Inference and Efficient Computation for Highly Structured Models with Applications

A thesis presented

by

Taeyoung Park

to

The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts

May 2006

Thesis Advisors: Professors David A. van Dyk and Xiao-Li Meng Taeyoung Park

Inference and Efficient Computation for Highly Structured Models with Applications

Abstract

This thesis presents efficient computational and inferential statistical techniques for highly structured and complex models with various applications. 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, Markov chain Monte Carlo samplers have been spotlighted because of their power to effectively generate samples from a high-dimensional distribution. However, their sometimes slow convergence has been a long standing complaint, especially when the complex models are fitted. In this thesis, we provide useful techniques to achieve quicker convergence with additional, but not substantial, human effort. In particular, we develop efficient Markov chain Monte Carlo samplers by generalizing the composition of conditional distributions used to construct the samplers. This allows the samplers to be constructed using a set of incompatible conditional distributions. Such incompatibility has been generally avoided in the construction of Markov chain Monte Carlo samplers because the resulting convergence properties are not well understood. We, however, capitalize on the set of incompatible conditional distributions to improve the convergence characteristics of a Markov chain Monte Carlo sampler, while maintaining the transition kernel of the Markov chain constructed by the sampler. This thesis mainly explores the utility of our strategy in a wide range of applications involving computational challenges.

Contents

List of Figures

List of Tables

Acknowledgments

It has been a long journey to write this dissertation. This would not have been possible without my advisor, Professor David A. van Dyk. Ever since I embarked upon my graduate study at Harvard, David has been my "American Idol" not because he is just a great singer but because he has always shown to me what a topnotch mentor is. I would like to sincerely and gratefully thank David for his continuous support, advice, and efforts that helped me survive at Harvard (and in the Unites States). He has been always willing to delve into the details of my work, give suggestions that guided my research in a better direction, and consult both academic and non-academic issues I encountered.

Special thanks also go to Professor Xiao-Li Meng who is my academic grandfather as well as another academic advisor at Harvard since David moved to California. He taught me the distinction between a statistician and a mathematician, and exemplified an excellent statistician. His intuition which I have admired helped me look through a problem rather than look at it.

I want to thank Professor Donald Rubin for valuable comments on my research and his collaboration on the German project. I was honored to work with Don who is not only a great statistician but also my academic great-grandfather. At Harvard, I was very lucky to have a unique experience to work with three generations in a row. I also sincerely appreciate the rest of the statistics faculty for their teaching and help during my graduate study at Harvard. I thank Betsey Cogswell, Vanessa LaBarca, Dale Rinkel, and Maureen Stanton for their administrative support.

I would also like to thank my primary research collaborators, Vinay Kashyap, Aneta Siemiginowska, and Andreas Zezas, for their contributions and collaborative efforts to various projects. My thanks also go to the rest of members of the California-Harvard AstroStatistics Collaboration (CHASC): Jim Chiang, Alanna Connors, Jeff Scargle, Nondas Sourlas, and Yaming Yu.

My peer (past or present) graduate students, Sue Ryung Chang, Hae Mi Choi, Gopi Goswami, Jim Greiner, Hongkai Ji, Hosung Kang, Dennis Lam, Chester Lee, Chenxin Li, Jingchen Liu, Xiaohong Shen, and Tingting Zhang made my academic life lively and enjoyable. I also acknowledge the friendship with Hui Jin and Qing Zhou who arrived at Harvard in the same year as me, and my officemates, Xiaodan Fan and Wei Zhang. I also thank old friends and the rest of the students in the Harvard statistics department who do not need to be named to know their importance to me. Without them, my life at Harvard would be like comedy without humor or, miserably, astrostatistics seminar without Chinese catering.

I am also grateful to Professors Jon Anderson and Engin Sungur at University of Minnesota, Morris, who urged me to keep on studying aboard while I was studying as an exchange student at University of Minnesota. Their advise and encouragements made me seriously think of being a statistician as a career.

I am indebted to my beloved parents, Kiho Park and Soonhee Lee, for their inestimable love and full support. Their encouragements and advice made all the difference and helped me follow my dream.

Finally, but the most importantly, I would like to thank my lovely wife, Youjin Je, for her true love and patience. She has had to endure my bad moods and share with me the pressures associated with my research and graduate study at Harvard. Without her support and help, I would not be able to complete this dissertation.

The research involved in this dissertation was partially supported by NSF grant DMS-04-06085 and by NASA Contracts NAS8-39073 and NAS8-03060 (Chandra X-ray Center).

Chapter 1

Introduction

1.1 A Harmony of Dissonance

A Gibbs sampler (Geman and Geman, 1984) is a simple but powerful sampling technique used to effectively sample from a (high-dimensional) joint distribution by iteratively sampling from its conditional distributions. Such a Gibbs sampler is generally expected to be composed of compatible conditional distributions that are defined on the same space. Despite its simplicity to implement and describe, however, the Gibbs sampler is criticized for its sometimes slow convergence, especially when it is used to fit complex models. In Chapter 2, we present partially marginalized Gibbs sampling strategies that improve the convergence characteristics of the Gibbs sampler by capitalizing on a set of incompatible conditional distributions that has inconsistent dependence structure. Such incompatibility has been simply avoided in the construction of Gibbs samplers because the resulting convergence properties are not well understood. We, however, introduce three prescriptive tools (marginalization, permutation, and trimming) which allow us to transform a Gibbs sampler into a partially marginalized Gibbs sampler with known stationary distribution and fast convergence. That is, one may create dis-

sonant sound using incompatible instruments, but the coordination of a conductor can make the discordant sound even more harmonious than harmonious sound using compatible instruments.

As an illustration, we apply our partially marginalized Gibbs sampling strategies to a variety of examples with complex models presented in Chapters 2, 3, and 4. In particular, Chapter 3 describes a highly structured multilevel spectral model to account for the distribution of the energies of photons emitted from an astronomical source with data contaminations by several non-trivial physical processes. The shape and structure of this distribution gives clues as to the composition, density, temperature, relative velocity, and distance of the source. Thus, spectral analysis is key to our understanding of the physical environment and structures of astronomical sources, the processes and laws which govern the births and deaths of planets, stars, and galaxies, and ultimately the structure and evolution of the universe. From a statistical point of view, a typical stellar spectrum can be formulated as a finite mixture distribution composed of one (or more) continuum terms and a set of emission line terms. While the continuum describes the general shape of a spectrum, each emission line represents a positive aberration from the continuum in a narrow band of energies. Emission lines are used to model the emission resulting from electrons falling to a lower energy shell in a particular ion. Thus, emission lines are important in the investigation of the composition of a source. The Doppler shift of the location of a known spectral line (such as a particular hydrogen line) can also be used to determine the relative velocity of a source. Thus, determining the precise location of emission lines is a critical task. In Chapter 3, we focus on a single narrow emission line that can be modeled with a Gaussian distribution or a delta function. Spectral data are typically contaminated by several non-trivial physical processes including non-homogeneous stochastic censoring, blurring of photon energies, and background contamination. Accounting for these processes leads us to construct a highly structured multilevel spectral model that is

formulated in terms of several layers of missing data. We devise several partially marginalized Gibbs samplers to fit the spectral model with narrow emission lines. We also test our models, methods, and computational strategies to simulated data under four different scenarios and apply them to the X-ray spectrum of the high redshift quasar, PG1634+706.

1.2 Joint Multiple Imputations

It is quite common to measure economic and demographic data on several geographical partitions of a fixed region, e.g., unemployment rates may be measured on different levels of political partitions to see the variation across the region. We focus on the case where the geographical partitions are not aligned, so that interpolating the data from one level of resolution to another is not obvious. Moreover, a response variable of interest can be multivariate and all the components of the response variable may not be observed on the same partitions. In Chapter 4, we develop three joint imputation models for non-nested data of this sort and illustrate the implementation of the imputation procedure for these models. In particular, we introduce the bivariate Gaussian model, bivariate lognormal model, and Poisson regression model for the non-nested data. To create joint imputations from these models, we consider three imputation methods. In the first method, we formulate a set of conditional distributions and iteratively impute one component of a multivariate response variable given the other components. The second method formulates the joint imputation procedure in terms of marginal and conditional distributions. Because of the misalignment of the partitions, however, it is not even feasible to write the correct marginal distribution and thus it is replaced with an incoherent marginal distribution. Lastly, the third method completely imputes one component of the multivariate response variable from the incoherent marginal distribution, and sequentially imputes the other components using each imputation

of the first component as an another covariate. We demonstrate our imputation methods with both simulated data and real German unemployment data.

1.3 BEHR: Bayesian Estimation of Hardness Ratios

Hardness ratios are summary statistics commonly used to characterize the spectra of faint X-ray sources whose low counts data prevent sophisticated spectral fitting. The classical approach to computing the hardness ratios uses a simple statistical technique based on the method of moments. The error bars associated with the classical method are computed using the delta method. Thus, the classical method relies on the Gaussian assumptions, so that it fails to provide realistic or reliable estimates for the hardness ratios of faint X-ray sources and to account for the Poissonian nature of low counts. In Chapter 5, we propose a new approach to modeling hardness ratios in Poisson limits and present statistically coherent Bayesian methods for computing the hardness ratios and their associated errors. Using the sophisticated Bayesian approaches, we calculate hardness ratios after explicitly modeling the detected photons as independent Poisson random variables. With a survey of X-ray sources, the hardness ratios can be used to investigate the spectral relationship and to cluster the X-ray sources based on the spectral characteristics. In this case, we relax the assumptions made for a single source and devise a hierarchical mixture model. Our simulation studies demonstrate the clear advantages of the Bayesian methods over the classical method and illustrate how to infer the spectral shape of an X-ray source based on the hardness ratios. Our clustering model is also applied to real Galaxy sources to answer a scientific question of interest.

Chapter 2

Using Incompatibility to Build Fast Gibbs Samplers

2.1 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 popularized Bayesian methods. More complex models that explicitly aim to incorporate application-specific stochastic features of a data generation mechanism are becoming ever 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, the topic of this chapter, it is well known that blocking or grouping steps (Liu *et al.*, 1994), nesting steps (van Dyk, 2000b), 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 a sampler. 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; Gelman *et al.*, 2006). The EM algorithm (Dempster *et al.*, 1977) can be used to compute the posterior mode of the parameters of a model 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.

Marginalization methods offer an example of the relationship between efficient EM-type algorithms and methods for improving the convergence of the Gibbs sampler. Marginalization 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 sampler is built. Liu *et al.* (1994) demonstrated the advantage of a special case of this strategy that they called collapsing. 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, i.e., 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 from 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 marginalization strategy described in the previous paragraph, both the ECME and AECM algorithms can be described as *partially marginalized 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 chapter. 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.

The Gibbs sampler begins with a joint posterior distribution of the unknown quantities and updates groups of these quantities by sampling them from their conditional distributions under the joint posterior distribution. The partially marginalized Gibbs (PMG) sampler replaces some of these conditional distributions with conditional distributions of some *marginal distributions* of the joint posterior distribution. This strategy is useful because it can result in samplers with significantly better convergence characteristics and it is interesting because it may require updating the parameters by sampling from a set of *incompatible* conditional distribu-

tions. That is, there may be no joint distribution that corresponds to this set of conditional distributions.

Our technique can also be viewed as a generalization of blocking in that the resulting conditional distributions can sometimes be combined in such a way as to arrive at a Gibbs sampler that is 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 the steps in this way. This is not always possible, however, and some partially marginalized Gibbs samplers can only be composed of draws from incompatible conditional distributions. In this regard, PMG samplers constitute a generalization of the Gibbs sampler, in that Gibbs samplers are generally expected to be constructed using the conditional distributions of some joint distribution. Like blocked samplers, however, PMG 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 PMG 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 in order to allow us to use the third tool, which is to *trim* sampled components from the various steps that can be removed from the sampler without altering the Markov transition kernel of the sampler. 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 of the steps is typically small (see however, Yu, 2005). Trimming, on the other hand, is explicitly designed to maintain the transition kernel of the Markov 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.

We demonstrate the utility of our strategy from both mathematical and empirical points of view. We illustrate the computational advantage using a general mixed effects model with proper prior distributions, a Merton's jump diffusion model in finance, and a piecewise-constant multivariate time series model used to jointly segment a number of time bins in astrophysics. These are all useful models that the authors came across in their applied work and that involve computational challenges that can be solved using PMG samplers.

The remainder of the chapter is divided into five sections. This chapter begins in Section 2.2 by describing a set of prototype two and four-step Gibbs samplers that we use to motivate and to illustrate our basic strategies and techniques. These illustrations are formalized in Section 2.3, where we describe in detail the three tools we use to construct PMG samplers: marginalization, permutation, and trimming. Section 2.4 presents mathematical arguments as to the advantage of PMG samplers in terms of their lag one autocorrelation and rate of convergence. Empirical results using several examples appear in Section 2.5. Concluding remarks are given in Section 2.6.

2.2 Motivating Examples

To illustrate PMG samplers in a transparent manner, we consider the simple random effects model given by

$$y_{ij} = \xi_i + \varepsilon_{ij} \text{ for } i = 1, \ldots, k \text{ and } j = 1, \ldots, n, \tag{2.1}$$

where $\xi_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^2)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ with y_{ij} 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. Under a Bayesian perspective, we are interested in the joint posterior distribution $p(\xi, \mu|Y)$ computed under the flat prior distribution $p(\mu) \propto$

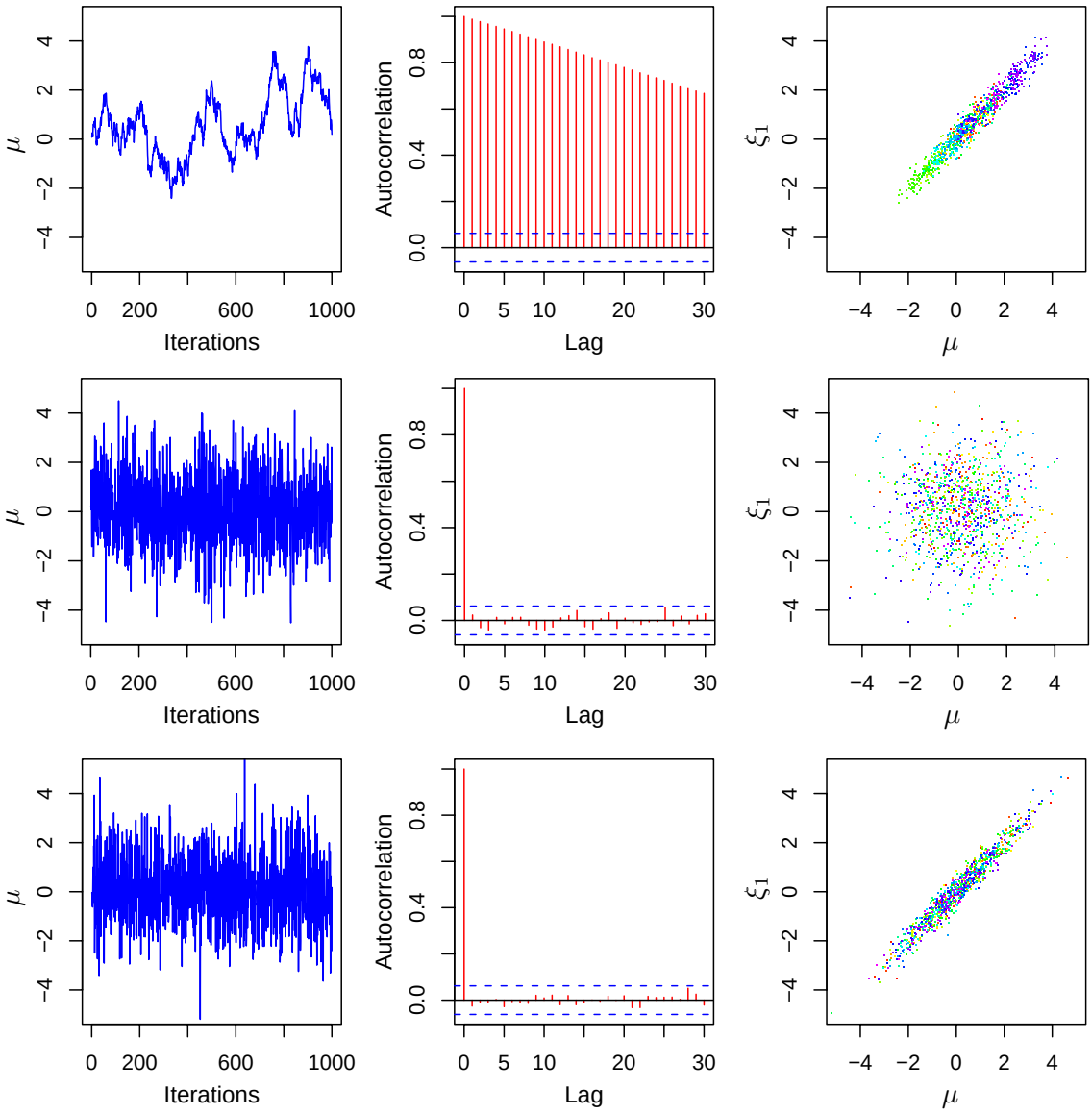

Figure 2.1: Comparison of Three Samplers for the Simple Random Effects Model. The first two columns show the mixing and autocorrelations of the subchain for μ and the last column the correlation structure between μ and ξ_1. The three rows represent the ordinary Gibbs sampler, the Gibbs sampler resulting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

1, where $\xi = (\xi_1, \xi_2, \ldots, \xi_k)$ and $Y = \{y_{ij}, i = 1, \ldots, k, j = 1, \ldots, n\}$. To fit this random effects model, we can use a prototype two-step Gibbs sampler that iterates between

STEP 1 : Draw $\xi^{(t)}$ from $p(\xi|\mu^{(t-1)}, Y)$, (Sampler 2.2.1)

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

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

with $\bar{Y}_{i\cdot}$ 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 2.2.1, the lag-one autocorrelation is the shrinkage parameter, $\sigma^2/(n\tau^2 + \sigma^2)$. Thus, the convergence rate of Sampler 2.2.1 depends on the relative magnitude of $n\tau^2$ and σ^2. If σ^2 is much greater than $n\tau^2$, the lag-one autocorrelation of Sampler 2.2.1 will be near one and the convergence slow. The first row of Figure 2.1 presents the output of Sampler 2.2.1 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 chapter, 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 $p(\mu|Y) = \int p(\xi, \mu|Y)d\xi$ of the target distribution $p(\xi, \mu|Y)$. We replace

STEP 2 of Sampler 2.2.1 with the trivial "conditional" distribution of this marginal distribution. This substitution yields a sampler that iterates between

STEP 1 : Draw $\xi^{(t)}$ from $p(\xi|\mu^{(t-1)}, Y)$, and (Sampler 2.2.2)

STEP 2 : Draw $\mu^{(t)}$ from $p(\mu|Y) = \mathrm{N}\left(\frac{\sum_i \sum_j y_{ij}}{nk}, \frac{n\tau^2 + \sigma^2}{nk}\right)$.

STEP 2 of Sampler 2.2.2 simulates μ directly from its marginal posterior distribution. The advantage of this strategy is clear in this toy example: We immediately obtain independent draws of μ from the target posterior distribution. However, the two conditional distributions used in Sampler 2.2.2, $p(\xi|\mu, Y)$ and $p(\mu|Y)$, are *incompatible* and imply inconsistent dependence structure. Even in this simple case, the incompatible conditional distributions improves the convergence characteristics of the sampler, but at the expense of the correlation structure of the target distribution. Indeed, because $\mu^{(t)}$ is sampled independently of $\xi^{(t)}$, the Markov chain has the stationary distribution $p(\xi|Y)p(\mu|Y)$ rather than $p(\xi, \mu|Y)$. This is illustrated in the second row of Figure 2.1, 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 ξ) is lost.

There is an obvious solution. Sampler 2.2.2 first draws ξ from its conditional posterior distribution $p(\xi|\mu, Y)$ and then draws μ from its marginal posterior distribution $p(\mu|Y)$, 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 $p(\mu|Y)$ and (Sampler 2.2.3)

STEP 2 : Draw $\xi^{(t)}$ from $p(\xi|\mu^{(t)}, Y)$.

Sampler 2.2.3 is constructed using a pair of incompatible conditional distributions and exhibits quicker convergence than Sampler 2.2.1, while maintaining the corre-

lation structure of the target distribution. Of course in this case, the PMG sampler (Sampler 2.2.3) is simply a blocked version of Sampler 2.2.1: STEPS 1 and 2 collapse into a single independent draw from the target distribution. As we shall illustrate, however, PMG samplers can be more general than blocked Gibbs samplers when there are more than two steps. The bottom row of Figure 2.1 illustrates the fast convergence of the subchain for μ and the correct correlation structure of μ and ξ_1.

Now we consider a more complex prototype four-step Gibbs sampler with target distribution $p(W, X, Y, Z)$. As the number of components in a Gibbs sampler increase, there are more ways to construct PMG samplers; here we focus on an example where partial marginalization does not correspond to blocking. (Generally this situation is even more complicated in that the sampled component may be vectors, and we may marginalize out certain subvectors.) We begin with the Gibbs sampler that iterates among

STEP 1 : Draw W from $p(W|X, Y, Z)$, (Sampler 2.2.4)

STEP 2 : Draw X from $p(X|W, Y, Z)$,

STEP 3 : Draw Y from $p(Y|W, X, Z)$, and

STEP 4 : Draw Z from $p(Z|W, X, Y)$.

Suppose it is possible to directly sample from $p(Y|X, Z)$ and $p(Z|X, Y)$, which are both conditional distributions of $\int p(W, X, Y, Z)dW$. By replacing STEPS 3 and 4 with draws from these two distributions, we are partially marginalizing W out of Sampler 2.2.4. 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. As we discuss above, this is illustrated by the loss of correlation structure in a sample generated with Sampler 2.2.2; see the last column of Figure 2.1. Nevertheless, we hope to capitalize

on the potential computational gain that partial marginalization offers. Thus, our goal is to formalize a procedure that allows us to introduce partially marginalized steps while ensuring the target stationary distribution is maintained. We illustrate our strategy in this example and formalize it in Section 2.3.

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 does not alter the stationary distribution of the chain or destroy the compatibility of the conditional distributions. For example, based upon the available reduced conditional distributions, we can sample W jointly with Y in STEP 3 and with Z in STEP 4. The resulting Gibbs sampler iterates among

STEP 1 : Draw $W^\star$ from $p(W|X, Y, Z)$, (Sampler 2.2.5)

STEP 2 : Draw X from $p(X|W, Y, Z)$,

STEP 3 : Draw $(W^\star, Y)$ from $p(W, Y|X, Z)$, and

STEP 4 : Draw (W, Z) from $p(W, Z|X, Y)$.

Here and elsewhere we use a superscript '$\star$' to designate an intermediate quantity that is sampled but is not the output of an iteration. Sampler 2.2.5 is a trivial generalization of what is typically considered to be a Gibbs sampler, in that W is sampled more than once during an iteration. For clarity, we sometimes use the term *simple Gibbs sampler* to refer to a sampler constructed using compatible conditional distributions in which each component is sampled exactly once in each iteration. We use the term *Gibbs sampler* to refer to a sampler with the same construction, except that some components may be updated more than once in each iteration. Occasionally we emphasize this distinction with the term *Gibbs sampler (not simple)*. Thus, Sampler 2.2.4 is a simple Gibbs sampler while Sampler 2.2.5 is a Gibbs sampler (not simple). Sampler 2.2.5 may be inefficient in that it draws 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 2.1 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 a 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 2.2.5, such permutation yields a Gibbs sampler that iterates among

STEP 1 : Draw $(W^\star, Y)$ from $p(W, Y|X, Z)$, (Sampler 2.2.6)

STEP 2 : Draw $(W^\star, Z)$ from $p(W, Z|X, Y)$,

STEP 3 : Draw W from $p(W|X, Y, Z)$, and

STEP 4 : Draw X from $p(X|W, Y, Z)$,

where the first two draws of W correspond to intermediate quantities that are not conditioned upon. This permutation alters the transition kernel, while maintaining the stationary distribution, and allows us to remove the two redundant draws of W, without changing the transition kernel. Removing the intermediate quantities $W^\star$ from Sampler 2.2.6 yields the PMG sampler that iterates among

STEP 1 : Draw Y from $p(Y|X, Z)$, (Sampler 2.2.7)

STEP 2 : Draw Z from $p(Z|X, Y)$,

STEP 3 : Draw W from $p(W|X, Y, Z)$, and

STEP 4 : Draw X from $p(X|W, Y, Z)$.

We can block STEPS 2 and 3 in Sampler 2.2.7 into a joint draw from $p(W, Z|X, Y)$, which yields

STEP 1 : Draw Y from $p(Y|X, Z)$, (Sampler 2.2.8)

STEP 2 : Draw W from $p(W, Z|X, Y)$, and

STEP 3 : Draw X from $p(X|W, Y, Z)$.

The three conditional distributions in Sampler 2.2.8 are incompatible. Thus, this PMG sampler does not simply correspond to a blocked version of Sampler 2.2.4. This illustrates that partial marginalization is a more general technique than blocking.

The resulting PMG sampler (e.g., Sampler 2.2.8) is not a Gibbs sampler in the ordinary sense. For example, permuting the draws in Sampler 2.2.8 may result in a transition kernel with unknown stationary distribution, whereas permuting the steps of a Gibbs sampler never affects its stationary distribution. Since the removal of intermediate quantities introduces incompatibility into the sampler, removal must be done with great care.

2.3 Basic Tools

Here we present three basic tools that we use to construct PMG 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

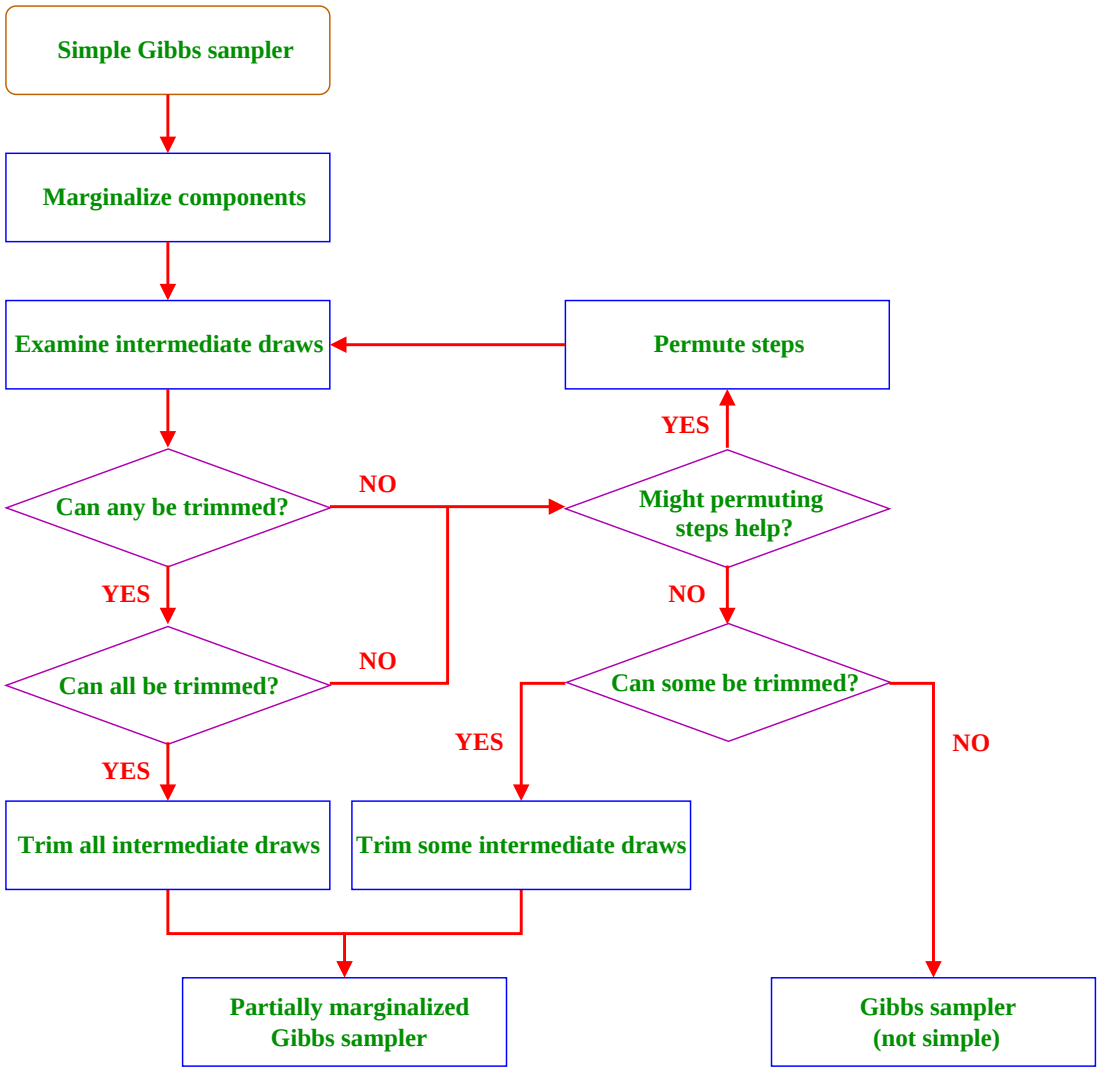

Figure 2.2: Flow Diagram for Deriving a PMG Sampler from a Simple Gibbs Sampler.

resulting PMG samplers converge quickly to the target distribution. We discuss the three tools, marginalization, permutation, and trimming, in the order that they are typically applied. Figure 2.2 presents a flow diagram that describes how the basic tools are applied to a simple Gibbs sampler in order to construct a PMG sampler which preserves a target stationary distribution. Each component of the flow diagram in Figure 2.2 is closely examined in the following subsections.

2.3.1 Marginalization

Suppose we aim to construct a PMG sampler with stationary distribution $p(X)$ where X is a vector quantity that we partition into J subvectors, $X = (X_1, X_2, \ldots, X_J)$. Consider the sequence of index sets of the components of X, $\mathcal{J} = \{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_P\}$, where $\mathcal{J}_p \subset \{1, 2, \ldots, J\}$ for $p = 1, 2, \ldots, P$ such that $\cup_{p=1}^{P} \mathcal{J}_p = \{1, 2, \ldots, J\}$. We denote the complement of $\mathcal{J}_p$ in $\{1, 2, \ldots, J\}$ by $\mathcal{J}_p^c$. Let $\mathcal{X}_{\mathcal{J}_p}$ denote the collection of components of X corresponding to the index set $\mathcal{J}_p$, i.e., $\mathcal{X}_{\mathcal{J}_p} = \{X_j : j \in \mathcal{J}_p\}$ for $p = 1, 2, \ldots, P$. STEP p of a P-step Gibbs sampler can be written as

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

where $\mathcal{X}_{\mathcal{J}_p^c}^{(t-1)} = \{(X_j^{(t)}, X_k^{(t-1)}) : j \in (\cup_{i=1}^{p-1} \mathcal{J}_i) \cap \mathcal{J}_p^c \text{ and } k \in \cap_{i=1}^{p} \mathcal{J}_i^c\}$, i.e., STEP p is conditional on the most recent draws of X not being sampled in the step. Notice that this is a Gibbs sampler using compatible conditional distributions, where some components of X may be updated in multiple steps; thus, this may not be a simple Gibbs sampler.

A simple Gibbs sampler updates each (vector) component of X only once during each iteration by sampling it from its complete conditional distribution. In our notation, this corresponds to the case where $\mathcal{J}$ is a partition of $\{1, 2, \ldots, J\}$. At the other extreme, suppose there exists an index k such that $k \in \mathcal{J}_p$ for each

p, then X_k is drawn in each step and is never conditioned upon; we say X_k has been (completely) marginalized out of the Gibbs sampler. In this case we can reformulate the Gibbs sampler completely in terms of the marginal distribution $\int \pi(X_1, X_2, \ldots, X_J)dX_k$, without altering the transition kernel for the other components of X.

The first step in constructing a PMG sampler is to marginalize some components of X out of some steps of the sampler. To do this we replace $\mathcal{J}_q$ with $\widetilde{\mathcal{J}}_q$ for some $q \in \{1, 2, \ldots, P\}$ where $\mathcal{J}_q$ is a proper subset of $\widetilde{\mathcal{J}}_q$. That is, in STEP q, we move some components of 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 2.4 for the theory on the improved rate of convergence. Then STEP q is conditional on fewer components of X and is given by

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

where $\mathcal{X}^{(t-1)}_{\widetilde{\mathcal{J}}^c_q} = \{(X^{(t)}_j, X^{(t-1)}_k) : j \in (\cup_{i=1}^{q-1}\mathcal{J}_i) \cap \widetilde{\mathcal{J}}^c_q \text{ and } k \in (\cap_{i=1}^{q-1}\mathcal{J}^c_i) \cap \widetilde{\mathcal{J}}^c_q\}$. Partially marginalizing out some components of X alters the transition kernel but not the stationary distribution for X or the compatibility of the conditional distributions. The improved rate of convergence for the resulting sampler is mainly attributed to this partial marginalization.

2.3.2 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 may change its transition kernel and interchange intermediate quantities with output quantities, but maintains the stationary distribution of the chain. Our goal in permuting the steps of a Gibbs sampler is to arrange the steps in such a way, 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 chapter, we tend to ignore the effect of permutation on convergence, but in some situations the permutation may critically affect convergence (Yu, 2005). Here we are merely interested in permutations because they can allow the removal of some intermediate quantities from the chain.

2.3.3 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 $\mathcal{X}_{\mathcal{J}_q}^{(t)}$ from $p(\mathcal{X}_{\mathcal{J}_q}|\mathcal{X}_{\widetilde{\mathcal{J}}_q^c}^{(t-1)})$.

The reduced conditional distribution sampled in this step is not typically compatible with the other conditional distributions sampled in the sampler. In particular, because $\mathcal{J}_q \cup \widetilde{\mathcal{J}}_q^c$ is not equal to $\mathcal{J}$ ($\mathcal{J}_q$ is a proper subset of $\widetilde{\mathcal{J}}_q$), this conditional distribution is not defined as the same space as the conditional distributions of the original sampler. Thus, it is trimming that introduces incompatibility into the conditional distributions of a PMG sampler. This means the resulting PMG 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 PMG sampler may result in a new Markov transition kernel with an unknown stationary distribution. Nonetheless, trimming is advantageous because each iteration is generally less computationally demanding. Indeed, as the piecewise-constant multivariate time series model illustrates in Section 2.5.3, trimming may render an intractable sampling step tractable.

Intermediate quantities are not part of the output of an iteration, but 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. On the other hand, trimming intermediate quantities that do impact the transition kernel can effect the correlation structure of the stationary distribution. Thus, care must be taken with the trimming of intermediate quantities in order to maintain the stationary distribution.

2.4 PMG Theory

In order to discuss the effect of partial marginalization on the convergence of a Gibbs sampler, we must 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 $h(X)$ such that $\int h^2(X)\pi(X)dX < \infty$. This set is a Hilbert space with inner product $\langle h, g\rangle = \mathrm{E}_\pi\{h(X)g(X)\}$, so that $\|h\|^2 = \mathrm{Var}_\pi(h)$. For a general Markov chain $\mathcal{M}_X = \{X^{(0)}, X^{(1)}, \dots\}$ with transition kernel $\mathcal{K}(X^{(1)} = X|X^{(0)} = X')$, we define the forward operator $\mathbf{F}$ on $L^2(\pi)$ for $\mathcal{M}_X$ by

$$\mathbf{F}h(X') = \int h(X)\mathcal{K}(X|X')dX = \mathrm{E}\{h(X^{(1)})|X^{(0)} = X'\}. \tag{2.2}$$

Let $L_0^2(\pi) = \big\{h : \mathrm{E}_\pi\{h(X)\} = 0, \mathrm{Var}_\pi\{h(X)\} = 1\big\}$. This is also a Hilbert space with the same inner product and is invariant under $\mathbf{F}$. We define $\mathbf{F}_0$ to be the forward operator on $L_0^2(\pi)$ induced by $\mathbf{F}$. If we define the norm of this forward operator by $\|\mathbf{F}_0\| = \sup_h \|\mathbf{F}_0 h(X)\|$ with the supremum taken over $h \in L_0^2(\pi)$, it can be shown that

$$\begin{aligned}
\|\mathbf{F}_0\| &= \sup_{h\in L_0^2(\pi)} \Big(\mathrm{Var}_\pi\big[\mathrm{E}\{h(X^{(1)})|X^{(0)}\}\big]\Big)^{1/2} && (2.3)\\
&= \sup_{h\in L_0^2(\pi)} \Big\{\mathrm{E}_\pi\Big(\big[\mathrm{E}\{h(X^{(1)})|X^{(0)}\}\big]^2\Big)\Big\}^{1/2} && (2.4)\\
&= \rho(X^{(1)}, X^{(0)}), && (2.5)
\end{aligned}$$

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

$$\begin{aligned} \rho(\vartheta, \varphi) &= \sup \mathrm{Corr}\{h(\vartheta), g(\varphi)\} && (2.6) \\ &= \sup_{h:\, \mathrm{Var}\{h(\vartheta)\}=1} \left(\mathrm{Var}_\pi\left[\mathrm{E}\{h(\vartheta)|\varphi\}\right]\right)^{1/2}, && (2.7) \end{aligned}$$

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 $\rho(X^{(1)}, X^{(0)})$ is computed under the stationary distribution of $\mathcal{M}_X$, and will also be denoted by $\rho(\mathcal{M}_X)$.

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

$$\lim_{n\to\infty} \|\mathbf{F}_0^n\|^{1/n} = r(\mathbf{F}_0) \tag{2.8}$$

and by the inequality

$$r(\mathbf{F}_0) \leq \|\mathbf{F}_0\|. \tag{2.9}$$

Along with the relationship between the maximum autocorrelation of $\mathcal{M}_X$ and $\|\mathbf{F}_0\|$, (2.8) and (2.9) justify the use of $\|\mathbf{F}_0\|$ in the analysis of the convergence behavior of $\mathcal{M}_X$.

Consider the P-step Gibbs sampler described in Section 2.3. We define a p-step-lagged Gibbs sampler for $p = 0, 1, \ldots, P-1$, as the Gibbs sampler with iteration that begins with STEP p+1, cycles through the steps in the same order as the original Gibbs sampler, and ends with STEP p. The forward operators of the P p-step-lagged Gibbs samplers have the same spectral radius, which we denote by r. They may, however, have different norms, hence different maximum autocorrelations. We denote the maximum autocorrelation $\rho(\mathcal{M}_X)$ of the p-step-lagged chain by ρ_p for $p = 0, 1, \ldots, P-1$. By (2.9), we have

$$r \leq \min_{p\in\{0,1,\ldots,P-1\}} \rho_p. \tag{2.10}$$

We will show that by marginalizing a component of 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 (2.10) on the spectral radius.

Because STEP p+1 is the first step of the p-step-lagged Gibbs sampler, we evaluate the effect of marginalizing a component of 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 marginalization, we consider the generic P-step Gibbs sampler introduced in Section 2.3 from which we marginalize some components of X in STEP 1. Thus, we wish to compare two sequences of index sets and their resulting transition kernels; namely $(\mathcal{X}_{\mathcal{J}_1}, \mathcal{X}_{\mathcal{J}_2}, \ldots, \mathcal{X}_{\mathcal{J}_P})$ and its kernel $\mathcal{K}(X|X')$ and $(\mathcal{X}_{\widetilde{\mathcal{J}}_1}, \mathcal{X}_{\widetilde{\mathcal{J}}_2}, \ldots, \mathcal{X}_{\widetilde{\mathcal{J}}_P})$ and its kernel $\widetilde{\mathcal{K}}(X|X')$, where $\mathcal{J}_p = \widetilde{\mathcal{J}}_p$ for $p = 2, \ldots, P$, but $\mathcal{X}_{\mathcal{J}_1} = \{x_1\}$ and $\mathcal{X}_{\widetilde{\mathcal{J}}_1} = \{x_1, x_2\}$ with $X = (x_1, x_2, x_3)$. Here (x_1, x_2, x_3) is an alternate partition of $X = (X_1, X_2, \ldots, X_J)$ introduced to simplify notation in the theorem. That is, $\mathcal{J}_1 \subset \widetilde{\mathcal{J}}_1 \subset \{1, 2, \ldots, J\}$, where both subsets are proper subsets, $x_1 = \{X_j : j \in \mathcal{J}_1\}$, $x_2 = \{X_j : j \in \widetilde{\mathcal{J}}_1 \setminus \mathcal{J}_1\}$, and $x_3 = \{X_j : j \in \widetilde{\mathcal{J}}_1^c\}$. In words, the two sequence of index sets represent identical samplers, except in STEP 1, where more components of X are drawn from the Gibbs sampler with kernel $\widetilde{\mathcal{K}}(X|X')$. In this case, we have the following result.

Theorem 1 *Sampling more components of X in the first step of a Gibbs sampler improves the resulting maximal autocorrelation, $\rho(\mathcal{M}_X)$.*

Proof: Let h be an arbitrary function of X with mean zero and finite variance under stationarity, i.e., $h \in L_0^2(\pi)$ with π the stationary distribution of X, then

$$\widetilde{\mathrm{E}}\{h(X)|x_3'\} = \int h(X)\mathcal{K}_{-1}(X|x_1, x_2, x_3')p(x_1, x_2|x_3')d\Xi_{-1}dx_1dx_2, \tag{2.11}$$

where $\widetilde{\mathrm{E}}$ represents expectation with respect to $\widetilde{\mathcal{K}}(X|X')$, $\mathcal{K}_{-1}(X|X')$ is the transition kernel implied by STEP 2 through STEP P of either sampler, and $\Xi_{-1} = (\mathcal{X}_{\mathcal{J}_2}, \ldots, \mathcal{X}_{\mathcal{J}_P})$ is the set of components updated in STEP 2 through STEP P, which

may include multiple copies of certain components of X. Now, the right-hand side of (2.11) can be written as

$$\begin{aligned}\widetilde{\mathrm{E}}\{h(X)|x_3'\} &= \int\left\{\int h(X)\mathcal{K}_{-1}(X|x_1,x_2',x_3')p(x_1|x_2',x_3')d\Xi_{-1}dx_1\right\}p(x_2'|x_3')dx_2' \\ &= \mathrm{E}_\pi[\mathrm{E}\{h(X)\,|\,x_2',x_3'\}\,|\,x_3'], \end{aligned} \tag{2.12}$$

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

$$\begin{aligned}\mathrm{E}_\pi\Big(\big[\widetilde{\mathrm{E}}\{h(X)\,|\,x_3'\}\big]^2\Big) &= \mathrm{E}_\pi\Big\{\big(\mathrm{E}_\pi[\mathrm{E}\{h(X)\,|\,x_2',x_3'\}\,|\,x_3']\big)^2\Big\} \\ &\leq \mathrm{E}_\pi\Big\{\mathrm{E}_\pi\big([\mathrm{E}\{h(X)\,|\,x_2',x_3'\}]^2\,|\,x_3'\big)\Big\} \\ &= \mathrm{E}_\pi\Big(\big[\mathrm{E}\{h(X)\,|\,x_2',x_3'\}\big]^2\Big). \end{aligned} \tag{2.13}$$

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

That is, the computational advantages can be achieved by successively marginalizing over the components of X in any single step of a Gibbs sampler. Thus, recursively using Theorem 1 provides the theoretical basis for the improved convergence characteristics of PMG samplers.

2.5 Applications

2.5.1 Mixed Effects Models With Proper Prior Distributions

Consider the general mixed effects model given by

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i \;\text{ with }\; b_i \overset{\text{iid}}{\sim} \mathrm{N}_q(0,\sigma^2D),\;\; \varepsilon_i \overset{\text{ind}}{\sim} \mathrm{N}_{n_i}(0,\sigma^2R_i),\;\text{ and }\; b_i \perp \varepsilon_i \tag{2.14}$$

where Y_i are $n_i \times 1$ vectors for group $i = 1,\ldots,k$, X_i and Z_i are known covariates of dimension $n_i \times p$ and $n_i \times q$, respectively, β is a $p \times 1$ vector of fixed effects, b_i are $q \times 1$ vectors of random effects, D is a $q \times q$ positive definite matrix, R_i are known $n_i \times n_i$ positive definite matrices, and ε_i are $n_i \times 1$ vectors of residuals. We aim to

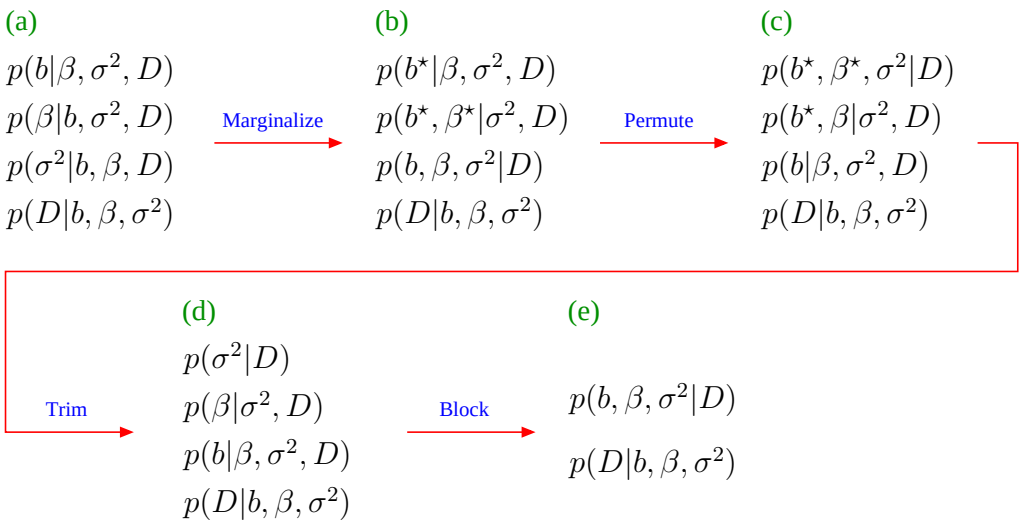

Figure 2.3: Illustration of Deriving a PMG Sampler for the Mixed Effects Model with Proper Prior Distributions. For clarity, conditioning on Y_{obs} for each sampling distribution is suppressed throughout.

generate samples from the posterior distribution $p(b,\beta,\sigma^2,T|Y_{\text{obs}})$ computed under the proper conjugate prior distributions,

$$\beta|\sigma^2 \sim \text{N}_p(\mu_\beta,\sigma^2\Sigma_\beta),\ \ \sigma^2 \sim \frac{\nu\sigma_0^2}{\chi^2_\nu},\ \text{ and }\ T \sim \text{Inv}-\text{Wishart}(\eta,T_0), \tag{2.15}$$

where $b = (b_1,b_2,\ldots,b_k)$ is a $q \times k$ matrix of random effects, $Y_{\text{obs}} = \{Y_i, i = 1,\ldots,k\}$, Inv$-$Wishart stands for an inverse Wishart distribution, and $W \sim \text{Inv}-\text{Wishart}(\nu,S)$ if $p(W) \propto |W|^{-(\nu+k+1)/2}\exp(-\frac{1}{2}\{SW^{-1}\})$ for $k \times k$ positive definite matrices S and W. To fit this mixed effects model, we can construct a simple Gibbs sampler. As exemplified in Section 2.2, however, the simple Gibbs sampler constructed for the mixed effects model can exhibit slow convergence. To facilitate computation in the context of an EM-type algorithm, van Dyk (2000a) suggests a reparameterization of the between group variance T in terms of the within group variance σ^2, i.e., $T = \sigma^2 D$. The utility of the parameterization (β,σ^2,D) was also noted by Schafer (1998) (see also Lindstrom and Bates, 1988). This reparameterization allows us to use a reduced conditional distribution when sampling σ^2. Using

this parameterization, the posterior distribution of interest is written as

$$\begin{aligned} p(b,\beta,\sigma^2,D|Y_{\text{obs}}) \;\propto\; & (\sigma^2)^{-((\nu+n+p)/2+1)}|\sigma^2 D|^{-(\eta+k+q+1)/2} \\ & \exp\Bigg(-\frac{1}{2\sigma^2}\Bigg[\nu\sigma_0^2+(\beta-\mu_\beta)^\top\Sigma_\beta^{-1}(\beta-\mu_\beta) \\ & +\sum\nolimits_{i=1}^{k}(Y_i-X_i\beta-Z_ib_i)^\top R_i^{-1}(Y_i-X_i\beta-Z_ib_i) \\ & +\text{tr}\Big\{D^{-1}\Big(T_0+\sum\nolimits_{i=1}^{k}b_ib_i^\top\Big)\Big\}\Bigg]\Bigg). \end{aligned} \tag{2.16}$$

Figure 2.3 illustrates the evolution from a simple Gibbs sampler to a PMG sampler by applying a sequence of the basic tools introduced in Section 2.3. In Figure 2.3(a), we begin with the four complete conditional distributions used in the simple Gibbs sampler. In particular, given $(\beta,\sigma^2,D,Y_{\text{obs}})$, $\{b_i, i=1,2,\ldots,k\}$ follow independent multivariate Gaussian distributions,

$$b_i|(\beta,\sigma^2,D,Y_{\text{obs}}) \;\sim\; \text{N}_q\Big(\widehat{b}_i(\beta,D),\sigma^2\big[D-DZ_i^\top U_i(D)Z_iD\big]\Big), \tag{2.17}$$

where $\widehat{b}_i(\beta,D) = DZ_i^\top U_i(D)(Y_i-X_i\beta)$ and $U_i(D)=(R_i+Z_iDZ_i^\top)^{-1}$; given $(b,\sigma^2,D,Y_{\text{obs}})$, β is also a multivariate Gaussian variate,

$$\beta|(b,\sigma^2,D,Y_{\text{obs}}) \;\sim\; \text{N}_p\Bigg(\widehat{\beta}(b),\sigma^2\Big[\Sigma_\beta^{-1}+\sum\nolimits_{i=1}^{k}X_i^\top R_i^{-1}X_i\Big]^{-1}\Bigg), \tag{2.18}$$

where $\widehat{\beta}(b) = \big(\Sigma_\beta^{-1}+\sum_{i=1}^{k}X_i^\top R_i^{-1}X_i\big)^{-1}(\Sigma_\beta^{-1}\mu_\beta+\sum_{i=1}^{k}X_i^\top R_i^{-1}(Y_i-Z_ib_i))$; given $(b,\beta,D,Y_{\text{obs}})$, σ^2 follows an inverse χ^2 distribution,

$$\begin{aligned} \sigma^2|(b,\beta,D,Y_{\text{obs}}) \;\sim\; & \Bigg(\nu\sigma_0^2+(\beta-\mu_\beta)^\top\Sigma_\beta^{-1}(\beta-\mu_\beta) \\ & +\sum\nolimits_{i=1}^{k}(Y_i-X_i\beta-Z_ib_i)^\top R_i^{-1}(Y_i-X_i\beta-Z_ib_i) \\ & \text{tr}\Big\{D^{-1}\Big(T_0+\sum\nolimits_{i=1}^{k}b_ib_i^\top\Big)\Big\}\Bigg)\Big/\chi^2_{\nu+n+p+q(\eta+k+q+1)}, \end{aligned} \tag{2.19}$$

where $n=\sum_{i=1}^{k}n_i$; and, given $(b,\beta,\sigma^2,Y_{\text{obs}})$, D follows an inverse Wishart distribution,

$$D|(b,\beta,\sigma^2,Y_{\text{obs}}) \;\sim\; \text{Inv}-\text{Wishart}\Bigg(\eta+k,\Big[T_0+\sum\nolimits_{i=1}^{k}b_ib_i^\top\Big]\Big/\sigma^2\Bigg). \tag{2.20}$$

In order to improve convergence, we can marginalize components out of some of these conditional distributions and construct a PMG sampler. For example, Figure 2.3(b) marginalizes b out of STEPS 2 and 3, and β out of STEP 3. Due to our choice of parametrization, the reduced conditional distributions are still in closed forms. Namely, the distribution of β given $(\sigma^2, D, Y_{\text{obs}})$ is a multivariate Gaussian distribution,

$$\beta|(\sigma^2, D, Y_{\text{obs}}) \;\sim\; \mathrm{N}_p\Big(\widetilde{\beta}(D), \sigma^2\Big[\Sigma_\beta^{-1} + \sum\nolimits_{i=1}^{k} X_i^\top U_i(D) X_i\Big]^{-1}\Big), \tag{2.21}$$

where $\widetilde{\beta}(D) = \big(\Sigma_\beta^{-1} + \sum_{i=1}^k X_i^\top U_i(D) X_i\big)^{-1}(\Sigma_\beta^{-1}\mu_\beta + \sum_{i=1}^k X_i^\top U_i(D) Y_i)$, while the distribution of σ^2 given (D, Y_{obs}) is inverse χ^2,

$$\begin{aligned}\sigma^2|(D, Y_{\text{obs}}) \;\sim\; \Big(&\nu\sigma_0^2 + (\widetilde{\beta}(D) - \mu_\beta)^\top \Sigma_\beta^{-1}(\widetilde{\beta}(D) - \mu_\beta) \\ &+ \sum\nolimits_{i=1}^{k}\Big[Y_i - X_i\widetilde{\beta}(D)\Big]^\top U_i(D)\Big[Y_i - X_i\widetilde{\beta}(D)\Big] \\ &+ \mathrm{tr}\Big\{D^{-1}T_0\Big\}\Big)\Big/\chi^2_{\nu+n+q(\eta+q+1)}.\end{aligned} \tag{2.22}$$

Permuting the steps in Figure 2.3(b) allows us to connect all of the marginalized quantities into the intermediate quantities that do not affect the transition kernel, see Figure 2.3(c). After the permutation, the marginalized quantities become redundant in the sampler. Trimming such marginalized quantities results in the PMG sampler in Figure 2.3(d). We can block the first three steps of Figure 2.3(d) into $p(b, \beta, \sigma^2|D, Y_{\text{obs}})$, thereby yielding the two-step Gibbs sampler in Figure 2.3(e). Thus, the PMG sampler corresponds to a blocked version of the original Gibbs sampler in Figure 2.3(a). We emphasize that if the PMG sampler in Figure 2.3(d) is implemented with another permutation of the steps, the resulting transition kernel may be different from that of the two-step Gibbs sampler in Figure 2.3(e).

