifiable given the so-called augmented data. Although these methods can result in enormous computational gains, their use has been somewhat limited due to the constrained framework they are constructed under and the necessary identification of a working parameter. This article proposes a new prescriptive framework that greatly expands the class of problems that can benefit from the key idea underlying these methods. In particular, we show how working parameters can automatically be introduced into any Gibbs sampler, and explore how they should be updated vis-à-vis the updating of the model parameters in order to either *fully or partially marginalize* them from the target distribution. A prior distribution is specified on the working parameters and the convergence properties of the Markov chain depend on this choice. Under certain conditions the optimal choice is improper and results in a non-positive recurrent joint Markov chain on the expanded parameter space. This leads to unexplored technical difficulties when one attempts to exploit the computational advantage in multi-step MCMC samplers, the very chains that might benefit most from this technology. In this article we develop strategies and theory that allow optimal marginal methods to be used in multi-step samplers. We illustrate the potential to dramatically improve the convergence properties of MCMC samplers by applying the marginal Gibbs sampler to a logistic mixed model.

Key words and phrases: Conditional data augmentation, Gibbs sampler, logistic mixed model, marginal data augmentation, MCMC, mixing rate, non-positive recurrent Markov chain, partial marginalization, stationary distribution, working parameters.

1. Expanding State Spaces in MCMC

Constructing a Markov chain on an expanded state space in the context of Monte Carlo sampling can greatly simplify the required component draws or lead to chains with better mixing properties. In a Bayesian context, we aim to construct a Markov chain with stationary distribution

$$p(\psi \mid Y) = \int p(\psi, \alpha \mid Y) d\alpha \propto \int p(Y \mid \psi, \alpha) p(\psi, \alpha) d\alpha, \tag{1.1}$$

where ψ is a vector of unobserved quantities of interest, perhaps including model parameters, latent variables, or missing data; Y represents observed or fixed quantities including the observed data; and α represents unobserved quantities introduced into the model for computational reasons.

In the statistical literature, the oldest and best known example of an expanded state space Markov chain Monte Carlo (MCMC) sampler is the data augmentation (DA) sampler (Tanner and Wong (1987)). The DA sampler introduces latent variables or missing data into the model, which are expressed as α in (1.1). In its simplest form, DA oscillates between sampling $\psi \sim p(\psi \mid Y, \alpha)$ and $\alpha \sim p(\alpha \mid Y, \psi)$. This is advantageous when the conditional draws are simple but sampling $p(\psi \mid Y)$ directly is complex or impossible under practical constraints. DA samplers can exhibit poor mixing and their traditional motivation is computational simplicity rather than speed.

Although equivalent to the DA sampler in its mathematical form, the use of auxiliary variables, is generally motivated by computational speed rather than simplicity (Edwards and Sokal (1988); Besag and Green (1993); Higdon (1998)). In particular, consider a situation where again $p(\psi \mid Y, \alpha)$ and $p(\alpha \mid Y, \psi)$ are easy to sample, but $p(\psi \mid Y)$ is not. In some cases, a Gibbs sampler can be used to sample $p(\psi \mid Y)$ by splitting ψ into a number of subcomponents. Like the DA sampler, the Gibbs sampler can exhibit poor mixing if the components of ψ are highly correlated. If the computational cost of conditioning on α is offset by the gain that stems from blocking ψ into one conditional step, the DA sampler is attractive for its speed. This is the situation that motivated auxiliary-variable methods such as the slice sampler (Neal (1997)). The slice sampler can be formulated as a special case of the DA sampler in which the target distribution can be written as

$$p(\psi \mid Y) = \pi(\psi \mid Y) \prod_{i=1}^{n} l_i(\psi \mid Y), \tag{1.2}$$

where any of the factors on the right-hand side might not depend on Y. The model is expanded via $p(\alpha \mid Y, \psi)$, a uniform distribution on the rectangle $[\{0, l_1(\psi \mid Y)\}, \cdots, \{0, l_n(\psi \mid Y)\}]$, in which case $p(\alpha \mid Y, \psi)$ is easy to sample and $p(\psi \mid Y, \alpha)$ may be easy to sample directly or via some Markov chain technique. Although this method is formally a special case of DA, it is more prescriptive than DA and has led to many useful samplers. In fact, Damien, Wakefield, and Walker (1999) used its easy implementation in a large class of application as the primary motivation for the method.

In this paper we focus on another strategy based on a special case of (1.1), namely, the method of working parameters. After being introduced by Meng and van Dyk (1997) and Liu, Rubin, and Wu (1998) in the context of the EM algorithm, working parameters have were used to improve the convergence of DA

samplers by Liu and Wu (1999) and Meng and van Dyk (1999). In a given model, a working parameter is not part of the standard model formulation and is not sampled in a typical DA sampler. Thus, if we let α represent the working parameter, the target posterior distribution is $p(\psi \mid Y, \alpha)$. Conditional augmentation methods aim to find the optimal value of α in terms of the rate of convergence of the resulting sampler. Marginal data augmentation (MDA), on the other hand, constructs Markov chains on the expanded parameter space of (ψ, α) and effectively marginalize out α by updating it in both steps. This results in draws from the marginal distribution $p(\psi \mid Y) = \int p(\psi, \alpha \mid Y) d\alpha$. To be sure that the likelihood is unaffected, it is required that α be unidentifiable in (1.1), i.e., $p(Y \mid \psi, \alpha) = p(Y \mid \psi)$. Thus, the first step in implementing these methods is the sometimes subtle task of finding a suitable working parameter—a task that we aim to simpify. Once we have the expanded model, however, two-step samplers based on a transformation of (ψ, α) can exhibit significant computational advantage, see Section 2.1. For example, new samplers for probit regression, multinomial probit models, t-models, random-effects models, and factor analysis illustrate the potential for marginal methods to dramatically improve the convergence of DA samplers. (Meng and van Dyk (1999); Liu and Wu (1999); van Dyk and Meng (2001); Imai and van Dyk (2005a,b); Gelman et al. (2008); Ghosh and Dunson (2009)).

Marginal MCMC methods construct a Markov chain with stationary distribution $p(\psi, \alpha \mid Y) \propto p(Y \mid \psi, \alpha) p(\psi) p(\alpha) = p(Y \mid \psi) p(\psi) p(\alpha) \propto p(\psi|Y) p(\alpha)$, where α is a working parameter. (More generally, we discuss replacing $p(\alpha)$ with $p(\alpha \mid \psi)$.) Because there is much more flexibility in the construction of MCMC samplers than in DA samplers, there is much more flexibility in how α is updated. One of the goals of this paper is to describe the possibilities and give general advice on efficient updating schemes. Because it is unidentified and introduced purely for computational reasons, we can choose the prior distribution on α to improve computation. Sometimes the optimal sampler in terms of the convergence of ψ occurs when $p(\alpha)$ is improper. Because α is unidentifiable, $p(\psi, \alpha \mid Y)$ is improper if $p(\alpha)$ is. Thus, although certain subchains may have the desired stationary distribution in this case, the resulting joint chain may not be positive recurrent since it has no (proper) stationary distribution. In particular, the subchain for ψ may not have the correct stationary distribution (Meng and van Dyk (1999)). Although these difficulties have been discussed for two-step samplers (Meng and van Dyk (1999); Liu and Wu (1999)), they have not yet been explored in multi-step samplers, where their potential for improving mixing is most needed. The primary aim of this article is to develop and illustrate theory and methods that allow these powerful techniques to be easily applied in complex MCMC samplers.

The article is organized into five sections. Section 2 gives background material on MDA and on a logistic mixed model used as a running example to illustrate marginal methods and their significant computational advantage. Sections 3 introduces partially marginalized Gibbs samplers and more general marginal MCMC methods. Theoretical results appear in Section 4. We conclude with a brief discussion in Section 5. An appendix gives technical details and supplemental illustrations.

2. Background

2.1. A simplified formulation of marginal data augmentation

Meng and van Dyk (1999) introduced MDA to improve performance of the two-step DA sampler; see Liu and Wu (1999) for a similar formulation of many of the same ideas. Here we introduce a simpler but more prescriptive formulation that can easily be generalized to multi-step Gibbs samplers and more general MCMC samplers. Starting with a target posterior distribution, $p(\psi \mid Y)$, with $\psi = (\psi_1, \psi_2)$, we introduce a working parameter α and define the joint posterior distribution

$$p(\psi, \alpha \mid Y) = p(\psi \mid Y)p(\alpha). \tag{2.1}$$

This joint model is always easy to specify and ensures that the working parameter is unidentifiable given the observed data. Because ψ and α are a posterori independent, we must introduce a joint transformation of ψ and α to construct a DA sampler that is substantively affected by the working parameter. To do this, we define a transformation of ψ_1 that is indexed by the working parameter α, $\tilde{\psi}_1 = \mathcal{D}_\alpha(\psi_1)$, where $\mathcal{D}_\alpha$ is an invertible and differentiable mapping and there exists a_0 such that $\mathcal{D}_{a_0}$ is an identity mapping. We consider constructing samplers of either the joint posterior distribution $p(\tilde{\psi}_1, \psi_2, \alpha \mid Y)$ or the marginal posterior distribution $p(\tilde{\psi}_1, \psi_2 \mid Y) = \int p(\tilde{\psi}_1, \psi_2, \alpha \mid Y)d\alpha$. These two strategies are called joint augmentation and marginal augmentation (van Dyk and Meng (2000, 2010)). Because we introduce a more general strategy that encompasses both, however, we blur the distinction and refer to these techniques generally as MDA. In particular, consider

Scheme 0: Sample $\tilde{\psi}_1 \sim p(\tilde{\psi}_1 \mid Y, \psi_2, \alpha = a_0)$ and $\psi_2 \sim p(\psi_2 \mid Y, \tilde{\psi}_1, \alpha = a_0)$.

MDA Scheme 1: Sample $(\tilde{\psi}_1, \alpha) \sim p(\tilde{\psi}_1, \alpha \mid Y, \psi_2)$ and $(\psi_2, \alpha) \sim p(\psi_2, \alpha \mid Y, \tilde{\psi}_1)$.

MDA Scheme 2: Sample $\tilde{\psi}_1 \sim p(\tilde{\psi}_1 \mid Y, \psi_2, \alpha)$ and $(\psi_2, \alpha) \sim p(\psi_2, \alpha \mid Y, \tilde{\psi}_1)$.

MDA Scheme 3: Sample $\tilde{\psi}_1 \sim p(\tilde{\psi}_1 \mid Y, \psi_2, \alpha)$, $\psi_2 \sim p(\psi_2 \mid Y, \tilde{\psi}_1, \alpha)$, and $\alpha \sim p(\alpha \mid Y, \tilde{\psi}_1, \psi_2)$.

When $\alpha = a_0$, $\tilde{\psi} = \psi$ so that SCHEME 0 is simply the standard DA sampler. By sampling α rather than conditioning on it in each step, SCHEME 1 constructs a *marginal* Markov chain on $(\tilde{\psi}_1, \psi_2)$ with stationary distribution $p(\tilde{\psi}_1, \psi_2 \mid Y) = \int p(\tilde{\psi}_1, \psi_2, \alpha \mid Y) d\alpha$. When this is done in some but not all steps (e.g., SCHEME 2), we use the term *partial marginalization*. Because it fully marginalizes out α, SCHEME 1 can be written: sample $\tilde{\psi}_1 \sim p(\tilde{\psi}_1, \mid Y, \psi_2)$ and $\psi_2 \sim p(\psi_2, \mid Y, \tilde{\psi}_1)$. Although this bares a striking resemblance to SCHEME 0, we illustrate how the introduction of α, transformation, and marginalization can dramatically improve convergence. An advantage of sampling α along with $(\tilde{\psi}_1, \psi_2)$ is the ability to transform to obtain a sample of (ψ_1, ψ_2).

Gaussian Example. We begin with a simple illustrative example that we return to several times to clarify ideas. This is not meant to introduce useful new samplers but rather to illustrate subtle features of the methods in a concrete example. We suppose ψ follows a bivariate Gaussian distribution with mean $\mu = (\mu_1, \mu_2)$ and variance Σ, and that we would like to sample $p(\psi \mid \mu, \Sigma)$. We parameterize Σ in terms of the marginal variances and correlation, i.e., σ_1^2, σ_2^2, and ϱ, respectively, and we introduce a scalar working parameter, α, with working prior distribution, $\alpha \sim \mathrm{N}(0, \omega^2 \sigma_1^2)$, independent of ψ. We could use any working prior distribution; we use this distribution to facilitate simple sampling. The working parameter enters the Gibbs sampler via the transformation $\tilde{\psi} = (\tilde{\psi}_1, \tilde{\psi}_2) = (\psi_1 + \alpha, \psi_2) = \mathcal{D}_\alpha(\psi)$; clearly $a_0 = 0$. We can easily compute the Gaussian joint distribution, $p(\tilde{\psi}, \alpha)$, and all of the relevant conditional and marginal distributions. The lag-one autocorrelation for ψ_2 is ϱ^2 for SCHEMES 0 and 3 and is $\varrho^2/(1+\omega^2)$ for SCHEMES 1 and 2. The first four rows of Figure 1 illustrate the convergence of the four sampling schemes, each run with $\varrho = 0.95$ and $\omega^2 = 25$, and the advantage of SCHEMES 1 and 2. (Appendix A gives an illustrative example of the advantage of SCHEME 1 in exploring multi-modal distributions.)

