Partially Collapsed Gibbs Samplers: Theory and Methods

David A. van Dyk[1] and Taeyoung Park

Ever increasing computational power along with ever more sophisticated statistical computing techniques is making it possible to fit ever more complex statistical models. Among the popular, computationally intensive methods, the Gibbs sampler (Geman and Geman, 1984) has been spotlighted because of its simplicity and power to effectively generate samples from a high-dimensional probability distribution. Despite its simple implementation and description, however, the Gibbs sampler is criticized for its sometimes slow convergence especially when it is used to fit highly structured complex models. Here, we present partially collapsed Gibbs sampling strategies that improve the convergence by capitalizing on a set of functionally incompatible conditional distributions. Such incompatibility is generally avoided in the construction of a Gibbs sampler because the resulting convergence properties are not well understood. We, however, introduce three basic tools (marginalization, permutation, and trimming) which allow us to transform a Gibbs sampler into a partially collapsed Gibbs sampler with known stationary distribution and faster convergence.

1 Introduction

The *ordinary Gibbs sampler* begins with the joint posterior distribution of a set of unknown quantities and updates groups of these quantities by sampling them from their conditional distributions under the joint posterior distribution. Each quantity is generally updated exactly once in each iteration. The partially collapsed Gibbs (PCG) sampler replaces some of these conditional distributions with conditional distributions under some *marginal distributions* of the joint posterior distribution. This strategy is useful because it can result in samplers with much better convergence properties and it is interesting because it may

[1] The authors gratefully acknowledge funding for this project partially provided by NSF grant DMS-01-04129, DMS-04-38240, and DMS-04-06085 and by NASA Contracts NAS8-39073 and NAS8-03060 (CXC).

result in sampling from a set of *incompatible* conditional distributions. I.e., there may be no joint distribution that corresponds to this set of conditional distributions.

Our technique can be viewed as a generalization of blocking (Liu *et al.*, 1994) in that the resulting conditional distributions can sometimes be combined to form a Gibbs sampler that is a blocked version of the original sampler. In such cases, we can recover a set of compatible conditional distributions and an ordinary Gibbs sampler by combining steps. This is not always possible, however, and some PCG samplers can only be composed of draws from incompatible conditional distributions. In this regard, PCG samplers constitute a generalization of the Gibbs sampler, in that ordinary Gibbs samplers are constructed using the conditional distributions of some joint distribution. Since both blocking and collapsing (Liu *et al.*, 1994) are special cases of PCG, our methods can be viewed as a generalizing and unifying framework for these important and efficient methods. Like blocked and collapsed samplers PCG samplers dominate their parent Gibbs samplers in terms of their convergence and maintain the target posterior distribution as their stationary distribution.

In order to transform a Gibbs sampler into a PCG sampler, we use three basic tools. The first tool is *marginalization* which entails 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 effect the convergence properties of the chain; marginalization can dramatically improve convergence, while 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.

Although our methods are motivated by computational challenges in applied problems

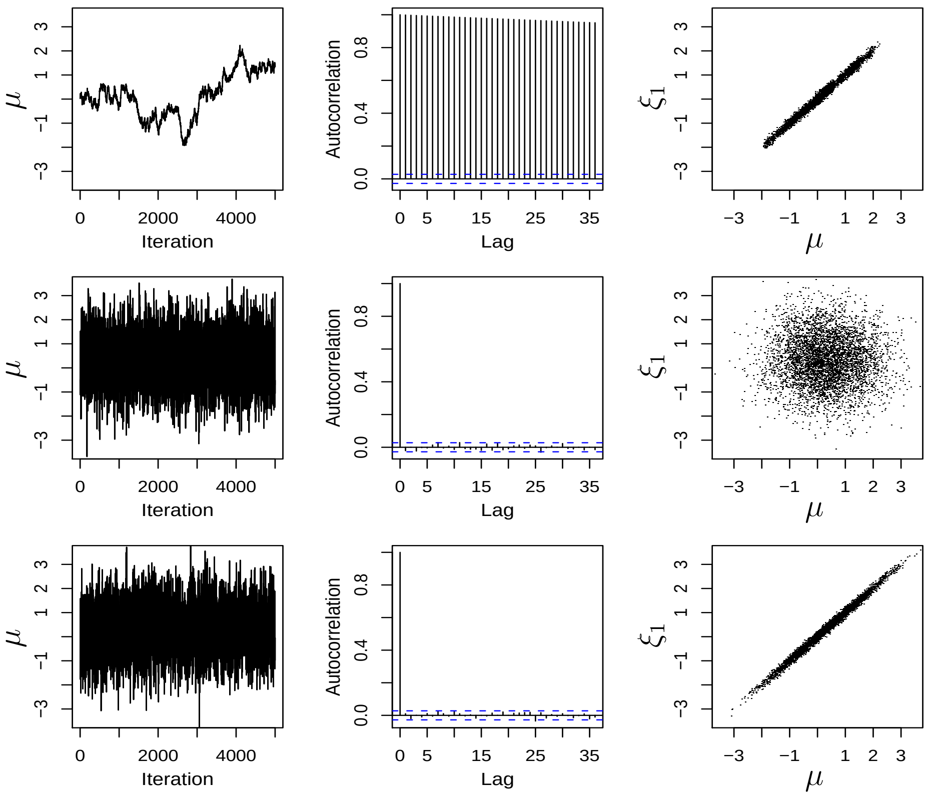

Figure 1: Comparison of Three Samplers for the Simple Random Effects Model. The first two columns show the mixing and autocorrelations of the subchain for μ and the last column the correlation structure between μ and ξ_1. The three rows represent the ordinary Gibbs sampler (Sampler 2.1), the Gibbs sampler resulting from the inappropriate substitution of a reduced conditional distribution (Sampler 2.2), and the PCG sampler (Sampler 2.3).

in astronomy and multiple imputation, to save space we focus here on our methods and their properties. Interested readers will find detailed applications in Park & van Dyk (2008).