We compare the convergence characteristics of the simple Gibbs sampler and PMG sampler via simulation study. We assume $k = 100$ groups and each group is of the same size $n_i = 2$. As for the model parameters, we use $\beta = (1\ 2)^\top$, $\sigma^2 = 1$, and

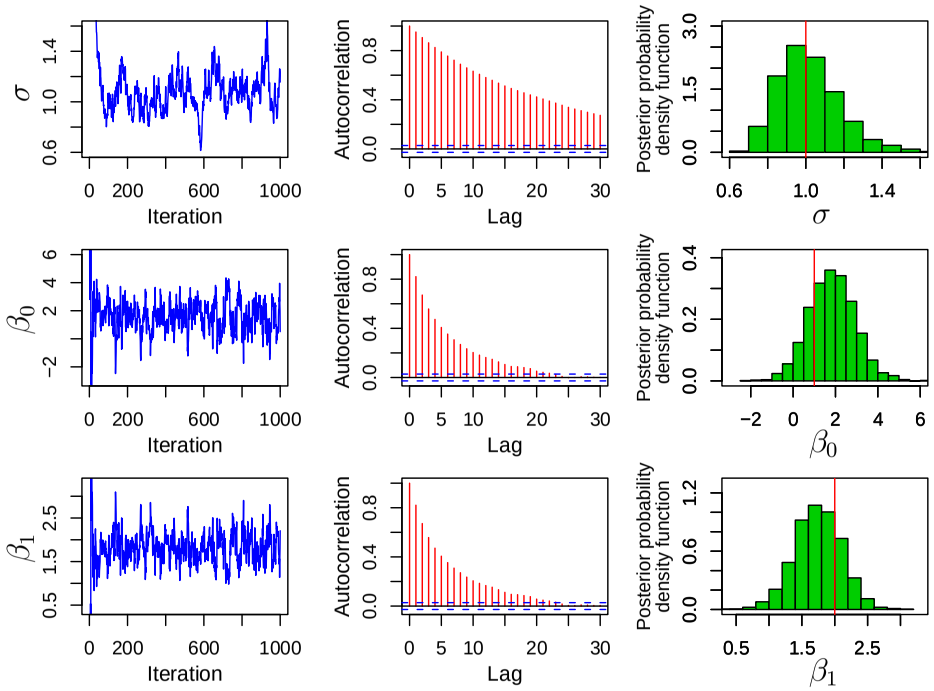

Figure 2.4: Mixing, Autocorrelation, and Marginal Posterior Distribution of Each Model Parameter Simulated by the Simple Gibbs Sampler Constructed for the Mixed Effects Model with Proper Prior Distributions.

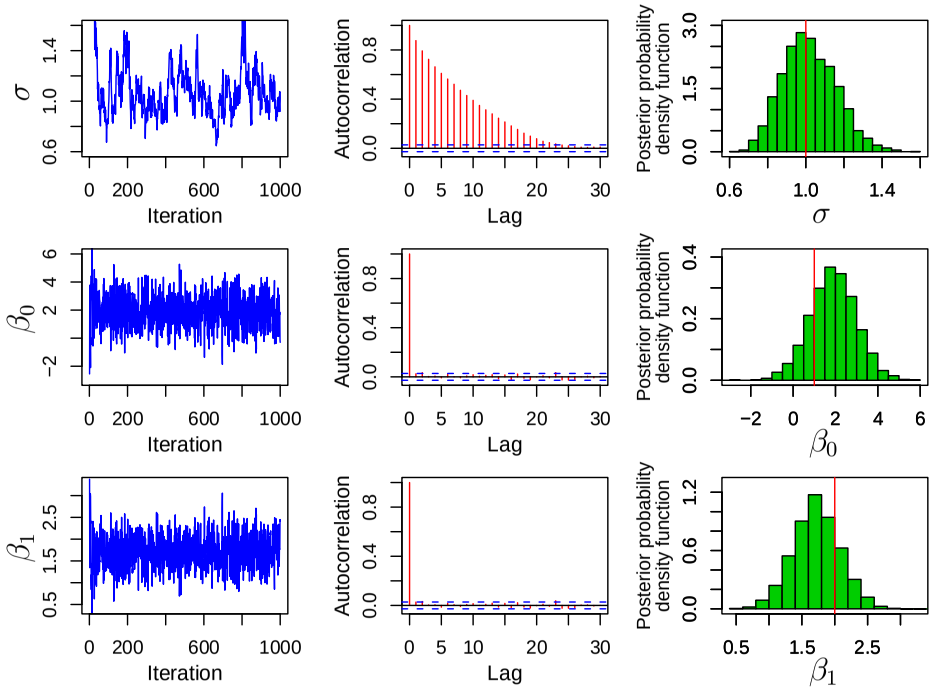

Figure 2.5: Mixing, Autocorrelation, and Marginal Posterior Distribution of Each Model Parameter Simulated by the PMG Sampler Constructed for the Mixed Effects Model with Proper Prior Distributions.

$T = \sigma^2 D = \begin{pmatrix} 25 & 45 \\ 45 & 100 \end{pmatrix}$. That is, we have $p = 2$, and the $q = 2$ components of the random effects are highly correlated. We decide the values of $(\mu_\beta, \Sigma_\beta, \nu, \sigma_0^2, \eta, T_0)$ such that the proper prior distributions are diffuse. To examine the convergence of the samplers, we run multiple chains of 10000 iterations each with different starting values, and compute the estimate of the potential scale reduction (Gelman and Rubin, 1992), denoted by $\sqrt{\widehat{R}}$, for all parameters of interest. If $\sqrt{\widehat{R}}$ is near 1 (e.g., below 1.2) for the parameters, we collect the second halves of the chains together and use those Monte Carlo draws for our inference; see Gelman and Rubin (1992) for a theoretical justification and discussion.

In Figures 2.4 and 2.5, we compare the convergence characteristics of selected model parameters whose conditional distributions are conditioning on less in the PMG sampler. The first two columns of Figures 2.4 and 2.5 show the convergence behaviors of the selected model parameters by using a single chain constructed by the Gibbs sampler and PMG sampler, respectively. By visually examining the mixing and autocorrelations of the parameters, we can compare the convergence of the chains. As we can confirm, the subchains for the parameters are mixed faster and have smaller autocorrelations with the PMG sampler than the Gibbs sampler: In order to make the distinction clear, the first column of Figure 2.4 and 2.5 is drawn with the last 1000 draws of a single chain, and the second column with the last 5000 draws. In particular, the computational gains of the partially marginalized methods are substantial for the subchains for β_0 and β_1, and the effect on σ is also evident although it is relatively small. The last column of Figures 2.4 and 2.5 presents the resulting marginal posterior distributions based on the second halves of the multiple chains. The vertical solid lines represents the true values used to simulate the test data. We note that the true values of the parameters are plausible in the posterior distributions, which demonstrates our fitting of the model is correct.

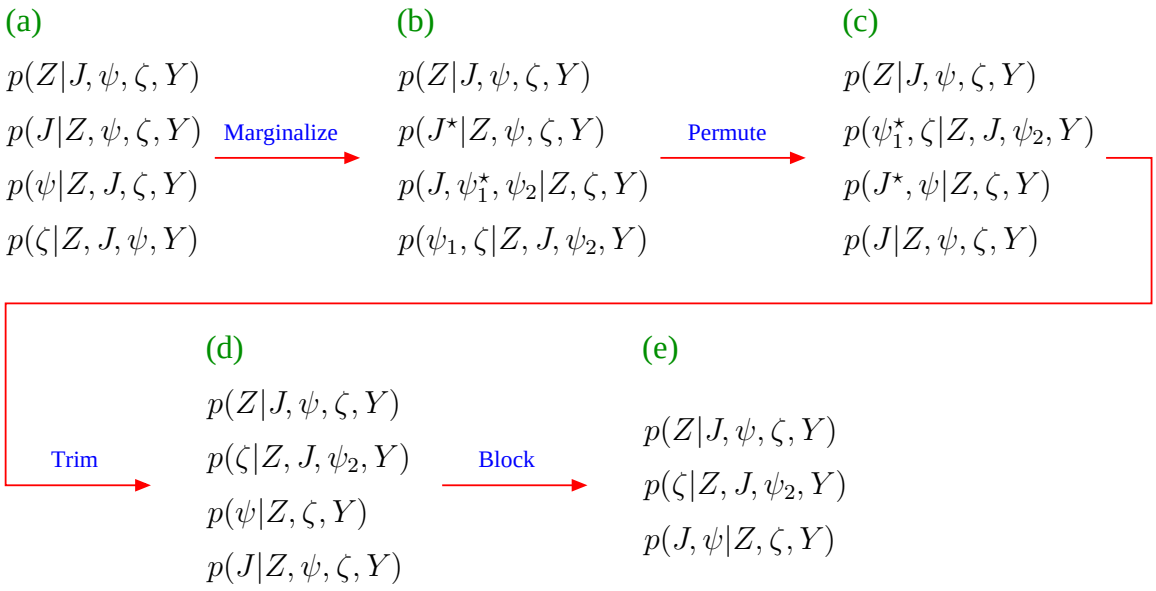

Figure 2.6: Illustration of Deriving a PMG Sampler for the Merton's Jump Diffusion Model with Proper Prior Distributions.

2.5.2 Merton's Jump Diffusion Model

Consider the Merton's jump diffusion model (Merton, 1976), which aims to model jumps in a stock price due to rare economic events or sudden news. The model is given by

$$\frac{dS_t}{S_{t-}} = \gamma d_t + \sigma dW_t + (e^{J_t} - 1)dN_t, \tag{2.23}$$

where S_t represents the stock price at time t, γ is the instantaneous expected return of the stock, σ is the instantaneous standard deviation of the stock's return, W_t is a Wiener process, the log-jump size J_t is a Gaussian random variable with mean μ_J and variance σ_J^2, and N_t is a Poisson process with arrival rate λ. Without the jump process, (2.23) is known as a geometric Brownian motion process and the successive log ratios of $\{S_t, t = 1, 2, \ldots, T\}$ are independent Gaussian random variables with mean γ and variance σ^2. When a jump occurs at time t, however, the process is no longer continuous; S_{t-} explicitly represents the discontinuity between jumps. In addition, we consider daily based stock prices, so that at most a single jump is

assumed to occur over each time interval, i.e.,

$$p(dN_t = 1) = p(N_{t+h} - N_t = 1) = \lambda h + o(h), \tag{2.24}$$

$$p(dN_t = 0) = p(N_{t+h} - N_t = 0) = 1 - \lambda h + o(h), \text{ and} \tag{2.25}$$

$$p(dN_t > 1) = p(N_{t+h} - N_t > 1) = o(h). \tag{2.26}$$

The e^{J_t} in (2.23) is a jump multiplier for S_t, so that the proportional net change for S_t is $e^{J_t} - 1$ when a jump occurs at time t.

By applying $Itô's\ Lemma$ for the stochastic differential equation of the jump diffusion model in (2.23), we obtain the jump diffusion process for the log-return given by

$$d \log S_t = \mu d_t + \sigma dW_t + J_t dN_t, \tag{2.27}$$

where the drift is reparameterized as $\mu = \gamma - \sigma^2/2$. Integrating (2.27) over a daily time increment $\Delta t = 1$ yields

$$Y_t = \mu + \sigma\varepsilon_t + J_t Z_t \text{ for } t = 1, 2, \ldots, T, \tag{2.28}$$

where $Y_t = \log(S_t/S_{t-1})$ is the difference in log return between time t and $t-1$, $\varepsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $J_t \overset{\text{iid}}{\sim} \mathrm{N}(\mu_J, \sigma_J^2)$, $J_t \perp \varepsilon_t$, and $Z_t \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\lambda)$. Statistically, (2.28) can be viewed as the mixture of two Gaussian random variables with random effects components, i.e.,

$$Y_t \sim \lambda \mathrm{N}(\mu + \mu_J, \sigma^2 + \sigma_J^2) + (1 - \lambda)\mathrm{N}(\mu, \sigma^2). \tag{2.29}$$

We parameterize the variance of the random effects in terms of the residual variance σ^2 to facilitate computation (e.g., van Dyk, 2000a; Schafer, 1998). Such parameterization, i.e., $\sigma_J^2 = \sigma^2\xi$, is key to marginalize the fixed and random effects out of the conditional distribution for σ^2 under proper conjugate prior distributions, and allows us to devise a PMG sampler; ξ quantifies the scale factor of the between group variance relative to the within group variance. Under the proper conjugate

prior distributions,

$$\begin{aligned}
&\mu|\sigma^2 \sim \mathrm{N}(a_0, \sigma^2 b_0),\ \sigma^2 \sim \frac{\nu\tau_0^2}{\chi^2_\nu},\\
&\mu_J|\sigma_J^2 \sim \mathrm{N}(a_J, \sigma_J^2 b_J),\ \sigma_J^2 \sim \frac{\eta\tau_J^2}{\chi^2_\eta},\ \text{and}\\
&\lambda \sim \mathrm{Beta}(\alpha, \beta),
\end{aligned} \tag{2.30}$$

we aim to generate a sample from the posterior distribution $p(Z, J, \mu, \mu_J, \sigma^2, \xi, \lambda|Y)$ that is given by

$$\begin{aligned}
p(Z, J, \mu, \mu_J, \sigma^2, \xi, \lambda|Y) \quad &\propto \quad \prod_{t=1}^{T} p(Y_t|Z_t, J_t, \mu, \sigma^2) p(J_t|\mu_J, \sigma^2, \xi) p(Z_t|\lambda) p(\mu, \mu_J, \sigma^2, \xi, \lambda)\\
&\propto \quad (\sigma^2)^{-\left((2T+\nu+\eta+4)/2+1\right)} \xi^{-\left((T+\eta+1)/2+1\right)}\\
&\quad \lambda^{\sum_{t=1}^{T} Z_t+\alpha-1}(1-\lambda)^{\sum_{t=1}^{T}(1-Z_t)+\beta-1}\\
&\quad \exp\Bigg(-\frac{1}{2\sigma^2}\Bigg[\sum\nolimits_{t=1}^{T}(Y_t-\mu-J_tZ_t)^2 + \frac{(\mu-a_0)^2}{b_0} + \nu\tau_0^2\Bigg]\\
&\quad -\frac{1}{2\sigma^2\xi}\Bigg[\sum\nolimits_{t=1}^{T}(J_t-\mu_J)^2 + \frac{(\mu_J-a_J)^2}{2b_J} + \eta\tau_J^2\Bigg]\Bigg).
\end{aligned} \tag{2.31}$$

For notational convenience, we define $\psi = (\psi_1, \psi_2)$ where $\psi_1 = (\mu, \mu_J)$ and $\psi_2 = \sigma^2$, and $\zeta = (\xi, \lambda)$, so that the posterior distribution of interest is rewritten as $p(Z, J, \psi, \zeta|Y)$.

Figure 2.6 illustrates five samplers designed to generate simulation from the target posterior distribution. We begin with a simple Gibbs sampler constructed with four complete conditional distributions that are listed in Figure 2.6(a) in the order in which they are sampled in each iteration. Each of these complete conditional distributions is a standard distribution. Given (J, ψ, ζ, Y), Z are T independent Bernoulli variables,

$$Z|(J, \psi, \zeta, Y) \quad \sim \quad \prod_{t=1}^{T} \mathrm{Bernoulli}\left(\frac{\lambda\phi(Y_t; \mu + J_t, \sigma^2)}{\lambda\phi(Y_t; \mu + J_t, \sigma^2) + (1-\lambda)\phi(Y_t; \mu, \sigma^2)}\right), \tag{2.32}$$

where $\phi(\,\cdot\,; \mu, \sigma^2)$ denotes a Gaussian density function centered at μ and with variance σ^2. Given (Z, ψ, ζ, Y), J is a multivariate Gaussian random variable with T

independent components, where

$$J_t|(Z,\psi,\zeta,Y) \;\sim\; \mathrm{N}\Big(\frac{Z_t(Y_t-\mu)+\mu_J/\xi}{Z_t+1/\xi}, \frac{\sigma^2}{Z_t+1/\xi}\Big), \tag{2.33}$$

for $t = 1, 2, \ldots, T$. The conditional distribution of ψ given (Z, J, ζ, Y) is decomposed into

$$p(\psi|Z,J,\zeta,Y) \;=\; p(\sigma^2|Z,J,\zeta,Y)p(\mu,\mu_J|\sigma^2,Z,J,\zeta,Y), \tag{2.34}$$

where σ^2 given (Z, J, ζ, Y) follows an inverse χ^2 distribution,

$$\begin{aligned}\sigma^2|(Z,J,\zeta,Y) \;\sim\; &\Big(\sum\nolimits_t (Y_t-\widehat{\mu}-J_tZ_t)^2+\frac{(\widehat{\mu}-a_0)^2}{b_0}+\nu\tau_0^2+\frac{1}{\xi}\sum\nolimits_t(J_t-\widehat{\mu}_J)^2\\ &+\frac{(\widehat{\mu}_J-a_J)^2}{\xi b_J}+\frac{\eta\tau_J^2}{\xi}\Big]\Big)\Big/\chi^2_{2T+\nu+\eta+2},\end{aligned} \tag{2.35}$$

each component of (μ, μ_J) given $(\sigma^2, Z, J, \zeta, Y)$ follows an independent Gaussian distribution,

$$\mu|(\sigma^2,Z,J,\zeta,Y) \;\sim\; \mathrm{N}\Big(\widehat{\mu}, \frac{\sigma^2}{T+1/b_0}\Big) \text{ and} \tag{2.36}$$

$$\mu_J|(\sigma^2,Z,J,\zeta,Y) \;\sim\; \mathrm{N}\Big(\widehat{\mu}_J, \frac{\sigma^2\xi}{T+1/b_J}\Big), \tag{2.37}$$

and $\widehat{\mu}$ and $\widehat{\mu}_J$ are given by

$$\widehat{\mu} \;=\; \frac{\sum_t(Y_t-J_tZ_t)+a_0/b_0}{T+1/b_0} \text{ and} \tag{2.38}$$

$$\widehat{\mu}_J \;=\; \frac{\sum_t J_t+a_J/b_J}{T+1/b_J}. \tag{2.39}$$

Finally, given (Z, J, ψ, Y), ξ follows an inverse χ^2 distribution,

$$\xi|(Z,J,\psi,Y) \;\sim\; \frac{\Big[\sum_t(J_t-\mu_J)^2+(\mu_J-a_J)^2/b_J+\eta\tau_J^2\Big]\Big/\sigma^2}{\chi^2_{T+\eta+1}}, \tag{2.40}$$

and λ follows a beta distribution,

$$\lambda|(Z,J,\psi,Y) \;\sim\; \mathrm{Beta}\Big(\sum\nolimits_t Z_t+\alpha, \sum\nolimits_t(1-Z_t)+\beta\Big). \tag{2.41}$$

By applying the tools described in Section 2.3, we can transform the Gibbs sampler in Figure 2.6(a) to a PMG sampler that is expected to exhibit better convergence. In Figure 2.6(b), we marginalize some components out of STEPS 3 and 4 of the Gibbs sampler given in Figure 2.6(a). Specifically, J is marginalized out of STEP 3 and $\psi_1 = (\mu, \mu_J)$ out of STEP 4. Our strategy in this marginalization is to be able to sample from standard conditional distributions even when we are conditioning on fewer components. For example, the distribution of ψ given (Z, ζ, Y), marginalized over J, is decomposed into

$$\begin{aligned} p(\psi|Z,\zeta,Y) &= \int p(\psi|Z,J,\zeta,Y)dJ \\ &= p(\sigma^2|Z,\zeta,Y)p(\mu,\mu_J|\sigma^2,Z,\zeta,Y), \end{aligned} \tag{2.42}$$

where σ^2 given (Z, ζ, Y) follows an inverse χ^2 distribution,

$$\sigma^2|(Z,\zeta,Y) \sim \frac{\sum_t \left[\frac{(Y_t-\widetilde{\mu}-\widetilde{\mu}_J Z_t)^2}{\xi Z_t+1}\right] + \frac{(\widetilde{\mu}-a_0)^2}{b_0} + \nu\tau_0^2 + \frac{1}{\xi}\left[\frac{(\widetilde{\mu}_J-a_J)^2}{b_J} + \eta\tau_J^2\right]}{\chi^2_{T+\nu+\eta+1}}, \tag{2.43}$$

and the joint distribution of μ and μ_J given (σ^2, Z, ζ, Y) is still bivariate Gaussian, i.e.,

$$\mu|(\sigma^2,\mu_J,Z,\zeta,Y) \sim \mathrm{N}\left(\widetilde{\mu}, \frac{\sigma^2}{\sum_t\left[\frac{1}{Z_t\xi+1}\right] + 1/b_0}\right) \text{ and} \tag{2.44}$$

$$\mu_J|(\sigma^2,Z,\zeta,Y) \sim \mathrm{N}\left(\widetilde{\mu}_J, \frac{\sigma^2}{\sum_t\left[\frac{Z_t}{Z_t\xi+1}\right] - \frac{\left(\sum_t\left[\frac{Z_t}{Z_t\xi+1}\right]\right)^2}{\sum_t\left[\frac{1}{Z_t\xi+1}\right]+1/b_0} + \frac{1}{\xi b_J}}\right), \tag{2.45}$$

and $\widetilde{\mu}$ and $\widetilde{\mu}_J$ are given by

$$\widetilde{\mu} = \frac{\sum_t\left[\frac{Y_t-Z_t\mu_J}{Z_t\xi+1}\right] + a_0/b_0}{\sum_t\left[\frac{1}{Z_t\xi+1}\right] + 1/b_0} \text{ and} \tag{2.46}$$

$$\widetilde{\mu}_J = \frac{\sum_t\left[\frac{Z_tY_t}{Z_t\xi+1}\right] - \frac{\sum_t\left[\frac{Z_t}{Z_t\xi+1}\right]\left(\sum_t\left[\frac{Y_t}{Z_t\xi+1}\right]+a_0/b_0\right)}{\sum_t\left[\frac{1}{Z_t\xi+1}\right]+1/b_0} + \frac{a_J}{\xi b_J}}{\sum_t\left[\frac{Z_t}{Z_t\xi+1}\right] - \frac{\left(\sum_t\left[\frac{Z_t}{Z_t\xi+1}\right]\right)^2}{\sum_t\left[\frac{1}{Z_t\xi+1}\right]+1/b_0} + \frac{1}{\xi b_J}}. \tag{2.47}$$

The distribution of ξ given (Z, J, ψ_2, Y), marginalized over ψ_1, is inverse χ^2,

$$\xi|(Z, J, \psi_2, Y) \quad \sim \quad \frac{\left[\sum_t (J_t - \widehat{\mu}_J)^2 + \frac{(\widehat{\mu}_J - a_J)^2}{b_J} + \eta\tau_J^2\right] \Big/ \sigma^2}{\chi^2_{T+\eta}}, \tag{2.48}$$

where $\widehat{\mu}_J$ is given in (2.39). We emphasize that the joint distribution after marginalization need not be a standard distribution since we hope to trim some components for each step.

In Figure 2.6(c), we permute the second and forth steps of the sampler in Figure 2.6(b), which makes the marginalized components redundant in the sampler. Thus, we are able to trim ψ_1 from STEP 2 and J from STEP 3 because these intermediate quantities are resampled in the following steps. This removal yields the PMG sampler in Figure 2.6(d). We can block the last two steps of the PMG sampler, as illustrated in Figure 2.6(e). However, the resulting set of conditional distributions remains incompatible, hence the PMG sampler is not simply a blocked Gibbs sampler.

To illustrate the computational gains we achieve, we conduct a simulation study for the jump diffusion model. For test data, we first simulate $T = 1000$ jump indicator variables Z based on $\lambda = 0.01$ that implies a jump is rare. When there is no jump, the data are simulated from a normal distribution with mean $\mu = 2$ and variance $\sigma^2 = 1$. When a jump occurs, another normal random variable with mean $\mu_J = 0$ and variance $\sigma_J^2 = \sigma^2\xi = 100$ is generated and added to the data. Using either the Gibbs sampler or PMG sampler, we run multiple chains of 10000 iterations each with different starting values to fit the Merton's jump diffusion model. The convergence of the chains is examined by computing the estimate of the potential scale reduction for all model parameters (Gelman and Rubin, 1992). The first two columns of Figures 2.7 and 2.8 compare the convergence characteristics exhibited by the Gibbs sampler and PMG sampler. We base our comparison in the first column with the last 1000 draws of a single chain, and the autocorrelations are plotted in the second column with the last 5000 draws of a single chain. From

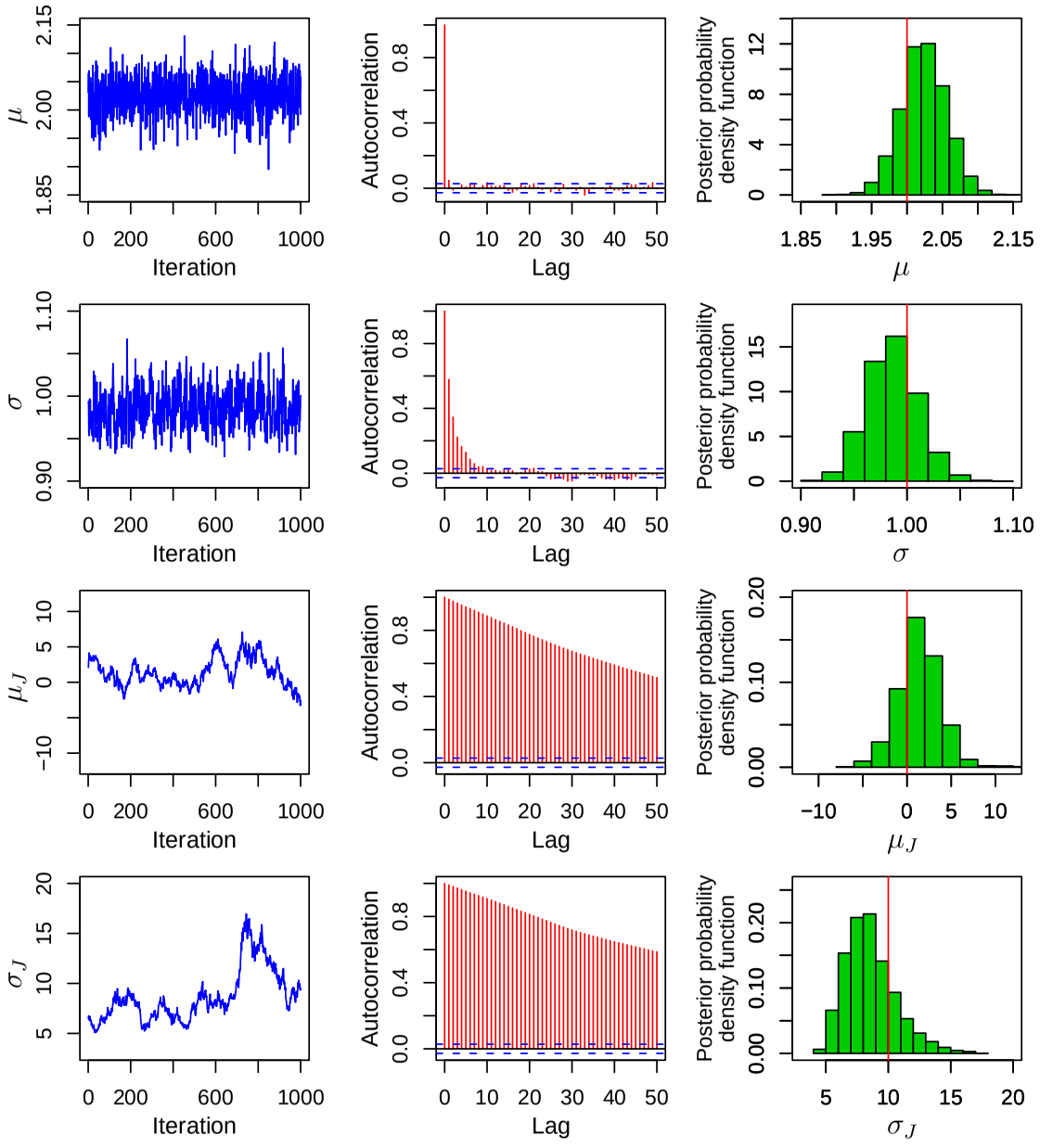

Figure 2.7: Mixing, Autocorrelation, and Marginal Posterior Distribution of Each Model Parameter Simulated by the Simple Gibbs Sampler Constructed for the Merton's Jump Diffusion Model with Proper Prior Distributions.

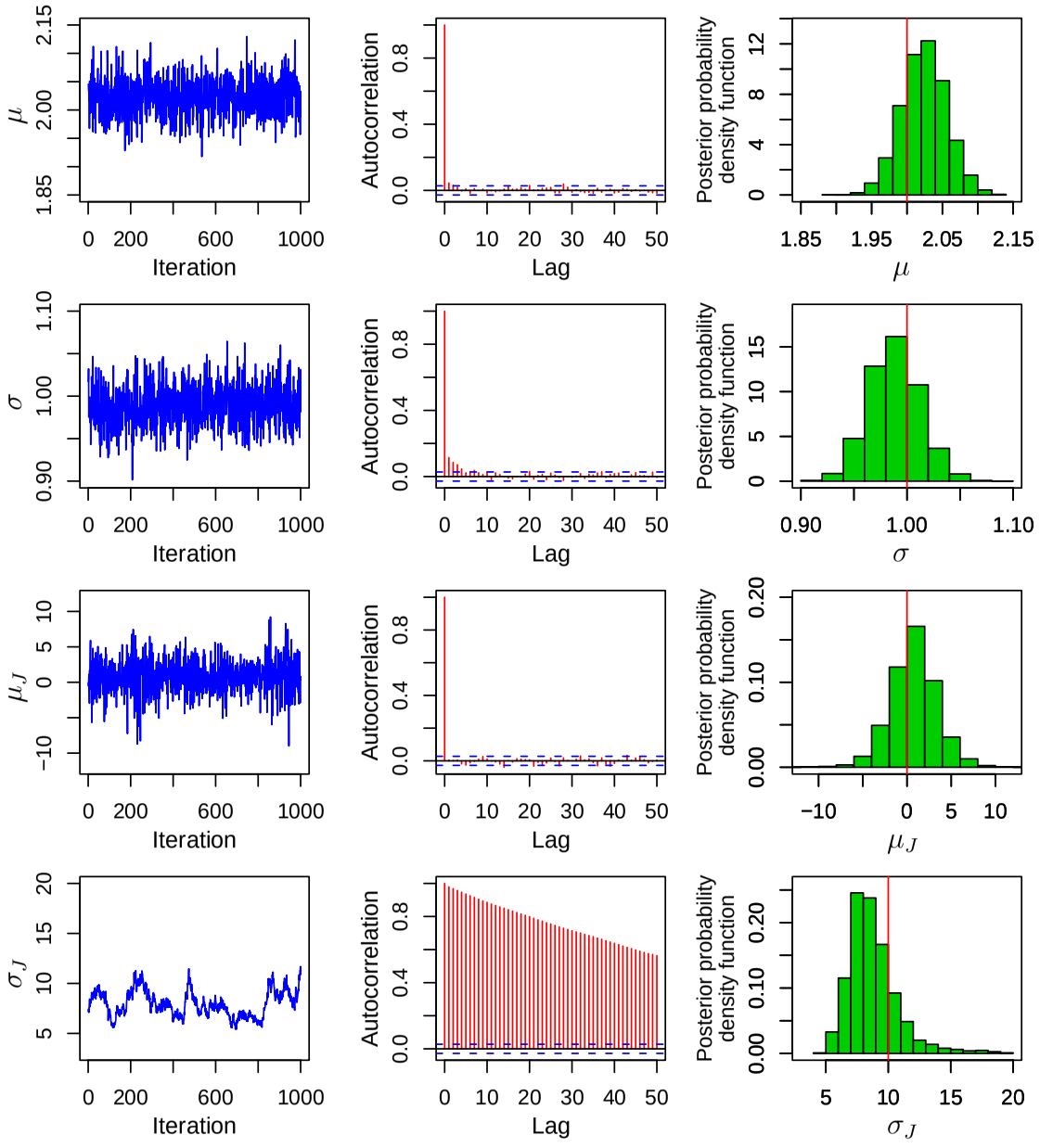

Figure 2.8: Mixing, Autocorrelation, and Marginal Posterior Distribution of Each Model Parameter Simulated by the PMG Sampler Constructed for the Merton's Jump Diffusion Model with Proper Prior Distributions.

the first two columns of Figures 2.7 and 2.8, we note the improved convergence characteristics exhibited by the PMG sampler. In particular, the effect of the partial marginalization on μ_J is substantial, and the convergence characteristics for σ is noticeably improved. The last column of Figures 2.7 and 2.8 presents the marginal posterior distributions of the selected model parameters using the second halves of the multiple chains for each sampler. The vertical solid lines represent the true values of the parameters used to simulate test data for this simulation study. We confirm that the true values are covered by the posterior distributions.

2.5.3 Multivariate Time Series Model for Joint Segmentation

The joint segmentation model for time series data from different signals in astrophysics (Dobigeon *et al.*, 2005) provides another example of a PMG sampler. Consider 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 constant Poisson intensities within time blocks constructed by sequentially combining the time bins. Thus, the likelihood for this model depends on a number of "unknown" time blocks. This difficulty makes the Gibbs sampler constructed for this model computationally infeasible. However, the PMG sampler can circumvent the computational difficulty by partially marginalizing over the Poisson intensities depending on the unknown time blocks. Thus, the partial marginalization essentially makes intractable sampling steps rather tractable with the expectation of quicker convergence. The advantage of this joint segmentation model over a single segmentation model is that we can impose our prior knowledge about the correlation structure for the joint probability of changing time blocks for different signals.

We begin by modeling the arrival of photons from different signals as an indepen-

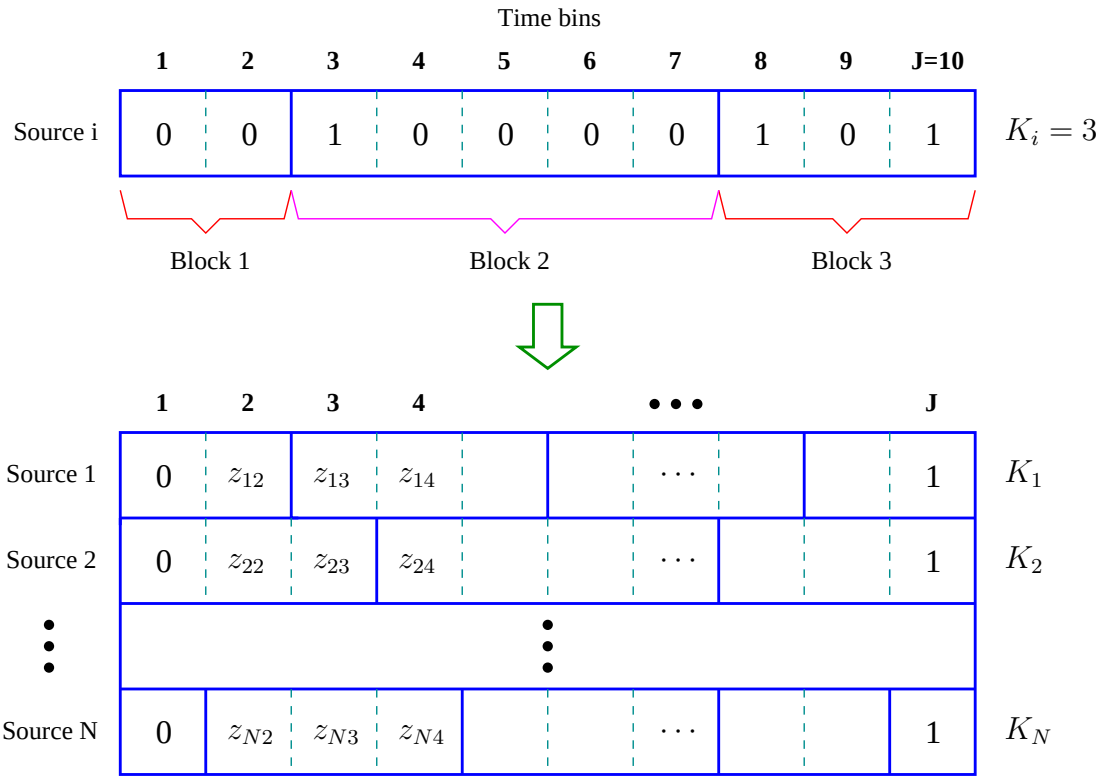

Figure 2.9: Block Indicator Matrix for Multivariate Time Series Data. The block indicator vector of signal i with $J = 10$ time bins and $K_i = 3$ time blocks is illustrated in the top. In the case of multiple signals, a $N \times J$ block indicator matrix is considered, as shown in the bottom. When the time bins are blocked by the solid lines, for example, we have $z_{12} = 0$, $z_{13} = 1$, $z_{14} = 0$, $z_{22} = 0$, $z_{23} = 0$, $z_{24} = 1$, $z_{N2} = 1$, $z_{N3} = 0$, and $z_{N4} = 0$. 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 $J - 2$ indicators are free for each signal.

dent inhomogeneous Poisson process, i.e.,

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_{ik}), \text{ for } i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K_i, \quad (2.49)$$

where Y_{ij} denotes the photon count from signal i that falls into time bin j and λ_{ik} represents the expected photon count from signal i in time block k. In words, photons from N signals are recorded in J equally spaced time bins which are segmented into K_i time blocks of signal i. Note that the Poisson intensity λ_{ik} depends on the unknown time blocks. We consider index sets of the time bins that are combined into time block k of signal i, i.e., $\{\mathcal{B}_{ik}, i = 1, 2, \ldots, N, k = 1, 2, \ldots, K_i\}$, where $\mathcal{B}_{ik} \subset \{1, 2, \ldots, J\}$ is the collection of bin indexes in block k for signal i and $\{\mathcal{B}_{ik}, k = 1, 2, \ldots, K_i\}$ for signal i are disjoint and sequential; n_{ik} denotes the cardinality of $\mathcal{B}_{ik}$.

Fitting the Poisson intensity would be greatly simplified if the time blocks were known. This leads us to consider fitting the model in the missing data framework. That is, we consider a $N \times J$ indicator matrix for the change points of blocks, $Z = (z_1, z_2, \ldots, z_J)$ where $z_j = (z_{1j}, z_{2j}, \ldots, z_{Nj})^\top$ for $j = 1, 2, \ldots, J$, and Z is treated as missing data. If $z_{ij} = 1$, there is a change point to a new block at bin j for signal i, and 0 otherwise. We set the indicator variables of the first and the last bins as $z_{i1} = 0$ and $z_{iJ} = 1$ for $i = 1, 2, \ldots, N$ to ensure that each row sum corresponds to the number of blocks for signal i, i.e., $\sum_{j=1}^{J} z_{ij} = K_i$. Figure 2.9 graphically illustrates how the indicator variables Z determine the time blocks. As a simple illustration, the block indicator vector of signal i with $J = 10$ time bins and $K_i = 3$ time blocks is shown in the top of Figure 2.9. When we have N signals, the block indicator vector becomes a $N \times J$ indicator matrix shown in the bottom of Figure 2.9.

Because z_{ij} is an indicator variable, each column vector z_j has 2^N possible configurations of 0 and 1, which is denoted by c_ℓ for $\ell = 1, 2, \ldots, 2^N$. For example, in the case of $N = 2$, we have $c_1 = (0\ 0)^\top$, $c_2 = (1\ 0)^\top$, $c_3 = (0\ 1)^\top$, and $c_4 = (1\ 1)^\top$. The probability corresponding to the configuration c_ℓ is denoted by $p_\ell = P(z_j = c_\ell)$,

and $\mathcal{P}$ is a set of the probabilities, i.e., $\mathcal{P} = \{p_\ell, \ell = 1, 2, \ldots, 2^N\}$ with $\sum_\ell p_\ell = 1$. Then, given $\mathcal{P}$, each column vector of Z follows an independent multinomial distribution,

$$z_j|\mathcal{P} \overset{\text{ind}}{\sim} \text{Multinomial}(1; \{p_\ell, \ell = 1, 2, \ldots, 2^N\}), \tag{2.50}$$

for $j = 1, 2, \ldots, J$. We denote by $S_\ell(Z)$ the number of the column vectors of Z such that $z_j = c_\ell$ for $j = 1, 2, \ldots, J$, i.e., $S_\ell(Z) = \sum_{j=1}^{J} 1_{\{z_j = c_\ell\}}$ where $1_{\{A\}}$ is equal to 1 if A is true and 0 otherwise. The Poisson intensity λ_{ik} is represented by a hierarchical structure, i.e.,

$$\lambda_{ik}|\gamma \overset{\text{iid}}{\sim} \text{Gamma}(\nu, \gamma), \text{ for } i = 1, 2, \ldots, N, \ k = 1, 2, \ldots, K_i, \tag{2.51}$$

where ν is fixed at 1 and γ is an adjustable hyperparameter. Using the conjugate Dirichlet prior distribution on $\mathcal{P}$ and the flat prior distribution on $\log \gamma$,

$$p(\mathcal{P}) \propto \prod_{\ell=1}^{2^N} p_\ell^{\alpha_\ell - 1} \quad \text{and} \quad p(\gamma) \propto \frac{1}{\gamma}, \tag{2.52}$$

the posterior distribution of interest is given by

$$\begin{aligned} p(Z, \mathcal{P}, \gamma, \lambda|Y) &\propto p(Y|Z, \lambda)p(Z|\mathcal{P})p(\lambda|\gamma)p(\mathcal{P})p(\gamma) \\ &\propto \left[\prod_{i=1}^{N}\prod_{k=1}^{K_i}\prod_{j\in\mathcal{B}_{ik}} \lambda_{ik}^{Y_{ij}} e^{-\lambda_{ik}}\right]\left[\prod_{\ell=1}^{2^N} p_\ell^{S_\ell(Z)+\alpha_\ell-1}\right]\left[\prod_{i=1}^{N}\prod_{k=1}^{K_i} \gamma e^{-\gamma\lambda_{ik}}\right]\frac{1}{\gamma} \\ &\propto \gamma^{\sum_{i=1}^{N} K_i - 1}\left[\prod_{i=1}^{N}\prod_{k=1}^{K_i} \lambda_{ik}^{\sum_{j\in\mathcal{B}_{ik}} Y_{ij}} e^{-(n_{ik}+\gamma)\lambda_{ik}}\right]\left[\prod_{\ell=1}^{2^N} p_\ell^{S_\ell(Z)+\alpha_\ell-1}\right]. \end{aligned} \tag{2.53}$$

However, because $\mathcal{P}$ can be completely marginalized out of the posterior distribution in (2.53), we can construct a simple Gibbs sampler to generate samples from the marginal distribution of the posterior distribution,

$$\begin{aligned} p(Z, \gamma, \lambda|Y) &= \int p(Z, \mathcal{P}, \gamma, \lambda|Y)d\mathcal{P} \\ &\propto \gamma^{\sum_{i=1}^{N} K_i - 1}\left[\prod_{i=1}^{N}\prod_{k=1}^{K_i} \lambda_{ik}^{\sum_{j\in\mathcal{B}_{ik}} Y_{ij}} e^{-(n_{ik}+\gamma)\lambda_{ik}}\right]\left[\frac{\prod_{\ell=1}^{2^N}\Gamma(S_\ell(Z)+\alpha_\ell)}{\Gamma\left(\sum_{\ell=1}^{2^N}(S_\ell(Z)+\alpha_\ell)\right)}\right]. \end{aligned} \tag{2.54}$$

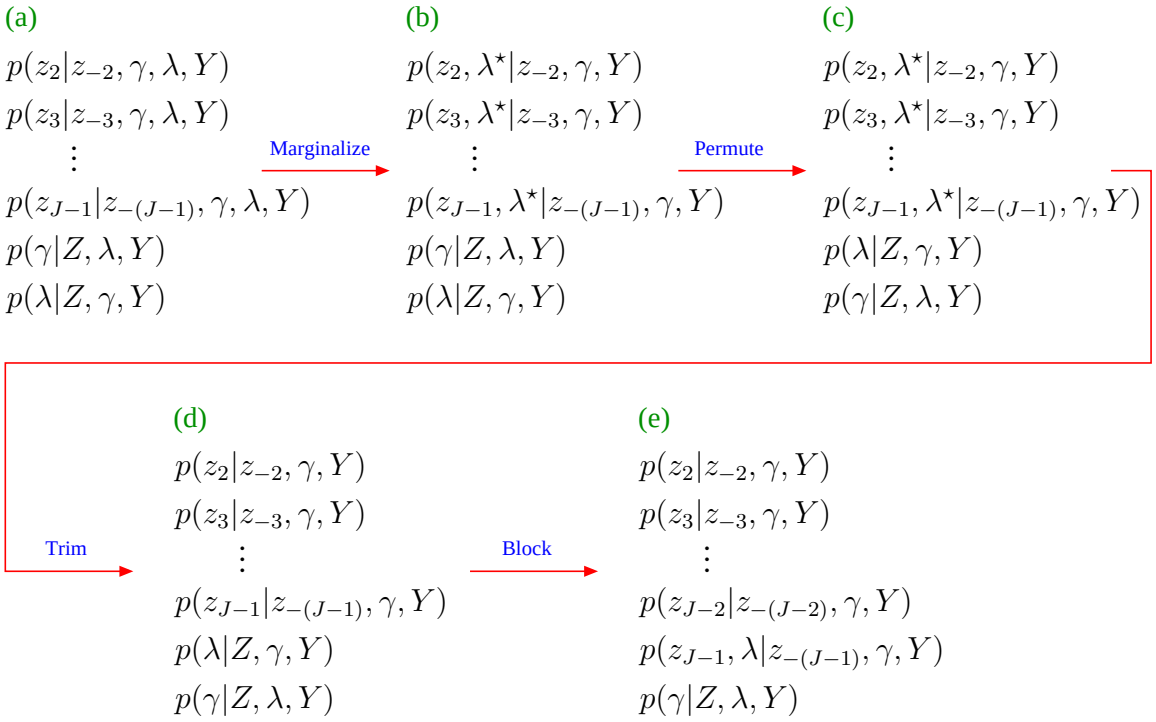

Figure 2.10: Illustration of Deriving a PMG Sampler in the Joint Segmentation Model for Multivariate Time Series Data.

