Journal of Computational and Graphical Statistics

Metropolis-Hastings within Partially Collapsed Gibbs Samplers

David A. van Dyk[a] & Xiyun Jiao[b]

[a] Department of Mathematics at Imperial College, London, SW7 2AZ

[b] postgraduate student in Statistics at Imperial College

Accepted author version posted online: 20 Jun 2014.

PLEASE SCROLL DOWN FOR ARTICLE

Metropolis-Hastings within Partially Collapsed Gibbs Samplers

David A. van Dyk and Xiyun Jiao[*]

Abstract

The Partially Collapsed Gibbs (PCG) sampler offers a new strategy for improving the convergence of a Gibbs sampler. PCG achieves faster convergence by reducing the conditioning in some of the draws of its parent Gibbs sampler. Although this can significantly improve convergence, care must be taken to ensure that the stationary distribution is preserved. The conditional distributions sampled in a PCG sampler may be incompatible and permuting their order may upset the stationary distribution of the chain. Extra care must be taken when Metropolis-Hastings (MH) updates are used in some or all of the updates. Reducing the conditioning in an MH within Gibbs sampler can change the stationary distribution, even when the PCG sampler would work perfectly if MH were not used. In fact, a number of samplers of this sort that have been advocated in the literature do not actually have the target stationary distributions. In this article, we illustrate the challenges that may arise when using MH within a PCG sampler and develop a general strategy for using such updates while maintaining the desired stationary distribution. Theoretical arguments provide guidance when choosing between different MH within PCG sampling schemes. Finally we illustrate the MH within PCG sampler and its computational advantage using several examples from our applied work.

Key Words: Astrostatistics; Blocking; Factor Analysis; Gibbs sampler; Incompatible Gibbs sampler; Metropolis-Hastings; Metropolis within Gibbs; Spectral Analysis.

1 Introduction

The popularity of the Gibbs sampler stems from its simplicity and power to effectively generate samples from a high-dimensional probability distribution. It can sometimes, however, be very slow to converge, especially when it is used to fit highly structured or complex models. The Partially Collapsed Gibbs (PCG) sampler offers a strategy for improving the convergence characteristics of a Gibbs sampler (van Dyk and Park, 2008; Park and van Dyk, 2009; van Dyk and Park, 2011). A PCG sampler achieves faster convergence by reducing the conditioning in some or all of the component draws of its parent Gibbs sampler. That is, one or more of the complete

[*]Professor David A. van Dyk holds a Chair in Statistics in the Department of Mathematics at Imperial College London, SW7 2AZ (dvandyk@imperial.ac.uk); Xiyun Jiao is a postgraduate student in Statistics at Imperial College.

conditional distributions is replaced by the corresponding conditional distribution of a multivariate marginal distribution of the target. For example, we might consider sampling $p(\psi_1|\psi_2)$ rather than $p(\psi_1|\psi_2,\psi_3)$, where $p(\psi_1|\psi_2)$ is a conditional distribution of the marginal distribution, $p(\psi_1,\psi_2)$, of the target $p(\psi_1,\psi_2,\psi_3)$. This strategy has already been proven useful in improving the convergence properties of numerous samplers (e.g., Bernardi *et al.*, 2013; Berrett and Calder, 2012; Caron *et al.*, 2014; Dobigeon and Tourneret, 2010; Hans *et al.*, 2012; Hu *et al.*, 2012, 2013; Kail *et al.*, 2010, 2011; Lin and Tourneret, 2010; Lindsten *et al.*, 2013; Park *et al.*, 2008; Park and van Dyk, 2009; Park, 2011; Park *et al.*, 2012a,b; Zhao and Lian, 2014, etc.).

Although the PCG sampler can be very efficient, it must be implemented with care to make sure that the stationary distribution of the resulting sampler is indeed the target. Unlike the ordinary Gibbs sampler, the conditional distributions sampled in a PCG sampler may be incompatible, meaning there is no joint distribution of which they are simultaneously the conditional distributions. In this case, permuting the order of the updates can change the stationary distribution of the chain.

As with an ordinary Gibbs sampler, we sometimes find that one or more of the conditional draws of a PCG sampler is not available in closed form and we may consider implementing such draws with the help of a Metropolis-Hastings (MH) sampler. Reducing the conditioning in one draw of an MH within Gibbs sampler, however, may alter the stationary distribution of the chain. This can happen even when the PCG sampler would work perfectly well if all of the conditional updates were available without resorting to MH updates. Examples arise even in a two-step MH within PCG sampler. Woodard *et al.* (2012), for example, points out this problem in certain samplers described in the literature for regression with functional predictors. Although they do not use the framework of PCG, these samplers are simple special cases of improper MH within PCG samplers. They first analyze the functional predictors in isolation of the regression and then use MH to update the regression parameters conditional on parameters describing the functional predictors. The first step effectively samples the functional parameters marginally and the second uses MH for sampling from the complete conditional of the regression parameters. In this article we pay special attention to this situation because it is both conceptually simple and important in practice. In Section 3.2 we propose two simple strategies that maintain the target distribution and in Section 4

we compare the performance of the two strategies theoretically.

In this article, we illustrate difficulties that may arise when using MH updates within a PCG sampler and develop a general strategy for using such updates while maintaining the target stationary distribution. We begin in Section 2 with two motivating examples that are chosen to review the subtleties of the PCG sampler, illustrate the complications that arise when MH is introduced into PCG, and set the stage for the methodological and theoretical contributions of this article. Section 2 ends by reviewing the method of van Dyk and Park (2008) for establishing the stationary distribution of a PCG sampler. The MH within PCG sampler is introduced in Section 3 along with methods for ensuring that its stationary distribution is the target distribution and several strategies for implementing the sampler while maintaining this target. Theoretical arguments are presented in Section 4 with the aim of guiding the choice between different implementations of the MH within PCG sampler. The proposed methods and theoretical results are illustrated in Section 5 in the context of several examples, including factor analysis and two examples from high-energy astrophysics. The factor analysis example contrasts the step-ordering constraints of MH within PCG and of the related ECME algorithm (Liu and Rubin, 1994). Final discussion appears in Section 6 and technical details of several of the MH within PCG samplers appear in an online appendix.

2 Background and Motivating Examples

2.1 Notation

We aim to sample from the target distribution, $p(\psi)$, by constructing a Markov chain $\{\psi^{(t)}, t = 1, 2, \dots\}$ with the stationary distribution $\pi(\psi)$, where ψ is a multivariate random variable. That is, we aim to construct a Markov chain such that $\pi(\psi) = p(\psi)$. We refer to a sampler as *proper* if it has a stationary distribution and that distribution coincides with the target, i.e., $\pi(\psi) = p(\psi)$; otherwise we call the sampler *improper*. Typically $p(\psi)$ is the posterior distribution in a Bayesian analysis, but this is not necessary. In data-driven examples, we use standard Bayesian notation.

To facilitate discussion of the relevant samplers, we divide ψ into J possibly multivariate non-overlapping subcomponents, i.e., $\psi = (\psi_1, \dots, \psi_J)$, and define $\mathscr{J} = \{1, 2, \dots, J\}$. The methods that we consider are Gibbs-type samplers that rely on the conditional distributions of either $p(\psi)$

or its multivariate marginal distributions. When conditional distributions cannot be sampled directly, we may use MH. For example, suppose we wish to sample the conditional distribution $p(\psi_{j_1}|\psi_{j_2})$ of the marginal distribution $p(\psi_{j_1}, \psi_{j_2})$, but cannot do so directly. In this case, we specify a jumping rule (i.e., a proposal distribution), denoted by $\mathcal{J}_{j_1|j_2}(\psi_{j_1}|\psi'_{j_1}, \psi'_{j_2}, \psi'_{j_3})$, where the subscript specifies the target conditional distribution and we use primes to indicate the current value of the subcomponents of ψ; notice that the jumping rule may depend on subcomponents other than ψ'_{j_1} and ψ'_{j_2}, namely, ψ'_{j_3}. In the MH update, we sample $\psi_{j_1}^{\text{prop}} \sim \mathcal{J}_{j_1|j_2}(\psi_{j_1}|\psi'_{j_1}, \psi'_{j_2}, \psi'_{j_3})$ and set $\psi_{j_1} = \psi_{j_1}^{\text{prop}}$ with probability $r = \min\left\{1, \dfrac{p(\psi_{j_1}^{\text{prop}}|\psi'_{j_2})\mathcal{J}_{j_1|j_2}(\psi'_{j_1}|\psi_{j_1}^{\text{prop}}, \psi'_{j_2}, \psi'_{j_3})}{p(\psi'_{j_1}|\psi'_{j_2})\mathcal{J}_{j_1|j_2}(\psi_{j_1}^{\text{prop}}|\psi'_{j_1}, \psi'_{j_2}, \psi'_{j_3})}\right\}$; otherwise the current value is retained, i.e., $\psi_{j_1} = \psi'_{j_1}$. This MH transition kernel, denoted by $\mathcal{M}_{j_1|j_2}(\psi_{j_1}|\psi'_{j_1}, \psi'_{j_2}, \psi'_{j_3})$, has stationary distribution $p(\psi_{j_1}|\psi_{j_2})$. We can also express the iterates explicitly. For instance, $\psi_2^{(t+1)} \sim \mathcal{M}_{2|1,3}(\psi_2|\psi_1^{(t+1)}, \psi_2^{(t)}, \psi_3^{(t)})$ is a typical expression for sampling from an MH transition kernel with stationary distribution $p(\psi_2|\psi_1^{(t+1)}, \psi_3^{(t)})$. Notice that this transition kernel depends on $\psi_2^{(t)}$ because the acceptance probability involves $\psi_2^{(t)}$ and because $\psi_2^{(t+1)}$ is set to $\psi_2^{(t)}$ if the proposal is rejected. Next we introduce two examples that illustrate the advantages and potential pitfalls that may arise when using PCG samplers when MH is required for some of their updates.

2.2 Spectral analysis in X-ray astronomy

We begin with an example from our applied work in X-ray astronomy that involves a spectral analysis model that can be fitted with the Data Augmentation algorithm and Gibbs-type samplers (van Dyk *et al.*, 2001; van Dyk and Meng, 2010). We use variants of this example as a running illustration of the methods we propose. The X-ray detectors used in astronomy are typically onboard space-based observatories and record the number of photons detected in each of a large number of energy bins. Spectral analysis aims to estimate the distribution of the photon energies. We use Poisson models for the recorded photon counts, where the expected count is parameterized as a function of the energy, E_i of bin i. A simple example is

$$X_i \stackrel{\text{ind}}{\sim} \text{Poisson}\left\{\Lambda_i = \alpha(E_i^{-\beta} + \gamma I\{i = \mu\})e^{-\phi/E_i}\right\}, \text{ for } i = 1, \ldots, n, \tag{1}$$

where X_i is the count in bin i; $\alpha, \beta, \gamma, \mu$ and ϕ are model parameters; $I\{\cdot\}$ is the indicator function; and n is the number of energy bins. The $\alpha E_i^{-\beta}$ term in (1) is a *continuum*—a smooth term that

extends over a wide range of energies. The $\alpha\gamma I\{i = \mu\}$ term is an *emission line*—a sharp narrow term that describes a distinct aberration from the continuum. The emission line in (1) is very narrow in that it is contained entirely in one energy bin. The parameters of the continuum and emission line describe the composition, temperature, and general physical environment of the source. The factor $e^{-\phi/E_i}$ in (1) accounts for absorption—lower energy photons are more likely to be absorbed by inter-stellar material and not be recorded by the detector. A typical spectral model might contain multiple summed continua and emission lines. We use a simple example here to focus attention on computational issues. Since α, β, γ and ϕ are often blocked in the samplers we discuss, we refer to them jointly as $\theta = (\alpha, \beta, \gamma, \phi)$. We assume that θ and μ are *a priori* independent and that μ is *a priori* uniform on $\{1, \ldots, n\}$.

In practice, we do not observe $X = (X_1, \ldots, X_n)$ directly because photon counts are subject to stochastic censoring, misclassification, and background contamination. First, because the sensitivity of the detector varies with energy, the probability that a photon is detected depends on its energy. Combining this with background contamination,

$$\tilde{X}_i \mid X_i \overset{\text{ind}}{\sim} \text{Binomial}\{X_i, A_i\} + \text{Poisson}(\xi_i), \quad \text{for} \ \ i = 1, \ldots, n, \tag{2}$$

where $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ are the photon counts, including background, that are not absorbed, $A = (A_1, \ldots, A_n)$ is the *effective area* of the detector which describes its sensitivity, and $\xi = (\xi_1, \ldots, \xi_n)$ is the expected background count. Second, misclassification occurs because a photon with energy E_i has probability P_{ij} of being recorded in bin j. Combining these effects, the conditional distribution of the observed photon counts $Y = (Y_1, \ldots, Y_n)$ given $\tilde{X}$ is

$$Y \mid \tilde{X} \overset{\text{ind}}{\sim} \sum_{i=1}^{n} \text{Multinomial}\Big\{\tilde{X}_i, \ (P_{i1}, \ldots, P_{in})\Big\}, \tag{3}$$

and marginally,

$$Y_j \overset{\text{ind}}{\sim} \text{Poisson}\Big\{ \sum_{i=1}^{n} P_{ij}(A_i\Lambda_i + \xi_i)\Big\}, \ \text{for} \ j = 1, \ldots, n, \tag{4}$$

where Λ_i is given by (1). While A and $P = \{P_{ij}\}$ are typically assumed known from instrumental calibration (see Lee *et al.*, 2011, and Xu *et al.*, 2014, for an exception), ξ is often specified in terms of a number of unknown parameters.

