Using Surrogate Distributions to Improve the Convergence Properties of Gibbs-type Samplers

a thesis presented for the degree of

Doctor of Philosophy of Imperial College London

and the

Diploma of Imperial College

by

Xiyun Jiao

Department of Mathematics
Imperial College
180 Queen's Gate, London SW7 2AZ

October 2016

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Xiyun Jiao

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Thesis advisor: Professor David A. van Dyk Xiyun Jiao

Using Surrogate Distributions to Improve the Convergence Properties of Gibbs-type Samplers

Abstract

Gibbs-type samplers are widely used tools for obtaining Monte Carlo samples from posterior distributions under complicated Bayesian models. Standard Gibbs samplers update component quantities of the parameter by sequentially sampling their conditional distributions under the target joint distribution. However, this strategy can be slow to converge if the components are highly correlated. We formalize a general strategy to construct more efficient samplers by replacing some of the conditional distributions with conditionals of a surrogate distribution. The surrogate distribution is designed to share certain marginal distributions with the target, but with lower correlations among its components. Although not necessarily recognized when they were introduced, a number of existing strategies for improving Gibbs can be formulated in this way (e.g., Marginal Data Augmentation, Partially Collapsed Gibbs sampling, Ancillarity-Sufficiency Interweaving Strategy, etc.). The use of surrogate distributions in Gibbs-type samplers may lead to incompatible conditional distributions and thus sensitivity to the order of the component draws. We propose a framework to combine different strategies involving surrogate distributions into a single coherent sampler that maintains the target stationary distribution and outperforms any of its component algorithms in terms of convergence. We use both theoretical arguments and numerical examples to illustrate the implementation and efficiency of our strategy. A problem in supernova cosmology has motivated our work and serves as a realistic testing ground for our methods. Finally, we correct two errors in the related Marginal Data Augmentation algorithms of Imai and van Dyk (2005) that are quite popular for fitting multinomial probit models.

I dedicate this thesis to my family,
for their endless support, encouragement, and love.

Acknowledgments

I would like to thank my supervisor, Prof. David A. van Dyk, for his generous donation of time, effort, and patience to instructing my PhD research. Also, I would like to thank Dr. Roberto Trotta for his constant instruction and help on my work in astrophysics. In addition, I would like to thank Prof. Axel Gandy and Dr. Leonardo Bottolo for their constructive suggestions on my research during early and late stage reviews. Finally, I would like to thank Hikmatali Shariff for the numerous helpful discussions on our project in supernova cosmology.

Xiyun Jiao

List of Publication

1. van Dyk, D. A. & Jiao, X. (2015). Metropolis-Hastings within Partially Collapsed Gibbs samplers. *Journal of Computational and Graphical Statistics*, **24**, 301–327.

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

2. Jiao, X., van Dyk, D. A., Trotta, R., & Shariff, H. (2015). The efficiency of next-generation Gibbs-type samplers: An illustration using a hierarchical model in cosmology. In *Proceedings of Joint Applied Statistics Symposium of International Chinese Statistical Association & Korean International Statistical Society, 2014*, to appear.

   Abstract: Supernovae occur when a stars life ends in a violent thermonuclear explosion, briefly outshining an entire galaxy before fading from view over a period of weeks or months. Because so-called Type Ia supernovae occur only in a particular physical scenario, their explosions have similar intrinsic brightnesses which allows us to accurately estimate their distances. This in turn allows us to constrain the parameters of cosmological models that characterize the expansion history of the universe. In this paper, we show how a cosmological model can be embedded into a Gaussian hierarchical model and fit using observations of Type Ia supernovae. The overall model is an ideal testing ground of new computational methods. Ancillarity- Sufficiency Interweaving Strategy (ASIS) and Partially Collapsed Gibbs (PCG) are effective tools to improve the convergence of Gibbs samplers. Besides using either of them alone, we can combine PCG and/or ASIS along with Metropolis-Hastings algorithm to simplify implementation and further improve convergence.We use four samplers to draw from the posterior distribution of the cosmological hierarchical model, and confirm the efficiency of both PCG and ASIS. Furthermore, we find that we can gain more efficiency by combining two or more strategies into one sampler.

3. Jiao, X. & van Dyk, D. A.: A corrected and more efficient suite of MCMC samplers for the multinomial probit model. *Journal of Econometrics*, to appear.

   Abstract: The multinomial probit (MNP) model is a useful tool for describing discrete-choice data and there are a variety of methods for fitting the model. Among them, the algorithms provided by Imai and van Dyk (2005a), based on Marginal Data Augmentation, are widely used, because they are efficient in terms of convergence and allow the prior distribution to be specified directly on identifiable parameters. Burgette and Nordheim (2012) modified a model and algorithm of Imai and van Dyk (2005a) to avoid an arbitrary choice that is often made to establish identifiability. There is an error in the algorithms of Imai and van Dyk (2005a), however, which affects both their algorithms and that of Burgette and Nordheim (2012). This error can alter the stationary distribution and the resulting fitted parameters as well as the efficiency of these algorithms. We propose a correction and use both a simulation study and real-data analyses to illustrate the difference between the original and corrected algorithms, both in terms of their estimated posterior distributions and their convergence properties. In some cases, the effect on the stationary distribution can be substantial.

4. Shariff, H., Jiao, X., Trotta, R., & van Dyk, D. A.: BAHAMAS: New SNIa analysis reveals inconsistencies with standard cosmology. *The Astrophysical Journal*, to appear.

   Abstract: We present results obtained by applying our BAyesian HierArchical Modeling for the Analysis of Supernova cosmology (BAHAMAS) software package to the 740 spectroscopically confirmed supernovae of type Ia (SNe Ia) from

the “Joint Light-curve Analysis” (JLA) data set. We simultaneously determine cosmological parameters and standardization parameters, including corrections for host galaxy mass, residual scatter, and object-by-object intrinsic magnitudes. Combining JLA and Planck data on the cosmic microwave background, we find significant discrepancies in cosmological parameter constraints with respect to the standard analysis: we find $\Omega_m = 0.399 \pm 0.027$, 2.8σ higher than previously reported, and $w = -0.910 \pm 0.045$, 1.6σ higher than the standard analysis. We determine the residual scatter to be $\sigma_{\text{res}} = 0.104 \pm 0.005$. We confirm (at the 95% probability level) the existence of two subpopulations segregated by host galaxy mass, separated at $\log(M/M_{\odot}) = 10$, differing in mean intrinsic magnitude by 0.055 ± 0.022 mag, lower than previously reported. Cosmological parameter constraints, however, are unaffected by the inclusion of corrections for host galaxy mass. We find $\sim 4\sigma$ evidence for a sharp drop in the value of the color correction parameter, $\beta(z)$, at a redshift $z_t = 0.662 \pm 0.055$. We rule out some possible explanations for this behavior, which remains unexplained.