Figure 2.10 illustrates the transformation of a simple Gibbs sampler into a PMG sampler for the joint segmentation time-series model, based on the marginalized target distribution in (2.54). Figure 2.10(a) shows the simple Gibbs sampler constructed using complete conditional distributions of the target posterior distribution in (2.54). Because the number of time blocks for signal i depends on the block indicator variables Z, the components of Z are not independent. However, each component of Z given $(z_{-j},\gamma,\lambda,Y)$ follows a multinomial distribution,

$$z_j|(z_{-j},\gamma,\lambda,Y) \;\sim\; \text{Multinomial}(1;\{\pi_\ell,\ell=1,2,\ldots,2^N\}), \tag{2.55}$$

where $z_{-j} = (z_1,\ldots,z_{j-1},z_{j+1},\ldots,z_J)$ and π_ℓ is proportional to $p(Z_j(\ell),\lambda,\gamma|Y)$ in (2.54) with $Z_j(\ell) = \{z_1,\ldots,z_{j-1},c_\ell,z_{j+1},\ldots,z_J\}$. Each of the other components also follows a standard distribution. That is, given (Z,λ,Y), γ follows a gamma

distribution,

$$\gamma|(Z,\lambda,Y) \;\sim\; \text{Gamma}\Big(\sum_{i=1}^{N} K_i, \sum_{i=1}^{N}\sum_{k=1}^{K_i}\lambda_{ik}\Big), \tag{2.56}$$

and given (Z,γ,Y), λ follows an independent gamma distribution,

$$\lambda_{ik}|(Z,\gamma,Y) \;\sim\; \text{Gamma}\Big(\sum_{j\in\mathcal{B}_{ik}} Y_{ij}+1, n_{ik}+\gamma\Big). \tag{2.57}$$

Unfortunately, the multinomial probabilities π_ℓ in (2.55) are not trivial to compute because changing c_ℓ in $Z_j(\ell)$ results in changing the number of blocks and thus λ_{ik} for the different number of blocks from the current K_i cannot be computed. Therefore, we emphasize that the Gibbs sampler in Figure 2.10(a) is not even feasible to implement. This difficulty, however, can be avoided by marginalizing over λ from the posterior distribution in (2.54). Namely, the conditional distribution of z_j given (z_{-j},γ,Y) is written as

$$z_j|(z_{-j},\gamma,Y) \;\sim\; \text{Multinomial}(1;\{\widetilde{\pi}_\ell, \ell=1,2,\ldots,2^N\}), \tag{2.58}$$

where the multinomial probabilities are given by

$$\begin{aligned}\widetilde{\pi}_\ell &\propto p(Z=Z_j(\ell)|\gamma,Y)\\ &\propto \frac{1}{\gamma}\left[\prod_{i=1}^{N}\prod_{k=1}^{K_i}\frac{\Gamma(\sum_{j\in\mathcal{B}_{ik}}Y_{ij}+1)}{(n_{ik}+\gamma)^{\sum_{j\in\mathcal{B}_{ik}}Y_{ij}+1}}\right]\left[\frac{\prod_{\ell=1}^{2^N}\Gamma(S_\ell(Z)+\alpha_\ell)}{\Gamma\big(\sum_{\ell=1}^{2^N}(S_\ell(Z)+\alpha_\ell)\big)}\right].\end{aligned} \tag{2.59}$$

Thus, $\widetilde{\pi}_\ell$ now depends only on γ that is free of the unknown time blocks. Such marginalization yields the Gibbs sampler in Figure 2.10(b). Here, all the marginalized quantities correspond to intermediate quantities, but not all are redundant. Because $p(\gamma|Z,\lambda,Y)$ is conditional on the intermediate quantity $\lambda^\star$, completely removing the intermediate quantities from the sampler would alter the transition kernel. To make the marginalized quantities redundant, $p(\gamma|Z,\lambda,Y)$ and $p(\lambda|Z,\gamma,Y)$ are interchanged in Figure 2.10(c). After the permutation, trimming the marginalized quantities yields the PMG sampler in Figure 2.10(d), which is not only feasible but also expected to exhibit better convergence characteristics than

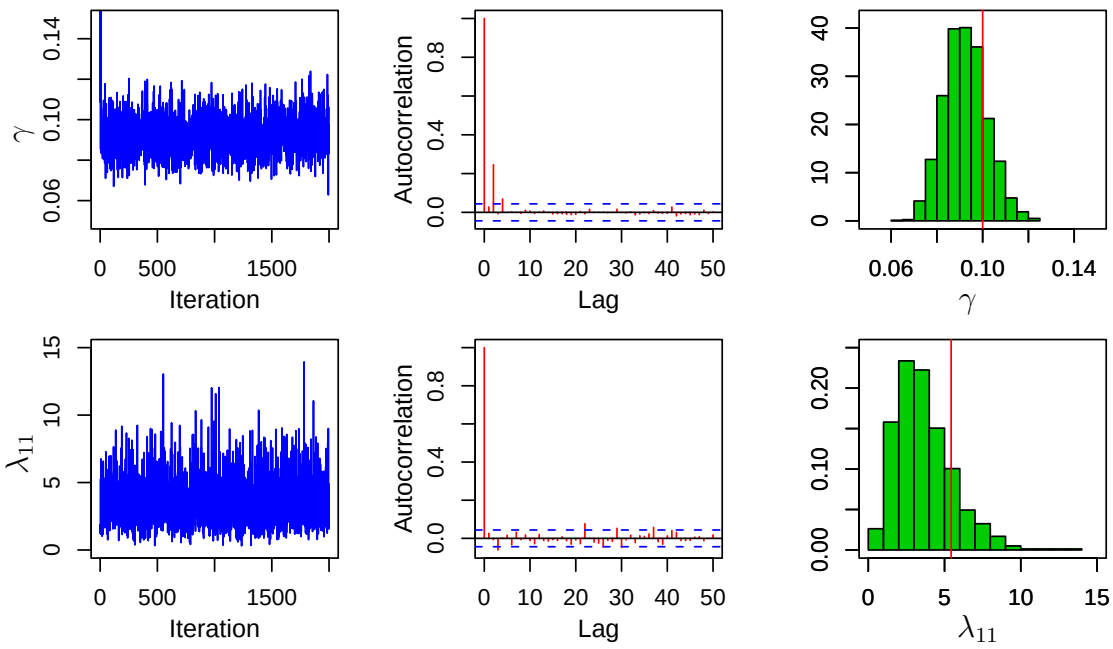

Figure 2.11: Mixing, Autocorrelation, and Marginal Posterior Distribution of γ and λ_{11} Simulated by the PMG Sampler in the Multivariate Time Series Model.

the simple Gibbs sampler in Figure 2.10(a). We can block STEPS $(J-2)$ and $(J-1)$ into $p(z_{J-1}, \lambda | z_{-(J-1)}, \gamma, Y)$, but the resulting set of conditional distributions in Figure 2.10(e) remains incompatible. Dobigeon *et al.* (2005) use the PMG sampler shown in Figure 2.10(e), and it is worthwhile noting that the sampler they construct is not a Gibbs sampler because changing the order of the steps may result in a transition kernel with unknown stationary distribution.

We conduct a simulation study for the piecewise-constant multivariate time-series model. In this simulation study, we assume a single signal (i.e., $N = 1$) and $J = 100$ time bins. We first construct $K_1 = 10$ time blocks by randomly sampling $K_1 - 1 = 9$ change points in the middle $J - 2 = 98$ time bins. For each of time blocks, we simulate λ_{1k} with $\gamma = 0.1$, and test data are generated from the model in (2.49). To fit this model, it is not possible to run the simple Gibbs sampler because of the computational difficulty, so that Figure 2.11 presents the output only from the PMG sampler. The convergence of the PMG sampler is examined by computing the

estimate of the potential scale reduction for the model parameters based on the multiple chains of 2000 iterations each with different starting values. The first two columns of Figure 2.11 illustrates the quick convergence of the model parameters, γ and λ_{11}. In the last column of Figure 2.11, the resulting marginal posterior distributions of these parameters cover the true values of the parameters based on which test data are simulated.

2.6 Discussion and Future Work

In this chapter, we present efficient Gibbs sampling techniques developed by generalizing the composition of conditional distributions in Monte Carlo samplers. Gibbs samplers are generally expected to be composed of compatible conditional distributions and, if any, simply avoid a set of incompatible conditional distributions. However, PMG samplers rather capitalize on such incompatible conditional distributions to improve the convergence characteristics of their Gibbs counterparts. The generalization of the composition comes at a price: the conditional distributions that consist of a PMG sampler need be performed in a certain order and the corresponding transition kernel has otherwise an unknown stationary distribution. For the purpose, we introduce three basic prescriptive tools, i.e., marginalization, permutation, and trimming, and by sequentially applying these tools, a Gibbs sampler is transformed into a PMG sampler. The resulting PMG sampler may be composed of a set of incompatible conditional distributions, although it may correspond to a blocked version of the original Gibbs sampler. The potential computational advantages exhibited by the PMG samplers are presented in Theorem 1 and illustrated by a couple of highly structured complex models. In terms of marginalization, PMG samplers also generalize efficient Gibbs samplers such as PX-DA schemes (Liu and Wu, 1999) and marginal data augmentation schemes (Meng and van Dyk, 1999), at least when these methods are used

with proper working prior distributions. Consequently, PMG samplers are more flexible to handle reduced conditional distributions for better convergence characteristics and can avoid computational difficulties that may be caused by fully marginalizing over some components.

Future work includes exploration into irregular but efficient MCMC samplers that can tackle the long-standing complaint of sometimes slow convergence for Gibbs samplers. By incorporating well-known frequentist tools and theorems into Bayesian methodology, we will be able to devise other efficient MCMC samplers.

Chapter 3

Fitting Narrow Spectral Lines in High Energy Astrophysics

3.1 Introduction

3.1.1 Scientific Background

X-ray observations show complex structures in both the spatial and spectral domains of astrophysical sources. However, quasars' nuclei remain spatially unresolved even with the highest-resolution X-ray telescopes. Most of the quasar energy is released within the unresolved core, and only spectral and timing information is available to study the nature of the X-ray emission.

Quasars are very luminous and compact. They emit most of their X-ray luminosity from their unresolved cores (< 1 pc). This emission is thought to originate from an accretion onto a supermassive black hole (Rees, 1978). However, the details of the accretion process or fuel supply are still uncertain, as well as the geometry of the accretion flow. It is also not clear which process is directly responsible for quasar

X-ray emission. Several components of the emission model have been identified, but the importance of their contributions to the X-ray emission may vary from source to source. The X-rays are usually associated with the non-thermal processes related to Comptonization of low-energy photons onto relativistic electrons. The Comptonization can take place in the accretion flow itself, a corona, or an outflow or innermost parts of a jet (see for example Markoff *et al.*, 2005; Sobolewska *et al.*, 2004; Sikora *et al.*, 1997).

Generally speaking, emission lines are an important part of the X-ray spectrum in that they can provide information as to the state of plasma; however, quasars' X-ray spectra are usually featureless as expected based on the Comptonization process. The only emission feature identified so far is the fluorescent Fe-K-alpha emission line. This line is thought to come directly from illuminated cold accretion flow as a fluorescent process (Fabian, 2005). The location of the line indicates the ionization state of iron in the emitting plasma, while the width of the line tells us the velocity of the plasma (Fabian, 2005).

Absorption features associated with the outflowing matter have also been observed in recent X-ray observations (Gallagher *et al.*, 2002). The location of absorption lines and the width of the lines provide information as to the velocity of the absorber and its distance from the quasar.

One of the goals of the X-ray data analysis is to understand the emission components present in the spectrum, and to obtain information about the emission and absorption features, as well as their locations and relation to the primary quasar emission. The detection of weak lines in noisy spectra is the main statistical problem in such analyses.

3.1.2 High Resolution High Energy Spectral Analysis

Testing for and identifying the location of the Fe-K-alpha line are statistically challenging tasks; see Protassov *et al.* (2002) for the method of posterior predictive p-value testing. High resolution data, such as those available with the *Chandra X-ray Observatory* carry much information as to the quasar's spectrum. Taking advantage of the high resolution spectral capacity of such instruments, however, requires careful statistical analysis. For example, the resolution of such instruments corresponds to a much finer discretization of the energy spectrum than was available with previous instruments. As a result, we expect lower counts in each bin of the X-ray spectrum. Such low-count data make the Gaussian assumptions that are inherent in traditional minimum χ^2 fitting inappropriate. A better strategy, which we employ, explicitly models photon arrivals as an inhomogeneous Poisson process (van Dyk *et al.*, 2001). In addition, data are subject to a number of processes that significantly degrade the source counts, e.g., the absorption, non-constant effective area, blurring of photons' energy, background contamination, and photon pile-up. Thus, we employ statistical models that directly account for these aspects of data collection. In particular, we design a highly structured multilevel spectral model with components for both the data collection processes and the complex spectral structures of the sources themselves. In this highly structured spectral model, a Bayesian perspective renders straightforward methods that can handle the complexity of *Chandra* data (van Dyk *et al.*, 2001; van Dyk and Kang, 2004; van Dyk *et al.*, 2006). As we shall illustrate, these methods allow us to use low-count high-resolution data, to search for the location of a narrow spectral line, to investigate its location uncertainty, and to construct statistical tests that measure the evidence in the data for including the spectral line in the source model.

3.1.3 A Statistical Model for the Spectrum

The energy spectrum of photons emitted from an astronomical source can be separated into two basic parts: a set of continuum terms and a set of emission lines. We begin with a standard spectral model that accounts for a single continuum term along with several spectral lines. Throughout this chapter, we use θ as a general representation of model parameters in the spectral model. The components of $\theta = (\theta^C, \theta^L, \theta^A, \theta^B)$ represent the collection of parameters for the Continuum, (emission) Lines, Absorption features, and Background contamination, respectively. (Notice that the roman letters in the superscripts serve as a mnemonic for these four processes.) Because the X-ray emission is measured by counting the arriving photons, we model the expected Poisson counts in energy bin $j \in \mathcal{J}$, where $\mathcal{J}$ is the set of energy bins, as

$$\Lambda_j(\theta) \;=\; \Delta_j f\big(\theta^C, E_j\big) + \sum_{k=1}^{K} \lambda_k \pi_j\big(\mu_k, \nu_k\big) \tag{3.1}$$

where Δ_j and E_j are the width and mean energy of bin j, $f(\theta^C, E_j)$ is the expected counts per unit energy due to the continuum term at energy E_j, θ^C is the set of free parameters in the continuum model, K is the number of emission lines, λ_k is the expected counts due to the emission line k, and $\pi_j(\mu_k, \nu_k)$ is the proportion of an emission line centered at energy μ_k and with width ν_k that falls into bin j. There are a number of smooth parametric forms to describe the continuum in some bounded energy range; in this chapter we parameterize the continuum term f as a power law, i.e., $f(\theta^C, E_j) = \alpha^C E_j^{-\beta^C}$ where α^C and β^C represent the normalization and the photon index, respectively. The emission lines can be modeled via the proportions $\pi_j(\mu_k, \nu_k)$ using Gaussian, Lorentzian, or delta function line profiles; the counts due to the emission line are distributed among the bins according to these proportions. While the Gaussian or Lorentzian function parameterizes an emission line in terms of center and width, the center is the only free parameter with a delta function; the width of the delta function line is necessarily zero.

While the model in (3.1) is of primary scientific interest, a more complex statistical model is needed to address the data collection processes mentioned in Section 3.1.2. We use the term *statistical model* to refer to the model that combines the *source* or *astrophysical model* with a model for the stochastic processes involved in data collection and recording. Thus, in addition to the source model, the statistical model describes such processes as instrument response and background contamination. Specifically, to account for the data collection processes, (3.1) is modified via

$$\Xi_l(\theta) \;=\; \sum_{j\in\mathcal{J}} M_{lj}\Lambda_j(\theta)d_j u(\theta^A, E_j) + \theta_l^B \tag{3.2}$$

where $\Xi_l(\theta)$ is the expected observed Poisson counts in detector channel $l \in \mathcal{L}$, $\mathcal{L}$ is the set of detector channels, M_{lj} is the probability that a photon that arrives with energy corresponding to bin j is recorded in detector channel l (i.e., M is the so-called redistribution matrix or RMF), d_j is the effective area (i.e., ARF) of bin j, $u(\theta^A, E_j)$ is the probability that a photon with energy E_j is *not* absorbed, θ^A is the collection of parameters for absorption features, and θ_l^B is a Poisson intensity of the background counts in channel l. While the scatter probability M_{lj} and the effective area d_j are presumed known from calibration, the absorption probability is parameterized using a smooth function; see van Dyk and Hans (2002) for details. For example, an important exponentiated Gaussian form of the absorption probability is expressed by setting (Freeman *et al.*, 1999)

$$u\big(\theta^A, E_j\big) = \exp\left\{-\theta_\lambda^A \exp\left[-\frac{(E_j-\theta_\mu^A)^2}{2\theta_\sigma^A}\right]\right\}, \tag{3.3}$$

where the set of free parameters $\theta^A = (\theta_\lambda^A, \theta_\mu^A, \theta_\sigma^A)$ represent the intensity, the center, and the spread of the absorption line. To quantify background contamination, a second data set is collected that is assumed to consist only of background counts; the background photon arrivals are also modeled as an inhomogeneous Poisson process.

The remainder of the chapter is organized into five sections. We begin in Section 3.2 with a toy example to illustrate the difficulties arising from the spectral model with a narrow emission line and a strategy for fitting the line. In Section 3.3, we devise a couple of efficient Gibbs samplers to fit the highly structured multilevel spectral model with either a delta function or narrow Gaussian emission line. In Section 3.4, a simulation study is performed to investigate the utility and limitation of the spectral models. Section 3.5 outlines how we fit a narrow emission line in the high redshift quasar PG1634+706, and test for the inclusion of the line in the spectral model. Concluding remarks appear in Section 3.6.

3.2 Toy Example

3.2.1 Missing Data Formulation

In order to illustrate the relevant details of fitting a line location, we designed a simplified example where we consider an *ideal instrument* that produces counts that are a mixture of continuum and emission line photons, but these counts are not subject to the data distortion processes described in Section 3.1.3. In addition, we postulate that the continuum is completely specified and one Gaussian emission line with a known width, ν_0, exists, so that the location of the emission line is the only free parameter to fit in the model. Accounting for the various forms of data distortion causes no conceptual difficulty, but obscures the ideas involved with fitting an emission line. As discussed in Section 3.3.1, the counts from the ideal instrument are one of the levels of missing data in our formulation of the model that does account for data distortion; we call these counts the *ideal counts*. We denote the ideal counts by $\mathbf{Y}_j^{\text{ideal}} = \mathbf{Y}_j^C + \mathbf{Y}_j^L$, where $\mathbf{Y}_j^{\text{ideal}}$, $\mathbf{Y}_j^C$, and $\mathbf{Y}_j^L$ are the total ideal counts, the counts due to the continuum, and the counts due to the emission line in bin j, respectively. Then we model the Poisson intensity of the

ideal counts in bin j as

$$\Lambda_j(\mu) = \Delta_j f(E_j) + \lambda \pi_j(\mu, \nu_0), \text{ for } j = 1, \ldots, J. \tag{3.4}$$

Given the ideal counts in this stylized example, it is easy to construct a simple Gibbs sampler to fit the line location μ; missing data are the ideal counts split into continuum and emission line counts, i.e., $\left\{(\mathbf{Y}_j^C, \mathbf{Y}_j^L),\ j = 1, \ldots, J\right\}$. Once the missing data are known, the line location can be fitted by using the photons due to the emission line.

This is an example of a simple finite mixture model where each observation is from one of two populations with a certain probability; more general finite mixture models have broad application in the social, biological, engineering, and physical sciences (Everitt and Hand, 1981; Aitkin and Rubin, 1985; Henna, 1985; Titterington *et al.*, 1985; Maine *et al.*, 1991; Lindsay, 1995; Pilla and Lindsay, 1996). It is well known that this finite mixture model can be embedded into a missing data model to simplify computation.

3.2.2 Simple Gibbs Sampling

In the toy example, the target joint posterior distribution of interest is given by

$$\begin{aligned} p(\mathbf{Y}^L, \mu \,|\, \mathbf{Y}^{\text{ideal}}) &\propto \left\{ \prod_{j=1}^{J} p\big(\mathbf{Y}_j^{\text{ideal}} \,|\, \mathbf{Y}_j^L, \mu\big) p\big(\mathbf{Y}_j^L \,|\, \mu\big) \right\} \cdot\ p(\mu) \\ &\propto \prod_{j=1}^{J} \left\{\lambda \pi_j(\mu, \nu_0)\right\}^{\mathbf{Y}_j^L} e^{-\lambda \pi_j(\mu, \nu_0)}, \end{aligned} \tag{3.5}$$

under the flat prior distribution, $p(\mu) \propto 1$. To generate samples from the joint distribution, we can construct a prototype two-step Gibbs sampler. Given the ideal counts, the missing counts due to the emission line in bin j follow a Binomial distribution. Thus, given the current draw of the line location, $\mu^{(t)}$, STEP 1 imputes the missing counts by

STEP 1 : Draw $(\mathbf{Y}^L)^{(t+1)}$ from $p(\mathbf{Y}^L \mid \mu^{(t)}, \mathbf{Y}^{\text{ideal}})$,

where $\mathbf{Y}_j^L \mid \mu^{(t)}, \mathbf{Y}_j^{\text{ideal}} \overset{\text{ind}}{\sim} \text{Binomial}\left(\mathbf{Y}_j^{\text{ideal}},\ \dfrac{\lambda\pi_j(\mu^{(t)}, \nu_0)}{\Delta_j f(E_j) + \lambda\pi_j(\mu^{(t)}, \nu_0)}\right)$

for $j = 1, 2, \ldots, J$.

In the next step, we draw the line location parameter given $(\mathbf{Y}^L, \mathbf{Y}^{\text{ideal}})$. To derive STEP 2, we model the energy of the photon i due to the emission line that falls into bin j as a Gaussian random variable,

$$Z_{ij} \overset{\text{iid}}{\sim} \text{N}(\mu, \nu_0^2) \text{ for } i = 1, \ldots, \mathbf{Y}_j^L \text{ and } j = 1, \ldots, J. \tag{3.6}$$

The photons due to the emission line in bin j are associated with the average of the energies of the photons, i.e.,

$$\bar{Z}_j \equiv \frac{1}{\mathbf{Y}_j^L} \sum_{i=1}^{\mathbf{Y}_j^L} Z_{ij} \overset{\text{ind}}{\sim} \text{N}\left(\mu, \frac{\nu_0^2}{\mathbf{Y}_j^L}\right) \text{ for } j = 1, \ldots, J. \tag{3.7}$$

Because of instrumental constraints, the energy of each photon that arrives at the detector is truncated to the interval corresponding to the bin, so that only the range of each photon energy is known. Because the binning is fine, however, $\bar{Z}_j$ can be approximated by the mean energy of the bin j, E_j. This simplification avoids explicitly including Z in our data augmentation scheme and focuses attention on the effects of using a narrow emission line model with binned data. In our actual data analysis Z is included in the data augmentation scheme, and exact algorithms are implemented; see van Dyk *et al.* (2001) and Section 3.3. The approximation connects (3.7) to the likelihood function $p(\mathbf{Y}_j^L|\mu)$, based on which STEP 2 draws the next iterate of the emission line location:

STEP 2 : Draw $\mu^{(t+1)}$ from $p(\mu \mid (\mathbf{Y}^L)^{(t+1)}, \mathbf{Y}^{\text{ideal}})$

$$= \text{N}\left(\frac{\sum_{j=1}^J E_j(\mathbf{Y}_j^L)^{(t+1)}}{\sum_{j=1}^J (\mathbf{Y}_j^L)^{(t+1)}},\ \frac{\nu_0^2}{\sum_{j=1}^J (\mathbf{Y}_j^L)^{(t+1)}}\right),$$

where the mean of the Gaussian distribution is the weighted average of the mean energies using the counts due to the line as weights.

3.2.3 Difficulty with Identifying Narrow Emission Lines

Although the Gibbs sampler described in Section 3.2.2 is simple, it breaks down even in this toy mixture problem if we replace the Gaussian emission line with a delta function line. Because the delta function line model fixes the width ν_0 at zero, the probability of the line that falls into bin j, $\pi_j(\mu, \nu_0)$, is 1 for the energy bin that contains the current iterate of the line location and 0 otherwise. Suppose the starting value of the line location $\mu^{(0)}$ is in bin k. Then, STEP 1 of the simple Gibbs sampler attributes all the line counts to bin k because $\pi_k(\mu, \nu_0) = 1$. Thus, STEP 2 updates the next iterate of the line location using only the $(\mathbf{Y}_k^L)^{(1)}$, so that the line location is necessarily in bin k as well:

$$\mu^{(1)} = \frac{E_k(\mathbf{Y}_k^L)^{(1)}}{(\mathbf{Y}_k^L)^{(1)}} = E_k, \tag{3.8}$$

i.e., $\mu^{(1)}$ is in the same bin k as $\mu^{(0)}$. Thus, every single state for the line location becomes an absorbing state, hence the line location parameter sticks at its starting value throughout the iteration. These absorbing states violate the irreducibility condition, so that the chain does not converge to a stationary distribution. The situation is not noticeably alleviated with a narrow Gaussian line model because STEP 2 does not result in a large change in the emission line location.

The difficulty can be avoided by devising algorithms that do not attribute photons to the emission line during the iteration. That is, we sample μ without conditioning on the missing counts due to the emission line. We thus evaluate the observed data posterior distribution $p(\mu \mid \mathbf{Y}^{\text{ideal}})$ at each possible value of the line location, and draw μ from the multinomial distribution given by

$$\mu^{(t)} \sim \text{Multinomial}\Big(1\,;\, \big\{p(\mu \mid \mathbf{Y}^{\text{ideal}})\big|_{\mu=E_m}, m = 1, 2, \ldots, J\big\}\Big), \tag{3.9}$$

where the multinomial probability in bin m is computed as

$$p(\mu \mid \mathbf{Y}^{\text{ideal}})|_{\mu=E_m} = \frac{\prod_{j=1}^{J} \Lambda_j(E_m)^{\mathbf{Y}_j^{\text{ideal}}} e^{-\Lambda_j(E_m)}}{\sum_{m=1}^{J} \prod_{j=1}^{J} \Lambda_j(E_m)^{\mathbf{Y}_j^{\text{ideal}}} e^{-\Lambda_j(E_m)}}. \tag{3.10}$$

Table 3.1: Data Augmentation in a Spectral Model.

Level	Variable	Notation
1.	The ideal data	$\ddot{\mathbf{Y}}_j^s$
2.	The **mixed** ideal data	$\ddot{\mathbf{Y}}_j^{\text{ideal}}$
3.	The mixed ideal data **after absorption**[1]	$\mathbf{Y}_j^+$
4.	The mixed and **blurred** ideal data after absorption	$\mathbf{Y}_l^+$
5.	The mixed and blurred ideal data after absorption and **background contamination**, i.e., the observed data	$\mathbf{Y}_l^{\text{obs}}$

For all variables, bin $j \in \mathcal{J}$, channel $l \in \mathcal{L}$, and $s \in S$.

Note that the toy example is illustrative and, in the actual data analysis, we construct several layers of missing data and introduce more model parameters.

3.3 A Full Spectral Analysis

3.3.1 Data Augmentation

In *Chandra* data, the photons are subject to the data distortion processes as discussed in Section 3.1.3. The method of data augmentation can simplify this convolved structure by using a hierarchical formulation. Table 3.1 describes the hierarchy of augmented data in the spectral model, which is essential to explain some stochastic features of the data collection mechanism. The data augmentation strategies described in this section follow van Dyk *et al.* (2001). In Table 3.1, the set S represents the collection of the continuum and emission line sources while a "+" in the superscript indicates the mixed photons of all components in S. To explicitly account for a series of the data contamination processes, we treat the ideal data from Level 1 to Level 4 in Table 3.1 as missing data. The mixed ideal data in Level 2 represents source counts under the ideal instrument considered in Section 3.2.1 (i.e., no response matrix, constant effective area, no absorption, and no

[1] In the statistical model, the effective area of the instrument is handled in exactly the same way as absorption. Thus, in this table absorption includes the effective area of the instrument.

background contamination). In real data, observed photons are degraded by these effects and thus need to be appropriately deconvolved into the ideal ones.

Going down one by one from Level 2 in Table 3.1 explains that the photon counts are contaminated by absorption and effective area, blurring, and background counts, respectively. For example, the effect of the absorption and effective area is accounted for modeling the data in Level 3 given the data in Level 2, which is formulated as

$$p(\dot{\mathbf{Y}}_j^+ \mid \ddot{\mathbf{Y}}_j^{\text{ideal}}, \theta) = \text{Binomial}\big(\ddot{\mathbf{Y}}_j^{\text{ideal}},\ d_j u(\theta^A, E_j)\big), \quad j \in \mathcal{J}. \tag{3.11}$$

The photon energies of the data in Level 3 are blurred by a series of multinomial distributions implied by the columns of the redistribution matrix M. Thus, the distribution of energy channel counts in Level 4 is the sum of the multinomial distributions over the bins, i.e.,

$$p(\mathbf{Y}^+ \mid \dot{\mathbf{Y}}^+, \theta) = \sum_{j \in \mathcal{J}} \text{Multinomial}\big(\dot{\mathbf{Y}}_j^+,\ M_j\big), \tag{3.12}$$

where $\mathbf{Y}^+ = \{\mathbf{Y}_l^+, l \in \mathcal{L}\}$, $\dot{\mathbf{Y}}^+ = \{\dot{\mathbf{Y}}_j^+, j \in \mathcal{J}\}$, and M_j is the jth column of M, $j \in \mathcal{J}$. We model the background counts as another independent Poisson variables, hence the observed data in Level 5 are modeled as

$$p(\mathbf{Y}_l^{\text{obs}} \mid \mathbf{Y}_l^+, \theta) = \mathbf{Y}_l^+ + \text{Poisson}(\theta_l^B), \quad l \in \mathcal{L}. \tag{3.13}$$

That is, given the model parameters the missing data can be sequentially modeled by using standard probability distributions, and given the missing data fitting the parameters is straightforward. Thus, when an emission line is modeled with a broad Gaussian distribution with fixed width, the standard EM, or expectation/maximization, algorithm (Dempster *et al.*, 1977) and DA, or data augmentation, scheme (Tanner and Wong, 1987) can be constructed to fit the spectral model; refer to van Dyk *et al.* (2001) and van Dyk and Kang (2004) for details. If the Gaussian line profile is replaced with a delta function or a narrow Gaussian distribution

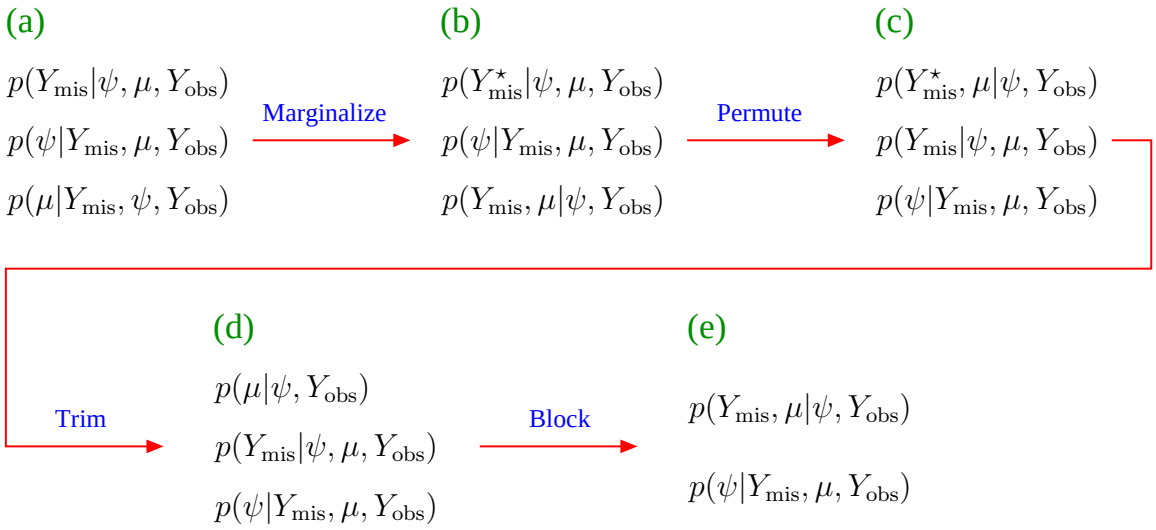

Figure 3.1: Illustration of Deriving PMG I for the Multilevel Spectral Model with Delta Function Emission Lines.

with unknown width, however, we need more sophisticated missing data algorithms due to the difficulty discussed in Section 3.2; refer to van Dyk and Park (2004) and Park (2004) for efficient EM-type algorithms.

3.3.2 Constructing Efficient Gibbs Samplers

Efficient Gibbs Samplers for the Delta Function Line Profile

The use of a narrow emission line gives rise to computational challenges that deteriorate the convergence rate of the standard missing data algorithms. With the delta function line model, we thus devise two variants of a PMG sampler, introduced in Chapter 2. Figures 3.1 and 3.2 illustrate the derivation of PMG I and PMG II, respectively.

In Figure 3.1(a), we begin by constructing the simple Gibbs sampler using the complete conditional distributions of target posterior distribution, $p(Y_{\text{mis}}|\psi,\mu,Y_{\text{obs}})$, where $Y_{\text{mis}} = \{(\ddot{\mathbf{Y}}^s_j, \ddot{\mathbf{Y}}^{\text{ideal}}_j, \dot{\mathbf{Y}}^+_j, \mathbf{Y}^+_l), j \in \mathcal{J}, l \in \mathcal{L}\}$, $Y_{\text{obs}} = \{\mathbf{Y}^{\text{obs}}_l, l \in \mathcal{L}\}$, and we denote by μ the line location parameter and by ψ the rest of the model parameters.

As illustrated in Section 3.2, however, the simple Gibbs sampler breaks down because the location of the delta function line does not move from its starting value. To improve the rate of convergence, we capitalize on a conditional distribution that marginalizes over the entire missing data Y_{mis} in STEP 3 of Figure 3.1(b). These marginalized missing data are part of the output quantities, hence they cannot be removed from the sampler. In Figure 3.1(c), the sampling steps in Figure 3.1(b) are permuted to make the marginalized quantities redundant in the sampler. Because the marginalized quantities do not correspond to the output quantities and are not conditioned upon in the subsequent steps, we remove Y_{mis} from STEP 1 in Figure 3.1(c). This removal results in PMG I in Figure 3.1(d). Specifically, STEP 1 draws μ from the multinomial distribution,

$$\mu^{(t+1)} \sim \text{Multinomial}\Big(1\,;\ \big\{p(\mu \mid \psi^{(t+1)}, Y_{\text{obs}})\big|_{\mu=E_j},\ j \in \mathcal{J}\big\}\Big), \tag{3.14}$$

by evaluating the observed posterior distribution with the flat prior distribution,

$$p(\mu \mid \psi^{(t+1)}, Y_{\text{obs}}) \propto \prod_{l \in \mathcal{L}} \Xi_l(\theta)^{\mathbf{Y}_l^{\text{obs}}}\ e^{-\Xi_l(\theta)}, \tag{3.15}$$

at the possible line locations. Then we note that STEPS 1 and 2 in Figure 3.1(d) can be combined into $p(Y_{\text{mis}}, \mu|\psi, Y_{\text{obs}})$, yielding the prototype two-step Gibbs sampler in Figure 3.1(e); in this case, the PMG sampler is a blocked version of the original Gibbs sampler in Figure 3.1(a).

In PMG I, marginalizing over the entire missing data involve additional evaluation of the observed posterior distribution evolving the large dimensional blurring matrix M. Thus, each iteration of PMG I is computationally expensive to compute and this difficulty persists even when sparse matrix techniques are implemented; PMG II is devised to avoid the costly procedure. As before, we begin with the simple Gibbs sampler in Figure 3.2(a). Then we marginalize over only part of missing data, $Y_{\text{mis2}} = \{(\ddot{\mathbf{Y}}_j^s, \ddot{\mathbf{Y}}_j^{\text{ideal}}), j \in \mathcal{J}\}$, to avoid the procedure accounting for the blurring in the evaluation of multinomial probabilities. This partial marginalization results in the sampler in Figure 3.2(b). The marginalized quantities Y_{mis2}, however,

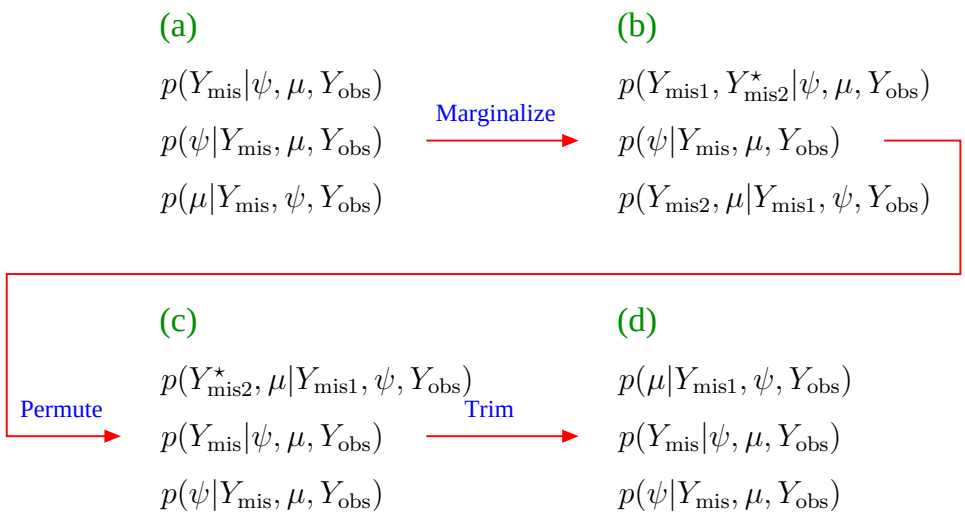

Figure 3.2: Illustration of Deriving PMG II for the Multilevel Spectral Model with Delta Function Emission Lines.

cannot be removed because doing so changes a transition kernel with the target posterior distribution. Permuting the steps in Figure 3.2(c) makes the marginalized quantities redundant in the sampler, hence they can be removed from STEP 1 without altering the transition kernel. Figure 3.2(d) shows the resulting PMG sampler, PMG II, which is constructed using a set of incompatible conditional distributions. In this case, the sampling steps in Figure 3.2(d) cannot be combined. In particular, STEP 1 of PMG II draws μ from the multinomial distribution,

$$\mu^{(t+1)} \sim \text{Multinomial}\Big(1\,;\ \big\{p(\mu\,|\,Y^{(t+1)}_{\text{mis1}},\psi^{(t+1)},Y_{\text{obs}})\big|_{\mu=E_j},\ j\in\mathcal{J}\big\}\Big), \tag{3.16}$$

where $Y_{\text{mis1}} = \{(\dot{\mathbf{Y}}^{+}_{j},\mathbf{Y}^{+}_{l}), j\in\mathcal{J}, l\in\mathcal{L}\}$ and the multinomial probabilities are computed by evaluating the marginalized posterior distribution with the flat prior distribution,

$$\begin{aligned} p(\mu\,|\,Y^{(t+1)}_{\text{mis1}},\psi^{(t+1)},Y_{\text{obs}}) &= \int p(Y_{\text{mis2}},\mu\,|\,Y^{(t+1)}_{\text{mis1}},\psi^{(t+1)},Y_{\text{obs}})dY_{\text{mis2}} \\ &\propto \prod_{j\in\mathcal{J}}\Big\{\Lambda_j(\theta)d_j u(\theta^A,E_j)\Big\}^{\dot{\mathbf{Y}}^{+}_{j}}\ e^{-\Lambda_j(\theta)d_j u(\theta^A,E_j)} \end{aligned} \tag{3.17}$$

at the possible line locations.

PMG I partially marginalizes over the entire missing data, while PMG II partially marginalizes over only part of the missing data. Thus, we expect PMG I to exhibit better convergence characteristics than PMG II when the chains of both samplers are run for the same number of iterations. In this spectral model, however, we must compare the computation time rather than the simple number of iterations, because each iteration of PMG I is much more costly than that of PMG II that avoids evolving the high dimensional redistribution matrix. Although both samplers performs similarly for the same amount of computation time, we employ PMG I to fit the spectral model with a delta function line throughout this chapter.

Efficient Gibbs Samplers for the Gaussian Line Profile

When high-resolution instruments such as the *Chandra X-ray Observatory* are used to collect photons, a narrow Gaussian distribution may be more appropriate to model emission lines because the binning of the energy spectrum is fine. However, when the location and width of an emission line are simultaneously fitted, the simple Gibbs sampler breaks down as the line gets narrower: That is, the current iterate of the line location parameter tends to get stuck at the previous iterate. This leads us to consider fitting the Gaussian line location and/or width without the missing data. In the case of the Gaussian emission line, we again present only two variants of a PMG sampler, PMG III and PMG IV, although many different samplers can be constructed with partial marginalization.

Figures 3.3 and 3.4 illustrate how to derive PMG III and PMG IV, respectively. To begin with, the simple Gibbs sampler, shown in Figures 3.3(a) and 3.4(a), is constructed from the target distribution $p(Y_{\rm mis}|\psi,\mu,\nu,Y_{\rm obs})$, where $Y_{\rm mis} = \{(\ddot{\mathbf{Y}}_j^s, \ddot{\mathbf{Y}}_j^{\rm ideal}, \dot{\mathbf{Y}}_j^+, \mathbf{Y}_l^+), j \in \mathcal{J}, l \in \mathcal{L}\}$, $Y_{\rm obs} = \{\mathbf{Y}_l^{\rm obs}, l \in \mathcal{L}\}$, and we denote by μ the line location parameter, by ν the line width, and by ψ the rest of the model parameters. To improve convergence characteristics, Figure 3.3(b) partially marginalizes

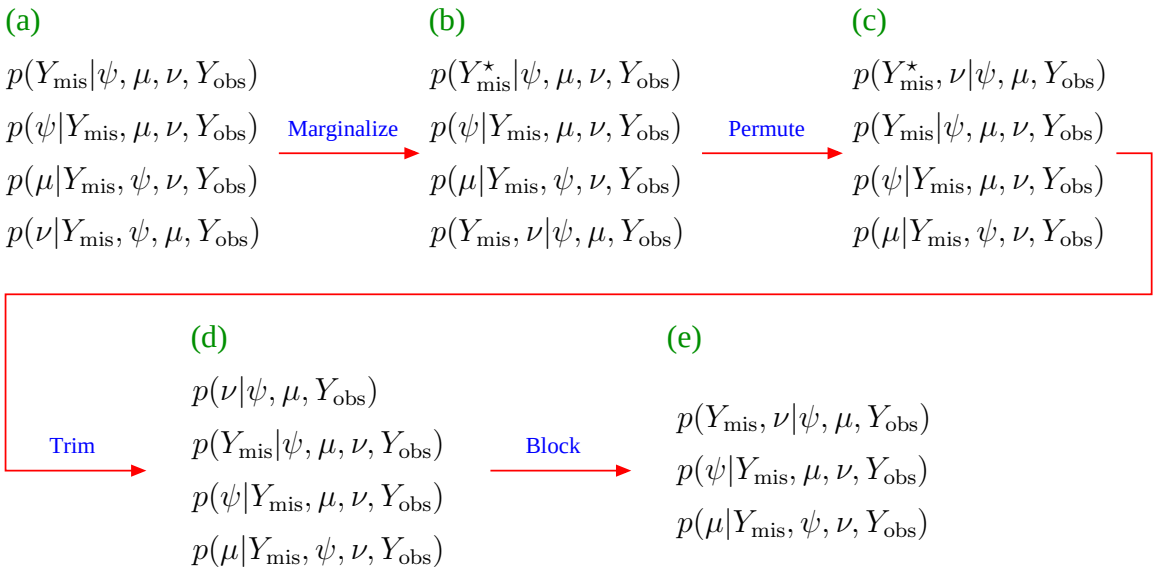

Figure 3.3: Illustration of Deriving PMG III for the Multilevel Spectral Model with Gaussian Emission Lines.

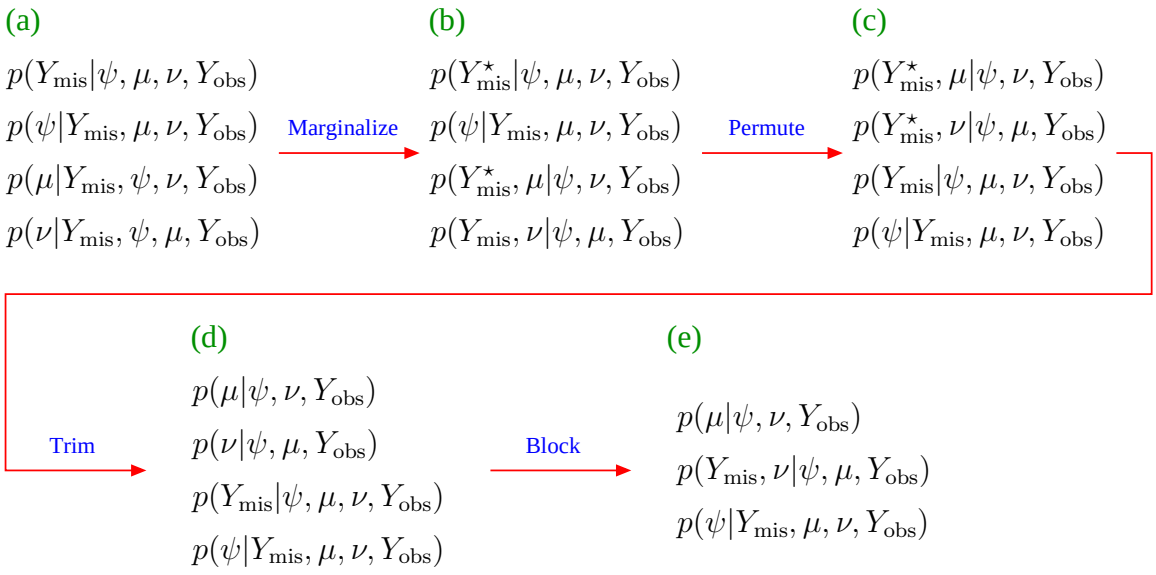

Figure 3.4: Illustration of Deriving PMG IV for the Multilevel Spectral Model with Gaussian Emission Lines.

the entire missing data out of only the sampling step for the line width, while the missing data are marginalized out of both steps for the line location and width in Figure 3.4(b). Permuting the steps in the resulting samplers ensures that the marginalized quantities correspond to the intermediate quantities that are not conditioned upon, as shown in both Figures 3.3(c) and 3.4(c). Removing these intermediate quantities from the samplers yields PMG III in Figure 3.3(d) and PMG IV in Figure 3.4(d). Notice that STEPS 1 and 2 in Figure 3.3(d) are combined into one sampling step, so that the four-step Gibbs sampler in Figure 3.3(a) is transformed into the three-step Gibbs sampler in Figure 3.3(e). However, after combining STEPS 2 and 3 in Figure 3.4(d), the three-step PMG sampler in Figure 3.4(e) is still composed of incompatible conditional distributions and thus does not correspond to a blocked version of the original Gibbs sampler in Figure 3.4(a).

