Chapter 1

Partially collapsed Gibbs sampling & path-adaptive Metropolis-Hastings in high-energy astrophysics

David A. van Dyk and Taeyoung Park

1.1 Introduction

As the many examples in this book illustrate, Markov chain Monte Carlo (MCMC) methods have revolutionized Bayesian statistical analyses. Rather than using off-the-shelf models and methods, we can use MCMC to fit application specific models that are designed to account for the particular complexities of a problem at hand. These complex multilevel models are becoming more prevalent throughout the natural, social, and engineering sciences largely because of the ease of using standard MCMC methods such as the Gibbs and Metropolis-Hastings (MH) samplers. Indeed, the ability to easily fit statistical models that directly represent the complexity of a data generation mechanism has arguably lead to the increased popularity of Bayesian methods in many scientific disciplines.

Although simple standard methods work surprisingly well in many problems, neither the Gibbs nor the MH sampler can directly handle problems with very high posterior correlations among the parameters. The marginal distribution of a given parameter is much more variable than the corresponding full conditional distribution in this case, causing the Gibbs sampler to take small steps. With MH a proposal distribution that does not account for the posterior correlation either has far too much mass in regions of low posterior probability or has such small marginal variances that only small steps are proposed, causing high rejection rates and/or high autocorrelations in the resulting Markov chains. Unfortunately, accounting for the posterior correlation requires more information about the posterior distribution than is typically available when the proposal distribution is constructed.

Much work has been devoted to developing computational methods that extend the usefulness of these standard tools in the presence of high correlations. For Gibbs sampling, for example, it is now well known that blocking or grouping steps (Liu et al., 1994), nesting steps (van Dyk, 2000), collapsing or marginalizing parameters (Liu et al., 1994; Meng and van Dyk, 1999), incorporating auxiliary variables (Besag and Green, 1993), certain parameter transformations (Gelfand et al., 1995; Yu, 2005), and parameter expansion (Liu and Wu, 1999) can all be used to improve the convergence of certain samplers. By embedding an MH sampler within the Gibbs sampler and updating one parameter at a time (i.e., the well-known Metropolis within Gibbs sampler) the same strategies can be used to improve MH samplers.

In this chapter we describe two newer methods that are designed to improve the performance of Gibbs and Metropolis within Gibbs samplers. The partially collapsed Gibbs sampler (Park and van Dyk, 2009; van Dyk and Park, 2008) takes advantage of the fact that we expect reducing conditioning to increase the variance of the complete conditional distributions of a Gibbs sampler. Thus, by replacing a subset of the complete conditional distributions by distributions that condition on fewer of the unknown quantities, i.e., conditional distributions of some marginal distributions of the target posterior distribution, we expect the sampler to take larger steps and its overall convergence characteristics to improve.

This strategy must be used with care, however, since the resulting set of conditional distributions may not be functionally compatible and changing the order of the draws can alter the stationary distribution of the chain. The second strategy involves updating the Metropolis proposal distribution to take account of what is known about the target distribution given an initial set of draws.

Although these are both general strategies with many potential applications, they were both motivated by a particular model fitting task in high-energy astrophysics. In recent years, technological advances have dramatically increased the quality and quantity of data available to astronomers. Multilevel statistical models are used to account for these complex data generation mechanisms which can include both the physical data sources and sophisticated instrumentation. Bayesian methods and MCMC techniques both find many applications among the many resulting statistical problems and are becoming evermore popular among astronomers. Examples include the search for planets orbiting distant stars (Gregory, 2005), the analysis of stellar evolution using sophisticated physics-based computer models (DeGennaro et al., 2008; van Dyk et al., 2009), the analysis of the composition and temperature distribution of stellar coronae (Kashyap and Drake, 1998), and the search for multi-scale structure in X-ray images (Connors and van Dyk, 2007; Esch et al., 2004), to name just a few. In this chapter we describe the partially collapsed Gibbs sampler and the path-adaptive Metropolis-Hastings sampler and show how they can dramatically improve the computational performance of MCMC samplers designed to search for narrow emission lines in high-energy astronomical spectral analysis.

1.2 Partially collapsed Gibbs sampler

Collapsing in a Gibbs sampler involves integrating a joint posterior distribution over a subset of unknown quantities to construct a marginal or *collapsed* posterior distribution under which a new collapsed Gibbs sampler is built (Liu et al., 1994). This strategy is similar the efficient data augmentation strategy used to improve the rate of convergence of the EM algorithm (van Dyk and Meng, 1997). Efficient data augmentation aims to construct an EM algorithm using as little missing data as possible. That is, a portion of the missing data is collapsed out of the distribution of unknown quantities. It is known that reducing the missing data in this way can only improve the rate of convergence of the EM algorithm (Meng and van Dyk, 1997).

Although these collapsing or marginalization strategies typically improve convergence, they may not be easy to implement. For example, the complete conditional distributions of the collapsed posterior distribution may be harder to work with than the conditional distributions of the original posterior distribution. The partially collapsed Gibbs sampler aims to take partial computational advantage of the collapsing strategy while maintaining simple implementation by mixing conditional distributions from the original posterior distribution with those of one or more collapsed posterior distribution. Thus, we use collapsing only in those conditional distributions where it does not complicate parameter updating. This strategy is analogous to the ECME and AECM algorithms which generalize EM by allowing different amounts of missing data when updating different model parameters (Liu and Rubin, 1994; Meng and van Dyk, 1997), see Park and van Dyk (2009) and van Dyk and Meng (2008) for discussion.

To see both the potential advantages and the potential pitfalls of partially collapsing a Gibbs sampler, consider a simple two-step sampler:

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

Step 2: Draw $\mathbf{Y}$ from $p(\mathbf{Y}|\mathbf{X})$.

Suppose that the marginal distribution of $\mathbf{Y}$ is a standard distribution so that we can easily eliminate the conditioning on $\mathbf{X}$ in Step 2:

Step 1: Draw $\mathbf{X}$ from $p(\mathbf{X}|\mathbf{Y})$, (Sampler 2)

Step 2: Draw $\mathbf{Y}$ from $p(\mathbf{Y})$.