The example illustrates the advantage of the simplified formulation given in (2.1). The working parameter, α, is not specified as an unidentifiable parameter that is identifiable given the augmented data. Indeed there is no data augmentation in this simple example.

The marginal chain for ψ_2 under both SCHEMES 0 and 1 is Markovian, and Meng and van Dyk (1999) showed the geometric rate of convergence of this marginal chain under SCHEME 1 dominates that of SCHEME 0. Moreover, while SCHEMES 1 and 2 have the same lag-one autocorrelation for linear combinations of ψ_2, the geometric rate of convergence of SCHEME 1 can be no worse than that of SCHEME 2 for the joint Markov chain; see Marchev and Hobert (2004) for a detailed analysis of the geometric convergence of marginal augmentation

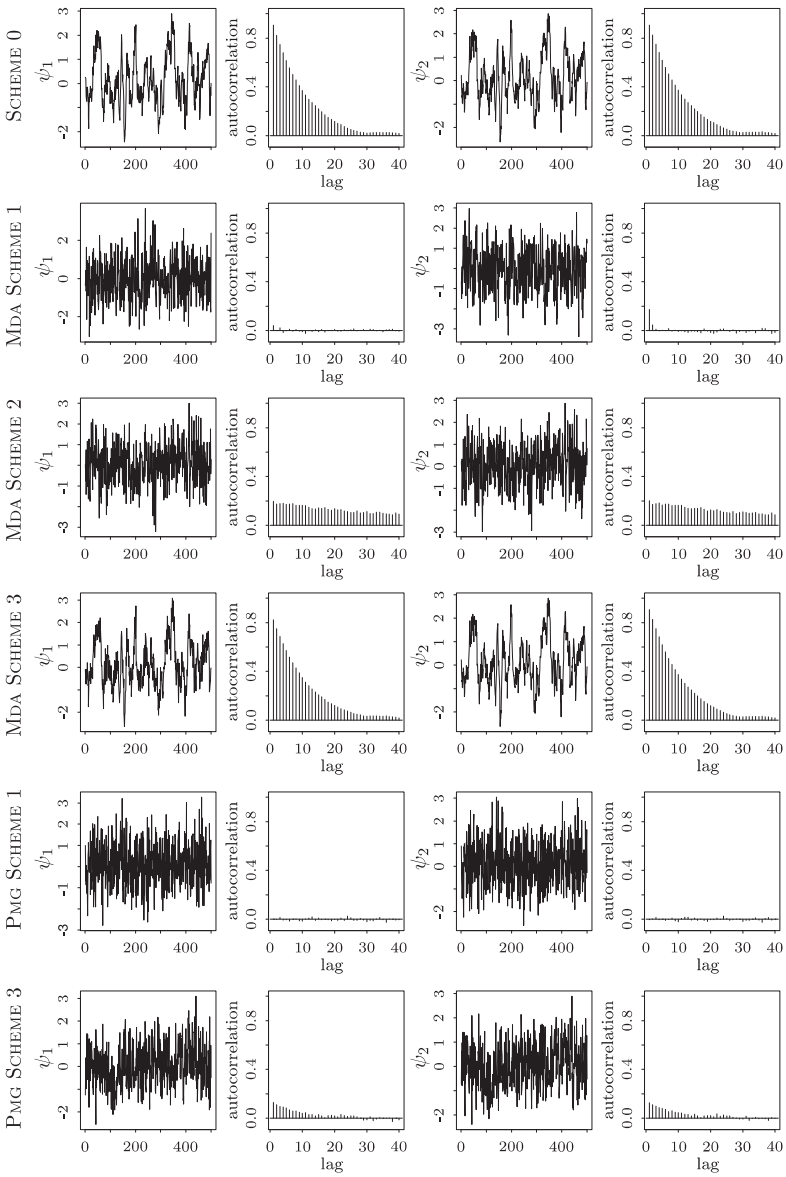

Figure 1. Six Sampling Schemes for the Gaussian Example. The figure gives time series plots and autocorrelation plots for ψ_1 and ψ_2 for each of six sampling schemes. MDA SCHEMES 1 and 2 outperform MDA SCHEMES 0 and 3, which are indistinguishable. (Although the slowly diminishing high-order autocorrelations of Scheme 2 make comparisons difficult, these correlations are all reduced as ω^2 increases.) Although MDA SCHEMES 1 and 2 have the same lag-one autocorrelation for ψ_2, MDA SCHEME 1 is clearly superior. By adding a second working parameter, PMG SCHEME 1 improves the convergence of MDA SCHEME 1; although MDA SCHEME 1 performs very well, PMG SCHEME 1 produces essentially independent draws of ψ_2. Finally, in contrast to MDA SCHEME 3, PMG SCHEME 3 offers marked improvement over SCHEME 0, illustrating the potential benefit of the optional steps.

in a two-step example. Because SCHEME 2 can be viewed as a blocked version of SCHEME 3, we expect it to preform better. The key to these results is the basic observation that less conditioning in any step of a Gibbs sampler tends to improve the overall convergence of the sampler; see van Dyk and Park (2008) for discussion of this principle in problems that do not involve working parameters. SCHEME 2 eliminates the conditioning on α in the second step of SCHEME 0 and thus improves convergence. SCHEME 1 further improves convergence by eliminating the conditioning on α in the first step. We aim to exploit this basic observation in multi-step samplers involving working parameters.

The computational advantage of all MDA methods relies on the ability to *jointly* sample components of the working parameter and of the transformed model parameters. If this were not possible, the advantage of MDA would vanish in the Gaussian example, see SCHEME 3. This key observation is of central importance when selecting the transformation $\mathcal{D}_\alpha$. While any transformation can in principle improve convergence, the improvement may not be realized if the working parameters cannot be at least partially marginalized through joint draws with components of the transformed model parameter; see, however, the discussion and Gaussian example in Section 4.1. In this regard selecting the transformations and working parameters is akin to selecting the partition of the model parameter used to construct a Gibbs sampler. The goal in both cases is a set of complete conditional distributions that can be easily sampled.

For SCHEMES 1 and 2 in the Gaussian example, the lag-one autocorrelation goes to zero as ω^2 goes to infinity. We generally expect more diffuse working prior distributions to result in samplers that mix better. In this, as in many examples, the optimal choice of prior distribution on α is improper. This complicates the situation because the joint target distribution, $p(\psi, \alpha \mid Y) = p(\psi \mid Y)p(\alpha)$, is also improper. Indeed, the first draw of SCHEME 1 is improper and under SCHEME 2, the joint Markov chain, $\mathcal{M}(\psi, \alpha) = \{(\psi^{(t)}, \alpha^{(t)}), t = 1, 2, \ldots\}$ is not positive recurrent; we use the notation $\mathcal{M}(x)$ for the chain $\{x^{(t)}, t = 1, 2, \ldots\}$. Meng and van Dyk (1999) showed, however, that in some cases the marginal chain for a component of ψ may still be positive recurrent with the corresponding marginal distribution as its stationary distribution. We generalize these results to multi-step chains in Section 4.2.

2.2. Motivating example: logistic mixed model

The MDA methods described in Section 2.1 are designed for two-step DA samplers. In this section, we introduce a multi-step slice sampler to motivate the extension of marginal methods to more complex and realistic situations. Consider the logistic mixed model,

$$y_{ij} \sim \text{Bernoulli}(p_{ij}) \text{ with } \text{logit}(p_{ij}) = x'_{ij}(\beta + b_i), \tag{2.2}$$

where y_{ij} is the binary response of unit j within group i, p_{ij} is the probability that y_{ij} equals one rather than zero, x_{ij} is a $q \times 1$ vector of observed covariates, β is a $q \times 1$ vector of fixed effects, and b_i is a $q \times 1$ vector of random effects, with $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, i.e., there are m groups with sizes $n_1, \ldots, n_m$. We assume the random effects are independently distributed, $b_i \sim N(0, T)$, with T a diagonal matrix with diagonal elements $(\tau_1^2, \ldots, \tau_q^2)$ and independent prior distributions $\tau_k^2 \sim \nu_0 \tau_{k,0}^2 / \chi_{\nu_0}^2$; we use a flat prior distribution on the fixed effects.

We focus on the posterior distribution of the unknown parameters (b, β, T) with $b = (b_1, \ldots, b_m)$,

$$p(b, \beta, T \mid Y) \propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left[\frac{\exp\{x_{ij}'(b_i + \beta)\}}{\exp\{x_{ij}'(b_i + \beta)\}+1} \right]^{y_{ij}} \left[1 - \frac{\exp\{x_{ij}'(b_i + \beta)\}}{\exp\{x_{ij}'(b_i + \beta)\}+1} \right]^{1-y_{ij}}$$
$$\times \prod_{i=1}^{m} |T|^{-1/2} \exp\left(-\frac{1}{2} b_i' T^{-1} b_i \right) \times |T|^{-(\nu_0/2+1)} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\nu_0 T_0 T^{-1}) \right\},$$

where $Y = (y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i)$. Since the posterior distribution is not a standard density function, it is common practice to use MCMC methods. We begin with an outline of a slice sampler based on the data augmentation scheme of Damien, Wakefield, and Walker (1999).

Following Damien, Wakefield, and Walker (1999), we suppose

$$y_{ij} = I\left[v_{ij} \leq g^{-1}\{x_{ij}'(\beta + b_i)\} \right], \tag{2.3}$$

where I is an indicator function, g^{-1} is the inverse logistic link function, and $v_{ij} \sim \mathrm{Unif}(0, 1)$. Model (2.3) reformulates (2.2) using the auxiliary variable $V = (v_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i)$; (2.2) represents $p(Y \mid \beta, b) = \int p(Y, V \mid \beta, b) dV$ and (2.3) represents $p(Y \mid V, \beta, b)$. Using (2.3) we construct:

Slice Sampler for the Logistic Mixed Model

Step 1: For each i and j, independently draw

$$v_{ij} \mid Y, b, \beta, T \sim \begin{cases} \mathrm{Unif}\left[0,\ g^{-1}\{x_{ij}'(\beta + b_i)\} \right] & \text{if } y_{ij} = 1 \\ \mathrm{Unif}\left[g^{-1}\{x_{ij}'(\beta + b_i)\},\ 1 \right] & \text{if } y_{ij} = 0. \end{cases} \tag{2.4}$$

Step 2: For $k = 1, \ldots, q$, sample $(b_{,k}, \beta_k, \tau_k^2)$ via two conditional draws, where $b_{,k} = (b_{1k}, \ldots, b_{mk})$; here b_{ik} and β_k represent component k of b_i and β. Although it is suppressed in the notation, we condition on Y, V, and all of the components except the kth of β and each b_i.

Cycle *k*: For $k = 1, \ldots, q$ repeat the following two substeps.

Substep 1: For each i, independently draw $b_{ik} \mid \beta_k, \tau_k^2 \sim N(0, \tau_k^2)$ subject to the constraint

$$\begin{cases} v_{ij} \le g^{-1}\{x'_{ij}(\beta + b_i)\} & \text{if } y_{ij} = 1 \\ v_{ij} > g^{-1}\{x'_{ij}(\beta + b_i)\} & \text{if } y_{ij} = 0, \end{cases} \quad \text{for} \quad j = 1, \ldots, n_i. \tag{2.5}$$

Substep 2: Draw $(\beta_k, \tau_k^2) \mid b_{,k}$ by independently drawing β_k given $b_{,k}$ uniformly subject to (2.5) and $\tau_k^2 \mid b_{,k} \sim (\nu_0 \tau_{k,0}^2 + \sum_{i=1}^m b_{ik})/\chi_{m+\nu_0}^2$.

This sampler consists of $2q + 1$ complete conditional draws; see Damien, Wakefield, and Walker (1999). To improve its convergence using marginal methods we must extend MDA. This is the topic of Section 3.

3. Marginal MCMC Methods

3.1. Model expansion

Suppose we wish to obtain a Monte Carlo sample from $p(\psi \mid Y)$, where $\psi = (\psi_1, \ldots, \psi_P)$ and each ψ_p may be multivariate. To focus attention on the working parameters, in the remainder of the paper we use ψ to represent all unobserved quantities except for the vector working parameter α. Thus, missing data and auxiliary variables are treated as components of ψ. We can construct a 'standard' Gibbs sampler beginning with an initial $\psi^{(0)}$, by iterating

Step 1: $\psi_1 \sim p(\psi_1 \mid Y, \psi_{-1})$,

$\vdots$

Step *P*: $\psi_P \sim p(\psi_P \mid Y, \psi_{-P})$,

where ψ_{-p} denotes $(\psi_1, \ldots, \psi_{p-1}, \psi_{p+1}, \ldots, \psi_P)$. We assume the standard Gibbs-sampler regularity conditions (Roberts (1996); Tierney (1994, 1996)) so the limiting distribution of $\psi^{(t)}$ is $p(\psi \mid Y)$.