5. Shariff, H., Dhawan. S., Jiao, X., Leibundgut, B., Trotta, R., & van Dyk, D. A.: Standardizing Type Ia supernovae using near infrared rebrightening timing. *Monthly Notices of the Royal Astronomical Society*, to appear.

   Abstract: Accurate standardization of Type Ia supernovae (SNIa) is instrumental to the usage of SNIa as distance indicators. We analyse a homogeneous sample of 22 low-z SNIa, observed by the Carnegie Supernova Project (CSP) in the optical and near infra-red (NIR). We study the time of the second peak in the NIR band due to re-brightening, t_2, as an alternative standardization parameter of SNIa peak brightness. We use BAHAMAS, a Bayesian hierarchical model for SNIa cosmology, to determine the residual scatter in the Hubble diagram. We find that in the absence of a colour correction, t_2 is a better standardization parameter compared to stretch: t_2 has a 1σ posterior interval for the Hubble residual scatter of $\sigma_{\Delta\mu} = \{0.250, 0.257\}$, compared to $\sigma_{\Delta\mu} = \{0.280, 0.287\}$ when stretch (x_1) alone is used. We demonstrate that when employed together with a colour correction, t_2 and stretch lead to similar residual scatter. Using colour, stretch and t_2 jointly as standardization parameters does not result in any further reduction in scatter, suggesting that t_2 carries redundant information with respect to stretch and colour. With a much larger SNIa NIR sample at higher redshift in the future, t_2 could be a useful quantity to perform robustness checks of the standardization procedure.

List of figures

List of tables

Contents

1

Introduction

Multilevel statistical models with hierarchical structures have become more and more popular in the physical, social, and engineering sciences, since they allow us to model complex phenomena and data-generating mechanisms. In Bayesian analyses, such models are generally fit by obtaining a Monte Carlo sample from their posterior distributions and using this sample to quantify likely values of model parameters, their correlations and uncertainties, and to make predictions under the models. Although it is often infeasible to obtain an independent Monte Carlo sample, a Markov chain can be constructed such that its stationary distribution equals the target posterior distribution. This is Markov Chain Monte Carlo (MCMC), for which Robert and Casella (2004) provided a brief history. Realizations of MCMC chains after sufficient burn-in approximate a correlated sample from the target posterior distribution.

Two of the most common strategies for constructing MCMC samples are the Gibbs sampler and Data Augmentation (DA) algorithm. The Gibbs sampler was introduced by Geman and Geman (1984) to cope with specific problems in Bayesian image analysis, and Smith and Roberts (1993) developed it for general Bayesian computation. The Gibbs sampler is an example of the method of model reduction, which uses a set of conditional distributions to construct a simple-to-implement and stable algorithm. In particular, it decomposes a high-dimensional joint distribution by using two or more steps, each of which samples a lower-dimensional component of the parameter from its *complete conditional distribution*, i.e., the distribution conditioning on the current values of all the other components. Gibbs is especially useful when the joint distribution cannot be easily sampled, while the complete conditional distributions are in closed form or are easy to sample. Furthermore, under mild conditions, as the number of iterations grows, the distribution of the joint Gibbs Markov chain converges to the target and distributions of the (perhaps non-Markovian) sub-chains converge to their corresponding marginal distributions, see Robert and Casella (2004) and Meyn and Tweedie (1993). Unfortunately, the Gibbs sampler, while easy to implement, is sometimes slow to converge.

While the DA algorithm can be viewed as a two-step Gibbs sampler, it has distinct history, motivation, and implementation strategy. Instead of decomposing the parameter space, data augmentation expands the dimension of the unknown quantities in such a way that the target distribution is preserved while enabling a simple Gibbs sampler on a larger set of unknowns. Tanner and Wong (1987) used the idea of data augmentation in posterior sampling and developed the DA algorithm. As with the Gibbs sampler, the main goal of the DA algorithm is to simplify implementation rather than to improve speed. The method of auxiliary variables (see, e.g., Besag and Green (1993) for an overview), which was developed independently of the DA algorithm, essentially utilized the same strategy but with the goal of speeding up convergence. See van Dyk and Meng (2010) for a detailed review and comparison of algorithms involving data augmentation.

A variety of extensions of the Gibbs sampler and the DA algorithm have been proposed to improve their convergence. The conditional data augmentation (CDA) and marginal data augmentation (MDA) algorithms (Meng and van Dyk, 1999), for example, were inspired by the idea of expanding the parameter space. Both methods introduce a working parameter, and either condition on the value of the working parameter leading to the best performance (in the case of CDA) or sample it in the iteration (in the case of MDA). Their effectiveness has been demonstrated in faster samplers for fitting factor analysis models (Ghosh and Dunson, 2009), capture-recapture models (Royle and Dorazio, 2012), Gaussian copula factor models (Murray *et al.*, 2013), probit models for independent or spatially correlated ordinal data (Schliep and Hoeting, 2015), etc. The MDA algorithm was developed independently by Liu and Wu (1999) under the name Parameter-Expanded Data Augmentation (PXDA). Both Meng and van Dyk (1999) and Liu and Wu (1999) pointed out the potential to further boost computational efficiency using generalized versions of their algorithms under limiting improper priors for the working parameter. Meng and van Dyk (1999) used empirical evidence to illustrate the possibility of obtaining positive recurrent Markov chains with better convergence by purposely constructing a non-positive recurrent chain in a larger parameter space using improper working priors. Liu and Wu (1999) proved that the PX-DA algorithm with a Haar measure prior (often improper) for the working parameter is optimal among a class of such acceleration algorithms. Hobert and Marchev (2008) unified the DA, PX-DA, and Haar PX-DA algorithms under the framework of "sandwiched" transition kernels, verified the computational advantage of (Haar) PX-DA using this framework, and provided a promising general approach for improving convergence of the DA algorithm. Hobert and Marchev (2008) also claimed that when group structure is present in the set of one-to-one mappings indexed by the working parameter, the MDA algorithm under limiting improper working priors is equivalent to the Haar PX-DA in regard to the sandwiched framework. Moreover, van Dyk (2010) introduced the technique of MDA into Gibbs samplers and broadened the class of problems that can benefit from MDA.

