Supplemental materials for this article are available through the JCGS web page at http://www.amstat.org/publications.

Partially Collapsed Gibbs Samplers: Illustrations and Applications

Taeyoung PARK and David A. VAN DYK

Among the computationally intensive methods for fitting complex multilevel models, the Gibbs sampler is especially popular owing to its simplicity and power to effectively generate samples from a high-dimensional probability distribution. The Gibbs sampler, however, is often justifiably criticized for its sometimes slow convergence, especially when it is used to fit highly structured complex models. The recently proposed Partially Collapsed Gibbs (PCG) sampler offers a new strategy for improving the convergence characteristics of a Gibbs sampler. A PCG sampler achieves faster convergence by reducing the conditioning in some or all of the component draws of its parent Gibbs sampler. Although this strategy can significantly improve convergence, it must be implemented with care to be sure that the desired stationary distribution is preserved. In some cases the set of conditional distributions sampled in a PCG sampler may be functionally incompatible and permuting the order of draws can change the stationary distribution of the chain. In this article, we draw an analogy between the PCG sampler and certain efficient EM-type algorithms that helps to explain the computational advantage of PCG samplers and to suggest when they might be used in practice. We go on to illustrate the PCG samplers in three substantial examples drawn from our applied work: a multilevel spectral model commonly used in high-energy astrophysics, a piecewise-constant multivariate time series model, and a joint imputation model for nonnested data. These are all useful highly structured models that involve computational challenges that can be solved using PCG samplers. The examples illustrate not only the computation advantage of PCG samplers but also how they should be constructed to maintain the desired stationary distribution. Supplemental materials for the examples given in this article are available online.

Key Words: AECM algorithm; Astrophysical data analysis; ECME algorithm; Incompatible Gibbs sampler; Marginal data augmentation; Multiple imputation; Spectral analysis.

1. INTRODUCTION

The development of Markov chain Monte Carlo (MCMC) methods over the past twenty years has revolutionized modern applied statistics, and has particularly influenced and

Taeyoung Park is Assistant Professor, Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260 (E-mail: *tpark@pitt.edu*). David A. van Dyk is Professor, Department of Statistics, University of California, Irvine, CA 92697 (E-mail: *dvd@uci.edu*).

Journal of Computational and Graphical Statistics, Volume 18, Number 2, Pages 283–305
DOI: 10.1198/jcgs.2009.08108

popularized Bayesian methods. More complex models that explicitly aim to incorporate application-specific stochastic features of a data generation mechanism are becoming more prevalent as a direct result of these sophisticated computational tools. Implementing MCMC samplers, however, is a nuanced business that often is as much a matter of intuition and art as it is a matter of science. Predicting the convergence characteristics of a sampler without making the large investment that is required to implement the sampler is often an impossible task. Indeed, accessing the convergence of a sampler after it has been implemented requires subtle diagnostics, and it is not difficult to be fooled into prematurely concluding that a sampler has fully explored a distribution.

Fortunately, much work has been devoted to developing practical strategies that serve to improve the convergence characteristics of MCMC samplers. In the context of Gibbs sampling, it is well known that blocking or grouping steps (Liu, Wong, and Kong 1994), nesting steps (van Dyk 2000b), collapsing or marginalizing parameters (Liu 1994; Meng and van Dyk 1999), incorporating auxiliary variables (Besag and Green 1993), certain parameter transformations (Gelfand, Sahu, and Carlin 1995; Liu 2003; Yu 2005), and parameter expansion (Liu and Wu 1999) can all be used to improve the convergence of certain samplers. Many of these strategies took their cue from or are analogous to similar techniques that are known to speed the convergence of EM-type algorithms (e.g., van Dyk and Meng 2001, 2009; Gelman et al. 2008). The EM algorithm (Dempster, Laird, and Rubin 1977) can be used to compute the posterior mode of the parameters by embedding the sampling distribution under the model into a joint distribution of the model parameters and a set of "latent variables" or "missing data" and performing iterative calculations based on the resulting conditional distributions of the parameters given the missing data and of the missing data given the parameters.

Collapsing and marginalization methods offer an example of the relationship between efficient EM-type algorithms and methods for improving the convergence of the Gibbs sampler. These methods integrate the joint posterior distribution of the unknown quantities, including unknown parameters, latent variables, and missing data, over some of these unknown quantities to construct a marginal posterior distribution under which a new collapsed Gibbs sampler is built (Liu 1994). In the context of EM algorithms, on the other hand, it is well known that the rate of convergence is improved by reducing the missing data in the model formulation, that is, by integrating the joint distribution over a portion of the missing data and deriving a new faster EM algorithm on the marginal distribution (Meng and van Dyk 1997; van Dyk 2000a). Of course, such strategies are generally only useful when the marginal distribution allows for the construction of simple closed form Gibbs samplers or EM algorithms.

Variants of the EM algorithm have been developed to take advantage of the basic idea behind marginalization even when a closed form EM algorithm is not available on the marginal distribution. The ECME algorithm (Liu and Rubin 1994), for example, allows one group of parameters to be updated using conditional distributions from the joint distribution and a second group to be updated using conditional distributions of the marginal distribution of the model parameters. The second group of parameters is updated by completely marginalizing out the latent variables and missing data and using a conditional distribution of the resulting marginal distribution. A generalization of the ECME algorithm, known as the AECM algorithm (Meng and van Dyk 1997), allows each of several groups of the parameters to be updated using conditional distributions of different margins of the joint posterior distribution. Relative to the collapsing strategy described in the previous paragraph, both the ECME and AECM algorithms can be described as *partially collapsed methods* in that they do not fully marginalize out any component of the missing data but rather marginalize out different components in different parts of the algorithm. This is the basic strategy that we aim to apply to the Gibbs sampler in this article. Because both the ECME and AECM algorithms have proved successful in a variety of applications, we expect from the onset that the resulting samplers will also exhibit improved convergence properties.

Van Dyk and Park (2008) developed the theory and methods necessary to apply this strategy when using a Gibbs sampler. The *partially collapsed Gibbs* (PCG) sampler replaces some of the conditional distributions of an ordinary Gibbs sampler with conditional distributions of some *marginal distributions* of the target joint posterior distribution. As with EM-type algorithms, this strategy is useful because it can result in algorithms with much better convergence properties and it is interesting because it may result in sampling from a set of *incompatible* conditional distributions. That is, there may be no joint distribution that corresponds to this set of conditional distributions. Although van Dyk and Park (2008) outlined the necessary methodology for the PCG sampler, their work contains only one simplified toy example. The primary goal of this article is to illustrate how the PCG sampler is used in real examples and to demonstrate the computational advantage of the strategy. In particular, we illustrate how the PCG sampler can be used to fit a multilevel spectral model commonly used in high-energy astrophysics, a piecewise-constant multivariate time series model, and a joint imputation model for nonnested data. These are all useful models that the authors came across in their applied work and that involve computational challenges that can be solved using PCG samplers.

The remainder of the article is divided into five sections. We begin in Section 2 by reviewing the motivation behind and basic strategies for PCG samplers. Sections 3–5 illustrate the implementation and computational advantage of PCG samplers in a sequence of three examples. Section 3 details the dramatic improvement in mixing that can be obtained using PCG samplers to fit a multilevel spectral model commonly used in high-energy astrophysics. The second example is described in Section 4 and illustrates how partial collapse makes intractable sampling steps tractable with the expectation of quicker convergence. Section 5 presents the third example where the data structure intrinsically suggests PCG sampling strategies. Concluding remarks are given in Section 6.

2. PARTIALLY COLLAPSED GIBBS SAMPLERS

Van Dyk and Park (2008) described three basic tools that can be used to transform a Gibbs sampler into a PCG sampler. The first tool is *marginalization* which involves moving a group of unknowns from being conditioned upon to being sampled in one or more steps of a Gibbs sampler; the marginalized group can differ among the steps. Second, we may need to *permute* the steps of the sampler to use the third tool, which is to *trim* sampled components from the various steps that can be removed from the sampler without altering its Markov transition kernel. Marginalization and permutation both trivially maintain the stationary distribution of a Gibbs sampler and both can affect the convergence properties of the chain; marginalization can dramatically improve convergence, whereas the effect of a permutation is typically small. Trimming, on the other hand, is explicitly designed to maintain the kernel of the chain. Its primary advantage is to reduce the complexity and the computational burden of the individual steps. It is trimming that introduces incompatibility into the sampler. In this section we use a simple schematic example to illustrate how the three tools are used and conclude with a review of the theoretical properties of PCG samplers. Details can be found in van Dyk and Park (2008).

We begin with the four-step Gibbs sampler:

$$\begin{aligned}
&\text{Step 1: Draw } \mathbf{W} \text{ from } p(\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z}),\\
&\text{Step 2: Draw } \mathbf{X} \text{ from } p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z}),\\
&\text{Step 3: Draw } \mathbf{Y} \text{ from } p(\mathbf{Y}|\mathbf{W},\mathbf{X},\mathbf{Z}), \text{ and}\\
&\text{Step 4: Draw } \mathbf{Z} \text{ from } p(\mathbf{Z}|\mathbf{W},\mathbf{X},\mathbf{Y}),
\end{aligned} \qquad \text{(Sampler 1)}$$

and suppose it is possible to directly sample from $p(\mathbf{Y}|\mathbf{X},\mathbf{Z})$ and $p(\mathbf{Z}|\mathbf{X},\mathbf{Y})$, which are both conditional distributions of $\int p(\mathbf{W},\mathbf{X},\mathbf{Y},\mathbf{Z})\,d\mathbf{W}$. If we were to simply replace Steps 3 and 4 with these conditional draws we would have no direct way of verifying that the stationary distribution of the resulting chain is the target joint distribution, $p(\mathbf{W},\mathbf{X},\mathbf{Y},\mathbf{Z})$. Thus, we use the three basic tools to reap the computational gains of partial collapse while ensuring that the resulting chain maintains the target distribution as its stationary distribution.