We introduce a working parameter by expanding the target posterior distribution, $p(\psi \mid Y)$ to

$$p(\psi, \alpha \mid Y) = p(\psi \mid Y) p(\alpha \mid \psi), \tag{3.1}$$

where $p(\alpha \mid \psi)$ is a prior distribution on α. The conditional independence assumed in (3.1) assures that α is a working parameter, i.e., $p(Y \mid \psi, \alpha) = p(Y \mid \psi)$. To construct a sampler we introduce a transformation of ψ_p that depends on α for each p: $\tilde{\psi}_p = \mathcal{D}_{\alpha,p}(\psi_p)$, where each $\mathcal{D}_{\alpha,p}$ is again an invertible and differentiable mapping and there exists a_0 such that each $\mathcal{D}_{a_0,p}$ is an identity mapping. Algorithms are constructed by sampling from a set of conditional distributions of $p(\tilde{\psi}, \alpha \mid Y)$, with $p(\tilde{\psi}, \alpha \mid Y)$ obtained with a change of variable of (3.1). As with MDA the primary goal when selecting the transformations, is to be sure that components of α and ψ can be jointly updated.

The fact that we do not require $p(\psi, \alpha) = p(\psi)p(\alpha)$ as in (2.1) has implications for statistical inference. In particular, the marginal posterior distribution of ψ,

$$p(\psi \mid Y) \propto p(Y \mid \psi) \int p(\psi, \alpha) d\alpha, \tag{3.2}$$

and the conditional posterior distribution of ψ given α,

$$p(\psi \mid Y, \alpha) \propto p(Y \mid \psi) p(\psi \mid \alpha), \tag{3.3}$$

may differ because of their respective prior distributions on ψ. Thus, samplers that explicitly or implicitly condition on α (or part of α) throughout the iteration have a different stationary distribution than those that sample α along with ψ. Although either of these distributions may be the target distribution, the goal is to construct samplers of $p(\psi \mid Y)$ with much better convergence properties than the corresponding samplers of $p(\psi \mid Y, \alpha)$. Of course if ψ and α are *a priori* independent, $p(\psi \mid Y) = p(\psi \mid Y, \alpha)$, and, thus we often assume such independence. In some cases, however, we can formulate the desired prior distribution on ψ as the corresponding marginal distribution of $p(\psi, \alpha)$; see McCulloch and Rossi (1994); Nobile (1998); Imai and van Dyk (2005a).

3.2. Partially marginalized Gibbs samplers

Generalizing MDA, we can construct a marginal Gibbs (MG) sampler using conditional distributions of the marginal distribution

$$p(\tilde{\psi}|Y) = \int p(\tilde{\psi}, \alpha \mid Y) d\alpha. \tag{3.4}$$

Because this marginal distribution is often difficult to work with, we may sample $p(\tilde{\psi}_p \mid Y, \tilde{\psi}_{-p})$ indirectly by sampling $p(\tilde{\psi}_p, \alpha \mid Y, \tilde{\psi}_{-p})$. This may involve first sampling $p(\alpha \mid Y, \tilde{\psi}_{-p})$ and then sampling $p(\tilde{\psi}_p \mid Y, \tilde{\psi}_{-p}, \alpha)$. The second step is computationally equivalent to sampling $p(\psi_p \mid Y, \psi_{-p}, \alpha)$, which is the same as STEP p of the standard Gibbs sampler when α and ψ are a priori independent, and transforming ψ_p to get $\tilde{\psi}_p$. This avoids the integration in (3.4).

Rather than marginalizing α out or sampling it in each step, an intermediate strategy is to sample a part of α while conditioning on the rest of α in each step. Because this strategy aims to partially accomplish the integration in (3.4), we call the resulting samplers *partially marginalizied* Gibbs (PMG) samplers. We use this term in the same way we did for MDA in Section 2.1. Full marginalization would involve fully sampling α in each step of the sampler. Because we are only partially accomplishing this, we use the term *partial marginalization*.

To describe the sampling scheme, we introduce partitions of α for each step of the sampler. Setting $\alpha = (\alpha_1, \ldots, \alpha_J)$, let $\mathcal{J} = \{\mathcal{J}_1, \ldots, \mathcal{J}_P\}$ be a set of

index sets, with $\mathcal{J}_p \subset \{1, \ldots, J\}$ for $p = 1, \ldots, P$. Let $\alpha_{(p)}$ be the collection of components of α corresponding to the index set $\mathcal{J}_p$, i.e., $\alpha_{(p)} = \{\alpha_j : j \in \mathcal{J}_p\}$. Finally let, $\mathcal{J}_p^c$ be the complement of $\mathcal{J}_p$ and $\alpha_{(p)}^c$ be the collection of components of α not in $\alpha_{(p)}$, $\alpha_{(p)}^c = \{\alpha_j : j \in \mathcal{J}_p^c\}$. To construct a PMG sampler, we replace STEP p of the standard Gibbs sampler with $(\tilde{\psi}_p, \alpha_{(p)}) \sim p(\tilde{\psi}_p, \alpha_{(p)} \mid Y, \tilde{\psi}_{-p}, \alpha_{(p)}^c)$, where we condition on the most recently sampled value of each element of $\tilde{\psi}_{-p}$ and $\alpha_{(p)}^c$. At the end of each iteration we compose $\tilde{\psi}^{(t+1)}$ and $\alpha^{(t+1)}$ of the most recently sampled values of their components.

In this setup we do not put any restrictions on the partitions of α, some components may be sampled in multiple steps or not at all. In any particular step, we may sample all of or none of α, so $\alpha_{(p)}^c$ or $\alpha_{(p)}$ may be empty. We also do not require that all of the components of α are sampled in at least one of the P steps because it may be advantageous to sample some components of α in separate steps, as in MDA SCHEME 3 in the example in Section 2.1. To accomplish this, we may add a set of P' optional steps to each iteration to sample components of α not sampled in other steps. Thus, we define another set of partitions of α, $\{\alpha_{(p)}, p = P+1, \ldots, P+P'\}$, in the PMG sampler.

Partially Marginalizied Gibbs Sampler:

STEP 1: $(\tilde{\psi}_1, \alpha_{(1)}) \sim p(\tilde{\psi}_1, \alpha_{(1)} \mid Y, \tilde{\psi}_{-1}, \alpha_{(1)}^c)$,

$\vdots$

STEP P: $(\tilde{\psi}_P, \alpha_{(P)}) \sim p(\tilde{\psi}_P, \alpha_{(P)} \mid Y, \tilde{\psi}_{-P}, \alpha_{(P)}^c)$,

STEP $P+1$ (optional): $\alpha_{(P+1)} \sim p(\alpha_{(P+1)} \mid Y, \tilde{\psi}^{(t+1)}, \alpha_{(P+1)}^c)$,

$\vdots$

STEP $P+P'$ (optional): $\alpha_{(P+P')} \sim p(\alpha_{(P+P')} \mid Y, \tilde{\psi}^{(t+1)}, \alpha_{(P+P')}^c)$,

STEP $P+P'+1$: Set $\psi_p^{(t+1)} = \mathcal{D}^{-1}_{\alpha^{(t+1)},p}(\tilde{\psi}_p^{(t+1)})$ for each p.

If $\alpha_{(p)} = \emptyset$ at each step and the optional steps are omitted, the result is a Gibbs sampler on the transformed parameter $\tilde{\psi}$, which implicitly conditions on α. In particular, if we condition on $\alpha = a_0$, this PMG sampler becomes what we call the *corresponding* standard Gibbs sampler. On the other hand, if $\alpha_{(p)} = \alpha$ for each step, α is removed from the Markov chain and the sampler is a true MG sampler that completely marginalizes out α in the sense that all draws are from conditional distributions of (3.4). In this case, we may replace each STEP p with $\tilde{\psi}_p \sim \int p(\tilde{\psi}_p, \alpha \mid Y, \tilde{\psi}_{-p}) d\alpha$, since α is not used in subsequent steps. Also in this

case the optional steps do not effect the transition kernel of, $\mathcal{M}(\tilde{\psi})$, and are not used. Generally, STEP $P + P' + 1$ is required to recover ψ, however, so we must draw α at least once in the iteration.

As discussed in Section 2.1, we expect that sampling more components of α in any step of a PMG sampler improves its convergence. Technical results to this effect are elusive owing to the non-Markovian character of the marginal chains of ψ and components of ψ in multi-step samplers. The theoretical development in van Dyk and Park (2008) applies directly to the joint chain of (ψ, α) and indicates that sampling more components of α in any step of a PMG sampler should improve the convergence of this chain. Insofar as we are interested only in the marginal chain of ψ, however, this may be of limited interest. Thus, we do not pursue a full exploration of application of these theoretical results in this setting. Nonetheless, this observation gives a hint of the theoretical advantage of marginalization in multi-step chains involving working parameters. When combined with empirical results, we believe that there is strong evidence of the advantage, see Section 4.3.

Gaussian Example: To illustrate the advantage of the PMG sampler over MDA even in a two-step sampler, we introduce a second working parameter and a second transformation in the Gaussian example. Suppose β is a second scalar working parameter with prior distribution, $\beta \sim \mathrm{N}(0, \omega\sigma_2^2)$, *a priori* independent of ψ and α. Because ψ and (α, β) are *a priori* independent, the samplers all have the same target distribution. We introduce β along with α into the model via the transformation $\widetilde{\psi} = (\widetilde{\psi}_1, \widetilde{\psi}_2) = (\psi_1 + \alpha, \psi_2 + \beta)$. The framework of the PMG sampler allows numerous sampling schemes with each of α and β being sampled or conditioned upon when sampling ψ_1 and ψ_2 and/or being sampled in optional steps. We start with MDA SCHEME 1 because it is the fastest and consider two possibilities for adding β to this sampler. The two samplers are named analogously,

PMG SCHEME 1:
Sample $(\widetilde{\psi}_1, \alpha, \beta) \sim p(\widetilde{\psi}_1, \alpha, \beta \mid \widetilde{\psi}_2)$ and $(\widetilde{\psi}_2, \alpha, \beta) \sim p(\widetilde{\psi}_2, \alpha, \beta \mid \widetilde{\psi}_1)$,

PMG SCHEME 2:
Sample $(\widetilde{\psi}_1, \alpha, \beta) \sim p(\widetilde{\psi}_1, \alpha, \beta \mid \widetilde{\psi}_2)$ and $(\widetilde{\psi}_2, \alpha) \sim p(\widetilde{\psi}_2, \alpha \mid \widetilde{\psi}_1, \beta)$.

The basic instinct to sample as many components of the working parameter as possible in each step suggests that PMG SCHEME 1 dominates both PMG SCHEME 2 and MDA SCHEME 1. The latter comparison stems from MDA SCHEME 1's implicit conditioning on $\beta = 0$ in both of its steps. Comparing MDA SCHEME 1 and PMG SCHEME 1 in Figure 1 illustrates the advantage of the PMG sampler. The autocorrelation of ψ_2 is essentially eliminated by adding β to the sampler.

3.3. Marginal MCMC methods

Section 4 discusses the asymptotic calculations required to verify the stationary distribution of a PMG chain when $p(\alpha)$ is improper. Segregating the computational complexity of the sampler into a number of simpler parts reduces the complexity of these calculations. For example, consider

The Nested MCMC *within Gibbs Sampler*

Step 1: $\psi_1^{(t+1)} \sim \mathcal{K}(\psi_1 \mid \psi_1'; \psi_{-1})$,

$$\vdots$$

Step P: $\psi_P^{(t+1)} \sim \mathcal{K}(\psi_P \mid \psi_P'; \psi_{-P})$,

where ψ_p' is generic notation for the previous draw of ψ_p, and $\mathcal{K}(\psi_p \mid \psi_p'; \psi_{-p})$ is a transition kernel for an irreducible aperiodic Markov chain with unique stationary distribution $p(\psi_p \mid Y, \psi_{-p})$. Clearly the resulting chain, $\mathcal{M}(\psi)$, is an irreducible aperiodic Markov chain with unique stationary distribution $p(\psi \mid Y)$. The kernels in the individual steps may be formulated as direct draws from the conditional distribution, $p(\psi_p|Y, \psi_{-1})$, or by using a PMG sampler with the correct stationary distribution for $\mathcal{M}(\psi_p)$. If an improper working prior distribution is used, the stationary distribution can be verified using the methods of Section 4. Other MCMC methods such as Metropolis-Hastings may be used for some steps. The advantage of this strategy is that we relegate the algorithmic complexity introduced with working parameters to a small subset of the draws, and thus simplify the asymptotic calculations required when using improper working prior distributions. This strategy is illustrated using the logistic mixed model in Sections 3.4 and 4.3.

3.4. Marginal slice sampling in the logistic mixed model

In this section we nest several PMG samplers within the slice sampler presented in Section 2.2. The samplers illustrate both the use and computational efficiency of marginal MCMC methods and how auxiliary and working parameters can be combined to create simple fast algorithms.

The data augmentation scheme used in this slice sampler has great potential primarily because it results in the only known general Gibbs sampler for the generalized linear mixed model that involves only standard distributions. Unfortunately, as illustrated in Section 4.3, the algorithm can be slow to converge. To improve computational efficiency, we suggest recentering the random effect using a working parameter, i.e., setting $\tilde{b}_i = \alpha + b_i$, where $\alpha = (\alpha_1, \ldots, \alpha_q)'$ is a $q \times 1$

working parameter. This transformation is motivated by the goal to *jointly* sample components of α and components of (b, β, T), and results in a reformulation of model (2.3) as

$$y_{ij} = I\left[v_{ij} \le g^{-1}\{x'_{ij}(\tilde{\beta} + \tilde{b}_i)\}\right], \text{ with } \tilde{b}_i \sim \mathrm{N}(\alpha, T), \tag{3.5}$$

where $\tilde{\beta} = \beta - \alpha$. Using the prior distribution, $\alpha_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \omega)$, we incorporate α_k into Cycle k of Step 2.