The Partially Collapsed Gibbs (PCG) sampler (van Dyk and Park, 2008; Park and van Dyk, 2009) is a useful tool to accelerate the convergence of Gibbs samplers. The PCG sampler, deploying the model reduction technique, improves the convergence of a Gibbs sampler by reducing the conditioning in some of its steps. This strategy has been very effective in practice, for example, in algorithms for fitting quantile regression models (Reed and Yu, 2009), spatial probit regression models (Berrett and Calder, 2012), a model for orthogonal component analysis of sparse representation (Dobigeon and Tourneret, 2010), a model determining the parameters of multi-path components for ultra-wide-band channels in engineering (Kail *et al.*, 2011), and a model for reconstructing the full three dimensional velocity field from observed distances and spectroscopic galaxy catalogues in astrophysics (Lavaux, 2016). PCG can be viewed as a generalization of blocking and collapsing (Liu *et al.*, 1994), which also improve the performance of Gibbs samplers.

A good choice of parameterization is essential to producing efficient sampling schemes. This is emphasized, for example, by Papaspiliopoulos *et al.* (2007) in the context of hierarchical models. Yu and Meng (2011) developed a strategy for boosting the efficiency of Gibbs-type samplers by interweaving two special parameterizations. As implied by its name, ASIS uses a pair of DA schemes: a sufficient and an ancillary augmentation. It is usually the case that the sampler constructed via one of these two augmentations is fast while the other is slow. ASIS takes advantage of this "beauty-and-beast" feature of the two DA algorithms by interweaving one into the other, producing a sampler that is more efficient than either alone. The computational benefits of ASIS have been illustrated in fitting, e.g., the multinomial logit model with Gaussian process priors (Filippone *et al.*, 2012), stochastic volatility models (Kastner and Fruhwirth-Schnatter, 2014), and the complex hierarchical model of infrared spectral energy distributions in astrophysics (Kelly *et al.*, 2012).

In a Gibbs-type sampler, when one or more of its component conditional distributions is not in closed form, we can use a Metropolis-Hastings (MH) update (Metropolis *et al.*, 1953; Hastings, 1970). While using MH within a Gibbs sampler is straightforward, embedding MH within a PCG sampler can alter the stationary distribution of the chain. This can happen even when the PCG sampler would work perfectly well if all of the conditional updates were available without resorting to MH updates. Examples arise even in a two-step MH within PCG sampler. Woodard *et al.* (2013), for example, pointed out this problem in certain samplers described in the literature for regression with functional predictors. Although they did not use the framework of PCG, these samplers are simple special cases of improper MH within PCG samplers. In this dissertation, we illustrate difficulties that may arise when using MH updates within a PCG sampler and develop a general procedure for deriving an MH within PCG sampler from the original Gibbs sampler while maintaining the target stationary distribution. We use theoretical arguments to guide the choice between different MH within PCG sampling schemes, and use numerical examples from our applied work to illustrate the MH within PCG sampler and its computational advantage. The work on embedding MH into PCG samplers has been published in van Dyk and Jiao (2015).

Among the many acceleration strategies, we focus on the MDA, ASIS, and PCG algorithms. In practice, we sometimes find that one of these strategies is efficient for improving convergence, but only for a subset of the parameters. At the same time, another strategy may only help the set of parameters unaffected by the first strategy. Thus, if there are multiple parameters that exhibit poor convergence, it can be difficult to decide which strategy to use. We propose constructing new samplers that use multiple acceleration strategies. This allows for more flexibility and power to produce samplers that are both easy to implement and fast to converge. There are already examples adopting this combining strategy in the literature. For example, as mentioned above, MH is often embedded into a Gibbs-type algorithm to facilitate sampling from one or more of its component conditional distributions (e.g., Gilks *et al.*, 1995; van Dyk and Jiao, 2015). In another example, van Dyk and Meng (2001) showed how the conditional and marginal data augmentation algorithms could be combined to derive efficient data augmentation schemes. Unless we are careful, however, combining strategies in this way may alter the stationary distribution of the overall sampler. Thus in this dissertation, we construct a general framework for combining different acceleration strategies into a single coherent sampler that maintains the target stationary distribution, and verify the efficiency of this framework for improving convergence using both theoretical arguments and numerical examples.

The essence of a number of existing acceleration strategies for improving the convergence rate of Gibbs samplers is to reduce the correlation between components within the Gibbs Markov chain. Inspired by this idea, we formalize a general strategy to construct more efficient samplers by replacing some of the conditional distributions with conditionals of a *surrogate distribution*. The surrogate distribution is designed to share certain marginal distributions with the target, but with lower correlations among its components. Although not necessarily recognized when they were introduced, MDA (including Haar PX-DA), PCG, and ASIS can all be formulated in terms of surrogate distributions. Thus in the combining strategy described above, when combining different strategies into a single sampler, we in fact replace some of the conditional distributions of the original sampler with conditionals of different surrogate distributions. As mentioned above, Hobert and Marchev (2008) has already adopted the idea of unifying different acceleration strategies under a general framework, which is promising to produce more efficient samplers. Their work, however, just focused on DA, PX-DA, and Haar PX-DA algorithms. The use of surrogate distributions in Gibbs-type samplers may lead to incompatible conditional distributions and thus sensitivity to the order of the component draws. In this dissertation, we demonstrate how to derive surrogate distributions from PCG, MDA, and ASIS algorithms, and use several examples to illustrate the potential of manipulating surrogate distributions to further improve the convergence of Gibbs-type samplers.