This is advantageous because we immediately obtain independent draws of $\mathbf{Y}$. The complication, however, is that the two "conditional" distributions in Sampler 2, $p(\mathbf{X}|\mathbf{Y})$ and

$p(\mathbf{Y})$ are incompatible and imply inconsistent dependence structure. Sampling $\mathbf{Y}$ from $p(\mathbf{Y})$ suggests that $\mathbf{X}$ and $\mathbf{Y}$ are independent, whereas sampling $\mathbf{X}$ from $p(\mathbf{X}|\mathbf{Y})$ suggests dependence. Because the stationary distribution of Sampler 2 is $p(\mathbf{X})p(\mathbf{Y})$ rather than the target distribution $p(\mathbf{X}, \mathbf{Y})$, information on the posterior correlation is lost. Of course, there is an obvious solution. If we simply change the order of the draws in Sampler 2, first sampling $\mathbf{Y}$ from its marginal distribution and then $\mathbf{X}$ from its conditional distribution, we obtain independent draws from the target posterior distribution. The conditional distributions remain incompatible, but the resulting Markov chain converges faster than that of Sampler 1 while maintaining the correlation of the target distribution. In this case, the partially collapsed sampler is simply a blocked version of Sampler 1: Sampling $p(\mathbf{Y})$ and then $p(\mathbf{X}|\mathbf{Y})$ combines into a single draw from the target distribution. As we shall illustrate, however, partially collapsed Gibbs samplers are more general than blocked Gibbs samplers.

This simple two-step sampler illustrates an important point: Care must be taken if we are to maintain the target stationary distribution when reducing the conditioning in some but not all of the steps of a Gibbs sampler. Van Dyk and Park (2008) describe three basic tools that can be used to transform a Gibbs sampler into a partially collapsed Gibbs sampler that mains the target distribution. 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. In Sampler 1 this involves replacing the sampling of $p(\mathbf{Y}|\mathbf{X})$ with the sampling of $p(\mathbf{X}, \mathbf{Y})$ in Step 2, see Figure 1.1(a) and (b). Notice that rather than simply reducing the conditioning as described above, we are moving $\mathbf{X}$ from being conditioned upon to being sampled. This preserves the stationary distribution of the Markov chain. The second tool is *permutation* of the steps. We may need to permute steps in order to use the third tool, which is to *trim* sampled components from steps if the components can be removed from the sampler without altering its Markov transition kernel. In Figure 1.1(c), we permute the steps so that we can trim $\mathbf{X}^\star$ from the sampler in (d). 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. Finally, we block the two steps in (e).

Both marginalization and permutation clearly maintain the stationary distribution of the

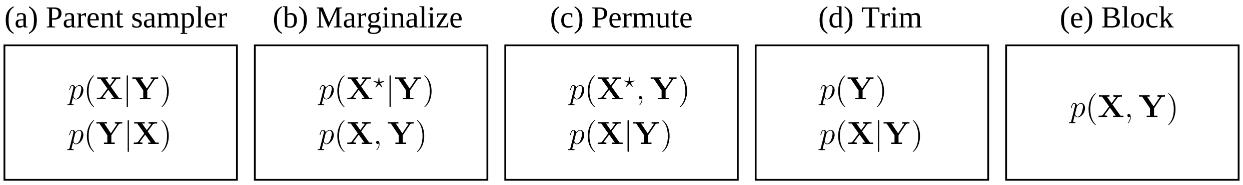
(a) Parent sampler
$p(\mathbf{X}|\mathbf{Y})$
$p(\mathbf{Y}|\mathbf{X})$
(b) Marginalize
$p(\mathbf{X}^\star|\mathbf{Y})$
$p(\mathbf{X},\mathbf{Y})$
(c) Permute
$p(\mathbf{X}^\star,\mathbf{Y})$
$p(\mathbf{X}|\mathbf{Y})$
(d) Trim
$p(\mathbf{Y})$
$p(\mathbf{X}|\mathbf{Y})$
(e) Block
$p(\mathbf{X},\mathbf{Y})$

Figure 1.1: Transforming Sampler 1 into a partially blocked Gibbs sampler using marginalization, permutation, and trimming. The sampler in (e) is a blocked version of Sampler 1.

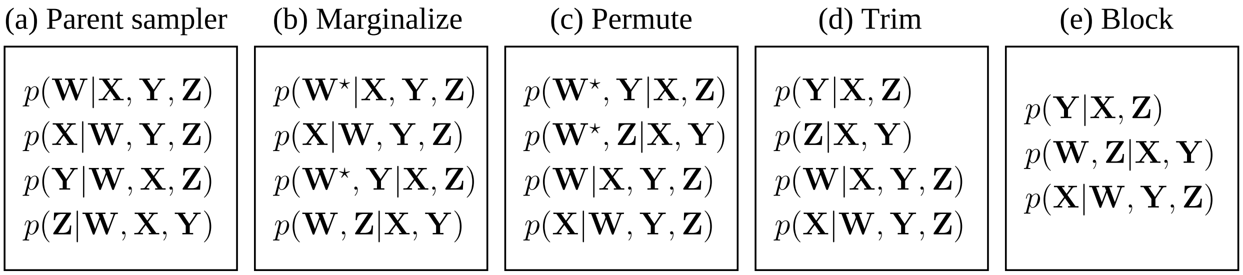
(a) Parent sampler
$p(\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z})$
$p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z})$
$p(\mathbf{Y}|\mathbf{W},\mathbf{X},\mathbf{Z})$
$p(\mathbf{Z}|\mathbf{W},\mathbf{X},\mathbf{Y})$
(b) Marginalize
$p(\mathbf{W}^\star|\mathbf{X},\mathbf{Y},\mathbf{Z})$
$p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z})$
$p(\mathbf{W}^\star,\mathbf{Y}|\mathbf{X},\mathbf{Z})$
$p(\mathbf{W},\mathbf{Z}|\mathbf{X},\mathbf{Y})$
(c) Permute
$p(\mathbf{W}^\star,\mathbf{Y}|\mathbf{X},\mathbf{Z})$
$p(\mathbf{W}^\star,\mathbf{Z}|\mathbf{X},\mathbf{Y})$
$p(\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z})$
$p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z})$
(d) Trim
$p(\mathbf{Y}|\mathbf{X},\mathbf{Z})$
$p(\mathbf{Z}|\mathbf{X},\mathbf{Y})$
$p(\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z})$
$p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z})$
(e) Block
$p(\mathbf{Y}|\mathbf{X},\mathbf{Z})$
$p(\mathbf{W},\mathbf{Z}|\mathbf{X},\mathbf{Y})$
$p(\mathbf{X}|\mathbf{W},\mathbf{Y},\mathbf{Z})$

Figure 1.2: Transforming a four-step Gibbs sampler into a partially blocked Gibbs sampler. The sampler in (e) is composed of incompatible conditional distributions, is not a blocked version of the sampler in (a), and is therefore not a Gibbs sampler *per se*.

chain and both can effect its convergence properties; marginalization can dramatically improve convergence, while the effect of a permutation is typically small. Reducing conditioning (i.e., marginalization) increases variance and hence the sizes of the sampling jumps, see van Dyk and Park (2008) for a technical treatment. Trimming is explicitly designed to maintain the kernel of the chain. The primary advantage of trimming is to reduce the complexity of the individual steps. In doing so, trimming may introduce incompatibility into a sampler.