Substep 1: Draw $\alpha_k^\star \sim \mathrm{N}(0, \omega)$ and, for each i, independently draw $b_{ik}^\star \mid \beta_k, \tau_k^2 \sim N(0, \tau_k^2)$ subject to (2.5) with b_i replaced with $b_i^\star$. Set $\tilde{b}_{\cdot k} = b_{\cdot k}^\star + \alpha_k^\star$. (Starred quantities are intermediate.)

Substep 2: Draw $(\beta_k, \tau_k^2, \alpha_k) \mid \tilde{b}_{\cdot k}$ by sampling

$$\tau_k^2 \sim \frac{\left\{\sum_{i=1}^m (\tilde{b}_{ik} - \tilde{b}_{\cdot k})^2 + \nu_k \tau_{k,0}^2\right\}}{\chi^2_{m+\nu_0-1}},$$

$\alpha_k \mid \tau_k^2 \sim \mathrm{N}(\tilde{b}_{\cdot k}, \tau_k^2/m)$, and $\tilde{\beta}_k$ uniformly subject to (2.5) with b_i replaced with $\tilde{b}_i$; here $\tilde{b}_{\cdot k} = \sum_{i=1}^m \tilde{b}_{ik}/m$. Transform to the original scale by setting $\beta_k = \tilde{\beta}_k + \alpha_k$ and $b_i = \tilde{b}_i - \alpha_k$.

Because α_k is updated in both substeps, it is completely marginalized out of the transition kernel $\mathcal{K}\{b_{\cdot k}, \beta_k, \tau_k^2 \mid b'_{\cdot k}, \beta'_k, (\tau_k^2)'\}$. (Here other model parameters are fixed.) Because each α_k is updated along with the model parameters, no optional steps are used. Before we investigate the performance of this sampler, we discuss convergence results that allow $p(\alpha)$ to be improper.

4. Theoretical Results

4.1. The advantage of the optional steps

Theorem 1 shows that using the optional steps of a PMG sampler to update components of α that are not updated along with ψ can sometimes be much less efficient than updating α along with ψ.

Theorem 1. *In a PMG sampler with* **(i)** ψ *and* α *a priori independent,* **(ii)** $\mathcal{D}_{\alpha,p}(\psi_p) = \psi_p$ *for* $p = 2, \ldots, P$, *and* **(iii)** $\alpha_{(p)} = \emptyset$ *for* $p = 1, \ldots, P$, *the Markov transition kernel for* ψ_{-1} *is identical to that of the corresponding standard Gibbs sampler regardless of* $\alpha_{(p)}$ *for* $p = P+1, \ldots, P'$.

The proof of this and other results appear in Appendix C. If we define $\psi_1^{(t)} = \mathcal{D}^{-1}_{\alpha^{(t-1)},1}(\tilde{\psi}_1^{(t)})$, it can also be shown that $\mathcal{M}(\psi)$ is Markovian with transition

kernel equal to that of the standard Gibbs sampler. Because suppositions 1 and 2 of Theorem 1 hold for MDA samplers, "SCHEME 3" of van Dyk and Meng (2001) is obsolete, at least for $\mathcal{M}(\psi_{-1})$. This is why MDA SCHEME 3 offers no advantage over SCHEME 0 in Figure 1. As we illustrate next, however, optional steps can be useful for PMG samplers that, unlike MDA, do not adhere to the suppositions of Theorem 1.

Gaussian Example: Consider a sampler that introduces both of the working parameters α and β into the bivariate Gaussian sampler using a sampling scheme that includes an optional step:

PMG SCHEME 3: Sample $\widetilde{\psi_1} \sim p(\widetilde{\psi_1} \mid \widetilde{\psi_2}, \alpha, \beta)$, $\widetilde{\psi_2} \sim p(\widetilde{\psi_2} \mid \widetilde{\psi_1}, \alpha, \beta)$, and $(\alpha, \beta) \sim p(\alpha, \beta \mid \widetilde{\psi_1}, \widetilde{\psi_2})$.

Figure 1 shows that PMG SCHEME 3 performs much better than MDA SCHEME 3. As the proof of Theorem 1 illustrates, the basic problem with MDA SCHEME 3 is that the transformation affects only ψ_1 and, even when we compute $\psi = \mathcal{D}_\alpha^{-1}(\tilde{\psi})$, has no affect on ψ_{-1}. With PMG SCHEME 3, on the other hand, the effects of the two working parameters are convolved within the sampler. Thus, optional steps can be useful in PMG samplers that are not MDA samplers.

4.2. Improving computational efficiency

As discussed in Section 2.1, when it is useful to marginalize out working parameters, more diffuse working prior distributions tend to result in better mixing samplers. In the limit when the working prior distribution becomes improper, however, technical difficulties may arise; the transition kernel may become improper or the Markov chain may become non-positive recurrent. Both of these difficulties may occur when the joint posterior distribution of (ψ, α) is improper. For example, if ψ and α are *a priori* independent, the posterior distribution of the working parameter is

$$p(\alpha \mid Y) = \int p(\psi, \alpha \mid Y) d\psi \propto p(\alpha) \int p(Y \mid \psi) p(\psi) d\psi \propto p(\alpha), \tag{4.1}$$

which is improper if $p(\alpha)$ is improper. Clearly, care must be taken when using improper working prior distributions. The example in Appendix B illustrates how a poor choice of an improper working prior distribution may upset the stationary distribution of the chain.

To address the technical difficulties associated with improper working prior distributions, we begin with a generalization of Lemma 1 of Liu and Wu (1999); the generalization accounts for the possibility that the stationary distribution depends on the choice of the working prior distribution. This is the case when ψ and α are not *a priori* independent, see also Imai and van Dyk (2005a).

Lemma 1. *Suppose we have a sequence of proper Markovian transition kernels,* $\mathcal{K}_m(\xi \mid \xi')$*, each with proper stationary distribution,* $\pi_m(\xi)$*. If* **(i)** $\mathcal{K}_\infty(\xi \mid \xi') = \lim_{m\to\infty} \mathcal{K}_m(\xi \mid \xi')$ *is a proper Markovian transition function, and* **(ii)** $\pi_\infty(\xi) = \lim_{m\to\infty} \pi_m(\xi)$ *represents a proper distribution, then* $\pi_\infty(\xi)$ *is the stationary distribution of* $\mathcal{K}_\infty(\xi \mid \xi')$*.*

The goal is to establish conditions for PMG samplers that guarantee that their stationary distribution is the target distribution when improper working prior distributions are used. We focus on verifying condition **(i)** of Lemma 1, leaving the verification of **(ii)** as the standard exercise of establishing that the posterior distribution is integrable under the (limiting) prior distribution $p(\psi) = \int p(\psi, \alpha) d\alpha$. Although Lemma 1 cannot be applied directly to the Markov chain $\mathcal{M}(\psi, \alpha)$, because the limiting kernel is improper (at least when ψ and α are *a priori* independent and $p(\alpha)$ is improper in the limit), it can be applied in some cases to $\mathcal{M}(\psi)$ or subchains of $\mathcal{M}(\psi)$.

In this section, we assume a sequence of PMG samplers with each sampler constructed with

1. ψ and α a priori independent,
2. $\mathcal{D}_{\alpha,p}(\psi_p) = \psi_p$ for $p = 2, \ldots, P$,
3. $\alpha_{(1)} = \alpha$,

and using a sequence of proper working prior distributions $p_m(\alpha)$. Because α is completely updated in STEP 1, the optional steps are unnecessary and we assume $P' = 0$. We also assume that the resulting transition kernels $\mathcal{K}_m(\tilde{\psi}, \alpha \mid \tilde{\psi}', \alpha')$ have proper stationary distributions $\pi_m(\psi, \alpha)$. Although these assumptions limit the use of improper working prior distributions, more general updating schemes can be used with proper working prior distributions. The next several results use Lemma 1 to verify that the limiting Markovian marginal transition kernels of $\mathcal{M}(\psi_{-1})$ and $\mathcal{M}(\psi)$ have the desired stationary distributions. We label these two results, i.e., the limiting behavior or $\mathcal{M}(\psi_{-1})$ and $\mathcal{M}(\psi)$, as R_1 and R_2. Meng and van Dyk (1999) and Liu and Wu (1999) only establish the stationary distribution of the marginal chain of one of the draws (R_1 in a two-step sampler). Thus, R_2 is more general than their result even in a two-step sampler. Figure 2 outlines the theoretical results. Corollary 1 establishes the sufficiency of conditions C_1^a and C_2^a for results R_1 and R_2, respectively; see below. Corollary 2 shows how a minor modification of the samplers along with the weaker condition, C_1^a, can establish the stronger result, R_2. Finally, Theorem 2 and Corollary 3 establish conditions C_1^b and C_2^b that imply C_1^a and C_2^a, respectively but are easier to verify. The final results also describe how to construct the optimal sampler. We begin

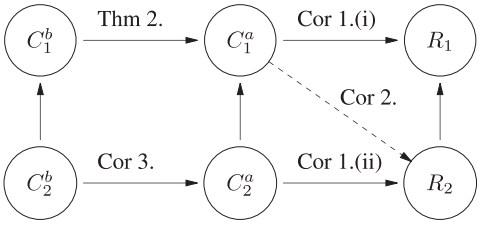

Figure 2. The Theoretical Results of Section 4.2. The conditions of parts **(i)** and **(ii)** of Corollary 1 are represented by C_1^a and C_2^a, respectively; the results of Corollary 1 are represented by R_1 and R_2. Theorem 2 and Corollary 3 provide conditions that may be easier to verify in practice; these are labeled C_1^b and C_2^b, respectively. Corollary 2 shows that for a modified PMG sampler R_2 follows from C_1^a. The vertical arrows indicate that the conditions and results in the second row are all stronger than those in the first row.

with Corollary 1, which applies Lemma 1 directly to the Markov chains $\mathcal{M}(\psi_{-1})$ and $\mathcal{M}(\psi)$.

Corollary 1. *If $\mathcal{M}_m(\tilde{\psi}, \alpha)$ is generated with a PMG sampler constructed with ψ and α a priori independent, $\mathcal{D}_{\alpha,p}(\psi_p) = \psi_p$ for $p = 2, \ldots, P$, $\alpha_{(1)} = \alpha$, and proper working prior distribution, $p_m(\alpha)$, and if $p(\psi \mid Y)$ is a proper distribution, then the subchains, $\mathcal{M}_m(\psi_{-1})$ and $\mathcal{M}_m(\psi)$ are Markovian with transition kernels $\mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ and $\mathcal{K}_m\{\psi \mid \psi'_{-1}\} \equiv \mathcal{K}_m(\psi \mid \psi')$ and stationary distributions $p(\psi_{-1} \mid Y)$ and $p(\psi \mid Y)$, respectively, for each m*[1]*. Thus,*

(i) *if $\mathcal{K}_\infty\{\psi_{-1} \mid \psi'_{-1}\} = \lim_{m\to\infty} \mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ is a proper Markovian transition kernel, then $p(\psi_{-1} \mid Y)$ is the stationary distribution of $\mathcal{M}_\infty(\psi_{-1})$, the Markov chain sampled under $\mathcal{K}_\infty\{\psi_{-1} \mid (\psi_{-1})'\}$, and,*

(ii) *if $\mathcal{K}_\infty\{\psi \mid \psi'_{-1}\} = \lim_{m\to\infty} \mathcal{K}_m\{\psi \mid \psi'_{-1}\}$ is a proper Markovian transition kernel, then $p(\psi \mid Y)$ is the stationary distribution of $\mathcal{M}_\infty(\psi)$, the Markov chain sampled under $\mathcal{K}_\infty\{\psi \mid \psi'_{-1}\}$.*

The proof of all results in this section are in Appendix C. Part **(i)** of Corollary 1 is useful when ψ_{-1} is of primary interest, e.g., when ψ_1 is an auxiliary variable. Integrating out ψ_1 and using Fatou's lemma, the supposition of **(i)** follows from the supposition of **(ii)**. If it is easier to verify the condition of **(i)** but if $p(\psi \mid Y)$ is the target distribution, we can alter the sampler by replacing the transformation in STEP $P + P' + 1$ with a direct draw of the conditional of ψ_1 given ψ_{-1}.