2 Motivating Examples

To illustrate PCG samplers, we consider the simple random effects model given by

$$y_{ij} = \xi_i + \varepsilon_{ij} \quad \text{for } i = 1, \ldots, k \text{ and } j = 1, \ldots, n, \tag{1}$$

where $\xi_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^2)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ are independent with y_{ij} being observation j in group i, n the number of units in each group, ξ_i the mean of group i, μ the mean of the group means, τ^2 the between group variance, σ^2 the within group variance, and τ^2 and σ^2 presumed known. From a Bayesian perspective, we are interested in the joint posterior distribution $p(\boldsymbol{\xi}, \mu|\boldsymbol{Y})$ computed under the flat prior distribution $p(\mu) \propto 1$, where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)$ and $\boldsymbol{Y} = \{y_{ij}, i = 1, \ldots, k, j = 1, \ldots, n\}$. To fit this random effects model, we can use a prototype two-step Gibbs sampler that iterates between

Step 1: Draw $\boldsymbol{\xi}^{(t)}$ from $p(\boldsymbol{\xi}|\mu^{(t-1)}, \boldsymbol{Y})$, (Sampler 2.1)

where $\xi_i|\mu^{(t-1)}, \boldsymbol{Y} \overset{\text{ind}}{\sim} \mathrm{N}\left(\dfrac{n\tau^2\bar{Y}_{i\cdot} + \sigma^2\mu^{(t-1)}}{n\tau^2 + \sigma^2}, \dfrac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right)$ for $i = 1, \ldots, k$, and

Step 2: Draw $\mu^{(t)}$ from $p(\mu|\boldsymbol{\xi}^{(t)}, \boldsymbol{Y}) = \mathrm{N}\left(\dfrac{\sum_i \xi_i^{(t)}}{k}, \dfrac{\tau^2}{k}\right)$,

with $\bar{Y}_{i\cdot} = \sum_{j=1}^{n} Y_{ij}/n$ being the mean of the observations in group i. (We emphasize that this is a toy example introduced for illustrative purposes. There is no need for Gibbs sampling when both the between and within group variances are known.) In the case of a two-step Gibbs sampler, the geometric rate of convergence is the same as the lag-one autocorrelation; in Sampler 2.1, the lag-one autocorrelation is the shrinkage parameter, $\sigma^2/(n\tau^2 + \sigma^2)$. To illustrate the potentially slow convergence of Sampler 2.1, we simulate test data with $\mu = 0$, $\sigma = 10$, and $\tau = 0.1$ when there are $n = 10$ observations in each of $k = 10$ groups. The test data are fit by running Sampler 2.1 with 5000 iterations; the resulting lag-one autocorrelation for μ is 0.999. The first row of Figure 1 presents the output of Sampler 2.1 and illustrates the poor mixing and high autocorrelations of the subchain for μ and the strong posterior correlation of μ and ξ_1.

To improve the convergence of a Markov chain constructed with a Gibbs sampler, we may replace a conditional distribution of the original Gibbs sampler with a conditional distribution of a *marginal distribution* of the target distribution. Throughout this article, such a conditional distribution that conditions upon fewer unknown components is referred to as a *reduced conditional distribution*. That is, reduced conditional distributions are conditional distributions of a marginal distribution of the target joint distribution. In the random effects model, we consider the marginal distribution $p(\mu|\boldsymbol{Y}) = \int p(\boldsymbol{\xi}, \mu|\boldsymbol{Y}) d\boldsymbol{\xi}$ of the target distribution $p(\boldsymbol{\xi}, \mu|\boldsymbol{Y})$. We replace Step 2 of Sampler 2.1 with the trivial "conditional" distribution of this marginal distribution. This substitution yields the sampler:

Step 1: Draw $\boldsymbol{\xi}^{(t)}$ from $p(\boldsymbol{\xi}|\mu^{(t-1)}, \boldsymbol{Y})$, and (Sampler 2.2)
Step 2: Draw $\mu^{(t)}$ from $p(\mu|\boldsymbol{Y}) = \mathrm{N}\left(\frac{\sum_i \sum_j y_{ij}}{nk}, \frac{n\tau^2 + \sigma^2}{nk}\right)$.

Step 2 of Sampler 2.2 simulates μ directly from its marginal posterior distribution. The advantage of this strategy is clear: We immediately obtain independent draws of μ from the target posterior distribution. However, the two conditional distributions used in Sampler 2.2, $p(\boldsymbol{\xi}|\mu, \boldsymbol{Y})$ and $p(\mu|\boldsymbol{Y})$, are *incompatible* and imply inconsistent dependence structure. Even in this simple case, the incompatible conditional distributions improve the convergence of the sampler, but at the expense of the correlation structure of the target distribution. Indeed, because $\mu^{(t)}$ is sampled independently of $\boldsymbol{\xi}^{(t)}$, the Markov chain has the stationary distribution $p(\boldsymbol{\xi}|\boldsymbol{Y})p(\mu|\boldsymbol{Y})$ rather than $p(\boldsymbol{\xi}, \mu|\boldsymbol{Y})$. This is illustrated in the second row of Figure 1, where we confirm that the subchain for μ converges immediately to its target marginal distribution, but the correlation structure between μ and ξ_1 (and all of $\boldsymbol{\xi}$) is lost.