To illustrate how the three tools are used in a more realistic setting we use the simple four-step example given in Figure 1.2(a). 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 the third and forth draws with draws from these conditional distributions we would have no direct way of verifying that the stationary distribution of the resulting chain is the the target joint distribution. Instead, we use the three basic tools to derive a partially collapsed Gibbs sampler. This allows us to reap the benefits of partial collapse while ensuring that the stationary distribution of the chain is the target posterior distribution.

In Figure 1.2(b), we use marginalization to move $\mathbf{W}$ from being conditioned upon to being sampled in the last two steps. 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 the most recently sampled valued of each quantity at the end of the iteration: $\mathbf{X}$ sampled in the second step, $\mathbf{Y}$ sampled in third step, and $(\mathbf{W}, \mathbf{Z})$ sampled in last step. Although sampling $\mathbf{W}$ three times may be inefficient, removing any two of the three draws affects the transition kernel of the chain: The draw in the first step is conditioned upon in the second step and the draw in the last step is part of the output of the iteration. In order to preserve the stationary distribution, we only remove 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. In Figure 1.2(c) we permute the steps so that two of the draws of $\mathbf{W}$ can be trimmed in (d). The intermediate draws of $\mathbf{W}$ sampled in the first and second steps of (c) are not used subsequently and both can be removed from the sampler. Finally, the middle two steps of (d) can be combined to derive the final sampler given in (e).

The samplers in (c) and (d) have the same stationary distribution because removing the intermediate quantities does not affect the transition kernel. Thus, we know the stationary distribution of (d) is the target posterior distribution. This illustrates how careful use of the three basic tools can lead to partially collapsed Gibbs samplers with the target stationary distribution. Notice that the samplers in (d) and (e) are not Gibbs sampler *per se*. The conditional distributions that are sampled in each are incompatible and permuting their order may alter the stationary distribution of the chain.

1.3 Path-adaptive Metropolis-Hastings sampler

The second computational method aims to improve the convergence of the MH sampler by updating the proposal distribution using information about the target distribution obtained from an initial run of the chain. Suppose a target distribution of interest has density $\pi(\mathbf{X})$. Given a current state $\mathbf{X}^{(t)}$, the MH sampler proposes a state $\mathbf{X}'$ using a proposal distribution

$p_1(\mathbf{X}'|\mathbf{X}^{(t)})$; we use a one in the subscript because we update this proposal distribution below. The move from $\mathbf{X}^{(t)}$ to $\mathbf{X}'$ is accepted with probability

$$q_1(\mathbf{X}'|\mathbf{X}^{(t)}) = \min\left\{1, \frac{\pi(\mathbf{X}')/p_1(\mathbf{X}'|\mathbf{X}^{(t)})}{\pi(\mathbf{X}^{(t)})/p_1(\mathbf{X}^{(t)}|\mathbf{X}')}\right\}$$

and $\mathbf{X}^{(t+1)}$ is set to $\mathbf{X}'$ with probability $q_1(\mathbf{X}'|\mathbf{X}^{(t)})$ and to $\mathbf{X}^{(t)}$ otherwise. Thus, for any $\mathbf{X}^{(t+1)} \neq \mathbf{X}^{(t)}$, the transition kernel of the MH sampler is

$$\mathcal{K}_1(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}) = p_1(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)})q_1(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}).$$

The path-adaptive Metropolis-Hastings (PAMH) sampler is an efficient MH sampler that uses an empirical distribution generated from an initial run of the chain (i.e., the path samples of the chain) as a second proposal distribution. This is used to construct a second transition kernel that is mixed with the original transition kernel in subsequent draws. In this way, we use the sample generated by MH to construct a proposal distribution that more closely resembles the target distribution. This can dramatically improve performance if the original MH sampler is either slow mixing or computationally demanding.

The PAMH sampler is a mixture of two MH samplers: With probability α, a proposal state $\mathbf{X}'$ is generated from $p_1(\mathbf{X}'|\mathbf{X}^{(t)})$ and accepted with probability $q_1(\mathbf{X}'|\mathbf{X}^{(t)})$; and with probability $1-\alpha$, a proposal state $\mathbf{X}'$ is generated from an empirical distribution $\hat{\pi}(\mathbf{X})$ and accepted with probability

$$q_2(\mathbf{X}'|\mathbf{X}^{(t)}) = \min\left\{1, \frac{\pi(\mathbf{X}')/\hat{\pi}(\mathbf{X}')}{\pi(\mathbf{X}^{(t)})/\hat{\pi}(\mathbf{X}^{(t)})}\right\}.$$

Thus, for any $\mathbf{X}^{(t+1)} \neq \mathbf{X}^{(t)}$, the transition kernel of the PAMH sampler is given by

$$\mathcal{K}_+(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}) = \alpha\mathcal{K}_1(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}) + (1-\alpha)\mathcal{K}_2(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}), \qquad (1.3.1)$$

where $\mathcal{K}_2(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}) = \hat{\pi}(\mathbf{X}^{(t+1)})q_2(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)})$.

The mixture proportion α is a tuning parameter that is set in advance. In effect the value of α is set to one during the initial run that uses only the original proposal distribution to generate samples from $\pi(\mathbf{X})$, and the path samples from the initial run are then used to compute an approximation to the target distribution, i.e., $\hat{\pi}(\mathbf{X})$. After the initial run, α is fixed at some value between 0 and 1 and the mixture kernel in (1.3.1) is used. In other words, the original MH sampler is run for the first N_1 iterations, and the PAMH sampler that mixes the original proposal distribution with an approximation to the target distribution is run for additional N_2 iterations. The number of iterations for the initial run, N_1 is usually set to a reasonable small number.

If the dimension of $\mathbf{X}$ is small, the empirical distribution, $\hat{\pi}(\mathbf{X})$ can be computed by discretizing the space into sufficiently small pixels and calculating the proportion of the initial N_1 draws that fall into each pixel. In some cases the approximation can be improved by discarding an initial burn-in from the N_1 draws. In this way, we approximate $\hat{\pi}(\mathbf{X})$ with a step function that is sampled by first selecting a pixel according to the empirical pixel probabilities and then sampling uniformly within the pixel. To get a more precise approximation to the target distribution, we can use more sophisticated non-parametric density estimation, such as kernel density estimation. This strategy is more efficient in higher dimensions and even in lower dimensions can improve the empirical approximation. Of course if the target distribution is discrete, no pixeling or smoothing is necessary.