Corollary 2. *Consider a sequence of PMG samplers as described in Corollary* 1*, but with the transformation STEP $P + P' + 1$ in each iteration of each sampler*

[1]The notation $\mathcal{K}_m\{\psi \mid \psi'_{-1}\}$ emphasizes that the kernel of $\mathcal{M}(\psi)$ does not depend on the previous value of ψ_1.

replaced with: Draw $\psi_1 \sim p(\psi_1 \mid Y, \psi_{-1})$. *Then* $\mathcal{M}_m(\psi)$ *is Markovian for each* m *and, if* $\mathcal{K}_\infty\{\psi_{-1} \mid \psi'_{-1}\} = \lim_{m\to\infty} \mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ *is a proper Markovian kernel, then* $p(\psi \mid Y)$ *is the stationary distribution of* $\mathcal{M}_\infty(\psi)$, *the Markov chain sampled under the limiting kernel.*

If ψ and α are not *a priori* independent, the results of Corollaries 1 and 2 still follow as long as condition **(ii)** of Lemma 1 holds. The following theorem, which extends Lemma 1 of van Dyk and Meng (2001), develops equivalent conditions that may be still easier to verify for part **(i)** of Corollary 1 and for Corollary 2. The theorem also describes how to construct the optimal sampler.

Theorem 2. *Consider the sequence of* PMG *samplers in Corollary* 1, *but with* $\alpha_{(2)} = \alpha$. *If*

(i) *there exists an improper working prior distribution* $p_\infty(\alpha)$ *such that* $p_m(\psi_p, \alpha_{(p)} \mid Y, \tilde{\psi}_{-p}, \alpha^c_{(p)}) \to p_\infty(\psi_p, \alpha_{(p)} \mid Y, \tilde{\psi}_{-p}, \alpha^c_{(p)})$ *as* $m \to \infty$ *for* $p = 2, \ldots, P$, *with* p_m *denoting the conditional distributions under the proper working prior distribution,* $p_m(\alpha)$, *and* p_∞ *denoting the same proper distributions under* $p_\infty(\alpha)$; *and*

(ii) $\int \mathcal{K}_\infty\{\psi_{-1}, \alpha \mid \mathcal{D}_{\alpha^\star,1}(\psi_1), \psi'_{-1}\} d\alpha$ *is invariant to* $\alpha^\star$, *where* $\mathcal{K}_\infty\{\psi_{-1}, \alpha \mid \tilde{\psi}_1, \psi'_{-1}\}$ *is the kernel for* STEPS *2-P of the* PMG *sampler run with the improper working prior distribution,* $p_\infty(\alpha)$,

then $\lim_{m\to\infty} \mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ *is the proper transition kernel of the same* PMG *sampler with working prior distribution* $p_\infty(\alpha)$, *except that* STEP 1 *is replaced with* STEP 1*: Sample* $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 \mid Y, \psi'_{-1}, \alpha = a_0\}$.

We can derive similar equivalent conditions for part **(ii)** of Corollary 1. This requires that the transition kernel $\mathcal{K}_\infty(\psi|\psi', \alpha^\star)$ be invariant to $\alpha^\star$, the draw of α in STEP 1. Taking account of the transformation in STEP $P + P' + 1$, supposition **(iii)** of Corollary 3 assures this invariance.

Corollary 3. *Consider a sequence of* PMG *samplers as in Corollary* 1, *but with* $\alpha_{(2)} = \alpha$. *If in addition to supposition* **(i)** *of Theorem* 2,

(iii) $\int p_m\left[\mathcal{D}^{-1}_{\alpha^\star,1}\{\mathcal{D}_{\alpha,1}(\psi_1)\} \mid \psi'_{-1}\right] \mid J\{\mathcal{D}_{\alpha,1}(\psi_1)|\alpha^\star\} \; J^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha\} \mid \times \mathcal{K}_\infty\{\psi_{-1}, \alpha \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi'_{-1}\} d\alpha$ *is invariant to* $\alpha^\star$,

then $\lim_{m\to\infty} \mathcal{K}_m(\psi \mid \psi')$ *is the proper transition kernel, that corresponds to implementing the same* PMG *sampler with improper working prior distribution* $p_\infty(\alpha)$, *except that* STEP 1 *is replaced with* STEP 1*: Sample* $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 \mid Y, \psi'_1, \alpha = a_0\}$.

Applying the change of variable $\psi^\star_1 = \mathcal{D}^{-1}_{\alpha^\star,1}\{\mathcal{D}_{\alpha,1}(\psi_1)\}$ to the density given by the integrand of condition **(iii)** implies supposition **(ii)** of Theorem 2. Thus,

the suppositions of the corollary are stronger than those of the theorem, and we can ignore **(ii)** when applying the corollary. We generally verify **(iii)** by verifying that $\mathcal{K}_\infty(\psi|\psi', \alpha^\star)$ does not depend on $\alpha^\star$. This strategy is illustrated in Appendix D where we verify the limiting kernel for the logistic mixed model.

Gaussian Example: In the Gaussian example with a single location working parameter on ψ_1, the transition kernel under MDA SCHEME 1 with proper working prior distribution can be represented as follows; for simplicity we set $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$. Given ψ_2':

STEP 1: Sample $\tilde{\psi}_1 \sim \mathrm{N}(\varrho\psi_2', 1 - \varrho^2 + \omega^2)$.

STEP 2: Sample $\psi_2 \mid \tilde{\psi}_1 \sim \mathrm{N}\left(\frac{\varrho}{1+\omega^2}\tilde{\psi}_1, \frac{1+\omega^2-\varrho^2}{1+\omega^2}\right)$ and
$\alpha \mid \tilde{\psi}_1, \psi_2 \sim \mathrm{N}\left\{\frac{\omega^2}{1+\omega^2}\tilde{\psi}_1 - \frac{\varrho\omega^2}{1+\omega^2-\varrho^2}\left(\psi_2 - \frac{\varrho}{1+\omega^2}\tilde{\psi}_1\right), \left(1 - \frac{\varrho^2\omega^4}{\omega^2(1+\omega^2-\varrho^2)}\right)\frac{\omega^2}{1+\omega^2}\right\}$.

STEP 3: Set $\psi_1 = \tilde{\psi}_1 - \alpha$.

In the limit as $\omega^2 \to \infty$, the draw in STEP 1 becomes improper, but (ψ_1, ψ_2) does not depend on $\tilde{\psi}_1$, and, thus the limiting transition kernel $\mathcal{K}(\psi \mid \psi')$ is proper. To see this we note that in the limit, STEP 2 becomes: Sample $\psi_2 \sim \mathrm{N}(0, 1)$ and $\alpha \mid \tilde{\psi}_1, \psi_2 \sim \mathrm{N}(\tilde{\psi}_1 - \varrho\psi_2, 1 - \varrho^2)$. The limiting kernel $\mathcal{K}_\infty(\psi_2 \mid \psi_2')$ is clearly proper. Thus, if we replace STEP 3 with $\psi_1 \sim p(\psi_1|\psi_2)$, Corollary 2 guarantees the stationary distribution of $\mathcal{M}(\psi)$ under the limiting kernel to be the target, which is evident. On the other hand, transforming the limiting $p(\alpha \mid \tilde{\psi}_1, \psi_2)$ via the transformation in STEP 3, we find $\psi_1 \mid \tilde{\psi}_1, \psi_2 \sim \mathrm{N}(\varrho\psi_2, 1 - \varrho^2)$. Thus the limiting kernel for $\mathcal{M}(\psi)$ is proper and by part **(ii)** of Corollary 1, has the desired stationary distribution, as is again evident.

4.3. Improper working prior distribution in the logistic mixed model

We now illustrate the computational advantage of PMG sampling when $p(\alpha)$ is improper. Starting with model (3.5), we use $p(\alpha_k) \propto 1$ for each k and replace CYCLE k of the slice sampler as follows.

SUBSTEP 1: For each i, draw $b_{ik}^\star \mid \beta_k, \tau_k^2 \overset{\text{indep.}}{\sim} \mathrm{N}(0, \tau_k^2)$ subject to (2.5) with b_i replaced with $b_i^\star$.

SUBSTEP 2: Draw $(\beta_k, \tau_k^2, \alpha_k) \mid b_{\cdot k}^\star$ by sampling

$$\tau_k^2 \sim \frac{\left\{\sum_{i=1}^m (b_{ik}^\star - b_{\cdot k}^\star)^2 + \nu_k \tau_{k,0}^2\right\}}{\chi^2_{m+\nu_0-1}},$$

$\alpha_k \mid \tau_k^2 \sim \mathrm{N}(b^\star_{\cdot k}, \tau_k^2/m)$, and $\tilde{\beta}_k$ uniformly subject to (2.5) with b_i replaced with $b^\star_i$; here $b^\star_{\cdot k} = \sum_{i=1}^m b^\star_{ik}/m$. Transform to the original scale by setting $\beta_k = \tilde{\beta}_k + \alpha_k$ and $b_i = b^\star_i - \alpha_k$.

This sampler's stationary distribution is verified in Appendix D using the method of Section 4.2.

We compare the convergence properties of the standard slice sampler and the PMG sampler for a logistic mixed model using a simulation with one covariate. We generated three data sets according to (2.2), each with $m = 11$, $\sum_i n_i = 54$, and n_i varying between 4 and 5. The covariates, x_{ij}, were independently generated as $x_{ij} \sim \mathrm{N}(0, 1)$, and the variance of the random effect was set to $\tau^2 = 0.5$. The three data sets differed in the magnitude of the fixed effect, which was set to $\beta = 0$, 1.5, and 10. The magnitude of the effect of the covariate determines the ability of the covariate to predict the outcome and can be an important factor in determining the relative efficiency of the DA and MDA samplers for probit regression; see van Dyk and Meng (2001). We generated Markov chains of length 10,000 using both the standard slice sampler and the PMG sampler with an improper working prior distribution. We fit model (2.2) to each of the three data sets using both samplers; each initialized at $\beta^{(0)} = 0.5, (\tau^2)^{(0)} = 1$. The first 500 draws of each chain is illustrated in Figure 3. The marginal algorithm significantly improves the autocorrelation of the Markov chains. Quantile-quantile plots comparing the samples generated by the two methods verify that they have the same stationary distribution; these plots are omitted.

To illustrate the effect of multiple working parameters, we simulated a data set using two random effects. The data was again generated according to (2.2) with $m = 11$, $\sum_i n_i = 54$, and n_i varying between 4 and 5. The covariates, x_{ij}, were independently generated as $x_{ij} \sim \mathrm{N}_2(0, I)$ with I the identity matrix. The variances of the random effect were set at $\tau_1^2 = \tau_2^2 = 0.5$ and the fixed effects were set at $\beta_1 = \beta_2 = 0$. We again generated Markov chains of length 10,000 using the standard slice sampler and three PMG samplers, the first PMG sampler with a working parameter for the first covariate, the second with a working parameter for the second covariate, and the third with both working parameters. Both working parameters were location parameters with flat working prior distributions, as described above. Each chain was initialized at $\beta_1^{(0)} = \beta_2^{(0)} = 0.5$ and $(\tau_1^2)^{(0)} = (\tau_2^2)^{(0)} = 1$; the first 500 draws of β_1 and β_2 from each chain are illustrated in Figure 4, which clearly illustrates the computational advantage of using both working parameters.

5. Concluding Remarks

The transformation that we use to introduce the working parameter into the model are componentwise transformations. That is, we insist on setting

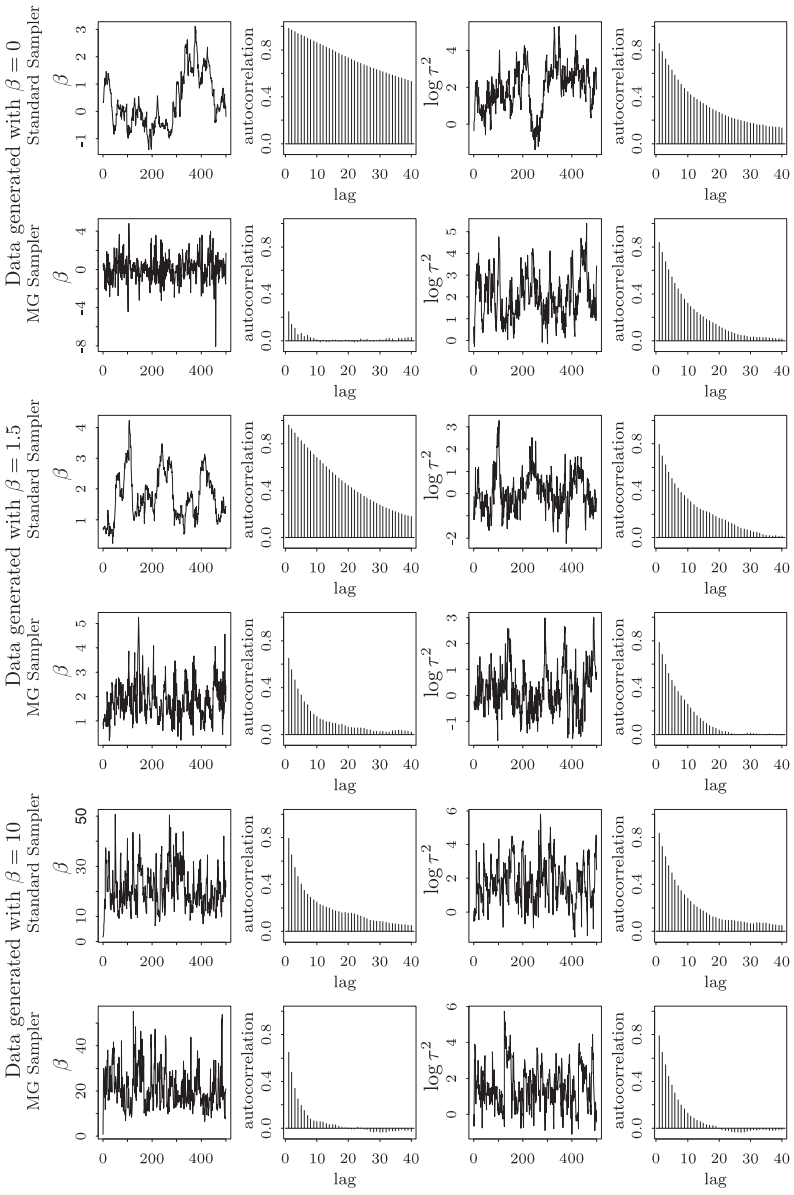

Figure 3. Convergence of Posterior Sampling Algorithms for Fitting a Logistic Regression Model with One Covariate. The first two rows compare the standard slice sampler with the PMG sampler for the data set generated with $\beta = 0$, the second two rows compare the samplers for the data generated with $\beta = 1.5$, and the final two rows compare the samplers for the data generated with $\beta = 10$. In all cases the PMG sampler performs better than the standard slice sampler. The improvement is especially strong for the fixed effect when the autocorrelation for the standard sampler is at its worst.