The first tool is marginalization. It allows us to move $\mathbf{W}$ from being conditioned upon to being sampled in Steps 3 and 4:

$$\begin{aligned}
&\text{Step 1: Draw } \mathbf{W}^\star \text{ from } p(\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z}),\\
&\text{Step 2: Draw } \mathbf{X} \text{ from } p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z}),\\
&\text{Step 3: Draw } (\mathbf{W}^\star,\mathbf{Y}) \text{ from } p(\mathbf{W},\mathbf{Y}|\mathbf{X},\mathbf{Z}), \text{ and}\\
&\text{Step 4: Draw } (\mathbf{W},\mathbf{Z}) \text{ from } p(\mathbf{W},\mathbf{Z}|\mathbf{X},\mathbf{Y}).
\end{aligned} \qquad \text{(Sampler 2)}$$

In each step we condition on the most recently sampled value of each quantity that is not sampled in that step. The output of the iteration consists of the most recently sampled value of each quantity at the end of the iteration: $\mathbf{X}$ sampled in Step 2, $\mathbf{Y}$ sampled in Step 3, and $(\mathbf{W},\mathbf{Z})$ sampled in Step 4. Here and elsewhere we use a superscript '$\star$' to designate an *intermediate quantity* that is sampled but is not part of the output of an iteration. Because $\mathbf{W}$ is sampled in multiple steps during an iteration, Sampler 2 is a simple generalization of an ordinary Gibbs sampler that updates each component once in an iteration. In this regard Sampler 2 is similar to the alternating subspace-spanning resampling algorithm (Liu 2003); it updates some components multiple times within each iteration. Marginalization does not affect the stationary distribution of the chain, but as we discuss later it is the source of the computational gain of PCG samplers over their parent Gibbs samplers.

Sampler 2 may be inefficient in that it samples $\mathbf{W}$ three times. Removing any two of the three draws, however, necessarily affects the transition kernel of the chain because the draw in Step 1 is conditioned on in Step 2 and the draw in Step 4 is part of the output of the iteration. Because we want to preserve the stationary distribution of the chain, we only consider removing intermediate quantities whose values are not conditioned upon subsequently. Permuting the steps of a Gibbs sampler does not alter its stationary distribution but can enable certain intermediate quantities to meet the criterion for removal. Thus, permutation is the second basic tool used in constructing a PCG sampler. In particular the sampler:

Step 1: Draw $(\mathbf{W}^{\star}, \mathbf{Y})$ from $p(\mathbf{W}, \mathbf{Y}|\mathbf{X}, \mathbf{Z})$,

Step 2: Draw $(\mathbf{W}^{\star}, \mathbf{Z})$ from $p(\mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{Y})$,

Step 3: Draw $\mathbf{W}$ from $p(\mathbf{W}|\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, and (Sampler 3)

Step 4: Draw $\mathbf{X}$ from $p(\mathbf{X}|\mathbf{W}, \mathbf{Y}, \mathbf{Z})$,

has the same stationary distribution as Sampler 2. In Sampler 3, however, the intermediate draws of $\mathbf{W}$ sampled in Steps 1 and 2 are not used in the subsequent steps because a new value of $\mathbf{W}$ is sampled in Step 3. Thus both of the intermediate draws of $\mathbf{W}$ can be removed (or trimmed) from the sampler. Replacing Step 2 in Sampler 3 with a draw from $p(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$ is equivalent to blocking Steps 2 and 3 into a joint draw of $\mathbf{W}$ and $\mathbf{Z}$ from $p(\mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{Y})$. Thus, using the third basic tool, that is, trimming the two intermediate quantities and combining Steps 2 and 3, we derive Sampler 4:

Step 1: Draw $\mathbf{Y}$ from $p(\mathbf{Y}|\mathbf{X}, \mathbf{Z})$,

Step 2: Draw $(\mathbf{W}, \mathbf{Z})$ from $p(\mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{Y})$, and (Sampler 4)

Step 3: Draw $\mathbf{X}$ from $p(\mathbf{X}|\mathbf{W}, \mathbf{Y}, \mathbf{Z})$.

Because removing the intermediate quantities does not affect the transition kernel of the chain, Sampler 4 has the same stationary distribution as Sampler 3 which we know has the target stationary distribution. Thus, by carefully using the three basic tools, we are guaranteed to arrive at a PCG sampler with the desired stationary distribution. We emphasize that Sampler 4 is not a Gibbs sampler per se. The three conditional distributions that are sampled are incompatible and permuting the order of the draws may alter the stationary distribution of the chain.

Using the basic tools sometimes leads to conditional distributions that can be combined to form a set of compatible conditional distributions. (This type of combination of steps occurred when we combined Steps 2 and 3 of Sampler 3 after trimming Step 2 to form Step 2 of Sampler 4.) When some conditional distributions are combined and the resulting set of conditional distributions remains compatible, the PCG sampler is simply a blocked version of its parent Gibbs sampler. As Sampler 4 illustrates, however, this does not always occur. In this regard PCG sampling can be viewed as a generalization of blocking.

The benefit of using PCG samplers is more efficient convergence. Van Dyk and Park (2008) showed that sampling more components in any set of steps of a Gibbs sampler (i.e., marginalization) tends to improve the convergence of the chain.[1] Rather than detailing these technical results, we illustrate the computational gain empirically in several examples. The example in Section 3, for example, shows a fast converging PCG sampler whose parent Gibbs sampler sticks at its starting value. Another potential advantage is illustrated in Section 4 where a conditional draw of the parent Gibbs sampler is intractable but the PCG sampler is easy to implement. We now turn to these empirical illustrations.

3. MULTILEVEL SPECTRAL MODEL IN HIGH-ENERGY ASTROPHYSICS

The data for our first example are collected to explore the distribution of the energies of photons emanating from the quasar PG 1634+706 (Park, van Dyk, and Siemiginowska 2008). (These datasets are available in the online supplemental materials folder to this article on the JCGS webpage.) The distribution can be formulated as a finite mixture distribution composed of a continuum term, which is a smooth function across a wide range of energies, and an emission line, which is a local feature highly focused on a narrow band of energies. Because photons are counted in a number of energy bins, the expected Poisson counts in energy bin j are modeled as

$$\Lambda_j(\boldsymbol{\theta}) \equiv f_j(\boldsymbol{\theta}^C) + \lambda\pi_j(\mu, \sigma^2) \quad \text{for } j = 1, \ldots, J, \tag{3.1}$$

where $\boldsymbol{\theta}$ is the set of model parameters, $f_j(\boldsymbol{\theta}^C)$ is the expected continuum counts in bin j, $\boldsymbol{\theta}^C$ is the set of free parameters in the continuum model, λ is the expected line counts, and $\pi_j(\mu, \sigma^2)$ is the proportion of an emission line with mean μ and variance σ^2 falling into bin j, which can be modeled with a narrow Gaussian distribution or a delta function. In this article, a power law model is used to describe the continuum term, that is, $f_j(\boldsymbol{\theta}^C) = \alpha^C E_j^{-\beta^C}$, where $\boldsymbol{\theta}^C = (\alpha^C, \beta^C)$ represent the normalization and photon index of the power law continuum, respectively, and E_j is the energy of bin j. Because the spectral data are subject to stochastic redistribution, stochastic censoring, and background contamination, we consider a more refined model. In particular, the observed photon counts and background photon counts in detector channel l are modeled with independent Poisson distributions,

$$\begin{aligned} Y_{\mathrm{src}\,l} &\sim \mathrm{Poisson}\Bigl(\sum_j M_{lj}\Lambda_j(\boldsymbol{\theta})u_j(\boldsymbol{\theta}^A) + \theta_l^B\Bigr) \qquad \text{and} \\ Y_{\mathrm{bkg}\,l} &\sim \mathrm{Poisson}(\kappa\theta_l^B) \quad \text{for } l = 1, \ldots, L, \end{aligned} \tag{3.2}$$

where $\mathbf{Y}_{\mathrm{obs}} = \{(Y_{\mathrm{src}\,l}, Y_{\mathrm{bkg}\,l}), l = 1, \ldots, L\}$ denotes a collection of the observed source and background counts, M_{lj} is the probability that a photon that arrives with energy corresponding to bin j is recorded in channel l, $u_j(\boldsymbol{\theta}^A)$ is the probability that a photon with energy corresponding to bin j is *not* absorbed and is parameterized by $\boldsymbol{\theta}^A$, θ_l^B is a Poisson

[1] In particular, van Dyk and Park studied the forward operator $\mathbf{F}_0$ induced by $\mathbf{F}h(\mathbf{X}') = \mathrm{E}\{h(\mathbf{X}^{(1)})|\mathbf{X}^{(0)} = \mathbf{X}'\}$ on the Hilbert space $\{h : \mathrm{E}_\pi\{h(\mathbf{X})\} = 0, \mathrm{var}_\pi\{h(\mathbf{X})\} < \infty\}$. The spectral radius of $\mathbf{F}_0$ typically governs the convergence of the Markov chain (Liu 2001). Van Dyk and Park derived a bound on this spectral radius and showed that their bound can only be reduced when more components are sampled in any step of a Gibbs sampler.

intensity of the background counts in channel l, and κ is a known correction factor to adjust for the difference between source and background exposures. In this article, the absorption probability is modeled as $u_j(\theta^A) = d_j \exp(-\theta^A X(E_j))$, where d_j is the known effective area of bin j and $X(E_j)$ is a tabulated value (Morrison and McCammon 1983). The finite mixture and data generation process can be described by a hierarchical structure of missing data and a Gibbs sampler can be constructed to fit the model (van Dyk et al. 2001). In particular, the set of missing data $\mathbf{Y}_{\text{mis}}$ is decomposed into $(\mathbf{Y}_{\text{mis 1}}, \mathbf{Y}_{\text{mis 2}})$, where $\mathbf{Y}_{\text{mis 1}}$ is the unobserved Poisson photon counts with expectation given in (3.1) and $\mathbf{Y}_{\text{mis 2}}$ is the unobserved mixture indicator variable for each of these counts under the finite mixture model given in (3.1).

3.1 DELTA FUNCTION EMISSION LINES

When an emission line is assumed to be narrow enough to fall within one energy bin of the detector, the narrow emission line can be modeled with a delta function; σ^2 in (3.1) is set to zero. In this case, however, the Gibbs sampler breaks down because the resulting subchain for μ does not move from its starting value, regardless of what the starting value is. This happens because the mixture indicator variable for the delta function emission line is zero for all of the energy bins but the one containing the previous iterate of the line location, and thus the next iterate of the line location is necessarily the same as the previous iterate. The only possibility for the line location to change is for no photons to be attributed to the line, a highly unlikely possibility even for a weak emission line. To improve convergence, we consider two variants of a PCG sampler constructed by partially marginalizing over the entire missing data or part of the missing data; we call the two PCG samplers PCG I and PCG II.

