

Marginal Markov Chain Monte Carlo Methods

David A. van Dyk *

Department of Statistics, University of California, Irvine, CA 92697

dvd@ics.uci.edu

Hosung Kang

Washington Mutual

hosung.kang@gmail.com

June 18, 2008

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. Since these samplers are typically used in a Bayesian framework, the working parameter requires a prior distribution and the convergence properties of the Markov chain depend on the choice of this distribution. Under certain conditions the optimal choice is improper and results in a joint Markov chain on the expanded parameter space that is not positive recurrent. 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.

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.

*Professor van Dyk's research was supported in part by NSF grants DMS-04-06085 and SES-05-50980.

In a Bayesian context, we aim to construct a Markov chain with stationary distribution,

$$p(\psi | Y) = \int p(\psi, \alpha | Y) d\alpha \propto \int p(Y | \psi, \alpha) p(\psi, \alpha) d\alpha, \quad (1)$$

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. (Throughout this article, whenever appropriate, all equalities and inequalities, such as in (1), are understood to hold almost surely with respect to the appropriate dominating measure.)

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). In its simplest form, the DA sampler oscillates between sampling $\psi \sim p(\psi | Y, \alpha)$ and $\alpha \sim p(\alpha | Y, \psi)$. The advantage of this strategy is clear when both conditional draws are simple but sampling $p(\psi | Y)$ directly is complex or impossible under practical constraints. It is well known, however, that DA samplers can exhibit poor mixing. Thus, the traditional motivation for data augmentation 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 $p(\psi | Y, \alpha)$ and $p(\alpha | Y, \psi)$ are easy to sample, but $p(\psi | Y)$ is not. In some cases, a Gibbs sampler can be used to sample $p(\psi | Y)$ 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 computational speed.

This is the situation that motivated the method of auxiliary variables, otherwise known as the slice sampler (Neal, 1997). The method can be formulated as a special case of the DA sampler in which the target distribution can be written,

$$p(\psi | Y) = \pi(\psi | Y) \prod_{i=1}^n l_i(\psi | Y), \quad (2)$$

where any of the factors on the left hand side might not depend on Y . The model is expanded via $p(\alpha | Y, \psi)$, a uniform distribution on the rectangle $[\{0, l_1(\psi | Y)\}, \dots, \{0, l_n(\psi | Y)\}]$, in which

case $p(\alpha | Y, \psi)$ is easy to sample and $p(\psi | Y, \alpha)$ may be easy to sample directly or via some Markov chain technique. Although this method is formally a special case of the DA sampler, it is more prescriptive than DA and has itself lead to many useful samplers. In fact, Damien *et al.* (1999) used its easy implementation in a large class of Bayesian application as their primary motivation for the method.

In this paper we develop another strategy based on a special case of (1), namely, the method of working parameters. First introduced by Meng and van Dyk (1997) and Liu *et al.* (1998) in the context of the EM algorithm working parameters were used to improve the convergence of DA samplers by Meng and van Dyk (1999) and Liu and Wu (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 $p(\psi | Y, \alpha)$. 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) and more general marginal methods, on the other hand, construct Markov chains on the expanded parameter space of (ψ, α) and effectively marginalize out α . This results in draws from the marginal distribution $p(\psi | Y) = \int p(\psi, \alpha | Y) d\alpha$. To be sure that this marginal distribution equals the target conditional distribution it is required that α be unidentifiable in (1), i.e., $p(Y | \psi, \alpha) = p(Y | \psi)$. Thus, the first step in implementing these methods is the sometimes subtle task of finding a suitable working parameter. 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 have been developed and illustrate the potential of marginal methods to dramatically improve the convergence properties of MCMC 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, 2008). See Liu (2003) and Yu and Meng (2008) for related strategies based on transformations of ψ alone.

Marginal MCMC methods construct a Markov chain with stationary distribution $p(\psi, \alpha | Y) = p(\psi | Y)p(\alpha | \psi)$, where α is a working parameter. 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 $p(\alpha | \psi)$ is improper.

Because α is unidentifiable, $p(\psi, \alpha | Y)$ is improper if $p(\alpha | \psi)$ 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 remainder of the article is organized into three sections. In Section 2 we introduce the marginal Gibbs sampler and the relevant background, theory, and methods for its use. We also discuss how our methods can be used in more general MCMC samplers. Section 3 applies the marginal Gibbs sampler to a logistic mixed model, and illustrates the significant computational advantage of the marginal sampler. We conclude with a brief discussion in Section 4; an appendix gives some technical details required to verify the optimal sampler for the logistic mixed model.

2 Marginal MCMC Methods

2.1 Marginal Data Augmentation

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

$$p(\psi, \alpha | Y) = p(\psi | Y)p(\alpha) \tag{3}$$

This joint model is always easy to specify and ensures that the working parameter is unidentifiable given the observed data. The working parameter is independent of ψ in both the prior and posterior distributions. Thus, not until we introduce a joint transformation of ψ and α can we construct a two-step Gibbs sampler (i.e., a DA sampler) that is substantively affected by the in-

roduction of 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}_α is an invertible and differentiable mapping and there exists \mathcal{I}_α such that $\mathcal{D}_{\mathcal{I}_\alpha}$ is an identity mapping. We consider constructing samplers for either the joint posterior distribution $p(\tilde{\psi}_1, \psi_2, \alpha | Y)$ or the marginal posterior distribution $p(\tilde{\psi}_1, \psi_2 | Y) = \int p(\tilde{\psi}_1, \psi_2, \alpha | Y) d\alpha$. These two strategies might formally be called joint augmentation and marginal augmentation, respectively (van Dyk and Meng, 2000, 2008). Because we introduce a more general strategy that encompasses both, however, we blur the distinction and refer to these techniques generally as MDA.

Example: Before we describe the computational advantages of MDA, we introduce a simple but illustrative example that we come back to several times in Section 2 to clarify ideas. We emphasize that this example is not meant to introduce new samplers that are useful in practice but rather to illustrate subtle features of our methods in a concrete example. We suppose ψ follows a bivariate Gaussian distribution with mean $\mu = (\mu_1, \mu_2)$ and variance Σ , that we parameterize in terms of the marginal variances and correlation, i.e., σ_1^2 , σ_2^2 , and ρ , respectively. We introduce a scalar working parameter, α , with working prior distribution, $\alpha \sim N(0, \omega^2 \sigma_1^2)$, 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 $\mathcal{I}_\alpha = 0$. We can easily compute the Gaussian joint distribution, $p(\tilde{\psi}, \alpha)$, and all of the relevant conditional and marginal distributions.

We consider four sampling schemes:

SCHEME 0: Sample $\tilde{\psi}_1 \sim p(\tilde{\psi}_1 | \psi_2, \alpha = \mathcal{I}_\alpha)$ and $\psi_2 \sim p(\psi_2 | \tilde{\psi}_1, \alpha = \mathcal{I}_\alpha)$.

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

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

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

When $\alpha = \mathcal{I}_\alpha = 0$, $\tilde{\psi}_1 = \psi$ so that SCHEME 0 is simply the standard DA sampler. The numbering of the SCHEMES 1, 2, and 3 corresponds to that given in the general framework of van Dyk and Meng (2001). It is easy to compute the lag-one autocorrelation for ψ_2 for each of the four samplers; for

SCHEMES 0 and 3 it is ϱ^2 and for SCHEMES 1 and 2 it is $\varrho^2/(1 + \omega^2)$. The convergence of the four sampling schemes, each run with $\varrho = 0.95$ and $\omega^2 = 25$ is illustrated in Figure 1, which shows the computational gain of SCHEMES 1 and 2. The advantage of these two sampling schemes stems from the fact that to differing degrees they marginalize out α in the complete conditional distributions; in fact, SCHEME 1, which marginalizes α out of both steps appears to outperform SCHEME 2. The term *marginalize* is used because SCHEME 1 effectively constructs a Markov Chain on $(\tilde{\psi}_1, \psi_2)$ with stationary distribution $p(\tilde{\psi}_1, \psi_2 | Y) = \int p(\tilde{\psi}_1, \psi_2, \alpha | Y) d\alpha$. This is accomplished by sampling rather than conditioning on α in each step. When you use this strategy in some but not all of the steps, we use the term *partial marginalization*.

From the formula for the lag-one autocorrelation, we expect more diffuse working prior distributions to result in samplers that mix better. If this prior distribution is improper, however, the first draw of SCHEME 1 is improper and neither SCHEME 1 nor SCHEME 2 has a stationary distribution. SCHEME 3 never samples α jointly with any component of ψ . This strategy fails to improve convergence in this case. We discuss this in detail in Section 2.4.