There is an obvious solution. Sampler 2.2 first draws $\boldsymbol{\xi}$ from its conditional posterior distribution $p(\boldsymbol{\xi}|\mu, \boldsymbol{Y})$ and then draws μ from its marginal posterior distribution $p(\mu|\boldsymbol{Y})$, rather than vice versa. If we simply exchange the order of the steps, we regain the correlation structure of the target distribution. The resulting Gibbs sampler iterates between

Step 1: Draw $\mu^{(t)}$ from $p(\mu|\boldsymbol{Y})$ and (Sampler 2.3)
Step 2: Draw $\boldsymbol{\xi}^{(t)}$ from $p(\boldsymbol{\xi}|\mu^{(t)}, \boldsymbol{Y})$.

Sampler 2.3 consists of two incompatible conditional distributions and converges quicker than Sampler 2.1, while maintaining the correlations of the target distribution. Of course in this case, the PCG sampler (Sampler 2.3) is simply a blocked version of Sampler 2.1: Steps 1 and 2 combine into a single independent draw from the target distribution. As we shall illustrate, however, PCG samplers can be more general than blocked Gibbs samplers when there are more than two steps. The bottom row of Figure 1 illustrates the fast convergence of the subchain for μ and the correct correlation structure of μ and ξ_1.

We now consider a more complex four-step prototype Gibbs sampler with target distribution $p(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. As the number of components in a Gibbs sampler increases, there are more ways to construct PCG samplers; here we focus on an example where partial collapse does not correspond to blocking. (Generally this situation is even more complicated when the sampled component are vectors, in that we may marginalize out certain subvectors, see Park and van Dyk (2008).) We begin with the Gibbs sampler that iterates among

Step 1: Draw $\boldsymbol{W}$ from $p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, (Sampler 2.4)

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

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

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

Suppose it is possible to directly sample from $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})$, which are both conditional distributions of $\int p(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})d\boldsymbol{W}$. By replacing Steps 3 and 4 with draws from these two distributions, we are partially collapsing $\boldsymbol{W}$ out of Sampler 2.4. Substituting the conditional distributions of a marginal distribution of the target distribution into a Gibbs sampler, however, may result in a transition kernel with unknown stationary distribution. This is illustrated by the loss of correlation structure when using Sampler 2.2; see the last column of Figure 1. Nevertheless, we hope to capitalize on the potential computational gain that partial collapse offers. Thus, our goal is to formalize a procedure that allows us to introduce partially collapsed steps while ensuring the target stationary distribution is maintained. We illustrate our strategy in this example and formalize it in Section 3.

Moving components in a step of a Gibbs sampler from being conditioned upon to being sampled can improve the convergence characteristics of the sampler. This neither alters the stationary distribution of the chain nor destroys the compatibility of the conditional distributions. For example, based upon the available reduced conditional distributions, we can sample $\boldsymbol{W}$ jointly with $\boldsymbol{Y}$ in Step 3 and with $\boldsymbol{Z}$ in Step 4 (e.g., $\boldsymbol{W}$ and $\boldsymbol{Y}$ can be sampled jointly by first sampling from $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and then from $P(\boldsymbol{W}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$, both of which we have assumed are tractable). The resulting Gibbs sampler iterates among

Step 1: Draw $\boldsymbol{W}^{\star}$ from $p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, (Sampler 2.5)

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

Step 3: Draw $(\boldsymbol{W}^{\star}, \boldsymbol{Y})$ from $p(\boldsymbol{W}, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, and

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

In each step we condition on the most recently sampled value of each variable that is not sampled in that step. Thus, in Step 2, we condition on the $\boldsymbol{W} = \boldsymbol{W}^{\star}$ drawn in Step 1. The output is composed of the most recently sampled values of each variable at the end of the iteration, i.e., $\boldsymbol{W}$ drawn in Step 4, $\boldsymbol{X}$ drawn in Step 2, $\boldsymbol{Y}$ drawn in Step 3, and $\boldsymbol{Z}$ drawn 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. Sampler 2.5 is a trivial generalization of what is typically considered to be an ordinary Gibbs sampler, in that $\boldsymbol{W}$ is sampled more than once during an iteration. Sampler 2.5 may be inefficient in that it draws $\boldsymbol{W}$ three times in each iteration. Removing any two draws from the iteration, however, necessarily affects the transition kernel because the first draw is conditioned upon in the next step and the third draw is part of the output of the sampler. As Figure 1 illustrates, such changes to the transition kernel can destroy the correlation structure of the stationary distribution or otherwise affect the convergence of the chain.

In general, we only consider removing draws of intermediate quantities from a sampler because removing draws of any part of the output quantities and replacing output quantities with corresponding intermediate quantities necessarily alters the transition kernel and may affect the stationary distribution. Moreover, we only remove draws of intermediate quantities if removing them from the iteration does not affect the transition kernel. Permuting the steps of a Gibbs sampler does not alter its stationary distribution, but sometimes enables us to meet these criteria for removing redundant draws. In the case of Sampler 2.5, such permutation yields a Gibbs sampler that iterates among

Step 1: Draw $(\boldsymbol{W}^\star, \boldsymbol{Y})$ from $p(\boldsymbol{W}, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, (Sampler 2.6)
Step 2: Draw $(\boldsymbol{W}^\star, \boldsymbol{Z})$ from $p(\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})$,
Step 3: Draw $\boldsymbol{W}$ from $p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, and
Step 4: Draw $\boldsymbol{X}$ from $p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z})$,

where the first two draws of $\boldsymbol{W}$ correspond to intermediate quantities that are not conditioned upon and are not part of the output. This permutation alters the transition kernel, while maintaining the stationary distribution, and allows us to remove the two redundant draws of $\boldsymbol{W}$, without changing the transition kernel. Removing the intermediate quantities $\boldsymbol{W}^\star$ from Sampler 2.6 yields the PCG sampler that iterates among

Step 1: Draw $\boldsymbol{Y}$ from $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, (Sampler 2.7)
Step 2: Draw $\boldsymbol{Z}$ from $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})$,
Step 3: Draw $\boldsymbol{W}$ from $p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, and
Step 4: Draw $\boldsymbol{X}$ from $p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z})$.