Detailed balance is satisfied by the mixture transition kernel in (1.3.1) because

$$\begin{aligned}\pi(\mathbf{X}^{(t)})\mathcal{K}_+(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}) &= \alpha \min\left\{\pi(\mathbf{X}^{(t)})p_1(\mathbf{X}^{(t+1)}|\mathbf{X}^{(t)}), \pi(\mathbf{X}^{(t+1)})p_1(\mathbf{X}^{(t)}|\mathbf{X}^{(t+1)})\right\} \\ &+ (1-\alpha)\min\left\{\pi(\mathbf{X}^{(t)})\hat{\pi}(\mathbf{X}^{(t+1)}), \pi(\mathbf{X}^{(t+1)})\hat{\pi}(\mathbf{X}^{(t)})\right\}\end{aligned}$$

is a symmetric function in terms of $\mathbf{X}^{(t)}$ and $\mathbf{X}^{(t+1)}$. Thus the resulting Markov chain is reversible with respect to $\pi(\mathbf{X})$. The PAMH sampler uses a mixture of two MH samplers rather than a single MH sampler with the mixture of two proposal distributions because the mixture of two MH samplers requires only the computation of one proposal distribution at each iteration. Thus, the PAMH sampler reduces the number of evaluations of $\pi(\mathbf{X})$. This significantly improves the overall computation in the example of Section 1.4 where this evaluation is computationally costly. See Tierney (1998) for a comparison of the asymptotic efficiency of these two strategies.

To illustrate the advantage of the PAMH sampling strategy, we introduce a simple example where both the Gibbs sampler and the MH sampler exhibit slow convergence. Consider the following bivariate distribution which has normal conditional distributions but is not a bivariate normal distribution,

$$p(X, Y) \propto \exp\left\{ -\frac{1}{2}\Big(8X^2Y^2 + X^2 + Y^2 - 8X - 8Y\Big)\right\}. \tag{1.3.2}$$

This is a special bimodal case of a parameterized family of distributions derived by Gelman and Meng (1991).

A Gibbs sampler can easily be constructed to simulate from (1.3.2):

Step 1: Draw X from $p(X|Y)$, where $X|Y \sim \mathrm{N}(4/(8Y^2+1), 1/(8Y^2+1))$,

Step 2: Draw Y from $p(Y|X)$, where $Y|X \sim \mathrm{N}(4/(8X^2+1), 1/(8X^2+1))$.

An MH sampler can be also constructed. An independent bivariate normal distribution is a simple choice for the proposal distribution. Given the current state $(X^{(t)}, Y^{(t)})$, we generate a proposal state $(X', Y') = (X^{(t)} + \epsilon_1, Y^{(t)} + \epsilon_2)$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$ for $i = 1, 2$, and accept the proposal state with probability $p(X', Y')/p(X^{(t)}, Y^{(t)})$. In this case, τ is a tuning parameter that is chosen in advance and affects the convergence of the resulting sampler. Too small of a value of τ produces small jumps which are often accepted but lead to a Markov chain that moves slowly. On the other hand, when τ is too large, the sampler will propose large jumps that are too often rejected. Thus, it is important to find a reasonable choice of τ between these two extremes. For illustration, we run three MH samplers, using $\tau = 0.5$, 1, and 2.

We ran the Gibbs sampler and the MH sampler with three different values of τ for 20000 iterations each. Convergence of the four samplers is described in the first four rows of Figure 1.3. The first two columns of Figure 1.3 show the trace plot of the last 5000 iterations and autocorrelation function computed using the last 10000 iterations of each chain of X. The last column compares each simulated marginal distribution of X based on the last 10000 draws (histogram) with the target distribution (solid line). Figure 1.3 illustrates the