This simple example illustrates the advantage of the working parameter formulation given in (3). These MDA samplers could not be constructed using the original working parameter paradigm. The parameter, α , is not an unidentifiable parameter that is identifiable given the augmented data. Indeed there is no data augmentation in this simple example. ■

SCHEMES 0–3 introduced in the example can be viewed as generic notation for general sampling schemes for MDA. Generally speaking, the marginal chain for ψ_2 under both SCHEME 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 in one two-step example. The key to all of these results is the basic observation that less conditioning in any step of a Gibbs sampler tends to improve the overall convergence of the resulting Markov chain, 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

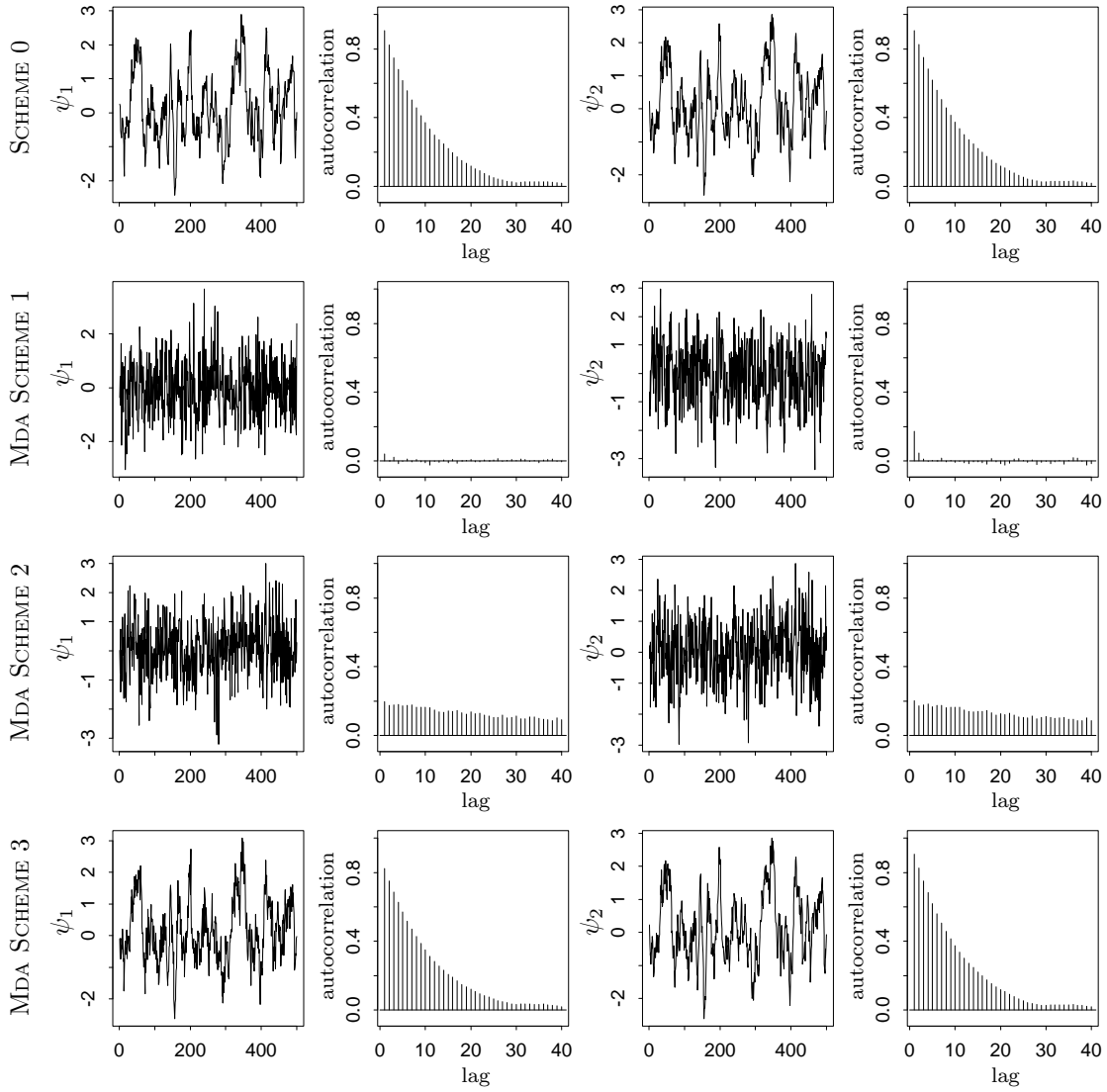


Figure 1: Mixing of MDA SCHEMES 0, 1, 2, and 3 in the Gaussian Example. The figure illustrate a time series plot and an autocorrelation plot for ψ_1 and ψ_2 for each of the four sampling schemes. MDA SCHEMES 1 and 2 both outperform mda SCHEMES 0 and 3, which are indistinguishable. Although mda SCHEMES 1 and 2 have the same lag-one autocorrelation for ψ_2 , MDA SCHEME 1 is clearly superior.

convergence by eliminating the conditioning on α in the first step. Our strategy is to exploit this basic observation in multi-step samplers involving working parameters.

The prior distribution on α can be regarded as a tuning parameter for algorithms involving marginal augmentation. In the example, we consider a family of prior distributions indexed by ω . The calculation of the lag-one autocorrelation illustrates how the choice of ω can affect the convergence of the resulting Markov chain. For MDA SCHEMES 2 and 3, the optimal chains occur when $\omega = \infty$ and the lag-one autocorrelation becomes zero. In this, as in many examples, the optimal limit occurs when the prior distribution on α becomes improper. This significantly complicates the situation because the joint target distribution, $p(\psi, \alpha | Y) = p(\psi | Y)p(\alpha)$ is also improper and the joint Markov chain, $\mathcal{M}_{(\psi, \alpha)} = \{(\psi^{(t)}, \alpha^{(t)}), t = 1, 2, \dots\}$ is not positive recurrent; we generally use the notation \mathcal{M}_x for the chain $\{x^{(t)}, t = 1, 2, \dots\}$. Meng and van Dyk (1999) showed, however, that in some cases the marginal chain for ψ , \mathcal{M}_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 2.5.

2.2 Model expansion

Suppose we wish to obtain a Monte Carlo sample from $p(\psi | Y)$ and partition ψ into (ψ_1, \dots, ψ_P) to construct a Gibbs sampler, where 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 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 the steps:

Standard Gibbs Sampler:

STEP 1: $\psi_1 \sim p(\psi_1 | Y, \psi_{-1})$,

\vdots

STEP P: $\psi_P \sim p(\psi_P | Y, \psi_{-P})$,

where ψ_{-p} is $(\psi_1, \dots, \psi_{p-1}, \psi_{p+1}, \dots, \psi_P)$, we condition on the most recently sampled value of each element of ψ_{-p} in each STEP p , and at the end of each iteration we compose $\psi^{(t+1)}$ of the

most recently sampled value of each of its components. Throughout, we assume the standard regularity conditions for the Gibbs sampler (see Roberts, 1996; Tierney, 1994, 1996) under which the limiting distribution of $\psi^{(t)}$ is $p(\psi | Y)$.

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

$$p(\psi, \alpha | Y) = p(\psi | Y)p(\alpha | \psi), \quad (4)$$

where $p(\alpha | \psi)$ is a prior distribution on α . The conditional independence assumed in (4) assures that α is a working parameter, i.e., $p(Y | \psi, \alpha) = p(Y | \psi)$. To construct a sampler we introduce a transformation of ψ_p for each p that depends on α : $\tilde{\psi}_p = \mathcal{D}_{\alpha,p}(\psi_p)$, where each $\mathcal{D}_{\alpha,p}$ is again an invertible and differentiable mapping and there exists \mathcal{I}_p such that $\mathcal{D}_{\mathcal{I}_p,p}$ is an identity mapping. Algorithms are constructed by sampling from a set of conditional distributions of $p(\tilde{\psi}, \alpha | Y)$, where $p(\tilde{\psi}, \alpha | Y)$ is obtained from (4) with a change of variable.

This formulation is more general than that used by Meng and van Dyk (1999) in that they assume $p(\psi, \alpha) = p(\psi)p(\alpha)$. The dependence of the prior distribution for α on ψ has some implications on statistical inference. In particular, the marginal posterior distribution of ψ ,

$$p(\psi | Y) \propto p(Y | \psi) \int p(\psi, \alpha) d\alpha, \quad (5)$$

and the conditional posterior distribution of ψ given α ,

$$p(\psi | Y, \alpha) \propto p(Y | \psi)p(\psi | \alpha), \quad (6)$$

may differ because of their respective prior distributions. Although either one of these quantities may be the target distribution, our goal is to construct samplers of $p(\psi | Y)$ with much better convergence properties than the corresponding samplers of $p(\psi | Y, \alpha)$. Of course if ψ and α are *a priori* independent, $p(\psi | Y) = p(\psi | Y, \alpha)$, 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).