We can block Steps 2 and 3 in Sampler 2.7 into a joint draw from $p(\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})$, thereby yielding

Step 1: Draw $\boldsymbol{Y}$ from $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, (Sampler 2.8)
Step 2: Draw $\boldsymbol{W}$ from $p(\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})$, and
Step 3: Draw $\boldsymbol{X}$ from $p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z})$.

The three conditional distributions in Sampler 2.8, however, remain incompatible. Thus, this PCG sampler does not simply correspond to a blocked version of Sampler 2.4. This illustrates that partial collapse is a more general technique than blocking.

The resulting PCG sampler (i.e., Sampler 2.8) is not an ordinary Gibbs sampler: Permuting its draws may result in a kernel with unknown stationary distribution. As in this case, however, we can sometimes verify that a PCG sampler is valid in that the resulting stationary distribution is equal to the target distribution. Still, since the removal of intermediate quantities introduces incompatibility, removal must be done with great care.

3 Basic Tools

Here we present three basic tools that we use to construct PCG samplers. Unless marginalized quantities are removed from the iteration with care, the resulting chain may not converge properly. Thus, the tools are designed to insure that the resulting PCG samplers converge quickly to the target distribution. We discuss the three tools, marginalization, permutation, and trimming, in the order that they are applied. *Of these trimming must be done with care, as it has the potential to alter the chain's stationary distribution.* Figure 2 presents a flowchart that describes how the basic tools are applied to an ordinary Gibbs sampler to construct a PCG sampler with the same stationary distribution. Each component of the flowchart in Figure 2 is closely examined in the following subsections.

3.1 Marginalization. We aim to construct a PCG sampler with stationary distribution $p(\boldsymbol{X})$ where $\boldsymbol{X}$ is a vector quantity that we partition into J subvectors, $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_J)$. Consider the sequence of index sets, $\boldsymbol{\mathcal{J}} = \{\boldsymbol{\mathcal{J}}_1, \boldsymbol{\mathcal{J}}_2, \ldots, \boldsymbol{\mathcal{J}}_P\}$, where $\boldsymbol{\mathcal{J}}_p \subset \{1, 2, \ldots, J\}$ for $p = 1, 2, \ldots, P$ such that $\cup_{p=1}^{P} \boldsymbol{\mathcal{J}}_p = \{1, 2, \ldots, J\}$. Let $\boldsymbol{\mathcal{J}}_p^c$ be the complement of $\boldsymbol{\mathcal{J}}_p$ in $\{1, 2, \ldots, J\}$ and $\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p}$ be the collection of components of $\boldsymbol{X}$ corresponding to the index set $\boldsymbol{\mathcal{J}}_p$, i.e., $\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p} = \{\boldsymbol{X}_j : j \in \boldsymbol{\mathcal{J}}_p\}$. Step p of a P-step Gibbs sampler can be written as

Step p**:** Draw $\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p}^{(t)}$ from $p(\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p} | \boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p^c}^{(t-1)})$, for $p = 1, 2, \ldots, P$,

where $\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p^c}^{(t-1)}$ consists of the most recently sampled values of each component of $\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_p^c}$. This is a Gibbs sampler that uses compatible conditional distributions, where some components of $\boldsymbol{X}$ may be updated in multiple steps; thus, it is not an ordinary Gibbs sampler.

An ordinary Gibbs sampler updates each (vector) component of $\boldsymbol{X}$ only once in an iteration. In our notation, this corresponds to the case where $\boldsymbol{\mathcal{J}}$ is a partition of $\{1, 2, \ldots, J\}$.