The model in (1) is a finite mixture model and can be fitted via the standard data augmentation scheme that sets $X_i = X_{iC} + X_{iL}$, where $X_{iC} \stackrel{\text{ind}}{\sim} \text{Poisson}\left(\alpha E_i^{-\beta} e^{-\phi/E_i}\right)$ and $X_{iL} \stackrel{\text{ind}}{\sim} \text{Poisson}\left(\alpha \gamma I\{i = \mu\} e^{-\phi/E_i}\right)$, are the photon counts in bin i generated from the continuum and emission line, respectively. We consider samplers that target $p(X, X_L, \theta, \mu | Y)$ rather than $p(\theta, \mu | Y)$ both because the ideal data, X, is of scientific interest and because its introduction simplifies the complete conditional distributions, especially in more complex models with multiple summed continua and spectral lines. Assuming ξ is known, this leads to a Gibbs sampler for (1)–(4):

Step 1: $(X^{(t+1)}, X_L^{(t+1)}) \sim p(X, X_L | Y, \theta^{(t)}, \mu^{(t)})$, $\qquad\qquad\qquad\qquad\qquad$ (Sampler 1)

Step 2: $\theta^{(t+1)} \sim p(\theta | Y, X^{(t+1)}, X_L^{(t+1)}, \mu^{(t)})$,

Step 3: $\mu^{(t+1)} \sim p(\mu | Y, X^{(t+1)}, X_L^{(t+1)}, \theta^{(t+1)})$,

where $X_L = (X_{1L}, \ldots, X_{nL})$. We separate μ and θ into two steps to facilitate derivation of the partially collapsed versions of this sampler. Because X_L completely specifies the line location, μ, $\text{Var}_\pi(\mu | X_L) = 0$, Sampler 1 is not irreducible, and $\mu^{(t)} = \mu^{(0)}$ for all t, for any choice of $\mu^{(0)}$. This problem can be solved by updating μ without conditioning on X_L. In particular, we can replace Step 3 of Sampler 1 with $(X_L^{(t+1)}, \mu^{(t+1)}) \sim p(X_L, \mu | Y, X^{(t+1)}, \theta^{(t+1)})$ and permute the steps to

Step 1: $(X_L^\star, \mu^{(t+1)}) \sim p(X_L, \mu | Y, X^{(t)}, \theta^{(t)})$, $\qquad\qquad\qquad\qquad$ (Sampler 2)

Step 2: $(X^{(t+1)}, X_L^{(t+1)}) \sim p(X, X_L | Y, \theta^{(t)}, \mu^{(t+1)})$,

Step 3: $\theta^{(t+1)} \sim p(\theta | Y, X^{(t+1)}, X_L^{(t+1)}, \mu^{(t+1)})$.

The sampled X_L in Step 1 is denoted by $X_L^\star$ because it is not an output of the Markov transition kernel; X_L is updated again in Step 2. In fact $X_L^\star$ is a redundant quantity in that it is not used at all subsequent to Step 1 and replacing Step 1 with $\mu^{(t+1)} \sim p(\mu | Y, X^{(t)}, \theta^{(t)})$ does not alter the Markov transition kernel of Sampler 2. The resulting sampler, that is,

Step 1: $\mu^{(t+1)} \sim p(\mu | Y, X^{(t)}, \theta^{(t)})$, $\qquad\qquad\qquad\qquad\qquad\qquad$ (Sampler 3)

Step 2: $(X^{(t+1)}, X_L^{(t+1)}) \sim p(X, X_L | Y, \theta^{(t)}, \mu^{(t+1)})$,

Step 3: $\theta^{(t+1)} \sim p(\theta | Y, X^{(t+1)}, X_L^{(t+1)}, \mu^{(t+1)})$,

is an example of a PCG sampler composed of incompatible conditional distributions. A variant of this sampler was discussed in Park and van Dyk (2009).

By its construction, the stationary distribution of Sampler 3 is $p(X, X_L, \theta, \mu | Y)$, see Section 2.4. Unlike an ordinary Gibbs sampler, however, permuting its steps may alter its stationary distribution. Suppose, for example, we obtain $(X^{(t)}, X_L^{(t)}, \theta^{(t)}, \mu^{(t)})$ from $p(X, X_L, \theta, \mu | Y)$ and update μ according to Step 1 of Sampler 3. The joint distribution of $(X^{(t)}, X_L^{(t)}, \theta^{(t)}, \mu^{(t+1)})$ would be

$$\int p(\mu^{(t+1)} | Y, X^{(t)}, \theta^{(t)}) p(X^{(t)}, X_L^{(t)}, \theta^{(t)}, \mu^{(t)} | Y) d\mu^{(t)} = p(X^{(t)}, \theta^{(t)}, \mu^{(t+1)} | Y) p(X_L^{(t)} | Y, X^{(t)}, \theta^{(t)}). \tag{5}$$

It is the conditional independence of $X_L^{(t)}$ and $\mu^{(t+1)}$ in (5) that makes Sampler 3 so much faster than Sampler 1; recall $\text{Var}_\pi(\mu | X_L) = 0$. Because the joint distribution of $\theta^{(t)}$ and $\mu^{(t+1)}$ in (5) is their posterior distribution and Step 2 conditions only on $\theta^{(t)}$ and $\mu^{(t+1)}$, the joint distribution of the unknowns after Step 2, that is, of $(X^{(t+1)}, X_L^{(t+1)}, \theta^{(t)}, \mu^{(t+1)})$, is again the target posterior distribution. Thus a cyclic permutation of the steps in Sampler 3 that ends either with Step 2 or Step 3 results in a proper sampler, but ending with Step 1 does not. With non-cyclic permutations, the stationary distribution is unknown.

2.3 A common error in the simplest PCG sampler

The potential pitfalls of introducing MH updates into a PCG sampler can be illustrated using the simplest possible PCG sampler. To see this, we start with a two-step Gibbs sampler with target distribution $p(\psi_1, \psi_2)$, where the second step relies on an MH update:

Step 1: $\psi_1^{(t+1)} \sim p(\psi_1 | \psi_2^{(t)})$, (Sampler 4)

Step 2: $\psi_2^{(t+1)} \sim \mathcal{M}_{2|1}(\psi_2 | \psi_1^{(t+1)}, \psi_2^{(t)})$.

While this sampler is proper, replacing Step 1 with $\psi_1^{(t+1)} \sim p(\psi_1)$ results in an improper sampler:

Step 1: $\psi_1^{(t+1)} \sim p(\psi_1)$, (Sampler 5)

Step 2: $\psi_2^{(t+1)} \sim \mathcal{M}_{2|1}(\psi_2 | \psi_1^{(t+1)}, \psi_2^{(t)})$.

The problem with Sampler 5 can be illustrated using a simulation study. Figure 1 compares 10,000

draws generated by Samplers 4 and 5 with $p(\psi_1, \psi_2)$ given by

$$
\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right]. \tag{6}
$$

The MH jumping rule in Step 2 of both samplers is a Gaussian distribution centered at the previous draw with variance equal to 3. Sampler 5 underestimates the correlation of the target distribution and overestimates the marginal variance of ψ_2. Of course, if we repeat Step 2 a sufficient number of times within each iteration of Sampler 5, it would deliver a draw (nearly) from its target, $p(\psi_2|\psi_1)$, and Sampler 5 would deliver (nearly) independent draws from $p(\psi_1, \psi_2)$. We discuss this strategy for constructing an approximately proper sampler in Section 3.2. Similarly, iterating Step 2 of Sampler 4 would (nearly) lead to a standard two-step Gibbs sampler.

The key to understanding the failure of Sampler 5 (without iterating Step 2) lies in the MH jumping rule used in Step 2 of both samplers. The kernel $\mathcal{M}_{2|1}$ depends on $\psi_2^{(t)}$ through its acceptance probability and its output if its proposal is rejected, thus $\mathcal{M}_{2|1}$ must be written as $\mathcal{M}_{2|1}(\psi_2|\psi_1^{(t+1)}, \psi_2^{(t)})$. Although $\mathcal{M}_{2|1}$ delivers a draw from $p(\psi_2|\psi_1^{(t+1)})$ if given a sample $(\psi_1^{(t+1)}, \psi_2^{(t)})$ from the target distribution, in Sampler 5, $\psi_1^{(t+1)}$ and $\psi_2^{(t)}$ are independent and $\mathcal{M}_{2|1}$ does not deliver a draw from $p(\psi_2|\psi_1^{(t+1)})$.

Unfortunately, there are several examples of samplers in the literature that have the same structure as the improper Sampler 5, for instance, Liu *et al.* (2009), Lunn *et al.* (2009), McCandless *et al.* (2010), and even in the popular WinBUGS package (Spiegelhalter, Thomas, Best and Lunn 2003), see Section 5.1. These samplers do not generally exhibit the desired stationary distributions.

2.4 Convergence of the Partially Collapsed Gibbs sampler

A three-phase framework for deriving proper PCG samplers is given in van Dyk and Park (2008). Consider the Gibbs sampler in Figure 2(a) that updates the components of $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)$ in three steps. In the first phase of the framework, one or more steps of the parent Gibbs sampler are replaced by steps that update, rather than condition on, some components of ψ. This is illustrated in Figure 2(b), where the update $\psi_1 \sim p(\psi_1|\psi_2', \psi_3', \psi_4')$ in Step 1 is replaced with $(\psi_1, \psi_3^\star) \sim p(\psi_1, \psi_3|\psi_2', \psi_4')$. Notice that in the modified step, ψ_3 is sampled rather than conditioned upon. This

conditioning reduction phase is key to the improved convergence properties of the PCG sampler. By conditioning on less, we expect to increase the variance of the updating distribution, at least on average. This is evident in Section 2.2 where the complete conditional for μ in Sampler 1 has zero variance, but its update with reduced conditioning in Sampler 2 readily allows μ to move across its parameter space. More formally, van Dyk and Park (2008) showed that sampling more unknowns in any set of steps of a Gibbs sampler can only reduce the so-called cyclic-permutation bound on the spectral radius of the sampler. The resulting substantial improvement in the rate of convergence is illustrated in the examples given in Bernardi *et al.* (2013), Berrett and Calder (2012), Caron *et al.* (2014), Dobigeon and Tourneret (2010), Hu *et al.* (2012), Hu *et al.* (2013), Kail *et al.* (2010, 2011), Lin and Tourneret (2010), Lindsten *et al.* (2013), Park *et al.* (2008), Park and van Dyk (2009), Park *et al.* (2012a), Park *et al.* (2012b), and Zhao and Lian (2014), among others. (*Conditioning reduction* was called *marginalization* by van Dyk and Park (2008).)

The conditioning reduction phase results in one or more components of ψ being updated in multiple steps; ψ_3 is updated in Steps 1 and 3 in Figure 2(b). If the same component is updated in two consecutive steps (without MH), the Markov transition kernel does not depend on the first update. We call quantities that are updated in a sampler, but do not affect its transition kernel *redundant quantities*—they must be updated subsequently or they would be part of the output of the iteration. The second phase of the framework is to *permute* the steps of the sampler with reduced conditioning to make as many of the updates redundant as possible. For example, we permuted the steps in Figure 2(b) so that ψ_3 is updated in Steps 2 and 3 of Figure 2(c) and $\psi_3^\star$ is redundant.

In the third phase, redundant quantities are removed or *trimmed* from the updating scheme. For example, Step 2 in Figure 2(d) does not update ψ_3. By construction, this does not affect the overall transition kernel. The resulting step samples from a conditional distribution of a marginal distribution of $p(\psi)$. For example, Step 2 in Figure 2(d) simulates from a conditional distribution of $p(\psi_1, \psi_2, \psi_4)$ rather than of $p(\psi_1, \psi_2, \psi_3, \psi_4)$. We refer to steps that sample or target such distributions as *reduced steps* and to steps that sample or target a complete conditional as *full steps*.

In some cases, the result of the three-phase framework is simply a blocked or collapsed (Liu *et al.*, 1994) version of the parent Gibbs sampler. In other cases, however, the resulting PCG sam-

pler is composed of draws from a set of incompatible conditional distributions (e.g., Sampler 3). Since all three phases preserve the stationary distribution of the parent sampler, we know that the resulting PCG sampler is proper. Because reducing the conditioning can significantly improve the rate of convergence of the sampler, while permutation typically has a minor effect, and trimming has no effect on the rate of convergence, we generally expect the PCG sampler to exhibit better and often much better convergence properties than its parent Gibbs sampler.