2.3 Partially marginalized Gibbs samplers

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

$$p(\tilde{\psi}|Y) = \int p(\tilde{\psi}, \alpha | Y) d\alpha. \quad (7)$$

Because this marginal distribution is often difficult to work with in real problems, we may sample $p(\tilde{\psi}_{(p)} | \tilde{\psi}_{(p)}^c, Y)$ indirectly by sampling $p(\tilde{\psi}_{(p)}, \alpha | \tilde{\psi}_{(p)}^c, Y)$. This strategy may involve first sampling $p(\alpha | \tilde{\psi}_{-p}, Y)$ and then sampling $p(\tilde{\psi}_p, | \tilde{\psi}_{-p}, \alpha, Y)$. The second step is computationally equivalent to sampling $p(\psi_p | \psi_{-p}, \alpha, Y)$, 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 strategy generally avoids the integration in (7).

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 (7), we call the resulting samplers partially marginalized Gibbs (PMG) samplers. To accomplish this, we must introduce partitions of α for each step of the Gibbs sampler. Setting $\alpha = (\alpha_1, \dots, \alpha_J)$, let $\mathcal{J} = \{\mathcal{J}_1, \mathcal{J}_2, \dots, \mathcal{J}_P\}$ be a sequence of index sets, where $\mathcal{J}_p \subset \{1, 2, \dots, J\}$ for $p = 1, 2, \dots, P$. 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 in $\{1, 2, \dots, J\}$ 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: $(\tilde{\psi}_p, \alpha_{(p)}) \sim p(\tilde{\psi}_p, \alpha_{(p)} | Y, \tilde{\psi}_{-p}, \alpha_{(p)}^c)$, 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 value of each 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. 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. In particular, we define another set of partitions of α , $\{\alpha_{(p)}, p = P + 1, \dots, P + P'\}$, in the PMG

sampler.

Partially Marginalized 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 $\psi_p^{(t+1)} = \mathcal{D}_{\alpha_{(t+1)}, p}^{-1}(\tilde{\psi}_p^{(t+1)})$ 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 $\alpha = \mathcal{I}_p$ in STEP p for $p = 1, \dots, P$, 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 resulting sampler is a true MG sampler that completely marginalizes out the working parameter in the sense that all draws are from conditional distributions of (7). 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 an PMG sampler to improve the overall convergence of the sampler. Technical results to this effect are illusive owing to the non-Markovian character of the marginal chain of ψ and of its subcomponents 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 , this may be of limited interest. Thus, we do not pursue a full exploration of application of these theoretical results of 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 3.

Example: To illustrate the advantage of PMG sampler over MDA even in a simple two step sampler, we introduce a second working parameter and a second transformation in the simple Gaussian example. In particular, suppose β is a second scalar working parameter with prior distribution, $\beta \sim N(0, \omega\sigma_2^2)$, *a priori* independent of both ψ and α . We introduce β along with α into the model via a transformation of ψ , namely $\widetilde{\psi} = (\widetilde{\psi}_1, \widetilde{\psi}_2) = (\psi_1 + \alpha, \psi_2 + \beta)$. 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 the optional steps. We start with MDA SCHEME 1 because it is the fastest to converge and consider two possibilities for adding β to this sampler. The two samplers below 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 $(\widetilde{\psi}_2, \alpha) \sim p(\widetilde{\psi}_2, \alpha \mid \widetilde{\psi}_1, \beta)$.

Our basic intuition that sampling as many components of the working parameter as possible in each step is beneficial suggests that PMG SCHEME 1 will dominate both PMG SCHEME 2 and MDA SCHEME 1. The latter comparison is based on the fact that MDA SCHEME 1 implicitly conditions on $\beta = 0$ in both of its steps. Comparing MDA SCHEME 1 in Figure 1 and PMG SCHEME 1 in Figure 2 illustrates the advantage of the PMG sampler. The lag-one autocorrelation of ψ_2 is essentially eliminated by adding the working parameter β to the sampler. ■

2.4 The Advantage of the Optional Steps

Generally, we use the optional steps of a PMG sampler to update components of the working parameter that are not updated in the theorem, it can be much less efficient than updating the working parameter along with ψ .

Theorem 1 *If a PMG sampler is constructed with*

1. ψ and α *a priori* independent,

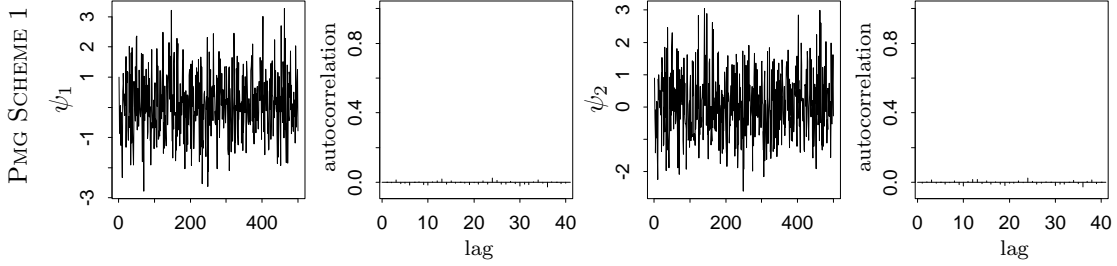


Figure 2: Adding a Second Working Parameter to the Gaussian Samplers. Comparing PMG SCHEME 1 with MDA SCHEME 1 in Figure 1, we see that adding the additional working parameter improves convergence; although MDA SCHEME 1 performs very well, PMG SCHEME 1 produces essentially independent draws of ψ_2 .

2. $\mathcal{D}_{\alpha,p}(\psi_p) = \psi_p$ for $p = 2, \dots, P$, and

3. $\alpha_{(p)} = \emptyset$ for $p = 1, \dots, P$

then the Markov transition kernel for ψ_{-1} , $\mathcal{K}\{\psi_{-1} \mid (\psi_{-1})'\}$, is identical to that of the corresponding standard Gibbs sampler, regardless of $\alpha_{(p)}$ for $p = P + 1, \dots, P'$.

Proof: We aim to show that the transition kernel of $\mathcal{M}_{\psi_{-1}}$,

$$\int \left[\int \mathcal{K} \left\{ \tilde{\psi}_1, \psi_{-1}, \alpha \mid \tilde{\psi}'_1, \psi'_{-1}, \alpha' \right\} d\alpha \right] d\tilde{\psi}_1 \quad (8)$$

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 (8) 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, (8) can be written

$$\begin{aligned} & \int \mathcal{K} \left(\tilde{\psi}_1, \psi_2, \psi_3 \mid \tilde{\psi}'_1, \psi'_2, \psi'_3, \alpha' \right) d\tilde{\psi}_1 \\ &= \int \tilde{p} \left(\tilde{\psi}_1 \mid \psi'_2, \psi'_3, \alpha' \right) p \left(\psi_2 \mid \tilde{\psi}_1, \psi'_3, \alpha' \right) p \left(\psi_3 \mid \tilde{\psi}_1, \psi_2, \alpha' \right) d\tilde{\psi}_1, \end{aligned} \quad (9)$$

where we use a tilde accent on p to emphasize that it represents the density of $\tilde{\psi}_1$ rather than ψ_1 . Rewriting \tilde{p} using the change of variable formula, (9) is equal to

$$\int p \left\{ \mathcal{D}_{\alpha',1}^{-1}(\tilde{\psi}_1) \mid \psi'_2, \psi'_3 \right\} \mid J(\tilde{\psi}_1 \mid \alpha') \mid p \left(\psi_2 \mid \tilde{\psi}_1, \psi'_3, \alpha' \right) p \left(\psi_3 \mid \tilde{\psi}_1, \psi_2, \alpha' \right) d\tilde{\psi}_1, \quad (10)$$

where $J(\tilde{\psi}_1 \mid \alpha)$ is the Jacobian of the inverse transformation $\mathcal{D}_{\alpha',1}^{-1}(\tilde{\psi}_1)$ or 1 if $\tilde{\psi}_1$ is discrete. Finally, by changing the variable of integration via $\psi_1^* = \mathcal{D}_{\alpha',1}^{-1}(\tilde{\psi}_1)$ and by the independence of ψ

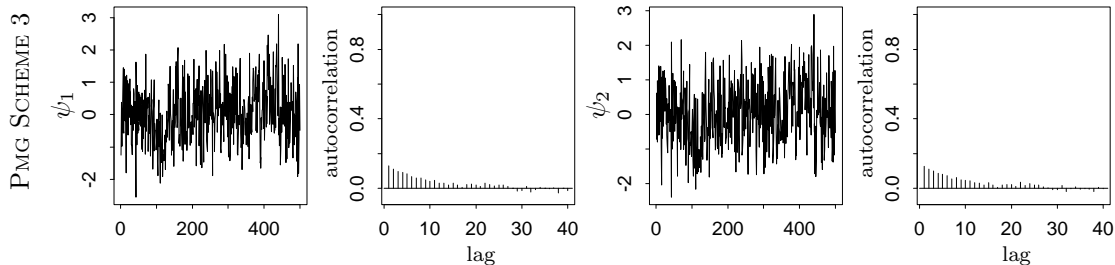


Figure 3: The Benefit of the Optional Steps. In contrast to MDA SCHEME 3, PMG SCHEME 3 offers marked improvement over SCHEME 0.

and α given Y (10) can be written

$$\int p(\psi_1^* | \psi'_2, \psi'_3) p(\psi_2 | \psi_1^*, \psi'_3) p(\psi_3 | \psi_1^*, \psi_2) d\psi_1^*, \quad (11)$$

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

If we define $\psi_1^{(t)} = \mathcal{D}_{\alpha^{(t-1)}, 1}^{-1}(\tilde{\psi}_1^{(t)})$, Theorem 1 can be taken one step further in that \mathcal{M}_ψ can be shown to be 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 in terms of the marginal chain $\mathcal{M}_{\psi_{-1}}$. This is why MDA SCHEME 3 offers no advantage over SCHEME 0 in Figure 1. As we illustrate next, however, the optional steps can be useful for PMG samplers that unlike MDA do not adhere to the suppositions of Theorem 1.