PCG I is constructed in the following way. In Figure 1(a), we begin by building a parent Gibbs sampler using a set of complete conditional distributions of the target posterior distribution, $p(\mathbf{Y}_{\text{mis}}, \boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$, where $\boldsymbol{\theta}$ can be partitioned into $\boldsymbol{\theta} = (\boldsymbol{\psi}, \mu)$ with μ being the location parameter for a delta function line and $\boldsymbol{\psi}$ being all the model parameters except the line location parameter. Each conditional distribution in Figure 1(a) follows a stan-

(a) Parent Gibbs sampler

$$p(\mathbf{Y}_{\text{mis}}|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\mu|\mathbf{Y}_{\text{mis}}, \boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$$

(b) Marginalization

$$p(\mathbf{Y}^{\star}_{\text{mis}}|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}, \mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$$

(c) Permute

$$p(\mathbf{Y}^{\star}_{\text{mis}}, \mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$$

(d) Trim

$$p(\mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$$

(e) Block

$$p(\mathbf{Y}_{\text{mis}}, \mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$$

Figure 1. Deriving PCG I for the multilevel spectral model with a delta function emission line. The resulting sampler in (e) is a blocked version of the parent Gibbs sampler in (a).

dard distribution, but the parent Gibbs sampler does not converge because of the effective absorbing state for μ. The convergence characteristics can be improved by marginalizing all of the missing data, $\mathbf{Y}_{\text{mis}}$, in Step 3 as is done in Figure 1(b). Although the resulting marginalized distribution is tractable (see below), the missing data sampled in Step 3 are part of the output quantities and cannot be removed from the sampler. Permuting the sampling steps as in Figure 1(c) allows us to remove $\mathbf{Y}^{\star}_{\text{mis}}$ from Step 1 of Figure 1(c) without affecting the transition kernel because $\mathbf{Y}^{\star}_{\text{mis}}$ is not part of the output quantities and is not conditioned upon in subsequent steps. The removal of the redundant marginalized quantities results in PCG I in Figure 1(d).

The marginalized distribution in Step 1 of Figure 1(d) draws μ from

$$\mu \sim \text{Multinomial}\big(1; \{p(\mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})|_{\mu=E_j}, j \in \mathcal{J}\}\big), \tag{3.3}$$

obtained by evaluating the observed posterior distribution at the midpoint of energy bin. Thus, we impose the natural resolution of the data on μ. Because Steps 1 and 2 in Figure 1(d) can be combined into $p(\mathbf{Y}_{\text{mis}}, \mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$, the PCG sampler corresponds to a blocked version of the original Gibbs sampler in Figure 1(a); see Figure 1(e).

In PCG I, sampling μ involves additional evaluation of the posterior distribution which can be computationally demanding because of the large dimensional blurring matrix $\mathbf{M} = \{M_{lj}\}$. To avoid this computational expense, we devise the PCG II sampler. This sampler is derived exactly the same as PCG I except that PCG II marginalizes over only part of the missing data, i.e., $\mathbf{Y}_{\text{mis }2}$; see Figure 2. After trimming the redundant intermediate quantities from the sampler, PCG II is given by Figure 2(d). In this case, the resulting set of conditional distributions are incompatible and cannot be combined. Thus, PCG II does not correspond to a blocked version of the parent Gibbs sampler in Figure 2(a). The marginalized distribution in Step 1 of PCG II draws μ from the multinomial distribution,

$$\mu \sim \text{Multinomial}\big(1; \{p(\mu|\mathbf{Y}_{\text{mis }1}, \boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})|_{\mu=E_j}, j \in \mathcal{J}\}\big), \tag{3.4}$$

and the multinomial probabilities are computed as in (3.3). By conditioning on $\mathbf{Y}_{\text{mis }1}$, however, we avoid accounting for $\mathbf{M}$, and significantly improve the speed of Step 1.

(a) Parent Gibbs sampler

$p(\mathbf{Y}_{\text{mis}}|\psi, \mu, \mathbf{Y}_{\text{obs}})$
$p(\psi|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$
$p(\mu|\mathbf{Y}_{\text{mis}}, \psi, \mathbf{Y}_{\text{obs}})$

(b) Marginalization

$p(\mathbf{Y}_{\text{mis}\,1}, \mathbf{Y}^{\star}_{\text{mis}\,2}|\psi, \mu, \mathbf{Y}_{\text{obs}})$
$p(\psi|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$
$p(\mathbf{Y}_{\text{mis}\,2}, \mu|\mathbf{Y}_{\text{mis}\,1}, \psi, \mathbf{Y}_{\text{obs}})$

(c) Permute

$p(\mathbf{Y}^{\star}_{\text{mis}\,2}, \mu|\mathbf{Y}_{\text{mis}\,1}, \psi, \mathbf{Y}_{\text{obs}})$
$p(\mathbf{Y}_{\text{mis}}|\psi, \mu, \mathbf{Y}_{\text{obs}})$
$p(\psi|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$

(d) Trim

$p(\mu|\mathbf{Y}_{\text{mis}\,1}, \psi, \mathbf{Y}_{\text{obs}})$
$p(\mathbf{Y}_{\text{mis}}|\psi, \mu, \mathbf{Y}_{\text{obs}})$
$p(\psi|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y}_{\text{obs}})$

Figure 2. Deriving PCG II for the multilevel spectral model with a delta function emission line. The resulting sampler in (d) is composed of incompatible distributions and is not a blocked version of the parent Gibbs sampler in (a).

Because PCG I marginalizes to a greater degree, we expect it to exhibit better convergence characteristics than PCG II; see theorem 1 of van Dyk and Park (2008). This is weighed against the extra computation and cost per iteration of (3.3) relative to (3.4): Each iteration of PCG I is about 10 times slower in computation time than that of PCG II. To empirically compare the computational performance, we use data for the X-ray spectrum of the high redshift quasar PG 1634+706 (Park, van Dyk, and Siemiginowska 2008). We run multiple chains of 10,000 and 100,000 iterations each with overdispersed starting values for PCG I and PCG II, respectively. The first two rows of Figure 3 compare the

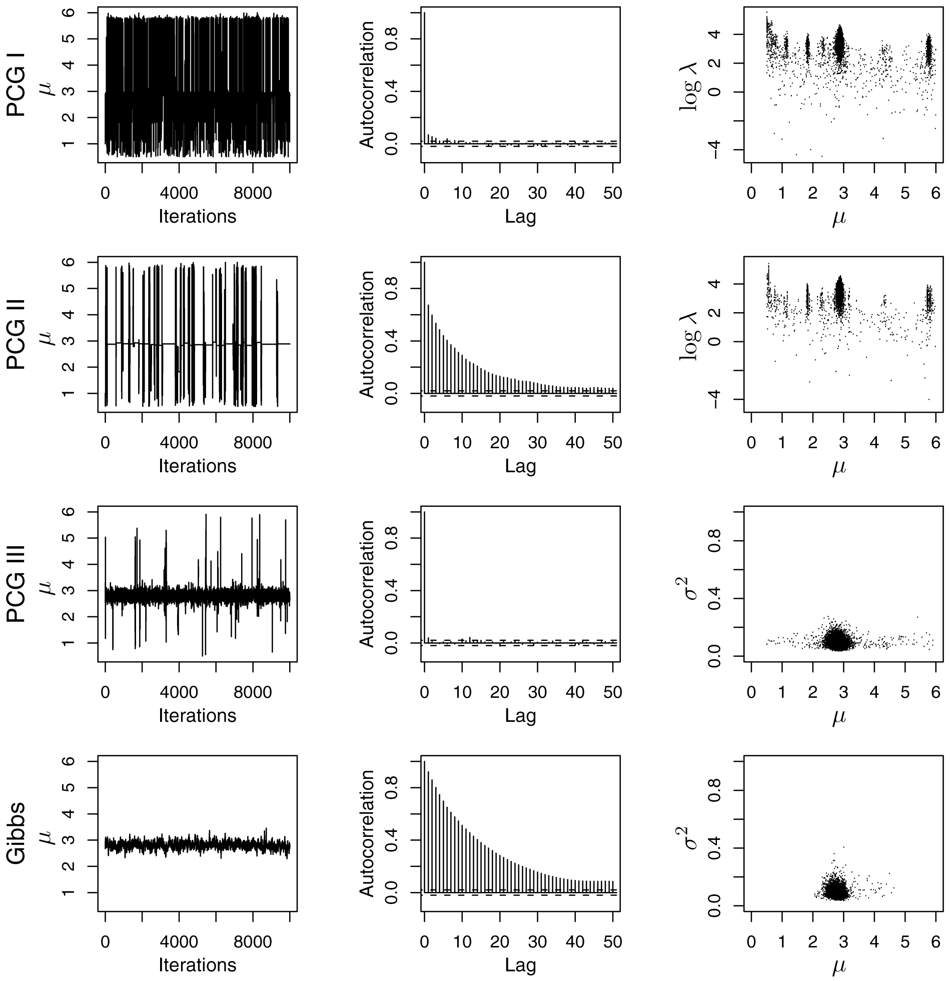

Figure 3. Comparing different samplers for spectral analysis. The top two rows show output from the PCG samplers, PCG I and PCG II, constructed for the spectral model with a delta function emission line. In this case, the parent Gibbs sampler does not converge, so that we omit its comparison. The bottom two rows compare the PCG and Gibbs samplers constructed for the spectral model with a Gaussian emission line. The first two columns show the mixing and autocorrelation plots for μ that are used to compare the convergence of different schemes. The last column presents a joint posterior distribution of μ and $\log\lambda$ for a delta function emission line and a joint posterior distribution of μ and σ^2 for a Gaussian emission line.

two PCG samplers that fit a delta function emission line in terms of their mixing, autocorrelation, and posterior distribution. After we detect convergence using the $\hat{R}^{1/2}$ statistic (Gelman and Rubin 1992), we collect the 10,000 posterior draws of μ and λ from the multiple chains to produce the joint posterior distribution given in the last column of Figure 3. As expected, PCG I exhibits quicker convergence than PCG II, but neither of the PCG samplers has difficulty jumping among the multiple modes.

3.2 Gaussian Emission Lines

Depending on the resolution of an X-ray detector and the true width of an emission line, the emission line may occupy more than one energy bin in the observed spectrum. To model an emission line with appreciable width, a Gaussian distribution can be used in place of a delta function. The Gibbs sampler constructed for fitting the Gaussian emission line profile may exhibit slow convergence because of the multimodal posterior distribution of the line location (Park, van Dyk, and Siemiginowska 2008). Thus, we consider marginalizing over all of the missing data when we sample the Gaussian line location and width parameters. This strategy results in a PCG sampler that we call PCG III and its derivation is presented in Figure 4.