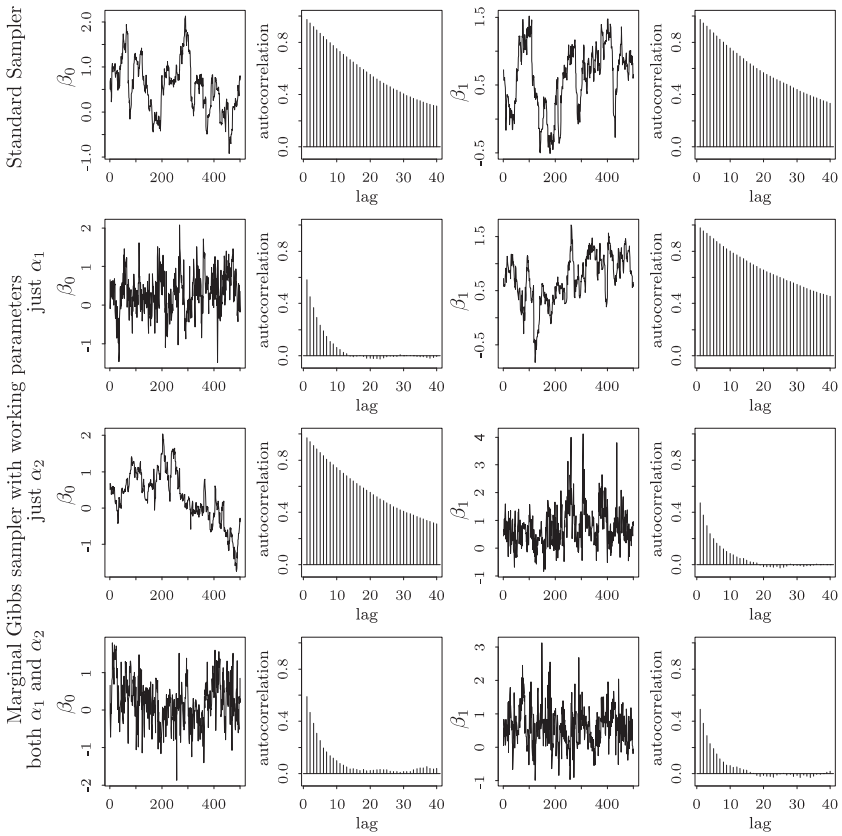

Figure 4. Convergence of Posterior Sampling Algorithms for Fitting a Logistic Regression Model with Two Covariates. The rows of the figure correspond to the standard slice sampler, a PMG sampler implemented with a location working parameter for the first random effect, a PMG sampler implemented with a location working parameter for the second random effect, and a PMG sampler implemented with both working parameters. Notice that each working parameter improves convergence, and including both working parameters produces the best sampler.

$\tilde{\psi}_p = \mathcal{D}_{\alpha,p}(\psi_p)$ for each p rather than considering the more general class of transformations $\tilde{\psi} = \mathcal{D}_\alpha(\psi)$. The reason for this can be illustrated using the simple bivariate Gaussian example once again. Consider transforming (ψ_1, ψ_2) to $\{\psi_1 - \mu_1 - \alpha(\psi_2 - \mu_2), \psi_2\}$. If we take $\alpha_{\text{opt}} = \rho\sigma_1/\sigma_2$, the two components of the transformation are independent, resulting in independent draws from the corresponding two-step Gibbs sampler. Relative to conditioning on $\alpha = \alpha_{\text{opt}}$, averaging over α slows the sampler down. In the general case, the transformation $\tilde{\psi} = \mathcal{D}_\alpha(\psi)$ can be viewed as a family of transformations of ψ indexed by α. In principle, we can then pick the optimal value of α to decorrelate the com-

ponents of $\tilde{\psi}$. Although this may be a useful strategy in practice, it involves a different strategy and different computational methods. Thus, we have chosen not to consider the general class of transformations here.

Generally speaking, one needs to find a good working prior distribution, both in terms of computational ease and efficiency. van Dyk and Meng (2001) introduce several criteria for choosing the working prior distribution in MDA samplers. These criteria recommend the distribution that results in the fastest EM algorithm using the same data augmentation scheme and conditional distributions. In principle a similar strategy can be employed in the multi-step regime by comparing the rates of convergence of the ECM or CM mode-finding algorithms (Meng and Rubin (1993)) as a function of the working prior distribution. These rates of convergence are mathematically more complex that that of EM, however, mitigating the attractiveness of such criteria. In practice it is generally easy enough to try a few different working prior distributions and observe the autocorrelation of a few quantities of interest in order to determine a good choice of the distribution, see for example Figure 6 in Appendix B.

Acknowledgement

The author gratefully acknowledges research support for this project funded in part by NSF grants DMS-01-04129, DMS-04-06085, and SES-05-50980, programming help provided by Hosung Kang while he was a graduate student at Harvard University, and many constructive suggestions from two anonymous referees and an associate editor.

Appendix

A. Using Marginalization to Jump Between Modes

As suggested by a reviewer, here we describe a simple example that illustrates how marginal methods can improve an MCMC sampler's ability to jump between modes of a distribution. Suppose the target distribution is the mixture of two bivariate Gaussian distributions,

$$p(\psi_1, \psi_2) = \frac{1}{2}\,\mathrm{N}_2\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right] + \frac{1}{2}\,\mathrm{N}_2\left[\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right].$$

A two step Gibbs sampler can be constructed by noting that the conditional distribution of ψ_1 give ψ_2 is univariate mixture of $\mathrm{N}(0,1)$ and $\mathrm{N}(\mu_1,1)$ with mixing weights proportional to $\phi(\psi_2)$ and $\phi(\psi_2-\mu_2)$, where $\phi(x)$ is the standard normal density function. We can construct a MG sampler, by introducing the transformation $\tilde{\psi}_1 = \psi_1 + \alpha$, with working prior distribution $\alpha \sim \mathrm{N}(0,\omega^2)$. Using SCHEME 1, we iterate between sampling $(\tilde{\psi}_1, \alpha) \sim p(\tilde{\psi}_1, \alpha \mid \psi_2)$ and $(\psi_2, \alpha) \sim$

$p(\psi_2, \alpha \mid \tilde{\psi}_1)$. The first step is accomplished by sampling α from its prior distribution, sampling $\psi_1 \sim p(\psi_1 \mid \psi_2)$ as in the standard Gibbs sampler, and computing $\tilde{\psi}_1 = \psi_1 + \alpha$. Standard probability calculations show that the second step requires sampling from a bivariate mixture of two Gaussian distributions,

$$p(\psi_2, \alpha \mid \tilde{\psi}_1) \propto \frac{1}{\sigma}\phi\left(\frac{\tilde{\psi}_1}{\sigma^2}\right) \mathrm{N}_2\left[\begin{pmatrix} 0 \\ \xi\tilde{\psi}_1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \xi \end{pmatrix}\right] + \frac{1}{\sigma}\phi\left(\frac{\tilde{\psi}_1 - \mu_1}{\sigma^2}\right) \mathrm{N}_2\left[\begin{pmatrix} \mu_2 \\ \xi(\tilde{\psi}_1 - \mu_1) \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \xi \end{pmatrix}\right],$$

where $\sigma^2 = 1+\omega^2$ and $\xi = \omega^2/(1+\omega^2)$. Since the ultimate goal is to sample from a mixture of two bivariate Gaussian distributions and the MG sampler requires a draw from this type of distribution, the example is only a toy example. Nonetheless, Figure 5 illustrates the ability of this MG sampler to jump between modes much more efficiently than its parent standard Gibbs sampler. The two samplers were applied to three mixture distributions with modes a varying distance apart. The first two rows show the results with $(\mu_1, \mu_2) = (3, 10)$. Although the two modes are 10 standard deviations apart in the ψ_2 direction they are much nearer in the ϕ_1 direction, enabling the standard Gibbs sampler to easily jump between modes. The next pair of rows corresponds to $(\mu_1, \mu_2) = (6, 10)$. Here the standard Gibbs sampler is only occasionally able to jump between the modes (19 times in 10,000 iterations); this results in high autocorrelations and a poor estimate of the relative size of the modes. The last two rows show that the standard sampler is completely unable to jump between modes when $(\mu_1, \mu_2) = (10, 10)$. Although the performance of the MG sampler also deteriorates as the modes grow more distant, it remains a viable sampler in all three simulations.

B. Improper Working Prior Distributions

Here we use the Gaussian example to illustrate the computational risks and benefits of using improper working prior distributions.

Gaussian Example. Returning to the simple Gaussian example, we introduce a new working parameter, $\phi \sim (\mathrm{Gamma}(\kappa_0))^{-1}$, independent of ψ and a transformation, $(\hat{\psi}_1, \hat{\psi}_2) = (\sqrt{\phi}\psi_1, \psi_2)$, where κ_0 is the shape parameter of the gamma distribution. We consider two sampling schemes.

Mda Scheme 1: Sample $(\hat{\psi}_1, \phi) \sim p(\hat{\psi}_1, \phi \mid \psi_2)$ and $(\psi_2, \phi) \sim p(\psi_2, \phi \mid \hat{\psi}_1)$,

Mda Scheme 2: Sample $\hat{\psi}_1 \sim p(\hat{\psi}_1 \mid \psi_2, \phi)$ and $(\psi_2, \phi) \sim p(\psi_2, \phi \mid \hat{\psi}_1)$.

Note that these are the same sampling schemes introduced in Section 2.1 but applied using a different expanded model. All of the steps in both schemes are

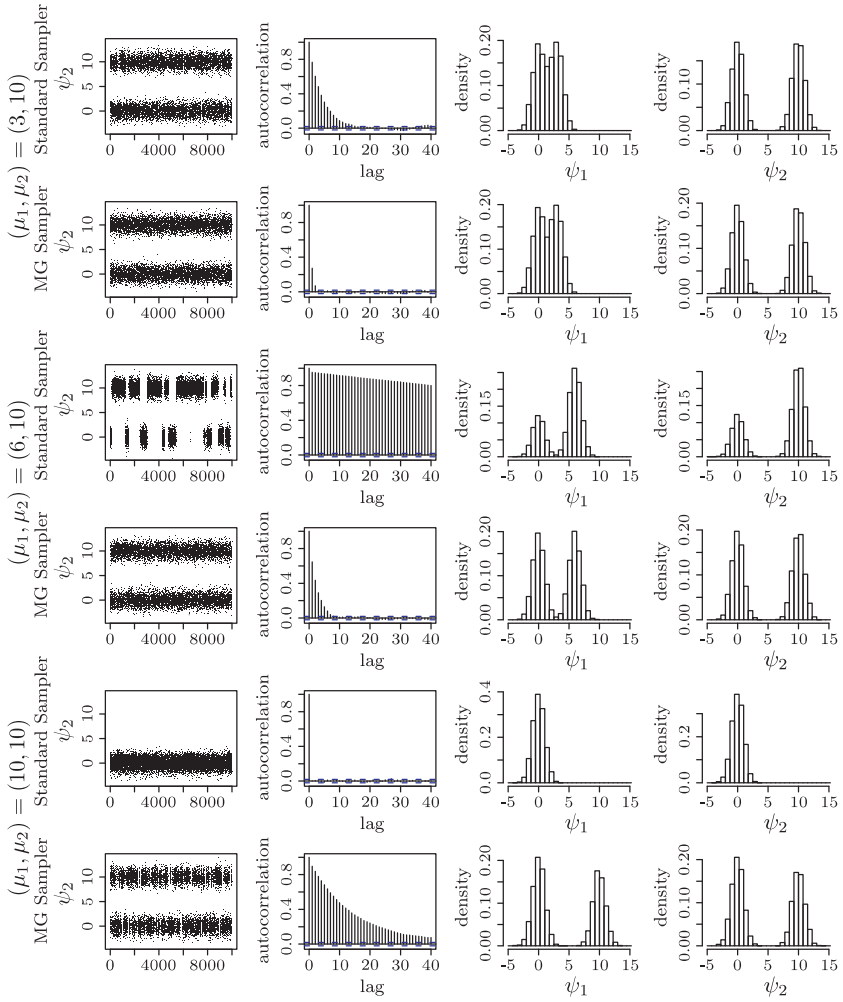

Figure 5. Sampling from a Mixture of Bivariate Gaussian Distributions. The first two rows illustrate the relative efficiencies of a standard Gibbs sampler and an MG sampler when the two Gaussian distributions are close in the ψ_1 direction. The columns correspond to a time series plot and an autocorrelation plot of the draws of ψ_2, and histograms of the draws of ψ_1 and ψ_2. The second pair of rows makes the same comparison when the modes are father apart in the ψ_2 direction, and the final pair when the modes are distant in both directions. The standard Gibbs sampler quickly becomes ineffective as the modes grow more distant. When $(\mu_1, \mu_2) = (6, 10)$ it badly misjudges the relative size of the modes, and when $(\mu_1, \mu_2) = (10, 10)$ it is unable to escape the mode where it begins. Although the MG sampler also becomes less efficient as the modes grow more distant, it gives acceptable results in all cases.