Example: Consider a sampler which 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 $\tilde{\psi}_1 \sim p(\tilde{\psi}_1 | \tilde{\psi}_2, \alpha, \beta)$, $\tilde{\psi}_2 \sim p(\tilde{\psi}_2 | \tilde{\psi}_1, \alpha, \beta)$, and $(\alpha, \beta) \sim p(\alpha, \beta | \tilde{\psi}_1, \tilde{\psi}_2)$.

Figure 3 illustrates the computational performance of PMG SCHEME 3 and shows that it performs much better than MDA SCHEME 3; see Figure 1. Thus, the optional steps can be useful in PMG samplers, that are not MDA samplers. ■

2.5 Optimal Computational Efficiency

As discussed in the Gaussian example in Section 2.1, when it is useful to marginalize out working parameters, more diffuse working prior distributions tend to result in samplers that mix better. 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 (marginal) posterior distribution of the working parameter is

$$p(\alpha | Y) = \int p(\psi, \alpha | Y) d\psi \propto p(\alpha) \int p(Y | \psi) p(\psi) d\psi \propto p(\alpha), \quad (12)$$

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

Example. Returning to the simple Gaussian example, we introduce a new working parameter, $\phi \sim (\text{Gamma}(a_0))^{-1}$ independent of ψ and a transformation $(\hat{\psi}_1, \hat{\psi}_2) = (\sqrt{\phi}\psi_1, \psi_2)$, where a_0 is the shape parameter of the gamma distribution. We consider two sampling schemes:

MDA SCHEME 1: Sample $(\hat{\psi}_1, \phi) \sim p(\hat{\psi}_1, \phi | \psi_2)$ and $(\psi_2, \phi) \sim p(\psi_2, \phi | \hat{\psi}_1)$,

MDA SCHEME 2: Sample $\hat{\psi}_1 \sim p(\hat{\psi}_1 | \psi_2, \phi)$ and $(\psi_2, \phi) \sim p(\psi_2, \phi | \hat{\psi}_1)$.

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 easy to accomplish. For example, to sample $p(\hat{\psi}_1, \phi | \psi_2) = p(\hat{\psi}_1 | \phi, \psi_2) p(\phi)$ we first sample ϕ from its working prior distribution and then sample ψ_1 from $p(\psi_1 | \psi_2)$ and compute $\hat{\psi}_1 = \sqrt{\phi}\psi_1$. 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 $p(\psi_2, \phi | \hat{\psi}_1)$ is proper. Routine calculations indicate that this holds so long as $a_0 > -1/2$.

To illustrate how these samplers work, we implement MDA SCHEME 1 with $a_0 = 5$ and MDA SCHEME 2 with $a_0 = 0$ and $a_0 = -0.2$ 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 4, where the values in curly brackets indicate the value of a_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

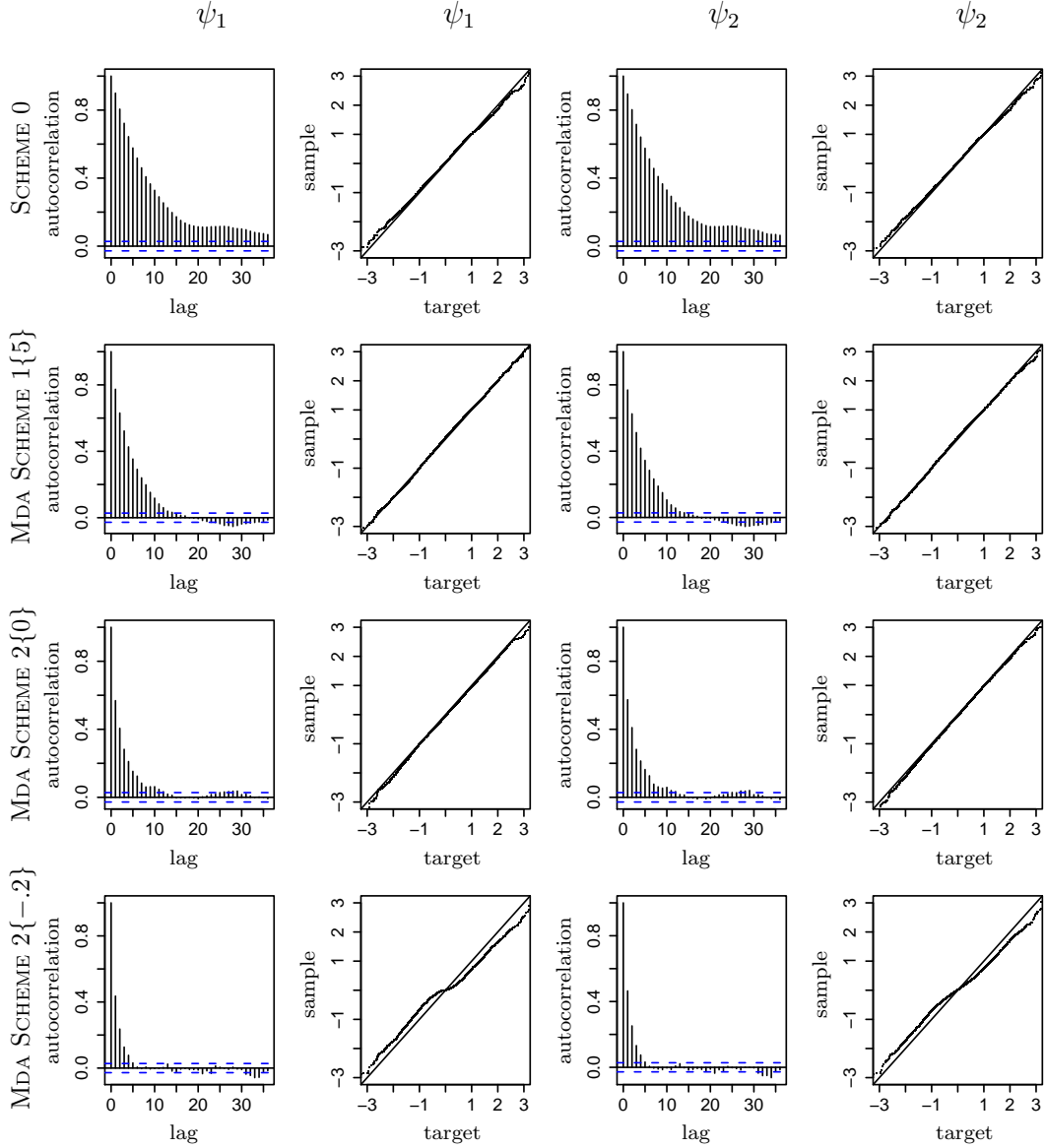


Figure 4: The Risks and Benefits of Using Improper Working Prior Distributions. The figure compares four sampling schemes in the simple Gaussian example. The four rows correspond to a standard sampler with no working parameters, an MDA sampler implemented with a proper working prior distribution and 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%.

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 15} 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. This section of the article 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. ■

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, $\mathcal{K}_m(\xi | \xi')$, each with proper stationary distribution, $\pi_m(\xi)$. If*

(i) $\mathcal{K}_\infty(\xi | \xi') = \lim_{m \rightarrow \infty} \mathcal{K}_m(\xi | \xi')$ *is a proper Markovian transition function and*

(ii) $\pi_\infty(\xi) = \lim_{m \rightarrow \infty} \pi_m(\xi)$ *represents a proper distribution*

then $\pi_\infty(\xi)$ is the stationary distribution of $\mathcal{K}_\infty(\xi | \xi')$.

Proof: By Fatou's lemma, we have,

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

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

We use the symbol ‘ ξ ’ in Lemma 1 because we apply the lemma in several ways in the construction of optimal PMG samplers. Our goal is to establish conditions for PMG samplers that guarantee their stationary distribution is the target posterior distribution when improper working prior distributions are used. We focus on verifying condition **(i)** of Lemma 1 leaving the verification of condition **(ii)** as the standard exercise of establishing that the posterior distribution is integratable under the 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 limit of the stationary distribution is improper (at least when ψ and α are *a priori* improper), it can be applied in some cases to either \mathcal{M}_ψ or subchains of \mathcal{M}_ψ .