As shown in Figure 4(a), we first construct a parent Gibbs sampler to sample from the target distribution $p(\mathbf{Y}_{\text{mis}}, \boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$, where $\boldsymbol{\theta}$ is now partitioned into $\boldsymbol{\theta} = (\boldsymbol{\psi}, \mu, \sigma^2)$, with μ being the location of the Gaussian line, σ^2 being the variance of the Gaussian line, and $\boldsymbol{\psi}$ being the remaining model parameters. The convergence of the parent Gibbs sampler can be improved by sampling the missing data along with μ and σ^2 in Steps 3 and 4, respectively, as shown in Figure 4(b). To make the multiply sampled quantities redundant in the sampler, we permute the steps in Figure 4(c) and remove the redundant quantities in Figure 4(d). Finally, Steps 2 and 3 are blocked, yielding the PCG sampler in Figure 4(e).

(a) Parent Gibbs sampler

$$p(\mathbf{Y}_{\text{mis}}|\boldsymbol{\psi}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\mu|\mathbf{Y}_{\text{mis}}, \boldsymbol{\psi}, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\sigma^2|\mathbf{Y}_{\text{mis}}, \boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$

(b) Marginalization

$$p(\mathbf{Y}^{\star}_{\text{mis}}|\boldsymbol{\psi}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}^{\star}_{\text{mis}}, \mu|\boldsymbol{\psi}, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}, \sigma^2|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$

(c) Permute

$$p(\mathbf{Y}^{\star}_{\text{mis}}, \mu|\boldsymbol{\psi}, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}^{\star}_{\text{mis}}, \sigma^2|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}|\boldsymbol{\psi}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$

(d) Trim

$$p(\mu|\boldsymbol{\psi}, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\sigma^2|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}|\boldsymbol{\psi}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$

(e) Block

$$p(\mu|\boldsymbol{\psi}, \sigma^2, \mathbf{Y}_{\text{obs}})$$
$$p(\mathbf{Y}_{\text{mis}}, \sigma^2|\boldsymbol{\psi}, \mu, \mathbf{Y}_{\text{obs}})$$
$$p(\boldsymbol{\psi}|\mathbf{Y}_{\text{mis}}, \mu, \sigma^2, \mathbf{Y}_{\text{obs}})$$

Figure 4. Illustration of deriving PCG III for the multilevel spectral model with a Gaussian emission line. The resulting sampler in (e) is composed of incompatible distributions and is not a blocked version of the parent Gibbs sampler in (a).

This sampler is composed of incompatible conditional distributions and is not a blocked version of the parent Gibbs sampler in Figure 4(a).

To implement the marginalization in Step 2 of Figure 4(d), the parameter space for σ^2 is treated as discrete. That is, σ^2 is taken to follow a multinomial distribution with probabilities evaluated on an equally spaced grid, in the same manner as with μ in (3.3).

The bottom two rows of Figure 3 compare the PCG III and parent Gibbs sampler constructed for the spectral model with a Gaussian emission line. These two samplers are run to fit the X-ray spectrum of quasar PG 1634+706, each with 10,000 and 100,000 iterations, respectively. Comparing the first two columns indicates that the PCG sampler has much quicker convergence than the Gibbs sampler, showing faster mixing and less autocorrelations. After ensuring convergence using multiple chains, we plot the joint posterior distribution with the 10,000 posterior draws of μ and σ^2 collected from the multiple chains. As confirmed in the last column of Figure 3, the Gibbs sampler requires more iterations to explore the entire surface of the joint distribution as fully as with PCG III.

4. JOINT SEGMENTATION OF MULTIVARIATE TIME SERIES DATA

The MCMC sampler provided by Dobigeon, Tourneret, and Scargle (2007) to fit the so-called joint segmentation model for Poisson time series data from multiple signals in astrophysics is an example of a PCG sampler. The model is designed to detect and characterize structure in two or more related Poisson time series. For the joint segmentation problem, Dobigeon, Tourneret, and Scargle (2007) constructed a Bayesian hierarchical model and proposed a Gibbs sampling strategy for model fitting. The MCMC sampler used to jointly segment the time series data from multiple signals is, however, a PCG sampler, not a Gibbs sampler per se. In fact, as we describe below, the parent Gibbs sampler for this model is degenerate and is not feasible to run. This is why Dobigeon, Tourneret, and Scargle (2007) devised an MCMC sampler that is Gibbs-like but is constructed with incompatible conditional distributions. They did not recognize the sampler as a PCG sampler or confirm that its stationary distribution is the target posterior distribution.

Consider multivariate time series data that are composed of photon counts from multiple signals observed in a number of equally spaced time bins. We assume that the data for each signal are generated from a Poisson process with constant intensity within each time block. The time blocks in turn are constructed by sequentially combining the time bins, as illustrated in Figure 5 and detailed below. This joint segmentation model has advantages over a segmentation model for a single signal because we can impose prior knowledge as to the correlation structure of the joint probability for the time block changes of different signals.

To formalize the model, we assume photons from each signal arrive following an independent inhomogeneous Poisson process, that is,

$$Y_{st} \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_{sb}) \qquad \text{for } s = 1, 2, \ldots, S, t = 1, 2, \ldots, T, b = 1, 2, \ldots, B_s, \tag{4.1}$$

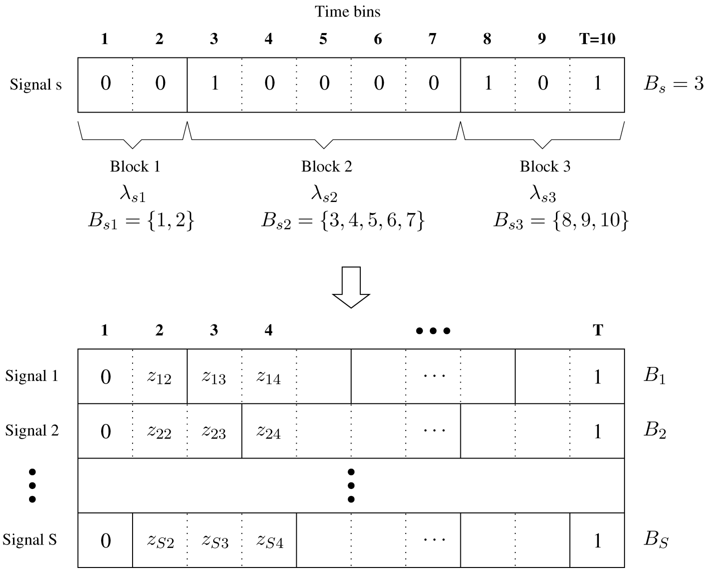

Figure 5. Block indicator matrix for multivariate time series data. The block indicator vector of signal s with $T = 10$ time bins and $B_s = 3$ time blocks is illustrated at the top. In the case of multiple signals, a $S \times T$ block indicator matrix is considered, as shown at the bottom. When the time bins are blocked by the solid lines, for example, we have $(z_{12}, z_{13}, z_{14}) = (0, 1, 0)$, $(z_{22}, z_{23}, z_{24}) = (0, 0, 1)$, and $(z_{S2}, z_{S3}, z_{S4}) = (1, 0, 0)$. Note that the indicator variables for the first and the last time bins are all fixed at 0 and 1, respectively, to match the row sum with the number of time blocks of the corresponding row; thus, only the middle $T - 2$ indicators are free for each signal.

where Y_{st} denotes the photon counts from signal s that fall into time bin t and λ_{sb} represents the expected photon counts per bin from signal s in time block b. In words, photons from each of the S signals are recorded in T equally spaced time bins which are grouped into B_s $(\leq T)$ time blocks for signal s. The Poisson intensities are modeled hierarchically, that is,

$$\lambda_{sb}|\gamma \overset{\text{iid}}{\sim} \Gamma(1, \gamma) \quad \text{for } s = 1, 2, \dots, S, b = 1, 2, \dots, B_s, \tag{4.2}$$

where γ is the rate parameter, that is, $\lambda \sim \Gamma(\alpha, \beta)$ if $p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$. For notational convenience, we consider an index set of time bins that are combined into time block b of signal s, that is, $\mathcal{B}_{sb} \subset \{1, 2, \dots, T\}$ is the collection of time bin indexes in block b for signal s. The sets $\{\mathcal{B}_{sb}, b = 1, 2, \dots, B_s\}$ for signal s are disjoint with union $\{1, 2, \dots, T\}$.

Fitting the Poisson intensity would be greatly simplified if the time blocks, $\mathcal{B} = \{\mathcal{B}_{sb}, s = 1, 2, \dots, S, b = 1, 2, \dots, B_s\}$, were known. This leads naturally to a Gibbs-type setup. In particular, we consider an $S \times T$ indicator matrix of block change points, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T)$ where $\mathbf{z}_t = (z_{1t}, z_{2t}, \dots, z_{St})^\top$ for each t, and $\mathbf{Z}$ is treated as missing data. The variable $z_{st} = 1$ if there is a change point to a new time block at bin t for signal s, and 0 otherwise. We set the indicator variables of the first and the last bins as $z_{s1} = 0$ and $z_{sT} = 1$ for each signal s so that each row sum corresponds to the number of time blocks for the signal s, that is, $\sum_{t=1}^{T} z_{st} = B_s$. Figure 5 graphically illustrates how the indicator variables $\mathbf{Z}$ determine the time blocks. As a simple illustration, the block indicator vector of signal s with $T = 10$ time bins and $B_s = 3$ time blocks is shown at the top of Figure 5. The bottom of Figure 5 illustrates the $S \times T$ matrix $\mathbf{Z}$. Because z_{st} is a binary variable, each column vector $\mathbf{z}_t$ has 2^S possible configurations of 0 and 1, which are denoted by $\mathbf{c}_l$ for $l = 1, 2, \ldots, 2^S$. For example, in the case of $S = 2$, we have $\mathbf{c}_1 = (0\ 0)^\top$, $\mathbf{c}_2 = (1\ 0)^\top$, $\mathbf{c}_3 = (0\ 1)^\top$, and $\mathbf{c}_4 = (1\ 1)^\top$. The probability of $\mathbf{z}_t$ corresponding to each configuration $\mathbf{c}_l$ is represented by $p_l = P(\mathbf{z}_t = \mathbf{c}_l)$, and $\mathbf{P}$ is the set of the probabilities, that is, $\mathbf{P} = \{p_l, l = 1, 2, \ldots, 2^S\}$ with $\sum_{l=1}^{2^S} p_l = 1$.