easy to accomplish. For example, because $p(\hat{\psi}_1, \phi \mid \psi_2) = p(\hat{\psi}_1 \mid \phi, \psi_2)p(\phi \mid \psi_2) = p(\hat{\psi}_1 \mid \phi, \psi_2)p(\phi)$ with the second equality following because ϕ and ψ are inde-

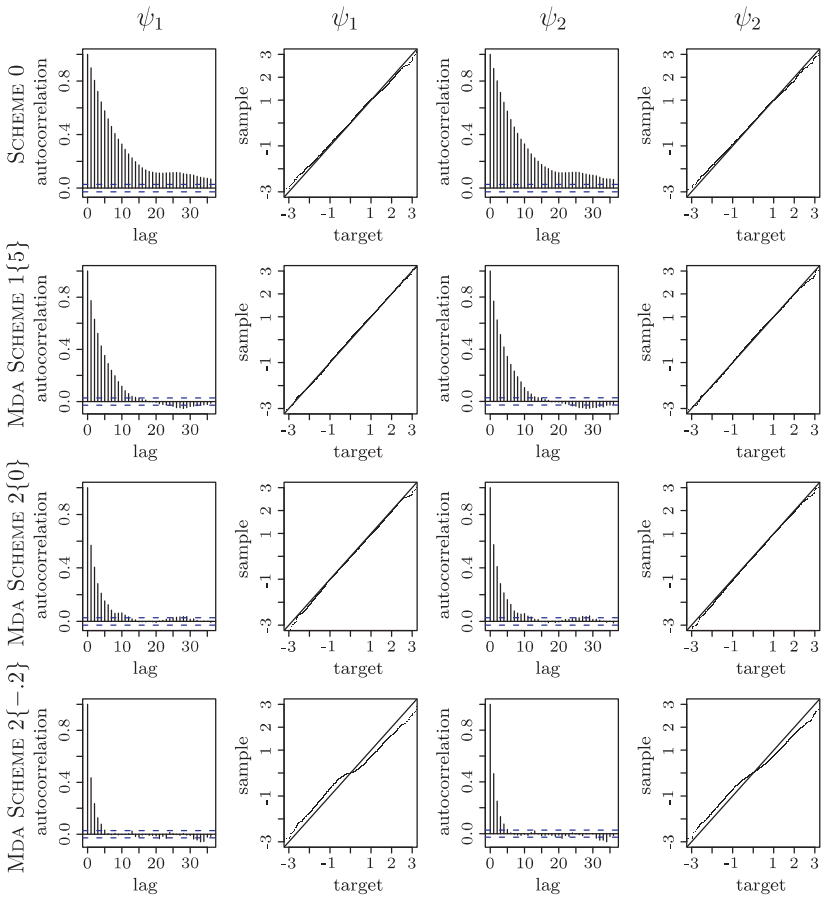

Figure 6. The Risks and Benefits of Using Improper Working Prior Distributions. The figure compares four sampling schemes in the simple Gaussian example. The four rows correspond to a standard sampler with no working parameters, an MDA sampler implemented with a proper working prior distribution and two MDA samplers implemented with improper working prior distributions. The values in curly brackets give the shape parameter for the gamma working prior distribution. The columns provide autocorrelation functions for both parameters and normal quantile plots that compare the Monte Carlo samples with the target standard normal distribution. The plots illustrate both the improvement in the autocorrelation functions resulting from the introduction of working parameters and improper working prior distributions, and the sensitivity of the stationary distribution of the chain to the choice of improper working prior distribution. The sampler illustrated in the bottom row underestimates the variance of ψ_1 by about 40%.

pendent, we can sample $p(\hat{\psi}_1, \phi \mid \psi_2)$ by first sampling ϕ from its working prior distribution and then sampling ψ_1 from $p(\psi_1|\psi_2)$ and computing $\hat{\psi}_1 = \sqrt{\phi}\psi_1$. This factorization shows that the first step in MDA Scheme 1 is improper if the working prior distribution is improper. Thus, we may only implement this scheme

with a proper working prior distribution. MDA SCHEME 2, on the other hand, can be implemented so long as $p(\psi_2, \phi \mid \hat{\psi}_1)$ is proper. Routine calculations indicate that this holds so long as $\kappa_0 > -1/2$.

To illustrate how these samplers work, we implement MDA SCHEME 1 with $\kappa_0 = 5$ and MDA SCHEME 2 with $\kappa_0 = 0$ and $\kappa_0 = -0.2$ and compare them with SCHEME 0. Both implementations of MDA SCHEME 2 use improper working prior distributions. (SCHEME 0 is unaffected by the working parameter model and is the same as described in Section 2.1.) The results appear in Figure 6, where the values in curly brackets indicate the value of κ_0 that was used in each run. The plots in the first and third columns show the autocorrelation functions of ψ_1 and ψ_2, respectively. Comparing MDA SCHEME 1{5} with SCHEME 0 illustrates that the introduction of the working parameter reduces the autocorrelations of the chains. Further comparing with the two implementations of MDA SCHEME 2 shows that using an improper working prior distribution does even better. The second and fourth columns compare the Monte Carlo samples with the target (standard normal) distribution using normal quantile-quantile plots. The improved autocorrelations of MDA SCHEME 1{5} and MDA SCHEME 2{0} result in a slightly better match than with SCHEME 0. The MDA SCHEME 2{−0.2} sampler, on the other hand, underestimates the variability of the target distributions of ψ_1 and ψ_2 by about 40% and 35%, respectively. Section 4.2 aims to develop theory for the PMG sampler that allows us to reap the computational benefits of improper working prior distributions, but with the assurance that the stationary distribution of the resulting chain is the target distribution.

C. Proofs

C.1. Proof of Theorem 1.

We aim to show that the transition kernel of $\mathcal{M}(\psi_{-1})$,

$$\mathcal{K}\{\psi_{-1} \mid (\psi_{-1})'\} = \int \left[\int \mathcal{K}\left\{ \tilde{\psi}_1, \psi_{-1}, \alpha \mid \tilde{\psi}_1', \psi_{-1}', \alpha' \right\} d\alpha \right] d\tilde{\psi}_1 \tag{C.1}$$

is equal to that of the corresponding standard Gibbs sampler where, by supposition 2, $\psi_{-1} = \tilde{\psi}_{-1}$. By construction the inner integral in (C.1) is simply $\mathcal{K}\left\{ \tilde{\psi}_1, \psi_{-1} \mid \tilde{\psi}_1', \psi_{-1}', \alpha' \right\}$. To simplify notation we suppress the dependency on Y and assume that $P = 3$. Thus, (C.1) can be written

$$\begin{aligned} &\int \mathcal{K}\left(\tilde{\psi}_1, \psi_2, \psi_3 \mid \tilde{\psi}_1', \psi_2', \psi_3', \alpha' \right) d\tilde{\psi}_1 \\ &\quad = \int \tilde{p}\left(\tilde{\psi}_1 \mid \psi_2', \psi_3', \alpha' \right) \; p\left(\psi_2 \mid \tilde{\psi}_1, \psi_3', \alpha' \right) \; p\left(\psi_3 \mid \tilde{\psi}_1, \psi_2, \alpha' \right) \; d\tilde{\psi}_1, \end{aligned} \tag{C.2}$$

where we use a tilde accent on p to emphasize that it represents the density of $\tilde{\psi}_1$ rather that of ψ_1. Rewriting $\tilde{p}$ using the change of variable formula, (C.2) is equal to

$$\int p\left\{\mathcal{D}^{-1}_{\alpha',1}(\tilde{\psi}_1) \mid \psi'_2, \psi'_3\right\} \mid J(\tilde{\psi}_1|\alpha') \mid p\left(\psi_2 \mid \tilde{\psi}_1, \psi'_3, \alpha'\right) p\left(\psi_3 \mid \tilde{\psi}_1, \psi_2, \alpha'\right) d\tilde{\psi}_1, \tag{C.3}$$

where $J(\tilde{\psi}_1 \mid \alpha)$ is the Jacobian of the inverse transformation $\mathcal{D}^{-1}_{\alpha,1}(\tilde{\psi}_1)$, or 1 if $\tilde{\psi}_1$ is discrete. Finally, by changing the variable of integration via $\psi^\star_1 = \mathcal{D}^{-1}_{\alpha',1}(\tilde{\psi}_1)$, and by the independence of ψ and α given Y, (C.3) can be written as

$$\int p\left(\psi^\star_1 \mid \psi'_2, \psi'_3\right)\ p\left(\psi_2 \mid \psi^\star_1, \psi'_3\right)\ p\left(\psi_3 \mid \psi^\star_1, \psi_2\right)\ d\psi^\star_1, \tag{C.4}$$

which is the Markovian transition kernel of the corresponding standard Gibbs sampler.

C.2. Proof of Lemma 1

Proof. By Fatou's lemma,

$$\begin{aligned}\int \pi_\infty(\xi')\mathcal{K}_\infty(\xi \mid \xi')d\xi' &= \int \lim_{m\to\infty} \pi_m(\xi')\mathcal{K}_m(\xi \mid \xi')d\xi' \\ &\le \lim_{m\to\infty} \int \pi_m(\xi')\mathcal{K}_m(\xi \mid \xi')d\xi' \\ &= \lim_{m\to\infty} \pi_m(\xi) = \pi_\infty(\xi)\end{aligned}$$

for every ξ. Because $\int\int \pi_\infty(\xi')\mathcal{K}_\infty(\xi \mid \xi')d\xi' d\xi = \int \pi_\infty(\xi)d\xi = 1$, the weak inequality must be an equality and thus $\int \pi_\infty(\xi')\mathcal{K}_\infty(\xi \mid \xi')d\xi' = \pi_\infty(\xi)$.

C.3. Proof of Corollary 2

Proof. The transition kernel for $\mathcal{M}_m(\psi)$ is $p(\psi_1 \mid Y, \psi_{-1})\mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$, where the second term is the transition kernel for $\mathcal{M}_m(\psi_{-1})$, as described in Corollary 1. It follows that $\mathcal{M}_m(\psi)$ is Markovian for each m and that $\lim_{m\to\infty} p(\psi_1 \mid Y, \psi_{-1})$ $\mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ is a proper transition kernel if $\lim_{m\to\infty} \mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ is.

C.4. Proof of Theorem 2

Proof. To simplify notation we suppress conditioning on Y and assume that $P = 3$. Then the marginal transition kernel of $(\psi_{-1}, \alpha^\star)$ with $\alpha^\star$ the draw of α from Step 1, $\mathcal{K}_m\{\psi_{-1}, \alpha^\star \mid \psi'_{-1}\}$, can be written as

$$\begin{aligned}\int \tilde{p}_m\left(\tilde{\psi}_1, \alpha^\star \mid \psi'_2, \psi'_3\right) p_m\left(\psi_2, \alpha^{\star\star}_{(3)}, \alpha^c_{(3)} \mid \tilde{\psi}_1, \psi'_3\right) \\ p_m\left(\psi_3, \alpha_{(3)} \mid \tilde{\psi}_1, \psi_2, \alpha^c_{(3)}\right) d\tilde{\psi}_1 d\alpha^{\star\star}_{(3)} d\alpha,\end{aligned} \tag{C.5}$$

where the accent on $\tilde{p}$ emphasizes that this is the conditional density of $\tilde{\psi}_1$ rather than ψ_1. Integrating out $\alpha^{\star\star}_{(3)}$, replacing $\tilde{p}_m(\tilde{\psi}_1, \alpha^\star \mid \psi'_2, \psi'_3)$ with $p_m(\alpha^\star)$ $p_m\{\mathcal{D}^{-1}_{\alpha^\star,1}(\tilde{\psi}_1) \mid \psi'_2, \psi'_3\}$ $|J(\tilde{\psi}_1 \mid \alpha^\star)|$, and changing the variable of integration via $\psi^\star_1 = \mathcal{D}^{-1}_{\alpha^\star,1}(\tilde{\psi}_1)$ in (C.5), yields

$$p_m(\alpha^\star) \int p_m\left(\psi^\star_1 \mid \psi'_2, \psi'_3\right) p_m\left\{\psi_2, \alpha^c_{(3)} \mid \mathcal{D}_{\alpha^\star,1}(\psi^\star_1), \psi'_3\right\} \\ p_m\left\{\psi_3, \alpha_{(3)} \mid \mathcal{D}_{\alpha^\star,1}(\psi^\star_1), \psi_2, \alpha^c_{(3)}\right\} d\psi^\star_1 d\alpha. \tag{C.6}$$

By Fatou's lemma and condition **(i)**, the limit as $m \to \infty$ of the integral in (C.6) is

$$\int p\left\{\psi^\star_1 \mid \psi'_{-1}\right\} \mathcal{K}_\infty\{\psi_{-1}, \alpha \mid \mathcal{D}_{\alpha^\star,1}(\psi^\star_1), \psi'_{-1}\} d\psi^\star_1 d\alpha. \tag{C.7}$$

By condition **(ii)**, we may replace $\alpha^\star$ under the integral by the identity value, a_0. Thus, in the limit $\alpha^\star$ and ψ_{-1} are independent, the kernel, $\mathcal{K}_\infty\{\psi_{-1} \mid \psi'_{-1}\}$ is the proper distribution given in (C.7) with $\alpha^\star$ replaced with a_0, and this kernel is identical to that resulting from implementing the PMG sampler with $p_\infty(\alpha)$, but with STEP 1 replaced with $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 \mid Y, \psi'_{-1}, \alpha = a_0\}$.