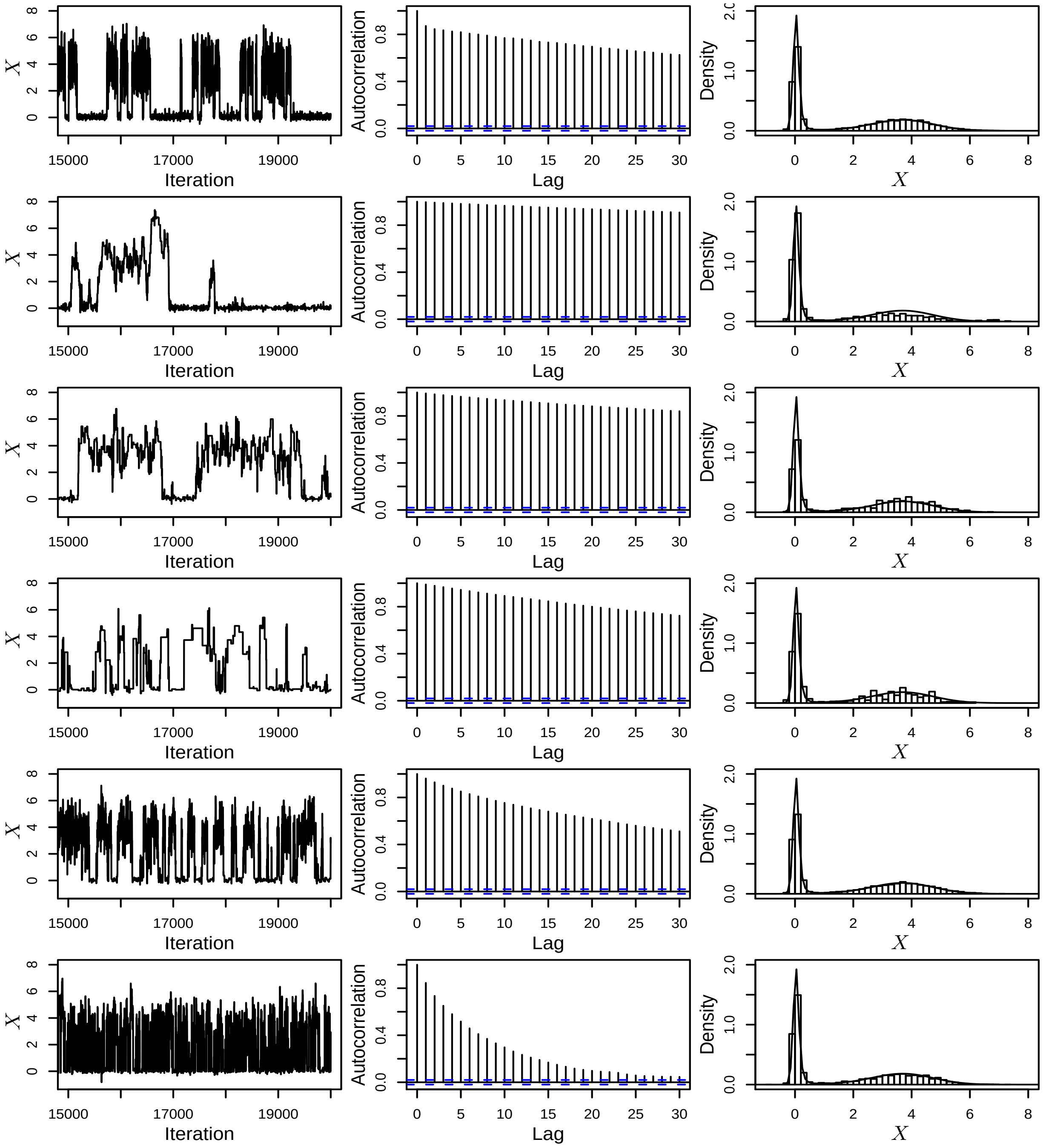

Figure 1.3: Comparing six MCMC samplers constructed for simulating a bivariate distribution that has normal conditional distributions but is not a bivariate normal distribution. The rows correspond to the Gibbs sampler, the MH sampler with $\tau = 0.5$, 1, and 2, the MH within PCG sampler run with $\tau = 1$, and the PAMH within PCG sampler run with $\tau = 1$ and $\alpha = 0.5$, respectively. The first column shows trace plots of the last 5000 iterations of each chain; the second column contains autocorrelation functions of the last 10000 draws; and the last column compares simulated marginal distributions of X based on the last 10000 draws (histograms) with the true marginal distribution (solid line).

slow convergence and high autocorrelation of all four MCMC samplers. Because of this poor convergence, the simulated marginal distributions do not approximate the target distribution precisely. Among the three MH samplers, the choice of $\tau = 1$ results in the best convergence.

We can use a PCG sampler to improve convergence as described in Section 1.2. In particular, if we could eliminate the conditioning on Y in Step 1 of the Gibbs sampler, we could generate independent draws by iterating between

Step 1: Draw X from $p(X)$, where

$$p(X) \propto \frac{1}{\sqrt{8X^2+1}} \exp\left\{ -\frac{1}{2}\Big(X^2 - 8X - \frac{16}{8X^2+1} \Big) \right\},$$

Step 2: Draw Y from $p(Y|X)$.

This PCG sampler would be a blocked one-step sampler if we could simulate $p(X)$ directly. Because we cannot, we consider indirect sampling in Step 1 using MH with a normal proposal distribution, i.e., $X'|X^{(t)} \sim \mathrm{N}(X^{(t)}, \tau^2)$. This results in an MH within PCG sampler that we implement with $\tau = 1$. To further improve convergence, we use PAMH sampling in Step 1. This results in a PAMH within PCG sampler, that we also implement with $\tau = 1$ and with the mixture proportion at $\alpha = 1$ for the first 1000 iterations and $\alpha = 1/2$ for the next 19000 iterations. We discretize the space into 200 bins equally spaced between -1 and 8, and approximate $\hat{\pi}(X)$ using bin proportions from the first 1000 iterations. The last two rows of Figure 1.3 illustrate the convergence of the MH and PAMH within Gibbs samplers, respectively. The PAMH within PCG sampler exhibits a dramatic improvement over the five other MCMC samplers.

1.4 Spectral analysis in high-energy astrophysics

We now turn to the illustration of PCG and PAMH in spectral analysis in high-energy astrophysics. In recent years technological advances have dramatically increased the quality and quantity of data available to astronomers. Instrumentation is tailored to data-collection challenges associated with specific scientific goals. These instruments provide massive new surveys resulting in new catalogs containing terabytes of data, high resolution spectrography and imaging across the electromagnetic spectrum, and incredibly detailed movies of dynamic and explosive processes in the solar atmosphere. The spectrum of new instruments is helping make impressive strides in our understanding of the universe, but at the same time generating massive data-analytic and data-mining challenges for scientists who study the data.

High-energy astrophysics is concerned with Ultraviolet-rays, X-rays, and γ-rays, i.e. photons with energies of a few electron volts (eV), a few keV, or greater than an MeV, respectively. Roughly speaking, the production of high-energy electromagnetic waves requires temperatures of millions of degrees and signals the release of deep wells of stored energy such as those in very strong magnetic fields, extreme gravity, explosive nuclear forces, and shock waves in hot plasmas. Thus, X-ray and γ-ray telescopes can map nearby stars with active magnetic fields, the remnants of exploding stars, areas of star formation, regions near the event horizon of a black hole, very distant turbulent galaxies, or even the glowing gas embedding a cosmic cluster of galaxies. The distribution of the energy of the electromagnetic emissions is called the spectrum and gives insight into these deep energy wells: the composition, density, and temperature/energy distribution of the emitting material; any chaotic or turbulent flows; and the strengths of the magnetic, electrical, or gravitational fields.

In this chapter we focus on X-ray spectral analysis. A typical spectrum can be formulated as a finite mixture distribution composed of one or more continuum terms, which are smooth functions across a wide range of energies, and one or more emission lines, which are local features highly focused on a narrow band of energies. For simplicity we focus on a the case when there is one continuum term and one emission line. Because of instrumental constraints, photons are counted in a number of energy bins. These photon counts are modeled as an inhomogeneous Poisson process with expectation in energy bin j modeled as

$$\Lambda_j(\boldsymbol{\theta}) = f_j\big(\boldsymbol{\theta}^C\big) + \lambda\pi_j\big(\mu, \sigma^2\big), \tag{1.4.1}$$

where $\boldsymbol{\theta}$ is the set of model parameters, $f_j\big(\boldsymbol{\theta}^C\big)$ is the expected continuum count in bin j, $\boldsymbol{\theta}^C$ is the set of free parameters in the continuum model, λ is the total expected line count, $\pi_j\big(\mu, \sigma^2\big)$

is the proportion of an emission line with location μ and width σ^2 falls into bin j. Various emission line profiles such as Gaussian distributions, t-distributions, and delta functions can be used to derive the emission line bin proportions as a function of μ and σ^2. In this chapter we focus on the use of a delta function which is parameterized only in terms of μ.