All of the results in this section are suppose a sequence of proper working prior distributions $p_m(\alpha)$ are used to construct a corresponding sequence of PMG samplers each with

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

and resulting transition kernel $\mathcal{K}_m(\tilde{\psi}, \alpha \mid \tilde{\psi}', \alpha')$ with proper stationary distribution, $\pi_m(\psi, \alpha)$. The next several results apply Lemma 1 to verify that the limiting Markovian marginal transition kernels of $\mathcal{M}_{\psi_{-1}}$ and \mathcal{M}_ψ have the desired stationary distribution. We label these two results, i.e., the limiting behavior of $\mathcal{M}_{\psi_{-1}}$ and \mathcal{M}_ψ , as R_1 and R_2 , respectively. Figure 5 is a schematic outline of these 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 be used to establish the stronger result, R_2 . Finally, Theorem 2 and Corollary 3 establish conditions C_1^b and C_2^b which imply C_1^a and C_2^a , respectively but are easier to verify in practice. The final results also describe how to construct the optimal sampler. We begin with Corollary 1, which applies Lemma 1 directly to the Markov chains $\mathcal{M}_{\psi_{-1}}$ and \mathcal{M}_ψ .

Corollary 1 *If $\mathcal{M}_{(\tilde{\psi}, \alpha)}(m)$ is sampled with a PMG sampler constructed with ϕ and α a priori independent, $\mathcal{D}_{\alpha, p}(\psi_p) = \psi_p$ for $p = 2, \dots, P$, $\alpha_{(1)} = \alpha$, and proper working prior distribution, $p_m(\alpha)$, and if $p(\psi \mid Y)$ is a proper distribution, then the subchains, $\mathcal{M}_{\psi_{-1}}(m)$ and $\mathcal{M}_\psi(m)$ are*

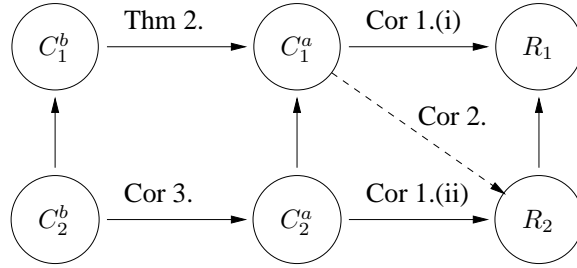


Figure 5: The Theoretical Results of Section 2.5. The conditions of parts **(i)** and **(ii)** of Corollary 1 are represented by C_1^a and C_2^a , respectively; the results of Corollary 1 are represented by R_1 and R_2 . Theorem 2 and Corollary 3 provide conditions that may be easier to verify in practice; 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.

both 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 . Thus,

(i) if $\mathcal{K}_\infty\{\psi_{-1} \mid \psi'_{-1}\} = \lim_{m \rightarrow \infty} \mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ is a proper Markovian transition kernel, then

$p(\psi_{-1} \mid Y)$ is the stationary distribution of $\mathcal{M}_{\psi_{-1}}(\infty)$, the Markov chain sampled under $\mathcal{K}_\infty\{\psi_{-1} \mid (\psi_{-1})'\}$, and,

(ii) if $\mathcal{K}_\infty\{\psi \mid \psi'_{-1}\} = \lim_{m \rightarrow \infty} \mathcal{K}_m\{\psi \mid \psi'_{-1}\}$ is a proper Markovian transition kernel, then

$p(\psi \mid Y)$ is the stationary distribution of $\mathcal{M}_\psi(\infty)$, the Markov chain sampled under $\mathcal{K}_\infty\{\psi \mid \psi'_{-1}\}$.

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 limiting condition for part **(i)** but $p(\psi \mid Y)$ is the target distribution, we can alter the sampler slightly, as described in the following Corollary.

Corollary 2 Consider a sequence of PMG samplers exactly as described in Corollary 1, but with the transformation $STEP\ P + P' + 1$ in each iteration of each sampler replaced with: Draw $\psi_1 \sim p(\psi_1 \mid Y, \psi_{-1})$. In this case, $\mathcal{M}_\psi(m)$ is Markovian for each m , and, if $\mathcal{K}_\infty\{\psi_{-1} \mid \psi'_{-1}\} = \lim_{m \rightarrow \infty} \mathcal{K}_m\{\psi_{-1} \mid \psi'_{-1}\}$ is a proper Markovian kernel, then $p(\psi \mid Y)$ is the stationary distribution of $\mathcal{M}_\psi(\infty)$, the Markov chain sampled under the limiting kernel.

Proof: The transition kernel for $\mathcal{M}_\psi(m)$ 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 $\mathcal{M}_{\psi_{-1}}(m)$ as described in Corollary 1. It follows that $\mathcal{M}_\psi(m)$

is Markovian for each m and that $\lim_{m \rightarrow \infty} p(\psi_1 | Y, \psi_{-1}) \mathcal{K}_m\{\psi_{-1} | \psi'_{-1}\}$ is a proper transition kernel if $\lim_{m \rightarrow \infty} \mathcal{K}_m\{\psi_{-1} | \psi'_{-1}\}$ is. ■

The special case of part (i) of Corollary 1 which results when $P = 2$ and $P' = 0$, was considered by Meng and van Dyk (1999), Liu and Wu (1999), and van Dyk and Meng (2001). 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 a sequence of PMG samplers as described in Corollary 1, but also with $\alpha_{(2)} = \alpha$. If*

- (i) *there exists a improper working prior distribution $p_\infty(\alpha)$ such that $p_m(\psi_p, \alpha_{(p)} | Y, \tilde{\psi}_{-p}, \alpha_{(p)}^c) \rightarrow p_\infty(\psi_p, \alpha_{(p)} | Y, \tilde{\psi}_{-p}, \alpha_{(p)}^c)$ as $m \rightarrow \infty$ for $p = 2, \dots, P$, 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) *$\int \mathcal{K}_\infty\{\psi_{-1}, \alpha | \mathcal{D}_{\alpha^*, 1}(\psi_1), \psi'_{-1}\} d\alpha$ is invariant to α^* , where $\mathcal{K}_\infty\{\psi_{-1}, \alpha | \tilde{\psi}_1, \psi'_{-1}\}$ denotes the transition kernel for STEP 2 through STEP P of the PMG sampler implemented with the improper working prior distribution, $p_\infty(\alpha)$,*

then $\lim_{m \rightarrow \infty} \mathcal{K}_m\{\psi_{-1} | \psi'_{-1}\}$ is the proper transition kernel, which 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 $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 | Y, \psi'_{-1}, \alpha = \mathcal{I}_1\}$.

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

$$\int \tilde{p}_m(\tilde{\psi}_1, \alpha^* | \psi'_2, \psi'_3) p_m(\psi_2, \alpha_{(3)}^{**}, \alpha_{(3)}^c | \tilde{\psi}_1, \psi'_3) p_m(\psi_3, \alpha_{(3)} | \tilde{\psi}_1, \psi_2, \alpha_{(3)}^c) d\tilde{\psi}_1 d\alpha_{(3)}^{**} d\alpha, \quad (13)$$

where the accent on \tilde{p} emphasizes that this is the conditional density of $\tilde{\psi}_1$ rather than ψ_1 . Integrating out $\alpha_{(3)}^{**}$, replacing $\tilde{p}_m(\tilde{\psi}_1, \alpha^* | \psi'_2, \psi'_3)$ with $p_m(\alpha^*) p_m\{\mathcal{D}_{\alpha^*,1}^{-1}(\tilde{\psi}_1) | \psi'_2, \psi'_3\} |J(\tilde{\psi}_1 | \alpha^*)|$, and changing the variable of integration via $\psi_1^* = \mathcal{D}_{\alpha^*,1}^{-1}(\tilde{\psi}_1)$ in (13) yields

$$p_m(\alpha^*) \int p_m(\psi_1^* | \psi'_2, \psi'_3) p_m\{\psi_2, \alpha_{(3)}^c | \mathcal{D}_{\alpha^*,1}(\psi_1^*), \psi'_3\} p_m\{\psi_3, \alpha_{(3)} | \mathcal{D}_{\alpha^*,1}(\psi_1^*), \psi_2, \alpha_{(3)}^c\} d\psi_1^* d\alpha. \quad (14)$$

By Fatou's lemma and condition **(i)**, the limit as $m \rightarrow \infty$ of the integral in (14) is

$$\int p\{\psi_1^* | \psi'_{-1}\} \mathcal{K}_\infty\{\psi_{-1}, \alpha | \mathcal{D}_{\alpha^*,1}(\psi_1^*), \psi'_{-1}\} d\psi_1^* d\alpha. \quad (15)$$

By condition **(ii)**, we may replace α^* under the integral by the identity value, \mathcal{I}_1 . Thus, in the limit α^* and ψ_{-1} are independent, the kernel, $\mathcal{K}_\infty\{\psi_{-1} | \psi'_{-1}\}$ is the proper distribution given in (15) with α^* replaced with \mathcal{I}_1 , and this kernel is identical to that resulting from implementing the PMG sampler with $p_\infty(\alpha)$, but with STEP 1 replaced with: $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 | Y, \psi'_{-1}, \alpha = \mathcal{I}_1\}$. ■

We can derive similar equivalent conditions for part **(ii)** of Corollary 1. Rather than integrating out $\tilde{\psi}_1$ in (13), we introduce the change of variable $\psi_1 = \mathcal{D}_{\alpha,1}^{-1}(\tilde{\psi}_1)$ and adjust the the proof of Theorem 2 appropriately, proving the following corollary. (A formal proof is given in Appendix B for the benefit of the referees.)