In both PMG samplers, the entire missing data are partially marginalized out, but PMG IV marginalizes Y_{mis} out of more steps than PMG III. Thus we expect PMG IV to outperform PMG III in terms of convergence, although each iteration of PMG IV is more costly because of additionally evaluating the observed posterior distribution to draw the line location without the missing data. Throughout this chapter, we employ PMG IV to fit the spectral model with a narrow Gaussian emission line.

3.4 Simulation Study

Our simulation study is conducted to assess the validity of the highly structured multilevel spectral model discussed in Section 3.1.3 and to illustrate computation methods to fit the model. We consider the following four CASES that we believe are representative of the cases that are of general interest:

CASE 1 : There is no emission line in the spectrum.

CASE 2 : There is a narrow and weak emission line at 2.85 keV with width 0.04 keV

Table 3.2: Posterior Modes of the Line Locations Fitted with the EM-type Algorithm.

Line profile model	Simulated data set	True line location (keV)	Posterior mode (keV)	Domain of convergence (keV)
Delta function line	CASE 1	·	3.945	0.6 – 1.9 and 2.1 – 2.2
		·	4.435	0.5, 2.0, and 2.3 – 6.0
	CASE 2	2.85	2.845	0.5 – 6.0
	CASE 3	3.40	3.385	0.5 – 6.0
	CASE 4	3.40	3.315	0.5 – 6.0
Gaussian line	CASE 1	·	4.045	0.5 – 6.0
	CASE 2	2.85	2.625	0.5 – 6.0
	CASE 3	3.40	3.405	0.5 – 6.0
	CASE 4	3.40	3.335	0.5 – 6.0

in the spectrum.

CASE 3 : There is a broad and strong emission line at 3.40 keV with width 0.207 keV in the spectrum.

CASE 4 : There is a narrow and strong emission line at 3.40 keV with width 0.04 keV in the spectrum.

For each of these CASES, we simulate a test data set with 1500 counts similar to the observed number of counts in the *Chandra* X-ray spectrum of PG1634+706 analyzed in Section 3.5, mimicking the real data situation. In the simulation, we assume no background contamination and use a power law continuum with $\alpha^C = 3.728e-5$ and $\beta^C = 1.8$. Our simulation is done with Sherpa (Freeman *et al.*, 2001) software in CIAO[2], assuming the *Chandra* responses (ARF and RMF files). The test data are used to fit the spectral model, then we compare the true values of the line location with its estimates and evaluate the evidence for including the line in the X-ray spectrum; refer to Section 3.5 for more details of our spectral analysis. This is the frequentist evaluation of a Bayesian program in that we test software by

[2]The software is publicly available on `http://cxc.harvard.edu/ciao/`.

applying it to data simulated with the *fixed* values of model parameters. To statistically assess the correctness of the Bayesian model-fitting software, we rather simulate the true values of model parameters from their proper prior distributions, fit a model for test data generated with each sample of the parameters, and examine estimated posterior quantiles to detect errors in the software; refer to Cook *et al.* (2006) for more details of this Bayesian simulation study.

We begin by running the Rotation(9) EM-type algorithm (van Dyk and Park, 2004) using different starting values to ensure that all of the important posterior modes are identified. Specifically, we use 56 starting values equally spaced between 0.5 keV and 6.0 keV, and run the algorithm twice for each starting value using the spectral models with both delta function and Gaussian lines. The results of the 56 runs for each of the four test data sets and both models are presented in Table 3.2. In particular, multiple modes are identified for the test data generated under CASE 1 when the delta function line is used in the spectral model. This results from the fact that there is no emission line in the spectrum and the variation around the continuum causes a multimodal posterior distribution of the delta function line location. Because we know the true line locations for CASES 2, 3, and 4, we can compare them with their estimates identified using the delta function and Gaussian lines. In Table 3.2, we notice that the line locations identified with the EM-type algorithm are near the true line locations in CASES 2, 3, and 4. The difference between the true values and the estimates can be calibrated using error bars on the estimates, the topic that we now turn to.

The use of Monte Carlo samplers allows us to investigate the surface for the posterior distribution of the line location. Figure 3.5 presents the resulting posterior distributions of the delta function and Gaussian line locations in the different CASES; the vertical solid lines represent the true line locations. When there is no emission line in the spectrum (i.e., CASE 1), the posterior distribution of the delta function line location tends to be multimodal and the Gaussian line location has a very dif-

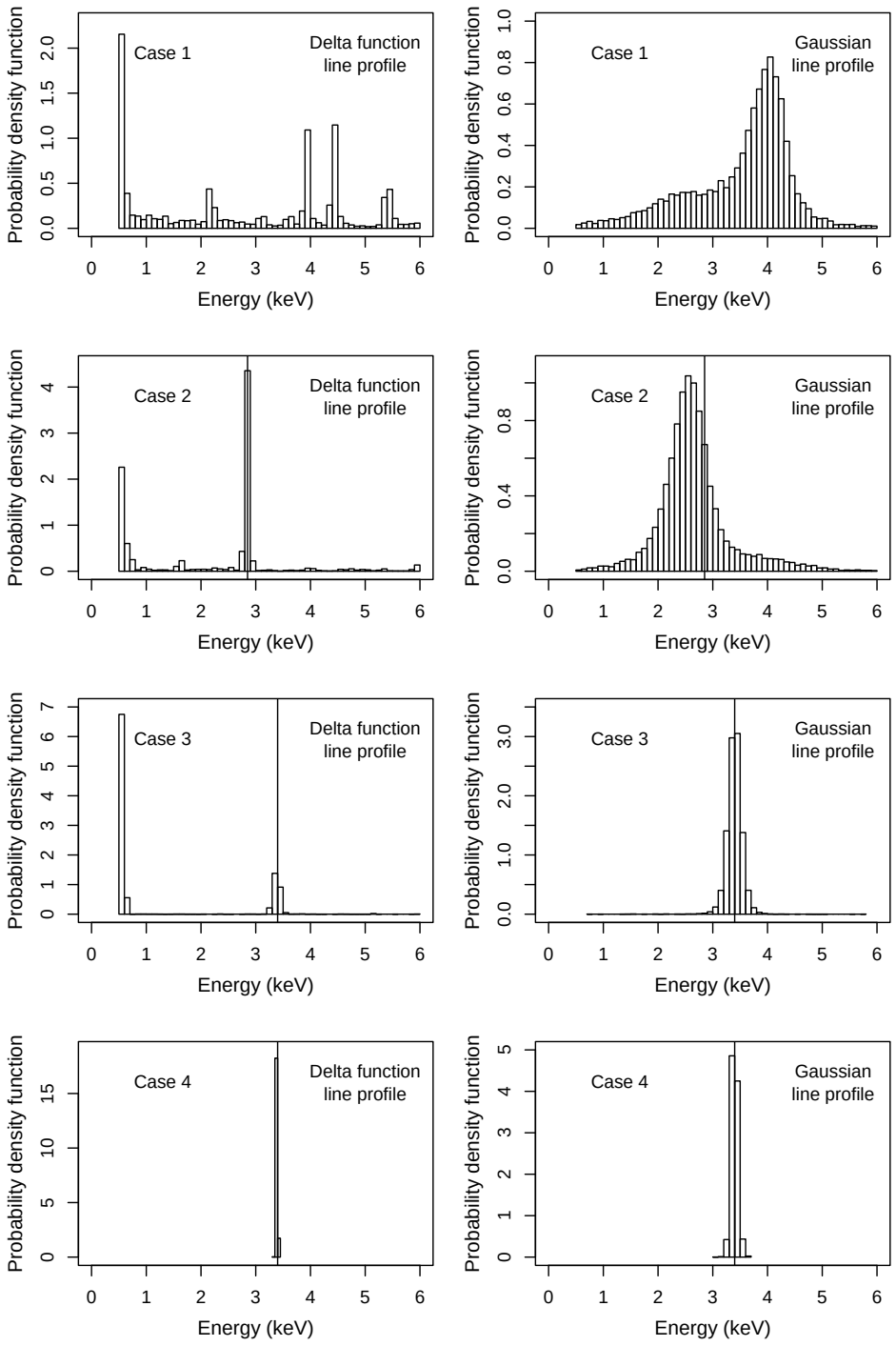

Figure 3.5: Posterior Distributions of the Line Location Using the Four Different CASES Fitted for the Spectral Models with the Delta Function and Gaussian Lines. The vertical solid lines represent true values of the line location for which test data are simulated. In CASE 1, there is no emission line in the spectrum. In CASES 2, 3, and 4, there is a significant posterior mode near the true value of the line location.

Table 3.3: 95% HPD Regions or 95% Posterior Intervals for the Line Location.

Line profile model	Simulated data set	True line location (keV)	95% HPD region	Posterior probability
Delta function line	CASE 1	·	(0.50, 1.45)	34.36%
		·	(2.02, 2.84)	11.55%
		·	(3.79, 4.20)	14.61%
		·	(4.33, 4.67)	15.65%
		·	(5.28, 5.64)	9.15%
	CASE 2	2.85	(0.50, 0.83)	31.27%
			(2.70, 2.97)	50.06%
	CASE 3	3.40	(0.52, 0.62)	72.29%
			(3.28, 3.47)	23.06%
	CASE 4	3.40	(3.36, 3.41)	98.25%
Gaussian line	CASE 1	·	(1.45, 4.80)	93.55%
	CASE 2	2.85	(1.44, 3.96)	92.18%
	CASE 3	3.40	(3.13, 3.67)	95.01%
	CASE 4	3.40	(3.28, 3.52)	95.32%

fuse posterior distribution, as shown in the first row of Figure 3.5. This implies that the identified line location(s) shown in Table 3.2 may not be well specified. When an emission line is narrow and weak (i.e., CASE 2), the spectral model with the delta function line seems better suited for fitting the test data than with the Gaussian line, as illustrated in the second row of Figure 3.5. The posterior mode of the delta function line location precisely estimates the true line location, but that of the Gaussian line location is much less precise. In the case of a broad and strong line (i.e., CASE 3), the spectral model with the Gaussian line seems better suited and in this case the posterior distribution of the delta function line location has a local mode near the true line location. With a narrow and strong line, spectral models with both lines agree with each other and produce the similar posterior distribution of the line location, as shown in the last row of Figure 3.5.

Based on the posterior distributions of the line location in Figure 3.5, we compute the highest posterior density (HPD) region to evaluate the uncertainty of the fitted

line location(s); when a posterior distribution is multimodal, the HPD region may consist of a number of disjoint intervals. Table 3.3 presents the 95% HPD region for the line location; the combined posterior probability in the last column may not add up to 95% because we list only the intervals whose posterior probabilities are greater than 5%. The true line locations are presented in the third column of Figure 3.5 and, in all CASES, are contained in the 95% HPD region.

Posterior predictive methods (Rubin, 1981, 1984; Gelman and Meng, 1996; Gelman *et al.*, 1996) can be employed to check the spectral model specification. The methods aim to check the self-consistency of the model, i.e., the ability of the fitted model to predict the data to which the model is fit. To quantify the evidence for the inclusion of the line in the spectrum, we consider the spectral model discussed in Section 3.1.3 with three different line models:

MODEL 0 : There is no emission line in the spectrum.

MODEL 1 : There is a delta function emission line in the spectrum.

MODEL 2 : There is a Gaussian emission line in the spectrum.

The posterior predictive distribution is used to generate simulated data $\left\{y_{\rm rep}^{(\ell)}, \ell = 1, 2, \ldots, 1000\right\}$ under MODEL 0. We compare the simulated data to the observed data via the likelihood ratio test statistic,

$$T_m(y_{\rm rep}^{(\ell)}) = \log\left\{\frac{\sup_{\theta\in\Theta_m} L(\theta|y_{\rm rep}^{(\ell)})}{\sup_{\theta\in\Theta_0} L(\theta|y_{\rm rep}^{(\ell)})}\right\}, \;\; m = 1,\, 2, \text{ and } \ell = 1, 2, \ldots, 1000, \qquad (3.18)$$

where Θ_0, Θ_1, and Θ_2 represent the parameter spaces under MODELS 0, 1, and 2, respectively. In particular, we generate 1000 samples from the posterior predictive distribution under MODEL 0, compute the test statistics for the posterior predictive samples, and compare the histogram of the resulting 1000 test statistics, $T_m(y_{\rm rep}^{(\ell)})$, to the observed test statistic, $T_m(Y_{\rm obs})$, under MODEL m for $m = 1,\, 2$. As illustrated in Figure 3.6, this comparison yields a posterior predictive p-value (or ppp-value)

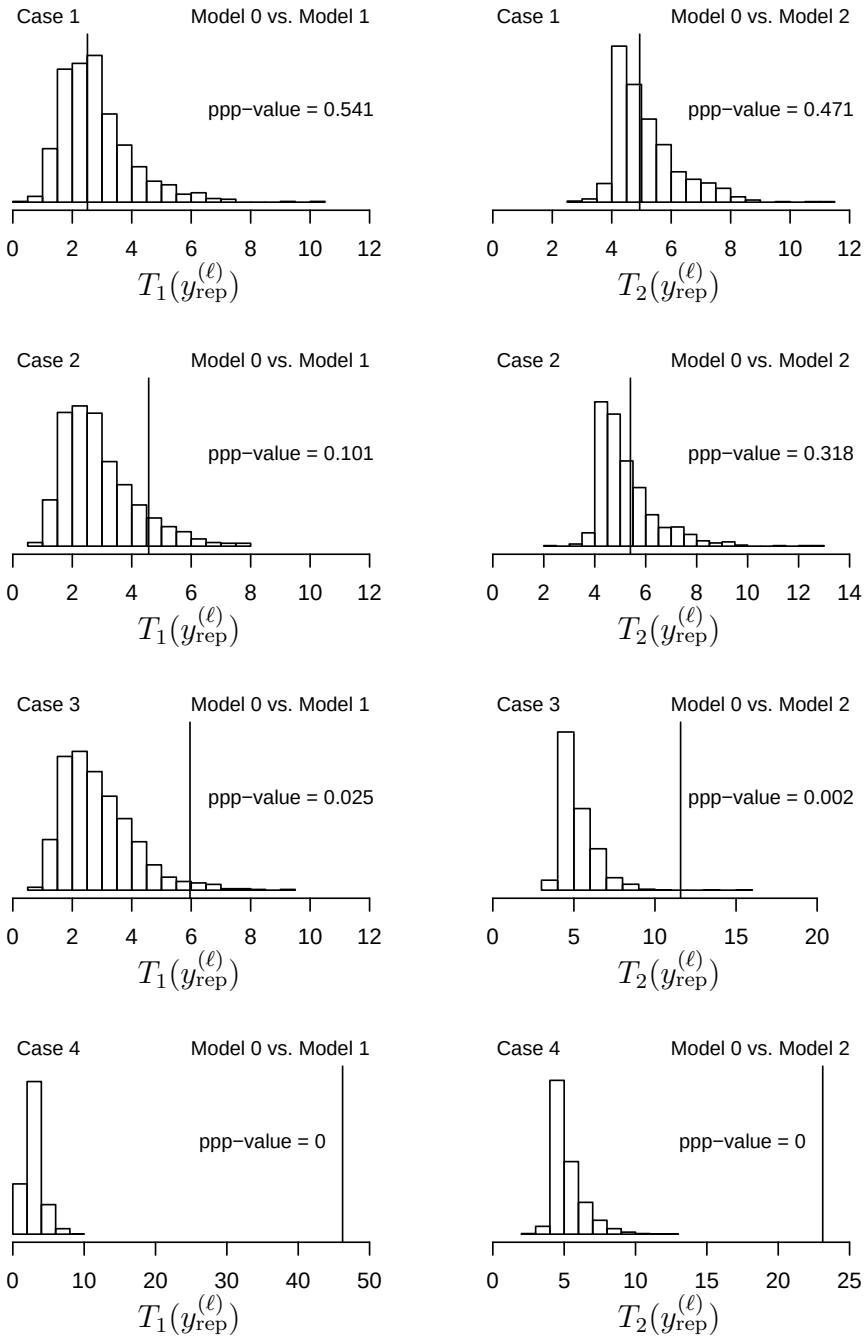

Figure 3.6: Posterior Predictive Checks of the Data Simulated Using the Four Different CASES. The left panels test for the evidence of the delta function emission line, and the right panels test for the Gaussian emission line. The vertical solid lines represent the observed test statistics which are compared with the test statistics from 1000 data sets simulated from the posterior predictive distribution under MODEL 0.

that is the proportion of the simulated test statistics that are as extreme as or more extreme than the observed test statistic (Rubin, 1984; Meng, 1994). Smaller ppp-values give more evidence for the alternative model, i.e., MODEL 1 or MODEL 2, thereby supporting the inclusion of the line in the spectrum in our case. As we can expect, there is not strong evidence of a line in CASE 1 because the first row of Figure 3.6 shows large ppp-values for CASE 1. In CASE 2, we have a narrow but weak emission line, which yields fairly strong evidence for the inclusion of the delta function line but not for the Gaussian line, as shown in the second row of Figure 3.6. For the spectral model with either line, the strong lines in CASES 2 and 3 give strong evidence for including a spectral line in the model, as illustrated in Figure 3.6.

The Fe-K-alpha emission line previously identified in the quasar's X-ray spectrum is narrow and weak, as in CASE 2. In this case, the simulation study illustrates that the line location fitted for the spectral model with the Gaussian line may not be precise and the model does not provide strong evidence of the narrow and weak line location in the spectrum. In the same case, however, the spectral model with the delta function line gives fairly strong evidence for the inclusion of the narrow and weak line in the spectrum. Based on this simulation study, we also notice that the spectral model with the delta function line more precisely identifies the true line locations, regardless of the width or strength. As such, we prefer fitting the delta function emission line for the real data of the high redshift quasar PG1634+706 analyzed in Section 3.5.

Table 3.4: Description of the *Chandra* Observations for PG1634+706

Observed data set	Exposure time (sec.)	Total counts
obs-id 47	5389.08	1651
obs-id 62	4854.57	1472
obs-id 69	4859.42	1457
obs-id 70	4859.68	1419
obs-id 71	4405.57	1356
obs-id 1269	10834.03	2216

3.5 Analysis of the Quasar PG1634+706

3.5.1 The High Redshift Quasar PG1634+706

PG1634+706 is a redshift ($z = 1.334$) radio quiet and optically bright quasar (Steidel and Sargent, 1991). The source was observed with *Chandra* ACIS-S detector (Weisskopf *et al.*, 2002) as a calibration target six times on March 23 and 24, 2000. Each observation lasted between 4.4 and 11 ksec. We used CIAO software to process the archival data and extracted the spectra assuming circular source regions of 1.8 arcsec radius. We used CALDB 2.24 calibration data. Table 3.4 lists each observation with its exposure time and the total counts in its spectrum.

The exact energy of the emission depends on the ionization state of the iron. The fluorescent Fe-K-alpha emission line has been observed in the quasar rest frame of near 6.4 keV, which corresponds to 2.74 keV in the observed frame of PG1634+706. We look for the line in the *Chandra* spectrum of PG1634+706 and test for its evidence in Section 3.5.2.

Table 3.5: Posterior Modes of the Line Locations Identified with the EM-type Algorithm. The line locations near 2.74 keV where the Fe-K-alpha emission line was identified are indicated in bold face.

Line profile model	Observed data set	Posterior mode (keV)	Domain of convergence (keV)
Delta function line	obs-id 47	**2.885**	0.5 – 6.0
	obs-id 62	**2.845**	0.5 – 6.0
	obs-id 69	1.805	0.5 – 6.0
	obs-id 70	**2.835**	0.5 – 6.0
	obs-id 71	**2.715**	0.5 – 3.7, 3.9 – 4.6, and 4.9 – 5.8
		5.605	3.8, 4.7 – 4.8, and 5.9 –6.0
	obs-id 1269	**2.905**	0.5 – 6.0
Gaussian line	obs-id 47	**2.765**	0.5 – 6.0
	obs-id 62	**2.595**	0.5 – 6.0
	obs-id 69	2.195	0.5 – 6.0
	obs-id 70	**2.705**	0.5 – 6.0
	obs-id 71	**2.605**	0.5 – 6.0
	obs-id 1269	**2.575**	0.5 – 6.0

3.5.2 Fitting a Spectral Model

We begin by searching for the posterior modes of the delta function and Gaussian emission lines for each of the six data sets. (For both lines, we use the Rotation(9) EM-type algorithm with 56 starting values equally spaced between 0.5 keV and 6.0 keV.) The results of the 56 runs for each data set appear in Table 3.5. Taking into account the quasar redshift a priori, as discussed in Section 3.5.1, we expect the most probable line to be located near 2.74 keV in the observed frame. The posterior modes listed in bold face in Table 3.5 are near 2.74 keV.

We use Monte Carlo methods to fully study the marginal posterior distributions of the delta function and Gaussian line locations. (In particular, we run several chains of PMG I and PMG IV, to examine convergence using multiple chains.) The solid lines in Figures 3.7 and 3.8 represent the resulting posterior distribution of the line location for each observation of PG1634+706; due to the fine binning of the *Chan-*

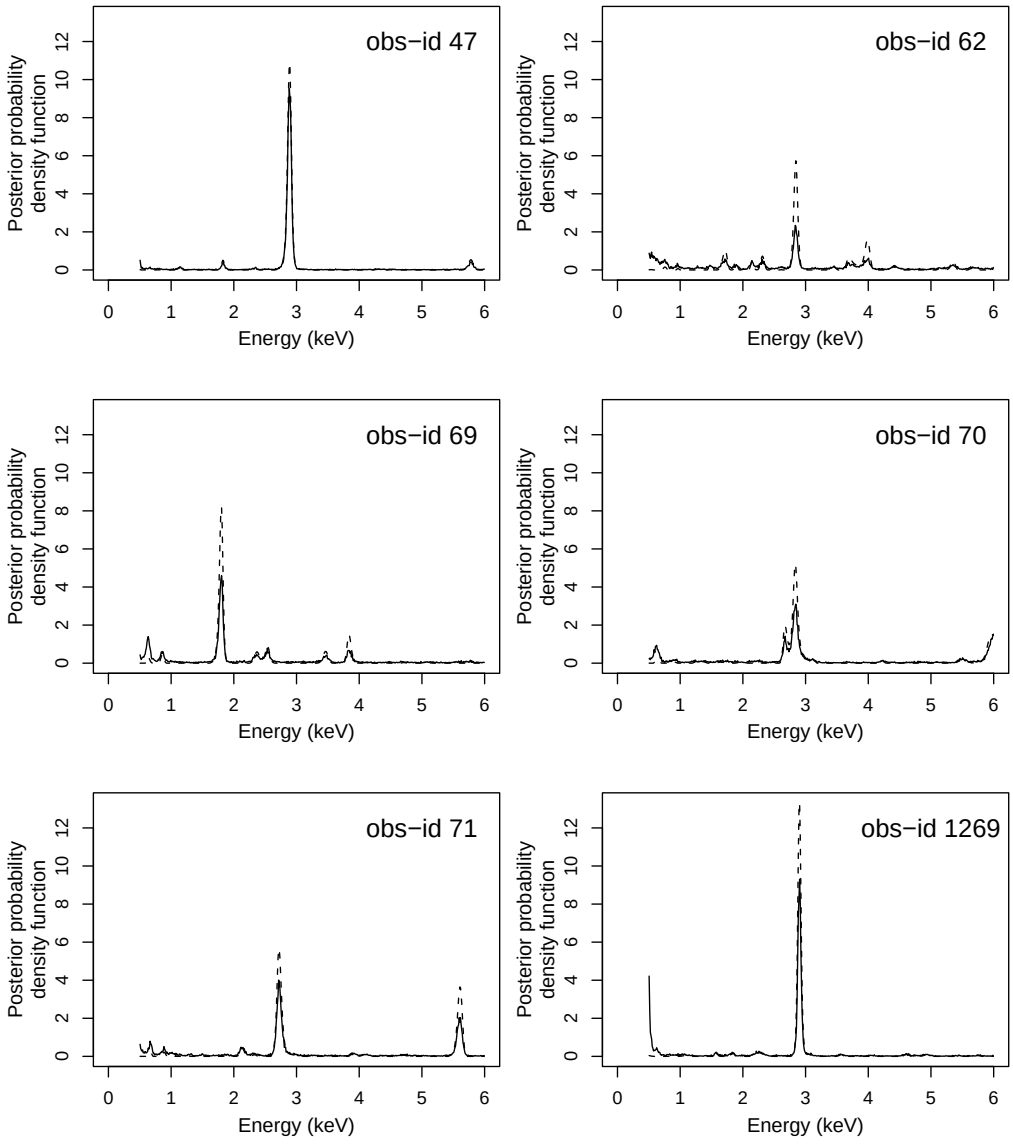

Figure 3.7: Posterior Distributions of the Delta Function Line Location μ for Different Observations of PG1634+706. The solid lines represent the marginal posterior distribution of the delta function line location, and the dashed lines represent the profile posterior distribution that is maximized over the nuisance parameters. For each data set, the marginal and profile posterior distributions agree as to the likely location of the emission line.

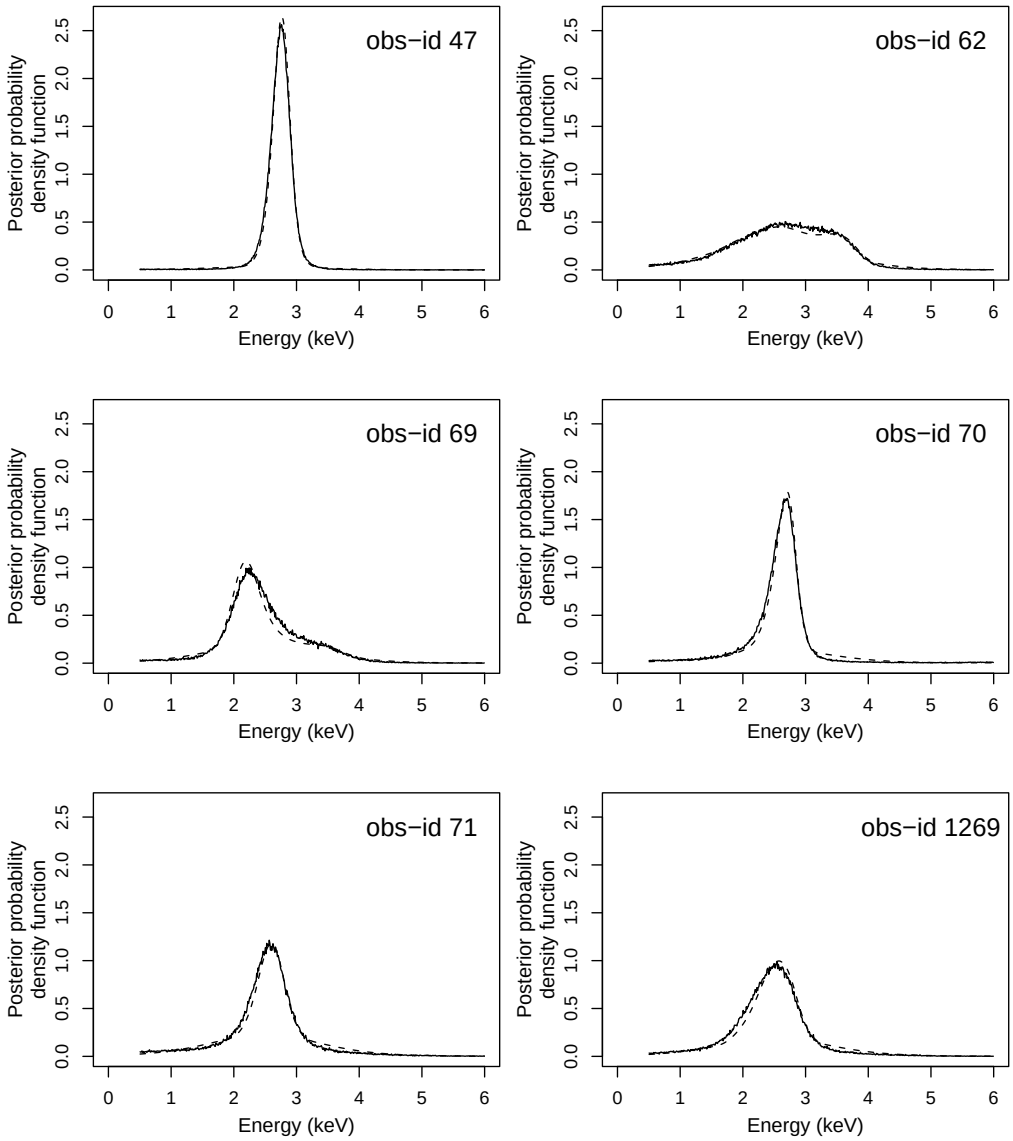

Figure 3.8: Posterior Distributions of the Gaussian Line Location μ for Different Observations of PG1634+706. The solid lines represent the marginal posterior distribution of the Gaussian line location, and the dashed lines represent the profile posterior distribution that is maximized over the nuisance parameters. For each data set, the marginal and profile posterior distributions agree as to the likely location of the emission line.

Table 3.6: 95% HPD Regions for the Delta Function Line Location Obtained with PMG I. The posterior modes of the line location near 2.74 keV where the Fe-K-alpha emission line was identified are indicated in bold face.

Observed data set	Posterior mode (keV)	95% HPD region	Posterior probability	Odds ratio
obs-id 47	**2.885**	(2.67, 3.07)	83.74%	·
obs-id 62	0.545	(0.50, 1.99)	35.21%	1.74
	2.305	(2.06, 2.43)	9.28%	0.33
	2.845	(2.45, 2.99)	23.78%	·
	3.985	(3.36, 4.10)	15.65%	0.59
obs-id 69	0.625	(0.50, 1.21)	23.64%	2.01
	1.805	(1.57, 1.97)	39.51%	4.25
	2.535	(2.20, 2.82)	13.33%	·
	3.835	(3.73, 3.98)	7.10%	0.50
obs-id 70	0.605	(0.50, 0.99)	13.14%	0.15
	2.845	(2.51, 3.19)	50.82%	·
	5.995	(5.73, 6.00)	14.36%	0.16
obs-id 71	0.665	(0.50, 1.19)	15.41%	0.25
	2.135	(2.02, 2.47)	7.23%	0.11
	2.715	(2.50, 3.07)	41.78%	·
	5.595	(5.36, 5.72)	21.99%	0.39
obs-id 1269	0.505	(0.50, 1.21)	18.25%	0.12
	2.905	(2.75, 3.12)	65.11%	·

dra observations, a marginal posterior distribution is presented by connecting the midpoint of each histogram bar. The corresponding profile posterior distribution is represented by dashed lines in each panel of Figures 3.7 and 3.8. Although the marginal and profile posterior distributions differ in treating nuisance parameters, the figures illustrate that both representations capture similar peaks.

The distribution shown in Figure 3.7 confirms that the delta function line location has a highly multimodal posterior distribution. When the posterior distribution is multimodal, basic summary statistics such as the mean, median, standard deviation, and percentiles can be difficult to interpret. In this case, the posterior distribution is better summarized by a list of the several posterior modes and the

Table 3.7: Summary Statistics for Selected Model Parameters Obtained with PMG I.

Observed data set	Parameter	Posterior mean	Posterior std. dev.	2.5%	Posterior median	97.5%
obs-id 47	λ	34.717	17.668	3.346	34.033	68.927
	α^C	3.4e-4	2.7e-5	3.0e-4	3.4e-4	4.0e-4
	β^C	1.754	0.093	1.577	1.752	1.942
	ω	0.332	0.166	0.032	0.323	0.671
obs-id 62	λ	19.764	23.372	0.610	13.590	92.996
	α^C	3.6e-4	2.9e-5	3.0e-4	3.6e-4	4.2e-4
	β^C	1.769	0.097	1.579	1.769	1.960
	ω	0.200	0.239	0.006	0.136	0.967
obs-id 69	λ	27.354	29.737	1.018	20.809	122.08
	α^C	3.5e-4	3.0e-5	2.9e-4	3.4e-4	4.1e-4
	β^C	1.741	0.098	1.552	1.741	1.935
	ω	0.129	0.130	0.005	0.097	0.554
obs-id 70	λ	24.627	22.578	1.242	20.342	88.06
	α^C	3.2e-4	2.6e-5	2.7e-4	3.2e-4	3.7e-4
	β^C	1.686	0.097	1.498	1.686	1.879
	ω	0.259	0.225	0.013	0.215	0.846
obs-id 71	λ	23.303	19.139	0.993	20.187	65.481
	α^C	4.0e-4	3.4e-5	3.3e-4	3.9e-4	4.7e-4
	β^C	1.826	0.103	1.629	1.825	2.029
	ω	0.244	0.204	0.010	0.213	0.670
obs-id 1269	λ	87.373	214.336	2.358	32.569	782.539
	α^C	2.7e-4	1.9e-5	2.4e-4	2.7e-4	3.1e-4
	β^C	1.985	0.084	1.821	1.985	2.150
	ω	0.920	2.130	0.016	0.279	7.736

corresponding HPD region. The posterior modes and the 95% HPD region of the line location are computed for each observation and are presented in Table 3.6. In the case of a multimodal distribution, each HPD region is typically composed of a number of intervals; only the intervals that have posterior probabilities greater than 5% are presented in Table 3.6. For example, the four intervals of obs-id 62 presented in Table 3.6 have a combined posterior probability of 83.92% and the other twenty one intervals not shown in the table have a posterior probability of about 11.08%, for a total of 95%. The posterior modes for the line location that are near 2.74 keV are indicated in bold face in Table 3.6. The posterior modes of the line location obtained by the Monte Carlo algorithm (i.e., PMG I) are somewhat different from those obtained by the mode finder (i.e., the Rotation(9) EM-type algorithm) due to the Monte Carlo errors of draws. The last column of Table 3.6 provides odds ratios, which compare each of the possible line locations with the line location nearest 2.74 keV,

$$\textbf{odds ratio} \;=\; \frac{p^L/(1-p^L)}{p^\star/(1-p^\star)}, \tag{3.19}$$

where p^L is the posterior probability of a particular line location and $p^\star$ is the posterior probability of the line location nearest 2.74 keV. Thus, a smaller odds ratio indicates that the particular line location is less probable than the location nearest 2.74 keV.

Table 3.7 presents basic summary statistics of the total expected counts due to the line and the power law continuum parameters for each observation of PG1634+706 when a delta function is used to model an emission line; the units of the power law normalization, α^C, are photons/(cm^2 $\cdot$ sec $\cdot$ keV), the photon index, β^C, is unitless, and the units of the expected counts due to the emission line, λ, are counts. Table 3.7 also presents the equivalent width, ω, which is a measure of line strength and is defined as

$$\omega \;=\; \frac{\lambda}{f(\theta^C, E_{j^\star}) \cdot \textsc{Area}_{j^\star} \cdot \textsc{Time}} \tag{3.20}$$

where λ is the expected line counts, $f(\theta^C, E_j)$ is the continuum intensity at the energy E_j, AREA_j is the effective area at the energy bin j in cm^2 units, TIME is the exposure time in sec units, and $j^\star$ is the index of the energy bin that contains the fitted line location; thus, ω is in keV units.

Because the posterior distributions for these parameters tend to be unimodal and symmetric, the basic summary statistics such as the mean, standard deviation, median, and various percentiles (e.g., 2.5%, 50%, and 97.5%) are meaningful. However, the posterior distribution of λ is sometimes highly right-skewed. This can be seen in Table 3.7 with obs-id 1269 where the posterior mean of λ is only 0.41 standard deviations away from zero and is much larger than the posterior median. The median and percentiles of the posterior distribution better represent such a skewed distribution than do the mean and standard deviation, which are more affected by extreme values.

Unlike the posterior distribution of the location of a delta function emission line, the posterior distribution of the Gaussian emission line location tends to be unimodal; see Figure 3.8. In this case, basic summary statistics become more relevant. Table 3.8 shows such summary statistics of the parameters (location, width, and total expected counts due to the line) of the Gaussian line profile and the power law continuum parameters for each observation of PG1634+706. In particular, the Gaussian line location and the power law continuum parameters exhibit unimodal, symmetric posterior distributions, which is confirmed by the agreement between the mean and the median. The Gaussian line width and the total expected line counts, however, tend to have right-skewed posterior distributions partly because these parameters are necessarily non-negative. In this case, we again prefer using the median and other percentiles to summarize the posterior distribution.

The six observations of PG1634+706 were independently observed with *Chandra*. Thus, under the flat prior distribution on μ, the posterior distribution of the line

Table 3.8: Summary Statistics for Selected Model Parameters Obtained with PMG IV.

Observed data set	Parameter	Posterior mean	Posterior std. dev.	2.5%	Posterior median	97.5%
obs-id 47	μ	2.730	0.251	2.275	2.745	3.115
	ν^2	0.198	0.111	0.078	0.168	0.504
	λ	90.956	38.237	20.541	88.925	171.868
	α^C	3.7e-4	2.0e-5	3.4e-4	3.7e-4	4.3e-4
	β^C	1.911	0.087	1.759	1.905	2.097
	ω	0.525	0.247	0.082	0.505	1.049
obs-id 62	μ	2.665	0.807	0.935	2.695	4.035
	ν^2	0.279	0.162	0.096	0.230	0.740
	λ	37.137	29.51	1.429	31.063	112.121
	α^C	3.8e-4	3.0e-5	3.4e-4	3.8e-4	4.5e-4
	β^C	1.874	0.100	1.708	1.865	2.096
	ω	0.340	0.269	0.011	0.287	0.988
obs-id 69	μ	2.475	0.639	1.205	2.385	3.885
	ν^2	0.278	0.167	0.090	0.230	0.757
	λ	56.079	40.394	1.985	49.302	150.499
	α^C	3.7e-4	3.0e-5	3.3e-4	3.7e-4	4.3e-4
	β^C	1.861	0.095	1.701	1.852	2.071
	ω	0.510	0.418	0.011	0.425	1.534
obs-id 70	μ	2.577	0.498	1.365	2.625	3.375
	ν^2	0.213	0.126	0.078	0.176	0.563
	λ	61.647	34.868	3.986	59.811	136.489
	α^C	3.5e-4	2.0e-5	3.2e-4	3.5e-4	4.0e-4
	β^C	1.844	0.083	1.700	1.837	2.023
	ω	0.565	0.353	0.030	0.527	1.346
obs-id 71	μ	2.513	0.628	0.945	2.545	3.865
	ν^2	0.245	0.146	0.090	0.203	0.672
	λ	46.621	31.241	2.601	42.089	117.568
	α^C	4.1e-4	4.0e-5	3.5e-4	4.1e-4	4.9e-4
	β^C	1.909	0.108	1.708	1.907	2.131
	ω	0.482	0.321	0.037	0.432	1.230
obs-id 1269	μ	2.436	0.585	1.055	2.465	3.675
	ν^2	0.253	0.141	0.090	0.221	0.640
	λ	62.337	43.516	3.248	54.766	162.727
	α^C	2.8e-4	2.0e-5	2.4e-4	2.8e-4	3.2e-4
	β^C	2.036	0.088	1.872	2.034	2.216
	ω	0.354	0.268	0.017	0.301	0.993

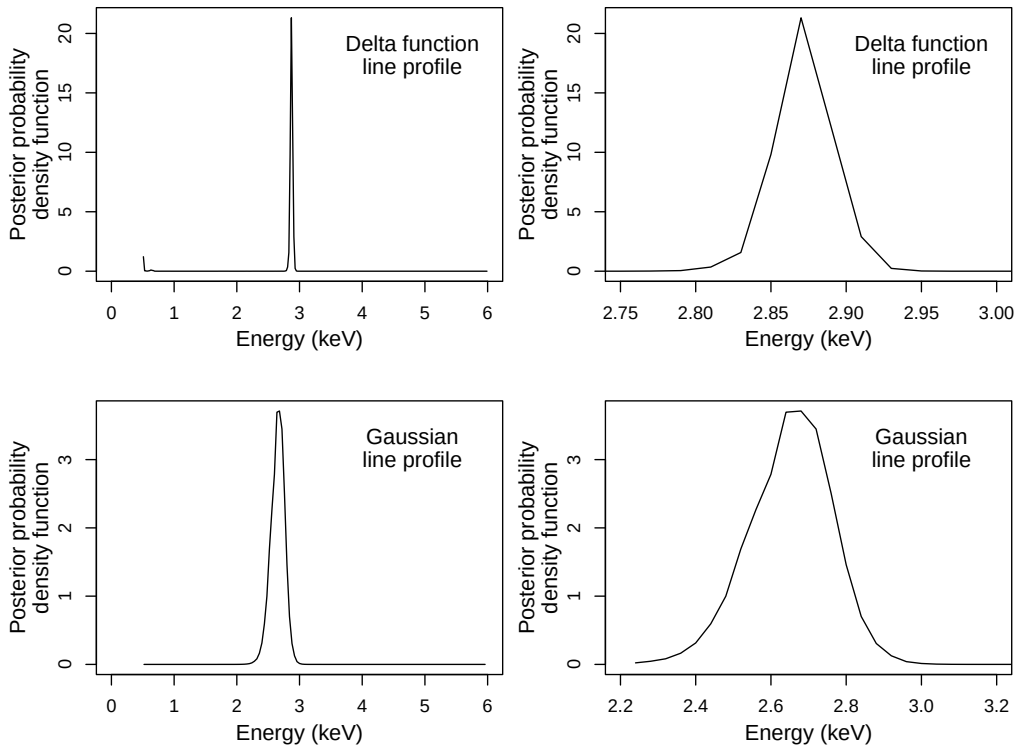

Figure 3.9: Posterior Distributions of the Line Location Given All of the Observations of PG1634+706. The panels in the left column are plotted over the entire energy range and those in the right column over the focused range near 2.74 keV.

location given all six observations is given by

$$\begin{aligned} p(\mu|y) &\propto \int \cdots \int \prod_{i=1}^{6} L(\mu, \psi_i|y_i) d\psi_1 \cdots d\psi_6 \\ &= \prod_{i=1}^{6} p(\mu|y_i), \end{aligned} \tag{3.21}$$

where $y = \{y_i, i = 1, \ldots, 6\}$ denotes the six observations, μ denotes the line location parameter, $\psi = \{\psi_i, i = 1, \ldots, 6\}$ denotes the set of model parameters other than μ for each of the six observations, and $L(\mu, \psi_i|y)$ represents a likelihood function of (μ, ψ_i) given y. (Here we allow ψ_i to vary among the six observations; i.e., we do not exclude the possibility that the six observations have somewhat different power law normalizations and photon indexes.) The values of the posterior distribution given one of the individual data set is sometimes indistinguishable from

Table 3.9: Summary Statistics for the Line Locations Given All Six Observations of PG1634+706. The posterior point estimates of the line location near 2.74 keV where the Fe-K-alpha emission line was identified are indicated in bold face.

Line profile model	Posterior mean	Posterior std. dev.	Posterior mode	95% HPD region	Posterior probability	Odds ratio
Delta function line	·	·	0.505	(0.50, 0.51)	2.2%	1.2e-3
	·	·	**2.865**	(2.83, 2.92)	94.8%	·
Gaussian line	**2.650**	0.111	·	·	·	·

zero because of numerical inaccuracies. Thus we add 1/20000 to the posterior probability of each energy bin and renormalize each of the posterior distributions. This allows the product given in (3.21) to be computed for each energy bin and is somewhat conservative as it increases the posterior uncertainty corresponding to each of the individual data sets. Figure 3.9 presents the posterior distributions of the delta function and Gaussian line locations given all six observations computed in this way; the panels of the left column examine the whole range of the line location while the panels of the right column focus on the range near 2.74 keV. Table 3.9 presents the posterior mode and HPD region for the delta function line location and the posterior mean and standard deviation for the Gaussian line location, based on the posterior distributions shown in Figure 3.9; the point estimates near 2.74 keV are reported in bold face. The odds ratios in Table 3.9 give us strong evidence of the delta function line location near 2.74 keV. The posterior mean and standard deviation of the line location under a delta function line profile are affected by the local mode near 0.5 keV, so that the posterior mode and HPD region can better represent the posterior distribution than the posterior mean and standard deviation. The posterior distribution of the Gaussian line location shown in Figure 3.9 closely follows a bell-shaped curve. Thus, we use the posterior mean and standard deviation as summary statistics. The most probable line is located

at $2.865^{+0.055}_{-0.035}$ keV for the delta function line profile; the posterior mean of the line location is 2.650 ± 0.222 keV for the Gaussian line profile. (Error bars correspond to 95% intervals.)

3.5.3 Model Checking and Evidence for the Emission Line

To check the self-consistency of the spectral model and evaluate the evidence for an emission line in the spectrum, van Dyk and Kang (2004) suggested the residual plots and posterior predictive methods for the spectral model used in this section; however, they use a Gaussian line with fixed width in the spectral model. Here the methods are applied to the spectral model with a delta function line.

With the *Chandra* observations of PG1634+706, we consider the same spectral model discussed in Section 3.1.3 except that we compare three models for the emission line:

MODEL 0 : There is no emission line in the spectrum.

MODEL 1 : There is a delta function emission line with fixed location but unknown intensity in the spectrum.

MODEL 2 : There is a delta function emission line with unknown location and intensity in the spectrum.

Taking the prior information for the Fe-K-alpha emission line location (near 2.74 keV) into consideration, MODEL 1 fixes the delta function line location near 2.74 keV. On the other hand, MODEL 2 uses no prior information for the line location. We begin with graphical model diagnostics to investigate whether the fitted models are consistent with the observed data. With the *Chandra* data set, obs-id 47, Figure 3.10(a) and (b) compare the observed data with the fitted models under MODELS 0 and 1 in the first and second column, respectively. In MODEL 1, we fix

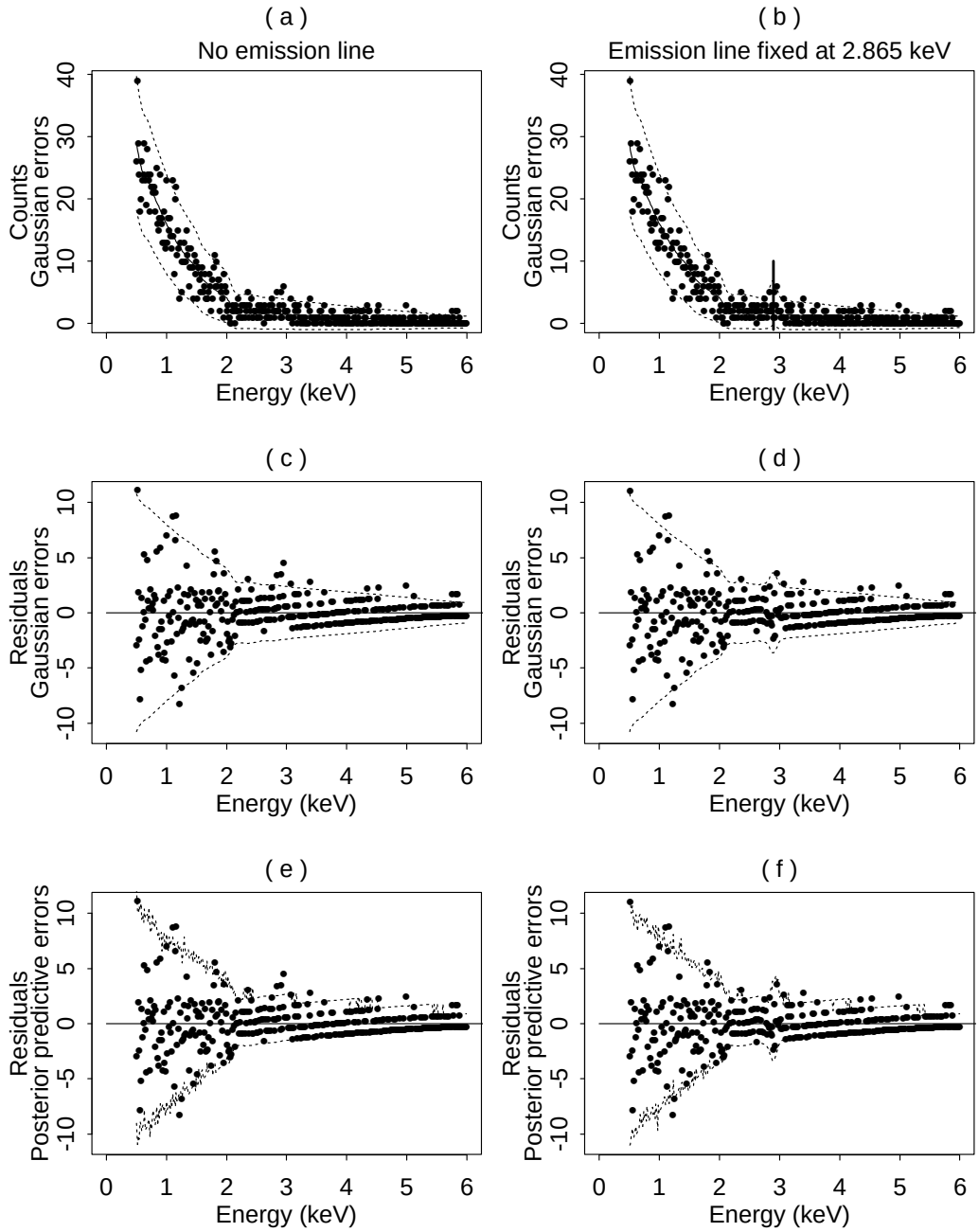

Figure 3.10: Model Diagnostic Plots with Obs-id 47. Panels (a) and (b) show the data with predictive errors based on a Gaussian approximation; panels (c) and (d) show the residuals with errors based on a Gaussian approximation; and panels (e) and (f) show the residuals with errors based on the posterior predictive distribution. The two columns of the figure correspond to MODELS 0 and 1 respectively. The excess counts near 2.865 keV are apparent for the panels (a) and (b), thereby indicating evidence for the inclusion of the emission line in the model; the location of the emission line is represented by a vertical line in the panel (b).

the line at 2.865 keV which is fitted near 2.74 keV for the spectral model with a delta function emission line. The expected counts per channel under the model, $\Xi_l(\hat{\theta})$, given in (3.2) are represented by a solid line and the predictive errors by a dotted line; $\hat{\theta}$ is the posterior mode of θ. The errors are computed as $\pm 2\sqrt{\Xi_l(\hat{\theta})}$, which is two standard deviations under the sampling model conditioning on $\hat{\theta}$. This corresponds to a 95% error bar under the Gaussian approximation. These errors do not account for the posterior uncertainty of θ. Figure 3.10(c) and (d) are mean subtracted versions of the panels (a) and (b), i.e., these panels are residual plots. To better account for the Poisson nature of the data and the posterior variability in θ, we can compute residual errors based on the posterior predictive distribution. These plots appear in Figure 3.10(e) and (f); the jagged nature of the posterior predictive residual errors is due to our Monte Carlo evaluation of this distribution. The advantage of the posterior predictive errors is evident for the low counts in the high energy tail of the spectra as shown in the residual plots of Figure 3.10. Comparing the two columns in Figure 3.10 near 2.865 keV also provides evidence for the inclusion of the emission line.