Combining strategies to improve the convergence of Gibbs samplers via the framework of surrogate distributions is motivated by computational challenges from our applied work in supernova cosmology. The Physics Nobel Prize (2011) was awarded for the discovery that the expansion of the universe is accelerating, a phenomenon attributed to the existence of "dark energy". Type Ia supernova (SN) observations have been instrumental in this discovery and remain an important tool to quantify the characteristics of dark energy (March *et al.*, 2011). Although details remain unclear, it is thought that a Type Ia SN occurs when a compact, carbon-oxygen white dwarf star accumulates extra material until its mass approaches a critical threshold ("Chandrasekhar threshold": $1.44 M_\odot$, where $M_\odot$ is the mass of the sun). Because of their common formation mechanism, all Type Ia SNe have similar absolute luminosity (which is measured in absolute magnitudes, i.e., the negative logarithm of flux). This means that their distance can be estimated from their apparent magnitude (i.e., their brightness as viewed from earth). We can also directly measure their redshift, a stretching of the wavelength of light emanating from objects moving away from us due to the expansion of the universe. The underlying cosmological models of interest predict the relationship between redshift and the difference between apparent and absolute magnitudes, called "distance modulus". We embed the cosmological model into a Gaussian hierarchical model that naturally represents the structure of the problem and dependence among its parameters. (See March *et al.* (2011) for more details on the formation of Type Ia SNe and their utility in fitting cosmological models.) When sampling from the posterior distribution of the cosmological hierarchical model, the standard Gibbs sampler converges slowly. This fact has largely motivated our work in developing more efficient samplers. Because of its complexity, the overall model has served as an ideal testing ground of our new computational methods. In this manuscript, we explore the relative efficiencies of four algorithms in fitting the hierarchical model and confirm the efficiency of both PCG and ASIS. Furthermore, we verify that we can gain more efficiency by combining two or more strategies into one sampler. In addition, we briefly describe our applied work in supernova cosmology, for which the details appear in Shariff *et al.* (2016).

We also consider the samplers for fitting the multinomial probit (MNP) model in some detail. The MNP model is widely used for modelling discrete-choice data in social sciences and transportation studies and it is typically fit from a Bayesian perspective using MCMC. Popular algorithms specify a set of latent Gaussian variables as augmented data, whose relative magnitudes determine the choices. McCulloch and Rossi (1994), for example, advocated a Gibbs sampler that was the first feasible Bayesian approach to fitting the MNP model. In their specification, however, the prior distribution for the identifiable parameters is only determined as a byproduct (Imai and van Dyk, 2005) (henceforth, IvD). Improvements to McCulloch and Rossi (1994) were introduced by Nobile (1998) and McCulloch *et al.* (2000), which are the bases for the comprehensive bayesm R package. Another set of samplers, based on the MDA algorithm, were introduced by IvD. Like McCulloch *et al.* (2000), IvD specified their model in terms of prior distributions that are set directly on the identifiable parameters, making the priors relatively easy to interpret. IvD also demonstrated that their algorithms tend to be faster than that of McCulloch *et al.* (2000). Because of their apparent advantages, IvD's algorithms have been widely used in practice to fit MNP models. Unfortunately, there are two errors in IvD's algorithms; both occur when sampling the variance-covariance matrix. First, IvD reparameterized the variables to facilitate the sampling of the variance-covariance matrix, and they made a mistake when transforming to the original parameterization. Second, when updating the variance-covariance matrix, a constraint on the matrix was overlooked. These errors can alter the stationary distribution and hence the fitted values and standard errors of the model parameters. They also can affect the efficiency of convergence. Burgette and Nordheim (2012) (henceforth, BN) modified the model of IvD by changing the manner in which unidentifiability in the scale is addressed. In particular, they fixed the trace of the variance-covariance matrix while IvD, like previous authors, fixed the first diagonal element. BN's algorithm for sampling from the posterior distribution builds upon Algorithm 1 of IvD. Thus the two errors made by IvD also affect BN's algorithm. BN made another mistake when updating the regression coefficient parameter, β. In this manuscript, we explain how to correct the errors in the algorithms of both IvD and BN, and use both a simulation study and real-data analyses to illustrate the difference between the original and the corrected algorithms in terms of their estimated posterior distributions and convergence properties. The corrections we propose are summarized in Jiao and van Dyk (2016), and will be implemented in the MNP R package.

The remainder of the manuscript is organized as follows. In Section 1.1, we introduce the notations used through the dissertation, in Section 1.2, we review the details of the MDA, ASIS, and PCG algorithms, and in Section 1.3, we describe various methods we use to compare relative efficiencies of different algorithms. In Chapter 2, we introduce the subtleties of embedding MH into PCG samplers, in Chapter 3, we construct the framework of combining different strategies into one single sampler, and in Chapter 4, we illustrate the potential of the framework of surrogate distribution strategy to further improve the performance of Gibbs-type samplers. In each of Chapters 2–4, we provide both theoretical arguments and numerical examples for illustration. We describe our applied work in supernova cosmology in Chapter 5. In a closing post-script, we correct the errors in algorithms of IvD and BN for fitting MNP models. Finally, we summarize our results and discuss future work in Chapter 7.

1.1 Markov Chain Background and Notations

1.1.1 Basic Notions of Markov Chains

We review the basic notions of Markov chains based on Robert and Casella (2004).

Definition of the Markov chain

Suppose $\{X^{(t)}, t = 0, 1, \dots\}$ is a sequence of random variables. The support of each $x^{(t)}$ is $\mathcal{X}$, an non-empty set equipped with a countably generated σ-algebra, $\mathcal{B}(\mathcal{X})$. A *transition kernel* $\mathcal{K}$ is a function defined on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$ such that

i) $\forall x \in \mathcal{X}$, $\mathcal{K}(\cdot|x)$ is a probability measure;

ii) $\forall A \in \mathcal{B}(\mathcal{X})$, $\mathcal{K}(A|\cdot)$ is measurable.

If $\mathcal{X}$ is discrete, the transition kernel $\mathcal{K}$ is simply a matrix K with elements,

$$K_{xx'} = \Pr(X^{(t+1)} = x'|X^{(t)} = x), \qquad \forall x,\, x' \in \mathcal{X}. \tag{1.1}$$

If $\mathcal{X}$ is continuous, the kernel denotes the conditional density $\mathcal{K}(x'|x)$ of the transition $\mathcal{K}(\cdot|x)$, that is, $\forall x \in \mathcal{X}$ and $\forall A \in \mathcal{B}(\mathcal{X})$,

$$\Pr(X \in A|x) = \int_A \mathcal{K}(y|x)\mathrm{d}y. \tag{1.2}$$

Given a transition kernel $\mathcal{K}$, the random variable sequence $\{X^{(t)}, t = 0, 1, \dots\}$ is a *Markov chain* if for any t, the distribution of $X^{(t+1)}$ given $X^{(t)} = x^{(t)}$, $X^{(t-1)} = x^{(t-1)}$, $\dots$, $X^{(0)} = x^{(0)}$ is the same as the distribution of $X^{(t+1)}$ given $X^{(t)} = x^{(t)}$, that is, $\forall A \in \mathcal{B}(\mathcal{X})$,

$$\Pr(X^{(t+1)} \in A | x^{(t)}, x^{(t-1)}, \dots, x^{(0)}) = \Pr(x^{(t+1)} \in A | x^{(t)}) = \int_A \mathcal{K}(x|x^{(t)}) \mathrm{d}x. \tag{1.3}$$