3   Using MH Algorithm within the PCG Sampler

3.1   Identifying the stationary distributions

We now consider the use of MH updates for some of the steps of a PCG sampler. As the example in Section 2.3 illustrates, introducing MH into a well behaved PCG sampler can destroy the sampler's stationary distribution. Thus, care must be taken to guarantee that an MH within PCG sampler is proper. Here we describe the basic complication that arises when MH is introduced into a PCG sampler and give advice as to how to ensure that the sampler is proper.

When deriving a PCG sampler (without MH), the conditioning reduction phase means some components of ψ are updated in multiple steps. If the same component is updated in consecutive steps, the Markov transition kernel does not depend on the first update. The first update is therefore redundant and can be omitted without affecting the stationary distribution of the chain.

This situation is more complicated when some of the steps of the PCG sampler require MH updates. Suppose, for example, we wish to sample from $p(\psi)$ with $\psi = (\psi_1, \psi_2, \psi_3)$ using a proper PCG sampler in which ψ_1 and ψ_2 are jointly updated in Step K via a draw from the conditional distribution $p(\psi_1, \psi_2 | \psi_3)$. Suppose also that ψ_2 is to be updated according to its full conditional distribution, $p(\psi_2 | \psi_1, \psi_3)$ in Step $K + 1$, but this cannot be done directly and we wish to use an MH update. The remaining unknowns, ψ_3, are updated in other steps of the sampler, which perhaps involve dividing ψ_3 into multiple subcomponents. That is, Steps K and $K + 1$ of the sampler are

Step K: $(\psi_1^{(t+1)}, \psi_2^{\star}) \sim p(\psi_1, \psi_2 | \psi_3'),$ (Sampler Fragment 1)

Step $K + 1$: $\psi_2^{(t+1)} \sim \mathcal{M}_{2|1,3}(\psi_2 | \psi_1^{(t+1)}, \psi_2^{\star}, \psi_3').$

If we were able to draw ψ_2 directly from its complete conditional distribution in Step $K + 1$, $\psi_2^\star$ would be redundant and we could remove it from the sampler by replacing the update in Step K with the reduced step $\psi_1^{(t+1)} \sim p(\psi_1|\psi_3')$. The MH update in Step $K + 1$, however, depends on $\psi_2^\star$ and replacing it with $\psi_2^{(t)}$ may change the chain's stationary distribution in an unpredictable way. In short, the MH update used in Step $K + 1$ means that we cannot reduce Step K. Generally speaking, an MH update in a step that follows a reduced step is problematic because reduced steps result in independences that do not exist in the target. (A reduced step that follows an MH step, however, is not inherently problematic.) More precisely, the kernel, $\mathcal{M}_{j_1|j_2}(\psi_{j_1}|\psi_{j_1}',\psi_{j_2}',\psi_{j_3}')$, can generally only be used if no component of $(\psi_{j_1},\psi_{j_2},\psi_{j_3})$ is trimmed in the previous step.

Luckily, the stationary distribution of an MH within PCG sampler can be verified using the same methods that are used for an ordinary PCG sampler. In particular, the three-phase framework of van Dyk and Park (2008) can be directly applied. The first two phases, conditioning reduction and permutation, apply equally well to MH within Gibbs samplers. Neither updating additional components of ψ in one or more steps nor permuting the order of the steps upsets the stationary distribution of an MH within Gibbs sampler. The final phase involves removing redundant updates. Because an MH step generally depends on the current draws of *all* of the components of ψ not reduced out in that step, there are fewer redundant draws when some steps involve MH. Nonetheless, any redundant updates that are identified can safely be removed in the trimming phase—by definition they do not affect the transition kernel. *The critical point is that unlike with an ordinary Gibbs sampler, we cannot simply replace some of the component draws of a PCG sampler with MH updates. Rather we must construct an MH within PCG sampler by applying the three-phase framework.*

Now suppose we wish to reduce the conditioning in an MH step. In Sampler Fragment 1, for example, if $p(\psi_3|\psi_1,\psi_2)$ is a standard distribution with known normalization, then we can evaluate $p(\psi_2|\psi_1) \propto p(\psi_1,\psi_2) = p(\psi_1,\psi_2,\psi_3)/p(\psi_3|\psi_1,\psi_2)$ and sample $\psi_2 \sim \mathcal{M}_{2|1}(\psi_2|\psi_1',\psi_2')$. Replacing Step $K + 1$ of Sampler Fragment 1 with this reduced MH step, however, can alter the chain's stationary distribution in unpredictable ways. Instead, we propose to replace the full MH step with the reduced MH step *followed immediately* by a direct draw from the complete conditional of the reduced quantities. In Sampler Fragment 1 this would entail replacing Step $K + 1$ with

Step $K + 1$ with Reduced Conditioning: $\psi_2^{(t+1)} \sim \mathcal{M}_{2|1}(\psi_2|\psi_1^{(t+1)}, \psi_2^\star)$ and $\psi_3 \sim p(\psi_3|\psi_1^{(t+1)}, \psi_2^{(t+1)})$.

This strategy ensures that the target stationary distribution is maintained. The expectation is that the updates of the reduced quantities will be trimmed after the steps are appropriately permuted and that the reduced MH step can be employed in the final sampler. We denote the transition kernel of the full step (i.e., the reduced MH step followed by the complete conditional of the reduced quantities) by $\mathcal{M}^\star$. In Sampler Fragment 1, we rewrite the step with reduced conditioning

Step $K + 1$ with Reduced Conditioning: $(\psi_2^{(t+1)}, \psi_3) \sim \mathcal{M}_{2,3|1}^\star(\psi_2, \psi_3|\psi_1^{(t+1)}, \psi_2^\star)$.

Notice that this full update is not formally an MH update and has the advantage that it does not depend on all of the components of ψ. Thus, this step can follow a step that reduces ψ_3 out.

We now illustrate the construction of a proper MH within PCG sampler for the spectral model given in (1). For simplicity, we assume that X is observed directly and we can ignore (2)–(4). Figure 3(a) gives a six-step Gibbs sampler. Three of its steps require MH updates; the details of all the steps are given in Appendix B. The conditioning in four of the steps is reduced in Figure 3(b), and the steps are permuted in Figure 3(c) to allow the redundant draws of $X_L^\star$ and $\alpha^\star$ to be trimmed in four steps. Sampler 6, the resulting proper MH within PCG sampler, appears in Figure 4.

3.2 Using MH following a reduced step

Using a full MH step immediately following a reduced step can be problematic. Sampler 5 illustrates this in its simplest form: a draw from a marginal distribution followed by an MH update of the conditional distribution of the remaining unknowns. As noted in Section 2.3 this is a particularly common problem in practice, even in its simplest form. In more complicated PCG samplers, the general phenomenon of introducing a full MH step immediately following a reduced step is the typical path by which introducing MH leads to an improper sampler. This is illustrated in Sampler Fragment 1, where we are unable to replace the update in Step K with the reduced step $\psi_1^{(t+1)} \sim p(\psi_1|\psi_3')$. Thus, this case is particularly important and we propose two alternate samplers that maintain the basic structure of the underlying PCG sampler while allowing a form of MH in the step following a reduced step. Both solutions are conceptually straightforward.

We begin by studying a special case that is useful for illustrating the two alternative samplers that we propose. We discuss the more general situation below. In particular we start in the general setting of Sampler Fragment 1, but consider a PCG sampler in which ψ_1 is updated in Step K via a direct draw from the conditional distribution $p(\psi_1|\psi_3)$ of the marginal distribution $p(\psi_1, \psi_3)$, i.e., a reduced step. Again suppose that an MH update is required to update ψ_2 in Step $K + 1$. That is, Steps K and $K + 1$ of the parent PCG sampler are

Step K: $\psi_1^{(t+1)} \sim p(\psi_1|\psi_3')$,   (Sampler Fragment 2)

Step $K + 1$: $\psi_2^{(t+1)} \sim p(\psi_2|\psi_1^{(t+1)}, \psi_3')$.

Because MH is needed for Step $K + 1$, these steps cannot be blocked.

One straightforward general solution to the intractability of $p(\psi_2|\psi_1^{(t+1)}, \psi_3')$ is simply to iterate the MH update within Step $K + 1$ to obtain a draw from the conditional distribution,

Iterated MH Strategy:

Step K: $\psi_1^{(t+1)} \sim p(\psi_1|\psi_3')$,   (Sampler Fragment 3)

Step $K + 1$: Sample $\psi_2^{(t+l/L)} \sim \mathcal{M}_{2|1,3}(\psi_2|\psi_1^{(t+1)}, \psi_2^{(t+(l-1)/L)}, \psi_3')$, for $l = 1, \ldots, L$, to obtain $\psi_2^{(t+1)} \overset{\text{approx}}{\sim}$
    $p(\psi_2|\psi_1^{(t+1)}, \psi_3')$ at the subiteration $l = L$.

We discuss methods for determining how large L must be in Sections 4.1 and 5.1. With sufficiently large L, the iterated MH strategy delivers a draw that approximately follows $p(\psi_2|\psi_1^{(t+1)}, \psi_3')$ and thus the sampler is *approximately proper*. In this special case the iterated MH strategy effectively blocks Steps K and $K + 1$ to (nearly) deliver an independent draw from $p(\psi_1, \psi_2|\psi_3')$.

Another solution to the intractability of $p(\psi_2|\psi_1^{(t+1)}, \psi_3')$ is a joint MH update on the blocked version of Steps K and $K + 1$,

Joint MH Strategy:

Step K: Update (ψ_1, ψ_2) jointly via an MH update with jumping rule $\mathcal{J}_{1,2|3}(\psi_1, \psi_2|\psi_2^{(t)}, \psi_3') = p(\psi_1|\psi_3') \mathcal{J}_{2|1,3}(\psi_2|\psi_1,$

Step $K + 1$: Omit.   (Sampler Fragment 4)

The jumping rule in Step K of Sampler Fragment 4 is exactly the concatenation of Step K and the jumping rule in Step $K + 1$ of Sampler Fragment 3. By concatenating we avoid iteration. In the

joint MH strategy, however, the proposed ψ_1 and ψ_2 are accepted or rejected together, whereas in the iterated MH strategy only the proposed ψ_2 is accepted or rejected.

The iterated MH strategy is in some sense a thinned version of the joint MH strategy. This, however, is an over simplification for two reasons. First, the iterated MH strategy updates ψ_1 only once for every L updates of ψ_2 whereas the joint MH strategy updates both together. Second, although the jumping rule in the joint MH strategy is the same as that used by the iterated MH strategy at its first subiteration, the acceptance probabilities differ. This results in a systematic difference in the performance of the resulting samplers, see Section 4.1.

Generalizing Sampler Fragment 2, Steps K and $K+1$ may not block even without MH. Suppose $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)$ and the parent PCG sampler contains the two steps

Step K: $\psi_1^{(t+1)} \sim p(\psi_1 | \psi_3^{(t)}, \psi_4')$,   (Sampler Fragment 5)

Step $K+1$: $(\psi_2^{(t+1)}, \psi_3^{(t+1)}) \sim p(\psi_2, \psi_3 | \psi_1^{(t+1)}, \psi_4')$,

where Step K is a reduced step and Step $K+1$ cannot be sampled directly. Here the conditional distributions cannot be blocked into a single step. We can still use the iterated MH strategy in Step $K+1$ to obtain a draw approximately from $p(\psi_2, \psi_3 | \psi_1^{(t+1)}, \psi_4')$ and an approximately proper sampler. Likewise we can implement the joint MH strategy, using the jumping rule $p(\psi_1 | \psi_3^{(t)}, \psi_4') \mathcal{J}_{2,3|1,4}(\psi_2, \psi_3 | \psi_1, \psi_2^{(t)}, \psi_3^{(t)}, \psi_4')$. The stationary distribution of the joint jumping rule is $p(\psi_1 | \psi_3^{(t)}, \psi_4') p(\psi_2, \psi_3 | \psi_1, \psi_4')$. Although a legitimate joint distribution on (ψ_1, ψ_2, ψ_3), this does not correspond to a conditional distribution of $p(\psi)$.

3.3   To block or not to block

Section 3.2 discusses the case where Step $K+1$ of Sampler Fragment 2 requires MH. We now consider the case where Step K requires MH. In particular,

Step K: $\psi_1^{(t+1)} \sim \mathcal{M}_{1|3}(\psi_1 | \psi_1^{(t)}, \psi_3')$,   (Sampler Fragment 6)

Step $K+1$: $\psi_2^{(t+1)} \sim p(\psi_2 | \psi_1^{(t+1)}, \psi_3')$.