Under the conjugate Dirichlet prior distribution on $\mathbf{P}$ and a flat prior distribution on $\log \gamma$, the posterior distribution of the unknown quantities is given by $p(\mathbf{Z}, \mathbf{P}, \gamma, \boldsymbol{\lambda} | \mathbf{Y}_{\text{obs}})$. Because of the conjugate choice of its prior distribution, $\mathbf{P}$ can be analytically integrated out of the posterior distribution. Thus, we begin with a parent Gibbs sampler with target distribution equal to the marginal distribution, $p(\mathbf{Z}, \gamma, \boldsymbol{\lambda} | \mathbf{Y}_{\text{obs}}) = \int p(\mathbf{Z}, \mathbf{P}, \gamma, \boldsymbol{\lambda} | \mathbf{Y}_{\text{obs}})\, d\mathbf{P}$. The parent Gibbs sampler is illustrated in Figure 6(a). Unfortunately, this Gibbs sampler cannot be implemented because $(\mathbf{Z}, \boldsymbol{\lambda})$ belongs to a space, $\{0, 1\}^{S \times T} \times \prod_{s=1}^{S} \Re_+^{B_s}$, whose dimension depends on the unknown time blocks $\{B_s, s = 1, 2, \ldots, S\}$: the distribution $p(\mathbf{z}_t | \mathbf{z}_{-t}, \gamma, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})$ is degenerate because the dimensionality of $\boldsymbol{\lambda}$ is specified by the number of time blocks in each time series, where $\mathbf{z}_{-t} = (\mathbf{z}_1, \ldots, \mathbf{z}_{t-1}, \mathbf{z}_{t+1}, \ldots, \mathbf{z}_T)$. To resolve this problem, we devise a PCG sampler that replaces $p(\mathbf{z}_t | \mathbf{z}_{-t}, \gamma, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})$ with $p(\mathbf{z}_t | \mathbf{z}_{-t}, \gamma, \mathbf{Y}_{\text{obs}})$ that integrates out the parameter vector whose dimension depends on $\mathbf{Z}$. Figure 6 illustrates the transformation of the parent Gibbs sampler into the PCG sampler. Starting with the parent Gibbs sampler in Figure 6(a), in Steps 1 through $(T-2)$ each $\mathbf{z}_t$ is sampled given $(\mathbf{z}_{-t}, \gamma, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})$ and follows a degenerate multinomial distribution. In Figure 6(b), we avoid these degenerate draws by jointly sampling $\boldsymbol{\lambda}$ in Steps 1 through $(T-2)$ along with a component of $\mathbf{Z}$. Because $p(\gamma | \mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})$ is conditional on the intermediate

(a) Parent Gibbs sampler

$$\begin{array}{l}
p(\mathbf{z}_2 | \mathbf{z}_{-2}, \gamma, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}}) \\
p(\mathbf{z}_3 | \mathbf{z}_{-3}, \gamma, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}}) \\
\qquad\vdots \\
p(\mathbf{z}_{T-1} | \mathbf{z}_{-(T-1)}, \gamma, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}}) \\
p(\gamma | \mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}}) \\
p(\boldsymbol{\lambda} | \mathbf{Z}, \gamma, \mathbf{Y}_{\text{obs}})
\end{array}$$

(b) Marginalization

$$\begin{array}{l}
p(\mathbf{z}_2, \boldsymbol{\lambda}^\star | \mathbf{z}_{-2}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\mathbf{z}_3, \boldsymbol{\lambda}^\star | \mathbf{z}_{-3}, \gamma, \mathbf{Y}_{\text{obs}}) \\
\qquad\vdots \\
p(\mathbf{z}_{T-1}, \boldsymbol{\lambda}^\star | \mathbf{z}_{-(T-1)}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\gamma | \mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}}) \\
p(\boldsymbol{\lambda} | \mathbf{Z}, \gamma, \mathbf{Y}_{\text{obs}})
\end{array}$$

(c) Permute

$$\begin{array}{l}
p(\mathbf{z}_2, \boldsymbol{\lambda}^\star | \mathbf{z}_{-2}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\mathbf{z}_3, \boldsymbol{\lambda}^\star | \mathbf{z}_{-3}, \gamma, \mathbf{Y}_{\text{obs}}) \\
\qquad\vdots \\
p(\mathbf{z}_{T-1}, \boldsymbol{\lambda}^\star | \mathbf{z}_{-(T-1)}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\boldsymbol{\lambda} | \mathbf{Z}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\gamma | \mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})
\end{array}$$

(d) Trim

$$\begin{array}{l}
p(\mathbf{z}_2 | \mathbf{z}_{-2}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\mathbf{z}_3 | \mathbf{z}_{-3}, \gamma, \mathbf{Y}_{\text{obs}}) \\
\qquad\vdots \\
p(\mathbf{z}_{T-1} | \mathbf{z}_{-(T-1)}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\boldsymbol{\lambda} | \mathbf{Z}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\gamma | \mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})
\end{array}$$

(e) Block

$$\begin{array}{l}
p(\mathbf{z}_2 | \mathbf{z}_{-2}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\mathbf{z}_3 | \mathbf{z}_{-3}, \gamma, \mathbf{Y}_{\text{obs}}) \\
\qquad\vdots \\
p(\mathbf{z}_{T-2} | \mathbf{z}_{-(T-2)}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\mathbf{z}_{T-1}, \boldsymbol{\lambda} | \mathbf{z}_{-(T-1)}, \gamma, \mathbf{Y}_{\text{obs}}) \\
p(\gamma | \mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})
\end{array}$$

Figure 6. Deriving a PCG sampler in the joint segmentation model for multivariate time series data.

quantity $\boldsymbol{\lambda}^\star$ sampled in Step $(T-2)$, we cannot completely remove the additionally sampled quantities from the sampler. To make the extra draws of $\boldsymbol{\lambda}$ redundant, $p(\gamma|\mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}_{\text{obs}})$ and $p(\boldsymbol{\lambda}|\mathbf{Z}, \gamma, \mathbf{Y}_{\text{obs}})$ are interchanged in the sampler shown in Figure 6(c). After the permutation, trimming the intermediate quantities yields the PCG sampler in Figure 6(d). The conditional distribution of $\mathbf{z}_t$ given $(\mathbf{z}_{-t}, \gamma, \mathbf{Y}_{\text{obs}})$ is still a multinomial distribution which depends only on γ and belongs to a space with a fixed dimension. As a result, the PCG sampler is feasible to implement with the expectation of faster convergence whereas the parent Gibbs sampler in Figure 6(a) breaks down due to the degenerate draws. Finally, we may block Steps $(T-2)$ and $(T-1)$ into $p(\mathbf{z}_{T-1}, \boldsymbol{\lambda}|\mathbf{z}_{-(T-1)}, \gamma, \mathbf{Y}_{\text{obs}})$, but the resulting set of conditional distributions in Figure 6(e) remains incompatible and the resulting MCMC sampler is not a Gibbs sampler per se. Dobigeon, Tourneret, and Scargle (2007) used the PCG sampler shown in Figure 6(e). They, however, did not recognize that the sampler is not a Gibbs sampler or that permuting the order of its steps would alter its stationary distribution.

We conduct a simulation study for the piecewise-constant multivariate time series model. In this simulation study, we assume $S = 2$ signals are observed, each with $T = 100$ time bins. In the interior $T - 2 = 98$ time bins, we set three block change points ($t = 20$, 50, and 80) for the first signal and one block change point ($t = 50$) for the second signal: $z_{st} = 1$ for $(s, t) \in \{(1, 20), (1, 50), (1, 80), (1, 100), (2, 50), (2, 100)\}$, and 0 otherwise. Thus, we have $B_1 = 4$ time blocks for the first signal and $B_2 = 2$ time blocks for the second signal. Using $\gamma = 0.1$, we simulate the Poisson intensities for the corresponding time blocks, that is, $(\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14}) = (3.9, 19.3, 7.3, 11.7)$ for the first signal and $(\lambda_{21}, \lambda_{22}) = (10.8, 15.9)$ for the second signal, and generate test data under the model in (4.1). The top two panels in Figure 7 show the test data for the two signals along with the true Poisson intensities used to simulate the data. Because the parent Gibbs sampler in Figure 6(a) is degenerate, the bottom four rows of Figure 7 present output only from the PCG sampler. The convergence of the PCG sampler is examined by computing the $\hat{R}^{1/2}$ statistic for the model parameters with the two chains of 1000 iterations. The output is based on 1000 draws from the second halves of the two chains. The two panels in the second row of Figure 7 present the posterior mean of block change points for the two signals and confirm that the true block change points are well estimated by the PCG sampler. Mixing and autocorrelation plots are shown in the bottom three rows of Figure 7, which illustrate the quick convergence of selected model parameters (γ, λ_{11}, and λ_{21}). The horizontal lines in the mixing plots represent the true values of the parameters, which are well covered by the posterior draws.

This simulation study is done using R. The R code for implementing the PCG sampler used to fit the joint segmentation model for multivariate Poisson time series data is available in this article's supplemental materials folder on the *JCGS* webpage.