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

$$\begin{aligned} \text{(iii)} \quad & \int p_m [\mathcal{D}_{\alpha^*,1}^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1)\} | \psi'_{-1}] | J\{\mathcal{D}_{\alpha,1}(\psi_1|\alpha^*)\} J^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1) | \alpha\} | \\ & \times \mathcal{K}_\infty\{\psi_{-1}, \alpha | \mathcal{D}_{\alpha,1}(\psi_1), \psi'_{-1}\} d\alpha \text{ is invariant to } \alpha^*, \end{aligned}$$

then $\lim_{m \rightarrow \infty} \mathcal{K}_m(\psi | \psi')$ is the proper transition kernel, which 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 $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 | Y, \psi'_{-1}, \alpha = \mathcal{I}_1\}$.

Applying the change of variable $\psi_1^* = \mathcal{D}_{\alpha^*,1}^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1)\}$ 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 supposition **(ii)** when applying the corollary. Supposition **(iii)** is an algebraic representation of the transition kernel, $\mathcal{K}_\infty(\psi|\psi', \alpha^*)$, under the PMG sampler, where α^* is the draw of α in STEP 1. Thus, we generally verify the supposition by

verifying that $\mathcal{K}_\infty(\psi|\psi', \alpha^*)$ does not depend on α^* . This strategy is illustrated in Section 3 where we verify the limiting kernel for the logistic mixed model.

Example: In the Gaussian example described in Section 2.1 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 $\tilde{\psi}_1 \sim N(\varrho\psi'_2, 1 - \varrho^2 + \omega^2)$.

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

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

In the limit as $\omega^2 \rightarrow \infty$, the draw in STEP 1 becomes improper, but (ψ_1, ψ_2) do 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 $\psi_2 \sim N(0, 1)$ and $\alpha \mid \tilde{\psi}_1, \psi_2 \sim N(\tilde{\psi}_1 - \varrho\psi_2, 1 - \varrho^2)$. 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}_ψ 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 $\psi_1 \mid \tilde{\psi}_1, \psi_2 \sim N(\varrho\psi_2, 1 - \varrho^2)$. Thus the limiting kernel for \mathcal{M}_ψ is proper and by part (ii) of Corollary 1, has the desired stationary distribution, as is again evident. ■

2.6 Marginal MCMC

Although the theoretical results in Section 2.5 are useful for establishing the stationary distribution of PMG samplers with improper working prior distributions, verifying the convergence of the marginal kernels can be difficult in practice, especially in samplers with many steps. Thus, we aim to construct samplers that segregates the computational complexity of the sampler into a number of simpler parts. In particular, consider:

Nested MCMC Gibbs Sampler

STEP 1: $\psi_1^{(t+1)} \sim \mathcal{K}(\psi_1 | \psi'_1; \psi_{-1}),$

\vdots

STEP P: $\psi_P^{(t+1)} \sim \mathcal{K}(\psi_P | \psi'_P; \psi_{-P}),$

where $\mathcal{K}(\psi_p | \psi'_p; \psi_{-p})$ is a transition kernel for an irreducible aperiodic Markov chain with unique stationary distribution $p(\psi_p | Y, \psi_{-p})$. As usual, we condition on the most recently sampled value of each component of ψ at each step.

Clearly the resulting chain, \mathcal{M}_ψ , is an irreducible aperiodic Markov chain with unique stationary distribution $p(\psi | Y)$. The kernels in the individual step may be formulated as direct draws from the conditional distribution, $p(\psi_p | Y, \psi_{-p})$, or by using a PMG sampler with the correct stationary distribution for \mathcal{M}_{ψ_p} . If an improper working prior distribution is used, this can be verified using part **(ii)** of Corollary 1 or using Corollary 2. 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 in Section 3.

3 Example: Logistic Mixed Model

3.1 Model Specification

In this section we nest several PMG samplers within a Gibbs sampler constructed using auxiliary variables to fit a logistic mixed model. 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 algorithms that are fast to converge.

Consider the logistic model

$$y_{ij} \sim \text{Bernoulli}(p_{ij}) \text{ with } \text{logit}(p_{ij}) = x'_{ij}(\beta + b_i), \quad (16)$$

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 $i = 1, \dots, m$ and $j = 1, \dots, n_i$, i.e., there m groups with sizes n_1, \dots, n_m . We assume the random effects are independently distributed, $b_i \sim N(0, T)$, with T a diagonal matrix with diagonal elements $(\tau_1^2, \dots, \tau_q^2)$ and independent prior distributions $\tau_k^2 \sim \nu_0 \tau_{k,0}^2 / \chi_{\nu_0}^2$; we use a flat prior on the fixed effects.

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

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

where $Y = (y_{ij}, i = 1, \dots, m, j = 1, \dots, n_i)$. 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 introduced by Damien, Wakefield, and Walker (1999) and then show how to improve computational performance using marginal methods.

3.2 The Standard Slice Sampler

Following Damien *et al.* (1999), we suppose

$$y_{ij} = I[v_{ij} \leq g^{-1}\{x'_{ij}(\beta + b_i)\}], \quad (17)$$

where, I is an indicator function, g^{-1} is the inverse logistic link function, and $v_{ij} \sim \text{Unif}(0, 1)$. Model (17) is simply a reformulation of (16), which introduces the auxiliary variable $V = (v_{ij}, i = 1, \dots, m, j = 1, \dots, n_i)$. Thus, (16) represents $p(Y | \beta, b) = \int p(Y, V | \beta, b) dV$ and (17) represents the integrand in the second expression. Using (17) we can construct a slice sampler:

Slice Sampler for the Logistic Mixed Model

STEP 1: For each i and j , independently draw

$$v_{ij} | Y, b, \beta, T \sim \begin{cases} \text{Unif}[0, g^{-1}\{x'_{ij}(\beta + b_i)\}] & \text{if } y_{ij} = 1 \\ \text{Unif}[g^{-1}\{x'_{ij}(\beta + b_i)\}, 1] & \text{if } y_{ij} = 0. \end{cases} \quad (18)$$

STEP 2: For $k = 1, \dots, q$, sample (b_k, β_k, τ_k^2) via two conditional draws, where $b_k = (b_{1k}, \dots, b_{mk})$;

here b_{ik} and β_k represent component k of b_i and β , respectively. Although it is suppressed

in our notation, we condition on Y , V , and all of the components except the k th of β and each b_i :

CYCLE k

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

$$\begin{cases} v_{ij} \leq g^{-1}\{x'_{ij}(\beta + b_i)\} & \text{if } y_{ij} = 1 \\ v_{ij} > g^{-1}\{x'_{ij}(\beta + b_i)\} & \text{if } y_{ij} = 0, \end{cases} \quad \text{for } j = 1, \dots, n_i. \quad (19)$$

SUBSTEP 2: Draw $(\beta_k, \tau_k^2) \mid b_{\cdot k}$ by independently drawing β_k given $b_{\cdot k}$ uniformly subject to (19) and $\tau_k^2 \mid b_{\cdot k} \sim (\nu_0 \tau_{k,0}^2 + \sum_{i=1}^m b_{ik}) / \chi_{m+\nu_0}^2$.

This sampler consists of $2q + 1$ complete conditional draws; see Damien *et al.* (1999) for details.

3.3 Marginal Slice Sampling in the Logistic Mixed Model

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 3.4, 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 $\alpha = (\alpha_1, \dots, \alpha_q)'$ is a $q \times 1$ working parameter. This transformation results in a reformulation of model (17),

$$y_{ij} = I \left[v_{ij} \leq g^{-1}\{x'_{ij}(\tilde{\beta} + \tilde{b}_i)\} \right], \quad \text{with } \tilde{b}_i \sim N(\alpha, T), \quad (20)$$

where $\tilde{\beta} = \beta - \alpha$. Using a flat working prior distribution, we can easily incorporate α_k into cycle k of STEP 2:

SUBSTEP 1: For each i , independently draw $b_{ik}^* \mid \beta_k, \tau_k^2 \sim N(0, \tau_k^2)$ subject to (19) with b_i replaced with b_i^* .