A chain $\{X^{(t)}, t = 0, 1, \dots\}$ is *time-homogeneous* if the distribution of $(X^{(t_1)}, \dots, X^{(t_k)})$ given $X^{(t_0)} = x^{(t_0)}$ is the same as the distribution of $(X^{(t_1 - t_0)}, \dots, X^{(t_k - t_0)})$ given $X^{(0)} = x^{(0)}$, $\forall k \in \mathbb{N}$ and each $(k+1)$-uplet $t_0 \leq t_1 \leq \cdots \leq t_k$. The structure of a time-homogeneous Markov chain is entirely determined by its transition kernel and initial state, $x^{(0)}$ (or initial distribution, μ). In this dissertation, we only consider time-homogeneous Markov chains.

Given $\mathcal{K}^1(A|x) = \mathcal{K}(A|x)$, $\forall x \in \mathcal{X}$ and $\forall A \in \mathcal{B}(\mathcal{X})$, the kernel for t ($t \in \mathbb{N}^+$) transitions is defined by

$$\mathcal{K}^t(A|x) = \int_{\mathcal{X}} \mathcal{K}^{t-1}(A|y) \mathcal{K}(y|x) \mathrm{d}y. \tag{1.4}$$

Chapman-Kolmogorov equations provide the convolution formulas for the transition kernels of the type, $\mathcal{K}^{t+s}$ ($\forall t, s \in \mathbb{N}$), that is, $\forall x \in \mathcal{X}$ and $\forall A \in \mathcal{B}(\mathcal{X})$,

$$\mathcal{K}^{t+s}(A|x) = \int_{\mathcal{X}} \mathcal{K}^s(A|y) \mathcal{K}^t(y|x) \mathrm{d}y. \tag{1.5}$$

If $\mathcal{X}$ is discrete, (1.5) is simply a matrix product.

Stability properties of Markov chains

Before recalling the properties of a Markov chain to ensure its convergence, we review some related basic notions. The first t for which the Markov chain $\{X^{(t)}, t = 0, 1, \dots\}$ enters a set $A \in \mathcal{B}(\mathcal{X})$ is denoted by $\tau_A = \inf\{t \geq 1 : X^{(t)} \in A\}$, and the *number of passages* in A is defined by $\eta_A = \sum_{t=1}^{\infty} I\{X^{(t)} \in A\}$, where $I\{\cdot\}$ is an indicator function, which equals to one if the condition in the brackets is satisfied, and equals to zero otherwise. Two related quantities, that is, the *average number of passages*, $\mathrm{E}(\eta_A|x)$, and *probability of returning to A in a finite number of steps*, $\Pr(\tau_A < +\infty|x)$, are of particular importance. A set $C \in \mathcal{B}(\mathcal{X})$ is *small* if there exists $t \in \mathbb{N}^+$ and a measure $\nu_t > 0$ such that, $\forall x \in C$ and $\forall A \in \mathcal{B}(\mathcal{X})$, $\mathcal{K}^t(A|x) \geq \nu_t(A)$. A σ-finite measure π is *invariant* for the transition kernel $\mathcal{K}(\cdot|\cdot)$ if

$$\pi(A) = \int_{\mathcal{X}} \mathcal{K}(A|x)\pi(x)\mathrm{d}x, \qquad \forall A \in \mathcal{B}(\mathcal{X}). \tag{1.6}$$

The invariant distribution is also referred to as *stationary* if π is a probability measure, since $X^{(0)} \sim \pi$ implies that $X^{(t)} \sim \pi$, for every t. The *total variation norm* for two arbitrary measures, μ_1 and μ_2, is defined by

$$||\mu_1 - \mu_2||_{\mathrm{TV}} = \sup\nolimits_{A \in \mathcal{B}(\mathcal{X})}|\mu_1(A) - \mu_2(A)|. \tag{1.7}$$

A Markov chain must enjoy good *stability* properties to guarantee an acceptance approximation of the simulated model. The first property we consider is *irreducibility*, which is the measure of the sensitivity of a Markov chain to the initial conditions ($x^{(0)}$ or μ). In the discrete case, the chain $\{X^{(t)}, t = 0, 1, \dots\}$ is *irreducible* if all states communicate, that is, if

$$\Pr(\tau_{x'} < \infty|x) > 0, \qquad \forall x,\, x' \in \mathcal{X}. \tag{1.8}$$

In the continuous case, the chain is *ξ-irreducible* for some measure ξ, if for every $A \in \mathcal{B}(\mathcal{X})$ with $\xi(A) > 0$, there exists t such that

$$\mathcal{K}^t(A|x) > 0, \qquad \forall x \in \mathcal{X}. \tag{1.9}$$

Irreducibility implies that every set $A \in \mathcal{B}(\mathcal{X})$ has a chance to be visited by the Markov chain $\{X^{(t)}, t = 0, 1, \dots\}$. However, this property is too weak to ensure that the trajectory of $\{X^{(t)}, t = 0, 1, \dots\}$ will enter A often enough. Thus, furthermore, we introduce the property of *recurrence.* In the general case, a set $A \in \mathcal{B}(\mathcal{X})$ is *recurrent* if $\mathrm{E}(\eta_A|x) = +\infty$ for every $x \in A$. The set A is *uniformly transient* if there exists a constant M such that $\mathrm{E}(\eta_A|x) < M$ for every $x \in A$; A is *transient* if there exists a countable collection of uniformly transient sets $\{B_i\}$ such that $A = \underset{i}{\cup} B_i$. A Markov chain $\{X^{(t)}, t = 0, 1, \dots\}$ is *recurrent* if

i) there exists a measure ξ such that $\{X^{(t)}, t = 0, 1, \dots\}$ is ξ-irreducible, and

ii) for every A in $\mathcal{B}(\mathcal{X})$ with $\xi(A) > 0$, $\mathrm{E}(\eta_A|x) = +\infty$ for every $x \in A$.