We now use ppp-values to compare the three models and quantify the evidence in the data for the delta function emission line as discussed in Section 3.4; see Protassov *et al.* (2002) for details. In the posterior predictive check, MODEL 1 fixes the line location at 2.74 keV using the prior information as to the Fe-K-alpha emission line. In order to combine the evidence for the line from all six observations with different exposure area and exposure time, we base our comparisons on the test statistic that is the sum of the loglikelihood ratio statistics for comparing MODEL m and MODEL 0, i.e.,

$$T_m(y_{\rm rep}^{(\ell)}) = \sum_{i=1}^{6} \log \left\{ \frac{\sup_{\theta\in\Theta_m} L\big(\theta|y_{{\rm rep}\,i}^{(\ell)}\big)}{\sup_{\theta\in\Theta_0} L\big(\theta|y_{{\rm rep}\,i}^{(\ell)}\big)} \right\}, \;\; m = 1,\, 2, \text{ and } \ell = 1, \ldots, 1000, \quad (3.22)$$

where Θ_0, Θ_1, and Θ_2 represent the parameter spaces under MODELS 0, 1, and 2, respectively, and $y_{\rm rep}^{(\ell)} = \big\{y_{{\rm rep}\,i}^{(\ell)}, i = 1, \ldots, 6\big\}$ denotes the collection of six data sets simulated under MODEL 0. Specifically, we generate 1000 samples from the poste-

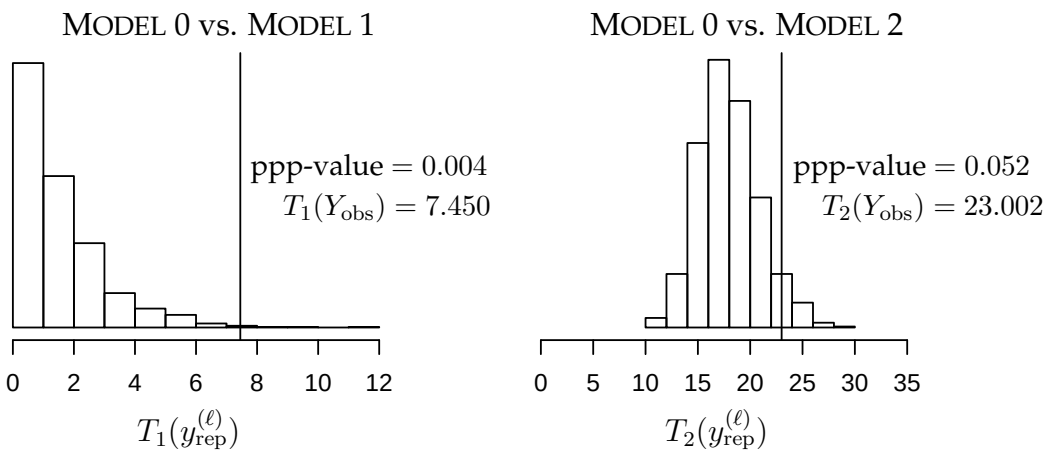

Figure 3.11: Posterior Predictive Checks Given All Six Observations of PG1634+706. In each of the two histograms, the observed test statistic (the vertical line) is compared with the test statistics from 1000 posterior predictive simulated data sets. The ppp-value is the proportion of the test statistics computed using the data simulated under MODEL 0 that are as extreme as or more extreme than the observed test statistic. Small ppp-values indicate stronger evidence of the emission line.

rior predictive distribution of $T_m(y_{\text{rep}}^{(\ell)})$ for $m = 1, 2$ under MODEL 0. Histograms of $T_1(y_{\text{rep}}^{(\ell)})$ and $T_2(y_{\text{rep}}^{(\ell)})$ appear in Figure 3.11; comparing these distributions with the observed values of the test statistics yields the ppp-values shown in Figure 3.11. Because there is strong evidence for the presence of the delta function spectral line given all six observations, MODELS 1 and 2 are preferable to MODEL 0. The comparison between MODELS 0 and 1 shows stronger evidence for the line location because we are using extra information about the plausible line location a priori.

3.6 Concluding Remarks

Identifying emission lines is an important problem to understand the physical environment and structures of astronomical sources. This chapter presents fully model-based Bayesian methods to detect the emission lines in X-ray spectra via

highly structured multilevel spectral models with the delta function and Gaussian lines. The models are fitted with efficient EM-type algorithms and Gibbs samplers designed to improve the convergence characteristics of their standard counterparts. The usefulness and comparison of the spectral models with two different line profiles are demonstrated in our simulation study. In particular, the advantage of the spectral model with the delta function line is that it can precisely identify narrow and weak lines and provide strong evidence for the inclusion of such lines in the spectrum, as compared to the spectral model with the Gaussian line. Thus, our simulation study illustrates that a delta function line can serve as a good starting line profile to identify the line location even in the case where the physical constraint of a spectral line is opposed to the delta function line.

Our model-based Bayesian methods are applied to the six different *Chandra* observations of PG1634+706 to identify a narrow emission line in the X-ray spectrum. The most probable line with a 95% error bar is identified at $2.865^{+0.055}_{-0.035}$ keV and 2.650 ± 0.222 keV in the observed frame with the delta function and Gaussian emission lines, respectively. These observed lines are redshifted into $6.69^{+0.128}_{-0.082}$ keV and 6.19 ± 0.518 keV in the quasar rest frame, respectively, and seem to indicate two opposite states of the ionization of the iron in the emission plasma.

In this chapter, we identify a single emission line and evaluate evidence of the line in the spectrum. If interested in more than one line (e.g., two lines), we can optimize or simulate another line location after putting the first line in its most probable location or simultaneously search for several line locations; BLoCXS (Bayesian fitting of Low Count X-ray Spectra) that is free statistical software and will soon be available on the CIAO contributed software page has both features for fitting multiple lines. For example, in one of the *Chandra* observations, the line location near 0.5 keV is also identified, as shown in Table 3.6. However, the energy 0.5 keV in the observed frame of PG1634+706 is transferred into 1.17 keV in the quasar rest frame where no line can be detected in quasars. Thus, the observed line near

0.5 keV is presumably due to the instrumental effects. In this case, we can run the algorithms to identify the second line after fixing the first line at near 0.5 keV or to search for both lines with spatial restrictions for the first line near 0.5 keV.

Chapter 4

Joint Imputation Models for Non-Nested Data

4.1 Introduction

A fixed geographical region is often divided into several different levels of political partitions. In the Unites States, for example, the country is sequentially divided into states, into counties, and partially into cities. In this case, one of the partitions contains or is completely nested within the others. In this chapter, however, we are particularly interested in data observed in Germany which also has different levels of geographical partitions that may not be aligned.

Unemployment data are measured on different non-nested geographic partitions of Germany, i.e., states, counties, agencies, and communities. The set of communities is the highest resolution partition, and the sets of counties and agencies are lower resolution partitions each of which consists of several communities. The two lower resolution partitions are not trivially defined, so that they are not generally nested one within the other.

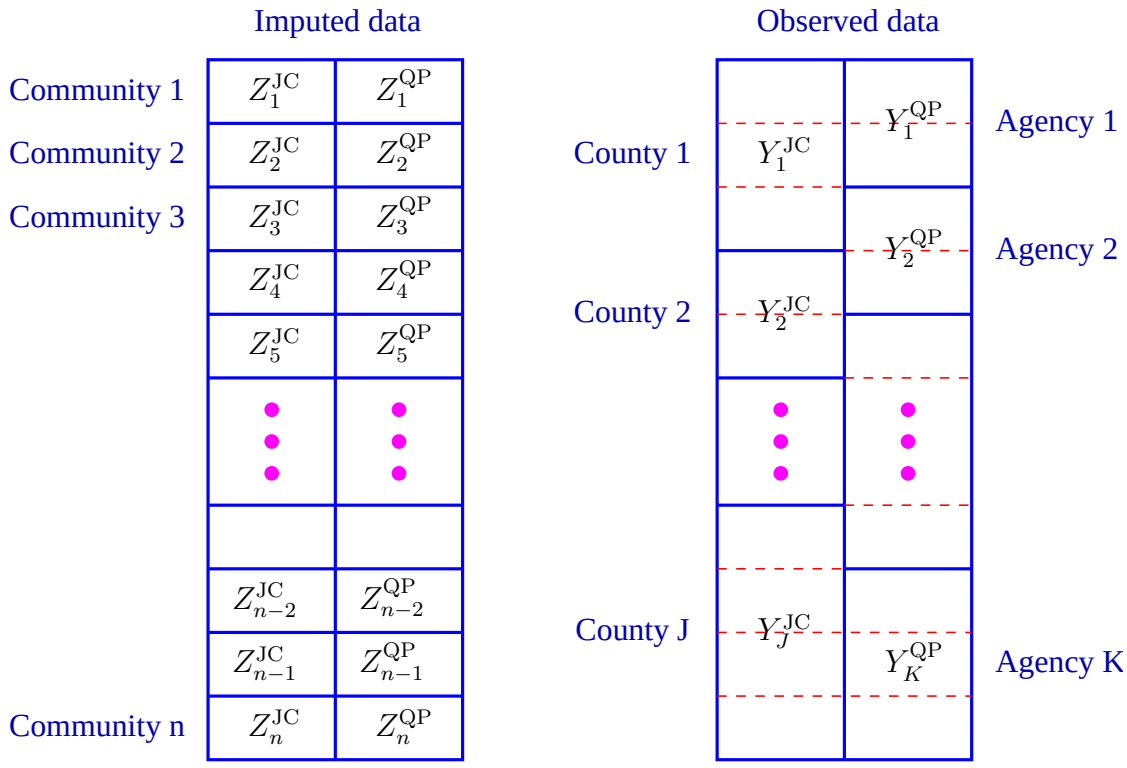

Figure 4.1: Data Structure We Would Like to Impute and We Observe.

Our initial data are composed of the bivariate response variable, *Job Creation* (JC) and *Qualification Program* (QP), and a number of covariates; Z_i^{JC} and Z_i^{QP} each denotes the number of German people who participate in the job creation and qualification programs in community i, respectively. Each component of the bivariate response variable is observed at either the county or agency level. There is also a set of covariates X that are fully observed at the community level. Because we assume "additivity" of the data, 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 at all of the levels of the partition. A difficulty arises when one component of the bivariate response variable is observed on a certain partition (e.g., the counties), and the other component is observed on a different partition (e.g., the agencies), which neither contains nor is nested within the first partition. Figure 4.1 illustrates the plausible misalignment of the observed

data, where $\{Y_j^{\mathrm{JC}}, j = 1, 2, \ldots, J\}$ denote the *Job Creation* data observed in J counties and $\{Y_k^{\mathrm{QP}}, k = 1, 2, \ldots, K\}$ denote the *Qualification Program* data observed in K agencies. With this misalignment, we aim to devise models to jointly impute the bivariate response variable at the highest resolution partition (i.e., the communities), shown in the left panel of Figure 4.1. Once we impute the bivariate response variable, the county-level or agency-level data that are missing can be recovered by aggregating the imputed community-level variable to the corresponding partition under additivity.

The remainder of this chapter is organized into five sections. In Section 4.2, we present three joint imputation models for data measured on misaligned partitions. Section 4.3 describes efficient computational algorithms used to fit the joint imputation models. A simulation study is conducted to examine the utility and limitation of the modeling and computational strategies in Section 4.4. In Section 4.5, we apply our strategies to the real German unemployment data measured on non-nested partitions. Concluding remarks appear in Section 4.6.

4.2 Joint Imputation Models for Non-Nested Data

4.2.1 Bivariate Gaussian Model

Given the partial sums of a bivariate response variable observed at the different non-nested levels of partition, we aim to create joint imputations of the bivariate response variable at the highest resolution partition, properly accounting for its correlation structure. Modeling the correlation structure would be simplified if we had the community-level data. Thus, a bivariate response variable at the highest level of resolution can be modeled with a bivariate Gaussian distribution, so that

our target model is given by

$$\begin{pmatrix} W_i^{\mathrm{JC}} \\ W_i^{\mathrm{QP}} \end{pmatrix} \overset{\text{ind}}{\sim} \mathrm{N}_2\left(\begin{pmatrix} X_i^\top \beta_1 \\ X_i^\top \beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right) \text{ for } i = 1, 2, \ldots, n, \tag{4.1}$$

where W_i^{JC} and W_i^{QP} denote the bivariate response variable of JC and QP in community i, respectively, X_i is a $p \times 1$ vector of known covariates in community i, β_1 and β_2 are $p \times 1$ vectors of coefficients for JC and QP, respectively, σ_1^2 and σ_2^2 are residual variances for W^{JC} and W^{QP}, respectively, σ_{12} represents a covariance between W^{JC} and W^{QP}, and n is the number of communities in a particular state or in all of the German states. We consider the Gaussian model despite the fact that our data (JC and QP) are non-negative integers because the model allows us to easily account for the correlation structure with bivariate (or multivariate) responses. The bivariate Gaussian model in (4.1) also serves as a guideline to fit the other more appropriate target models that we consider.

As illustrated in Figure 4.1, 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 $\cup_{j=1}^{J} \mathcal{J}_j = \{1, 2, \ldots, n\}$ and $\mathcal{J}_j \cap \mathcal{J}_k = \emptyset$ for $j \neq k$. Likewise, we define by $\mathcal{K}_k$ the set of indices of the communities that are nested within agency k, for $k = 1, 2, \ldots, K$. Because of the additivity of the variables, the response variable JC in county j, Y_j^{JC}, consists of the sum of the n_j values of W_i^{JC} in county j, i.e., $Y_j^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} W_i^{\mathrm{JC}}$ for $j = 1, 2, \ldots, J$. Under the target model in (4.1), the marginal distribution of Y_j^{JC} is thus given by

$$Y_j^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} W_i^{\mathrm{JC}} \overset{\text{ind}}{\sim} \mathrm{N}\left(\Big(\sum_{i \in \mathcal{J}_j} X_i \Big)^\top \beta_1, \sigma_1^2 n_j \right) \text{ for } j = 1, 2, \ldots, J, \tag{4.2}$$

where $n_j = \sum_{i \in \mathcal{J}_j} 1$. Similarly, the response variable QP in agency k is the sum of the community-level response variable QP, i.e., $Y_k^{\mathrm{QP}} = \sum_{i \in \mathcal{K}_k} W_i^{\mathrm{QP}}$ for $k = 1, 2, \ldots, K$. Thus, the target model in (4.1) also implies

$$Y_k^{\mathrm{QP}} = \sum_{i \in \mathcal{K}_k} W_i^{\mathrm{QP}} \overset{\text{ind}}{\sim} \mathrm{N}\left(\Big(\sum_{i \in \mathcal{K}_k} X_i \Big)^\top \beta_2, \sigma_2^2 m_k \right) \text{ for } k = 1, 2, \ldots, K, \tag{4.3}$$

where $m_k = \sum_{i \in \mathcal{K}_k} 1$. Fitting of the two marginal distributions of the target model is straightforward, given the observed data. Fitting the correlation structure, however, is more challenging.

4.2.2 Bivariate Lognormal Model

To address the non-negativity of the real data while capitalizing on the flexibility of the Gaussian distribution for modeling the correlation structure, we consider the bivariate lognormal model given by

$$\begin{pmatrix} W_i^{\mathrm{JC}} \\ W_i^{\mathrm{QP}} \end{pmatrix} \equiv \begin{pmatrix} \log Z_i^{\mathrm{JC}} \\ \log Z_i^{\mathrm{QP}} \end{pmatrix} \overset{\text{ind}}{\sim} \mathrm{N}_2\left(\begin{pmatrix} (\log X_i)^\top \gamma_1 \\ (\log X_i)^\top \gamma_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \right) \text{ for } i = 1, \ldots, n, \quad (4.4)$$

where W_i^{JC} and W_i^{QP} are defined as log-transformed Z_i^{JC} and Z_i^{QP}, respectively, $\log X_i$ is a $p \times 1$ vector of known covariates in community i, γ_1 and γ_2 are $p \times 1$ vectors of coefficients for JC and QP, respectively, τ_1^2 and τ_2^2 are residual variances for W^{JC} and W^{QP}, respectively, and τ_{12} represents a covariance between W^{JC} and W^{QP}. Under the target model in (4.4), the observed data that are the partial sums of the community-level data do not follow a standard probability distribution. That is, we observe $Y_j^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} Z_i^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} e^{W_i^{\mathrm{JC}}}$ rather than $Y_j^\star \equiv \sum_{i \in \mathcal{J}_j} \log Z_i^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} W_i^{\mathrm{JC}}$, and $Y_j^\star$ cannot be computed from the observed Y_j^{JC} because $\log Y_j^{\mathrm{JC}} \neq Y_j^\star$ for $n_j > 1$. Thus, although the lognormal model appears to better account for the underlying nature of the real data, it is computationally much harder to fit than the bivariate Gaussian model in (4.1).

4.2.3 Poisson Regression Model

As an alternative model, we may model the number of people in the job creation or qualification program in community i as an inhomogeneous Poisson process.

We consider the target models given by

$$Z_i^{\mathrm{JC}} \sim \mathrm{Poisson}\big(\lambda_i^{\mathrm{JC}}\big), \text{ where } \log \lambda_i^{\mathrm{JC}} = X_i^\top \delta_1 \tag{4.5}$$

$$Z_i^{\mathrm{QP}} | Z_i^{\mathrm{JC}} \sim \mathrm{Poisson}\big(\lambda_i^{\mathrm{QP}}\big), \text{ where } \log \lambda_i^{\mathrm{QP}} = \big(X_i \; Z_i^{\mathrm{JC}}\big)^\top \delta_2 \tag{4.6}$$

or

$$Z_i^{\mathrm{JC}} | Z_i^{\mathrm{QP}} \sim \mathrm{Poisson}\big(\lambda_i^{\mathrm{JC}}\big), \text{ where } \log \lambda_i^{\mathrm{JC}} = \big(X_i \; Z_i^{\mathrm{QP}}\big)^\top \delta_1 \tag{4.7}$$

$$Z_i^{\mathrm{QP}} \sim \mathrm{Poisson}\big(\lambda_i^{\mathrm{QP}}\big), \text{ where } \log \lambda_i^{\mathrm{QP}} = X_i^\top \delta_2. \tag{4.8}$$

Given the observed data, however, fitting the first target model in (4.5) and (4.6) or the second target model in (4.7) and (4.8) is not straightforward because, for example,

$$Y_j^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} Z_i^{\mathrm{JC}} \sim \mathrm{Poisson}\bigg(\sum_{i \in \mathcal{J}_j} \lambda_i^{\mathrm{JC}} \bigg). \tag{4.9}$$

but $\sum_{i \in \mathcal{J}_j} \lambda_i^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} e^{X_i^\top \delta_1}$ does not correspond to the mean of a standard log linear model, i.e., $\exp\big((\sum_{i \in \mathcal{J}_j} X_i)^\top \delta_1 \big) \neq \sum_{i \in \mathcal{J}_j} e^{X_i^\top \delta_1}$. Thus, the observed data cannot be directly modeled as a Poisson loglinear model. We thus devise computationally intensive methods to fit the target models in Section 4.3. With the Poisson regression model in (4.5) and (4.6) or in (4.7) and (4.8), however, we lose the power of analytically modeling the correlation structure of the bivariate response variable, and the model fitting depends on the order in which we set up the conditional distributions of the joint models.

4.3 Computation

4.3.1 Overview of Computational Methods

Based on the joint imputation models described in Section 4.2, we aim to create joint imputations of a bivariate response variable from its joint posterior predictive distribution, e.g., $p(Z^{\mathrm{JC}}, Z^{\mathrm{QP}} | Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$. Because the posterior predictive distribution is implicitly integrated over the model parameters θ, the joint

imputation procedure can be accomplished by iteratively drawing the missing community-level variable from $p(Z^{\mathrm{JC}}, Z^{\mathrm{QP}}|\theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ and the parameters from $p(\theta|Z^{\mathrm{JC}}, Z^{\mathrm{QP}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$. Because the bivariate response variable is measured on misaligned partitions, however, it is not even feasible to analytically write the joint imputation procedure $p(Z^{\mathrm{JC}}, Z^{\mathrm{QP}}|\theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, which leads us to consider three methods to create joint imputations.

To circumvent the difficulty, the first method formulates the joint imputation procedure in terms of two complete conditional distributions, $p(Z^{\mathrm{JC}}|Z^{\mathrm{QP}}, \theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ and $p(Z^{\mathrm{QP}}|Z^{\mathrm{JC}}, \theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, because the conditional distribution of one component of the bivariate response variable given the other component is rather available. We note that these two conditional distributions are compatible for Gaussian models. In the case of non-Gaussian models, however, these two conditional distributions are not necessarily compatible because they are not generally constructed from a consistent model. In this case, the corresponding Markov chain may not have a (known) stationary distribution. That is, one could use the two conditional distributions for the Poisson regression model, but we emphasize that they may be incompatible because they are not derived from a common joint distribution.

In the second method, the joint imputation procedure is also embedded into the MCMC sampler, but we consider formulating the joint imputation procedure in terms of marginal and conditional distributions, e.g., $p(Z^{\mathrm{JC}}|\theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ and $p(Z^{\mathrm{QP}}|Z^{\mathrm{JC}}, \theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$. Unfortunately, however, the marginal distribution $p(Z^{\mathrm{JC}}|\theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ is not expressed in a closed form because of the misalignment of partitions. We thus suggest using an incoherent marginal distribution $p(Z^{\mathrm{JC}}|\theta, Y^{\mathrm{JC}})$ as if it were $p(Z^{\mathrm{JC}}|\theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$. That is, we do not use a coherent model in this method, but each imputation should be better behaved because we use the marginal and conditional distributions instead of the two conditional distributions. Our simulation study in Section 4.4 shows that this approximation works reasonably well.

Lastly, after we completely impute one component of a bivariate response variable from a marginal distribution, the other component is sequentially imputed using the imputed variable as an another covariate. Because of the same difficulty of the misaligned partitions, the incoherent marginal distribution is substituted for the correct marginal distribution. Although it is also based on the incoherent imputation procedure, this strategy not only produces better behaved imputations, but reduces model complexity by focusing on one component at a time. Because the imputation procedure for each component is separated, we can easily use a sampling importance resampling (SIR) method (Rubin, 1987) that corrects the inconsistent imputation procedure toward the consistent one. In addition to SIR, we also use a MCMC sampler when imputing one component at a time. We note that the MCMC strategy requires $M + 1$ different chains to create M joint imputations: We run a single chain to create the M imputations of the first component, and then M chains for the second component using each of the M imputations as an additional covariate in each chain.

In this chapter, we employ the second imputation method for the bivariate Gaussian model to create joint imputations and devise an incompatible MCMC sampler introduced in Chapter 2. For the bivariate lognormal and Poisson regression models, the third imputation method is used after formulating a joint distribution in terms of marginal and conditional distributions. In particular, we use both the SIR and MCMC methods to create joint imputations from the bivariate lognormal model, while the Poisson regression model is fitted by the MCMC method.

4.3.2 Creating Joint Multiple Imputations

Fitting the Bivariate Gaussian Model

The bivariate Gaussian model could be easily fitted if the community-level variables were known. This leads us to consider the method of data augmentation

where we treat the community-level variables as missing data. Using the multivariate Jeffrey's flat prior distribution,

$$p(\theta) \;\propto\; \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{vmatrix}^{-3/2}, \tag{4.10}$$

the target posterior distribution of interest is given by

$$\begin{aligned} p(W^{\mathrm{JC}}, W^{\mathrm{QP}}, \theta | Y^{\mathrm{JC}}, Y^{\mathrm{QP}}) \;\propto\; & \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{vmatrix}^{-\frac{n+3}{2}} \\ \times\; \exp\Bigg(& -\frac{1}{2} \sum_{i=1}^{n} \begin{pmatrix} W_i^{\mathrm{JC}} - X_i^\top \beta_1 \\ W_i^{\mathrm{QP}} - X_i^\top \beta_2 \end{pmatrix}^\top \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} W_i^{\mathrm{JC}} - X_i^\top \beta_1 \\ W_i^{\mathrm{QP}} - X_i^\top \beta_2 \end{pmatrix} \Bigg), \end{aligned} \tag{4.11}$$

where θ denotes the model parameters, i.e., $\theta = (\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \sigma_{12})$. Using the complete conditional distributions of the target distribution in (4.11), we can easily construct the MCMC sampler that iterates among

STEP 1 : Draw β_1 from $p(\beta_1 | \beta_2, \sigma_1^2, \sigma_2^2, \sigma_{12}, W^{\mathrm{JC}}, W^{\mathrm{QP}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, (Sampler 4.3.1)

STEP 2 : Draw β_2 from $p(\beta_2 | \beta_1, \sigma_1^2, \sigma_2^2, \sigma_{21}, W^{\mathrm{JC}}, W^{\mathrm{QP}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$,

STEP 3 : Draw $(W^{\mathrm{JC}}, W^{\mathrm{QP}})$ from $p(W^{\mathrm{JC}}, W^{\mathrm{QP}} | \theta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, and

STEP 4 : Draw $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ from $p\left(\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \Big| \beta_1, \beta_2, W^{\mathrm{JC}}, W^{\mathrm{QP}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}} \right)$.

As introduced in Chapter 2, however, we can improve the convergence characteristics of the MCMC sampler by partially marginalizing over $(W^{\mathrm{JC}}, W^{\mathrm{QP}})$. Specifically, we first marginalize $(W^{\mathrm{JC}}, W^{\mathrm{QP}})$ out of STEPS 1 and 2, then trim the draws because they do not affect the transition kernel of the Markov chain constructed by the MCMC sampler. This results in the partially marginalized MCMC sampler that iterates among

STEP 1 : Draw β_1 from $p(\beta_1 | \beta_2, \sigma_1^2, \sigma_2^2, \sigma_{12}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, (Sampler 4.3.2)

STEP 2 : Draw β_2 from $p(\beta_2 | \beta_1, \sigma_1^2, \sigma_2^2, \sigma_{21}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$,

STEP 3 : Draw $(W^{\text{JC}}, W^{\text{QP}})$ from $p(W^{\text{JC}}, W^{\text{QP}}|\theta, Y^{\text{JC}}, Y^{\text{QP}})$, and

STEP 4 : Draw $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ from $p\left(\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \middle| \beta_1, \beta_2, W^{\text{JC}}, W^{\text{QP}}, Y^{\text{JC}}, Y^{\text{QP}}\right)$.

Note that Sampler 4.3.2 is constructed using a set of incompatible conditional distributions, maintaining the same transition kernel as Sampler 4.3.1; see Chapter 2 for more details.

To obtain the conditional distributions marginalized over $(W^{\text{JC}}, W^{\text{QP}})$ in STEPS 1 and 2 of Sampler 4.3.2, we consider the reparameterization of $(\sigma_1^2, \sigma_2^2, \sigma_{12})$. To do so, we introduce a $(J+K) \times (2n)$ indicator matrix $\mathcal{M}$ where the first J rows and the first n columns indicate which communities are contained in each county, and the next K rows and the second n columns indicate which communities are in each agency; the other components of $\mathcal{M}$ are set to zero. Then, observed data are given by multiplying the vector of community-level variables by the indicator matrix $\mathcal{M}$, i.e.,

$$\begin{pmatrix} Y^{\text{JC}} \\ Y^{\text{QP}} \end{pmatrix} \;=\; \mathcal{M} \cdot \begin{pmatrix} W^{\text{JC}} \\ W^{\text{QP}} \end{pmatrix} \;\equiv\; \mathcal{M} \cdot \begin{pmatrix} W_1^{\text{JC}} \\ \vdots \\ W_n^{\text{JC}} \\ W_1^{\text{QP}} \\ \vdots \\ W_n^{\text{QP}} \end{pmatrix}, \tag{4.12}$$

where Y^{JC} is a $J \times 1$ vector containing the observed partial sums of W^{JC}, Y^{QP} is a $K \times 1$ vector containing the observed partial sums of W^{QP}, and W^{JC} and W^{QP} are $n \times 1$ vectors of the community-level variables JC and QP, respectively. Using the

notation in (4.12), we can rewrite the target model in (4.1) as

$$\begin{pmatrix} Y^{\mathrm{JC}} \\ Y^{\mathrm{QP}} \end{pmatrix} \sim \mathrm{N}_{J+K}\left(\begin{pmatrix} (\sum_{i\in\mathcal{J}_1} X_i)^\top\beta_1 \\ \vdots \\ (\sum_{i\in\mathcal{J}_J} X_i)^\top\beta_1 \\ (\sum_{i\in\mathcal{K}_1} X_i)^\top\beta_2 \\ \vdots \\ (\sum_{i\in\mathcal{K}_K} X_i)^\top\beta_2 \end{pmatrix}, \mathcal{M}\cdot\begin{pmatrix} \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} \\ \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 \end{pmatrix}\cdot\mathcal{M}^\top\right)$$

$$\equiv \mathrm{N}_{J+K}\left(\begin{pmatrix} (X^{\mathrm{JC}})^\top\beta_1 \\ (X^{\mathrm{QP}})^\top\beta_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right), \tag{4.13}$$

where X^{JC} is a $p \times J$ matrix containing the partial sums of the covariates in the counties, X^{QP} is a $p \times K$ matrix containing the partial sums of the covariates in the agencies, Σ_{11} is a $J \times J$ diagonal matrix of the variances of Y^{JC}, Σ_{22} is a $K \times K$ diagonal matrix of the variances of Y^{QP}, and $\Sigma_{12} = \Sigma_{21}^\top$ is a $J \times K$ covariance matrix of Y^{JC} and Y^{QP}. Based on the prior distribution in (4.10) and the model in (4.13), the conditional distribution in STEP 1 of Sampler 4.3.2 is given by

$$\begin{aligned} &p(\beta_1|\beta_2,\sigma_1^2,\sigma_2^2,\sigma_{12},Y^{\mathrm{JC}},Y^{\mathrm{QP}}) \\ &\quad\propto\ p(\beta_1,\beta_2,\sigma_1^2,\psi_2^2,\psi_{12}|Y^{\mathrm{JC}},Y^{\mathrm{QP}})\cdot(\sigma_1^2)^{-2} \\ &\quad\propto\ \exp\left(-\frac{1}{2\sigma_1^2}\begin{pmatrix} W^{\mathrm{JC}}-(X^{\mathrm{JC}})^\top\beta_1 \\ W^{\mathrm{QP}}-(X^{\mathrm{QP}})^\top\beta_2 \end{pmatrix}^\top\Upsilon\begin{pmatrix} W^{\mathrm{JC}}-(X^{\mathrm{JC}})^\top\beta_1 \\ W^{\mathrm{QP}}-(X^{\mathrm{QP}})^\top\beta_2 \end{pmatrix}\right), \end{aligned} \tag{4.14}$$

where $(\sigma_1^2)^{-2}$ is the Jacobian of the transformation $(\sigma_2^2,\sigma_{12}) \mapsto (\psi_2^2,\psi_{12})$ and $\Upsilon \equiv \begin{pmatrix} \Upsilon_{11} & \Upsilon_{12} \\ \Upsilon_{21} & \Upsilon_{22} \end{pmatrix} = \sigma_1^2\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1}$ is a function of ψ_2^2 and ψ_{12}. The transformation allows us to draw β_1 from the reduced conditional distribution that has a standard form. Based on the posterior distribution in (4.14), STEP 1 of Sampler 4.3.2 is implemented by

$$\beta_1|(\beta_2,\sigma_1^2,\sigma_2^2,\sigma_{12},Y^{\mathrm{JC}},Y^{\mathrm{QP}}) \sim \mathrm{N}_p\Big(\hat{\beta}_1,\sigma_1^2\Big[X^{\mathrm{JC}}\Upsilon_{11}(X^{\mathrm{JC}})^\top\Big]^{-1}\Big), \tag{4.15}$$

where $\hat{\beta}_1 = \Big[X^{\mathrm{JC}}\Upsilon_{11}(X^{\mathrm{JC}})^\top\Big]^{-1}\Big(X^{\mathrm{JC}}\Upsilon_{11}Y^{\mathrm{JC}} + X^{\mathrm{JC}}\Upsilon_{12}\Big[Y^{\mathrm{QP}} - (X^{\mathrm{QP}})^\top\beta_2\Big]\Big)$. Due to the symmetry, STEP 2 of Sampler 4.3.2 is implemented in the same manner as STEP 1, except using $p(\beta_2|\beta_1,\sigma_1^2,\sigma_2^2,\sigma_{21},Y^{\mathrm{JC}},Y^{\mathrm{QP}}) \propto p(\beta_2|\beta_1,\psi_1^2,\sigma_2^2,\psi_{21},Y^{\mathrm{JC}},Y^{\mathrm{QP}})$ where $\psi_1^2 = \sigma_1^2/\sigma_2^2$ and $\psi_{21} = \sigma_{12}/\sigma_2^2$.

Next, STEP 3 imputes the community-level variables given the observed sums and parameters. Because of the difficulty with the non-nested data, we first formulate the joint imputation procedure in terms of $p(W^{\text{JC}}|\theta, Y^{\text{JC}})$ and $p(W^{\text{QP}}|W^{\text{JC}}, \theta, Y^{\text{JC}}, Y^{\text{QP}})$ rather than $p(W^{\text{JC}}|\theta, Y^{\text{JC}}, Y^{\text{QP}})$ and $p(W^{\text{QP}}|W^{\text{JC}}, \theta, Y^{\text{JC}}, Y^{\text{QP}})$, as discussed in Section 4.3.1. Then, given the observed sum, the community-level variables within each county or agency follow a multivariate truncated normal distribution. To implement this step, we use a Gibbs sampling technique by drawing each community-level variable from its complete conditional distribution of the joint distribution. For example, to draw the community-level variables from $p(\{W_i^{\text{JC}}, i \in \mathcal{J}_j\}|\theta, Y_j^{\text{JC}})$, we notice that if we arbitrarily drop one component W_ℓ^{JC}, the remaining components and Y_j^{JC} jointly follow a multivariate normal distribution. Thus it is easy to compute $p(\{W_i^{\text{JC}}, i \in \mathcal{J}_j, i \neq \ell\}|\theta, Y_j^{\text{JC}})$ that is a multivariate truncated normal distribution. We draw each component of $\{W_i^{\text{JC}}, i \in \mathcal{J}_j, i \neq \ell\}$ from the corresponding complete conditional distribution of $p(\{W_i^{\text{JC}}, i \in \mathcal{J}_j, i \neq \ell\}|\theta, Y_j^{\text{JC}})$, and finally we set $W_\ell^{\text{JC}} = Y_j^{\text{JC}} - \sum_{i \in \mathcal{J}_j, i \neq \ell} W_i^{\text{JC}}$. We repeat this Gibbs sampling procedure for some iterations to obtain better imputations for $\{W_i^{\text{JC}}, i \in \mathcal{J}_j, i \neq \ell\}$, so that the nested Gibbs sampling steps are embedded into the MCMC sampler.

With the community-level variables imputed in STEP 3, the covariance matrix of the target model in (4.1) can be easily updated. Specifically, given $(\beta_1, \beta_2, W^{\text{JC}}, W^{\text{QP}}, Y^{\text{JC}}, Y^{\text{QP}})$, STEP 4 draws $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ from an Inverse Wishart distribution,

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \Big| (\beta_1, \beta_2, W^{\text{JC}}, W^{\text{QP}}, Y^{\text{JC}}, Y^{\text{QP}}) \\ \sim \text{Inv}-\text{Wishart}\left(n, \sum_{i=1}^{n} \begin{pmatrix} W_i^{\text{JC}} - X_i^\top \beta_1 \\ W_i^{\text{QP}} - X_i^\top \beta_2 \end{pmatrix} \begin{pmatrix} W_i^{\text{JC}} - X_i^\top \beta_1 \\ W_i^{\text{QP}} - X_i^\top \beta_2 \end{pmatrix}^\top \right). \quad (4.16)$$

Fitting the Bivariate lognormal Model via MCMC

Based on the bivariate lognormal model given in (4.4), we begin our joint imputation procedure by formulating the joint distribution $p(W^{\mathrm{JC}}, W^{\mathrm{QP}}|\vartheta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ in terms of incoherent marginal and conditional distributions, i.e., $p(W^{\mathrm{JC}}|\vartheta_1, Y^{\mathrm{JC}})$ and $p(W^{\mathrm{QP}}|W^{\mathrm{JC}}, \vartheta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, where $\vartheta = (\vartheta_1, \vartheta_2)$, $\vartheta_1 = (\gamma_1, \tau_1^2)$, and $\vartheta_2 = (\gamma_2, \tau_2^2, \tau_{12})$. Note that we use the incoherent marginal distribution as if it were $p(W^{\mathrm{JC}}|\vartheta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, because the correct marginal distribution is unknown from the non-nested data. To create joint imputations, we completely impute W^{JC} from $p(W^{\mathrm{JC}}|\vartheta_1, Y^{\mathrm{JC}})$, then each imputation of W^{JC} is used as an additional covariate to sequentially impute W^{QP} from $p(W^{\mathrm{QP}}|W^{\mathrm{JC}}, \vartheta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$.

To illustrate the componentwise imputation, we first describe the marginal imputation procedure for W^{JC}. We notice that only $n_j - 1$ components of $\{W_i^{\mathrm{JC}}, i \in \mathcal{J}_j\} = (W_{1,j}^{\mathrm{JC}}, W_{2,j}^{\mathrm{JC}}, \ldots, W_{n_j,j}^{\mathrm{JC}})$ are free given the observed sum Y_j^{JC} in county j, which enables us to find the joint distribution $p\big(W_{1,j}^{\mathrm{JC}}, W_{2,j}^{\mathrm{JC}}, \ldots, W_{n_j-1,j}^{\mathrm{JC}}, Y_j^{\mathrm{JC}}|\vartheta_1\big)$ using the transformation $W_{n_j,j}^{\mathrm{JC}} = \log\big(Y_j^{\mathrm{JC}} - \sum_{l=1}^{n_j-1} e^{W_{l,j}^{\mathrm{JC}}}\big)$. Then, using the flat prior distribution on $(\gamma_1, \log\tau_1)$, i.e., $p(\vartheta_1) \propto (\tau_1^2)^{-1}$, the target posterior distribution of interest is given by

$$
\begin{aligned}
&p\big(\{W_{l,j}^{\mathrm{JC}}, l = 1, \ldots, n_j - 1, j = 1, \ldots, J\}, \vartheta_1 \big| Y^{\mathrm{JC}}\big) \\
&\quad \propto \ \exp\Bigg(-\frac{1}{2\tau_1^2} \sum\nolimits_{j=1}^{J} \Bigg\{ \sum\nolimits_{l=1}^{n_j-1} \Big[W_{l,j}^{\mathrm{JC}} - (\log X_{l,j})^\top \gamma_1 \Big]^2 \\
&\qquad + \Bigg[\log\Big(Y_j^{\mathrm{JC}} - \sum\nolimits_{l=1}^{n_j-1} e^{W_{l,j}^{\mathrm{JC}}} \Big) - (\log X_{n_j,j})^\top \gamma_1 \Bigg]^2 \Bigg\} \Bigg) \\
&\qquad \times (\tau_1^2)^{-(n/2+1)} \prod\nolimits_{j=1}^{J} \Big(Y_j^{\mathrm{JC}} - \sum\nolimits_{l=1}^{n_j-1} e^{W_{l,j}^{\mathrm{JC}}} \Big)^{-1}. \qquad (4.17)
\end{aligned}
$$

Using the complete conditional distributions of the target distribution in (4.17), we can construct the MCMC sampler for county j, which iterates between

STEP 1 : Draw $W_{l,j}^{\mathrm{JC}}$ from $p(W_{l,j}^{\mathrm{JC}}|W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}})$ (Sampler 4.3.3)

for $l = 1, 2, \ldots, n_j - 1$,

STEP 2 : Draw ϑ_1 from $p(\vartheta_1 | W_{1,j}^{\mathrm{JC}}, W_{2,j}^{\mathrm{JC}}, \ldots, W_{n_j-1,j}^{\mathrm{JC}}, Y^{\mathrm{JC}})$,

where $W_{-l,j}^{\mathrm{JC}} = (W_{1,j}^{\mathrm{JC}}, \ldots, W_{l-1,j}^{\mathrm{JC}}, W_{l+1,j}^{\mathrm{JC}}, \ldots, W_{n_j-1,j}^{\mathrm{JC}})$. Because the conditional distribution in STEP 1 of Sampler 4.3.3 is not a standard distribution, i.e.,

$$\begin{aligned} p(W_{l,j}^{\mathrm{JC}} | W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}}) \;\propto\; & \exp\Bigg(-\frac{1}{2\tau_1^2}\Bigg\{ \Big[W_{l,j}^{\mathrm{JC}} - (\log X_{l,j})^\top \gamma_1 \Big]^2 \\ & + \Bigg[\log\Big(Y_j^{\mathrm{JC}} - \textstyle\sum_{l=1}^{n_j-1} e^{W_{l,j}^{\mathrm{JC}}} \Big) - (\log X_{n_j,j})^\top \gamma_1 \Bigg]^2 \Bigg\} \Bigg) \\ & \times \Big(Y_j^{\mathrm{JC}} - \textstyle\sum_{l=1}^{n_j-1} e^{W_{l,j}^{\mathrm{JC}}} \Big)^{-1}, \end{aligned} \tag{4.18}$$

we use a Metropolis-Hastings algorithm (Metropolis and Ulam, 1949; Hastings, 1970) with the truncated Gaussian distribution,

$$J_t(W_{l,j}^{\mathrm{JC}} | W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}}) \;\propto\; \mathrm{N}\Big((\log X_{l,j})^\top \gamma_1, \tau_1^2 \Big) \cdot 1_{\left\{ W_{l,j}^{\mathrm{JC}} \,<\, Y_j^{\mathrm{JC}} - \sum e^{W_{-l,j}^{\mathrm{JC}}} \right\}}, \tag{4.19}$$

as the jumping rule at iteration t, where $1_{\{A\}}$ is 1 if A is true and 0 otherwise. We accept the draw $(W_{l,j}^{\mathrm{JC}})^*$ from $J_t(W_{l,j}^{\mathrm{JC}} | W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}})$ with probability

$$\frac{p\big((W_{l,j}^{\mathrm{JC}})^* \big| W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}}\big) \big/ J_t\big((W_{l,j}^{\mathrm{JC}})^* \big| W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}}\big)}{p\big((W_{l,j}^{\mathrm{JC}})^{(t-1)} \big| W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}}\big) \big/ J_t\big((W_{l,j}^{\mathrm{JC}})^{(t-1)} \big| W_{-l,j}^{\mathrm{JC}}, \vartheta_1, Y^{\mathrm{JC}}\big)}. \tag{4.20}$$

After STEP 1 of Sampler 4.3.3, the last component of $\{W_i^{\mathrm{JC}}, i \in \mathcal{J}_j\}$ in county j is recovered via $W_{n_j,j}^{\mathrm{JC}} = \log\big(Y_j^{\mathrm{JC}} - \sum_{l=1}^{n_j-1} e^{W_{l,j}^{\mathrm{JC}}}\big)$. With the draws of $\{W_{l,j}^{\mathrm{JC}}, l = 1, \ldots, n_j - 1, j = 1, \ldots, J\}$ from STEP 1, the conditional distribution in STEP 2 of Sampler 4.3.3 is rewritten as

$$\begin{aligned} & p(\vartheta_1 | \{W_{l,j}^{\mathrm{JC}}, l = 1, \ldots, n_j - 1, j = 1, \ldots, J\}, Y^{\mathrm{JC}}) \\ & \quad = \; p(\vartheta_1 | W^{\mathrm{JC}}) \\ & \quad \propto \; (\tau_1^2)^{-(n/2+1)} \exp\Bigg(-\frac{1}{2\tau_1^2} \sum_{i=1}^{n} \Big[W_i^{\mathrm{JC}} - (\log X_i)^\top \gamma_1 \Big]^2 \Bigg). \end{aligned} \tag{4.21}$$

That is, we can easily draw $\vartheta_1 = (\gamma_1, \tau_1^2)$ using standard probability distributions derived from (4.21). After monitoring the convergence of the chain, we create multiple imputations for W^{JC}, which are used as additional covariates in the conditional imputations for W^{QP}.