C.5. Proof of Corollary 3

Proof. To simplify notation we again suppress conditioning on Y and assume that $P = 3$. Then the marginal transition kernel of $(\psi, \alpha^\star)$ with $\alpha^\star$ the draw of α from STEP 1, $\mathcal{K}_m\{\psi, \alpha^\star \mid \psi'_{-1}\}$, can be written as

$$\int \tilde{p}_m\left\{\mathcal{D}_{\alpha,1}(\psi_1), \alpha^\star \mid \psi'_2, \psi'_3\right\} \ \left| J^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha\}\right| \\ \times p_m\left\{\psi_2, \alpha^{\star\star}_{(3)}, \alpha^c_{(3)} \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi'_3\right\} p_m\left\{\psi_3, \alpha_{(3)} \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi_2, \alpha^c_{(3)}\right\} d\alpha^{\star\star}_{(3)} d\alpha, \tag{C.8}$$

where the accent on $\tilde{p}$ emphasizes that this is the conditional density of $\tilde{\psi}_1$ rather than ψ_1. Integrating out $\alpha^{\star\star}_{(3)}$ and replacing $\tilde{p}_m(\tilde{\psi}_1, \alpha^\star \mid \psi'_2, \psi'_3)$ with $p_m(\alpha^\star)$ $p_m\{\mathcal{D}^{-1}_{\alpha^\star,1}(\tilde{\psi}_1) \mid \psi'_2, \psi'_3\}$ $|J(\tilde{\psi}_1 \mid \alpha^\star)|$ in (C.8), yields

$$p_m(\alpha^\star) \int p_m\left[\mathcal{D}^{-1}_{\alpha^\star,1}\{\mathcal{D}_{\alpha,1}(\psi_1)\} \mid \psi'_2, \psi'_3\right] \ \left| J\left\{\mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha^\star\right\} J^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha\}\right| \\ \times p_m\left\{\psi_2, \alpha^c_{(3)} \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi'_3\right\} \ p_m\left\{\psi_3, \alpha_{(3)} \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi_2, \alpha^c_{(3)}\right\} \ d\alpha. \tag{C.9}$$

By Fatou's lemma and condition **(i)** of Theorem 2, the limit as $m \to \infty$ of the integral in (C.9) is

$$\int p_m\left[\mathcal{D}^{-1}_{\alpha^\star,1}\{\mathcal{D}_{\alpha,1}(\psi_1)\} \mid \psi'_2, \psi'_3\right] \ \left| J\left\{\mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha^\star\right\} J^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha\}\right| \\ \times \mathcal{K}_\infty\{\psi_{-1}, \alpha \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi'_{-1}\} d\alpha. \tag{C.10}$$

By condition **(iii)**, we may replace $\alpha^\star$ under the integral by the identity value, a_0. Thus, in the limit $\alpha^\star$ and ψ_1 are independent, the kernel, $\mathcal{K}_\infty\{\psi \mid \psi'_{-1}\}$ is the proper distribution given in (C.10) with $\alpha^\star$ replaced with a_0, and this kernel is identical to that resulting from implementing the PMG sampler with $p_\infty(\alpha)$, but with STEP 1 replaced with $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 \mid Y, \psi'_{-1}, \alpha = a_0\}$.

To verify that condition **(iii)** implies condition **(ii)** of Theorem 2, we apply the change of variable $\psi_1^\star = \mathcal{D}^{-1}_{\alpha^\star,1}\{\mathcal{D}_{\alpha,1}(\psi_1)\}$ in the density given in (C.10); the Jacobian of the transformation is $\left| J^{-1}\{\mathcal{D}_{\alpha^\star,1}(\psi_1^\star) \mid \alpha^\star\}\, J\{\mathcal{D}_{\alpha^\star,1}(\psi_1^\star) \mid \alpha\} \right|$.

D. The Optimal Sampler for the Logistic Mixed Model

Here we verify that the marginal slice sampler given in Section 4.3 for the logistic mixed model has the target posterior distribution as its stationary distribution. Because the sampler is constructed by nesting PMG samplers within a larger Gibbs sampler, we need only verify the limiting kernel for each of the PMG samplers. We consider a sequence of transition kernels, $\mathcal{K}_\omega\{b_{,k}, \beta_k, \tau_k^2 \mid b'_{,k}, \beta'_k, (\tau_k^2)'\}$, constructed using a two-step PMG sampler with complete conditional distributions corresponding to the standard slice sampler, $\alpha_{(1)} = \alpha_{(2)} = \alpha_k$, and working prior distribution, $\alpha \sim \mathrm{N}(0, \omega^2 I)$, with I the identity matrix. We verify that the stationary distribution of the transition kernel, $\lim_{\omega\to\infty} \mathcal{K}_\omega\{b_{,k}, \beta_k, \tau_k^2 \mid b'_{,k}, \beta'_k, (\tau_k^2)'\}$, is $p(b_{,k}, \beta_k, \tau_k^2)$; here and throughout the appendix we suppress conditioning on Y, V, and the components other than the kth of β and each b_i. We use Corollary 3 and must verify condition **(i)** of Theorem 2 and condition **(iii)** of Corollary 3. We begin by explicitly deriving the stochastic mapping of CYCLE k of STEP 2.
SUBSTEP1: Sample $\tilde{b}_{,k}, \alpha_k \mid \beta_k, \tau_k^2$ by independently sampling $\alpha_k^\star \sim \mathrm{N}(0, \omega^2)$ and $b_i^\star \sim \mathrm{TN}\{0, \tau_k^2, L(b_{ik}), U(b_{ik})\}$ for each i, where

$$L(b_{ik}) = \max_{j:(y_{ij}-1/2)x_{ij}>0} \left\{ \frac{\mathrm{logit}(v_{ij}) - S_{ij,-k}}{x_{ijk}} - \beta_k \right\},$$

$$U(b_{ik}) = \min_{j:(y_{ij}-1/2)x_{ij}<0} \left\{ \frac{\mathrm{logit}(v_{ij}) - S_{ij,-k}}{x_{ijk}} - \beta_k \right\},$$

with $\mathrm{TN}\{\mu, \sigma^2, L, U\}$ denoting a $\mathrm{N}(\mu, \sigma^2)$ distribution truncated to the interval (L, U), $S_{ij,-k} = \sum_{l\neq k} x_{ijl}(\beta_l + b_{il})$, and x_{ijl} represents component l of x_{ij}. Finally, set $\tilde{b}_{,k} = b^\star_{,k} + \alpha_k^\star$; here we use a star in the superscript to indicate an intermediate quantity. In the limit the distribution of $\alpha_k^\star$ becomes improper; we show, however, that the limiting transition kernel does not depend on $\alpha_k^\star$.

Substep 2: Sample $(\tilde{\beta}_k, \alpha_k, \tau_k^2) \mid \tilde{b}_{,k}$; for finite ω this has density,

$$p_\omega(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{,k})$$
$$\propto \prod_{ij} \left(I\left[v_{ij} \le g^{-1}\{x'_{ij}(\beta + b_i)\}\right] \right)^{y_{ij}} \left(I\left[v_{ij} > g^{-1}\{x'_{ij}(\beta + b_i)\}\right] \right)^{1-y_{ij}}$$
$$\times (\tau_k^2)^{-(\nu_k+m)/2-1} \exp\left[-\frac{1}{2\tau_k^2} \left\{ \sum_{i=1}^{m} (\tilde{b}_{ik} - \tilde{b}_{\cdot k})^2 + m(\tilde{b}_{\cdot k} - \alpha_k)^2 + \nu_k \tau_{k,0}^2 \right\} - \frac{\alpha_k^2}{2\omega^2} \right],$$

where $\tilde{b}_{\cdot k} = (1/m) \sum_{i=1}^{m} \tilde{b}_{ik}$. Clearly, $p_\omega(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{,k}) \to p_\infty(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{,k})$ as $\omega \to \infty$, where p_∞ represents the conditional distribution under the limiting improper prior distribution, $p(\omega) \propto 1$. This satisfies condition **(i)** of Theorem 2. We can simulate $p_\infty(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{,k})$ by sampling

$$\tau_k^2 \sim \frac{\left(\sum_{i=1}^{m} (b_{ik}^\star - b_{\cdot k}^\star)^2 + \nu_k \tau_{k,0}^2 \right)}{\chi^2_{m+\nu_0-1}}, \ \delta_1^\star \sim \mathrm{N}(b_{\cdot k}^\star, \frac{\tau_k^2}{m}), \quad \text{and} \quad \delta_2^\star \sim \mathrm{Unif}(L(\beta_k), U(\beta_k)),$$

and setting $\alpha_k = \delta_1^\star + \alpha_k^\star$, and $\tilde{\beta}_k = \delta_2^\star - \alpha_k^\star$, where

$$L(\beta_k) = \max_{\{i,j;(y_{ij}-1/2)x_{ij}>0\}} \left\{ \frac{\mathrm{logit}(v_{ij}) - S_{ij,-k}}{x_{ijk}} - b_{ik}^\star \right\},$$

$$U(\beta_k) = \min_{\{i,j;(y_{ij}-1/2)x_{ij}<0\}} \left\{ \frac{\mathrm{logit}(v_{ij}) - S_{ij,-k}}{x_{ijk}} - b_{ik}^\star \right\}.$$

We complete the iteration by transforming back to the original parameterization, $\beta_k = \tilde{\beta}_k + \alpha_k = \delta_1^\star + \delta_2^\star$ and $b_{ik} = \tilde{b}_{ik} - \alpha_k = b_{ik}^\star - \delta_1^\star$ for each i. Because $(b_{,k}, \beta_k, \tau_k^2)$ does not depend on $\alpha_k^\star$, condition **(iii)** of Corollary 3 is satisfied and we have the desired result.

References

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55**, 25-37.

Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Roy. Statist. Soc. Ser. B* **61**, 331-344.

Edwards, R. and Sokal, A. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Rev. Ser. D* **38**, 2009-2012.

Gelman, A., van Dyk, D. A., Huang, Z. and Boscardin, W. J. (2008). Transformation and parameter-expanded Gibbs samplers for multilevel and generalized linear models. *J. Comput. Graph. Statist.* **17**, 95-122.

Ghosh, J. and Dunson, D. (2009). Default priors and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Statist.* **18**, 306-320.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93**, 585-595.

Imai, K. and van Dyk, D. A. (2005a). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econom.* **124**, 311-334.

Imai, K. and van Dyk, D. A. (2005b). MNP: R package for fitting multinomial the probit model. *J. Statist. Software* **14**, Issue 5.

Liu, C., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion for EM acceleration – the PXEM algorithm. *Biometrika* **75**, 755-770.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94**, 1264-1274.

Marchev, D. and Hobert, J. P. (2004). Geometric ergodicity of van Dyk and Mengś algorithm for the multivariate student's t model. *J. Amer. Statist. Assoc.* **99**, 228-238.

McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *J. Econom.* **64**, 207-240.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267-278.

Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm - an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301-320.

Neal, R. M. (1997). Markov chain Monte Carlo methods based on 'slicing' the density function. Technical Report No. 9722, Department of Statistics, University of Toronto .

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statist. Comput.* **8**, 229-242.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 45-57. Chapman & Hall, London.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528-550.

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

Tierney, L. (1996). Introduction to general state-space markov chain theory. In *Markov Chain Monte Carlo in Practice* (Editors: W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), 59-74. Chapman & Hall, London.

van Dyk, D. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Amer. Statist. Assoc.* **103**, 790-796.

van Dyk, D. A. and Meng, X.-L. (2000). Algorithms based on data augmentation. In *Computing Science and Statistics: Proceedings of the 31st Symposium on the Interface* (Edited by M. Pourahmadi and K. Berk), 230-239. Interface Foundation of North America, Fairfax Station, VA.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* **10**, 1-111.

van Dyk, D. A. and Meng, X.-L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statist. Sci.*, in press.

Department of Statistics, University of California, Irvine, CA 92697-1250, U.S.A.
E-mail: dvd@uci.edu

(Received June 2008; accepted June 2009)