The chain $\{X^{(t)}, t = 0, 1, \dots\}$ is *transient* if it is ξ-irreducible and $\mathcal{X}$ is transient. Note that a ξ-irreducible Markov chain is either recurrent or transient. The property of recurrence can be strengthened by requiring not only an infinite average number of visits to every set but also an infinite number of visits to every path of a Markov chain. The stronger requirement leads to the property of *Harris recurrence.* A set $A \in \mathcal{B}(\mathcal{X})$ is *Harris recurrent* if

$$\Pr(\tau_A < +\infty|x) = 1 \tag{1.10}$$

for every $x \in \mathcal{X}$. A Markov chain $\{X^{(t)}, t = 0, 1, \dots\}$ is *Harris recurrent* if there exists a measure ξ such that $\{X^{(t)}, t = 0, 1, \dots\}$ is ξ-irreducible and every set $A \in \mathcal{B}(\mathcal{X})$ with $\xi(A) > 0$ is Harris recurrent.

An increased level of stability for a Markov chain $\{X^{(t)}, t = 0, 1, \dots\}$ is attained by the existence of an invariant measure. If $\{X^{(t)}, t = 0, 1, \dots\}$ is a recurrent Markov chain, there exists an invariant σ-finite measure, π, which is unique up to a multiplicative factor. If π is an invariant probability measure (i.e., stationary distribution), the chain is *positive*; otherwise, the chain is called *null recurrent.* The stationary distribution of a Markov chain can be identified by the *detailed balance condition.* A Markov chain $\{X^{(t)}, t = 0, 1, \dots\}$ with transition kernel $\mathcal{K}$ satisfies the *detailed balance condition* if there exists a function f such that, for every $(x, y) \in \mathcal{X}$,

$$\mathcal{K}(x|y)f(y) = \mathcal{K}(y|x)f(x). \tag{1.11}$$

If a Markov chain $\{X^{(t)}, t = 0, 1, \ldots\}$ with transition kernel $\mathcal{K}$ satisfies the detailed balance condition with f a probability density function, then

i) the density f is the stationary distribution of the chain, and

ii) the chain is *reversible*, that is, the distribution of $X^{(t+1)}$ given $X^{(t+2)} = x$ is the same as the distribution of $X^{(t+1)}$ given $X^{(t)} = x$.

With the existence of the kernel $\mathcal{K}$ and the stationary distribution f, detailed balance is equivalent to reversibility.

The behaviour of a Markov chain $\{X^{(t)}, t = 0, 1, \ldots\}$ may sometimes be restricted by deterministic constraints on the transitions from $X^{(t)}$ to $X^{(t+1)}$. These constraints are formalized in terms of *period*. Generally, a ξ-irreducible Markov chain $\{X^{(t)}, t = 0, 1, \ldots\}$ has a *cycle of length* d if there exists a small set C, an integer M, and a probability distribution ν_M such that d is the greatest common denominator (g.c.d.) of

$$\{t \geq 1 : \exists\ \delta_t > 0 \text{ such that } C \text{ is small for } \nu_t \geq \delta_t \nu_M\}. \tag{1.12}$$

The *period* of $\{X^{(t)}, t = 0, 1, \ldots\}$ is defined by the largest d satisfying (1.12) and $\{X^{(t)}, t = 0, 1, \ldots\}$ is *aperiodic* if $d = 1$.

Limiting behaviour of Markov chains

It is important to explore the limiting behaviour of a Markov chain $\{X^{(t)}, t = 0, 1, \ldots\}$. The invariant distribution π is a natural candidate for the limiting distribution of $\{X^{(t)}, t = 0, 1, \ldots\}$ because of its existence and uniqueness. We review sufficient conditions for $\{X^{(t)}, t = 0, 1, \ldots\}$ to be asymptotically distributed as its invariant distribution π without depending on the initial conditions, that is, *ergodicity.* First, if $\{X^{(t)}, t = 0, 1, \ldots\}$ is Harris positive recurrent and aperiodic, then for every initial distribution μ,

$$\lim_{t \to +\infty} \left|\left| \int \mathcal{K}^t(\cdot|x)\mu(x)\mathrm{d}x - \pi \right|\right|_{\mathrm{TV}} = 0. \tag{1.13}$$

In particular, MCMC algorithms lead to aperiodic chains, recurrence or even Harris recurrence holds for most MCMC algorithms, and moreover, the chain produced by an MCMC algorithm naturally possess an invariant distribution due to its construction.

1.1.2 Notations

We denote the generic observed data and unknown parameter by Y_{obs} and ψ, respectively. We wish to sample from the posterior distribution of ψ, i.e., $p(\psi|Y_{\text{obs}})$. The MCMC algorithms we consider are Gibbs-type samplers that rely on the complete conditional distributions of either $p(\psi|Y_{\text{obs}})$ or its multivariate marginal distributions. Thus, we divide ψ into N possibly multivariate non-overlapping components, that is, $\psi = (\psi_1, \dots, \psi_N)$ with $N \geq 2$. The Gibbs sampler updates each ψ_j in each iteration by sampling its complete conditional distribution, see Figure 1.1(a) for an example of a Gibbs sampler. In this dissertation, we only consider systematic-scan Gibbs samplers Liu *et al.* (1995). That is, in each iteration, the components of ψ are updated in a fixed ordering and each component is updated exactly once. The DA algorithm divides ψ into only two components. Owing to the different motivation of DA and following existing literature on DA, we denote these two components by θ and Y_{mis}. In the DA scenario, the distribution of interest is usually $p(\theta|Y_{\text{obs}})$ and the augmented data, Y_{mis}, is introduced to enable Gibbs sampling by iteratively updating $p(Y_{\text{mis}}|\theta, Y_{\text{obs}})$ and $p(\theta|Y_{\text{mis}}, Y_{\text{obs}})$. The MDA (Haar PX-DA), ASIS, and PCG samplers all start with a Gibbs (or DA) sampler, and transform it in some way to improve its convergence properties. Henceforth, we refer to the original Gibbs sampler as the *parent Gibbs sampler*.