Due to instrumental constraints, the photon counts are subject to blurring of the individual photon energies, stochastic censoring with energy dependent rates, and background contamination. To account for these processes, we embed the scientific model (1.4.1) in a more complex observed-data model. In particular, the observed photon counts in detector channel l are modeled with a Poisson distribution,

$$Y_{\mathrm{obs}\,l} \sim \mathrm{Poisson}\left(\sum\nolimits_j M_{lj}\Lambda_j(\boldsymbol{\theta})u_j(\boldsymbol{\theta}^A) + \theta_l^B\right), \tag{1.4.2}$$

where 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 observed, and θ_l^B is the expected background counts in channel l. A multilevel model can be constructed to incorporate both the finite mixture distribution of the spectral model and the complexity of the data generation mechanism. Using a missing-data/latent-variable setup, a standard Gibbs sampler can be constructed to fit the model (van Dyk et al., 2001; van Dyk and Kang, 2004).

1.5 Efficient MCMC in spectral analysis

As a specific example, we consider data collected using the *Chandra X-ray Observatory* in an observation of the quasar PG1634+706 (Park et al., 2008). Quasars are extremely distant astronomical objects that are believed to contain supermassive black holes with masses exceeding that of our Sun by a million fold. Because quasars are very distant, the universe was a fraction of its current age when the light we now see as a quasar was emitted. They are also very luminous and therefore give us a way to study the "young" universe. Thus, the study of quasars is important for cosmological theory and their spectra can give insight into their composition, temperature, distance, and velocity.

(a) Parent sampler	(b) Marginalize	(c) Permute	(d) Trim	(e) Block
$p(\mathbf{Y}_{\rm mis}\vert\boldsymbol{\psi},\mu,\mathbf{Y})$ $p(\boldsymbol{\psi}\vert\mathbf{Y}_{\rm mis},\mu,\mathbf{Y})$ $p(\mu\vert\mathbf{Y}_{\rm mis},\boldsymbol{\psi},\mathbf{Y})$	$p(\mathbf{Y}^{\star}_{\rm mis}\vert\boldsymbol{\psi},\mu,\mathbf{Y})$ $p(\boldsymbol{\psi}\vert\mathbf{Y}_{\rm mis},\mu,\mathbf{Y})$ $p(\mathbf{Y}_{\rm mis},\mu\vert\boldsymbol{\psi},\mathbf{Y})$	$p(\mathbf{Y}^{\star}_{\rm mis},\mu\vert\boldsymbol{\psi},\mathbf{Y})$ $p(\mathbf{Y}_{\rm mis}\vert\boldsymbol{\psi},\mu,\mathbf{Y})$ $p(\boldsymbol{\psi}\vert\mathbf{Y}_{\rm mis},\mu,\mathbf{Y})$	$p(\mu\vert\boldsymbol{\psi},\mathbf{Y})$ $p(\mathbf{Y}_{\rm mis}\vert\boldsymbol{\psi},\mu,\mathbf{Y})$ $p(\boldsymbol{\psi}\vert\mathbf{Y}_{\rm mis},\mu,\mathbf{Y})$	$p(\mathbf{Y}_{\rm mis},\mu\vert\boldsymbol{\psi},\mathbf{Y})$ $p(\boldsymbol{\psi}\vert\mathbf{Y}_{\rm mis},\mu,\mathbf{Y})$

Figure 1.4: Transforming the parent Gibbs sampler into PCG I. The PCG I sampler in (e) is constructed by partially collapsing out the missing data and corresponds to a blocked version of its parent sampler in (a).

We are particularly interested in an emission feature of the quasar's spectrum, which is a narrow Fe-K-alpha emission line whose location indicates the ionization state of iron in the emitting plasma. To fit the location of a narrow emission line, we model the emission line with a delta function, so that the entire line falls within one data bin.

Unfortunately the standard Gibbs sampler described in van Dyk et al. (2001) breaks down when delta function are used to model emission lines. Using the method of data augmentation, the standard Gibbs sampler is constructed in terms of missing data that include unobserved Poisson photon counts with expectation given in (1.4.1) and unobserved mixture indicator variables for the mixture given in (1.4.1). To see why the standard sampler fails, we examine how the mixture indicator variables and line location are updated. The components of the mixture indicator variable are updated for each photon within each bin as a Bernoulli variable with probability of being from an emission line,

$$\frac{\lambda\pi_j(\mu)}{f_j(\boldsymbol{\theta}^C)+\lambda\pi_j(\mu)} \tag{1.5.1}$$

in energy bin j. (We supress the width, σ^2, of the emission line $\pi_j(\mu,\sigma^2)$, because delta functions have no width. Because the delta function is contained in a single bin, $\pi_j(\mu)=1$ if μ is within bin j, and 0 otherwise. This means that the probability in (1.5.1) is zero for all energy bins except the one containing the current line location, μ. Thus, in each iteration of the standard Gibbs sampler, the only bin that can have photons attributed to the emission line is the bin that contains the current iterate of the line location. When the line location is updated using the photons attributed to the emission line, it is necessarily set to the same value as the current iterate. The line location is thus fixed throughout the iterations at the midpoint of a bin containing its starting value. As a result, the standard Gibbs sampler is not positive recurrent and does not converge to the target distribution. Although this sampler works fine with emission lines of appreciable with, it fails for delta functions (Park and van Dyk, 2009; van Dyk and Park, 2004).

To understand the computational challenges of fitting this model we must go into some of the technical details of the Gibbs sampler. Let $\mathbf{Y} = \{Y_{\mathrm{obs}\,l}\}$ be the observed data modeled in (1.4.2), $\mathbf{Y}_{\mathrm{mis}} = (\mathbf{Y}_{\mathrm{mis}\,1}, \mathbf{Y}_{\mathrm{mis}\,2})$ be a collection of missing data, where $\mathbf{Y}_{\mathrm{mis}\,1}$ denotes the unobserved Poisson photon counts with expectation given in (1.4.1) and $\mathbf{Y}_{\mathrm{mis}\,2}$ the unobserved mixture indicator variable for each photon under the finite mixture model given in (1.4.1), μ be the delta function line location, and $\boldsymbol{\psi}$ be the model parameters other than μ. To sample from the target distribution $p(\mathbf{Y}_{\mathrm{mis}}, \boldsymbol{\psi}, \mu|\mathbf{Y}_{\mathrm{obs}})$, the parent Gibbs sampler is constructed by iteratively sampling from its conditional distributions, as shown in Figure 1.4(a). This a special case of the "standard" Gibbs sampler discussed above and derived by van Dyk et al. (2001). We devise a PCG sampler to improve the convergence of the parent Gibbs sampler. To construct a PCG sampler, we eliminate the conditioning on all or some of $\mathbf{Y}_{\mathrm{mis}}$ in the step that updates μ. In combination with PAMH, this results in three new efficient samplers.