SUBSTEP 2: Draw $(\beta_k, \tau_k^2, \alpha_k) \mid b_{\cdot k}^*$ by sampling $\tau_k^2 \sim \left\{ \sum_{i=1}^m (b_{ik}^* - b_{\cdot k}^*)^2 + \nu_k \tau_{k,0}^2 \right\} / \chi_{m+\nu_0-1}^2$, $\alpha_k \mid \tau_k^2 \sim N(b_{\cdot k}^*, \tau_k^2/m)$, and $\tilde{\beta}_k$ uniformly subject to (19) with b_i replaced with b_i^* , here $b_{\cdot k}^* = \sum_{i=1}^m b_{ik}^*/m$. Finally we transform to the original scale by setting $\beta_k = \tilde{\beta}_k + \alpha_k$ and $b_i = b_i^* - \alpha_k$.

The stationary distribution of this sampler is verified in Appendix A; its computational advantage is illustrated in Section 3.4.

3.4 Empirical Results

We illustrate the convergence properties of the 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 (16), each with $m = 11$, $\sum_i n_i = 54$, and n_i varying between 4 and 5. The covariates, x_{ij} , were independently generated via $x_{ij} \sim N(0, 1)$ 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 $\beta = 0, 1.5, \text{ and } 10$. 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 a location working parameter and improper working prior distribution as described above. We fit model (16) to each of the three data sets using both samplers; each chain was initialized at $\beta^{(0)} = 0.5, (\tau^2)^{(0)} = 1$. The first 500 draws of each chain is illustrated in Figure 6. The marginal algorithm significantly improves the the autocorrelation of the Markov chains. Quantile-quantile plot comparing the Markov chains generated with the two samplers verify that they have the same stationary distributions; 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 (16) with $m = 11$, $\sum_i n_i = 54$, and n_i varying between 4 and 5. The covariates, x_{ij} , were independently generated via $x_{ij} \sim N_2(0, I)$ 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 7, which clearly illustrates the computational advantage of using both working parameters.

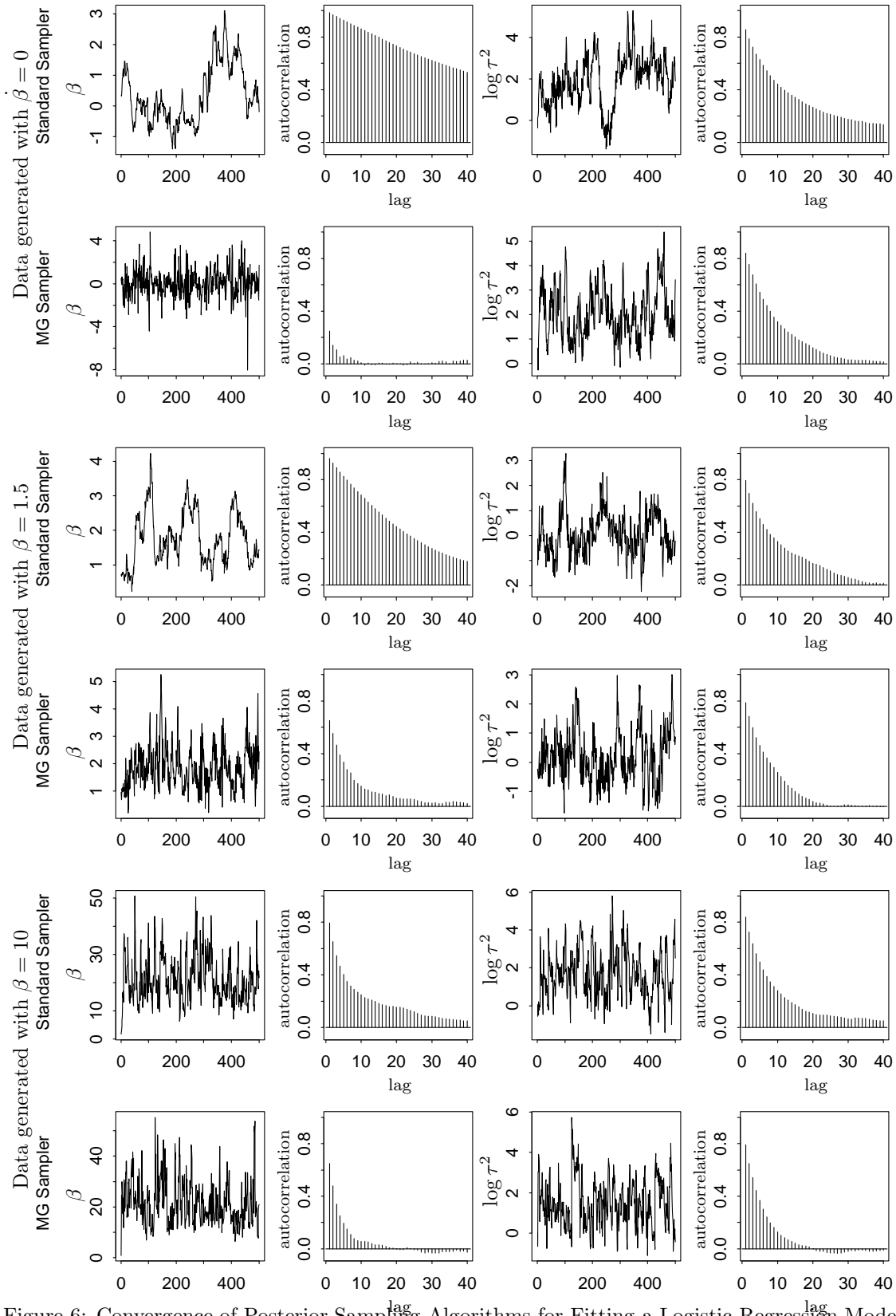


Figure 6: Convergence of Posterior Sampling Algorithms for Fitting a Logistic Regression Model with One Covariate. The first two rows compare the standard slice sampler with the PMG sampler for the data set generated with $\beta = 0$, the second two rows compare the samplers for the data generated with $\beta = 1.5$, and the the final two rows compare the samplers for the 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.

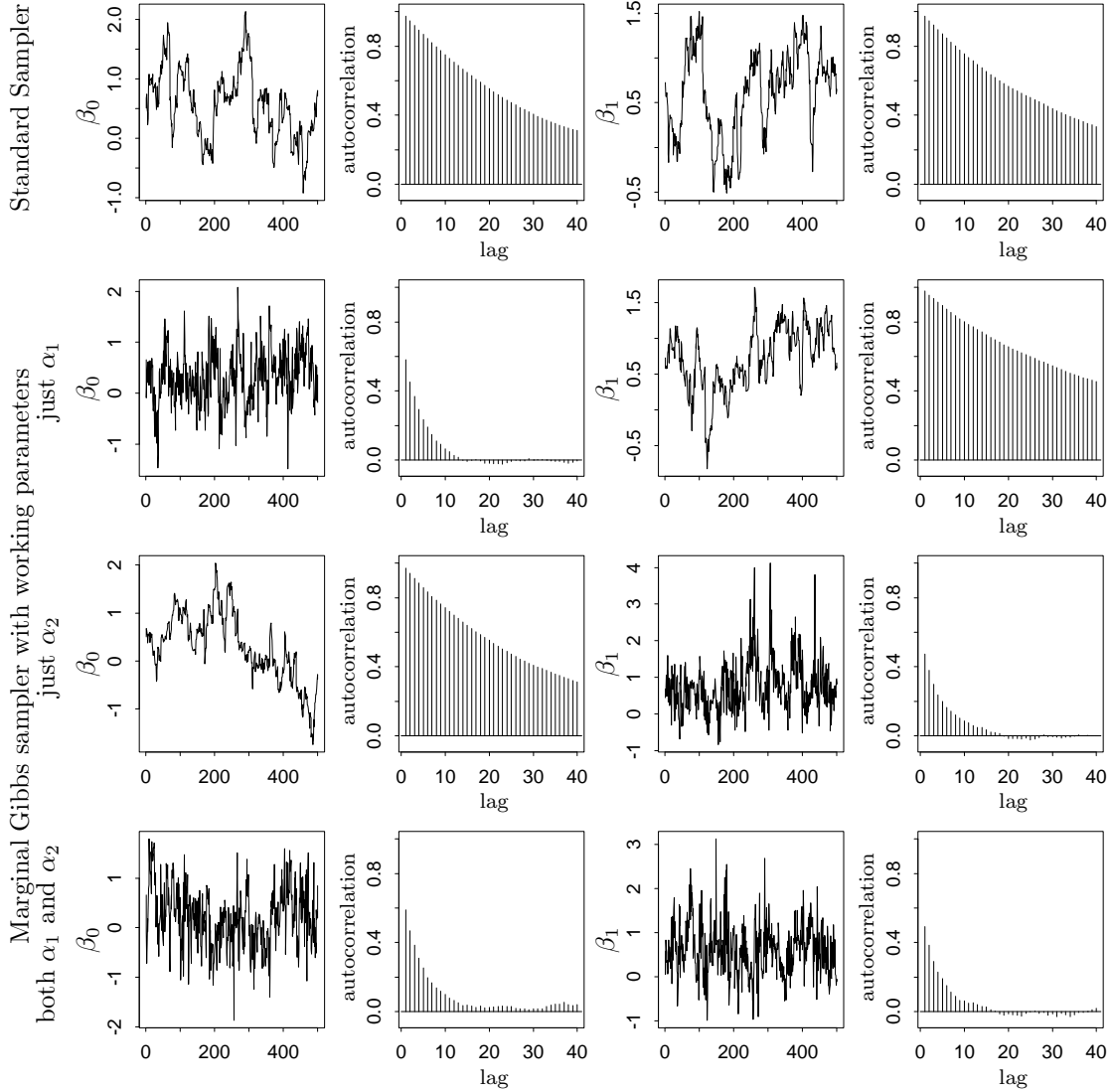


Figure 7: Convergence of Posterior Sampling Algorithms for Fitting a Logistic Regression Model with Two Covariates. The rows of the of the figure correspond to the standard slice sampler, a PMG sampler implemented with a location working parameter for the first random effect, a PMG sampler implemented 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.