We wish to obtain a Monte Carlo sample from the generic target distribution, $p(\psi)$. We achieve this by using an MCMC sampler to construct a Markov chain $\{\psi^{(t)}, t = 1, 2, \dots\}$ with the stationary distribution $\pi(\psi)$. We refer to a sampler as *proper* if it has a stationary distribution and that distribution coincides with the target, i.e., $\pi(\psi) = p(\psi)$; otherwise we call the sampler *improper*. In a typical Bayesian setting, $p(\psi) \equiv p(\psi|Y_{\text{obs}})$. A Markov chain with transition kernel $\mathcal{K}(\psi|\psi')$ means that the conditional distribution of $\psi^{(t)}$ given $\psi^{(t-1)}$ is $\mathcal{K}\left(\psi^{(t)}|\psi^{(t-1)}\right)$. In this dissertation, we focus on cases where the transition kernel is parameterized as $\mathcal{K}(\psi|\psi';\zeta)$, where ζ is some parameter that is not part of the Markov chain but affects its kernel and perhaps also its stationary distribution. Sometimes, the transition kernel may not depend on ψ' or may depend on only some components of ψ'. These two cases are indicated by $\mathcal{K}(\psi|\ ;\zeta)$ and $\mathcal{K}(\psi_1,\psi_2|\psi_2';\zeta)$, respectively. We denote the transition kernel of an MCMC sampler updating $p(\psi)$ by $\mathcal{K}(\psi|\psi')$, where the prime indicates the current state of a parameter. The joint Gibbs Markov chain, $\{\psi^{(t)}, t=1,2,\dots\}$, has stationary distribution equal to $p(\psi|Y_{\text{obs}})$, and each sub-chain $\{\psi_j^{(t)}, t=1,2,\dots\}$, can be viewed, upon convergence, as a sample from its corresponding marginal posterior distribution, i.e., $p(\psi_j|Y_{\text{obs}})$.

The transition kernel of a Gibbs sampler updating $p(\psi)$ can be written as the product of N component transition kernels of lower-dimensional Markov chains, that is,

$$\mathcal{K}(\psi|\psi') = \prod_{j=1}^{N} \mathcal{K}_j[\psi_j|\ ;\mathcal{F}_j(\psi,\psi')], \tag{1.14}$$

where $\mathcal{K}_j$ is the transition kernel for a Markov chain with stationary distribution $p(\psi_j|\mathcal{F}_j(\psi,\psi'))$. In a standard Gibbs sampler, $\mathcal{K}_j[\psi_j|\ ;\mathcal{F}_j(\psi,\psi')] = p[\psi_j|\mathcal{F}_j(\psi,\psi')]$, which does not depend on ψ_j'; $\mathcal{F}_j(\psi,\psi')$ is the sub-vector of (ψ,ψ') taking the components of ψ already updated in the current iteration and components of ψ' that are not. Suppose, for example, $\psi=\{\psi_1,\psi_2,\psi_3\}$, and the transition kernel $\mathcal{K}(\psi|\psi')$ is equal to $\prod_{j=1}^{3} p[\psi_j|\mathcal{F}_j(\psi,\psi')]$, where $\mathcal{F}_1(\psi,\psi')=(\psi_2',\psi_3')$, $\mathcal{F}_2(\psi,\psi')=(\psi_1,\psi_3')$, and $\mathcal{F}_3(\psi,\psi')=(\psi_1,\psi_2)$. If a conditional distribution is not in closed form, we update it with MH; this alters the transition kernel. Suppose, for example, we cannot sample $p(\psi_1|\psi_2,\psi_3)$ directly and use MH. In this case, we specify a jumping rule (i.e., a proposal distribution), denoted by $\mathcal{J}_{1|2,3}(\psi_1|\psi_1',\psi_2',\psi_3')$, where the subscript specifies the target conditional distribution. In the MH update, we sample $\psi_1^{\text{prop}} \sim \mathcal{J}_{1|2,3}(\psi_1|\psi_1',\psi_2',\psi_3')$ and set $\psi_1 = \psi_1^{\text{prop}}$ with probability $r = \min\left\{1,\ \dfrac{p(\psi_1^{\text{prop}}|\psi_2',\psi_3')\mathcal{J}_{1|2,3}(\psi_1'|\psi_1^{\text{prop}},\psi_2',\psi_3')}{p(\psi_1'|\psi_2',\psi_3')\mathcal{J}_{1|2,3}(\psi_1^{\text{prop}}|\psi_1',\psi_2',\psi_3')}\right\}$; otherwise the current value is retained, i.e., $\psi_1=\psi_1'$. Thus the component transition kernel of ψ_1 becomes $\mathcal{K}_1[\psi_1|\psi_1';\mathcal{F}_1(\psi,\psi')=(\psi_2',\psi_3')]$, which depends on ψ_1' because the acceptance probability of the MH algorithm involves ψ_1', and because the new iterate of

ψ_1 is set to ψ_1' if the proposal of ψ_1 is rejected. We denote this MH transition kernel by $\mathcal{M}_{1|2,3}(\psi_1|\psi_1', \psi_2', \psi_3')$, where $\mathcal{M}$ emphasizes the use of MH and the subscript specifies its stationary distribution.

1.2 Review of the Algorithms

1.2.1 Marginal Data Augmentation

Formally, for a DA scheme to be legitimate, the augmented-data model $p(Y_{\text{mis}}, Y_{\text{obs}}|\theta)$ must be a *completion* (see Definition 10.3 of Robert and Casella (2004)) of the observed-data model $p(Y_{\text{obs}}|\theta)$, i.e.,

$$\int p(Y_{\text{mis}}, Y_{\text{obs}}|\theta)\mathrm{d}Y_{\text{mis}} = p(Y_{\text{obs}}|\theta). \tag{1.15}$$

The MDA algorithm expands the augmented-data model $p(Y_{\text{mis}}, Y_{\text{obs}}|\theta)$ by introducing a working parameter α. The expanded model $\tilde{p}(\tilde{Y}_{\text{mis}}, Y_{\text{obs}}|\theta, \alpha)$ must also be a completion of $p(\theta|Y_{\text{obs}})$, that is,

$$\int \tilde{p}(\tilde{Y}_{\text{mis}}, Y_{\text{obs}}|\theta, \alpha)\mathrm{d}\tilde{Y}_{\text{mis}} = p(Y_{\text{obs}}|\theta). \tag{1.16}$$

By (1.16), α is not identifiable under the observed-data model $p(Y_{\text{obs}}|\theta)$. A general method for introducing α into an augmented-data model is to construct a one-to-one and differentiable mapping between Y_{mis} and $\tilde{Y}_{\text{mis}}$. Specifically, following Liu and Wu (1999), we assume that

LW-1: the working parameter α indexes a "data-transformation" mechanism with $\tilde{Y}_{\text{mis}} = \mathcal{G}_\alpha(Y_{\text{mis}})$, that is, for any fixed α, $\mathcal{G}_\alpha$ induces a one-to-one and differentiable mapping (i.e., a diffeomorphism) between Y_{mis} and $\tilde{Y}_{\text{mis}}$. (Note that Liu and Wu (1999) defined the "data transformation" via $Y_{\text{mis}} = t_\alpha(\tilde{Y}_{\text{mis}})$. Thus their t_α is $\mathcal{G}_\alpha^{-1}$ here.)