First, PCG I is constructed by eliminating the conditioning on all of $\mathbf{Y}_{\mathrm{mis}}$ in the step that updates μ. Figure 1.4 shows how the parent Gibbs sampler shown in (a) is transformed into PCG I shown in (e) by partially collapsing $\mathbf{Y}_{\mathrm{mis}}$ out of the sampler. In Figure 1.4(b), $\mathbf{Y}_{\mathrm{mis}}$ is moved from being conditioned upon to being sampled in the step that updates μ. The steps are then permuted in Figure 1.4(c) in order to make one of the two updates of $\mathbf{Y}_{\mathrm{mis}}$ redundant. This allows us to trim the unused sample of $\mathbf{Y}^{\star}_{\mathrm{mis}}$ from the first step in Figure 1.4(d). Finally, we can combine the first two steps into a single sampling of $p(\mathbf{Y}_{\mathrm{mis}}, \mu|\boldsymbol{\psi}, \mathbf{Y})$. The resulting PCG sampler in Figure 1.4(e) is a blocked version of the parent Gibbs sampler.

Because the likelihood function is flat in each bin as a function of μ, we can consider μ to be a discrete parameter. Its distribution given ψ and $\mathbf{Y}_{\mathrm{obs}}$ follows a multinomial distribution with values corresponding to the data bins and probability vector proportional to the product of the Poisson distributions given in (1.4.2). (We use a flat prior distribution

(a) Parent sampler	(b) Marginalize	(c) Permute	(d) Trim
$p(\mathbf{Y}_{\text{mis}}\|\boldsymbol{\psi}, \mu, \mathbf{Y})$	$p(\mathbf{Y}_{\text{mis}\,1}, \mathbf{Y}^{\star}_{\text{mis}\,2}\|\boldsymbol{\psi}, \mu, \mathbf{Y})$	$p(\mathbf{Y}^{\star}_{\text{mis}\,2}, \mu\|\mathbf{Y}_{\text{mis}\,1}, \boldsymbol{\psi}, \mathbf{Y})$	$p(\mu\|\mathbf{Y}_{\text{mis}\,1}, \boldsymbol{\psi}, \mathbf{Y})$
$p(\boldsymbol{\psi}\|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y})$	$p(\boldsymbol{\psi}\|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y})$	$p(\mathbf{Y}_{\text{mis}}\|\boldsymbol{\psi}, \mu, \mathbf{Y})$	$p(\mathbf{Y}_{\text{mis}}\|\boldsymbol{\psi}, \mu, \mathbf{Y})$
$p(\mu\|\mathbf{Y}_{\text{mis}}, \boldsymbol{\psi}, \mathbf{Y})$	$p(\mathbf{Y}_{\text{mis}\,2}, \mu\|\mathbf{Y}_{\text{mis}\,1}, \boldsymbol{\psi}, \mathbf{Y})$	$p(\boldsymbol{\psi}\|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y})$	$p(\boldsymbol{\psi}\|\mathbf{Y}_{\text{mis}}, \mu, \mathbf{Y})$

Figure 1.5: Transforming the parent Gibbs sampler into PCG II. This PCG sampler is constructed by partially collapsing out part of the missing data, is composed of a set of incompatible conditional distributions, and is not a blocked version of the sampler in (a).

on μ.) This probability vector must be computed at each iteration of the sampler, which is computational expensive owing to the large dimension of the blurring matrix $\mathbf{M} = \{M_{lj}\}$ and the large number of energy bins. Because sampling from $p(\mu|\boldsymbol{\psi}, \mathbf{Y}_{\text{obs}})$ is so expensive, we consider a second PCG sampler that avoids this update. In particular, we consider eliminating only the mixture indicator variables, $\mathbf{Y}_{\text{mis}\,2}$, from the step that updates μ in the derivation of PCG II. Because the resulting update for μ conditions on $\mathbf{Y}_{\text{mis}\,1}$, $\boldsymbol{\psi}$, and $\mathbf{Y}_{\text{obs}}$, its distribution is multinomial with probability vector proportional to the product of the Poisson distributions given in (1.4.1). This distribution does not involve the large dimension of the blurring matrix and is much quicker to compute.

Figure 1.5 illustrates the construction of PCG II which is identical to that of PCG I except that only $\mathbf{Y}_{\text{mis}\,2}$ is moved from being conditioned upon to being sampled in the step that updates μ. Unlike PCG I, however, PCG II consists of a set of incompatible conditional distributions and does not correspond to a blocked version of the parent Gibbs sampler, see Figure 1.5(d). Due to the greater degree of collapsing, PCG I is expected to have better convergence characteristics than PCG II, see van Dyk and Park (2008). The tradeoff, however, is that an iteration of PCG II is much faster to compute than one of PCG I. A numerical comparison of the two samplers appears below.

In order to further improve computational speed, we consider using an MH step to update μ in an effort to avoid the expense of computing a lengthy probability vector at each iteration. This requires us to evaluate only two components of the multinomial probability vector, the components corresponding to the current value and the proposed value of the line location.