Sampler Fragment 6 does not lead to convergence problems because the inputs to Step $K+1$ follow the correct distribution; Figure A in the online appendix verifies the stationary distribution of its parent chain.

We might consider blocking the two steps in Sampler Fragment 6 into a single MH update as

Step K: Update (ψ_1, ψ_2) jointly via the MH jumping rule $\mathcal{J}_{1,2|3}(\psi_1, \psi_2 | \psi_1^{(t)}, \psi_3') = \mathcal{J}_{1|3}(\psi_1 | \psi_1^{(t)}, \psi_3') p(\psi_2 | \psi_1, \psi_3')$,

Step $K + 1$: Omit.   (Sampler Fragment 7)

The jumping rule in Sampler Fragment 7 is exactly the concatenation of the jumping rules in the two steps of Sampler Fragment 6. There is a fundamental difference, however, in that the transition kernel corresponding to the concatenated jumping rule depends on $\psi_2^{(t)}$: if the MH proposal is rejected, $(\psi_1^{(t+1)}, \psi_2^{(t+1)}) = (\psi_1^{(t)}, \psi_2^{(t)})$, whereas neither of the steps in Sampler Fragment 6 depends on $\psi_2^{(t)}$. This means that care must be taken to ensure blocking in this way does not upset the stationary distribution of the chain.

Steps 3 and 4 of Sampler 6 are an example of Sampler Fragment 6, with $\psi_1 = \beta$, $\psi_2 = \alpha$ and $\psi_3 = (\gamma, \mu, \phi)$. Blocking Steps 3 and 4 of Sampler 6 results in Sampler 7, see the second panel of Figure 4. Unfortunately, this is an improper sampler, which we verify using a simulation study. We begin by generating an artificial data set consisting of $n = 550$ bins with $\alpha = 37.62$, $\beta = 1$, $\gamma = 40/37.62$, $\mu = 250$, and $\phi = 0.2$, see Figure 5. We run two versions of Sampler 7. Sampler 7(a) uses the concatenated jumping rule given in Sampler Fragment 7 to update (α, β), while Sampler 7(b) uses an independent bivariate normal jumping rule centered at the current value of (α, β). We use a uniform prior distribution for each parameter, and run 30,000 iterations of Samplers 6, 7(a), and 7(b) using the same starting values ($\alpha = 30$, $\beta = 3$, $\gamma = 1$, $\mu = 10$ and $\phi = 0.5$). Scatter plots of (α, β, ϕ) for the last 10,000 draws from the three samplers appear in Figure 6, which shows that Samplers 7(a) and 7(b) underestimate the correlations of the target distribution; this effect is especially dramatic for Sampler 7(b). Figure 7 compares the marginal distributions of α, β, and ϕ generated with Samplers 6 and 7(b), and shows that Sampler 7(b) underestimates the marginal variances of all three parameters. (The marginals generated with Sampler 7(a) are more similar to those generated with Sampler 6.)

The problem with Sampler 7 can be understood in the terms of Section 3.2. Blocking the updates for α and β results in an MH step that follows directly after a pair of reduced steps (the updates of μ and ϕ). If μ and ϕ were known, and Steps 1 and 2 were removed, both versions of Sampler 7 would be proper. As it is, the stationary distribution of Sampler 7 cannot be verified

with the three-phase framework.

The comparison between Sampler Fragments 6 and 7 is similar to that between the iterated and joint MH strategies in Section 3.2. Theoretical perspectives on these choices appear in Sec. 4.

4 Theory

4.1 Comparing the iterated and joint MH strategies

In this section we compare the iterated and joint MH strategies in terms of their acceptance probabilities. Although it is generally recognized that an acceptance probability of 20% to 40% is best for a symmetric Metropolis jumping rule (Roberts *et al.*, 1997), we argue that the better choice between the two strategies is determined by maximizing the acceptance probability. This is because both the iterated and joint MH strategies start with the *same proposal*—they are numerically identical. The rule of thumb for tuning the acceptance probability to between 20% and 40% is based on comparing *different proposal distributions* with an eye on avoiding high acceptance rates because they typically correspond to jumping rules that propose very small steps. In this case the initial step sizes are the same and we aim to reduce correlation by increasing the jumping probability. We begin with theoretical results and then illustrate them numerically.

To simplify notation we suppress the conditioning on ψ_3 in Sampler Fragments 3 and 4. This is equivalent to a formal comparison of the iterated and joint MH strategies as alternatives to the improper two-step Sampler 5. We assume that (i) the sampler has been verified to be proper so that $\pi = p$ and (ii) the jumping rule used to update ψ_2 does not depend on ψ_1, i.e., $\mathcal{J}_{2|1}(\psi_2|\psi_1', \psi_2') = \mathcal{J}_{2|1}(\psi_2|\psi_2')$. While the transition kernel $\mathcal{M}_{2|1}(\psi_2|\psi_1', \psi_2')$ will typically depend on ψ_1', the jumping rule often will not, for example, a symmetric Metropolis-type jumping rule does not.

The acceptance probability of the first draw in Step $K + 1$ of the iterated MH strategy is

$$r_{\text{iter}} = \frac{p(\psi_2^{\text{prop}}|\psi_1^{(t+1/L)})\mathcal{J}_{2|1}(\psi_2^{(t)}|\psi_2^{\text{prop}})}{p(\psi_2^{(t)}|\psi_1^{(t+1/L)})\mathcal{J}_{2|1}(\psi_2^{\text{prop}}|\psi_2^{(t)})}, \tag{7}$$

where $\psi_1^{(t+1/L)} \sim p(\psi_1)$ and $\psi_2^{\text{prop}} \sim \mathcal{J}_{2|1}(\psi_2|\psi_2^{(t)})$. With the joint MH strategy, it is

$$r_{\text{joint}} = \frac{p(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})\{p(\psi_1^{(t)})\mathcal{J}_{2|1}(\psi_2^{(t)}|\psi_2^{\text{prop}})\}}{p(\psi_1^{(t)}, \psi_2^{(t)})\{p(\psi_1^{\text{prop}})\mathcal{J}_{2|1}(\psi_2^{\text{prop}}|\psi_2^{(t)})\}} = \frac{p(\psi_2^{\text{prop}}|\psi_1^{\text{prop}})\mathcal{J}_{2|1}(\psi_2^{(t)}|\psi_2^{\text{prop}})}{p(\psi_2^{(t)}|\psi_1^{(t)})\mathcal{J}_{2|1}(\psi_2^{\text{prop}}|\psi_2^{(t)})}, \tag{8}$$

where $\psi_1^{\text{prop}} \sim p(\psi_1)$ and $\psi_2^{\text{prop}} \sim \mathcal{J}_{2|1}(\psi_2|\psi_2^{(t)})$.

Lemma 4.1 *In the setting described in the previous paragraph,*

$$\mathrm{E}_\pi[r_{\text{iter}}/r_{\text{joint}}] \geq 1. \tag{9}$$

The expectation in (9) is under the common stationary distribution, π, of both chains and is conditional on the random seed used at the start of each iteration. That is, since $(\psi_1^{(t+1/L)}, \psi_2^{\text{prop}})$ sampled under the iterated MH strategy and $(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})$ sampled under the joint MH strategy are drawn in exactly the same way, we assume these quantities are numerically equal. Expression (9) asserts that while both strategies start with the same proposal—$(\psi_1^{(t+1/L)}, \psi_2^{\text{prop}})$ under the iterated MH strategy and $(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})$ under the joint—the iterated MH strategy is on average more likely to accept ψ_2. (The iterated MH strategy *always* accepts ψ_1.)

Proof: With the numerical equality of the proposals,

$$\frac{r_{\text{iter}}}{r_{\text{joint}}} = \frac{p(\psi_2^{(t)}|\psi_1^{(t)})}{p(\psi_2^{(t)}|\psi_1^{(t+1/L)})}, \tag{10}$$

where $(\psi_1^{(t)}, \psi_2^{(t)}, \psi_1^{(t+1/L)}) \sim \pi(\psi_1^{(t)}, \psi_2^{(t)})\pi_1(\psi_1^{(t+1/L)})$ with π_1 the ψ_1 marginal distribution of π. Because $(\psi_1^{(t)}, \psi_2^{(t)}) \sim \pi$ and $\pi = p$, the numerator of (10) is the conditional density of ψ_2 evaluated at $\psi_2^{(t)}$. This is not true of the denominator because $\psi_2^{(t)}$ is independent of $\psi_1^{(t+1/L)}$. Thus, we might expect that the numerator of (10) is typically larger than the denominator, as claimed in (9).

Recalling that $\pi = p$, substituting (10) into (9), and applying Jensen's inequality, we need only verify that

$$\int \log\left[\pi(\psi_2|\psi_1)\right] \pi(\psi_1, \psi_2)d\psi_1 d\psi_2 \geq \int \log\left[\pi(\psi_2|\psi_1)\right] \pi(\psi_1)\pi(\psi_2)d\psi_1 d\psi_2. \tag{11}$$

Expression (11) can be verified using a standard property of entropy along with the Kullback-Leiber (KL) divergence. In particular, because KL is nonnegative,

$$\int \log\left[\pi(\psi_2)\right] \pi(\psi_1)\pi(\psi_2)d\psi_1 d\psi_2 \geq \int \log\left[\pi(\psi_2|\psi_1)\right] \pi(\psi_1)\pi(\psi_2)d\psi_1 d\psi_2. \tag{12}$$

(The standard KL expression can be recovered by adding $\int \log [\pi(\psi_2)] \pi(\psi_1)\pi(\psi_2)d\psi_1 d\psi_2$ to both sides of (12).) But a standard property of entropy (e.g., Ebrahimi *et al.*, 1999) is

$$\int \log [\pi(\psi_2|\psi_1)] \pi(\psi_1, \psi_2)d\psi_1 d\psi_2 \geq \int \log [\pi(\psi_2)] \pi(\psi_1)\pi(\psi_2)d\psi_1 d\psi_2. \tag{13}$$

Combining (12) and (13) gives (11) and hence the desired result. ∎

We now return to the bivariate Gaussian simulation of Section 2.3 to compare the computational performance of the iterated and joint MH strategies. Again we sample ψ_1 from its marginal distribution and use the same MH jumping rule to update ψ_2 according to its conditional distribution. The iterated strategy is run with $L = 7$, in order to return $\psi_2^{(t+1)}$ that is essentially independent of $\psi_2^{(t)}$. The value of L was set using an initial MH run of $5,000$ iterations and inspecting the autocorrelation function. The initial MH sampler delivers essentially independent draws after 7 iterations, see Figure 8(a). Of course, the computational cost per iteration of the iterated MH strategy depends on L. With $L = 7$, each iteration requires eight univariate normal draws, whereas the joint strategy requires two. The autocorrelation functions of ψ_2 for both the iterated and joint MH strategies appear in Figure 8(b)–(c) and show the clear computational advantage of the iterated MH strategy. It returns essentially independent draws, whereas the joint MH strategy requires almost thirty iterations to obtain nearly independent draws.

In practice, it is important to check that the value of L used in Sampler Fragment 3 delivers samples that are essentially independent of the starting value of the iterated MH strategy. Fortunately, a simple diagnostic is available through the autocorrelation function of $\psi_2^{(t)}$ in Sampler Fragment 3, e.g., Figure 8(b). If the lag one autocorrelation is not essentially zero, the run should be repeated with a larger value of L. If ψ_2 is updated elsewhere in the sampler, the efficacy of the iterated MH strategy can be isolated by computing the correlation between the initial input of ψ_2 and the final output after iteration of the MH update in Step $K + 1$ of Sampler Fragment 3. Xu *et al.* (2014) propose testing the hypothesis that $\psi_2^{(t)}$ and $\psi_2^{(t+1)}$ are uncorrelated by computing,

$$\hat{\rho} = \frac{\sum_{t=T_0+1}^{T}(\psi_2^{(t)} - \hat{\psi}_2)(\psi_2^{(t+1)} - \hat{\psi}_2)}{\sum_{t=T_0+1}^{T}(\psi_2^{(t)} - \hat{\psi}_2)^2}, \quad \text{where} \quad \hat{\psi}_2 = \sum_{t=T_0+1}^{T} \psi_2^{(t)},$$

where T_0 is the burn in size. If L is large enough, then for large T, $(T - T_0)\hat{\rho}^2/(1 - \hat{\rho}^2) \overset{\text{approx}}{\sim} N(0, 1)$.