The imputation procedure for W^{QP} is the same as above, except that we have an additional covariate, W^{JC}. For computational convenience, we use the transformation $\widetilde{\vartheta} \mapsto \vartheta$ where $\widetilde{\vartheta} = (\vartheta_1, \widetilde{\vartheta}_2) = (\gamma_1, \tau_1^2, \gamma_{2|1}, \tau_{2|1}^2, \alpha_{2|1})$, $\gamma_{2|1} = \gamma_2 - \alpha_{2|1}\gamma_1$, $\tau_{2|1}^2 = \tau_2^2 - \alpha_{2|1}^2\tau_1^2$, and $\alpha_{2|1} = \tau_{12}/\tau_1^2$. Because the imputations of W^{QP} are of our main interest, we can rewrite the imputation procedure $p(W^{\mathrm{QP}}|W^{\mathrm{JC}}, \vartheta, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ as $p(W^{\mathrm{QP}}|W^{\mathrm{JC}}, \widetilde{\vartheta}_2, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$. Using the flat prior distribution on $(\alpha_{2|1}, \gamma_{2|1}, \log\tau_{2|1})$, i.e., $p(\widetilde{\vartheta}_2) \propto (\tau_{2|1}^2)^{-1}$, the second target posterior distribution is given by

$$
\begin{aligned}
&p\big(\{W_{l,k}^{\mathrm{QP}}, l = 1, \ldots, m_k - 1, k = 1, \ldots, K\}, \widetilde{\vartheta}_2 \big| W^{\mathrm{JC}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}}\big) \\
&\quad\propto\ \exp\Bigg(-\frac{1}{2\tau_{2|1}^2} \sum\nolimits_{k=1}^{K} \Bigg\{ \sum\nolimits_{l=1}^{m_k-1} \Big[W_{l,k}^{\mathrm{QP}} - (\log X_{l,k})^\top \gamma_{2|1} - W_{l,k}^{\mathrm{JC}} \alpha_{2|1} \Big]^2 \\
&\qquad + \left[\log\left(Y_k^{\mathrm{QP}} - \sum\nolimits_{l=1}^{m_k-1} e^{W_{l,k}^{\mathrm{QP}}} \right) - (\log X_{m_k,k})^\top \gamma_{2|1} - W_{m_k,k}^{\mathrm{JC}} \alpha_{2|1} \right]^2 \Bigg\} \Bigg) \\
&\qquad \times (\tau_{2|1}^2)^{-(n/2+1)} \prod\nolimits_{k=1}^{K} \left(Y_k^{\mathrm{QP}} - \sum\nolimits_{l=1}^{m_k-1} e^{W_{l,k}^{\mathrm{QP}}} \right)^{-1}. \qquad (4.22)
\end{aligned}
$$

Using the complete conditional distributions of the target distribution in (4.22), we construct another MCMC sampler for agency k, which iterates between

STEP 1 : Draw $W_{l,k}^{\mathrm{QP}}$ from $p(W_{l,k}^{\mathrm{QP}}|W_{-l,k}^{\mathrm{QP}}, W^{\mathrm{JC}}, \widetilde{\vartheta}_2, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$ (Sampler 4.3.4)

for $l = 1, 2, \ldots, m_k - 1$,

STEP 2 : Draw $\widetilde{\vartheta}_2$ from $p(\widetilde{\vartheta}_2|W_{1,k}^{\mathrm{QP}}, W_{2,k}^{\mathrm{QP}}, \ldots, W_{m_k-1,k}^{\mathrm{QP}}, W^{\mathrm{JC}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$,

where $W_{-l,k}^{\mathrm{QP}} = (W_{1,k}^{\mathrm{QP}}, \ldots, W_{l-1,k}^{\mathrm{QP}}, W_{l+1,k}^{\mathrm{QP}}, \ldots, W_{m_k-1,k}^{\mathrm{QP}})$. Specifically, the conditional distribution in STEP 1 of Sampler 4.3.4 is given by

$$
\begin{aligned}
&p(W_{l,k}^{\mathrm{QP}}|W_{-l,k}^{\mathrm{QP}}, W^{\mathrm{JC}}, \widetilde{\vartheta}_2, Y^{\mathrm{JC}}, Y^{\mathrm{QP}}) \\
&\quad\propto\ \exp\Bigg(-\frac{1}{2\tau_{2|1}^2} \Bigg\{ \Big[W_{l,k}^{\mathrm{QP}} - (\log X_{l,k})^\top \gamma_{2|1} - W_{l,k}^{\mathrm{JC}} \alpha_{2|1} \Big]^2 \\
&\qquad + \left[\log\left(Y_k^{\mathrm{QP}} - \sum\nolimits_{l=1}^{m_k-1} e^{W_{l,k}^{\mathrm{QP}}} \right) - (\log X_{m_k,k})^\top \gamma_{2|1} - W_{m_k,k}^{\mathrm{JC}} \alpha_{2|1} \right]^2 \Bigg\} \Bigg) \\
&\qquad \times \left(Y_k^{\mathrm{QP}} - \sum\nolimits_{l=1}^{m_k-1} e^{W_{l,k}^{\mathrm{QP}}} \right)^{-1}. \qquad (4.23)
\end{aligned}
$$

Because $W_{l,k}^{\mathrm{QP}}$ does not follow a standard distribution, as shown in (4.23), we again use a Metropolis-Hastings algorithm with the truncated Gaussian distribution,

$$\begin{aligned} & J_t(W_{l,k}^{\mathrm{QP}} | W_{-l,k}^{\mathrm{QP}}, W^{\mathrm{JC}}, \widetilde{\vartheta}_2, Y^{\mathrm{JC}}, Y^{\mathrm{QP}}) \\ & \quad \propto \mathrm{N}\Big((\log X_{l,k})^\top \gamma_{2|1} + W_{l,k}^{\mathrm{JC}} \alpha_{2|1}, \tau_{2|1}^2 \Big) \cdot 1_{\left\{ W_{l,k}^{\mathrm{QP}} < Y_k^{\mathrm{QP}} - \sum e^{W_{-l,k}^{\mathrm{QP}}} \right\}}, \end{aligned} \tag{4.24}$$

as the jumping rule at iteration t. After STEP 1 of Sampler 4.3.4, the last component of $\{W_i^{\mathrm{QP}}, i \in \mathcal{K}_k\}$ for agency k is also recovered from $W_{m_k,k}^{\mathrm{QP}} = \log\big(Y_k^{\mathrm{QP}} - \sum_{l=1}^{m_k-1} e^{W_{l,k}^{\mathrm{QP}}}\big)$. Once we complete STEP 1, the conditional distribution in STEP 2 of Sampler 4.3.4 is simplified as

$$\begin{aligned} & p(\widetilde{\vartheta}_2 | \{W_{l,k}^{\mathrm{QP}}, l = 1, \ldots, m_k - 1, k = 1, \ldots, K\}, W^{\mathrm{JC}}, Y^{\mathrm{JC}}, Y^{\mathrm{QP}}) \\ & \quad = p(\widetilde{\vartheta}_2 | W^{\mathrm{JC}}, W^{\mathrm{QP}}) \\ & \quad \propto (\tau_{2|1}^2)^{-(n/2+1)} \exp\Bigg(- \frac{1}{2\tau_{2|1}^2} \sum\nolimits_{i=1}^{n} \Big[W_i^{\mathrm{QP}} - (\log X_i)^\top \gamma_{2|1} - W_i^{\mathrm{JC}} \alpha_{2|1} \Big]^2 \Bigg). \end{aligned} \tag{4.25}$$

Thus, the regression parameters $\widetilde{\vartheta}_2 = (\gamma_{2|1}, \tau_{2|1}^2, \alpha_{2|1})$ can be drawn from standard probability distributions derived from (4.25).

Fitting the Bivariate Lognormal Model Using SIR

As an another approach to fitting the bivariate lognormal model in (4.4), we consider the componentwise imputations using SIR. Under the target model in (4.4), the marginal distribution of Z^{JC} is given by

$$\log Z_i^{\mathrm{JC}} \overset{\mathrm{ind}}{\sim} \mathrm{N}\Big((\log X_i)^\top \gamma_1, \tau_1^2 \Big) \text{ for } i = 1, 2, \ldots, n, \tag{4.26}$$

which implies that

$$Y_j^\star \equiv \sum_{i \in \mathcal{J}_j} \log Z_i^{\mathrm{JC}} \overset{\mathrm{ind}}{\sim} \mathrm{N}\Bigg(\Big(\sum_{i \in \mathcal{J}_j} \log X_i \Big)^\top \gamma_1, \tau_1^2 n_j \Bigg) \text{ for } j = 1, 2, \ldots, J. \tag{4.27}$$

Our goal is to impute the variable at the community level. Ideally, if we observed $Y^\star = (Y_1^\star, Y_2^\star, \ldots, Y_J^\star)$, we could easily compute $p(\{Z_i^{\mathrm{JC}}, i \in \mathcal{J}_j\} | Y_j^\star, \vartheta_1)$ using properties of the multivariate normal distribution and impute the community-level

response variable Z^{JC}, as illustrated by fitting the Bivariate Gaussian model in Section 4.3.2. Unfortunately, however, we do not observe $Y^{\star}$ and cannot compute it from the observed $Y^{\text{JC}} = \{\sum_{i \in \mathcal{J}_j} Z_i^{\text{JC}}, j = 1, 2, \ldots, J\}$ because $\log Y_j^{\text{JC}} = \log \sum_{i \in \mathcal{J}_j} Z_i^{\text{JC}} \neq \sum_{i \in \mathcal{J}_j} \log Z_i^{\text{JC}} = Y_j^{\star}$ for $n_j > 1$. Thus, directly fitting the bivariate lognormal model is intractable based on our observed data.

To circumvent this difficulty, we can impute a large number of "proposal" data sets under a *working model* that is relatively tractable, and resample several imputations according to the importance ratios computed using the posterior predictive distributions under the target and working models, i.e., $p(Z^{\text{JC}}|Y^{\star})/p(Z^{\text{JC}}|Y^{\text{JC}})$. The importance ratio measures the likelihood of a particular value of the response variables in each community under the target model in (4.26) relative to its likelihood under the working model. By resampling the proposal data sets according to the importance ratios, we are correcting the working model toward the target model: Proposal data sets that are more likely under the working model than under the target model tend to be discarded, whereas proposal data sets that are more likely under the target model tend to be retained. As a tractable working model, we consider a truncated normal model, i.e.,

$$Z_i^{\text{JC}} \overset{\text{ind}}{\sim} \text{N}(X_i^{\top} \widetilde{\gamma}_1, \widetilde{\tau}_1^2) 1_{\{Z_i^{\text{JC}} > 0\}} \text{ for } i = 1, 2, \ldots, n, \tag{4.28}$$

where Z_i^{JC} is truncated to be positive.

The importance resampling procedure begins by fitting the working model. Under the working model in (4.28), the observed Y_j^{JC} in county j also follow a truncated normal distribution,

$$Y_j^{\text{JC}} \overset{\text{ind}}{\sim} \text{N}\Big(\Big(\sum_{i \in \mathcal{J}_j} X_i \Big)^{\top} \widetilde{\gamma}_1, \widetilde{\tau}_1^2 n_j \Big) 1_{\{Y_i^{\text{JC}} > 0\}}, \text{ for } j = 1, 2, \ldots, J, \tag{4.29}$$

due to the additivity of the variables. Based on (4.29), we simulate M draws of $\widetilde{\vartheta}_1 = (\widetilde{\gamma}, \widetilde{\tau}_1^2)$ from the posterior distribution $p(\widetilde{\vartheta}_1|Y)$. Because the likelihood implied by (4.29) involves a truncated normal distribution under the working model, a

least squares method based on a normal distribution cannot be directly used to fit the parameters. Instead, we tentatively impute the supposed truncated negative values of the response variable given the corresponding covariates and the current draw of the parameters, add them to the observed data set with the covariates, and use weighted least squares to update the parameters $\widetilde{\vartheta}_1$. Specifically, we draw the number of negative values of Y_j^{JC} from a geometric distribution,

$$T_j \sim \mathrm{Geometric}\big(p(Y_j^{\mathrm{JC}} < 0)\big) \text{ for } j = 1, 2, \ldots, J, \tag{4.30}$$

and impute the negative response variables via a truncated normal distribution,

$$V_{\ell j}^{\mathrm{JC}} \overset{\mathrm{ind}}{\sim} \mathrm{N}\bigg(\Big(\sum_{i \in \mathcal{J}_j} X_i\Big)^{\top} \widetilde{\gamma}_1, \widetilde{\tau}_1^2 n_j\bigg) 1_{\{V_{\ell j}^{\mathrm{JC}} < 0\}}, \text{ for } \ell = 1, \ldots, T_j \text{ and } j = 1, \ldots, J. \tag{4.31}$$

This procedure is iteratively repeated in order to obtain a sample from the posterior distribution of $\widetilde{\vartheta}_1$. We emphasize that the supposed truncated negative values of the response variable are introduced for computational convenience. They have no interpretation outside of the sampling algorithm. Given the posterior draws of $\widetilde{\vartheta}_1$ and the observed sums Y^{JC}, we impute M data sets of Z^{JC} at the community level by drawing from $p(Z^{\mathrm{JC}}|\widetilde{\vartheta}_1, Y^{\mathrm{JC}})$. In particular, the joint conditional distribution of $\{Z_i^{\mathrm{JC}}, i \in \mathcal{J}_j\}$ given the parameters $\widetilde{\vartheta}_1$ and the observed sum Y_j^{JC} is a multivariate truncated normal distribution that is truncated to an interval between 0 and the total sum Y_j^{JC}. Directly drawing from the multivariate truncated normal distribution is not straightforward, so that we take advantage of the nested Gibbs sampling technique used when the bivariate Gaussian model is fitted.

Let $Z_i^{\mathrm{JC}(m)}$ denote the mth data set imputed in community i by using the mth posterior draw of $\widetilde{\vartheta}_1$ for $i = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, M$. Once we obtain the M imputed data sets at the community level, we must compute importance ratios that are used to compare the target and working models and to resample several imputations based on the importance ratios. To do this, the posterior predictive distributions under the target and working models are evaluated at each of the M imputed data sets. That is, the objective is to compute $p(\log Z^{\mathrm{JC}(m)}|Y^{\star})$ under the

target model and $p(Z^{\mathrm{JC}(m)}|Y^{\mathrm{JC}})$ under the working model, for $m = 1, 2, \ldots, M$. Unfortunately, $p(\log Z^{\mathrm{JC}(m)}|Y^{\star})$ cannot be evaluated directly, because $Y^{\star}$ is not observed. To circumvent this difficulty, we instead compute $p(\log Z^{\mathrm{JC}(m)}|Y^{\star(m)})$ where $Y^{\star(m)} = \sum_{i \in \mathcal{J}_j} \log Z_i^{\mathrm{JC}(m)}$. Thus, for each $Z^{\mathrm{JC}(m)}$, we evaluate the posterior predictive distribution under the target model,

$$\begin{aligned} p(\log Z^{\mathrm{JC}(m)}|Y^{\star(m)}) &= \int p(\log Z^{\mathrm{JC}(m)}|\vartheta_1) p(\vartheta_1|Y^{\star(m)}) d\vartheta_1 \\ &\approx \frac{1}{L} \sum_{\ell=1}^{L} p(\log Z^{\mathrm{JC}(m)}|\vartheta_1^{(\ell)}), \end{aligned} \tag{4.32}$$

where $\vartheta_1^{(\ell)}$ for $\ell = 1, 2, \ldots, L$ represents the posterior samples of ϑ_1 from $p(\vartheta_1|Y^{\star(m)})$ under the target model with $Y^{\star}$ replaced by its value under the mth imputed data set. The L draws of ϑ_1 are used to numerically integrate ϑ_1 out of the joint distribution $p(\log Z^{\mathrm{JC}(m)}, \vartheta_1|Y^{\star(m)})$. The posterior predictive distribution under the working model is evaluated via

$$\begin{aligned} p(Z^{\mathrm{JC}(m)}|Y) &= \int p(Z^{\mathrm{JC}(m)}|\widetilde{\vartheta}_1) p(\widetilde{\vartheta}_1|Y) d\widetilde{\vartheta}_1 \\ &\approx \frac{1}{L} \sum_{\ell=1}^{L} p(Z^{\mathrm{JC}(m)}|\widetilde{\vartheta}_1^{(\ell)}), \end{aligned} \tag{4.33}$$

where $\widetilde{\vartheta}_1^{(\ell)}$ for $\ell = 1, 2, \ldots, L$ represents the posterior samples from $p(\widetilde{\vartheta}_1|Y)$ under the working model. The L draws of $\widetilde{\vartheta}_1$ are used to numerically integrate $\widetilde{\vartheta}_1$ out of the joint distribution $p(Z^{\mathrm{JC}(m)}, \widetilde{\vartheta}_1|Y)$. With the two evaluated posterior predictive distributions in hand, we can compute the importance ratios (IR) for the M imputed data sets, as

$$\begin{aligned} \mathrm{IR}_m \equiv \frac{p(Z^{\mathrm{JC}(m)}|Y^{\star})}{p(Z^{\mathrm{JC}(m)}|Y)} &\approx \frac{p(Z^{\mathrm{JC}(m)}|Y^{\star(m)})}{p(Z^{\mathrm{JC}(m)}|Y)} \\ &= \frac{p(\log Z^{\mathrm{JC}(m)}|Y^{\star(m)})|J_m|}{p(Z^{\mathrm{JC}(m)}|Y)}, \text{ for } m = 1, 2, \ldots, M, \end{aligned} \tag{4.34}$$

where $|J_m|$ is the Jacobian of the transformation $Z^{\mathrm{JC}} \mapsto \log Z^{\mathrm{JC}}$ for the mth imputed data set given by

$$|J_m| = \prod_{i=1}^{n} \frac{1}{Z_i^{\mathrm{JC}(m)}} = \exp\left(-\sum_{i=1}^{n} \log Z_i^{\mathrm{JC}(m)} \right). \tag{4.35}$$

Using probabilities that are proportional to the importance ratios, we resample several sets of community-level imputations without replacement.

Once we obtain M imputations of Z^{JC}, we sequentially impute the second component Z^{QP} using each of the M imputations as an additional covariate. That is, the above procedure is repeated for each imputation of Z^{JC} to obtain the corresponding imputation of Z^{QP}, after we replacing the target imputation model $p(Z^{\mathrm{JC}}|\vartheta_1)$ with $p(Z^{\mathrm{QP}}|Z^{\mathrm{JC}}, \vartheta)$.

Fitting the Poisson Regression Model

To create joint imputations from the Poisson regression model in (4.5) and (4.6), we employ the componentwise imputation method. That is, the first component Z^{JC} is imputed from the incoherent marginal distribution $p(Z^{\mathrm{JC}}|\delta_1, Y^{\mathrm{JC}})$, then we impute the second component Z^{QP} from the conditional distribution $p(Z^{\mathrm{QP}}|Z^{\mathrm{JC}}, \delta_1, \delta_2, Y^{\mathrm{JC}}, Y^{\mathrm{QP}})$, using each imputation of Z^{JC} as an additional covariate for each imputation of Z^{QP}.

We begin our imputation method by illustrating the marginal imputation procedure of Z^{JC}. Based on the Poisson regression model in (4.5), the observed sum in county j, $Y_j^{\mathrm{JC}} = \sum_{i \in \mathcal{J}_j} Z_i^{\mathrm{JC}}$, follows a Poisson distribution,

$$Y_j^{\mathrm{JC}} \overset{\mathrm{ind}}{\sim} \mathrm{Poisson}\Big(\sum\nolimits_{i \in \mathcal{J}_j} \lambda_i^{\mathrm{JC}}\Big) = \mathrm{Poisson}\Big(\sum\nolimits_{i \in \mathcal{J}_j} e^{X_i^\top \delta_1}\Big), \tag{4.36}$$

for $j = 1, 2, \ldots, J$. Thus, under the flat prior distribution $p(\delta_1) \propto 1$, the observed-data log posterior distribution is given by

$$\log p(\delta_1|Y^{\mathrm{JC}}) = \text{constant} + \sum_{j=1}^{J} \left\{ Y_j^{\mathrm{JC}} \log \left(\sum_{i \in \mathcal{J}_j} e^{X_i^\top \delta_1} \right) \right\} - \sum_{j=1}^{J} \sum_{i \in \mathcal{J}_j} e^{X_i^\top \delta_1}. \tag{4.37}$$

Because the posterior distribution of δ_1 is not a standard distribution, we sample δ_1 using a Metropolis-Hastings algorithm. As a jumping rule, we choose the multivariate Gaussian distribution whose mode and curvature are computed using

a multivariate Newton-Raphson method and matched to the observed posterior distribution. To compute the posterior mode, $\hat{\delta}_1$, we run the following Newton-Raphson iteration from a starting value $\delta_1^{(0)}$ until a pre-specified precision is met:

$$\delta_1^{(t)} = \delta_1^{(t-1)} - \left[L''\left(\delta_1^{(t-1)}\right)\right]^{-1} L'\left(\delta_1^{(t-1)}\right), \text{ for } t = 1, 2, \ldots, \tag{4.38}$$

where $L(\delta_1) \equiv \log p(\delta_1 | Y^{\mathrm{JC}})$, $L'(\delta_1)$ is a $p \times 1$ vector of first derivatives of the log posterior distribution, and $L''(\delta_1)$ is a $p \times p$ matrix of second derivatives of the log posterior distribution.

Given the posterior draws of δ_1, we impute the community-level data, Z^{JC}, subject to the observed sums, Y^{JC}. In particular, the conditional distribution $p(\{Z_i^{\mathrm{JC}}, i \in \mathcal{J}_j\} | \delta_1, Y_j^{\mathrm{JC}})$ follows a multinomial distribution, i.e.,

$$\{Z_i^{\mathrm{JC}}, i \in \mathcal{J}_j\} \big| (\delta_1, Y_j^{\mathrm{JC}}) \;\sim\; \text{Multinomial}\left(Y_j^{\mathrm{JC}}, \frac{\left\{\exp(X_i^\top \delta_1), i \in \mathcal{J}_j\right\}}{\sum_{i \in \mathcal{J}_j} \exp(X_i^\top \delta_1)}\right). \tag{4.39}$$

After completely imputing Z^{JC}, we use the imputations of Z^{JC} to sequentially impute Z^{QP} in the model (4.6). Thus, we repeat the same imputation procedure as above to create the imputations of Z^{QP}, except that each imputation of Z^{JC} is used as an additional covariate.

4.4 Simulation Study

To test and compare our models and computational methods, we conduct four simulation studies. We simulate data for a number of communities in a fixed region, and then aggregate the community-level data to the county or agency level. In our simulation studies, we have the same number of communities (i.e., $n = 120$), 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.

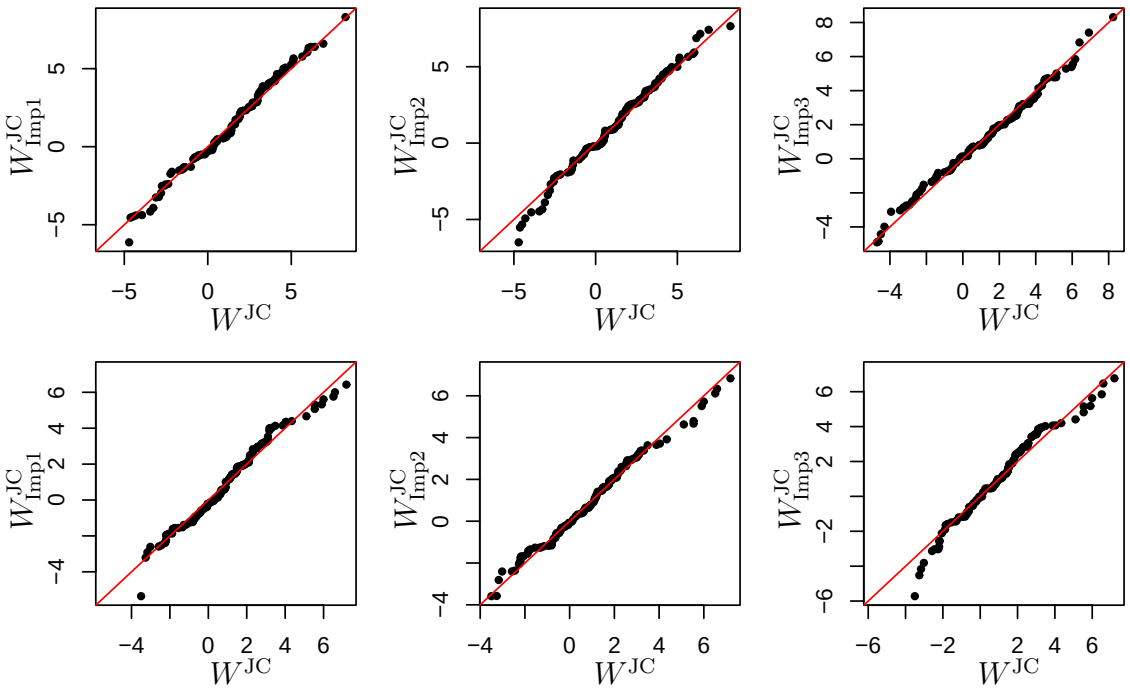

Figure 4.2: Quantile-Quantile Plots of Joint Multiple Imputations and Original Data Simulated Under the Bivariate Gaussian Model. The first row of the figure is drawn for the Job Creation variable, W^{JC}, and the second row for the Qualification Program variable, W^{QP}.

For the bivariate Gaussian model, we generate data using the regression coefficients, $\beta_1 = (1.0\ 0.5)^\top$ and $\beta_2 = (1.0\ 0.2)^\top$, and the covariance matrix, $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 1.8 \\ 1.8 & 4 \end{pmatrix}$. The values of these model parameters are used to simulate $n = 120$ community-level data. Our observed data are given by taking the partial sums of every five Z^{JC} variables (i.e, Y^{JC}) and every eight Z^{QP} variables (i.e, Y^{QP}). To compare data imputed under the model to the original full data, we construct a quantile-quantile plot. If the multiple imputations are from the same distribution as the original data, the quantile-quantile plot will be linear, i.e., the points in the plot follow a 45-degree line. As confirmed in Figure 4.2, each quantile-quantile plot follows almost exactly the 45-degree line, so that our imputations are very close to the original data. That is, the partially marginalized MCMC sampler performs rea-

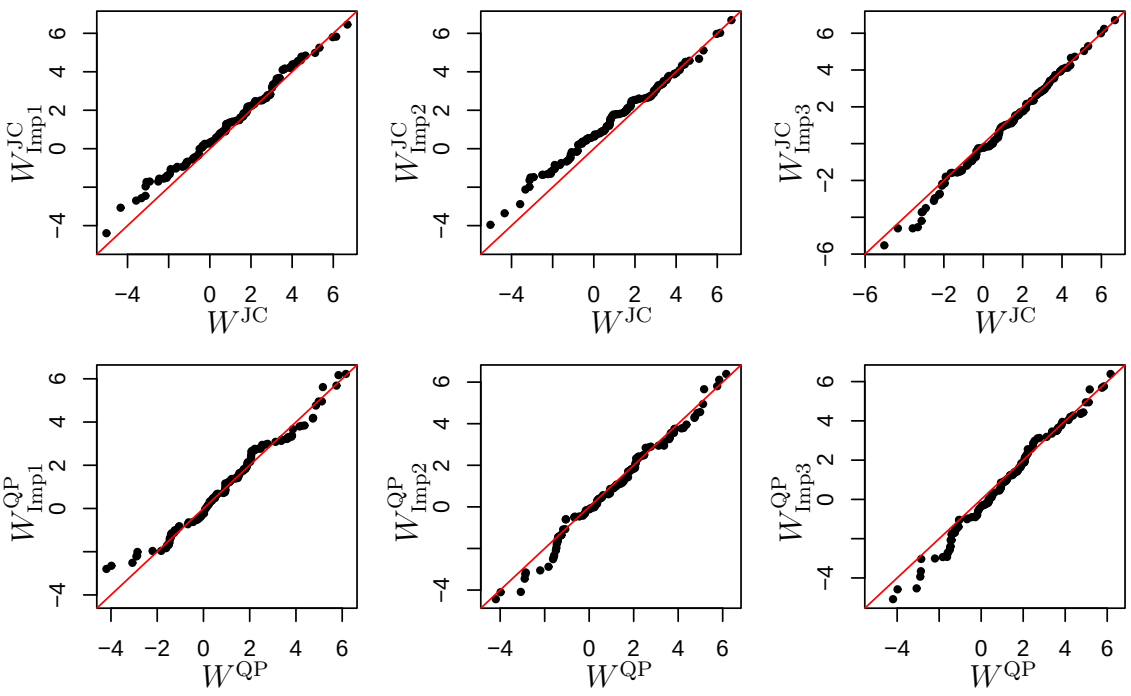

Figure 4.3: Quantile-Quantile Plots of Joint Multiple Imputations and Original Data Simulated Under the Bivariate Lognormal Model. The first row of the figure is drawn for the Job Creation variable, $W^{\mathrm{JC}} = \log Z^{\mathrm{JC}}$, and the second row for the Qualification Program variable, $W^{\mathrm{QP}} = \log Z^{\mathrm{QP}}$.

sonably well, although the incoherent marginal distribution is substituted for the correct marginal distribution.

As the set up for the second simulation study, we fix the number of communities at $n = 120$ and construct the counties and agencies the same way as before. We use the same values of the model parameters, i.e., the regression coefficients are $\gamma_1 = (1.0\ 0.5)^\top$ and $\gamma_2 = (1.0\ 0.2)^\top$, and the covariance matrix is $\begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 1.8 \\ 1.8 & 4 \end{pmatrix}$. As described in Section 4.3.2, we fit the test data simulated under the bivariate lognormal model by using MCMC samplers. In particular, we first create multiple imputations for one component of a bivariate response variable using an MCMC sampler and then, using each imputation as an additional covariate, we create the corresponding imputation for the other component

using another MCMC sampler. The number of sampling steps that compose each MCMC sampler increases as more communities are nested in each of the counties or agencies. That is, if the number of communities is much greater than the number of counties or agencies, we have more community-level data to impute. This may cause inefficiency in the algorithm and unreliable imputations because our observed information is much less than missing information. Moreover, if the residual variances are large, the computation may be unstable because σ_1^2 and σ_2^2 are the variances on the log scale of Z^{JC} and Z^{QP}, respectively. In a moderate situation with reasonably large number of communities in each county or agency and with relatively small residual variances, however, we expect this strategy to be efficient and reliable. Figure 4.3 shows the comparison of the joint multiple imputations and the original data simulated with the true values of the parameters. The quantile-quantile plots are almost linear for both W^{JC} and W^{QP}.

As another way to create joint multiple imputations, we can fit the bivariate lognormal model using the SIR algorithm. However, when a truncated normal distribution is considered as a working model, it fails to cover the thick right tail probability of the lognormal distribution in the target model. To illustrate how the SIR method performs, we generate $n = 120$ community-level response variables from a univariate lognormal distribution. Our observed data are composed of $J = 120/5 = 24$ county-level variables, by combining every five of the community-level variables. For the model parameters, we use $\gamma_1 = (1.0\ 0.5)^\top$ and $\tau_1^2 = 1$. After imputing 200 community-level variables under the working model, we resample the $M = 6$ best imputations according to the importance ratios. Figure 4.4 shows the quantile-quantile plots comparing these multiple imputations with the original data and illustrates the working model fails to generate the extreme values observed in the original data.

Due to this limitation, we may consider another tractable but more flexible working model with a much thicker tail than a truncated normal distribution. That is,

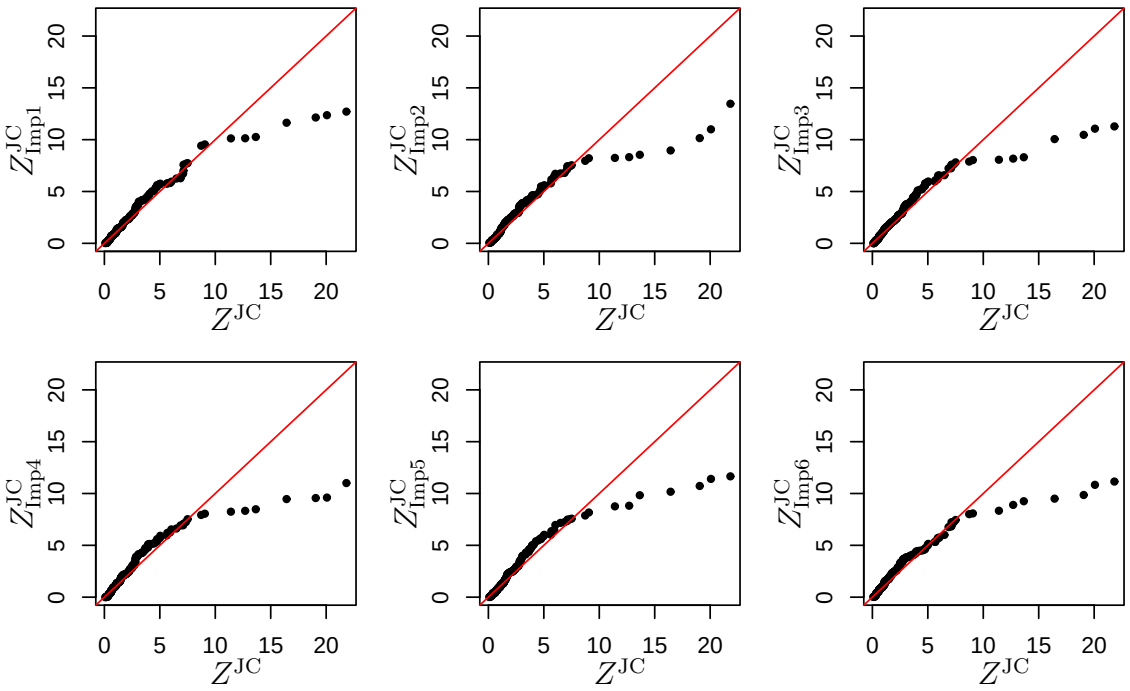

Figure 4.4: Quantile-Quantile Plots of Multiple Imputations and Original Data Simulated Under the Lognormal Model. Because the truncated normal distribution of the working model has a thinner right tail probability then the lognormal distribution, all the six imputations clearly do not cover the right tail of the original data simulated under the lognormal model. This simulation study is done only for one component of the bivariate response variable, i.e., Z^{JC}, but it suffices for illustrating the problem.

we can use a truncated t model with degrees of freedom (df) adjusted to allow for more or less thickness in the tails. For the truncated t model, we use the same procedure as with a truncated normal model, except that we replace the weighted least squares fitting for the parameters with a weighted t-regression routine (i.e., iteratively reweighted least squares fitting). However, another simulation study not shown here illustrates that a truncated t model also fails to cover extreme values under a lognormal distribution.

Lastly, we fit the Poisson regression model for non-nested data. In our simulation study, we use $\delta_1 = (1.0\ 0.5)^\top$ and $\delta_2 = (1.0\ 0.2)^\top$. It is difficult to model the correla-

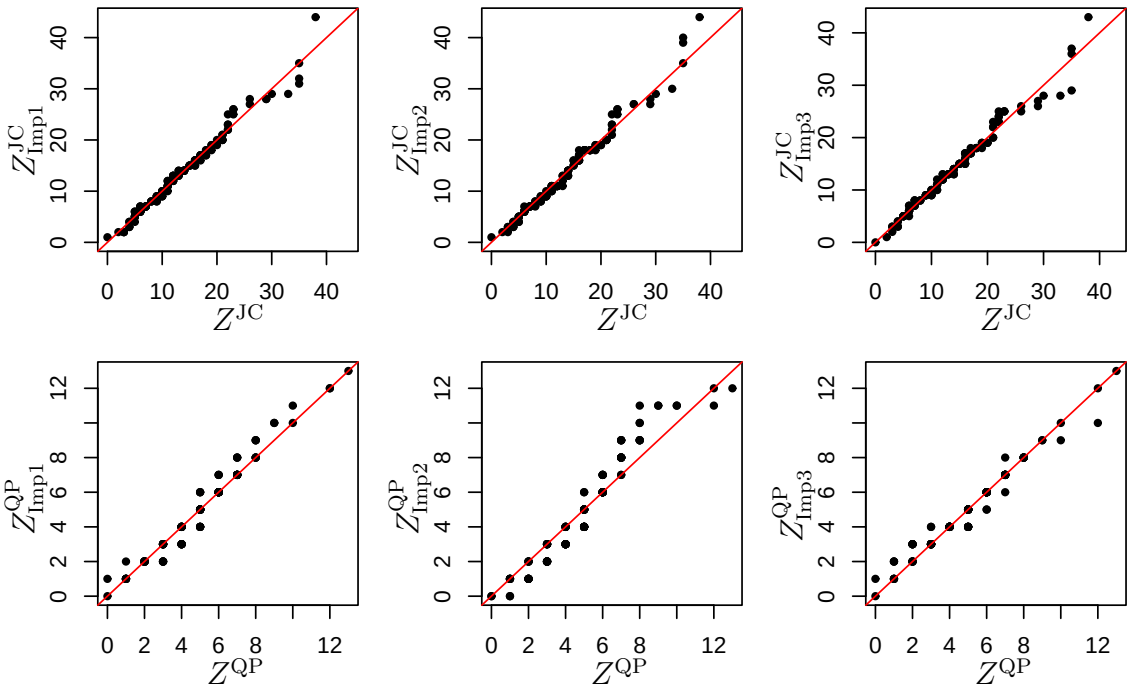

Figure 4.5: Quantile-Quantile Plots of Joint Multiple Imputations and Original Data Simulated Under the Poisson Regression Model for Non-Nested Data. The first row of the figure is drawn for the Job Creation variable, Z^{JC}, and the second row for the Qualification Program variable, Z^{QP}.

tion structure of a bivariate response variable using the Poisson regression model, but practically the joint imputation model performs very well. Fitting the Poisson regression model is less computationally intensive compared to the other models, and is more robust to the extreme cases. (This is verified through other simulation studies with many different values of the parameters.) In Figure 4.5, we confirm that our multiple imputations are very close to the original data simulated with known parameters.

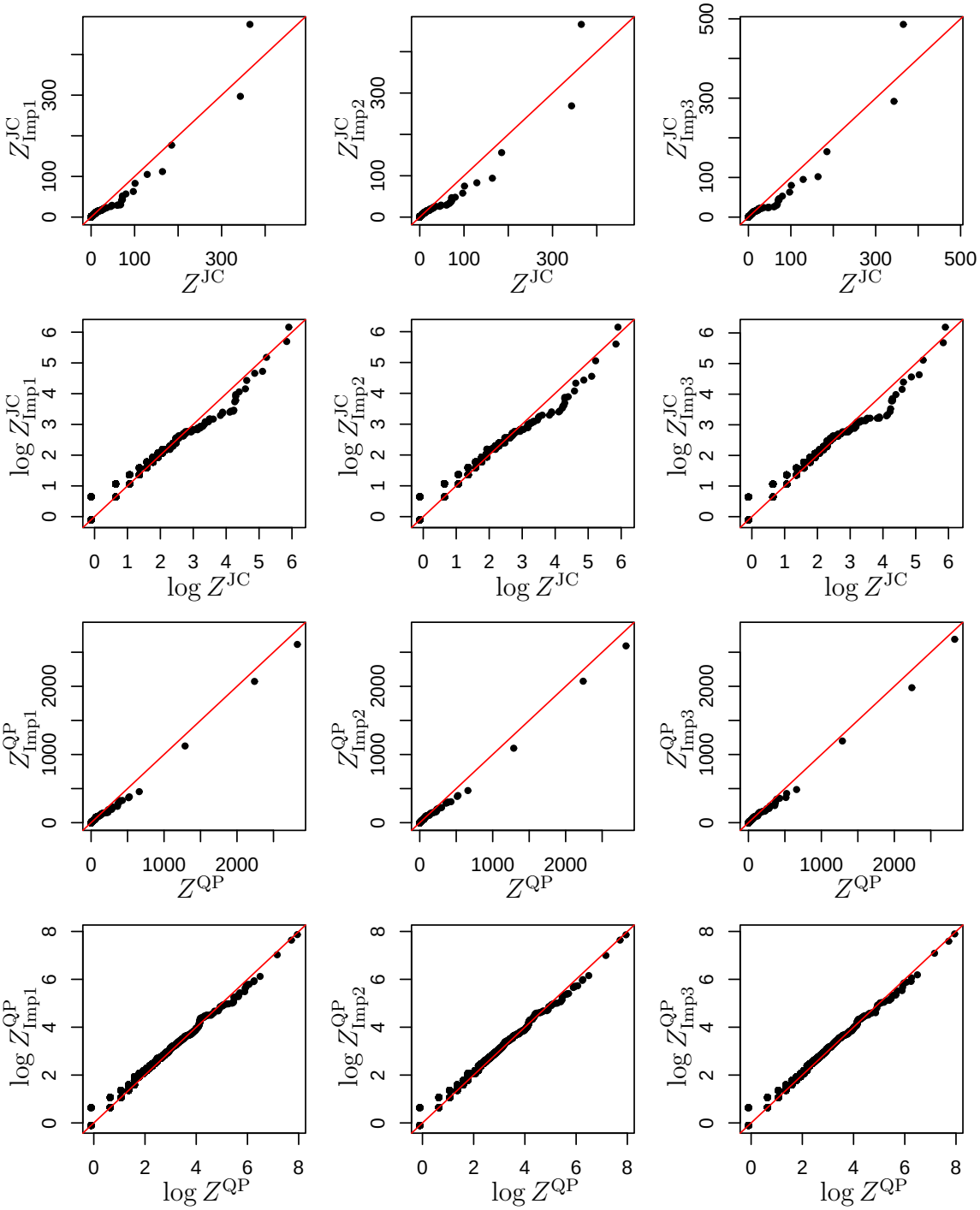

Figure 4.6: Quantile-Quantile Plots of Joint Multiple Imputations and True Unemployment Data Measured in the German State of Barvaria in the Year 2001. The first two rows of the figure are drawn for the Job Creation variable, Z^{JC}, and its log transformation, $\log Z^{JC}$, respectively. The last two rows correspond to the Qualification Program variable, Z^{QP}, and its log transformation, $\log Z^{QP}$, respectively. The bivariate response variable is jointly imputed by using the Poisson regression mode for non-nested data.

4.5 Analysis of German Unemployment Data

As a real data example, we use the German unemployment data measured on the German state of Barvaria. The state of Barvaria consists of 2056 communities, which are contained by both 96 counties and 27 agencies. Each of the counties contains 1 to 58 communities, while 38 to 124 communities are nested in the agencies. (We note that n_i and m_k are large.) To impute the community-level bivariate response variable in Barvaria, we can consider either the bivariate lognormal model or Poisson regression model, eliminating the bivariate Gaussian model because the response variable is non-negative.

In the case of the bivariate lognormal model, we can fit the target model using the MCMC sampler or SIR method. However, these computational methods may not be appropriate to impute the German unemployment data. First, the MCMC sampler constructed for the bivariate lognormal model may not be efficient because each agency or county is composed of so many communities. For example, our observed data consist of 27 agency-level variables, but 2029 (= 2056 – 27) community-level variables are missing and need to be imputed. With the SIR method, the truncated normal model as a working model fails to cover the thick right tail of the target lognormal model, as illustrated in Figure 4.4. Indeed, the distribution of the bivariate response variable in the German data is highly right skewed (see the x-axis of panels in the first and third rows of Figure 4.6), so that the distribution cannot be well approximated by the truncated normal distribution or truncated t distribution. Thus, unless the working model has sufficient thickness in the right tail, our imputation will not be correct.

Instead, we consider fitting the Poisson regression model for non-nested data because this approach is relatively robust to the data with few observations and many to impute. In the year 2001, the bivariate response variable, JC and QP, was measured on both the agencies and communities in Bavaria. A number of descriptive

variables were also measured on the communities, so that we consider them as covariates in the imputation model: the (male and female) population, employed (male and female) population, living employed (male and female) population, size of area, and GNP. Because these descriptive variables tend to be highly correlated with one another, however, they are transformed into a set of five orthogonal variables by using principle components analysis after a log transformation. The resulting principle components serve as covariates in the Poisson regression model for non-nested data. We aim to create ten imputations of the community-level data only using the agency-level response variable, and compare the joint multiple imputations to the true community-level data. By doing so, we are able to see how good our imputations are in the real data situation. In particular, we use the Poisson regression model given in (4.5) and (4.6). That is, the variable Z^{JC} is first imputed, then we impute the variable Z^{QP} given Z^{JC} as an additional covariate. Figure 4.5 shows the quantile-quantile plots comparing the joint multiple imputations with the true data measured in the communities. Because of the large original scale for Z^{JC} and Z^{QP}, we also draw the quantile-quantile plots on a log scale in the second and fourth rows of Figure 4.5. These plots illustrate that our imputations appear to be close to the true data.

The three largest communities in Barvaria in terms of population are outliers in the bivariate response variable (Z^{JC}, Z^{QP}). They are the community of $\mathrm{Städt\ München}$ with $Z^{\mathrm{JC}} = 343$ and $Z^{\mathrm{QP}} = 2827$, the community of $\mathrm{Städt\ Nürnberg}$ with $Z^{\mathrm{JC}} = 365$ and $Z^{\mathrm{QP}} = 2240$, and the community of $\mathrm{Städt\ Augsburg}$ with $Z^{\mathrm{JC}} = 185$ and $Z^{\mathrm{QP}} = 1288$. Excluding these three communities, we randomly select three other communities and compare the true values of Z^{JC} and Z^{QP} with a histogram of the ten imputations. The three randomly selected communities are representative of all the communities excluding $\mathrm{Städte\ München}$, $\mathrm{Nürnberg}$, and $\mathrm{Augsburg}$ in that the multiple imputations cover the true values. Figure 4.7 presents this comparison and illustrates that our multiple imputations seem reasonable except the

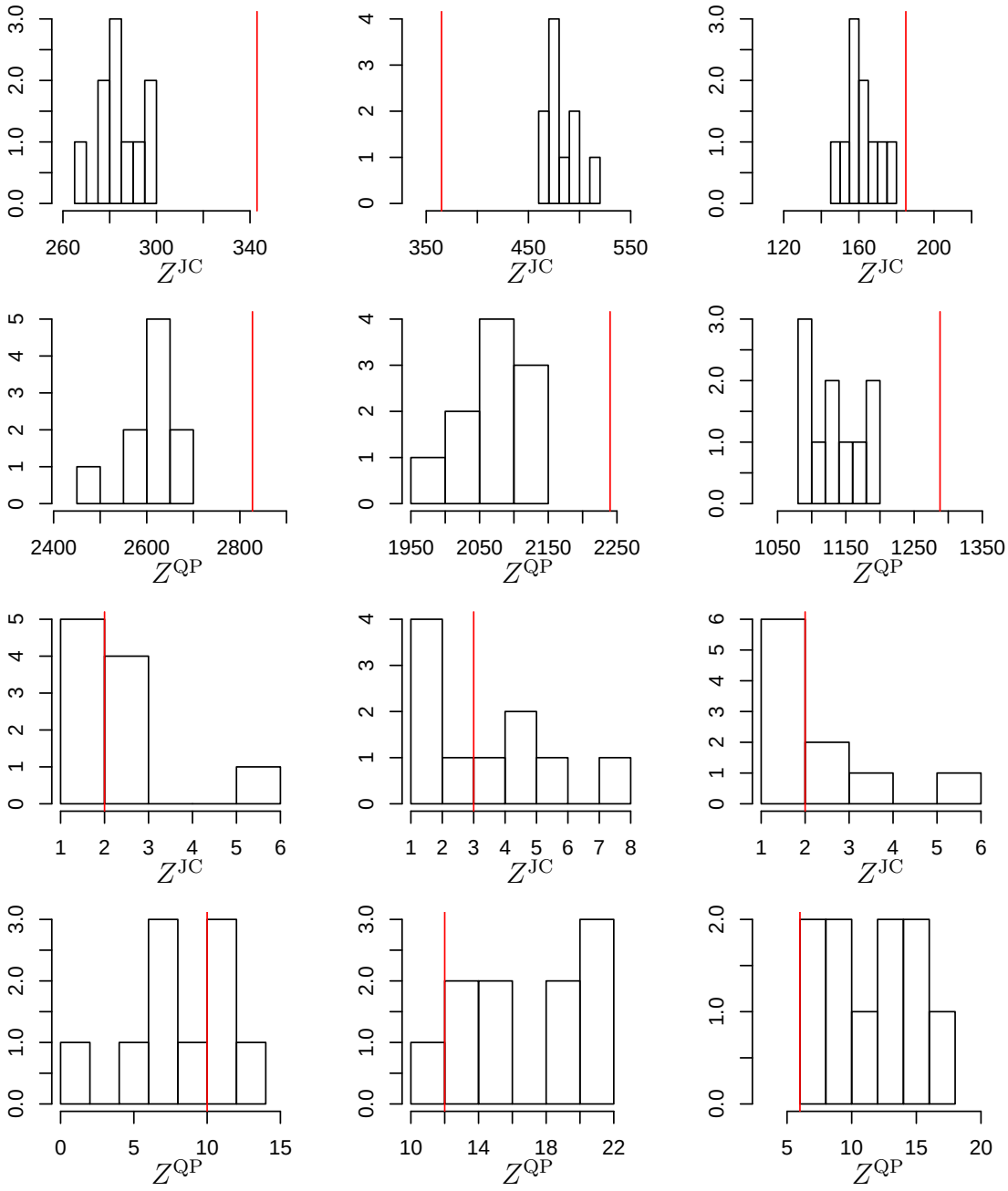

Figure 4.7: Histograms of Multiple Imputations with Indication of True Unemployment Data Measured in the Six Communities of Barvaria in the Year 2001. The first two rows of the figure correspond to the bivariate response variable for the communities with outlier response variables. For the first two rows, the three columns correspond to the communities of Städte München, Nürnberg, and Augsburg, respectively. The last two rows correspond to the multiple imputation of Z^{JC} and Z^{QP} for the randomly selected communities. The vertical solid lines represent the true data measured in the communities.

communities of Städte München and Augsburg for which our imputations always underpredict.

4.6 Concluding Remarks

We present three different joint imputation models considered for a bivariate variable measured on the misaligned geographical partitions. To create joint imputations while properly accounting for the correlation structure of the bivariate response variable, we devise efficient Monte Carlo based algorithms used to fit the joint imputation models. Our simulation studies illustrate the advantages and limitations of the models and the corresponding computational algorithms. Among the joint imputation models, we prefer the Poisson regression model for non-nested data because its underlying Poisson assumptions agree with the real unemployment data, and the computational method to fit the model is robust and easy to implement. We apply our modeling and computational strategies to the real German unemployment data, which illustrates the multiple imputations created from the Poisson regression model are close to the real data measured on the highest level of resolution.

Chapter 5

Computing Hardness Ratios with Poissonian Errors

5.1 Introduction

X-rays are high energy electromagnetic waves, i.e., photons, which are emitted by an astronomical source. The detector aboard the space-based *Chandra X-ray Observatory* measures the energy, sky coordinates, and arrival time for each emanating photon that arrives at the detector; because of the digital nature of the instrument, each of these measurements is necessarily discrete. In particular, the distribution of the photons as a function of their energies, called a spectrum, is of main interest because the shape of the energy spectrum of an astronomical source is highly informative as to the physical processes at the surface of the source. The insufficient number of photons detected for a faint X-ray source, however, makes any sophisticated spectral analysis infeasible and generally leads us to consider an alternative description for the spectrum. That is, a *hardness ratio*, which is based on aggregate photon counts in two sub-energy bands, becomes a useful tool to quantify and

characterize the source spectrum. When the energy is split into two sub-energy bands, the lower energy end of the X-ray spectrum is called the *soft band* and the higher energy end is called the *hard band*. Based on the soft and hard counts that are aggregate photons detected in the soft and hard bands, respectively, a hardness ratio is defined as either the ratio of the soft counts and hard counts or a monotone function of the ratio; the choice of which to use is determined by the astronomical field of application. We consider three types of a hardness ratio, i.e.,

$$\begin{aligned} \text{the simple counts ratio,} \quad & \mathrm{R} = \frac{S}{H}, \\ \text{the X-ray color,} \quad & \mathrm{C} = \log_{10} \frac{S}{H}, \text{ and} \\ \text{the fractional difference hardness ratio,} \quad & \mathrm{HR} = \frac{H-S}{H+S}, \end{aligned} \tag{5.1}$$

where S and H represent the photon counts in the soft and hard bands, respectively. These classical definitions of the hardness ratio are simply based on the observed source counts and fail to account for the underlying Poissonian nature of the counts especially when we study faint X-ray sources. Here, we present Bayesian methods that correctly deal with the non-Gaussian nature of low count data and asymptotically agree with (5.1) under Gaussian assumptions for high counts data.