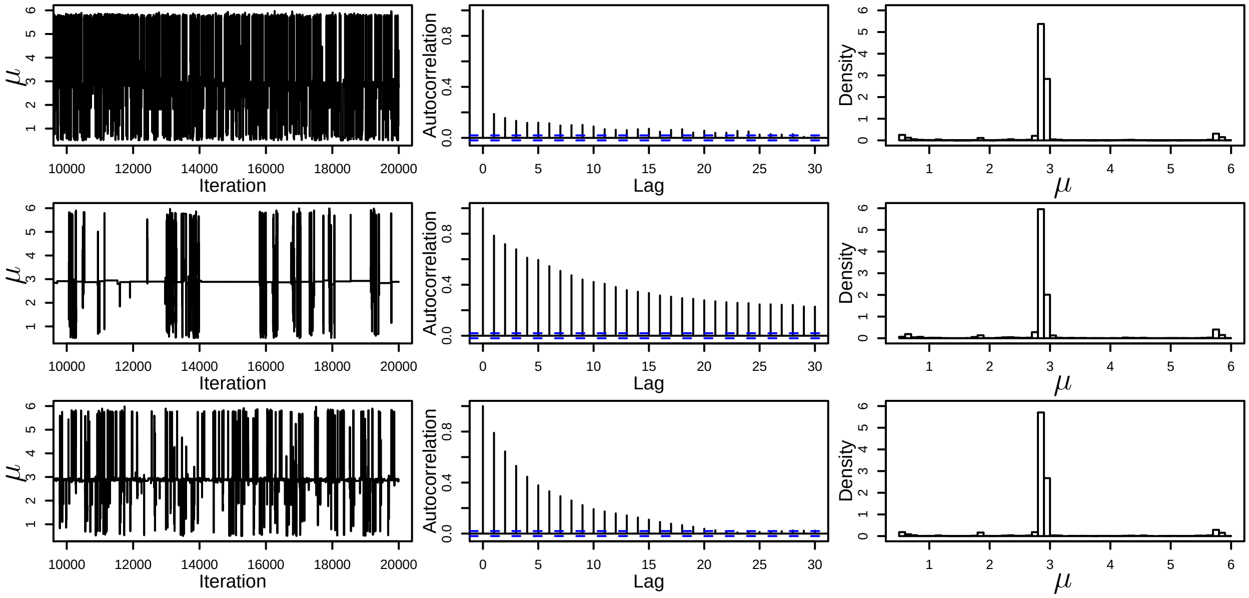

Figure 1.6: Comparing three efficient MCMC samplers constructed for spectral analysis. The rows correspond to the PCG I, PCG II, and the PAMH within PCG I samplers, respectively. The PAMH sampler was run with $\tau = 2$ and $\alpha = 0.5$. The first column shows trace plots of the last 10000 iterations of each chain; the second column contains autocorrelation plots of the last 10000 draws; and the last column presents a simulated marginal distributions of the line location μ based on the last 10000 draws. Although PCG I mixes more quickly than PAMH within PCG I, it requires about 50 times more computing time for 20000 draws.

Although this can significantly reduce computation time per iteration, it is difficult to find a good proposal distribution because the posterior distribution of the line location can be highly multimodal. A proposal distribution with relatively high variance is required to allow jumping among the modes, but this leads to many rejected proposals in regions between the modes. To improve the convergence of the MH sampler, we consider the PAMH sampling strategy to update μ. We use a normal proposal distribution with standard deviation , $\tau = 2$, for the initial MH within Gibbs sampler, which allows jumps across the range of energies (5.5 keV). After 1000 iterations we use an estimate of the discrete marginal posterior distribution of μ in a second MH transition kernel that we mix with the original kernel in a fifty-fifty mixture for the remaining draws of the PAMH within PCG I sampler.

The convergence characteristics of the three samplers are compared in Figure 1.6. Each of the three samplers is run for 20000 iterations. The rows correspond to the PCG I, PCG II, and the PAMH within PCG I samplers, respectively; the columns correspond to the trace and autocorrelation plots of the last 10000 iterations, and the simulated marginal posterior distribution of the line location based on the last 10000 iterations, respectively.

Comparing the first two columns of Figure 1.6 illustrates that PCG I has the quickest convergence among the three PCG samplers, but the other two PCG samplers also have fairly fast convergence. When sampling μ from its multinomial conditional distribution, however, PCG I requires significantly more computing time than PCG II, which in turn takes significantly more time than PAMH within PCG I . The total computing time for 20000 iterations of PCG I, PCG II, and PAMH within PCG I on a UNIX machine is 15 hours 35 minutes, 1 hour 55 minutes, and 19 minutes, respectively. PAMH within PCG I offers a dramatic improvement in computation time with very good mixing. The fast computing allows more interactive model fitting and enables us to easily run chains longer when necessary.

1.6 Conclusion

In this chapter we illustrate the use of two computational techniques to improve the performance of MCMC samplers in a particular example from high-energy astrophysics. There are many other computational techniques and variants on MCMC samplers that can be applied to the myriad of complex model fitting challenges in astronomy. Puzzling together the appropriate computational and statistical methods for the numerous outstanding data-analytic problems offers a gold mine for methodological researchers. We invite all interested readers to join us in this seemingly endless but ever enjoyable endever!

Acknowledgments

The authors gratefully acknowledge funding for this project partially provided by NSF grants DMS-04-06085 and SES-05-50980 and by NASA Contract NAS8-39073 and NAS8-03060 (CXC). The spectral analysis example stems from work of the California-Harvard Astrostatistics Collaboration (CHASC, URL `www.ics.uci.edu/∼dvd/astrostats.html`).

Bibliography

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B, Methodological*, 55:25–37.

Connors, A. and van Dyk, D. A. (2007). How to win with non-Gaussian data: Poisson goodness-of-fit. In *Statistical Challenges in Modern Astronomy IV* (Editors: E. Feigelson and G. Babu), Astronomical Society of the Pacific Conference Series, 371:101–117, San Francisco.

DeGennaro, S., von Hippel, T., Jefferys, W. H., Stein, N., van Dyk, D. A., and Jeffery, E. (2008). Inverting Color-Magnitude Diagrams to Access Precise Cluster Parameters: A New White Dwarf Age for the Hyades. *Submitted to The Astrophysical Journal.*

Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A. (2004). An image reconstruction technique with error estimates. *The Astrophysical Journal*, 610:1213–1227.

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 Meng, X.-L. (1991). A note on bivariate distributions that are conditionally normal. *The American Statistician*, 45:125–126.

Gregory, P. C. (2005). A Bayesian Analysis of Extrasolar Planet Data for HD 73526. *The Astrophysical Journal*, 631:1198–1214.

Kashyap, V. and Drake, J. J. (1998). Markov-Chain Monte Carlo Reconstruction of Emission Measure Distributions: Application to Solar Extreme-Ultraviolet Spectra. *The Astrophysical Journal*, 503:450–466.

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., 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.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274.

Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, 59:511–567.

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

Park, T. and van Dyk, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Technical Report*.

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*, page under review.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8:1–9.

van Dyk, D. and Park, T. (2004). Efficient EM-type algorithms for fitting spectral lines in high-energy astrophysics. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: Contributions by Donald Rubin's Statistical Family* (Editors: A. Gelman and X.-L. Meng). Wiley & Sons, New York.

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. (2000). Nesting EM algorithms for computational efficiency. *Statistical Sinica*, 10:203–225.

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

van Dyk, D. A., DeGennaro, S., Stein, N., Jefferys, W. H., and von Hippel, T. (2009). Statistical Analysis of Stellar Evolution. *Submitted to The Annals of Applied Statistics*.

van Dyk, D. A. and Kang, H. (2004). Highly structured models for spectral analysis in high-energy astrophysics. *Statistical Science*, 19:275–293.

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.

van Dyk, D. A. and Meng, X.-L. (2008). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science*, under review.

Yu, Y. (2005). *Three Contributions to Statistical Computing*. PhD thesis, Department of Statistics, Harvard University.