4 Concluding remarks

The transformation that we use to introduced the working parameter into the model are componentwise transformations. That is we insist on setting $\tilde{\psi}_p = \mathcal{D}_{\alpha,p}(\psi_p)$ 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 Gaussian example once again. Consider transforming (ψ_1, ψ_2) to $\{\psi_1 - \mu_1 - \alpha(\psi_2 - \mu_2), \psi_2\}$ in the bivariate Gaussian example. 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 components 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.

In the more general framework, 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 than 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 to determine a good choice of the distribution. This is the strategy that we used, for example in Figure 4.

Appendix

A The Optimal Sampler for the Logistic Mixed Model

Here we verify that the marginal slice sampler given in Section 3.3 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, $\mathcal{K}_\omega\{b_{\cdot,k}, \beta_k, \tau_k^2 \mid b'_{\cdot,k}, \beta'_k, (\tau_k^2)'\}$, 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, $\alpha \sim N(0, \omega^2 I)$, with I the identity matrix. We verify that the stationary distribution of the transition kernel, $\lim_{\omega \rightarrow \infty} \mathcal{K}_\omega\{b_{\cdot,k}, \beta_k, \tau_k^2 \mid b'_{\cdot,k}, \beta'_k, (\tau_k^2)'\}$, is $p(b_{\cdot,k}, \beta_k, \tau_k^2)$; here and throughout the appendix we suppress conditioning on Y, V , and the components other than the k th 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:

SUBSTEP 1: Sample $\tilde{b}_{\cdot,k}, \alpha_k \mid \beta_k, \tau_k^2$ by independently sampling $\alpha_k^* \sim N(0, \omega^2)$ and $b_i^* \sim \text{TN}\{0, \tau_k^2, L(b_{ik}), U(b_{ik})\}$ for each i , where

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

and

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

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

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

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

where $\tilde{b}_{\cdot,k} = \frac{1}{m} \sum_{i=1}^m \tilde{b}_{ik}$. Clearly, $p_\omega(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{\cdot,k}) \rightarrow p_\infty(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{\cdot,k})$ as $\omega \rightarrow \infty$, 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 $p_\infty(\tilde{\beta}_k, \tau_k^2, \alpha_k \mid \tilde{b}_{\cdot,k})$ by sampling

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

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

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

and

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

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

B 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 (ψ, α^*) with α^* the draw of α from STEP 1, i.e., $\mathcal{K}_m\{\psi, \alpha^* \mid \psi'_{-1}\}$ can be written

$$\begin{aligned} \int \tilde{p}_m \{ \mathcal{D}_{\alpha,1}(\psi_1), \alpha^* \mid \psi'_2, \psi'_3 \} \mid J^{-1}\{ \mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha \} \mid p_m \{ \psi_2, \alpha_{(3)}^{**}, \alpha_{(3)}^c \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi'_3 \} \\ \times p_m \{ \psi_3, \alpha_{(3)} \mid \mathcal{D}_{\alpha,1}(\psi_1), \psi_2, \alpha_{(3)}^c \} \mid d\alpha_{(3)}^{**} d\alpha, \end{aligned} \quad (21)$$

where the accent on \tilde{p} emphasizes that this is the conditional density of $\tilde{\psi}_1$ rather than ψ_1 . Integrating out $\alpha_{(3)}^{**}$ and replacing $\tilde{p}_m(\tilde{\psi}_1, \alpha^* \mid \psi'_2, \psi'_3)$ with $p_m(\alpha^*) p_m\{\mathcal{D}_{\alpha^*,1}^{-1}(\tilde{\psi}_1) \mid \psi'_2, \psi'_3\} \mid J(\tilde{\psi}_1 \mid \alpha^*) \mid$, in (21) yields

$$\begin{aligned} p_m(\alpha^*) \int p_m \left[\mathcal{D}_{\alpha^*,1}^{-1}\{ \mathcal{D}_{\alpha,1}(\psi_1) \} \mid \psi'_2, \psi'_3 \right] \mid J \{ \mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha^* \} \mid J^{-1}\{ \mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha \} \mid \\ \times p_m \left\{ \psi_2, \alpha_{(3)}^c \mid \mathcal{D}_{\alpha,1}(\psi_1^*), \psi'_3 \right\} \mid p_m \left\{ \psi_3, \alpha_{(3)} \mid \mathcal{D}_{\alpha,1}(\psi_1^*), \psi_2, \alpha_{(3)}^c \right\} \mid d\alpha. \end{aligned} \quad (22)$$

By Fatou's lemma and condition **(i)** of Theorem 2, the limit as $m \rightarrow \infty$ of the integral in (22) is

$$\int p_m \left[\mathcal{D}_{\alpha^*,1}^{-1}\{ \mathcal{D}_{\alpha,1}(\psi_1) \} \mid \psi'_2, \psi'_3 \right] \mid J \{ \mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha^* \} \mid J^{-1}\{ \mathcal{D}_{\alpha,1}(\psi_1) \mid \alpha \} \mid$$

$$\times \mathcal{K}_\infty\{\psi_1^c, \alpha \mid \mathcal{D}_{\alpha,1}(\psi_1^*), (\psi_1^c)'\} d\alpha. \quad (23)$$

By condition **(iii)**, we may replace α^* under the integral by the identity value, \mathcal{I}_1 . Thus, in the limit α^* and ψ_1 are independent, the kernel, $\mathcal{K}_\infty\{\psi \mid \psi_{-1}pr\}$ is the proper distribution given in (23) with α^* replaced with \mathcal{I}_1 , and this kernel is identical to that resulting from implementing the PMG sampler with $p_\infty(\alpha)$, but with STEP 1 replaced with: $\tilde{\psi}_1 \sim p\{\tilde{\psi}_1 \mid Y, \psi'_{-1}, \alpha = \mathcal{I}_1\}$. ■

To verify that condition **(iii)** implies condition **(ii)** of Theorem 2, we apply the change of variable, $\psi_1^* = \mathcal{D}_{\alpha^*,1}^{-1}\{\mathcal{D}_{\alpha,1}(\psi_1)\}$ in the density given in (23); the Jacobian of the transformation is $\left| J^{-1}\{\mathcal{D}_{\alpha^*,1}(\psi_1^*) \mid \alpha^*\} J\{\mathcal{D}_{\alpha^*,1}(\psi_1^*) \mid \alpha\} \right|$.

References

- Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 25–37.
- Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B* **61**, 331–344.
- Edwards, R. and Sokal, A. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review Series 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. *Journal of Computational and Graphical Statistics* **17**, 95–122.
- Ghosh, J. and Dunson, D. (2008). Default priors and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics* To appear.
- Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association* **93**, 585–595.
- Imai, K. and van Dyk, D. A. (2005a). A Bayesian analysis of the multinomial probit model using marginal augmentation. *The Journal of Econometrics* **124**, 311–334.
- Imai, K. V. and van Dyk, D. A. (2005b). MNP: R package for fitting multinomial probit models. *The Journal of Statistical Software* **15**, Issue 5.
- Liu, C. (2003). Alternating subspace-spanning resampling to accelerate Markov Chain Monte Carlo simulation. *Journal of the American Statistical Association* **98**, 110–117.
- 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 scheme for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.
- Marchev, D. and Hobert, J. P. (2004). Geometric ergodicity of van Dyk and Mengs algorithm for the multivariate students t model. *Journal of the American Statistical Association* **99**, 228–238.

- McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **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). *Journal of the Royal Statistical Society, Series B, Methodological* **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. *Statistics and Computing* **8**, 229–242.
- Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (Editors: W. R. Gilks, S. Richardson, and D. J. Spiegelhalter). Chapman & Hall, London.
- 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.
- 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). Chapman & Hall, London.
- 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* (Editors: 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). *The Journal of Computational and Graphical Statistics* **10**, 1–111.
- van Dyk, D. A. and Meng, X.-L. (2008). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science*, under review.
- van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, to appear.
- Yu, Y. and Meng, X. L. (2008). Espousing classical statistics with modern computation: Sufficiency, ancillarity, and an interweaving generation of MCMC. *Technical Report, Department of Statistics, Harvard University* .