Advanced X-ray instruments such as the *Chandra X-ray Observatory* allow us to detect more faint sources with low counts. For such low counts data, hardness ratios are typically used to extract spectral properties of X-ray sources, although a hardness ratio is the coarsest description of a spectrum. For example, a low value of R (equivalently, a low value of C and a high value of HR) implies that we expect relatively more hard counts than soft counts. In this case, the source is more likely to be classified into a hard one that has high temperature generated in flares. The spectral shape of an X-ray source can be also inferred by comparing the hardness ratio with its theoretical values under the power-law or thermal models.

In the context of *Chandra* data, three sub-energy bands are often used, called the soft band (0.3 – 0.9 keV), the medium band (0.9 – 2.5 keV), and the hard band (2.5 – 8.0 keV). In this case, a source spectrum can be described by using two hardness ratios based on the soft and medium counts or the medium and hard counts, and drawing a so-called X-ray color-color diagram. Sometimes, the soft and medium bands are also merged into a single band, yielding another soft band (0.3 – 2.5 keV). Thus, the number of sub-energy bands is flexible and the choice of energy ranges is not particularly confined.

The remainder of this chapter is organized into six sections. The chapter begins in Section 5.2 with recapitulating the classical method of computing hardness ratios with background contamination under Gaussian assumptions. Section 5.3 models the observed counts with Poissonian errors and redefine the hardness ratios by using parameters. In Section 5.4, we describe prior specification and introduce new Bayesian methods corresponding to the Poisson model. Section 5.5 compares the classical method with the new Bayesian approach through a simulation study. Section 5.6 outlines various applications of the hardness ratios computed by the Bayesian methods. Discussion and future work follow in Section 5.7.

5.2 The Classical Method

The conventional hardness ratios are computed as (5.1) when no background sources are present in a source area. In the presence of some X-ray sources other than the one of interest, the detected photons are subject to background contamination. To quantify the background contamination, we take another observation around, but some distance away from the source of interest in a region of space that contains no apparent X-ray source. As illustrated in Figure 5.1, source counts are obtained in a source area represented by the smaller circle and background counts are collected in the annulus around the source area to account for back-

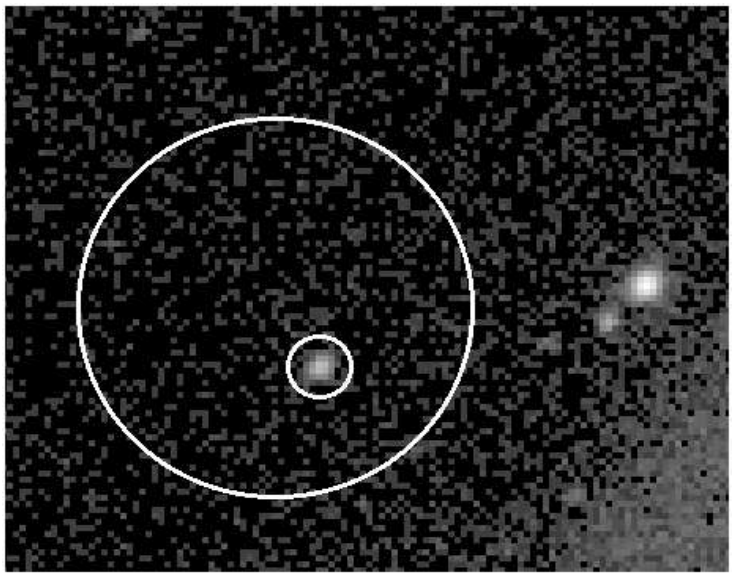

Figure 5.1: Illustration of How to Collect Photon Counts in the Sky. The smaller circle represents a source area where the soft and hard source counts are detected, and background counts are recorded in the annulus around the smaller circle. The brightness of each pixel corresponds to the total number of photon counts detected in the pixel.

ground contamination in the source area. Each detected photon is classified into either the soft or hard count according to its own energy. In addition to the difference in the exposure area of source and background observations, we account for the difference in the exposure time because more photon counts are expected to be observed with longer exposure time. The overall difference in the exposure area and exposure time is summarized and denoted by a known constant r. With the background counts in the soft band (B_S) and the hard band (B_H) collected in a background area that is r times the source area, the conventional hardness ratio is generalized to

$$\begin{aligned} \text{the simple counts ratio,} \quad & \mathrm{R} = \frac{S - B_S/r}{H - B_H/r}, \\ \text{the X-ray color,} \quad & \mathrm{C} = \log_{10}\left(\frac{S - B_S/r}{H - B_H/r}\right), \text{ and} \\ \text{the fractional difference hardness ratio,} \quad & \mathrm{HR} = \frac{(H - B_H/r) - (S - B_S/r)}{(H - B_H/r) + (S - B_S/r)}. \end{aligned} \tag{5.2}$$

In words, the background counts adjusted for the the difference in the exposure area and exposure time are *directly* subtracted from the counts observed in the source area, simply based on the method of moments. The uncertainty of the hardness ratios is propagated under Gaussian assumptions, i.e., the delta method is used to compute their standard errors such as

$$\begin{aligned}
\sigma_{\mathrm{R}} &= \frac{S}{H}\sqrt{\frac{\sigma_S^2+\sigma_{B_S}^2/r^2}{(S-B_S/r)^2}+\frac{\sigma_H^2+\sigma_{B_H}^2/r^2}{(H-B_H/r)^2}}, \\
\sigma_{\mathrm{C}} &= \frac{1}{\ln(10)}\sqrt{\frac{\sigma_S^2+\sigma_{B_S}^2/r^2}{(S-B_S/r)^2}+\frac{\sigma_H^2+\sigma_{B_H}^2/r^2}{(H-B_H/r)^2}}, \text{ and} \\
\sigma_{\mathrm{HR}} &= \frac{2\sqrt{(H-B_H/r)^2\big(\sigma_S^2+\sigma_{B_S}^2/r^2\big)+(S-B_S/r)^2\big(\sigma_H^2+\sigma_{B_H}^2/r^2\big)}}{\big[(H-B_H/r)+(S-B_S/r)\big]^2},
\end{aligned} \tag{5.3}$$

where σ_S, σ_H, σ_{B_S}, and σ_{B_H} are the standard errors of S, H, B_S, and B_H, respectively. These errors are typically approximated by the Gehrels prescription (Gehrels, 1986),

$$\sigma_X \approx \sqrt{X+0.75}+1, \tag{5.4}$$

where X is one of S, H, B_S, and B_H: One standard error in (5.4) away from the observed counts approximates the 16^{th} and 84^{th} percentiles for the Poisson counts, which corresponds to a 68% error bar under the Gaussian assumptions.

5.3 Modeling the Hardness Ratios

Hardness ratios are often used to describe faint X-ray sources with very low counts; it is not uncommon for either or both of the soft and hard counts to be zero. The classical method presented in Section 5.2 generally fails to account for the asymmetric nature of the Poisson counts from such low intensity sources and is not even applicable when the background subtraction results in a negative estimate for the source counts. Thus, instead of the Gaussian assumptions on the detected photons, we directly model photon arrivals as an inhomogeneous Poisson process.

The observed counts in a source area are a convolution of the source counts (η) and the background counts (β), hence we can write $S = \eta_S + \beta_S$ and $H = \eta_H + \beta_H$. (Note that Greek letters indicate unobserved quantities and Roman letters observed ones throughout this chapter, except for the hardness ratios, i.e., R, C, and HR.) We model the source counts as independent Poisson random variables,

$$\eta_S \sim \text{Poisson}(\lambda_S) \quad \text{and} \quad \eta_H \sim \text{Poisson}(\lambda_H), \tag{5.5}$$

where λ denotes the expected source counts in the source exposure area, and the background counts in a source area are modeled as independent Poisson random variables,

$$\beta_S \sim \text{Poisson}(\xi_S) \quad \text{and} \quad \beta_H \sim \text{Poisson}(\xi_H), \tag{5.6}$$

where ξ denotes the expected background counts in the source exposure area. Thus we can simply write $S \sim \text{Poisson}(\lambda_S + \xi_S)$ and $H \sim \text{Poisson}(\lambda_H + \xi_H)$ because the sum of two independent Poisson random variables is a Poisson random variable with the sum of two Poisson intensities. The observed background counts collected in the annulus around an X-ray source are modeled as independent Poisson random variables,

$$B_S \sim \text{Poisson}(r \cdot \xi_S) \quad \text{and} \quad B_H \sim \text{Poisson}(r \cdot \xi_H), \tag{5.7}$$

where the intensity ξ is scaled by the known correction factor r that accounts for the difference in source and background areas. Given the source intensities (λ_S and λ_H), it is legitimate for the hardness ratios to be rewritten as

$$\begin{aligned} \text{the simple counts ratio,} \qquad & \text{R} = \frac{\lambda_S}{\lambda_H}, \\ \text{the X-ray color,} \qquad & \text{C} = \log_{10} \frac{\lambda_S}{\lambda_H}, \text{ and} \\ \text{the fractional difference hardness ratio,} \qquad & \text{HR} = \frac{\lambda_H - \lambda_S}{\lambda_H + \lambda_S}, \end{aligned} \tag{5.8}$$

which are parameter-driven rather than data-driven as in (5.2). As such, the rewritten hardness ratios are of more direct scientific interest. In particular, the classical hardness ratios in (5.2) are simply the method-of-moments estimates of the rewritten hardness ratios in (5.8). Also notice that the background contamination is explicitly taken into account via direct modeling, as shown in (5.6) and (5.7), so that we implicitly eliminate the case of negative source counts sometimes resulting from the background subtraction in (5.2).

5.4 Bayesian Approach

5.4.1 A Bayesian Model for a Single Source

With the underlying Poisson likelihood functions in (5.5) and (5.6), we assign independent conjugate gamma prior distributions for both source and background intensities. That is, we assign independent gamma prior distributions for λ and ξ,

$$\lambda_S \sim \text{Gamma}\big(\psi_{S_1}, \psi_{S_2}\big) \quad \text{and} \quad \lambda_H \sim \text{Gamma}\big(\psi_{H_1}, \psi_{H_2}\big) \tag{5.9}$$

$$\xi_S \sim \text{Gamma}\big(\psi_{S_3}, \psi_{S_4}\big) \quad \text{and} \quad \xi_H \sim \text{Gamma}\big(\psi_{H_3}, \psi_{H_4}\big) \tag{5.10}$$

where $\mu \sim \text{Gamma}(\alpha, \beta)$ if $p(\mu) \propto \mu^{\alpha-1} \exp(-\beta\mu)$ for $\alpha > 0$ and $\beta > 0$, and the values of ψ are calibrated according to our prior knowledge (and uncertainty) about the parameters; see Section 5.4.3 for the discussion of choosing a prior distribution. Here, we assume that the soft and hard intensities are a priori independent because we cannot compute the correlation between λ_S and λ_H with a single source. With a survey of sources, however, we relax the independence assumption and devise a hierarchical mixture model for X-ray sources; refer to Section 5.4.2.

Monte Carlo Integration: Gibbs Sampling

As discussed in Section 5.3, the detected photons in the source region are a convolution of the source and background counts, i.e., $S = \beta_S + \eta_S$ and $H = \beta_H + \eta_H$. In the Bayesian missing data model, we treat the background counts in the source exposure area (β_S and β_H) as missing data; the net source counts (η_S and η_H) are fully determined when β_S and β_H are known. Intuitively, it is straightforward to estimate the Poisson intensities if detected photons are split into the source and background counts. Based on this setting, the Gibbs sampler produces Monte Carlo draws of λ_S and λ_H along with stochastically imputing the missing data (β_S and β_H). Due to the conditional independence we assume, both soft and hard bands also have exactly the same sampling steps, hence we illustrate the Gibbs sampler only for the soft band. First, the joint posterior distribution of λ_S, ξ_S, and β_S is given by

$$\begin{aligned} p(\lambda_S, \xi_S, \beta_S | S, B_S) &\propto p(S|\lambda_S, \beta_S) p(B_S|\xi_S) p(\beta_S|\xi_S) p(\lambda_S) p(\xi_S) \\ &\propto \frac{1}{(S-\beta_S)!\beta_S!} \lambda^{S-\beta_S+\psi_{S_1}-1} \xi^{B+\beta_S+\psi_{S_3}-1} \\ &\quad \exp\Big(-(1+\psi_{S_2})\lambda_S - (1+c+\psi_{S_4})\xi_S \Big). \end{aligned} \tag{5.11}$$

That is, conditional on the total soft counts (S), the unobserved background counts in the source exposure area (β_S) follows a binomial distribution: Given the current iterates of the parameters, $\lambda_S^{(t)}$ and $\xi_S^{(t)}$, STEP 1 is given by

STEP 1 : Draw $\beta_S^{(t+1)}$ from $p(\beta_S|\lambda_S^{(t)}, \xi_S^{(t)}, S, B_S) = \text{Binomial}\left(S,\ \dfrac{\xi_S^{(t)}}{\lambda_S^{(t)} + \xi_S^{(t)}}\right)$,

where the binomial probability is the relative magnitude of the source intensity and the combined intensity. Next, STEPS 2 and 3 draw the source and background intensities from the gamma distributions. In particular, STEPS 2 and 3 find the next iterates of the intensities using

STEP 2 : Draw $\lambda_S^{(t+1)}$ from $p(\lambda_S|\xi_S^{(t)}, \beta_S^{(t+1)}, S, B_S)$

$$= \text{Gamma}\big(S - \beta_S^{(t+1)} + \psi_{S_1},\ 1 + \psi_{S_2}\big)$$

STEP 3 : Draw $\xi_S^{(t+1)}$ from $p(\xi_S|\lambda_S^{(t+1)}, \beta_S^{(t+1)}, S, B_S)$

$$= \text{Gamma}\big(B_S + \beta_S^{(t+1)} + \psi_{S_3},\ 1 + r + \psi_{S_4}\big).$$

To accomplish one Gibbs iteration, these steps are implemented for the soft and hard bands. After iterating the Gibbs sampler T times, we collect a posterior sample $\{\lambda_S^{(t)},\ \lambda_H^{(t)},\ t = t_0 + 1, \ldots, T\}$ for a sufficiently long burn-in period t_0. The analytical calculation to determine burn-in is far from computationally feasible in most situations. However, visual inspection of plots of the Monte Carlo output is commonly used for determining burn-in. More formal tools for determining t_0, called convergence diagnostics, have been proposed; for a recent review, see Cowles and Carlin (1996). Under a monotone transformation of the posterior samples, $(T - t_0)$ Monte Carlo draws for each hardness ratio are obtained, which enables us to find its point estimates and the corresponding error bar. Because the Monte Carlo draws are free of transformation, the posterior distribution of each type of hardness ratio can be computed by transforming the Monte Carlo draws of λ_S and λ_H, according to (5.8).

Numerical Integration: Gaussian Quadrature

By analytically obtaining the marginal posterior distribution of the source intensity, we can more precisely compute the posterior distribution of each hardness ratio. Because the models for the hard and soft bands are symmetric, we again illustrate the computation only for the soft source intensity λ_S. To begin with, we write the joint posterior distribution of λ_S and ξ_S as

$$\begin{aligned} p(\lambda_S, \xi_S|S, B_S) &= \frac{p(\lambda_S)p(\xi_S)p(S|\lambda_S, \xi_S)p(B_S|\xi_S)}{\int_0^\infty\int_0^\infty p(\lambda_S)p(\xi_S)p(S|\lambda_S, \xi_S)p(B_S|\xi_S)d\xi_S\, d\lambda_S} \\ &= \frac{(\lambda_S + \xi_S)^S \lambda_S^{\psi_{S_1}-1} \xi_S^{B_S+\psi_{S_3}-1} e^{-(1+\psi_{S_2})\lambda_S - (1+r+\psi_{S_4})\xi_S}}{\int_0^\infty\int_0^\infty (\lambda_S + \xi_S)^S \lambda_S^{\psi_{S_1}-1} \xi_S^{B_S+\psi_{S_3}-1} e^{-(1+\psi_{S_2})\lambda_S - (1+r+\psi_{S_4})\xi_S} d\xi_S d\lambda_S}. \end{aligned} \quad (5.12)$$

Then, the binomial expansion to $(\lambda_S + \xi_S)^S$, i.e.,

$$(\lambda_S + \xi_S)^S = \sum_{j=0}^{S} \frac{\Gamma(S+1)}{\Gamma(j+1)\Gamma(S-j+1)} \lambda_S^j \, \xi_S^{S-j}, \tag{5.13}$$

enables us to analytically obtain the marginal posterior distribution of λ_S by integrating ξ_S out of the joint distribution in (5.12). Using the binomial expansion and integrating ξ_S out of the joint posterior distribution in (5.12), the marginal posterior distribution of λ_S is computed as

$$p(\lambda_S|S, B_S) = \frac{\displaystyle\sum_{j=0}^{S} \frac{\Gamma(S-j+B_S+\psi_{S_3})}{\Gamma(j+1)\Gamma(S-j+1)(1+r+\psi_{S_4})^{S-j+B_S+\psi_{S_3}}} \lambda_S^{j+\psi_{S_1}-1} e^{-(1+\psi_{S_2})\lambda_S}}{\displaystyle\sum_{j=0}^{S} \frac{\Gamma(S-j+B_S+\psi_{S_3})}{\Gamma(j+1)\Gamma(S-j+1)(1+r+\psi_{S_4})^{S-j+B_S+\psi_{S_3}}} \cdot \frac{\Gamma(j+\psi_{S_1})}{(1+\psi_{S_2})^{j+\psi_{S_1}}}}. \tag{5.14}$$

Here a priori independence of λ_S and λ_H decomposes the joint posterior distribution of these two intensities into a product of their marginal posterior distributions, i.e.,

$$p(\lambda_S, \lambda_H|S, H, B_S, B_H) = p(\lambda_S|S, B_S)p(\lambda_H|H, B_H). \tag{5.15}$$

Based on the joint posterior distribution of λ_S and λ_H in (5.15), we compute the posterior distribution of each hardness ratio as follows: The posterior distribution of R is obtained by integrating λ_H out of $p(\mathrm{R}, \lambda_H|S, H, B_S, B_H)$, i.e.,

$$\begin{aligned} p(\mathrm{R}|S, H, B_S, B_H) &= \int p(\mathrm{R}, \lambda_H|S, H, B_S, B_H) d\lambda_H \\ &= \int p(\lambda_S, \lambda_H|S, H, B_S, B_H) \left| \frac{\partial(\lambda_S, \lambda_H)}{\partial(\mathrm{R}, \lambda_H)} \right| d\lambda_H \\ &= \int p(\mathrm{R}\lambda_H, \lambda_H|S, H, B_S, B_H) \lambda_H \, d\lambda_H, \end{aligned} \tag{5.16}$$

where λ_S is substituted with $\mathrm{R}\lambda_H$ in (5.16); the posterior distribution of C is obtained by integrating λ_H out of $p(\mathrm{C}, \lambda_H|S, H, B_S, B_H)$, i.e.,

$$\begin{aligned} p(\mathrm{C}|S, H, B_S, B_H) &= \int p(\mathrm{C}, \lambda_H|S, H, B_S, B_H) d\lambda_H \\ &= \int p(\lambda_S, \lambda_H|S, H, B_S, B_H) \left| \frac{\partial(\lambda_S, \lambda_H)}{\partial(\mathrm{C}, \lambda_H)} \right| d\lambda_H \\ &= \int p(10^{\mathrm{C}}\lambda_H, \lambda_H|S, H, B_S, B_H) 10^{\mathrm{C}} \ln(10) \lambda_H \, d\lambda_H, \end{aligned} \tag{5.17}$$

where λ_S is substituted with $10^{\mathrm{C}}\lambda_H$ in (5.17); and the posterior distribution of HR is obtained by integrating $\omega = \lambda_S + \lambda_H$ out of $p(\mathrm{HR}, \omega|S, H, B_S, B_H)$, i.e.,

$$\begin{aligned} p(\mathrm{HR}|S, H, B_S, B_H) &= \int p(\mathrm{HR}, \omega|S, H, B_S, B_H) d\omega \\ &= \int p(\lambda_S, \lambda_H|S, H, B_S, B_H) \left| \frac{\partial(\lambda_S, \lambda_H)}{\partial(\mathrm{HR}, \omega)} \right| d\omega \\ &= \int p\left(\frac{(1-\mathrm{HR})\omega}{2}, \frac{(1+\mathrm{HR})\omega}{2} \Big| S, H, B_S, B_H \right) \frac{\omega}{2}\, d\omega, \end{aligned} \tag{5.18}$$

where λ_S and λ_H are substituted with $(1-\mathrm{HR})\omega/2$ and $(1+\mathrm{HR})\omega/2$ in (5.18). Computing the marginal posterior distribution of each hardness ratio involves with integrating over a nuisance parameter. We thus employ Gaussian quadrature to precisely evaluate the marginal posterior distribution via numerical integration; refer to Wichura (1989) for details of the computational technique. To approximate the distribution, we treat each hardness ratio as a discrete variable and evaluate its marginal posterior distribution at each of abscissas equally spaced over the finite range of the hardness ratio. Our inferences are based on the approximate posterior distributions.

5.4.2 A Bayesian Hierarchical Model for Clustering

With a survey of X-ray sources, we are interested in the relationship between hardness ratios across different sources, which can be used to cluster the X-ray sources. Thus, we relax the independence assumption and devise a hierarchical mixture model. With a slight modification for the likelihood functions in (5.5), (5.6), and (5.7), i.e.,

$$S_i = \eta_{S,i} + \beta_{S,i} \sim \mathrm{Poisson}(\lambda_{S,i} + \xi_{S,i}), \tag{5.19}$$

$$H_i = \eta_{H,i} + \beta_{H,i} \sim \mathrm{Poisson}(\lambda_{H,i} + \xi_{H,i}), \tag{5.20}$$

$$B_{S,i} \sim \mathrm{Poisson}(r_i \cdot \xi_{S,i}), \text{ and} \tag{5.21}$$

$$B_{H,i} \sim \mathrm{Poisson}(r_i \cdot \xi_{H,i}), \tag{5.22}$$

we assign bivariate lognormal prior distributions to both λ and ξ. In particular, we assume two mixture components on λ but a single component on ξ, i.e.,

$$\begin{pmatrix} \log_{10}\lambda_{S,i} \\ \log_{10}\lambda_{H,i} \end{pmatrix} \Big| \, \mu_\lambda, \Sigma_\lambda, \alpha \;\sim\; (1-\alpha)\cdot \mathrm{N}_2\big(\mu_{\lambda,0},\, \Sigma_{\lambda,0}\big) + \alpha \cdot \mathrm{N}_2\big(\mu_{\lambda,1},\, \Sigma_{\lambda,1}\big) \text{ and} \tag{5.23}$$

$$\begin{pmatrix} \log_{10}\xi_{S,i} \\ \log_{10}\xi_{H,i} \end{pmatrix} \Big| \, \mu_\xi, \Sigma_\xi \;\sim\; \mathrm{N}_2\big(\mu_\xi,\, \Sigma_\xi\big), \tag{5.24}$$

where μ_{λ,δ_i} is a 2×1 vector of means of source intensities for component δ_i, $\Sigma_{\lambda,\delta_i}$ is a 2×2 positive definite covariance matrix of source intensities for component δ_i, δ_i is a mixture components indicator variable for source i with $\delta_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\alpha)$, μ_ξ is a 2×1 vector of means of background intensities, Σ_ξ is a 2×2 positive definite covariance matrix of background intensities, and $i = 1, 2, \ldots, n$ with n the number of X-ray sources in a survey. For the hyper-parameters, the conjugate prior distributions are assigned as

$$\Sigma_{\lambda,\delta_i} \;\sim\; \mathrm{Inv{-}Wishart}(d_0, D_0), \tag{5.25}$$

$$\mu_{\lambda,\delta_i} \,|\, \Sigma_{\lambda,\delta_i} \;\sim\; \mathrm{N}_2\big(a_0, \Sigma_{\lambda,\delta_i}/b_0\big), \tag{5.26}$$

$$\Sigma_\xi \;\sim\; \mathrm{Inv{-}Wishart}(d_1, D_1), \tag{5.27}$$

$$\mu_\xi \,|\, \Sigma_\xi \;\sim\; \mathrm{N}_2\big(a_1, \Sigma_\xi/b_1\big), \text{ and} \tag{5.28}$$

$$\alpha \;\sim\; \mathrm{Beta}(c_0, c_1), \tag{5.29}$$

where a, b, c, d, and D are fixed constants.

Gibbs Sampling with Metropolis-Hastings Steps

With the hierarchical mixture model, we treat both the background counts in the source exposure area, β, and the mixture components indicator variables, δ. Then based on the joint posterior distribution, $p(\beta, \delta, \lambda, \xi, \mu, \Sigma, \alpha | Y_{\text{obs}})$, a Gibbs sampler is constructed as follows. Given the current iterates of the parameters $\theta^{(t)} = (\lambda^{(t)}, \xi^{(t)}, \mu^{(t)}, \Sigma^{(t)}, \alpha^{(t)})$, STEP 1 draws the net background counts in the source exposure area for source i via

STEP 1.1 : Draw $\beta_S^{(t+1)}$ from $p(\beta_{S,i}|\theta^{(t)}, S, B_S) = \text{Binomial}\left(S_i,\ \dfrac{\xi_{S,i}^{(t)}}{\lambda_{S,i}^{(t)} + \xi_{S,i}^{(t)}}\right)$, and

STEP 1.2 : Draw $\beta_H^{(t+1)}$ from $p(\beta_{H,i}|\theta^{(t)}, H, B_H) = \text{Binomial}\left(H_i,\ \dfrac{\xi_{H,i}^{(t)}}{\lambda_{H,i}^{(t)} + \xi_{H,i}^{(t)}}\right)$.

Next, STEP 2 updates the mixture indicator variables via Bernoulli distributions

STEP 2 : Draw $\delta_i^{(t+1)}$ from $p(\delta_i|\theta^{(t)})$

$$= \text{Bernoulli}\left(\frac{\alpha \text{N}_2\big(\log_{10}\lambda_i^{(t)}; \mu_{\lambda,1}^{(t)}, \Sigma_{\lambda,1}^{(t)}\big)\pi_1}{(1-\alpha)\text{N}_2\big(\log_{10}\lambda_i^{(t)}; \mu_{\lambda,0}^{(t)}, \Sigma_{\lambda,0}^{(t)}\big)\pi_0 + \alpha \text{N}_2\big(\log_{10}\lambda_i^{(t)}; \mu_{\lambda,1}^{(t)}, \Sigma_{\lambda,1}^{(t)}\big)\pi_1}\right),$$

where $\text{N}_2(\ \cdot\ ; \mu, \Sigma)$ denotes a bivariate Gaussian density function with mean μ and covariance matrix Σ, and π_k is the prior distribution of $(\mu_{\lambda,k}, \Sigma_{\lambda,k})$ in (5.26) and (5.27), i.e., $\pi_k = p(\mu_{\lambda,k}, \Sigma_{\lambda,k})$ for $k = 0, 1$. Then, STEP 3 jointly draws the mean vector and covariance matrix for component δ_i. Because the models for two components are symmetric, we illustrate the sampling step only for component 1:

STEP 3.1 : Draw $\Sigma_{\lambda,1}^{(t+1)}$ from $p\Big(\Sigma_{\lambda,1}\big|\delta_i^{(t+1)}, \lambda^{(t)}\Big)$

$$= \text{Inv}-\text{Wishart}\Big(d_0 + \textstyle\sum_i \delta_i^{(t+1)}, \Upsilon_{\lambda,1}\Big),$$

where $\Upsilon_{\lambda,1}$ is computed as

$$\Upsilon_{\lambda,1} = D_0 + \sum_{i=1}^{n} \delta_i\big(\log_{10}\lambda_i - \widetilde{\lambda}\big)\big(\log_{10}\lambda_i - \widetilde{\lambda}\big)^\top + \frac{b_0 \sum_i \delta_i}{b_0 + \sum_i \delta_i}\big(\widetilde{\lambda} - a_0\big)\big(\widetilde{\lambda} - a_0\big)^\top,$$

$\lambda_i = (\lambda_{S,i}\ \lambda_{H,i})^\top$, and $\widetilde{\lambda} = (\sum_i \delta_i \log\lambda_i)/\sum_i \delta_i$. Then, we draw the next iterate of the the mean vector $\mu_{\lambda,1}$ given $\Sigma_{\lambda,1}^{(t+1)}$, i.e.,

STEP 3.2 : Draw $\mu_{\lambda,1}^{(t+1)}$ from $p\Big(\mu_{\lambda,1}\big|\delta_i^{(t+1)}, \Sigma_{\lambda,1}^{(t+1)}, \lambda^{(t)}\Big)$

$$= \text{N}_2\Big(\widetilde{\mu}_{\lambda,1}, \Sigma_{\lambda,1}^{(t+1)}/(b_0 + \textstyle\sum_i \delta_i)\Big),$$

where $\widetilde{\mu}_{\lambda,1}$ is given by

$$\widetilde{\mu}_{\lambda,1} = \frac{b_0 a_0 + \sum_i \delta_i \widetilde{\lambda}}{b_0 + \sum_i \delta_i}.$$

In STEP 4, the mean vector and covariance matrix of background intensities are drawn in the same fashion as STEP 3. Namely, we first draw Σ_ξ, then draw μ_ξ given Σ_ξ:

STEP 4.1 : Draw $\Sigma_\xi^{(t+1)}$ from $p\Big(\Sigma_\xi \big| \xi^{(t)}\Big) = \text{Inv}-\text{Wishart}\Big(d_1 + n, \Upsilon_\xi\Big)$,

where Υ_ξ is computed as

$$\Upsilon_\xi = D_1 + \sum_{i=1}^{n} \delta_i \big(\log_{10} \xi_i - \bar{\xi}\big)\big(\log_{10} \xi_i - \bar{\xi}\big)^\top + \frac{b_1 n}{b_1 + n}\big(\bar{\xi} - a_1\big)\big(\bar{\xi} - a_1\big)^\top,$$

$\xi_i = (\xi_{S,i}\ \xi_{H,i})^\top$ and $\bar{\xi} = (\sum_i \log \xi_i)/n$. Then, we draw the next iterate of the mean vector μ_ξ given $\Sigma_\xi^{(t+1)}$, i.e.,

STEP 4.2 : Draw $\mu_\xi^{(t+1)}$ from $p\Big(\mu_\xi \big| \Sigma_\xi^{(t+1)}, \xi^{(t)}\Big) = \text{N}_2\Big(\widetilde{\mu}_\xi, \Sigma_\xi^{(t+1)}/(b_1 + n)\Big)$,

where $\widetilde{\mu}_\xi$ is given by

$$\widetilde{\mu}_\xi = \frac{b_1 a_1 + n\bar{\xi}}{b_1 + n}.$$

STEP 5 updates the mixture proportion α given $\delta_i^{(t+1)}$, i.e.,

STEP 5 : Draw $\alpha^{(t+1)}$ from $p(\alpha|\delta^{(t+1)}) = \text{Beta}\Big(\sum_i \delta_i^{(t+1)} + c_0, \sum_i \big(1 - \delta_i^{(t+1)}\big) + c_1\Big)$.

Lastly, STEP 6 draws the source and background intensities, and Metropolis-Hastings steps (Metropolis and Ulam, 1949; Hastings, 1970) are used in this step because their conditional distributions are not standard distributions. Because the distributions are symmetric, we again illustrate its implementation only for the soft intensities:

STEP 6.1 : Draw $\lambda_{S,i}^{(t+1)}$ from $p\Big(\lambda_{S,i}\big|\beta_S^{(t+1)}, \delta^{(t+1)}, \mu_\lambda^{(t+1)}, \Sigma_\lambda^{(t+1)}, \lambda_H^{(t)}\Big)$

$$\propto \lambda_{S,i}^{S_i-\beta_{S,i}-1} \exp\left(-\lambda_{S,i} - \frac{\Big(\log_{10}\lambda_{S,i} - \mu_{\lambda_S,\delta_i} - \rho_{\lambda,\delta_i}\frac{\sigma_{\lambda_S,\delta_i}}{\sigma_{\lambda_H,\delta_i}}\big(\log_{10}\lambda_{H,i} - \mu_{\mu_H,\delta_i}\big)\Big)^2}{2\big(1-\rho_{\lambda,\delta_i}^2\big)\sigma_{\lambda_S,\delta_i}^2}\right),$$

where the covariance matrix $\Sigma_{\lambda,\delta_i}$ is written as

$$\Sigma_{\lambda,\delta_i} = \begin{pmatrix} \sigma_{\lambda_S,\delta_i}^2 & \rho_{\lambda,\delta_i}\cdot\sigma_{\lambda_S,\delta_i}\cdot\sigma_{\lambda_H,\delta_i} \\ \rho_{\lambda,\delta_i}\cdot\sigma_{\lambda_S,\delta_i}\cdot\sigma_{\lambda_H,\delta_i} & \sigma_{\lambda_H,\delta_i}^2 \end{pmatrix}, \text{ and}$$

STEP 6.2 : Draw $\xi_{S,i}^{(t+1)}$ from $p\Big(\xi_{S,i}\big|\beta_S^{(t+1)}, \mu_\xi^{(t+1)}, \Sigma_\xi^{(t+1)}, \xi_H^{(t)}\Big)$

$$\propto \xi_{S,i}^{B_{S,i}+\beta_{S,i}-1} \exp\left(-(1+r_i)\xi_{S,i} - \frac{\Big(\log_{10}\xi_{S,i} - \mu_{\xi_S} - \rho_\xi\frac{\sigma_{\xi_S}}{\sigma_{\xi_H}}\big(\log_{10}\xi_{H,i} - \mu_{\mu_H}\big)\Big)^2}{2\big(1-\rho_\xi^2\big)\sigma_{\xi_S}^2}\right),$$

where the covariance matrix Σ_ξ is written as

$$\Sigma_\xi = \begin{pmatrix} \sigma_{\xi_S}^2 & \rho_\xi\cdot\sigma_{\xi_S}\cdot\sigma_{\xi_H} \\ \rho_\xi\cdot\sigma_{\xi_S}\cdot\sigma_{\xi_H} & \sigma_{\xi_H}^2 \end{pmatrix}.$$

For the Metropolis-Hastings steps, jumping rules are bivariate Gaussian distributions whose mode and curvature are matched to the target distributions in STEP 6. If the mode of a target distribution is below zero, however, an exponential distribution whose moments are matched given the previous draws is used as the jumping rules. Then $\lambda_S^{(t+1)}$ and $\xi_S^{(t+1)}$ are updated with $\lambda_S^* \sim J_{\lambda_S,t+1}(\lambda_S^*|\lambda_S^{(t)})$ and $\xi_S^* \sim J_{\xi_S,t+1}(\xi_S^*|\xi_S^{(t)})$ with probabilities

$$\frac{p(\lambda_S^*|\theta_{-\lambda_S})/J_{\lambda_S,t+1}(\lambda_S^*|\lambda_S^{(t)})}{p(\lambda_S^{(t)}|\theta_{-\lambda_S})/J_{\lambda_S,t+1}(\lambda_S^{(t)}|\lambda_S^*)} \text{ and } \frac{p(\xi_S^*|\theta_{-\xi_S})/J_{\lambda_S,t+1}(\xi_S^*|\xi_S^{(t)})}{p(\xi_S^{(t)}|\theta_{-\xi_S})/J_{\xi_S,t+1}(\xi_S^{(t)}|\xi_S^*)}, \tag{5.30}$$

respectively, where $\theta_{-\psi}$ denotes model parameters θ other than ψ.

5.4.3 Prior Specification

If there is a strong belief as to the hardness ratio (location or spread), we can incorporate the information as a prior distribution, which is called an informative prior

distribution. The Bayesian method produces the posterior distribution, which may be used as an informative prior distribution for future observation of the same source. In particular, the Gibbs sampler for a single source directly produces the gamma posterior distributions of source and background intensities, hence they can serve as informative prior distributions.

With no prior information available, however, we typically use a flat (or non-informative) prior distribution that minimizes the effect of a prior distribution on posterior inferences. In the Poisson likelihood of a single source, we generally consider three sorts of a flat prior distributions for the positive intensity: when $X|\theta \sim \text{Poisson}(\theta)$,

1. a flat prior distribution on the original scale, i.e.,

$$p(\theta) \propto 1,$$

2. a Jeffrey's flat prior distribution, i.e.,

$$p(\theta) \propto I_\theta^{1/2},$$

3. a flat prior distribution under a log transformation, i.e.,

$$p(\log \theta) \propto 1,$$

where $I_\theta = \text{E}[-\partial^2 \log p(X|\theta)/\partial\theta^2|\theta]$ is the expected Fisher information (Casella and Berger, 1990). In words, the first flat prior distribution is flat between 0 and ∞; the second flat prior distribution is proportional to the square root of the Fisher information; and the third flat prior distribution is flat under a log transformation. The functional forms of these prior distributions are generalized to $p(\theta) \propto \theta^{\phi-1}$, where we call ϕ an index: The first flat prior distribution corresponds to $\phi = 1.0$; the second flat prior distribution corresponds to $\phi = 0.5$; and the third flat prior

distribution corresponds to $\phi = 0.0$. We notice that these three flat prior distributions are all improper, i.e., non-integrable. An improper prior distribution may cause an improper posterior distribution on which no inferences can be made. In our case, as long as ϕ is strictly positive, a posterior distribution is proper. Thus, in the third case, we adopt values of ϕ that are strictly positive but close to 0, e.g., $\phi = 10^{-1}$ that is denoted by 0.0^{+}.

Although these prior distributions are all flat, they slightly differ in specific assumptions based on which they are constructed. Thus, in the case of low counts data, the posterior distribution of a hardness ratio may somewhat vary with the choice of the flat prior distribution; however, we expect the posterior distributions under different flat prior distributions to be almost identical when observed data contain sufficient information (Appendix A). Based on our simulation study, we prefer using the Jeffrey's flat prior distribution with a single source. In the case of the hierarchical mixture model in Section 5.4.2, we also employ diffuse prior distributions that mimic the multivariate Jeffrey's flat prior distributions, i.e.,

$$d_k \to -1,\ |D_k| \to 1,\ b_k \to 0,\ \text{ and }\ c_k \to 0,$$

for $k = 0, 1$ for (5.26), (5.27), (5.28), (5.29), and (5.29).

5.5 Verification

5.5.1 Comparison with the Classical Method

The Gaussian assumptions are inherent in the classical method of computing hardness ratios. However, the assumptions are inappropriate for faint X-ray sources with low counts which hardness ratios are typically used to describe. The detected photons are non-negative integers, so that it is also not valid to model these counts with a Gaussian distribution that is continuous over the real numbers. In this case,

a better strategy, which we adopt, directly models the arrival of photons as an inhomogeneous Poisson process, as discussed in Section 5.3. The error propagation method in (5.3) is an asymptotic result for large samples. If counts are not large enough, the use of the error propagation method cannot be justified. The standard error in (5.4) is also approximated by simulation, which cannot be generalized either.

In the classical method, the background contamination is typically accounted for by directly subtracting the adjusted background counts from the observed source counts; the result is analyzed as if it were a source observation free of background contamination. This procedure is clearly questionable, especially when the detected photons are low. It can lead to the rather embarrassing problem of negative resulting counts, where no statistical inferences can be made. In our Bayesian methods, we model the counts in the source and background observations as independent Poisson random variables, one with the sum of the source and background intensities and the other with the scaled background intensity (van Dyk, 2003), so that no data are discarded due to the negative resulting counts.

5.5.2 Simulation Study

In order to compare the classical method with our Bayesian methods, a simulation study for the case of a single source is designed to calculate frequentist coverage rates for a true value of each hardness ratio. Given the true values of parameters, source and background counts test data are generated and then used to construct 95% intervals of each hardness ratio with the classical and Bayesian methods. We summarize the computed intervals in terms of two statistics, the coverage rate and mean length. The coverage rate is the percentage of the intervals that contain the true value of the hardness ratio, while the mean length is the average of the ranges of the intervals. In addition to these summary statistics, we compute and

Table 5.1: Comparison between the Classical and Bayesian Methods.

	Method	Hardness Ratio	Coverage Rate	Mean Length	Mean Square Error	
					by mode	by mean
CASE I	Classical Method	R	95.0%	1.24	0.065	
		C	98.5%	0.54	0.013	
		HR	99.0%	0.61	0.016	
	Gibbs Sampler	R	94.0%	1.03	0.056	0.069
		C	96.0%	0.44	0.012	0.012
		HR	95.0%	0.49	0.016	0.015
	Gaussian Quadrature	R	94.5%	1.03	0.055	0.069
		C	96.0%	0.43	0.012	0.012
		HR	94.5%	0.49	0.016	0.015
CASE II	Classical Method	R	97.5%	192.44	93.27	
		C	100.0%	6.02	0.27	
		HR	100.0%	3.70	0.21	
	Gibbs Sampler	R	97.0%	8.96	0.317	85.482
		C	99.5%	1.52	0.078	0.113
		HR	95.0%	1.23	0.184	0.083
	Gaussian Quadrature	R	97.0%	8.18	0.394	20.338
		C	99.5%	1.51	0.074	0.112
		HR	95.0%	1.23	0.187	0.083

compare the mean square error of point estimates from the classical and Bayesian methods. The mean square error of the point estimate $\hat{\theta}$ of θ is defined as the sum of the variance and squared bias for an estimator, i.e., $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + [\mathrm{E}(\hat{\theta}) - \theta]^2$. A method that constructs shorter intervals with the similar coverage rate and produces a point estimate with a lower mean square error is generally preferred. The entire simulation was repeated with different magnitudes of the source intensities, λ_S and λ_H. Intrinsically, we are interested in the following two prototypical cases:

CASE I : hardness ratios for high counts, i.e.,

$$\lambda_S = \lambda_H = 30,\ \xi_S = \xi_H = 0.1,\ \text{and}\ r = 100;$$

CASE II : hardness ratios for low counts, i.e.,

$$\lambda_S = \mu_H = 3,\ \xi_S = \xi_H = 0.1,\ \text{and}\ r = 100.$$

This simulation study illustrates two typical cases, i.e., high counts and low counts sources. CASE I represents high counts sources with which Poisson assumptions asymptotically agree with Gaussian assumptions; CASE II represents low counts sources where the Gaussian assumptions are inappropriate.

Table 5.1 presents the results of 200 imaginary sources for each case. The data are used to compute point estimates and 95% intervals by using the classical method, Gibbs sampler, and Gaussian quadrature. The Bayesian methods use flat prior distributions on λ and ξ on an original scale, i.e., $\phi = 1.0$; refer to Section 5.4.3. In CASE I, the posterior distributions of the hardness ratios agree with the corresponding Gaussian approximation of the classical method. The results of CASE II, however, indicate that the Gaussian assumptions clearly fail in the classical method that yields too wide intervals and point estimates with large mean square errors. In particular, the hardness ratio HR is defined between –1 and 1, so that the maximum length of an interval must be 2. However, the mean length of intervals computed by the classical method is 3.70, which is by no means informative. This comparison between the classical and Bayesian methods is illustrates in Figures 5.2 and 5.3. Since the Gaussian quadrature tends to yield more precise intervals, we generally prefer the Gaussian quadrature to the Gibbs sampler. However, because of the summation inside the posterior density in (5.14), the Gaussian quadrature tends to be computationally more expensive as the detected source counts are bigger; on the other hand, the Gibbs sampler is very quick with the reasonable length of a chain, no matter how big the source counts are. Thus, we recommend using the Gaussian quadrature for relatively low counts data and the Gibbs sampler for high counts data.

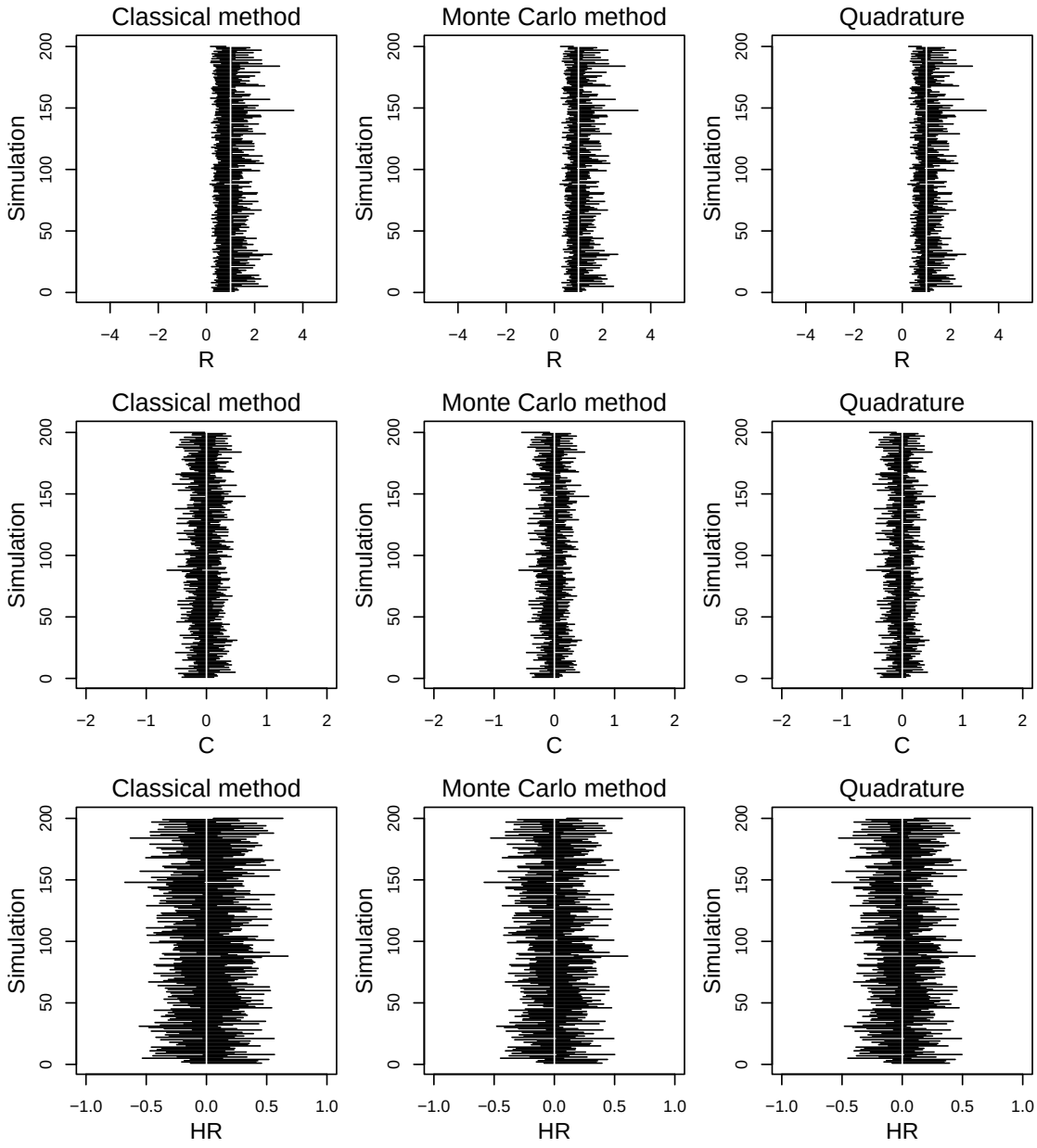

Figure 5.2: Simulation Results of CASE I using the Bayesian and Classical Methods. Three columns correspond to the classical method, Gibbs sampler, and Gaussian quadrature, respectively. The horizontal lines are the 95% intervals computed for each set of test data, and the vertical white lines represent true values of hardness ratios. Notice that all the different methods exhibit similar performance in this case.