4.2 Comparing the samplers with and without blocking

To compare the blocking strategy in Sampler Fragment 7 with Sampler Fragment 6, we compute its acceptance rate, again suppressing the conditioning on ψ_3 for simplicity, as

$$r_{\text{blocked}} = \frac{p(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})\mathcal{J}_1(\psi_1^{(t)}|\psi_1^{\text{prop}})p(\psi_2^{(t)}|\psi_1^{(t)})}{p(\psi_1^{(t)}, \psi_2^{(t)})\mathcal{J}_1(\psi_1^{\text{prop}}|\psi_1^{(t)})p(\psi_2^{\text{prop}}|\psi_1^{\text{prop}})} = \frac{p(\psi_1^{\text{prop}})\mathcal{J}_1(\psi_1^{(t)}|\psi_1^{\text{prop}})}{p(\psi_1^{(t)})\mathcal{J}_1(\psi_1^{\text{prop}}|\psi_1^{(t)})} = r_{\text{not blocked}}, \tag{14}$$

where $r_{\text{not blocked}}$ is the acceptance probability of Step K in Sampler Fragment 6, where there is no blocking. This means that Sampler Fragments 6 and 7 are identical in terms of their update of ψ_1, but whereas Sampler Fragment 6 updates ψ_2 with a new value at every iteration, blocking causes ψ_2 to only be updated if ψ_1 is updated. Thus, we expect the blocking strategy of Sampler Fragment 7 to reduce the efficiency of the sampler, and contrary to general advice regarding blocking (e.g., Liu *et al.*, 1994), the blocking strategy of Sampler Fragment 7 should be avoided.

Together, the results of Sections 4.1 and 4.2 should be taken to discourage the combining of an MH update and a direct draw from a conditional distribution into a single MH update.

5 Examples

5.1 The simplest MH within PCG sampler

MH within PCG samplers are useful for fitting multi-component models in which part of the model must be fitted off-line. Consider a two-step sampler that updates ψ_1 and ψ_2 each in turn, but for computational reasons, we wish to update ψ_1 off-line. This may, for example, stem from the use of computer models that involve some costly evaluations in the update of ψ_1. As an illustration, we consider the problem of accounting for calibration uncertainty in high-energy astrophysics (Lee *et al.*, 2011) using a special case of model (4) in Section 2.2:

$$Y_j \sim \text{Poisson}\{A_j \alpha E_j^{-\beta}\}, \text{ for } j = 1, \ldots, n. \tag{15}$$

Here we consider the case where the effective area vector $A = (A_1, \ldots, A_n)$ is not known, and must be estimated along with α and β. In-space calibration and sophisticated modelling of the instrument result in a representative sample of possible A values. Lee *et al.* (2011) shows how a Principal

Component Analysis (PCA) of this sample can be used to derive a degenerate multivariate normal prior for A. In particular, we can write $A(Z) = A_0 + QZ$, where A_0 ($n \times 1$) and Q ($n \times q$) are known, the components of the ($q \times 1$) vector, Z, are independent standard normal variables, and $q \ll n$. Since A is a deterministic function of Z, we can confine attention to the parameter (Z, α, β). With the expectation that Y would be relatively noninformative for $A(Z)$ and to simplify computation, Lee *et al.* (2011) suggests adopting $p(Z)p(\alpha, \beta | Z, Y)$ as the target distribution for statistical inference, an approximation that they call *Pragmatic Bayes*. Thus, the target can be sampled by first drawing $Z \sim p(Z)$ and then updating α and β given Z. Using a uniform prior for α and β: $p(\alpha, \beta) \propto 1$, the complete conditional for α is in closed form, but β requires MH.

One might be tempted to implement the following improper MH within PCG sampler:

Step 1: $Z^{(t+1)} \sim p(Z)$,   (Sampler 8)

Step 2: $\beta^{(t+1)} \sim \mathcal{M}_{\beta | Y, \alpha, A(Z)}(\beta | \alpha^{(t)}, \beta^{(t)}, A(Z^{(t+1)}))$,

Step 3: $\alpha^{(t+1)} \sim p(\alpha | Y, \beta^{(t+1)}, A(Z^{(t+1)}))$.

This update of α and β reflects the simple form of (15). Methods for fitting more general spectral models were considered by Lee *et al.* (2011). To derive an (approximately) proper sampler, we can remove the conditioning on α and implement the iterated MH strategy in Step 2:

Step 1: $Z^{(t+1)} \sim p(Z)$,   (Sampler 9)

Step 2: $\beta^{(t+l/L)} \sim \mathcal{M}_{\beta | Y, A(Z)}(\beta | \beta^{(t+(l-1)/L)}, A(Z^{(t+1)}))$, for $l = 1, \ldots, L$,

Step 3: $\alpha^{(t+1)} \sim p(\alpha | Y, \beta^{(t+1)}, A(Z^{(t+1)}))$.

As suggested in Section 4.1, we determine L using an initial MH run of $1,000$ iterations and inspecting its autocorrelation function. We found that the component MH sampler delivers essentially independent draws of β after 20 iterations and thus set $L = 20$ in Step 2 of Sampler 9.

We use a simulation study to illustrate the impropriety of Sampler 8. The data are simulated using $n = 1,078$ energy bins ranging from 0.225 to 10.995 keV, $q = 7$, $Z_j = 1.5$ ($j = 1, \ldots, q$), $\alpha = 30$ and $\beta = 1$. For each sampler, a chain of length 20,000 is run with a burnin of 10,000 from the starting values $Z_j = 0$ ($j = 1, \ldots, q$), $\alpha = 1$ and $\beta = 1$. Figure 9 shows that using $L = 20$ in

Sampler 9 is sufficiently large and that Sampler 8 both underestimates the correlation of Z_2 and β and the marginal variability of both α and (more dramatically) β.

While Lee *et al.* (2011) recognized the hazard of Sampler 8 and proposed Sampler 9, there are other examples in the literature where MH is used within a PCG sampler incorrectly, resulting in improper samplers. Liu *et al.* (2009), for example, proposed a sampler very similar to Sampler 8 in structure, but in a completely different setting. To predict the temperature of a particular device at a certain time point, the parameters describing the physical properties of the device were linked to the other parameters via a computationally expensive computer model. One of the approaches described in Liu *et al.* (2009) for sampling all the model parameters from their posterior distribution was to update the physical-property parameters from their prior distributions first, and then sample the remaining parameters conditioning on the prior-generated values of the physical-property parameters. This approach was expected to reduce the confoundedness between the parameters and thus improve the mixture of the Markov chain. Since the updates of the other parameters relied on MH, this approach is problematic as illustrated in Section 2.3. In analogy to Figure 9, Liu *et al.* (2009) showed that the marginal distributions of the other parameters sampled via this approach were more variable than via the full Bayesian analysis or some other approaches. Other examples of improper samplers that are similar in structure to Sampler 8 were proposed in Lunn *et al.* (2009), McCandless *et al.* (2010), and even the popular WinBUGS package (Spiegelhalter, Thomas, Best and Lunn 2003), see Woodard *et al.* (2012) for discussion.

5.2 Spectral analysis with narrow lines in high-energy astrophysics

Section 3.3 uses a simulation study to illustrate a potential problem with Sampler Fragment 7, that is, how the blocking of an MH update and a direct draw from a conditional distribution can result in an improper sampler. Here we use the same simulation study to illustrate the improved convergence properties of three proper MH within PCG samplers relative to their parent Gibbs sampler. The only difference is that for each sampler here, a chain of 20,000 iterations is run with a burnin of 10,000 iterations.

As pointed out in Section 2.2, the standard Gibbs sampler for the spectral model (1) breaks down since the resulting subchain for μ does not move from its starting value (Park and van Dyk,

2009). To solve this problem, we sample μ without conditioning on X_L and obtain an MH within PCG sampler, i.e., Sampler 10, given in the first panel of Figure 10. Sampler 6 in Figure 4 is another MH within PCG sampler but with a higher degree of partial collapsing, by which we mean more quantities are reduced out in Sampler 6 than in Sampler 10. Not only does Sampler 6 update μ without conditioning on X_L, but it also reduces α out of its first three steps, whereas Sampler 10 does not remove α from any step. Sampler 11 attempts to further improve Sampler 6 by blocking the MH updates of β and ϕ, see the second panel of Figure 10. Unlike Sampler 7 which also blocks 2 steps of Sampler 6, Sampler 11 is proper, see Figure 11. Thus Samplers 6, 10 and 11 are all proper MH within PCG samplers with common parent Gibbs sampler given in Figure 3(a), but with different degrees of partial collapsing. (The derivation of Sampler 6 appears in Figure 3 and that of Sampler 10 is omitted to save space.)

The convergence characteristics of α, β, and ϕ using Samplers 10 and 11 are compared in Figure 12; γ and μ converge well for all three samplers. All three MH within PCG samplers outperform the parent Gibbs sampler, since the latter does not converge to the target. Sampler 11 performs much better than Sampler 10 in terms of the mixing and autocorrelations of α, β, and ϕ. The performance of Sampler 6 is better than Sampler 10, but not as good as Sampler 11. (To save space, the results of the intermediate Sampler 6 are omitted in Figure 12.) These results show that proper MH within PCG samplers outperform their parent Gibbs sampler in computational efficiency and a higher degree of partial collapsing can improve the convergence even further.

5.3 Relating ECME with Newton-type updates to MH within PCG samplers

The Expectation-Maximization (EM) algorithm is a frequently used technique for computing maximum likelihood or maximizing a posterior estimate. The Expectation/Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993) extends the EM algorithm by replacing the M-step of each EM iteration with a sequence of CM-steps, each of which maximizes the *constrained* expected complete-data loglikelihood function. Liu and Rubin (1994) further generalized ECM with the Expectation/Conditional Maximization Either (ECME) algorithm by replacing some of its CM-steps with steps that maximize the corresponding constrained *actual* likelihood function. ECME can converge substantially faster than either EM or ECM while maintaining the stable monotone

convergence and basic simplicity of its parent algorithms. The Gibbs sampler can be viewed as the stochastic counterpart of ECM, see van Dyk and Meng (2010). PCG extends Gibbs sampling in a manner analogous to ECME's extension of ECM: both PCG and ECME reduce conditioning in a subset of their parameter updates (Park and van Dyk, 2009). The analogy is not perfect, however. In ECME, for example, the CM-steps maximizing the constrained actual likelihood must be last to guarantee monotone convergence (Meng and van Dyk, 1997). On the other hand, with PCG, the corresponding partially collapsed steps must be the first to guarantee a proper sampler.

For ECME, numerical methods, such as Newton-Raphson, may be used to maximize the actual likelihood if no closed-form solution is available. In the context of PCG samplers, these Newton-Raphson steps can often be implemented using MH updates.

Here we illustrate how this is done by using an ECME algorithm developed for a factor analysis model by Liu and Rubin (1998). They derived EM and ECME algorithms and showed that ECME with Newton-type updates converges more quickly than EM. Analogously, it is natural to expect that when fitting this model under a Bayesian framework, a proper MH within PCG sampler will be more efficient than its parent Gibbs sampler. Liu and Rubin (1998) considered the model,

$$Y_i \sim \mathrm{N}_p\Big[Z_i\beta, \Sigma = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)\Big], \text{ for } i = 1, \ldots, n, \tag{16}$$

where Y_i is the $(1 \times p)$ vector for observation i, Z_i is the $(1 \times q)$ vector of the q factors, σ_j^2 is component j of the diagonal variance-covariance matrix, and β is the $(q \times p)$ matrix of factor loadings. We use β_j to represent column j of β and set $Y = \left(Y_1^{\mathrm{T}}, \ldots, Y_n^{\mathrm{T}}\right)^{\mathrm{T}}$ and $Z = \left(Z_1^{\mathrm{T}}, \ldots, Z_n^{\mathrm{T}}\right)^{\mathrm{T}}$. We use $\mathrm{N}_q(0, I)$ as the prior for Z_i $(i = 1, \ldots, n)$ and specify noninformative priors for β and Σ, that is, $p(\sigma_j^2) = \mathrm{Inv\text{-}Gamma}(0.01, 0.01)$ and $p(\beta_j) = \mathrm{N}_q\left[0, V = \mathrm{Diag}(100, \ldots, 100)\right]$ $(j = 1, \ldots, p)$. Ghosh and Dunson (2009) discuss this model and its priors in detail.

Sampler 12 (see top panel of Figure 13) is a standard Gibbs sampler in which each complete conditional distribution can be sampled directly. To improve its convergence, we construct a proper MH within PCG sampler, Sampler 13, which is also given in Figure 13. Because Z is highly correlated with $\sigma_2^2, \ldots, \sigma_5^2$, Sampler 13 updates $\sigma_2^2, \ldots, \sigma_5^2$ without conditioning on Z. Since σ_1^2 converges well with the standard Gibbs sampler in the simulation described below, we do not alter its update in Sampler 13. The reduced updates of $\sigma_2^2, \ldots, \sigma_5^2$ require MH steps. The derivation

of Sampler 13 from its parent Gibbs sampler, i.e., Sampler 12, using the three-phase framework appears in Figure 14.