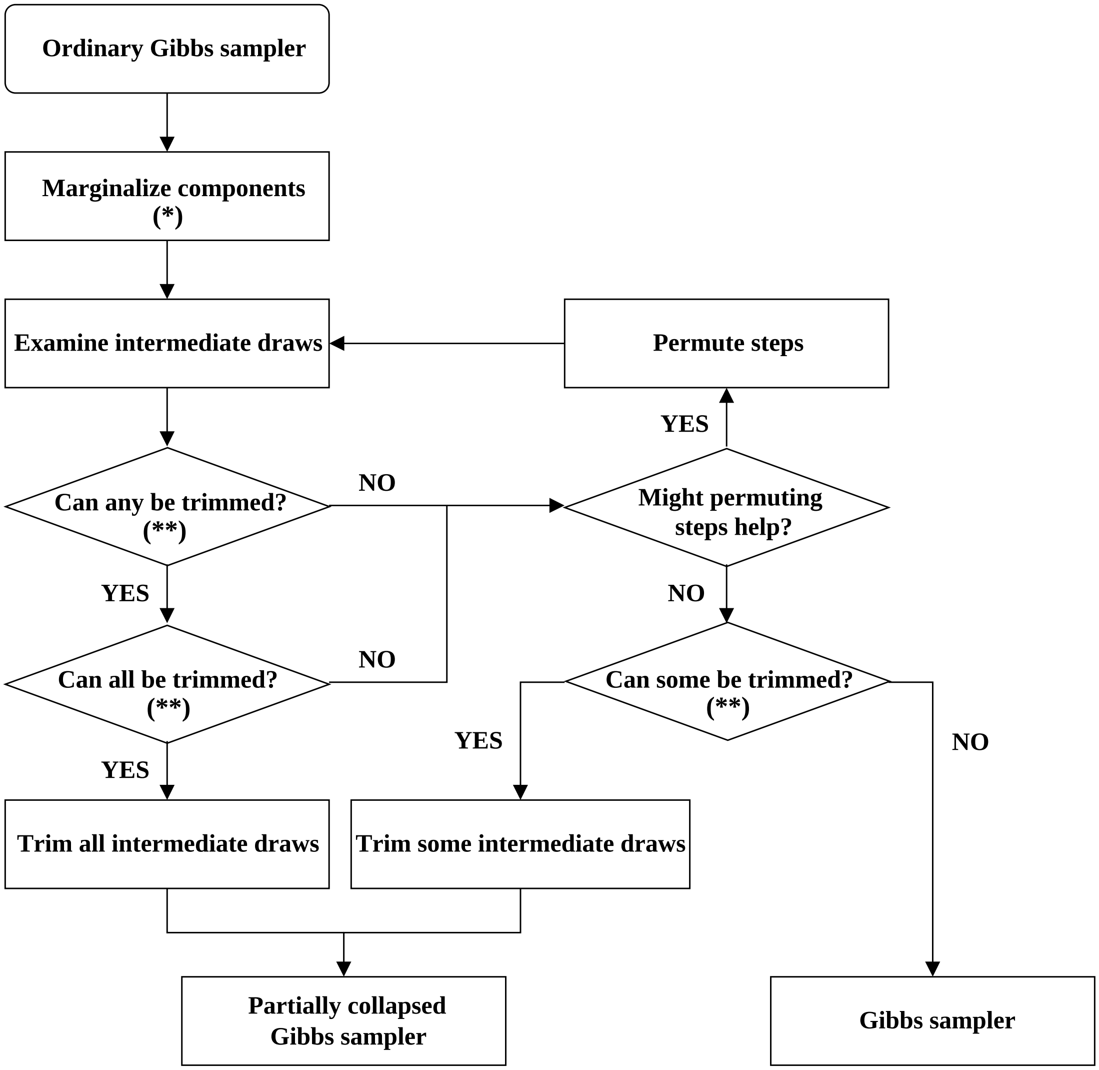
Ordinary Gibbs sampler
Marginalize components (*)
Examine intermediate draws
Permute steps
YES
Can any be trimmed? (**)
NO
Might permuting steps help?
YES
NO
Can all be trimmed? (**)
NO
Can some be trimmed? (**)
YES
YES
NO
Trim all intermediate draws
Trim some intermediate draws
Partially collapsed Gibbs sampler
Gibbs sampler

Figure 2: Flow Diagram for Deriving a PCG Sampler from an Ordinary Gibbs Sampler.

$(*)$ As many components should be marginalized in as many steps as possible.

$(**)$ Intermediate draws should only be trimmed if they are not conditional upon in subsequent draws.

At the other extreme, suppose there exists an index k such that $k \in \boldsymbol{\mathcal{J}}_p$ for each p, then $\boldsymbol{X}_k$ is drawn in each step and is never conditioned upon; we say $\boldsymbol{X}_k$ has been (completely) collapsed out of the Gibbs sampler. In this case we can reformulate the Gibbs sampler in terms of the marginal distribution $\int \pi(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_J) d\boldsymbol{X}_k$, without altering its kernel.

The first step in constructing a PCG sampler is to marginalize some components of $\boldsymbol{X}$ out of some steps of the sampler. To do this we replace $\boldsymbol{\mathcal{J}}_q$ with $\widetilde{\boldsymbol{\mathcal{J}}}_q$ for some $q \in \{1, 2, \ldots, P\}$ where $\boldsymbol{\mathcal{J}}_q$ is a proper subset of $\widetilde{\boldsymbol{\mathcal{J}}}_q$. That is, in Step q, we move some components of $\boldsymbol{X}$ from being conditioned upon to being sampled. As we shall see, the marginalization can improve the convergence properties of the Gibbs sampler; see Section 4 for the theory on the improved rate of convergence. Then Step q is conditional on fewer components of $\boldsymbol{X}$:

Step q: Draw $\boldsymbol{\mathcal{X}}^{(t)}_{\widetilde{\boldsymbol{\mathcal{J}}}_q}$ from $p(\boldsymbol{\mathcal{X}}_{\widetilde{\boldsymbol{\mathcal{J}}}_q} | \boldsymbol{\mathcal{X}}^{(t-1)}_{\widetilde{\boldsymbol{\mathcal{J}}}^c_q})$.

Marginalizing out some components of $\boldsymbol{X}$ alters the transition kernel but not the stationary distribution for $\boldsymbol{X}$ or the compatibility of the conditional distributions. The improved convergence of the sampler is mainly attributed to marginalization.

3.2 Permutation. In the case of a P-step Gibbs sampler, the steps can be reordered into $P!$ possible permutations. Permuting the compatible conditional distributions of a Gibbs sampler typically changes its transition kernel and interchanges intermediate quantities with output quantities, but maintains the stationary distribution of the chain. Our goal in permuting the steps is to arrange them, so that as many of marginalized components as possible are intermediate quantities that are not conditioned upon in subsequent steps.