5. JOINT IMPUTATION MODEL FOR NONNESTED DATA

Our last example deals with a joint imputation model used to create imputations for a multivariate response variable that is observed on *misaligned* partitions. For the joint imputation model, we construct a PCG sampler by taking advantage of the misaligned data

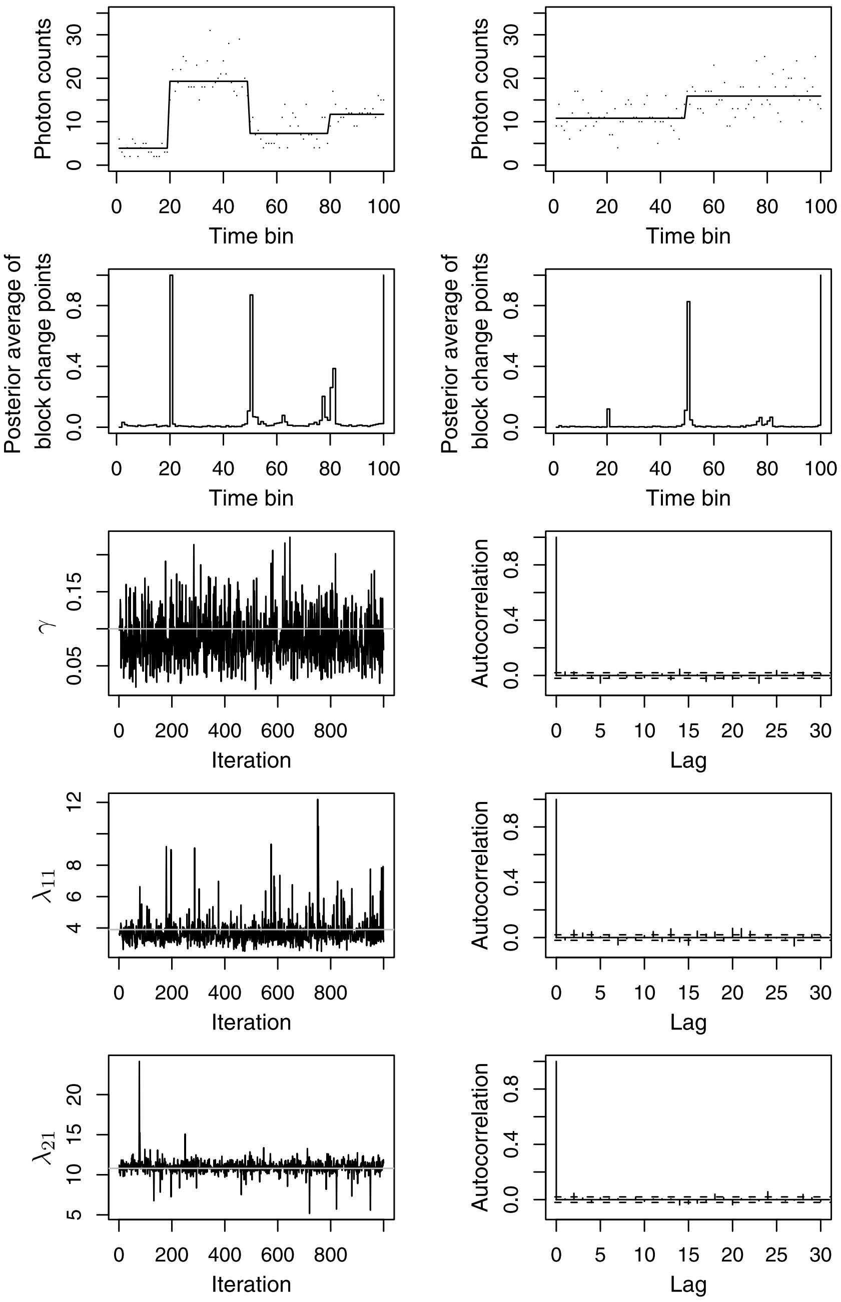

Figure 7. Simulated data and output from the PCG sampler in the multivariate time series model. The top two panels show the observed counts of the two signals with a representation of the expected counts (solid lines). The panels in the second row show the posterior mean of the block indicator matrix **Z** for two signals. The bottom three rows show the mixing and autocorrelation plots of selected model parameters (γ, λ_{11}, and λ_{21}), which illustrate the fast convergence of the PCG sampler. The solid horizontal lines in the mixing plots represent the true values of the model parameters used to simulate the data.

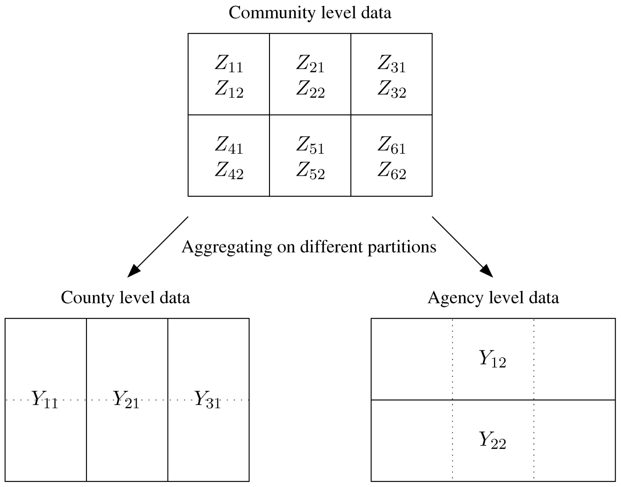

Figure 8. Illustration of a misaligned data structure in a hypothetical state. The highest resolution partition (i.e., the set of communities) is combined into lower resolution partitions (i.e., the sets of counties and agencies) that are *not* aligned. Each component of a bivariate response variable associated with the highest resolution partition is aggregated to one of the two lower resolution partitions. Only the county and agency level data are observed; the community level data are missing. For example, when a bivariate response variable on $n = 6$ communities is given by $\{(Z_{i1}, Z_{i2}), i = 1, 2, \ldots, 6\}$, the first component is aggregated to $J = 3$ counties ($Y_{11} = Z_{11} + Z_{41}$, $Y_{21} = Z_{21} + Z_{51}$, and $Y_{31} = Z_{31} + Z_{61}$) and the second component is aggregated to $K = 2$ agencies ($Y_{12} = Z_{12} + Z_{22} + Z_{32}$ and $Y_{22} = Z_{42} + Z_{52} + Z_{62}$). The joint imputation model creates imputations for the response variable that is missing on the lower resolution partitions, that is, $\widetilde{Y}_{11} = Z_{11} + Z_{21} + Z_{31}$, $\widetilde{Y}_{21} = Z_{41} + Z_{51} + Z_{61}$, $\widetilde{Y}_{12} = Z_{12} + Z_{42}$, $\widetilde{Y}_{22} = Z_{22} + Z_{52}$, and $\widetilde{Y}_{32} = Z_{32} + Z_{62}$.

structure and reparameterization. A fixed geographical region is often divided into several different levels of political partitions. In the United States, for example, the country is sequentially divided into states, into counties, and partially into cities. In this case, one of the partitions generally contains or is completely nested within others. Unlike the United States, however, Germany is composed of different nonnested geographic partitions, that is, states, counties, agencies, and communities. The set of communities is the highest resolution partition, the sets of counties and agencies are lower resolution partitions each of which consists of several communities, and each state is divided into both counties and agencies. The two partitions between states and communities do not generally nest one within the other; see Figure 8. In this example, we consider such misaligned partitions of a certain state. A multivariate response variable is associated with each community but observed only at the level of either the counties or the agencies. That is, only certain partial sums of each component of the multivariate response variable are available on one of the two lower resolution partitions. Here we assume additivity of the data because they are generally cumulative counts. A difficulty arises when some components of the response variable are observed on a certain lower resolution partition (e.g., the counties), and the other components are observed on a different lower resolution partition (e.g., the agencies), which neither contains nor is nested within the first partition. Given the observed data, our joint imputation method aims to create joint imputations for the multivariate response variable that is missing on the lower resolution partitions, properly accounting for its correlation structure. Figure 8 illustrates simplified misaligned partitions where the communities are grouped into either counties or agencies, but the two lower resolution partitions are not nested within each other; details are given below.

For clarity, we consider a simplified joint imputation model as follows. At the community level, we have a bivariate response variable that is denoted by $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ where $\mathbf{Z}_1 = \{Z_{i1}, i = 1, 2, \ldots, n\}$ and $\mathbf{Z}_2 = \{Z_{i2}, i = 1, 2, \ldots, n\}$ with n being the number of communities in a particular state. There is also a set of covariates $\mathbf{X}$ that is fully observed at the community level. Due to the additivity assumption, a variable observed at the community level can be aggregated to recover the corresponding variable at both the county and agency levels. Thus, the covariates are available on all of the levels of the partition. To simplify modeling of the correlation structure, we assume that the community level data follow a bivariate Gaussian distribution, so that our complete-data model is given by

$$\begin{pmatrix} Z_{i1} \\ Z_{i2} \end{pmatrix} \overset{\text{ind}}{\sim} \mathrm{N}_2 \left(\begin{pmatrix} \mathbf{X}_i^\top \boldsymbol{\beta}_1 \\ \mathbf{X}_i^\top \boldsymbol{\beta}_2 \end{pmatrix}, \boldsymbol{\Sigma} \right) \quad \text{for } i = 1, 2, \ldots, n, \tag{5.1}$$

where $\mathbf{X}_i$ is a $p \times 1$ vector of known covariates in community i, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p \times 1$ vectors of coefficients, and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{smallmatrix}\right)$ denotes the 2×2 covariance matrix with σ_1^2 and σ_2^2 being residual variances for Z_{i1} and Z_{i2}, respectively, and σ_{12} being the residual covariance between Z_{i1} and Z_{i2}. The model in (5.1) allows us to explicitly account for the correlation structure of the bivariate response variable and easily can be generalized for a multivariate response variable. The set of communities is partitioned into J disjoint counties and into K disjoint agencies. Let $\mathcal{J}_j$ be the set of indices of the communities that are nested within county j, for $j = 1, 2, \ldots, J$. Then, we have $\mathcal{J}_j \subset \{1, 2, \ldots, n\}$ such that $\bigcup_{j=1}^J \mathcal{J}_j = \{1, 2, \ldots, n\}$ and $\mathcal{J}_j \cap \mathcal{J}_k = \emptyset$ for $j \neq k$, and the cardinality of $\mathcal{J}_j$ is $|\mathcal{J}_j| = n_j$ with $n = \sum_{j=1}^J n_j$. Likewise, we define by $\mathcal{K}_k$ the set of indices of the communities that are nested within agency k for $k = 1, 2, \ldots, K$, so that $|\mathcal{K}_k| = m_k$ and $n = \sum_{k=1}^K m_k$. Here, we assume that the first component of the bivariate response variable is observed only on the counties and the second component only on the agencies. That is, let $\mathbf{Y}_1 = \{Y_{j1}, j = 1, 2, \ldots, J\}$ denote the first component of the bivariate response variable observed on J counties and $\mathbf{Y}_2 = \{Y_{k2}, k = 1, 2, \ldots, K\}$ denote the second component observed on K agencies. Because of the additivity of the variables, Y_{j1} consists of the sum of the n_j values of Z_{i1} for county j, that is, $Y_{j1} = \sum_{i \in \mathcal{J}_j} Z_{i1}$ for $j = 1, 2, \ldots, J$. Likewise, Y_{k2} consists of the sum of the m_k values of Z_{i2} for agency k, that is, $Y_{k2} = \sum_{i \in \mathcal{K}_k} Z_{i2}$ for $k = 1, 2, \ldots, K$. Thus, the observed data consist of $\mathbf{Y}_{\text{obs}} = (\mathbf{Y}_1, \mathbf{Y}_2)$, and the missing data are denoted by $\widetilde{\mathbf{Y}}_1 = \{\widetilde{Y}_{k1}, k = 1, 2, \ldots, K\}$ and $\widetilde{\mathbf{Y}}_2 = \{\widetilde{Y}_{j2}, j = 1, 2, \ldots, J\}$, where $\widetilde{Y}_{k1} = \sum_{i \in \mathcal{K}_k} Z_{i1}$ and $\widetilde{Y}_{j2} = \sum_{i \in \mathcal{J}_j} Z_{i2}$. Once we impute the bivariate response variable $\mathbf{Z}$, the missing response variable on the lower resolution partitions, $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1, \widetilde{\mathbf{Y}}_2)$, can be recovered by aggregating the imputed community-level variable $\mathbf{Z}_{\text{imp}} = (\mathbf{Z}_{\text{imp }1}, \mathbf{Z}_{\text{imp }2})$.