We use a simulation study to illustrate the improved convergence of the MH within PCG sampler over its parent Gibbs sampler. In particular, we set $p = 5$, $q = 2$, and $n = 100$; σ_j^2 ($j = 1, \ldots, 5$) are generated from Inv-Gamma$(1, 0.25)$ and β_{hj} ($h = 1, 2; j = 1, \ldots, 5$) from N$(0, 3^2)$. We run 20,000 iterations for each sampler with a burnin of 10,000 using the same starting values ($Z_i = [1, 1]^{\text{T}}$, $\beta_{hj} = 1$, and $\sigma_j^2 = 1$). Figure 15 compares Samplers 12 and 13 in terms of mixing, autocorrelation, and density estimation of σ_2^2 and σ_3^2; the first two columns correspond to Sampler 12, and the last two columns correspond to Sampler 13; σ_1^2 converges well for both samplers, and σ_4^2 and σ_5^2 behave similarly as σ_2^2 and σ_3^2. The computational advantage of Sampler 13 is evident. More importantly, the MH within PCG sampler delivers a much more trustworth estimate of the marginal posterior distributions as illustrated in the histograms in Figure 15.

We repeated the simulation with $p = 50$ and $q = 30$ and found that Sampler 13 again outperformed Sampler 12 in a manner similar to what is reported in Figure 15. When run with $p = 50$ and $q = 2$, however, both samplers delivered nearly uncorrelated draws.

6 Discussion

Since its introduction in 2008, the PCG sampler has been deployed to improve the convergence properties of numerous Gibbs-type samplers in a variety of applied settings. As with ordinary Gibbs samplers, MH updates are sometimes required within PCG samplers. Ensuring that the target stationary distribution is maintained in this situation involves subtleties that do not arise in ordinary MH within Gibbs samplers. This has led to the proposal of a number of improper samplers in the literature. This article elucidates these subtleties, offers a strategy for guaranteeing that the target stationary distribution is maintained, and provides advice as to how best to implement MH within PCG samplers. Some of this advice applies equally to ordinary MH within Gibbs samplers. It is commonly understood, for example, that blocking steps within a Gibbs sampler should improve its convergence. We find, however, that this may not be true if MH is involved.

Reducing conditioning in one or more steps of a Gibbs sampler as prescribed by PCG can only improve convergence. If MH is required to implement the reduced steps, however, the overall performance of the algorithm may deteriorate, especially if a poor choice is made for MH jumping rule. Thus, there is a natural trade-off between the computational complexity of MH and the reduced correlation afforded by partial collapsing. Generally speaking, some trial and error may be needed to negotiate this trade-off. In practice we often start with an MH within Gibbs sampler, which already involves MH and can be improved by partial collapsing without any added complexity. We expect our strategies to extend the application of PCG samplers in practice and to provide researchers with additional tools to improve the convergence of Gibbs-type samplers.

Supplemental Mateirals: Details of Samplers 6 and 10–13 appear in an online appendix.

References

Bernardi, M., Gayraud, G., and Petrella, L. (2013). Bayesian inference for CoVaR. Preprint (ArXiv: 1306.2834 [stat.ME]).

Berrett, C. and Calder, C. A. (2012). Data augmentation strategies for the Bayesian spatial probit regression model. *Computational Statistics and Data Analysis* **56**, 478–490.

Caron, F., Teh, Y. W., and Murphy, T. B. (2014). Bayesian nonparametric plackett-luce models for the analysis of preferences for college degree programmes. *Annals of Applied Statistics* under revision.

Dobigeon, N. and Tourneret, J.-Y. (2010). Bayesian orthogonal component analysis for sparse representation. *IEEE Transactions on Signal Processing* **58**, 2675–2685.

Ebrahimi, N., Maasoumi, E., and Soofi, E. (1999). Measuring informativeness of data by entropy and variance. In *Advances in Econometrics, Income Distribution, and Methodology of Science (Essays in Honor of Camilo Dagum)*. Springer-Verlag.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics* **18**, 306–320.

Hans, C., Allenby, G. M., Craigmile, P. F., Lee, J., MacEachern, S. N., and Xu, X. (2012). Covariance

decompositions for accurate computation in Bayesian scale-usage models. *Journal of Computational and Graphical Statistics* **21**, 538–557.

Hu, Y., Gramacy, R. B., and Lian, H. (2012). Bayesian quantile regression for single-index models. *Statistics and Computing* **22**.

Hu, Y., Zhao, K., and Lian, H. (2013). Bayesian quantile regression for partially linear additive models. Preprint (ArXiv: 1307.2668 [stat.CO]).

Kail, G., Tourneret, J.-Y., Hlawatsch, F., and Dobigeon, N. (2010). A partially collapsed Gibbs sampler for parameters with local constraints. In *IEEE International Conference on Acoustics Speech and Signal Processing* (ICASSP), 3886–3889.

Kail, G., Witrisal, K., and Hlawatsch, F. (2011). Direction-resolved estimation of multipath parameters for UWB channels: A partially collapsed Gibbs sampler method. In *IEEE International Conference on Acoustics Speech and Signal Processing* (ICASSP), 3484–3487.

Lee, H., Kashyap, V. L., van Dyk, D. A., Connors, A., Drake, J. J., Izem, R., Meng, X. L., Min, S., Park, T., Ratzlaff, P., Siemiginowska, A., and Zezas, A. (2011). Accounting for calibration uncertainties in X-ray analysis: Effective areas in spectral fitting. *The Astrophysical Journal* **731**, 126–144.

Lin, C. and Tourneret, J.-Y. (2010). P- and T-wave delineation in the ECG signals using a Bayesian approach and a partially collapsed Gibbs sampler. *IEEE Transactions on Biomedical Engineering* **57**, 2840–2849.

Lindsten, F., Schona, T. B., and Jordan, M. I. (2013). Bayesian semiparametric Wiener system identification. *Automatica* **49**, 2053–2063.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–648.

Liu, C. and Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica* **8**, 729–747.

Liu, F., Bayarri, M. J., and Berger, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis* **4(1)**, 119–150.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with sequential PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics* **36**, 19–38.

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *International Journal of Biostatistics* **6(2)**, Article 16.

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.

Park, T. (2011). Bayesian analysis of individual choice behavior with aggregate data. *Journal of Computational and Graphical Statistics* **20**, 158–173.

Park, T., Jeong, J.-H., and Lee, J. W. (2012a). Bayesian nonparametric inference on quantile residual life function: Application to breast cancer data. *Statistics in Medicine* **31**, 1972–1985.

Park, T., Krafty, R. T., and Sánchez, A. I. (2012b). Bayesian semi-parametric analysis of Poisson change-point regression models: application to policy-making in Cali, Colombia. *Journal of Applied Statistics* **39**, 2285–2298.

Park, T. and van Dyk, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics* **18**, 283–305.

Park, T., van Dyk, D. A., and Siemiginowska, A. (2008). Searching for narrow emission lines in X-ray spectra: Computation and methods. *The Astrophysical Journal* **688**, 807–825.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability* **7**, 110–120.

van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal* **548**, 224–243.

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. *Statistical Science* **25**, 429–449.

van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association* **103**, 790–796.

van Dyk, D. A. and Park, T. (2011). Partially collapsed Gibbs sampling and path-adaptive Metropolis-Hastings in high-energy astrophysics. In *Handbook of Markov Chain Monte Carlo* (Editors: S. Brooks, A. Gelman, G. Jones, and X.-L. Meng), 383–399. Chapman & Hall/CRC Press.

Woodard, D. B., Crainiceanu, C., and Ruppert, D. (2012). Hierarchical adaptive regression kernels for regression with functional predictors. *The Journal of Computational and Graphical Statistics,* in press.

Xu, J., van Dyk, D. A., Kashyap, V. L., Siemiginowska, A., Connors, A., Drake, J., Meng, X. L., Ratzlaff, P., and Yu, Y. (2014). A fully Bayesian method to simultaneously fit calibration products and model parameters in X-ray spectral analyses. *Technical Report.*

Zhao, K. and Lian, H. (2014). Bayesian tobit quantile regression with single-index models. *Journal of Statistical Computation and Simulation* to appear.
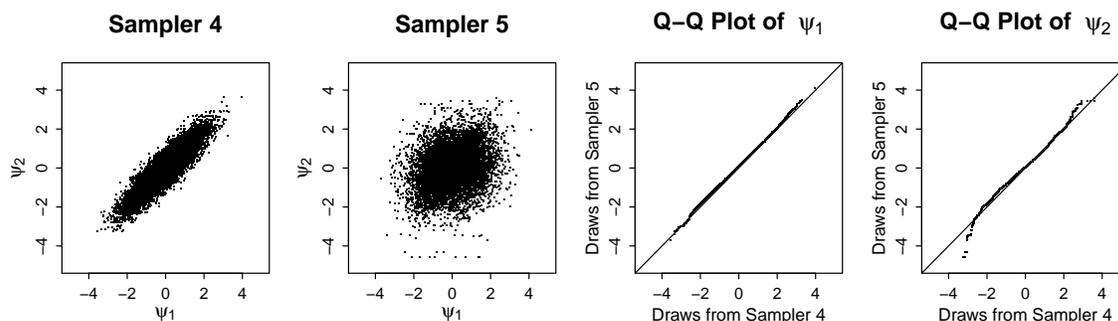
Figure 1: Proper and improper samplers for the bivariate normal target distribution. The first two panels give scatter plots of ψ_1 and ψ_2 for 10,000 draws from Samplers 4 and 5, respectively. The marginal distributions of the two samplers are compared in the two quantile-quantile plots. The improper Sampler 5 severely underestimates the correlation between ψ_1 and ψ_2, and slightly overestimates the variance of ψ_2.

(a) Parent Gibbs Sampler	(b) Reduce Conditioning	(c) Permute	(d) Trim
$\psi_1 \sim p(\psi_1\|\psi_2', \psi_3', \psi_4')$	$(\psi_1, \psi_3^\star) \sim p(\psi_1, \psi_3\|\psi_2', \psi_4')$	$\psi_2 \sim p(\psi_2\|\psi_1', \psi_3', \psi_4')$	$\psi_2 \sim p(\psi_2\|\psi_1', \psi_3', \psi_4')$
$\psi_2 \sim p(\psi_2\|\psi_1, \psi_3', \psi_4')$	$\psi_2 \sim p(\psi_2\|\psi_1, \psi_3^\star, \psi_4')$	$(\psi_1, \psi_3^\star) \sim p(\psi_1, \psi_3\|\psi_2, \psi_4')$	$\psi_1 \sim p(\psi_1\|\psi_2, \psi_4')$
$(\psi_3, \psi_4) \sim p(\psi_3, \psi_4\|\psi_1, \psi_2)$	$(\psi_3, \psi_4) \sim p(\psi_3, \psi_4\|\psi_1, \psi_2)$	$(\psi_3, \psi_4) \sim p(\psi_3, \psi_4\|\psi_1, \psi_2)$	$(\psi_3, \psi_4) \sim p(\psi_3, \psi_4\|\psi_1, \psi_2)$

Figure 2: A three-phase framework for deriving a proper PCG sampler. The parent Gibbs sampler appears in (a). The sampler in (b) reduces the conditioning in Step 1 by updating ψ_3 rather than conditioning on it. The steps of this sampler are permuted in (c) to allow the redundant draw of $\psi_3^\star$—in Step 2 of (c)—to be trimmed in the PCG sampler in (d).