The permutation of the steps may affect the convergence of a sampler, but its influence is typically small as compared to that of marginalization (van Dyk and Meng, 1997). In this article, we tend to ignore the effect of permutation on convergence. Here we are merely interested in permutations because they can allow the removal of some intermediate quantities.

3.3 Trimming. By trimming we mean discarding a subset of the components that were to be sampled in one or more steps of a Gibbs sampler. In the P-step Gibbs sampler, for example, trimming the marginalized components of X in Step q yields

Step q: Draw $\boldsymbol{\mathcal{X}}^{(t)}_{\boldsymbol{\mathcal{J}}_q}$ from $p(\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_q}|\boldsymbol{\mathcal{X}}^{(t-1)}_{\widetilde{\boldsymbol{\mathcal{J}}}^c_q})$.

The reduced conditional distribution sampled in this step is not typically compatible with the other conditional distributions sampled in the sampler. In particular, because $\boldsymbol{\mathcal{J}}_q \cup \widetilde{\boldsymbol{\mathcal{J}}}^c_q$ is not equal to $\boldsymbol{\mathcal{J}}$ ($\boldsymbol{\mathcal{J}}_q$ is a proper subset of $\widetilde{\boldsymbol{\mathcal{J}}}_q$), this conditional distribution is not defined on the same space as the conditional distributions of the original sampler. Thus, it is trimming that introduces incompatibility into the conditional distributions of a PCG sampler. This means the resulting PCG sampler may no longer be a Gibbs sampler, per se, since Gibbs samplers are generally expected to be constructed with compatible conditional distributions. Unlike a Gibbs sampler, permuting the steps of a PCG sampler may result in a new Markov transition kernel with an unknown stationary distribution. Nonetheless, trimming is advantageous because each iteration is less computationally demanding. Indeed trimming may render an intractable sampling step tractable, see Park and van Dyk (2008).

We emphasize that trimming must be done carefully because it has the potential to alter the stationary distribution of the chain. *Intermediate quantities may be conditioned upon in subsequent draws. Thus, we can only trim intermediate quantities that are not conditioned upon if we hope to maintain the transition kernel.* Trimming intermediate quantities that are conditioned upon in subsequent steps can impact the transition kernel and the correlation structure of the stationary distribution. Thus, care must be taken when trimming intermediate quantities in order to maintain the stationary distribution.

4 PCG Theory

To discuss the effect of partial collapse on convergence, we introduce some technical concepts concerning Markov chains. (We follow the notation of Liu, 2001, Section 6.7.) Let

$L^2(\pi)$ denote the set of all functions $h(\boldsymbol{X})$ such that $\int h^2(\boldsymbol{X})\pi(\boldsymbol{X})d\boldsymbol{X} < \infty$. This set is a Hilbert space with inner product $\langle h, g\rangle = \mathrm{E}_\pi\{h(\boldsymbol{X})g(\boldsymbol{X})\}$, so that $\|h\|^2 = \mathrm{Var}_\pi(h)$. For a general Markov chain $\boldsymbol{\mathcal{M}}_{\boldsymbol{X}} = \{\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \dots\}$ with transition kernel $\mathcal{K}(\boldsymbol{X}^{(1)} = \boldsymbol{X}|\boldsymbol{X}^{(0)} = \boldsymbol{X}')$, we define the forward operator $\boldsymbol{F}$ on $L^2(\pi)$ for $\boldsymbol{\mathcal{M}}_{\boldsymbol{X}}$ by

$$\boldsymbol{F}h(\boldsymbol{X}') = \int h(\boldsymbol{X})\mathcal{K}(\boldsymbol{X}|\boldsymbol{X}')d\boldsymbol{X} = \mathrm{E}\{h(\boldsymbol{X}^{(1)})|\boldsymbol{X}^{(0)} = \boldsymbol{X}'\}.$$

Let $\boldsymbol{L}_0^2(\pi) = \big\{h : \mathrm{E}_\pi\{h(\boldsymbol{X})\} = 0, \mathrm{Var}_\pi\{h(\boldsymbol{X})\} < \infty\big\}$. This is also a Hilbert space with the same inner product and is invariant under $\boldsymbol{F}$. We define $\boldsymbol{F}_0$ to be the forward operator on $L_0^2(\pi)$ induced by $\boldsymbol{F}$. If we define the norm of this forward operator by $\|\boldsymbol{F}_0\| = \sup_h \|\boldsymbol{F}_0 h(\boldsymbol{X})\|$ with the supremum taken over $h \in L_0^2(\pi)$, it can be shown that

$$\begin{aligned}\|\boldsymbol{F}_0\| &= \sup_{h\in \boldsymbol{L}_0^2(\pi)} \Big(\mathrm{Var}_\pi\big[\mathrm{E}\{h(\boldsymbol{X}^{(1)})|\boldsymbol{X}^{(0)}\}\big]\Big)^{1/2} \\ &= \sup_{h\in \boldsymbol{L}_0^2(\pi)} \Big\{\mathrm{E}_\pi\Big(\big[\mathrm{E}\{h(\boldsymbol{X}^{(1)})|\boldsymbol{X}^{(0)}\}\big]^2\Big)\Big\}^{1/2} = \rho(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)}),\end{aligned}$$

where $\rho(\boldsymbol{\vartheta}, \boldsymbol{\varphi})$ is the maximum correlation of ϑ and φ,

$$\rho(\boldsymbol{\vartheta}, \boldsymbol{\varphi}) = \sup \mathrm{Corr}\{h(\boldsymbol{\vartheta}), g(\boldsymbol{\varphi})\} = \sup_{h:\, \mathrm{Var}\{h(\boldsymbol{\vartheta})\}=1} \big(\mathrm{Var}_\pi\big[\mathrm{E}\{h(\boldsymbol{\vartheta})|\boldsymbol{\varphi}\}\big]\big)^{1/2},$$

where the first sup is over all non-constant scalar functions h and g with finite variance; see, e.g., Liu *et al.* (1994). Here the maximum autocorrelation $\rho(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)})$ is computed under the stationary distribution of $\boldsymbol{\mathcal{M}}_{\boldsymbol{X}}$, and is also denoted by $\rho(\boldsymbol{\mathcal{M}}_{\boldsymbol{X}})$.