Figure 8 presents a simplified example where $n = 6$ communities are combined into either $J = 3$ counties or $K = 2$ agencies, and both lower resolution partitions are misaligned. In this example, the observed data consist of the first component aggregated to the counties and the second component aggregated to the agencies: Y_{11}, Y_{21}, Y_{31}, Y_{12}, and Y_{22}. On the other hand, the first component aggregated to the agencies and the second component

(a) Parent Gibbs sampler

$$\begin{array}{l}
p(\mathbf{Z}_1|\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_2|\mathbf{Z}_1,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\beta}_1|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\beta}_2|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\Sigma}|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{Y}_{\mathrm{obs}})
\end{array}$$

(b) Marginalization

$$\begin{array}{l}
p(\mathbf{Z}_1^{\star},\mathbf{Z}_2^{\star}|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_2^{\star}|\mathbf{Z}_1,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_1^{\star},\mathbf{Z}_2^{\star},\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_2|\boldsymbol{\beta}_1,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\Sigma}|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{Y}_{\mathrm{obs}})
\end{array}$$

(c) Permute

$$\begin{array}{l}
p(\mathbf{Z}_1^{\star},\mathbf{Z}_2^{\star},\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_1^{\star},\mathbf{Z}_2^{\star},\boldsymbol{\beta}_2|\boldsymbol{\beta}_1,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_1,\mathbf{Z}_2^{\star}|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_2|\mathbf{Z}_1,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\Sigma}|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{Y}_{\mathrm{obs}})
\end{array}$$

(d) Trim

$$\begin{array}{l}
p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_1|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_2|\mathbf{Z}_1,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\Sigma}|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{Y}_{\mathrm{obs}})
\end{array}$$

(e) Block

$$\begin{array}{l}
p(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_2|\boldsymbol{\beta}_1,\boldsymbol{\Sigma},\mathbf{Y}_{\mathrm{obs}})\\
p(\boldsymbol{\Sigma}|\mathbf{Z}_1,\mathbf{Z}_2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{Y}_{\mathrm{obs}})
\end{array}$$

Figure 9. Deriving a PCG sampler in the joint imputation model for nonnested data.

aggregated to the counties are not observed and must be imputed: $\widetilde{Y}_{11} = Z_{11} + Z_{21} + Z_{31}$, $\widetilde{Y}_{21} = Z_{41} + Z_{51} + Z_{61}$, $\widetilde{Y}_{12} = Z_{12} + Z_{42}$, $\widetilde{Y}_{22} = Z_{22} + Z_{52}$, and $\widetilde{Y}_{32} = Z_{32} + Z_{62}$.

In the joint imputation model, the posterior distribution of the unknown quantities is given by $p(\mathbf{Z}_1, \mathbf{Z}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}|\mathbf{Y}_{\mathrm{obs}})$ under the conjugate noninformative prior distribution on $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma})$. We can create multiple imputations for $(\mathbf{Z}_1, \mathbf{Z}_2)$ by running a parent Gibbs sampler constructed using the conditional distributions in Figure 9(a). To improve the convergence of the parent Gibbs sampler, we capitalize on some marginal distributions of the target distribution that are suggested by the misaligned data structure. In Figure 9(b), we sample $\mathbf{Z}_2$ along with $\mathbf{Z}_1$ in Step 1, and $(\mathbf{Z}_1, \mathbf{Z}_2)$ along with $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in Steps 3 and 4, respectively. Permuting the conditional distributions in Figure 9(c) leaves the additionally sampled quantities unused in the subsequent steps of the sampler. Thus, we can safely remove these intermediate draws from the sampler, which yields a PCG sampler in Figure 9(d). Finally, the six steps in Figure 9(d) can be blocked into the three incompatible steps in Figure 9(e). Details are given in the Appendix.

To illustrate the improved convergence of the PCG sampler over the parent Gibbs sampler, a simulation study is conducted as follows. First, we assume $n = 120$ communities in a single state, and we construct the counties and agencies by combining every $n_j = 5$ communities and every $m_k = 8$ communities, respectively. This means that there are $J = 120/5 = 24$ counties and $K = 120/8 = 15$ agencies. Then, under the model in (5.1), we generate test data $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ on $n = 120$ communities using covariates, $\mathbf{X}$, generated from $\mathrm{N}(0, 5^2)$, the regression coefficients, $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11})^{\top} = (1\ 0.5)^{\top}$ and $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12})^{\top} = (1\ 0.2)^{\top}$, and the covariance matrix, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1.8 \\ 1.8 & 4 \end{pmatrix}$. The observed data $\mathbf{Y}_{\mathrm{obs}} = (\mathbf{Y}_1, \mathbf{Y}_2)$ are constructed by taking the partial sums of every five $\mathbf{Z}_1$ variables and every eight $\mathbf{Z}_2$ variables, respectively. After running two chains each with 2000 iterations, the convergence of both the parent Gibbs sampler and PCG sampler is examined by com-

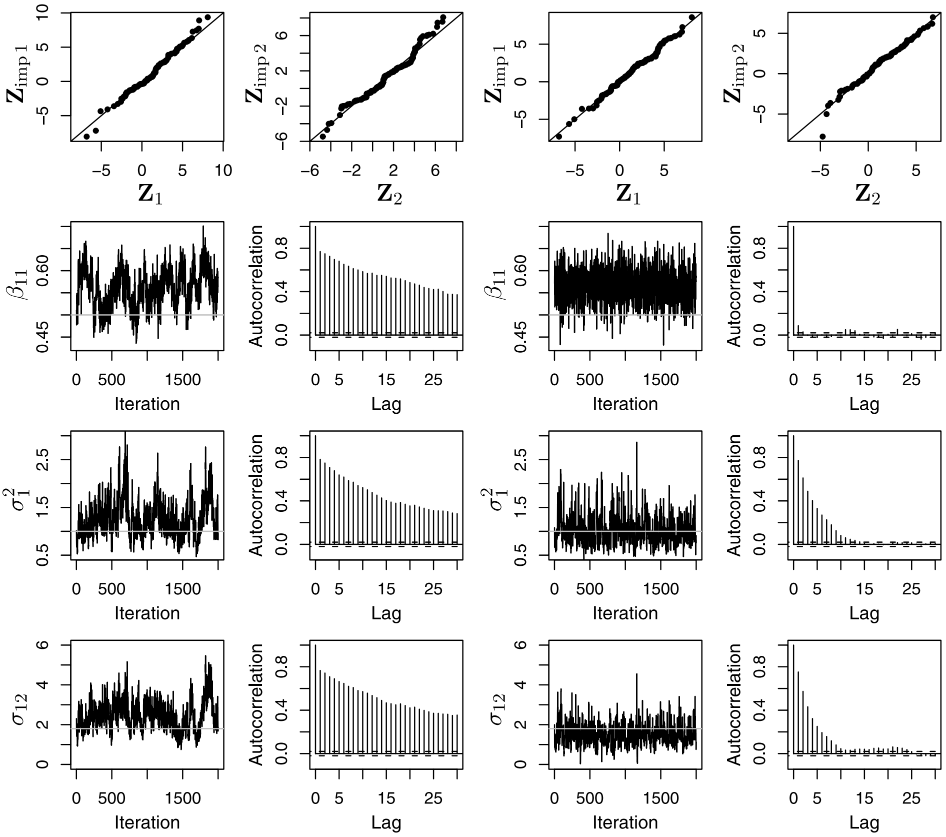

Figure 10. Comparison of the parent Gibbs sampler and PCG sampler devised for the joint imputation model with nonnested data. The first two columns correspond to the parent Gibbs sampler and the last two columns the PCG sampler. The first row shows the quantile–quantile plots for true community-level data $\mathbf{Z}$ and the imputed community-level data $\mathbf{Z}_{\text{imp}}$. The convergence characteristics of selected model parameters (β_{11}, σ_1^2, and σ_{12}) are compared using the mixing and autocorrelation plots. The PCG sampler has much quicker convergence than the parent Gibbs sampler. The solid horizontal lines in the mixing plots represent the true values of the selected model parameters.

puting the $\hat{R}^{1/2}$ statistic for all model parameters. Because all the $\hat{R}^{1/2}$ statistics are less than 1.1, our inference is based on 2000 draws from the second halves of the two chains for each sampler. Figure 10 compares the two samplers in terms of mixing and autocorrelation; the first two columns correspond to the parent Gibbs sampler, and the last two columns the PCG sampler. The top four panels in Figure 10 present the quantile–quantile plots for the test data, $\mathbf{Z}$, and imputed data, $\mathbf{Z}_{\text{imp}}$, under both the parent Gibbs sampler and PCG sampler. Because the data in each of the quantile–quantile plots follow a 45° line, they illustrate that our imputations created by both samplers are very close to the test data. In the bottom three rows of Figure 10, the convergence characteristics are compared between the parent Gibbs sampler and PCG sampler for the selected model parameters (β_{11}, σ_1^2, and σ_{12}). It is evident that the posterior draws of the model parameters obtained by the PCG sampler exhibit faster mixing behavior and lower autocorrelations than those obtained by the parent Gibbs sampler, so that the PCG sampler outperforms the parent Gibbs sampler.

This simulation study is done using R. The R code for implementing the PCG sampler used to create joint imputations for a bivariate response variable that is observed on misaligned partitions is available in this article's supplemental materials folder on the *JCGS* webpage.

6. CONCLUDING REMARKS

In this article, we discuss the relationship between the EM algorithm and the Gibbs sampler with respect to efficient computational strategies, and describe the recently proposed PCG sampler as a stochastic counterpart to the ECME and AECM algorithms. Using different data augmentation schemes in the conditional maximization steps of the ECM algorithm (Meng and Rubin 1993) corresponds to constructing each sampling step of the Gibbs sampler by using different marginal distributions of a target distribution. This may result in conditional distributions that are functionally incompatible. The PCG sample capitalizes on this incompatibility to improve convergence while maintaining the common target stationary distribution.