(a) Parent MH within Gibbs Sampler

Step 1: $p(X_L|X, \alpha', \beta', \gamma', \mu', \phi')$
Step 2: $p(\alpha|X, X_L, \beta', \gamma', \mu', \phi')$
Step 3: $\mathcal{M}_{\beta|X,X_L,\alpha,\gamma,\mu,\phi}(\beta|X_L, \alpha, \beta', \gamma', \mu', \phi')$
Step 4: $p(\gamma|X, X_L, \alpha, \beta, \mu', \phi')$
Step 5: $\mathcal{M}_{\mu|X,X_L,\alpha,\beta,\gamma,\phi}(\mu|X_L, \alpha, \beta, \gamma, \mu', \phi')$
Step 6: $\mathcal{M}_{\phi|X,X_L,\alpha,\beta,\gamma,\mu}(\phi|X_L, \alpha, \beta, \gamma, \mu, \phi')$

(b) Reduce Conditioning

Step 1: $p(X_L^\star|X, \alpha', \beta', \gamma', \mu', \phi')$
Step 2: $p(\alpha^\star, X_L^\star|X, \beta', \gamma', \mu', \phi')$
Step 3: $\mathcal{M}_{\beta,X_L,\alpha|X,\gamma,\mu,\phi}^\star(\beta, X_L^\star, \alpha^\star|\beta', \gamma', \mu', \phi')$
Step 4: $p(\gamma|X, X_L^\star, \alpha^\star, \beta, \mu', \phi')$
Step 5: $\mathcal{M}_{\mu,X_L,\alpha|X,\beta,\gamma,\phi}^\star(\mu, X_L^\star, \alpha^\star|\beta, \gamma, \mu', \phi')$
Step 6: $\mathcal{M}_{\phi,X_L,\alpha|X,\beta,\gamma,\mu}^\star(\phi, X_L, \alpha|\beta, \gamma, \mu, \phi')$

(c) Permute

Step 1: $\mathcal{M}_{\mu,X_L,\alpha|X,\beta,\gamma,\phi}^\star(\mu, X_L^\star, \alpha^\star|\beta', \gamma', \mu', \phi')$
Step 2: $\mathcal{M}_{\phi,X_L,\alpha|X,\beta,\gamma,\mu}^\star(\phi, X_L^\star, \alpha^\star|\beta', \gamma', \mu, \phi')$
Step 3: $\mathcal{M}_{\beta,X_L,\alpha|X,\gamma,\mu,\phi}^\star(\beta, X_L^\star, \alpha^\star|\beta', \gamma', \mu, \phi)$
Step 4: $p(\alpha, X_L^\star|X, \beta, \gamma', \mu, \phi)$
Step 5: $p(X_L|X, \alpha, \beta, \gamma', \mu, \phi)$
Step 6: $p(\gamma|X, X_L, \alpha, \beta, \mu, \phi)$

(d) Trim

Step 1: $\mathcal{M}_{\mu|X,\beta,\gamma,\phi}(\mu|\beta', \gamma', \mu', \phi')$
Step 2: $\mathcal{M}_{\phi|X,\beta,\gamma,\mu}(\phi|\beta', \gamma', \mu, \phi')$
Step 3: $\mathcal{M}_{\beta|X,\gamma,\mu,\phi}(\beta|\beta', \gamma', \mu, \phi)$
Step 4: $p(\alpha|X, \beta, \gamma', \mu, \phi)$
Step 5: $p(X_L|X, \alpha, \beta, \gamma', \mu, \phi)$
Step 6: $p(\gamma|X, X_L, \alpha, \beta, \mu, \phi)$

Figure 3: Three-phase framework used to derive Sampler 6 from its parent MH within Gibbs sampler. The parent sampler appears in (a) with Steps 3, 5 and 6 requiring MH updates. The conditioning in Steps 2, 3, 5, and 6 is reduced in (b). The steps are permuted in (c) to allow redundant draws of $X_L^\star$ and $\alpha^\star$ to be trimmed in Steps 1–4. The resulting proper MH within PCG sampler, i.e., Sampler 6, appears in (d).

Sampler 6
Step 1: $\mu \sim \mathcal{M}_{\mu|X,\beta,\gamma,\phi}(\mu|\beta', \gamma', \mu', \phi')$,

Step 2: $\phi \sim \mathcal{M}_{\phi|X,\beta,\gamma,\mu}(\phi|\beta', \gamma', \mu, \phi')$,

Step 3: $\beta \sim \mathcal{M}_{\beta|X,\gamma,\mu,\phi}(\beta|\beta', \gamma', \mu, \phi)$,

Step 4: $\alpha \sim p(\alpha|X, \beta, \gamma', \mu, \phi)$,

Step 5: $X_L \sim p(X_L|X, \alpha, \beta, \gamma', \mu, \phi)$,

Step 6: $\gamma \sim p(\gamma|X, X_L, \alpha, \beta, \mu, \phi)$.

Sampler 7
Step 1: $\mu \sim \mathcal{M}_{\mu|X,\beta,\gamma,\phi}(\mu|\beta', \gamma', \mu', \phi')$,

Step 2: $\phi \sim \mathcal{M}_{\phi|X,\beta,\gamma,\mu}(\phi|\beta', \gamma', \mu, \phi')$,

Step 3: $(\alpha, \beta) \sim \mathcal{M}_{\alpha,\beta|X,\gamma,\mu,\phi}(\alpha, \beta|\alpha', \beta', \gamma', \mu, \phi)$,

Step 4: $X_L \sim p(X_L|X, \alpha, \beta, \gamma', \mu, \phi)$,

Step 5: $\gamma \sim p(\gamma|X, X_L, \alpha, \beta, \mu, \phi)$.

Figure 4: Samplers 6 and 7. Figure 3 verifies the propriety of Sampler 6, an MH within PCG sampler for fitting the spectral model in (1). Sampler 7 blocks Steps 3 and 4 of Sampler 6 into a single MH step. Unfortunately, this results in an improper sampler, see Section 3.3. Details of Sampler 6 are given in Section B.1 of the online appendix.
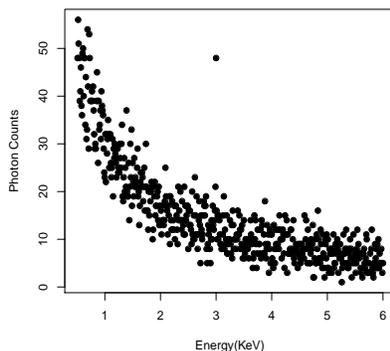
Figure 5: A dataset simulated under the spectral model (1) and used in the simulation study in Section 3.3.
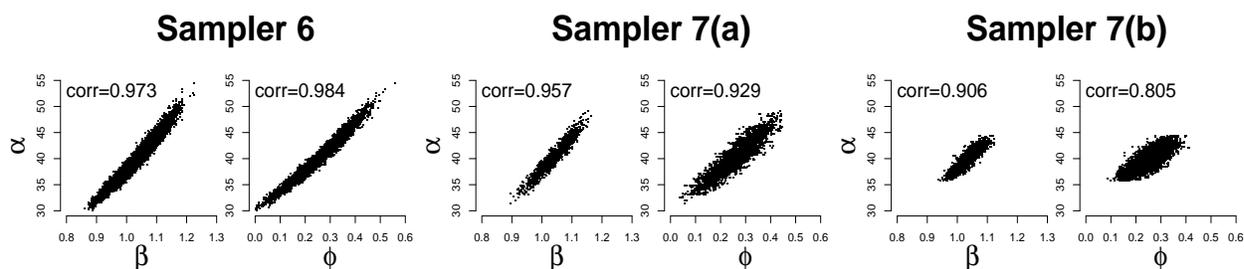

Figure 6: Scatter plots of α, β and ϕ for 10,000 draws from Samplers 6, 7(a) and 7(b) respectively. The two versions of Sampler 7 block the two steps of Sampler 6 that update α and β. Unfortunately, this results in an improper sampler. When updating (α, β), Sampler 7(a) uses the concatenation of Sampler 6's jumping rules for α and β, while Sampler 7(b) uses an independent bivariate normal jumping rule. The impropriety of Sampler 7(b) is especially dramatic.

Q–Q Plot of α **Q–Q Plot of** β **Q–Q Plot of** ϕ
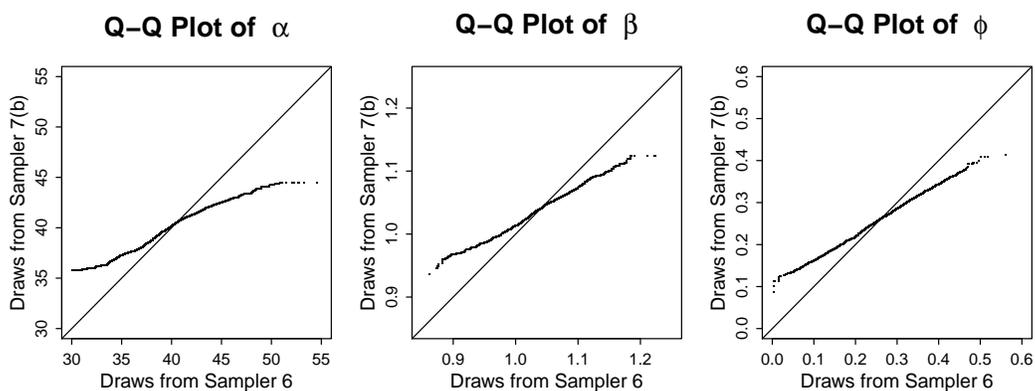
Figure 7: Quantile-quantile plots of α, β and ϕ corresponding to draws generated with Samplers 6 and 7(b). Sampler 7(b) severely underestimates the marginal variances of all three parameters.
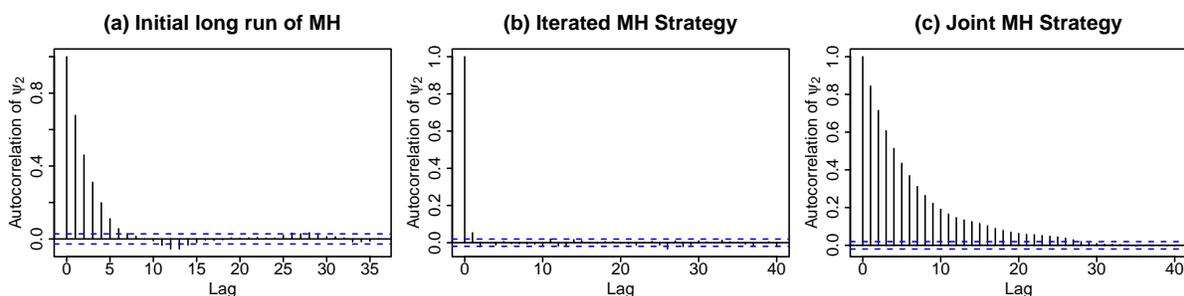
(a) Initial long run of MH   **(b) Iterated MH Strategy**   **(c) Joint MH Strategy**

Figure 8: Autocorrelation functions of ψ_2 for (a) an initial MH run of Step 2 of Sampler 5 with ψ_1 fixed, (b) the iterated MH strategy, and (c) the joint MH strategy, all under the bivariate normal simulation of Section 2.3. Panel (a) shows that the initial MH runs deliver essentially independent draws after 7 iterations, so that iterated MH strategy was run with $L = 7$. Panels (b) and (c) show that the iterated strategy outperforms the joint strategy.
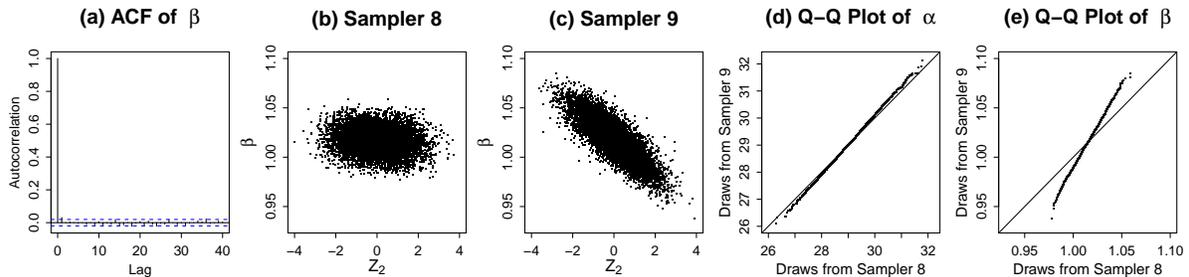
Figure 9: Numerical evaluation of Samplers 8 and 9 using data simulated under model (15). (a): the diagnostic plot suggested in Section 4.1 for the choice of $L = 20$ in Sampler 9. Since the lag-one autocorrelation of $\beta^{(t)}$ is essentially zero, L is sufficiently large. (b) and (c): scatter plots of Z_2 and β from Samplers 8 and 9 respectively. (d) and (e): quantile-quantile plots of α and β respectively. Sampler 9 is (approximately) proper while Sampler 8 is improper and underestimates the correlation between Z_2 and β and also the marginal variability of both α and β.

Figure 10: Samplers 10 and 11. Sampler 10 is the proper MH within PCG sampler for the spectral model (1) with the lowest degree of partial collapsing, while Sampler 11 is that with the highest degree of partial collapsing. Details of Samplers 10 and 11 appear in Section B.1 of the online appendix.