The spectral radius of $\boldsymbol{F}_0$, $r(\boldsymbol{F}_0)$, typically governs the convergence of $\boldsymbol{\mathcal{M}}_{\boldsymbol{X}}$ (Liu, 2001), and is related to the norm by

$$\lim_{n\to\infty} \|\boldsymbol{F}_0^n\|^{1/n} = r(\boldsymbol{F}_0) \quad \text{and by the inequality} \quad r(\boldsymbol{F}_0) \le \|\boldsymbol{F}_0\|. \tag{2}$$

Along with the relationship between the maximum autocorrelation of $\boldsymbol{\mathcal{M}}_{\boldsymbol{X}}$ and $\|\boldsymbol{F}_0\|$, (2) justifies the use of $\|\boldsymbol{F}_0\|$ in the analysis of the convergence behavior of $\boldsymbol{\mathcal{M}}_{\boldsymbol{X}}$.

Consider the P-step Gibbs sampler described in Section 3. We define a p-step-lagged Gibbs sampler for $p = 0, 1, \dots, P-1$, as the Gibbs sampler with iteration that begins with Step $p{+}1$, cycles through the steps in the order as the original sampler, and ends with Step p. The forward operators of the P p-step-lagged Gibbs samplers (with $p = 0, 1, \dots, P-1$) have the same spectral radius, which we call r. They may, however, have different norms and maximum autocorrelations. We denote the maximum autocorrelation $\rho(\boldsymbol{\mathcal{M}}_{\boldsymbol{X}})$ of the p-step-lagged chain by ρ_p for $p = 0, 1, \dots, P-1$. By (2), we have

$$r \le \min_{p\in\{0,1,\dots,P-1\}} \rho_p, \tag{3}$$

where we call $\min_{p\in\{0,1,\dots,P-1\}} \rho_p$ the *cyclic-permutation bound* on the spectral radius. Below we show that by marginalizing a component of $\boldsymbol{X}$ in Step $p{+}1$ (the first step of the p-step-lagged Gibbs sampler) we reduce ρ_p, thereby reducing the bound given in (3) on the spectral radius. This leads to the following result.

Theorem 1 *Sampling more components of $\boldsymbol{X}$ in any set of steps of a Gibbs sampler can only reduce the cyclic-permutation bound on the spectral radius of that Gibbs sampler.*

It remains only to show that marginalizing a component of $\boldsymbol{X}$ in Step $p+1$ reduces ρ_p. Because Step $p{+}1$ is the first step of the p-step-lagged Gibbs sampler, we evaluate the effect of marginalizing a component of $\boldsymbol{X}$ in Step $p{+}1$ on ρ_p. This is because the theorem below evaluates the effect of marginalization in the first step of a Gibbs sampler. To illustrate the computational advantages of the partial collapse, we consider the generic P-step Gibbs sampler introduced in Section 3 from which we marginalize some components of $\boldsymbol{X}$ in Step 1. Thus, we wish to compare two sequences of index sets and their resulting transition kernels; namely $(\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_1}, \boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_2}, \dots, \boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_P})$ and its kernel $\mathcal{K}(\boldsymbol{X}|\boldsymbol{X}')$ and $(\boldsymbol{\mathcal{X}}_{\widetilde{\boldsymbol{\mathcal{J}}_1}}, \boldsymbol{\mathcal{X}}_{\widetilde{\boldsymbol{\mathcal{J}}_2}}, \dots, \boldsymbol{\mathcal{X}}_{\widetilde{\boldsymbol{\mathcal{J}}_P}})$ and its kernel $\widetilde{\mathcal{K}}(\boldsymbol{X}|\boldsymbol{X}')$, where $\boldsymbol{\mathcal{J}}_p = \widetilde{\boldsymbol{\mathcal{J}}}_p$ for $p = 2, \dots, P$, but $\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{J}}_1} = \{\boldsymbol{x}_1\}$ and $\boldsymbol{\mathcal{X}}_{\widetilde{\boldsymbol{\mathcal{J}}_1}} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ with $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$. Here $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ is an alternate partition of $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_J)$ introduced to simplify notation in the theorem. That is, $\boldsymbol{\mathcal{J}}_1 \subset \widetilde{\boldsymbol{\mathcal{J}}}_1 \subset \{1, 2, \dots, J\}$, where both subsets are proper subsets, $\boldsymbol{x}_1 = \{\boldsymbol{X}_j : j \in \boldsymbol{\mathcal{J}}_1\}$, $\boldsymbol{x}_2 = \{\boldsymbol{X}_j : j \in \widetilde{\boldsymbol{\mathcal{J}}}_1 \setminus \boldsymbol{\mathcal{J}}_1\}$, and $\boldsymbol{x}_3 = \{\boldsymbol{X}_j : j \in \widetilde{\boldsymbol{\mathcal{J}}}_1^c\}$. In words, the two sequence of index sets represent identical samplers, except in Step 1, where more components of $\boldsymbol{X}$ are drawn in the Gibbs sampler with kernel $\widetilde{\mathcal{K}}(\boldsymbol{X}|\boldsymbol{X}')$. In this case, we have the following result.