PCG samplers that are composed of incompatible conditional distributions are illustrated by three examples with substantive computational challenges. In the first example, the Gibbs sampler exhibits very slow convergence and, in the extreme, it does not converges to the target posterior distribution. In this case, the PCG sampler dramatically improves the convergence of its parent Gibbs sampler by marginalizing missing data out of some conditional distributions. The second example illustrates the convenience of the PCG sampler by showing its ability to avoid the need of jumping between spaces of different dimensions. By marginalizing over a certain parameter vector whose dimension is not fixed, the PCG sampler avoids use of reversible jump MCMC algorithm (Green 1995). Although this marginalization results in a set of incompatible conditional distributions, the PCG sampler maintains the target stationary distribution as its parent Gibbs sampler. In the last example, the misaligned data structure makes it easy to find a set of incompatible conditional distributions. Taking advantage of these incompatible distributions, the PCG sampler provides faster convergence than its parent Gibbs sampler.

APPENDIX: IMPLEMENTING THE PCG SAMPLER IN FIGURE 9(d)

In this appendix, we describe the details of the sampling steps of the PCG sampler for fitting our joint imputation model for nonnested data. The marginalized distributions in Steps 1, 2, and 3 of the PCG sampler in Figure 9(d) can be obtained by making use of the misaligned data structure and reparameterization. Due to the additivity of variables, it is easy to obtain the joint distribution of the $(J+K)\times 1$ vector of $(\mathbf{Y}_1,\mathbf{Y}_2)$, which is a multivariate Gaussian distribution given by

$$\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix} \sim \mathrm{N}_{J+K}\left(\begin{pmatrix}\mathbf{X}_1^\top\boldsymbol{\beta}_1\\ \mathbf{X}_2^\top\boldsymbol{\beta}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Upsilon}_{11} & \boldsymbol{\Upsilon}_{12}\\ \boldsymbol{\Upsilon}_{21} & \boldsymbol{\Upsilon}_{22}\end{pmatrix}\right), \tag{A.1}$$

where $\mathbf{X}_1 = \{\sum_{i \in \mathcal{J}_j} \mathbf{X}_i, j = 1, 2, \ldots, J\}$ is a $p \times J$ matrix of known covariates for counties, $\mathbf{X}_2 = \{\sum_{i \in \mathcal{K}_k} \mathbf{X}_i, k = 1, 2, \ldots, K\}$ is a $p \times K$ matrix of known covariates for agencies, and $\binom{\boldsymbol{\Upsilon}_{11} \boldsymbol{\Upsilon}_{12}}{\boldsymbol{\Upsilon}_{21} \boldsymbol{\Upsilon}_{22}}$ is a $(J + K) \times (J + K)$ covariance matrix between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ which is a function of $\boldsymbol{\Sigma}$. Based on the likelihood in (A.1), the conditional distribution in Step 1 of the PCG sampler in Figure 9(d) is

$$\begin{aligned} p(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2, \boldsymbol{\Sigma}, \mathbf{Y}_{\text{obs}}) &\propto p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \psi_2^2, \psi_{12} | \mathbf{Y}_{\text{obs}}) \cdot (\sigma_1^2)^{-2} \\ &= \mathrm{N}_p(\hat{\boldsymbol{\beta}}_1, \sigma^2 [\mathbf{X}_1 \boldsymbol{\Psi}_{11} \mathbf{X}_1^\top]^{-1}), \end{aligned} \tag{A.2}$$

where $\psi_2^2 = \sigma_2^2/\sigma_1^2$, $\psi_{12} = \sigma_{12}/\sigma_1^2$, $(\sigma_1^2)^{-2}$ is the Jacobian of the transformation $(\sigma_2^2, \sigma_{12}) \mapsto (\psi_2^2, \psi_{12})$, $\binom{\boldsymbol{\Psi}_{11} \boldsymbol{\Psi}_{12}}{\boldsymbol{\Psi}_{21} \boldsymbol{\Psi}_{22}} = \sigma^2 \binom{\boldsymbol{\Upsilon}_{11} \boldsymbol{\Upsilon}_{12}}{\boldsymbol{\Upsilon}_{21} \boldsymbol{\Upsilon}_{22}}^{-1}$, and $\hat{\boldsymbol{\beta}}_1 = [\mathbf{X}_1 \boldsymbol{\Psi}_{11} \mathbf{X}_1^\top]^{-1} (\mathbf{X}_1 \boldsymbol{\Psi}_{11} \mathbf{Y}_1 + \mathbf{X}_1 \boldsymbol{\Psi}_{12} [\mathbf{Y}_2 - \mathbf{X}_2^\top \boldsymbol{\beta}_2])$. Due to symmetry, the marginalized distribution in Step 2 of the PCG sampler in Figure 9(d) can be derived using the reparameterization of $\psi_1^2 = \sigma_1^2/\sigma_2^2$ and $\psi_{21} = \sigma_{12}/\sigma_2^2$. In Step 3, we make a county-wise update of $\mathbf{Z}_1$, for example, $\{Z_{i1}, i \in \mathcal{J}_j\}$. When we arbitrarily drop one component Z_{l1} out of $\{Z_{i1}, i \in \mathcal{J}_j\}$, the remaining components along with Y_j jointly follow a multivariate Gaussian distribution. Thus, it is easy to compute $p(\{Z_{i1}, i \in \mathcal{J}_j, i \neq l\} | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}, \mathbf{Y}_{\text{obs}})$ which is also a multivariate Gaussian distribution. We sequentially draw each component of $\{Z_{i1}, i \in \mathcal{J}_j, i \neq l\} = \{Z_{s1}, s = 1, 2, \ldots, n_j - 1\}$ from the corresponding conditional distributions that factorize $p(\{Z_{i1}, i \in \mathcal{J}_j, i \neq l\} | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}, \mathbf{Y}_{\text{obs}})$, and finally we set $Z_{l1} = Y_j - \sum_{i \in \mathcal{J}_j, i \neq l} Z_{i1}$. More specifically, the conditional distribution for the sth component, $p(Z_{s1} | Z_{11}, \ldots, Z_{(s-1)1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}, \mathbf{Y}_{\text{obs}})$, is an extended skew-normal distribution (Azzalini 2005) because $p(Z_{s1}, Z_{s2} | Z_{11}, \ldots, Z_{(s-1)1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}, \mathbf{Y}_{\text{obs}})$ is a bivariate Gaussian distribution and Z_{s2} is truncated above at an upper bound that is implied by $\mathbf{Y}_2$. That is, we draw Z_{s1} from the extended skew-normal distribution,

$$Z_{s1} \sim \mu_s + \nu_s \, \mathrm{SN}(\alpha, \tau_s) \quad \text{for } s = 1, 2, \ldots, n_j - 1, \tag{A.3}$$

where $\mu_s = \mathbf{X}_s^\top \boldsymbol{\beta}_1 + (Y_j - 1_{\{s>1\}} \sum_{i=1}^{s-1} Z_{i1} - \sum_{i \in \mathcal{J}_j} \mathbf{X}_i^\top \boldsymbol{\beta}_1)/(n_j - s + 1)$, $\nu_s = \sigma_1^2 (n_j - s)/(n_j - s + 1)$, $Z \sim \mathrm{SN}(\alpha, \tau)$ if Z follows an extended skew-normal distribution with shape parameter α and upper bound τ, that is, $p(Z) = \phi(-Z) \Phi(\tau \sqrt{1 + \alpha^2} - \alpha Z)/\Phi(\tau)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a standard Gaussian random variable, respectively, $\alpha = \sigma_{12}/\sqrt{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2}$, and τ_s is approximated by the standardized value of $\sum_{i \in \mathcal{J}_j} \hat{Z}_{i2}$, where $\hat{Z}_{i2} = Y_{k2}/m_k$ is the maximum likelihood estimate for Z_{i2} when $i \in \mathcal{K}_k$.

SUPPLEMENTAL MATERIALS

Datasets for the quasar PG 1634+706: Archive containing the data and calibration files used for the spectral analysis in Section 3. (quasardata.tar.gz, GNU zipped tar file)

R code for joint segmentation: R code for the PCG sampler used to fit the piecewise-constant multivariate time series model in Section 4. (pcg1.r, R file)

R code for joint imputation: R code for the PCG sampler used to create joint imputations of a bivariate variable that is observed on misaligned partitions as described in Section 5. (pcg2.r, R file)

ACKNOWLEDGMENTS

The authors thank an anonymous referee and the associate editor for helpful comments and suggestions. They also thank a member of the Editorial Board who served as editor for this article. The authors gratefully acknowledge funding for this project partially provided by NSF Grants DMS-04-38240, DMS-04-06085, and SES-05-50980 and by NASA Contracts NAS8-39073 and NAS8-03060 (CXC).

[Received August 2008. Revised March 2009.]

REFERENCES

Azzalini, A. (2005), "The Skew-Normal Distribution and Related Multivariate Families," *Scandinavian Journal of Statistics*, 32, 159–188.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–37.

Dobigeon, N., Tourneret, J.-Y., and Scargle, J. (2007), "Joint Segmentation of Multivariate Astronomical Time Series: Bayesian Sampling With a Hierarchical Model," *IEEE Transactions on Signal Processing*, 55, 414–423.

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

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

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.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

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

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Applications to Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966.

——— (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

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

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.

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*, Ser. B, 59, 511–567.

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

Morrison, R., and McCammon, D. (1983), "Interstellar Photoelectric Absorption Cross Sections, 0.03–10 keV," *The Astrophysical Journal*, 270, 119–122.

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.

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

——— (2000b), "Nesting EM Algorithms for Computational Efficiency," *Statistica Sinica*, 10, 203–225.

van Dyk, D. A., and Meng, X.-L. (2001), "The Art of Data Augmentation," *The Journal of Computational and Graphical Statistics*, 10, 1–111.

——— (2009), "Cross-Fertilizing Strategies for Better EM Mountain Climbing and DA Field Exploration: A Graphical Guide Book," *Statistical Science*, under review.

van Dyk, D., and Park, T. (2008), "Partially Collapsed Gibbs Samplers: Theory and Methods," *Journal of the American Statistical Association*, 103, 790–796.

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.

Yu, Y. (2005), "Three Contributions to Statistical Computing," Ph.D. thesis, Department of Statistics, Harvard University.