(a) Parent MH within Gibbs Sampler

Step 1: $p(X_L | X, \alpha', \beta', \gamma', \mu', \phi')$
Step 2: $p(\alpha | X, X_L, \beta', \gamma', \mu', \phi')$
Step 3: $\mathcal{M}_{\beta | X, X_L, \alpha, \gamma, \mu, \phi}(\beta | X_L, \alpha, \beta', \gamma', \mu', \phi')$
Step 4: $p(\gamma | X, X_L, \alpha, \beta, \mu', \phi')$
Step 5: $\mathcal{M}_{\mu | X, X_L, \alpha, \beta, \gamma, \phi}(\mu | X_L, \alpha, \beta, \gamma, \mu', \phi')$
Step 6: $\mathcal{M}_{\phi | X, X_L, \alpha, \beta, \gamma, \mu}(\phi | X_L, \alpha, \beta, \gamma, \mu, \phi')$

(b) Reduce Conditioning

Step 1: $p(X_L^\star | X, \alpha', \beta', \gamma', \mu', \phi')$
Step 2: $p(\alpha^\star, X_L^\star | X, \beta', \gamma', \mu', \phi')$
Step 3: $\mathcal{M}_{\beta, X_L, \alpha, \phi | X, \gamma, \mu}^\star(\beta^\star, X_L^\star, \alpha^\star, \phi^\star | \beta', \gamma', \mu', \phi')$
Step 4: $p(\gamma | X, X_L^\star, \alpha^\star, \beta^\star, \mu', \phi^\star)$
Step 5: $\mathcal{M}_{\mu, X_L, \alpha | X, \beta, \gamma, \phi}^\star(\mu, X_L^\star, \alpha^\star | \beta^\star, \gamma, \mu', \phi^\star)$
Step 6: $\mathcal{M}_{\phi, X_L, \alpha, \beta | X, \gamma, \mu}^\star(\phi, X_L, \alpha, \beta | \beta^\star, \gamma, \mu, \phi^\star)$

(c) Permute

Step 1: $\mathcal{M}_{\mu, X_L, \alpha | X, \beta, \gamma, \phi}^\star(\mu, X_L^\star, \alpha^\star | \beta', \gamma', \mu', \phi')$
Step 2: $\mathcal{M}_{\phi, X_L, \alpha, \beta | X, \gamma, \mu}^\star(\phi^\star, X_L^\star, \alpha^\star, \beta^\star | \beta', \gamma', \mu, \phi')$
Step 3: $\mathcal{M}_{\beta, X_L, \alpha, \phi | X, \gamma, \mu}^\star(\beta, X_L^\star, \alpha^\star, \phi | \beta^\star, \gamma', \mu, \phi^\star)$
Step 4: $p(\alpha, X_L^\star | X, \beta, \gamma', \mu, \phi)$
Step 5: $p(X_L | X, \alpha, \beta, \gamma', \mu, \phi)$
Step 6: $p(\gamma | X, X_L, \alpha, \beta, \mu, \phi)$

(d) Trim

Step 1: $\mathcal{M}_{\mu | X, \beta, \gamma, \phi}(\mu | \beta', \gamma', \mu', \phi')$
Step 2: $\mathcal{M}_{\beta, \phi | X, \gamma, \mu}(\beta^\star, \phi^\star | \beta', \gamma, \mu, \phi')$
Step 3: $\mathcal{M}_{\beta, \phi | X, \gamma, \mu}(\beta, \phi | \beta^\star, \gamma', \mu, \phi^\star)$
Step 4: $p(\alpha | X, \beta, \gamma', \mu, \phi)$
Step 5: $p(X_L | X, \alpha, \beta, \gamma', \mu, \phi)$
Step 6: $p(\gamma | X, X_L, \alpha, \beta, \mu, \phi)$

Figure 11: Three-phase framework used to derive Sampler 11. The parent MH within Gibbs sampler appears in (a). The conditioning in Steps 2, 3, 5, and 6 is reduced in (b) and the steps are permuted in (c) to allow redundant draws of $X_L^\star$ and $\alpha^\star$ to be trimmed in Steps 1–4. This leaves Steps 2 and 3 of (d) identical; one of them is omitted in Sampler 11, see Figure 10. (In (d), $\beta^\star$ and $\phi^\star$ are not redundant because they are among the inputs of Step 3.)
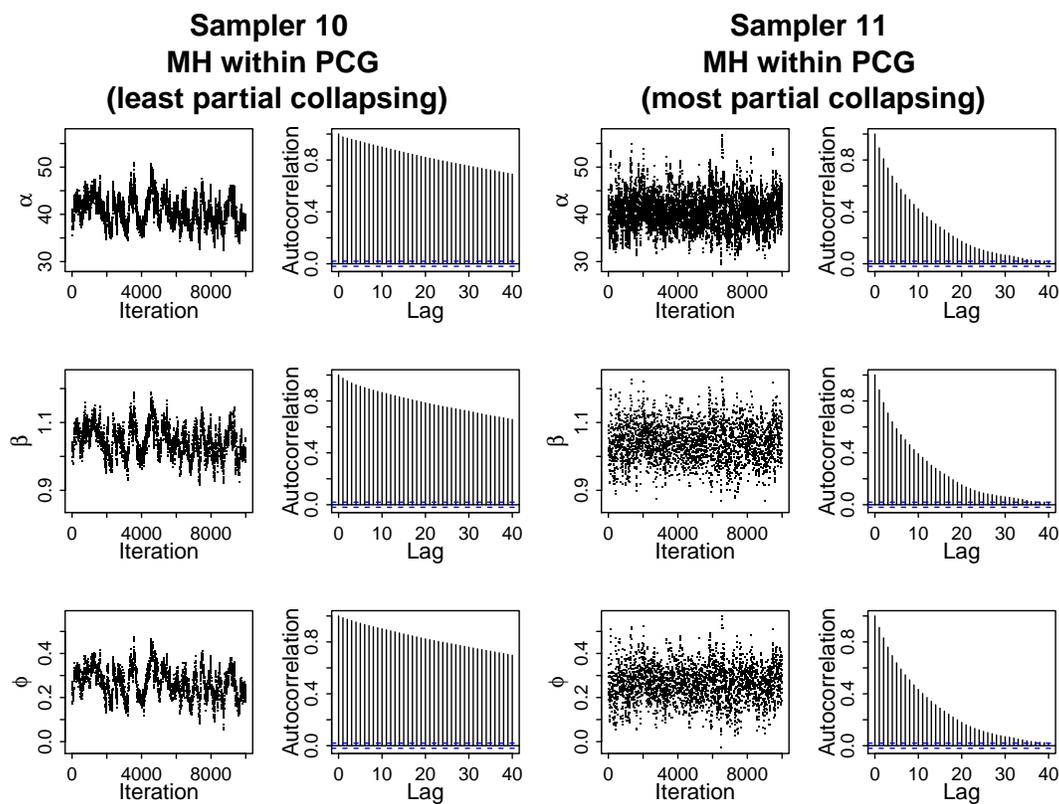
Figure 12: Comparing Samplers 10 and 11 using data simulated under model (1). The first two columns are the time-series and autocorrelation plots for the posterior draws of α, β, and ϕ respectively from Sampler 10, while the last two columns are those from Sampler 11. Sampler 11 performs significantly better than Sampler 10.

Sampler 12

Step 1: $Z_i \sim p(Z_i|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$,

Step 2: $\sigma_j^2 \sim p(\sigma_j^2|Y, Z, \beta')$, for $j = 1, \ldots, 5$,

Step 3: $\beta_j \sim p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 5$.

Sampler 13

Step 1: $\sigma_1^2 \sim p(\sigma_1^2|Y, Z', \beta')$,

Step j: $\sigma_j^2 \sim \mathcal{M}_{\sigma_j^2|Y,\beta,\sigma_1^2,\ldots,\sigma_{j-1}^2,\sigma_{j+1}^2,\ldots,\sigma_5^2}(\sigma_j^2|\beta', \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_j^{2\prime}, \ldots, \sigma_5^{2\prime})$, for $j = 2, \ldots, 5$,

Step 6: $Z_i \sim p(Z_i|Y, \beta', \Sigma)$, for $i = 1, \ldots, 100$,

Step 7: $\beta_j \sim p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 5$.

Figure 13: Two samplers for fitting (16). Sampler 12 is a standard Gibbs sampler and Sampler 13 is a proper MH within PCG sampler. Notice that Sampler 13 does not condition on Z in its updates of $\sigma_2^2, \ldots, \sigma_5^2$. Details of Samplers 12 and 13 appear in Section B.2 of the online appendix

(a) Parent Gibbs Sampler (Sampler 12)

Step 1: $p(Z_i|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$
Step 2: $p(\sigma_j^2|Y, Z, \beta')$, for $j = 1, \ldots, 5$
Step 3: $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 5$

(b) Reduce Conditioning

Step 1: $p(Z_i^\star|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$
Step 2: $p(\sigma_1^2|Y, Z^\star, \beta')$
Step 1 + j: $\mathcal{M}^\star_{\sigma_j^2, Z|Y, \beta, \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_{j+1}^2, \ldots, \sigma_5^2}(\sigma_j^2, Z^\star|\beta', \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_j^{2\prime}, \ldots, \sigma_5^{2\prime})$, for $j = 2, 3, 4$

Step 6: $\mathcal{M}^\star_{\sigma_5^2, Z|\beta, \sigma_1^2, \ldots, \sigma_4^2}(\sigma_5^2, Z|\beta', \sigma_1^2, \ldots, \sigma_4^2, \sigma_5^{2\prime})$
Step 7: $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 5$

(c) Permute

Step 1: $p(\sigma_1^2|Y, Z', \beta')$
Step j: $\mathcal{M}^\star_{\sigma_j^2, Z|Y, \beta, \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_{j+1}^2, \ldots, \sigma_5^2}(\sigma_j^2, Z^\star|\beta', \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_j^{2\prime}, \ldots, \sigma_5^{2\prime})$, for $j = 2, \ldots, 5$
Step 6: $p(Z_i|Y, \beta', \Sigma)$, for $i = 1, \ldots, 100$
Step 7: $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 5$

(d) Trim (Sampler 13)

Step 1: $p(\sigma_1^2|Y, Z', \beta')$,
Step j: $\mathcal{M}_{\sigma_j^2|Y, \beta, \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_{j+1}^2, \ldots, \sigma_5^2}(\sigma_j^2|\beta', \sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_j^{2\prime}, \ldots, \sigma_5^{2\prime})$, for $j = 2, \ldots, 5$,
Step 6: $p(Z_i|Y, \beta', \Sigma)$, for $i = 1, \ldots, 100$,
Step 7: $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 5$.

Figure 14: Using the three-phase framework to derive Sampler 13 from its parent Gibbs sampler, i.e., Sampler 12. The parent Gibbs sampler is in (a); the conditioning in Steps 3–6 is reduced in (b); and the steps are permuted in (c) to allow redundant draws of $Z^\star$ to be trimmed in Steps 2–5. The resulting proper Sampler 13 is in (d).
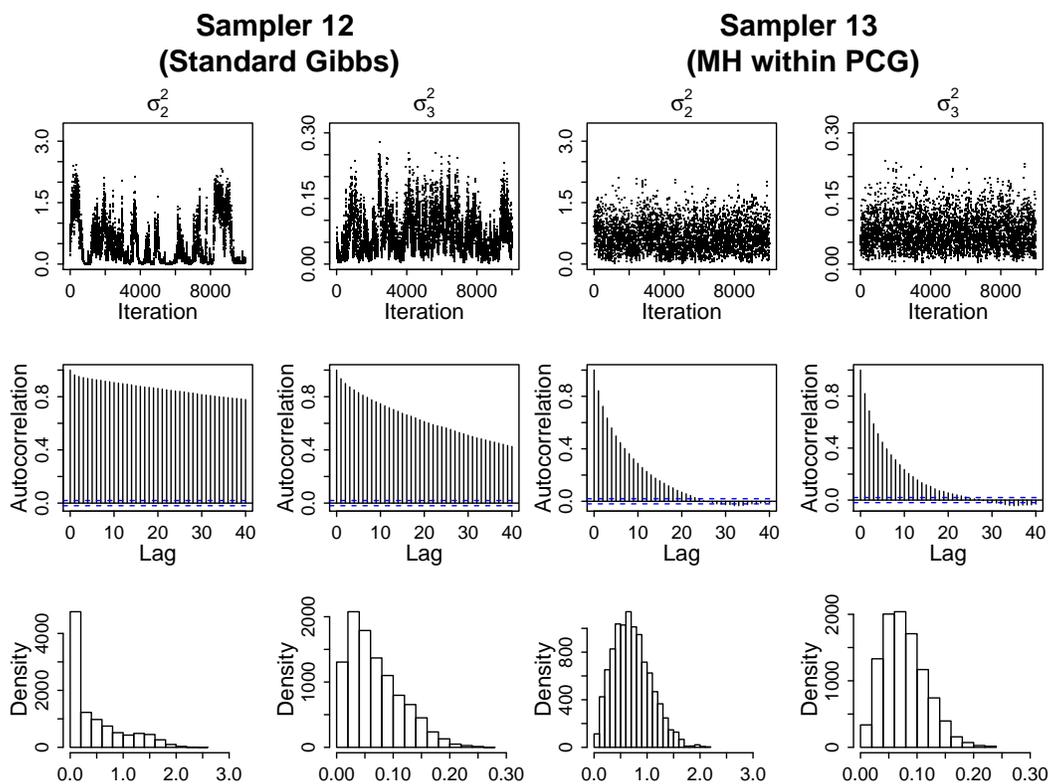
Figure 15: Comparing Samplers 12 and 13 using data simulated under the factor analysis model (16). The first two columns are the time-series, autocorrelation, and histogram plots for the posterior draws of σ_2^2 and σ_3^2 respectively from Sampler 12, while the last two columns are those from Sampler 13. Sampler 13 performs significantly better than Sampler 12 both in terms of convergence properties and in its estimates of the marginal posterior distributions.

——— **End of Document** ———

(The following is included to establish labels for the sections of the online appendix.)

A Stationary Distribution of Sampler Fragment 6

B Details of the Steps in the Gibbs-type Samplers

B.1 Details of the steps in the Gibbs-type samplers based on model (1)

B.2 Details of the steps in the Gibbs-type samplers based on model (16)