Theorem 2 *Sampling more components of $\boldsymbol{X}$ in the first step of a Gibbs sampler improves the resulting maximal autocorrelation, $\rho(\boldsymbol{\mathcal{M}}_X)$.*

Proof: Let h be an arbitrary function of $\boldsymbol{X}$ with mean zero and finite variance under stationarity, i.e., $h \in \boldsymbol{L}_0^2(\pi)$ with π the stationary distribution of $\boldsymbol{X}$, then

$$\widetilde{\mathrm{E}}\{h(\boldsymbol{X})|\boldsymbol{x}_3'\} = \int h(\boldsymbol{X})\mathcal{K}_{-1}(\boldsymbol{X}|\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3')p(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{x}_3')d\Xi_{-1}d\boldsymbol{x}_1 d\boldsymbol{x}_2, \tag{4}$$

where $\widetilde{\mathrm{E}}$ represents expectation with respect to $\widetilde{\mathcal{K}}(\boldsymbol{X}|\boldsymbol{X}')$, $\mathcal{K}_{-1}(\boldsymbol{X}|\boldsymbol{X}')$ is the transition kernel implied by Step 2 through Step P of either sampler, and $\boldsymbol{\Xi}_{-1} = (\boldsymbol{\mathcal{X}}_{\mathcal{J}_2}, \ldots, \boldsymbol{\mathcal{X}}_{\mathcal{J}_P})$ is the set of components updated in Step 2 through Step P, which may include multiple copies of certain components of $\boldsymbol{X}$. Now, the right-hand side of (4) can be written as

$$\begin{aligned}\widetilde{\mathrm{E}}\{h(\boldsymbol{X})|\boldsymbol{x}_3'\} &= \int \left\{ \int h(\boldsymbol{X})\mathcal{K}_{-1}(\boldsymbol{X}|\boldsymbol{x}_1, \boldsymbol{x}_2', \boldsymbol{x}_3')p(\boldsymbol{x}_1|\boldsymbol{x}_2', \boldsymbol{x}_3')d\Xi_{-1}d\boldsymbol{x}_1 \right\} p(\boldsymbol{x}_2'|\boldsymbol{x}_3')d\boldsymbol{x}_2' \\ &= \mathrm{E}_\pi[\mathrm{E}\{h(\boldsymbol{X}) \,|\, \boldsymbol{x}_2', \boldsymbol{x}_3'\} \,|\, \boldsymbol{x}_3'],\end{aligned}$$

where the inner expectation is with respect to $\mathcal{K}(\boldsymbol{X}|\boldsymbol{X}')$. Thus,

$$\begin{aligned}\mathrm{E}_\pi\Big(\big[\widetilde{\mathrm{E}}\{h(\boldsymbol{X}) \,|\, \boldsymbol{x}_3'\}\big]^2\Big) &= \mathrm{E}_\pi\Big\{\big(\mathrm{E}_\pi[\mathrm{E}\{h(\boldsymbol{X}) \,|\, \boldsymbol{x}_2', \boldsymbol{x}_3'\} \,|\, \boldsymbol{x}_3']\big)^2\Big\} \\ &\le \mathrm{E}_\pi\Big\{\mathrm{E}_\pi\big([\mathrm{E}\{h(\boldsymbol{X}) \,|\, \boldsymbol{x}_2', \boldsymbol{x}_3'\}]^2 \,|\, \boldsymbol{x}_3'\big)\Big\} = \mathrm{E}_\pi\Big(\big[\mathrm{E}\{h(\boldsymbol{X}) \,|\, \boldsymbol{x}_2', \boldsymbol{x}_3'\}\big]^2\Big).\end{aligned}$$

But since $\mathrm{Var}_\pi[h(\boldsymbol{X})]$ is the same for both kernels, the maximal autocorrelation induced by $\widetilde{\mathcal{K}}(\boldsymbol{X}|\boldsymbol{X}')$ is bounded above by that of $\mathcal{K}(\boldsymbol{X}|\boldsymbol{X}')$. ■

That is, the computational advantages can be achieved by successively marginalizing over the components of $\boldsymbol{X}$ in any single step of a Gibbs sampler. Thus, repeatedly applying Theorem 2 provides the theoretical basis for the improved convergence of PCG samplers.

5 Concluding Remarks

In this article, we present efficient Gibbs sampling techniques developed by generalizing the composition of the conditional distributions in Gibbs samplers. Unlike ordinary Gibbs samplers, PCG samplers use incompatible conditional distributions to improve the convergence characteristics of the samplers. This generalization comes at a price: the conditional distributions that compose a PCG sampler need be performed in a certain order to maintain the target stationary distribution. We introduce three basic prescriptive tools, marginalization, permutation, and trimming, and show how sequentially applying these tools can transform a Gibbs sampler into a PCG sampler. The resulting PCG sampler may be composed of a set of incompatible conditional distributions and generally exhibits superior convergence characteristics. This strategy is illustrated in the fitting of three models stemming from our applied work in astronomy and multiple imputation in Park and van Dyk (2008).

References

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag, New York.

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.

Park, T. and van Dyk, D. A. (2008). Partially collapsed Gibbs samplers: Illustrations and applications. *Technical Report* .

van Dyk, D. A. and Meng, X.-L. (1997). Some findings on the orderings and groupings of conditional maximizations within ECM-type algorithms. *The Journal of Computational and Graphical Statistics* **6**, 202–223.