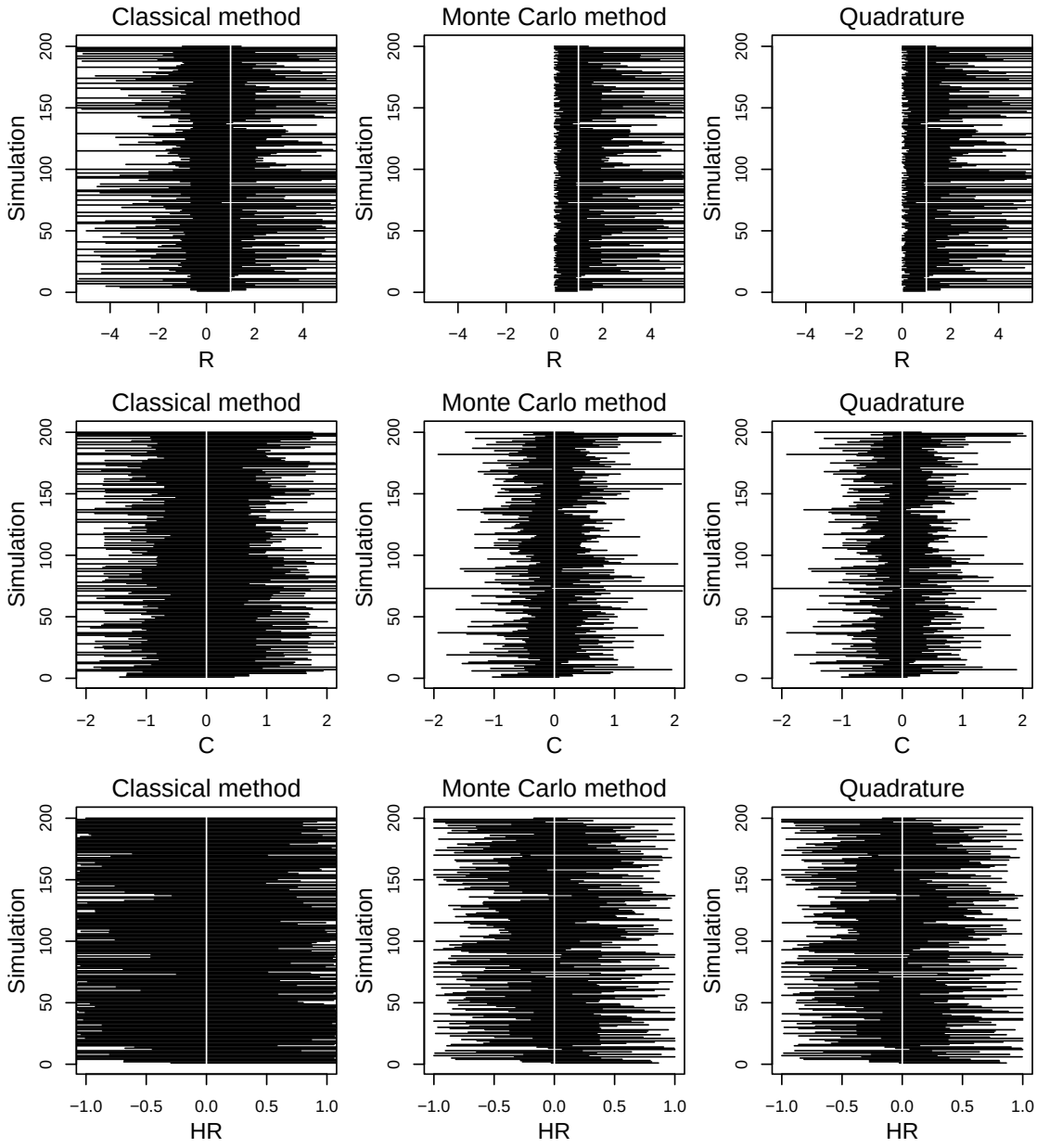

Figure 5.3: Simulation Results of CASE II using the Bayesian and Classical Methods. Three columns correspond to the classical method, Gibbs sampler, and Gaussian quadrature, respectively. The horizontal lines are the 95% intervals computed for each set of test data, and the vertical white lines represent true values of hardness ratios. Notice that, in the case of low counts data, the Bayesian methods dramatically outperform the classical method.

Because the inferences for each hardness ratio are based on its posterior distribution, we consider two point estimates, the posterior mode and posterior mean. The last two columns of Table 5.1 present the comparison of the posterior mean and mode in terms of a mean square error. A posterior mode is the most likely value of a posterior distribution and is invariant to transformation, so that it is expected to represent a posterior distribution better than any other estimates. In the case of HR, however, a posterior mean seems to be a better estimate than a posterior mode because the posterior mean is more robust to the boundary effects that HR has. Thus, a posterior mode is generally preferable for both R and C, but a posterior mean for HR.

5.6 Applications

5.6.1 Characterizing Source Spectra

A color-color diagram is a popular graphical summary for the spectra of faint X-ray sources. We can divide an energy spectrum into three sub-energy bands, i.e., the soft, middle, and hard bands, as described in Section 5.1. Then, the soft X-ray color (C_S) is computed with the soft and middle source intensities, while the hard X-ray color (C_H) is computed with the middle and hard source intensities: $\mathrm{C}_S = \log_{10}(\lambda_S/\lambda_M)$ and $\mathrm{C}_H = \log_{10}(\lambda_M/\lambda_H)$. Then the color-color diagram is simply a scatter plot of the soft and hard colors for different X-ray sources. Figure 5.4 shows the ideal locations of soft and hard X-ray colors for a source spectrum which follows either the power-law or thermal model with specified parameter values. Based on the color-color diagram, we aim to infer the spectrum of a source with uncertainty from the location of its colors on the diagram, to differentiate the spectral shapes of several different sources, and to compare the spectral shape of several difference observations of the same source. Here we illustrate the application of the Bayesian method to extract spectral shape of a faint X-ray source.

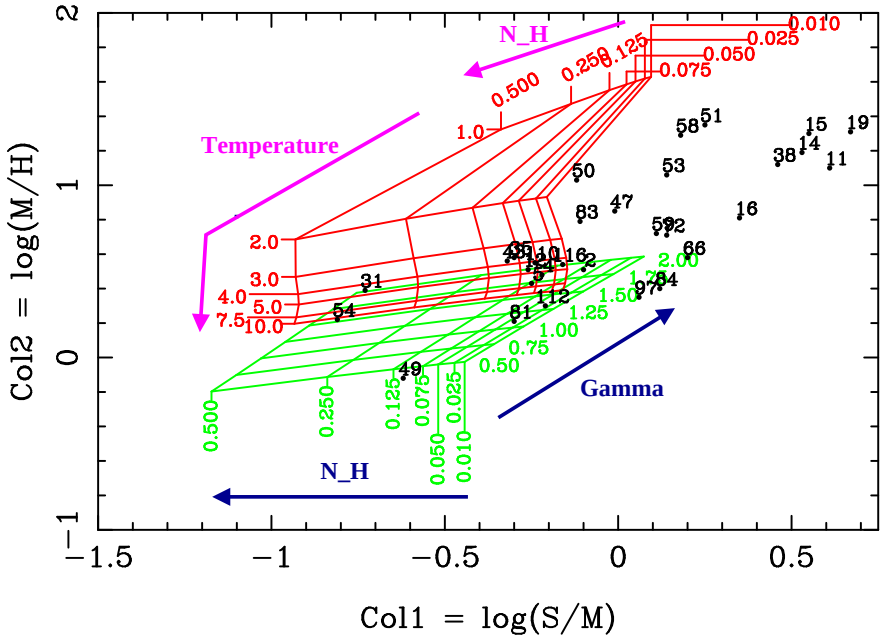

Figure 5.4: A X-ray Color-Color Diagram with the Grids of the Power-Law and Thermal Models. The power-law model is parameterized in terms of N_H and Γ, while the thermal model is parameterized in terms of N_H and Temperature. The grids are drawn for an ideal detector response.

We simulate data for one faint source with $\lambda_S = 10$, $\lambda_M = 20$, and $\lambda_H = 10$, which puts this source into the sector of $N_H = (0.075 - 0.125)$ and $\Gamma = (1.25 - 1.50)$ of the power-law model; the simulated data set consists of $S = 8$, $M = 18$, and $H = 7$, assuming no background contamination. When the classical method is applied to the data, we obtain point estimates with one-dimensional errors bars, as shown in the top left panel of Figure 5.5. However, C_S and C_H are not independent but negatively correlated by construction, so that two marginal error bars do not yield the compact summary of uncertainty. On the other hand, the Bayesian method enables us to directly obtain a joint posterior distribution of the soft and hard colors; we use the flat prior distribution, $\phi = 1.0$. Because we assume independence among source intensities in the sub-energy bands with a single source, the Gibbs sampler runs two independent chains for one with S and M and the other with M and H. Based on the Monte Carlo draws of the source intensities, we obtain a joint posterior distribution of C_S and C_H. The top right panel of Figure 5.5 shows the

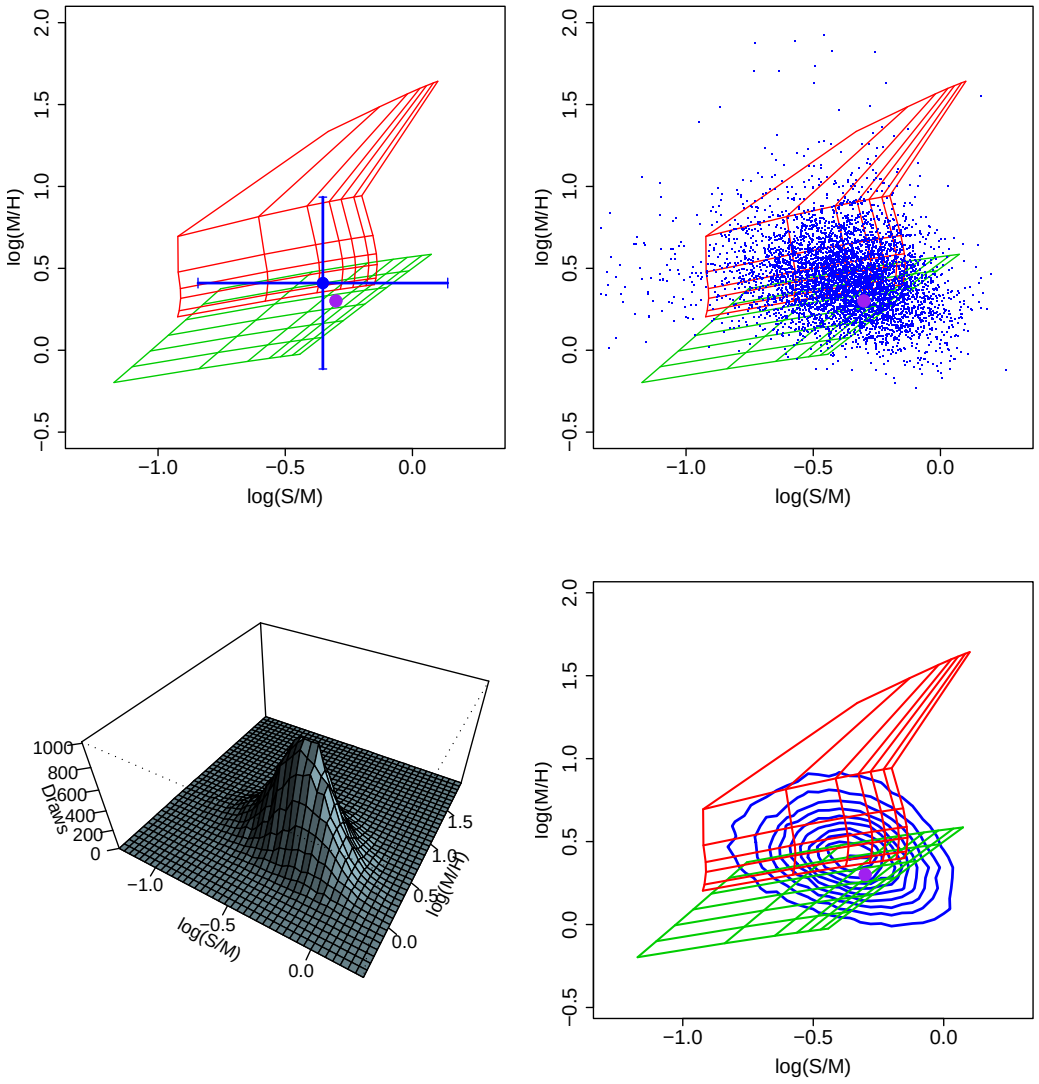

Figure 5.5: X-ray Colors Fitted by the Classical and Bayesian Methods. The top left panel shows the point estimates of colors with marginal error bars fitted by the classical method. In the top right panel, posterior draws of the X-ray colors simulated by the Bayesian method are superimposed on the grids. The bottom panels show three-dimensional graphical summaries for the posterior draws. The large dot in the panels except the bottom left one represents the true values of X-ray colors.

Table 5.2: Posterior Probabilities for the Grid of the Power-Law Model. The 95% posterior region is indicated in bold face.

		N_H					
		0.250–0.500	0.125–0.250	0.075–0.125	0.050–0.075	0.025–0.050	0.010–0.025
Γ	1.75–2.00	**11.36%**	**13.93%**	**3.35%**	**1.00%**	0.53%	0.24%
	1.50–1.75	**5.56%**	**13.70%**	**5.99%**	**2.34%**	**1.70%**	0.67%
	1.25–1.50	**1.80%**	**7.76%**	**5.61%**	**3.11%**	**2.82%**	**1.56%**
	1.00–1.25	0.38%	**2.71%**	**2.87%**	**2.26%**	**2.33%**	**1.58%**
	0.75–1.00	0.07%	0.54%	**0.82%**	0.75%	**1.00%**	0.81%
	0.50–0.75	0.01%	0.09%	0.15%	0.18%	0.23%	0.17%

joint posterior draws of C_S and C_H resulting from the Gibbs sampler; a large dot in the diagram represents the true values of the X-ray colors. In the bottom row of Figure 5.5 presents the three-dimensional histogram of the draws to the left and the contour plot to the right.

Because the Monte Carlo draws are superimposed on the grids of the power-law and thermal models in the color-color diagram, we can reversely infer the parameters of the models by computing posterior probabilities corresponding to each section split by the grids. Table 5.2 presents the normalized posterior probabilities of the X-ray colors in the grid of the power-law model. The 95% highest joint posterior density (HJPD) region is shown in bold face. If the power-law model is believed for this source, the most likely parameter values are $\widehat{N}_H = (0.125 - 0.250)$ and $\widehat{\Gamma} = (1.75 - 2.00)$.

5.6.2 Cluster Analysis for Galaxy Sources

With a survey of X-ray sources, hardness ratios can be used to answer scientific questions of interest. For example, the negative relationship between the soft band X-ray flux (λ_S) and the reciprocal of the simple hardness ratio ($1/\mathrm{R} = \lambda_H/\lambda_S$) is of interest; in this case, the energy spectrum is divided into two sub-energy bands.

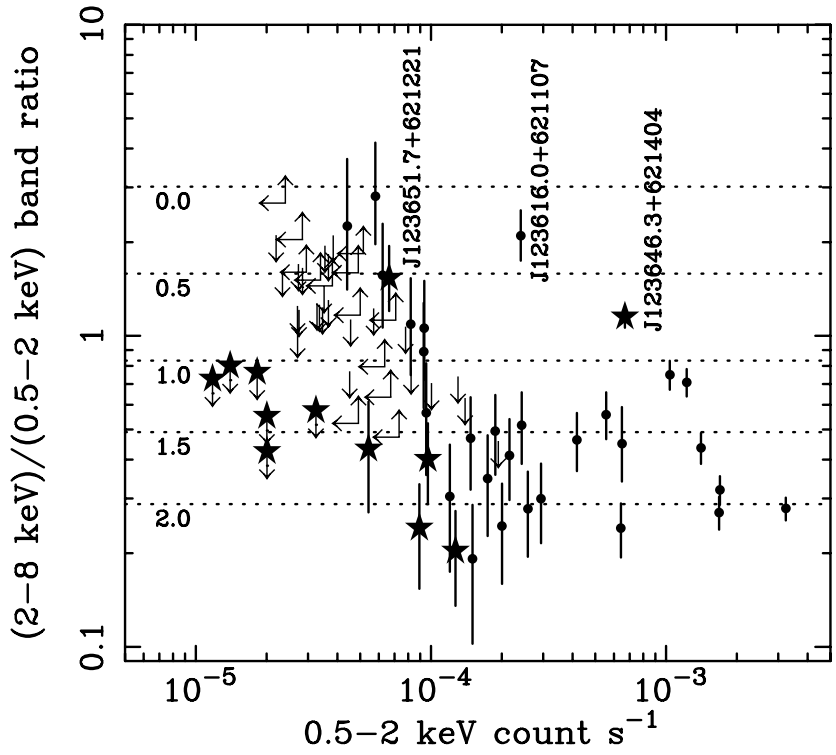

Figure 5.6: Band Ratio As a Function of Soft-Band (0.5 – 2.0 keV) Count Rate for *Chandra* Sources. By this figure, Brandt *et al.* (2001) reported that fainter sources tend to have more hard-band (2.0 – 8.0 keV) counts per unit soft count.

This scientific question specifically means that sources with fewer soft counts tend to have more hard counts per unit soft count. Brandt *et al.* (2001) report this negative relationship on a log scale, based on the method of moments. However, the correlation between $\log_{10}\lambda_S$ and $\log_{10}(\lambda_H/\lambda_S)$ is analytically decomposed into

$$\mathrm{Corr}\left(\log_{10}\lambda_S, \log_{10}\frac{\lambda_H}{\lambda_S}\right) = \frac{\mathrm{Corr}(\log_{10}\lambda_S, \log_{10}\lambda_H)\frac{\sqrt{\mathrm{Var}(\log_{10}\lambda_H)}}{\sqrt{\mathrm{Var}(\log_{10}\lambda_S)}} - 1}{\sqrt{\mathrm{Var}(\log_{10}\lambda_H - \log_{10}\lambda_S)}/\sqrt{\mathrm{Var}(\log_{10}\lambda_S)}}, \quad (5.31)$$

and its sign is negative if and only if the numerator is less than zero. In other words, the correlation of scientific interest becomes negative when the slope for regressing $\log_{10}\lambda_H$ on $\log_{10}\lambda_S$ is less than one, i.e.,

$$\varphi \equiv \mathrm{Corr}(\log_{10}\lambda_S, \log_{10}\lambda_H)\frac{\sqrt{\mathrm{Var}(\log_{10}\lambda_H)}}{\sqrt{\mathrm{Var}(\log_{10}\lambda_S)}} < 1. \quad (5.32)$$

Thus, the scientific question must be re-formalized in terms of the regression slope φ. If the regression slope is zero, knowing $\log_{10}\lambda_S$ does not help explain the vari-

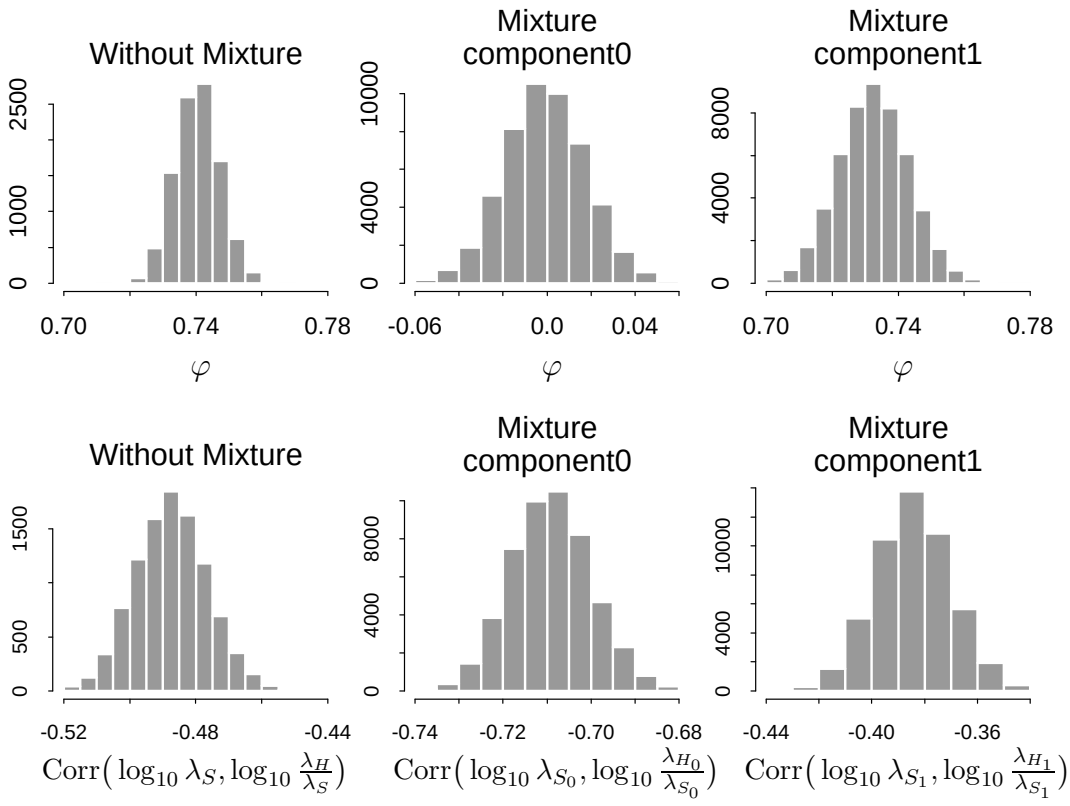

Figure 5.7: Posterior Distributions of the Regression Slopes and Overall Correlations Fitted by MODEL 1 and MODEL 2. The top row shows the posterior distributions of φ, while the bottom row shows the posterior distributions of the correlation between $\log_{10}\lambda_S$ and $\log_{10}(\lambda_H/\lambda_S)$.

ation in $\log_{10}\lambda_H$. However, a zero regression slope results in a negative overall correlation in (5.31), thereby misleading its interpretation.

To address the scientific question, we use real X-ray source data from *Chandra Multiwavelength Project* (ChaMP). To begin with, we consider two hierarchical mixture models introduced in Section 5.4.2:

MODEL 1 : X-ray sources are from one underlying component.

MODEL 2 : X-ray sources are from two underlying components.

In order words, MODEL 1 has no mixture component, while MODEL 2 has two mixture components. Using the Gibbs sampler described in Section 5.4.2, we fit both

models to obtain the posterior distribution of the regression slope. The top row of Figure 5.7 presents the posterior distribution of the regression slope φ. Under MODEL 1, the posterior distribution of φ is below one, so that the overall correlation between $\log_{10} \lambda_S$ and $\log_{10}(\lambda_H/\lambda_S)$ in (5.31) is negative. In the case of MODEL 2, the component-wise regression slopes are also less than one and imply the negative overall correlations, but they differ in how much the slope is less than one. Using the mixture model, the posterior mode of the mixture proportion is around 34%. Thus, almost one third of X-ray sources in ChampX have an almost zero regression slope, implying that the soft and hard intensities may not be correlated with each other. The second row of Figure 5.7 shows the overall correlation in (5.31). As expected, the posterior distributions of the overall correlation are below zero. Interestingly, the overall correlation fitted by MODEL 1 is distributed between the posterior distributions of two well-separated component-wise overall correlations fitted by MODEL 2. This, the overall correlation of MODEL 1 seems to serve as a pooled estimate of the two component-wise overall correlations of MODEL 2.

5.7 Discussion

5.7.1 R versus C versus HR

With low counts data, the posterior distribution of the counts ratio, R, tends to be skewed to the right because of the Poissonian nature of data; R is necessarily positive. The X-ray color, $\mathrm{C} = \log_{10} \mathrm{R}$, is a log transformation of R, which makes the skewed distribution more symmetric. The fractional difference hardness ratio, $\mathrm{HR} = (1 - \mathrm{R})/(1 + \mathrm{R})$, is a monotonically decreasing transformation of R, so that HR approaches to 1 as R tends to 0 (i.e., a source gets harder) and to –1 as R tends to ∞ (i.e., a source gets softer). The monotone transformation makes the values of HR bounded below by –1 and above by 1, thereby reducing asymmetry of the skewed distribution. R and HR are bounded on one side or two sides, while C is

unbounded due to the log transformation.

The posterior distribution of any hardness ratio becomes more symmetric as both soft and hard intensities increase. Regardless of the size of intensities, however, the X-ray color has the most symmetric posterior distribution among the popular definitions of a hardness ratio. Figure 5.8 illustrates the effect of the magnitude of source intensities on the symmetry of the posterior distribution of each hardness ratio; the posterior distribution of C is confirmed to have the most symmetric posterior distribution. In the figure, we fix $\mathrm{R} = 2$ and the soft and hard intensities are determined by beginning with $\lambda_S = 2$ and $\lambda_H = 1$ and increasing the intensities by a factor of 5 in each subsequent column, assuming no background contamination in the simulation.

5.7.2 Advantages

A significant improvement that our Bayesian methods provide over the classical way of computing hardness ratios is that we use the correct Poisson distribution throughout and do not make asymptotic Gaussian assumptions for high counts data. Thus, while the classical method works well only with high counts data and fails to give reliable results with low counts data, our methods are valid in all regimes. Because the observed counts are non-negative integers, moreover, it is not appropriate to model the counts with Gaussian random variables which are defined on a real line.

Because our methods are based on Poisson assumptions in a fully model based statistical approach, we need not rely on the data-driven estimates of hardness ratios. Instead, we compute the posterior probability distribution of each hardness ratio, which provides reliable estimates and correct error bars even when either or both soft and hard counts are very low. In particular, our methods are not limited to "detectable" counts, requiring no minimum number of counts. Even with high

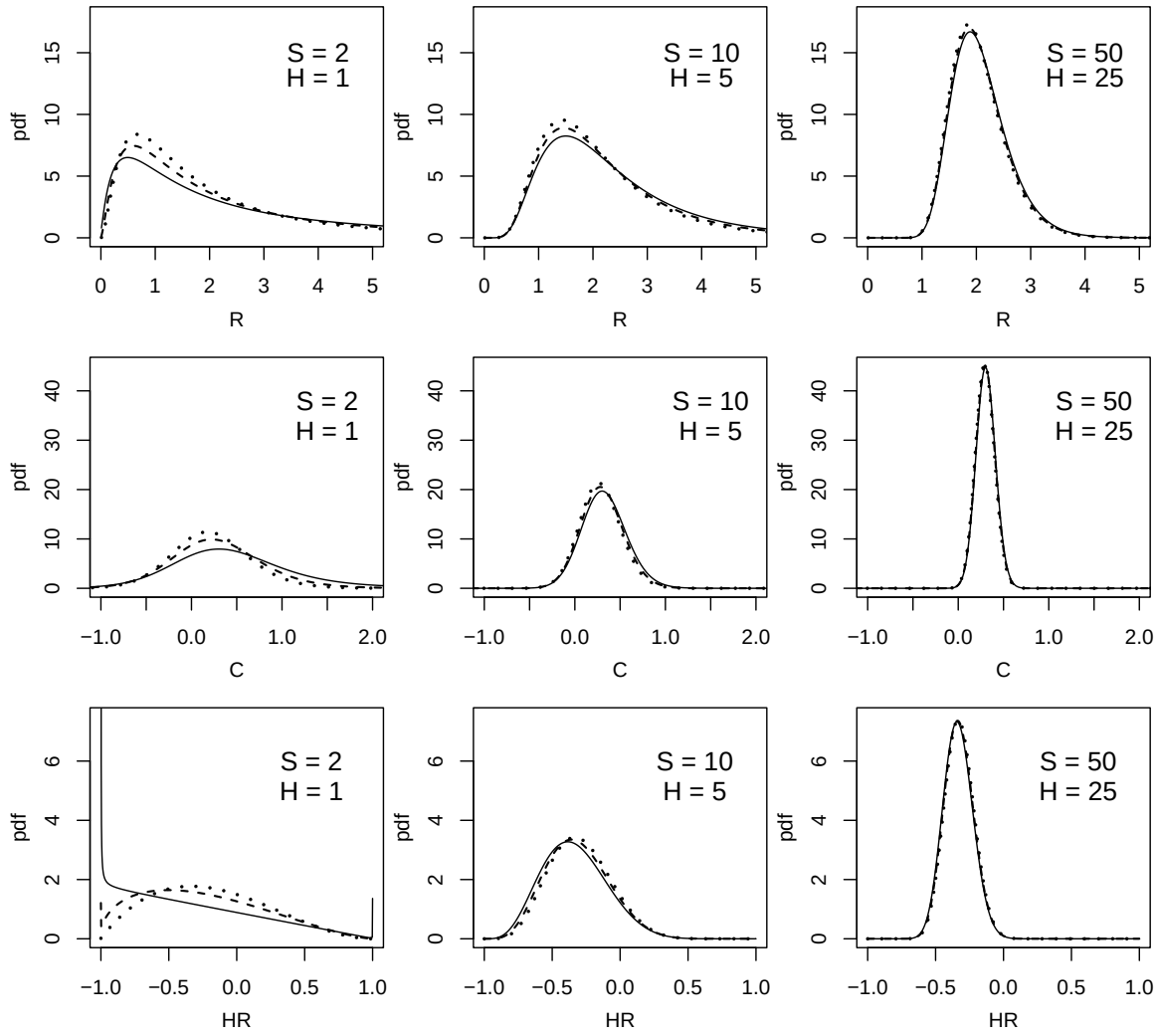

Figure 5.8: Posterior Distributions of R (top row), C (middle row), and HR (bottom row), with Different Source Intensities and Flat Prior Distributions. The solid lines represent the flat prior distribution $\phi = 0.0^{+}$, the dashed lines $\phi = 0.5$, and the dotted lines $\phi = 1.0$. At small counts shown in the left column, the non-symmetric shape of the posterior distribution for each hardness ratio is clear as does the effect of the choice of flat prior distributions. At higher counts shown in the right column, the posterior distributions tend to be symmetric and the effect of the prior distributions on the posterior distribution is minimal.

counts data, our Bayesian methods give more precise estimates and error bars for the hardness ratios than the classical method, although both methods yield similar results. Moreover, a priori information can be embedded into our Bayesian model and can be updated by observed data, thereby producing more accurate results as we collect more data.

5.7.3 Limitations

Because the Bayesian methods do not allow for a simple analytic solution similar to standard error propagation in the Gaussian case, the computational methods used to obtain the probability distributions become important. In this chapter, we implement the Markov chain Monte Carlo scheme (e.g., the Gibbs sampler) and Gaussian quadrature. The Gibbs sampler is based on Monte Carlo simulation, hence the convergence behavior need to be closely examined, especially for the Gibbs sampler with Metropolis-Hastings steps. On the other hand, the Gaussian quadrature precisely computes the posterior distribution as long as the number of bins is large enough; however, its computation becomes expensive as the observed source counts become large. In general, the Gibbs sampler is computationally much quicker than the method based on Gaussian quadrature, but care must be taken to ensure that the number of iterations is sufficient to ensure convergence of a Markov chain. Generally we recommend using the Gibbs sampler for high counts data and the Gaussian quadrature for low counts data.

Appendix

A. Effect of Flat Prior Distributions

For comparison of the effect of these flat prior distributions, we simulate the data sets for different magnitudes of source intensities assuming no background contamination. When a Poisson intensity is small, the shape of a hardness ratio is susceptible to the choice of flat prior distributions because observed data do not contain much information about the intensity. On the other hand, the distribution of a hardness ratio for large counts data hardly depends on the different prior distributions. Table 5.4 illustrates the posterior behavior of the color, C, depending on the magnitudes of intensities (λ_S and λ_H) and the choice of flat prior distributions ($\phi = 0.0^+, 0.5, 1.0$). In particular, Table 5.4 presents the coverage rate and mean length of the 95% posterior intervals computed for the color with 1000 test data simulated with each pair of λ_S and λ_H in the grid of intensities; the legend key is given in Figure 5.3. The coverage rate is the percentage of the simulations that produce 95% posterior intervals of the color actually containing its true value computed with each (λ_S, λ_H) pair, while the mean length is the average of the range for the 95% posterior intervals for the color. We compare the effect of three flat prior distributions on these posterior quantities for the color; we carry out this calculation for three choices of the prior index, $\phi = 0.0^+, 0.5, 1.0$, corresponding to the top,

Table 5.3: Legend Key for Table 5.4.

	Hard band source intensity, λ_H	
Soft band source intensity, λ_S	Coverage rate of intervals with $\phi = 0.0^+$	Average length of intervals with $\phi = 0.0^+$
	Coverage rate of intervals with $\phi = 0.5$	Average length of intervals with $\phi = 0.5$
	Coverage rate of intervals with $\phi = 1.0$	Average length of intervals with $\phi = 1.0$

Table 5.4: Coverage of the X-ray Color (C) Using the Bayesian Method with Different Indexes (0.0^{+}, 0.5, and 1.0).

		λ_H															
		0.5		1.0		2.0		4.0		8.0		16.0		32.0		64.0	
λ_S	0.5	100 %	23.99	100 %	22.67	100 %	21.19	100 %	19.02	100 %	14.07	100 %	11.32	100 %	11.14	100 %	11.11
		100 %	5.10	100 %	4.88	99.9 %	4.51	99.9 %	3.95	100 %	3.35	99.9 %	3.09	99.7 %	3.02	100 %	3.03
		100 %	2.87	100 %	2.71	99.0 %	2.50	98.4 %	2.25	98.4 %	2.02	97.2 %	1.90	97.0 %	1.82	98.4 %	1.83
	1.0	100 %	22.95	100 %	21.59	100 %	19.61	100 %	17.40	100 %	12.24	100 %	10.03	100 %	9.84	100 %	9.51
		100 %	4.91	100 %	4.69	99.7 %	4.31	100 %	3.77	99.9 %	3.04	99.8 %	2.80	99.7 %	2.74	99.8 %	2.71
		100 %	2.73	100 %	2.58	99.2 %	2.37	98.9 %	2.14	98.5 %	1.85	97.3 %	1.73	97.7 %	1.67	98.0 %	1.65
	2.0	100 %	21.42	100 %	19.57	99.8 %	17.52	99.7 %	15.04	99.6 %	10.54	99.3 %	8.45	99.9 %	7.95	99.8 %	7.82
		99.9 %	4.52	99.9 %	4.30	99.7 %	3.89	99.6 %	3.27	99.5 %	2.56	99.2 %	2.33	99.0 %	2.21	99.4 %	2.21
		99.1 %	2.51	99.5 %	2.37	99.7 %	2.14	98.8 %	1.87	97.8 %	1.60	98.1 %	1.49	97.3 %	1.40	97.7 %	1.39
	4.0	100 %	18.60	100 %	16.80	99.9 %	15.18	99.2 %	13.11	97.9 %	8.87	97.9 %	6.06	97.0 %	5.85	96.8 %	5.63
		100 %	3.92	100 %	3.68	99.6 %	3.30	99.2 %	2.64	98.5 %	1.89	98.0 %	1.60	97.9 %	1.53	96.6 %	1.49
		98.4 %	2.24	99.2 %	2.09	99.0 %	1.87	99.1 %	1.59	97.7 %	1.29	97.3 %	1.14	97.1 %	1.08	97.4 %	1.04
	8.0	100 %	13.66	100 %	12.28	99.3 %	10.93	98.3 %	8.74	99.2 %	5.13	98.2 %	2.50	97.8 %	2.26	97.4 %	2.14
		99.9 %	3.29	100 %	3.04	99.7 %	2.64	97.8 %	1.86	98.8 %	1.25	97.6 %	0.96	97.2 %	0.86	97.4 %	0.83
		97.7 %	1.99	98.7 %	1.85	98.7 %	1.63	97.5 %	1.28	98.1 %	1.01	97.3 %	0.83	96.3 %	0.75	97.1 %	0.71
	16.0	100 %	11.37	100 %	9.96	99.8 %	8.49	97.7 %	6.11	98.6 %	2.41	97.9 %	0.73	96.7 %	0.60	95.5 %	0.54
		99.9 %	3.10	99.8 %	2.79	99.2 %	2.34	97.8 %	1.60	98.0 %	0.95	96.6 %	0.65	95.6 %	0.55	94.4 %	0.50
		97.5 %	1.91	98.0 %	1.73	98.2 %	1.49	97.2 %	1.14	96.7 %	0.83	96.6 %	0.63	95.3 %	0.54	93.7 %	0.49
	32.0	100 %	10.82	100 %	9.56	99.3 %	8.00	97.7 %	5.74	97.3 %	2.33	95.7 %	0.59	97.1 %	0.43	93.8 %	0.37
		99.9 %	2.98	99.6 %	2.71	99.0 %	2.21	97.8 %	1.51	96.5 %	0.88	94.4 %	0.55	97.2 %	0.43	94.3 %	0.36
		98.2 %	1.80	97.4 %	1.65	96.0 %	1.40	97.1 %	1.06	96.0 %	0.75	94.5 %	0.54	97.2 %	0.42	94.5 %	0.36
	64.0	100 %	10.99	100 %	9.52	99.7 %	7.84	97.7 %	5.69	97.7 %	2.11	96.3 %	0.53	94.6 %	0.36	96.5 %	0.29
		99.9 %	3.02	99.6 %	2.72	99.0 %	2.23	98.1 %	1.52	97.7 %	0.83	94.8 %	0.50	94.8 %	0.36	96.5 %	0.29
		98.5 %	1.82	97.4 %	1.66	97.1 %	1.40	97.8 %	1.05	96.6 %	0.71	94.2 %	0.49	94.9 %	0.36	96.5 %	0.28

middle, and bottom elements in each cell of Table 5.4. In order to simplify the comparison, we simulate test data using the soft and hard source intensities assuming no background contamination. Based on Table 5.4, we can consider the following four different scenarios, i.e,

1. when both λ_S and λ_H are small (e.g., $\lambda_S = \lambda_H = 0.5$),
2. when both λ_S and λ_H are large (e.g., $\lambda_S = \lambda_H = 64$),
3. when λ_S is much smaller than λ_H (e.g., $0.5 = \lambda_S \ll \lambda_H = 64$), and
4. when λ_H is much smaller than λ_S (e.g., $64 = \lambda_S \gg \lambda_H = 0.5$).

The first scenario is the case of low counts data, while the second scenario is the case of large counts data. A lower value of ϕ makes the posterior distributions of both λ_S and λ_H more leaning toward zero, which results in thicker left and right tails of the posterior distribution of the color, respectively; this is because a smaller λ_S makes C smaller and a smaller λ_H makes C larger. Thus, as ϕ tends to be small with low counts data, the mean length of the 95% posterior intervals becomes larger, while maintaining high coverage rates for different values of ϕ. The second scenario is the case of high counts data, which is opposite to the first scenario in terms of the number of counts. In this case, because observed data contain more information about the intensities, the effect of ϕ on the posterior distribution of the color is minimal, and we have almost identical results no matter what flat prior distributions are used. When one intensity is much smaller than the other, the posterior behavior of the color can be dramatically changed. That is, in the third scenario, a small value of ϕ affects only λ_S, making the color even smaller. Namely, small ϕ makes the posterior distribution of the color have an elongated left tail, but the right tail is barely affected. In the case of the fourth scenario, the exactly opposite situation occurs: the left tail remains similar for different values of ϕ, but

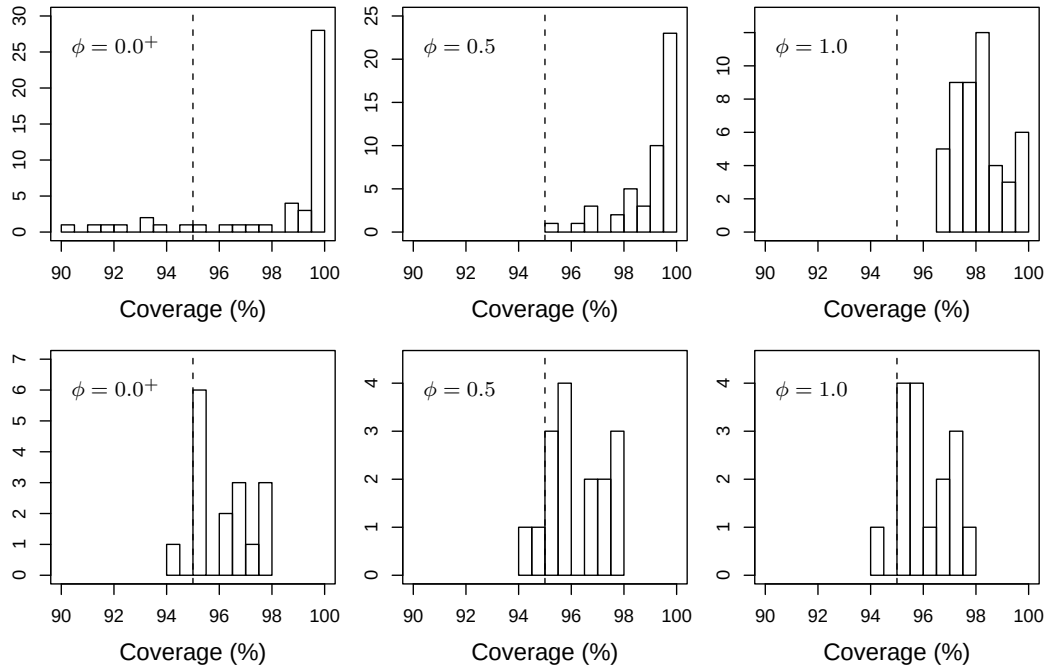

Figure 5.9: Empirical Distributions of Coverage Rates with Different Indexes. Each histogram represents the empirical distribution of a coverage rate for the color. The vertical dotted lines represent the theoretical coverage rate 95%. The top row corresponds to the cases where at least one of λ_S and λ_H is less than or equal to 4 in Table 5.4, while the bottom row to the cases where both λ_S and λ_H are greater than or equal to 8.

the right tail is thicker with a small value of ϕ. Thus, smaller ϕ can result in the bigger mean length of the 95% posterior intervals, as we confirm in Table 5.4.

Considering all different scenarios in Table 5.4, we suggest using the Jeffrey's flat prior distribution (i.e., $\phi = 0.5$) especially when either soft or hard intensity is low. As it turns out, $\phi = 0.5$ is a conservative choice because it maintains fairly good coverage rates and yields the reasonably large mean length of the 95% posterior intervals, regardless of different values of the (λ_S, λ_H) pair. The top row of Figure 5.9 shows the empirical distributions of the coverage rates when at least one of λ_S and λ_H is less than or equal to 4. When we do not have much information about the source intensities, we prefer conservative results with high coverage rates and, at the same time, with reasonably large mean length of the intervals. Among the

three choices of the flat prior distributions, Figure 5.9 suggests the Jeffrey's flat prior distribution as the most conservative flat prior distribution in that sense. With high counts data, i.e., when both λ_S and λ_H are greater than or equal to 8 in Table 5.4, all flat prior distributions seem equivalent because the resulting posterior distributions are very similar to one another, as shown in the bottom row of Figure 5.9. Due to the sampling errors, we expect most coverage rates to be between 93% and 97% which are three standard deviations away from 95% because the standard deviation of the coverage probability for the 95% posterior intervals is given by $\sqrt{(0.95)(0.05)/1000} = 0.0069$.

References

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B, Methodological* **47**, 67–75.

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

Brandt, W., Hornschemeier, A., Alexander, D., Garmire, G., Schneider, D., Broos, P., Townsley, L., Bautz, M., Feigelson, E., and Griffiths, R. (2001). The Chandra deep survey of the hubble deep field north area. IV. an ultradeep image of the HDF-N. *The Astronomical Journal* **122**, 1–20.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Cook, S., Gelman, A., and Rubin, D. B. (2006). Validation of software for bayesian models using posterior quantiles. *The Journal of Computational and Graphical Statistics, to appear*.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* **91**, 883–904.

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, Series B, Methodological* **39**, 1–37.

Dobigeon, N., Tourneret, J.-Y., and Scargle, J. (2005). Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *IEEE Trans. Signal Processing* to appear.

Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions*. Chapman & Hall, London.

Fabian, A., ed. (2005). *X-ray reflections on AGN*, Spain. astro-ph/0511537.

Freeman, P., Doe, S., and Siemiginowska, A. (2001). Sherpa: a mission-independent data analysis application. *Proceedings of SPIE, Astronomical Data Analysis, Jean-Luc Starck; Fionn D. Murtagh; Eds.* **4477**, 76–87.

Freeman, P., Graziani, C., Lamb, D., Loredo, T., Fenimore, E., Murakami, T., and Yashida, A. (1999). Statistical analysis of spectral line candidates in gamma-ray burst GRB 870303. *The Astrophysical Journal* **524**, 753–771.

Gallagher, S. C., Brandt, W. N., Chartas, G., and Garmire, G. P. (2002). X-ray spectroscopy of quasi-stellar objects with broad ultraviolet absorption lines. *The Astrophysical Journal* **567**, 37–41.

Gehrels, N. (1986). Confidence limits for small number of event in astrophysical data. *The Astrophysical Journal* **303**, 336–346.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parameterization for normal linear mixed models. *Biometrika* **82**, 479–488.

Gelman, A., Huang, Z., van Dyk, D. A., and Boscardin, W. J. (2006). Transformed and parameter-expanded Gibbs samplers for multilevel linear and generalized linear models. *Submitted to The Journal of Computational and Graphical Statistics*.

Gelman, A. and Meng, X.-L. (1996). Model checking and model improvement. In *Markov Chain Monte Carlo in Practice* (Editors: W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), 189–201. Chapman & Hall, New York.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness (with discussion). *Statistica Sinica* **6**, 733–807.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences (with discussion). *Statistical Science* **7**, 457–472.

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.

Hastings, W. K. (1970). Monte Carlo sampling methods usings Markov chains and their applications. *Biometrika* **57**, 97–109.

Henna, J. (1985). On estimation of countable mixtures of continuous distributions. *Journal of the Japanese Statistical Society* **15**, 75–82.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. The Institute of Mathematical Statistics and the American Statistical Association, Hayward, CA and Alexandria, VA.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measure data. *Journal of the American Statistical Association* **83**, 1014–1022.

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. (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.

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

Maine, M., Boullion, T., and Rizzuto, G. T. (1991). Detecting the number of components in a finite mixture having normal components. *Communication in Statistics – Theory & Methods* **20**, 611–620.

Markoff, S., Nowak, A. N., and Wilms, J. (2005). Going with the flow: Can the base of jets subsume the role of compact accretion disk coronae? *The Astrophysical Journal* **635**, 1203–1216.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142–1160.

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.

Merton, R. (1976). Option pricing when the underlying stock returns are discontinuous. *Journal of Financial Economics* **3**, 125–144.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association* **44**, 335–341.

Park, T. (2004). Spectral analysis with delta functions emission lines. *Ph.D. Qualifying Paper, Department of Statistics, Harvard University*.

Pilla, R. S. and Lindsay, B. G. (1996). Faster EM methods in high-dimensional finite mixtures. *Proceedings of the Statistical Computing Section of the American Statistical Association* 166–171.

Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2002). Statistics: Handle with care – detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal* **571**, 545–559.

Rees, M. J. (1978). Accretion and the quasar phenomenon. *Physica Scripta* **17**, 193–200.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151–1172.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. *Journal of the American Statiatical Association* **82**, 543–546.

Schafer, J. L. (1998). *Some Improved Procedures for Linear Mixed Models*. Technical Report 98-28, The Methodology Center, The Pennsylvania State University, available at http://methcenter.psu.edu/publications/index.html.

Sikora, M., Madejski, G., Moderski, R., and Poutanen, J. (1997). Learning about active galactic nucleus jets from spectral properties of blazars. *The Astrophysical Journal* **484**, 108.

Sobolewska, M. A., Siemiginowska, A., and Zycki, P. T. (2004). High-redshift radio-quiet quasars: Exploring the parameter space of accretion models. Part II. Patchy corona model. *The Astrophysical Journal* **617**, 102–112.

Steidel, C. C. and Sargent, W. L. W. (1991). Emission-line and continuum properties of 92 bright QSOs - luminosity dependence and differences between radio-selected and optically selected samples. *The Astrophysical Journal* **382**, 433–465.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley & Sons, New York.

van Dyk, David A. Connors, A., Esch, D. N., Freeman, P., Kang, H., Karovska, M., Kashyap, V., Siemiginowska, A., and Zezas, A. (2006). Deconvolution in high-energy astrophysics: Science, instrumentation, and methods. *Bayesian Analysis* **1**, 139–236.

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. A. (2000a). Fitting mixed-effects models using efficient EM-type algorithms. *The Journal of Computational and Graphical Statistics* **9**, 78–98.

van Dyk, D. A. (2000b). Nesting EM algorithms for computational efficiency. *Statistical Sinica* **10**, 203–225.

van Dyk, D. A. (2003). Hierarchical models, data augmentation, and Markov chain Monte Carlo with discussion. In *Statistical Challenges in Modern Astronomy III* (Editors: E. Feigelson and G. Babu), 41–56. Springer–Verlag, New York.

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

van Dyk, D. A. and Hans, C. M. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray Observatory. In *Spatial Cluster Modelling* (Editors: D. Denison and A. Lawson), 175–198. CRC Press, London.

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

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. (2001). The art of data augmentation. *The Journal of Computational and Graphical Statistics* **10**, 1–111.

Weisskopf, M. C., Brinkman, B., Canizares, C., Garmire, G., Murray, S., and Van Speybroeck, L. P. (2002). An overview of the performance and scientific results from the chandra x-ray observatory. *The Publications of the Astronomical Society of the Pacific* **114**, 1–24.

Wichura, M. J. (1989). An algorithm for patterson gaussian quadrature. *Technical Report No. 257, Department of Statistics, The University of Chicago*.

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