Meng and van Dyk (1999) pointed out that $\mathcal{G}$ can operate on the parameter θ as well, but in this dissertation we focus on the case where $\mathcal{G}$ depends only on α. For each $\mathcal{G}$, there typically exists a scalar e such that $\mathcal{G}_e(Y_{\text{mis}}) = Y_{\text{mis}}$.

When the prior distribution of the working parameter α is proper, the MDA algorithm proceeds as

Step 1: $(\tilde{Y}_{\text{mis}}^{(t+1)}, \alpha^{\star}) \sim \tilde{p}(\tilde{Y}_{\text{mis}}, \alpha | \theta^{(t)}, Y_{\text{obs}})$, (Sampler 1.1)

Step 2: $(\theta^{(t+1)}, \alpha^{(t+1)}) \sim \tilde{p}(\theta, \alpha | \tilde{Y}_{\text{mis}}^{(t+1)}, Y_{\text{obs}})$; set $Y_{\text{mis}}^{(t+1)} = \mathcal{G}^{-1}_{\alpha^{(t+1)}}\big(\tilde{Y}_{\text{mis}}^{(t+1)}\big)$.

In Sampler 1.1, α is sampled in both steps and the first update of α is not part of the final output. We refer to updates which are sampled but not included in the output of each iteration as *intermediate quantities*, following van Dyk and Park (2008), and indicate them with a "$\star$" in their superscript. We refer to Sampler 1.1 as a *collapsed DA sampler*, since it can be viewed as sampling the complete conditional distributions of $\tilde{p}(\tilde{Y}_{\text{mis}}, \theta | Y_{\text{obs}}) = \int p(\tilde{Y}_{\text{mis}}, \theta, \alpha | Y_{\text{obs}}) \mathrm{d}\alpha$. In this regard, Sampler 1.1 is equivalent to a standard DA sampler constructed from $\tilde{p}(\tilde{Y}_{\text{mis}}, \theta | Y_{\text{obs}})$. Thus the marginal Markov chain, $\{\theta^{(t)}, t = 0, 1, \dots\}$, produced by Sampler 1.1 is reversible with $p(\theta | Y_{\text{obs}})$ as its stationary distribution (Liu *et al.*, 1994). Collapsing α out increases the (expected) variance of the conditional distributions used in Sampler 1.1. This allows for bigger jumps and thus faster convergence, see Meng and van Dyk (1999) and van Dyk and Meng (2001) for both theoretical and practical illustrations.

Meng and van Dyk (1999) showed that the rate of convergence of MDA depends on the choice of prior on α and argued for optimizing the rate of convergence within a class of certain priors. In some cases, the optimal choice is an improper prior on α. Because α does not appear in the likelihood of (1.16), however, an improper prior results in an improper posterior. Thus we cannot marginalize α out in Step 1 of Sampler 1.1. Suppose we instead condition on the previous iteration of α when sampling $\tilde{Y}_{\text{mis}}$:

Step 1: $\tilde{Y}_{\text{mis}}^{(t+1)} \sim \tilde{p}(\tilde{Y}_{\text{mis}} | \theta^{(t)}, \alpha^{(t)}, Y_{\text{obs}})$, (Sampler 1.2)

Step 2: $(\theta^{(t+1)}, \alpha^{(t+1)}) \sim \tilde{p}(\theta, \alpha | \tilde{Y}_{\text{mis}}^{(t+1)}, Y_{\text{obs}})$; set $Y_{\text{mis}}^{(t+1)} = \mathcal{G}^{-1}_{\alpha^{(t+1)}}\big(\tilde{Y}_{\text{mis}}^{(t+1)}\big)$.

With a proper prior on α, Sampler 1.2 is also a standard DA sampler but with stationary distribution $\tilde{p}(\tilde{Y}_{\text{mis}}, \alpha, \theta | Y_{\text{obs}})$. We can use an improper prior for α if we can verify that the marginal chain $\{\theta^{(t)}, t = 0, 1, \ldots\}$ is a positive recurrent Markov chain with $p(\theta|Y_{\text{obs}})$ as its stationary distribution, despite of the fact that the joint chain $\{(\alpha^{(t)}, \theta^{(t)}), t = 0, 1, \ldots\}$ is null recurrent. This occurs when the two conditions in Lemma 1 of van Dyk and Meng (2001) hold, that is, letting $\tilde{p}_\infty(\alpha, \theta|\tilde{Y}_{\text{mis}}, Y_{\text{obs}})$ be the joint distribution of θ and α conditioning on $\tilde{Y}_{\text{mis}} = \mathcal{G}_\alpha(Y_{\text{mis}})$ under an improper working prior $p_\infty(\alpha)$,

vDM-1: there exists a sequence of proper working priors indexed by m, i.e., $\{p_m(\alpha)\}$ and an m_∞ such that the corresponding $\tilde{p}_m(\alpha, \theta|\tilde{Y}_{\text{mis}}, Y_{\text{obs}})$converges to $\tilde{p}_\infty(\alpha, \theta|\tilde{Y}_{\text{mis}}, Y_{\text{obs}})$ as $m \to m_\infty$;

vDM-2: the conditional distribution of the expanded posterior with the improper prior, $\tilde{p}(\theta|\mathcal{G}_\alpha(Y_{\text{mis}}), Y_{\text{obs}})$, is free of α.

Under Conditions vDM-1 and vDM-2, the sub-chain $\{\theta^{(t)}, t = 0, 1, \ldots\}$ induced by Sampler 1.2 with the working prior $p_\infty(\alpha)$ is Markovian, and its transition kernel $p_\infty(\theta^{(t+1)}|\theta^{(t)}, \alpha^{(t)})$ is the limit of the kernel $p_m(\theta^{(t+1)}|\theta^{(t)})$ induced by Sampler 1.1 with the working prior $p_m(\alpha)$ as $m \to m_\infty$, see Lemma 1 of van Dyk and Meng (2001). Thus Conditions vDM-1 and vDM-2 are sufficient to guarantee that $\{\theta^{(t)}, t = 0, 1, \ldots\}$ induced by Sampler 1.2 under $p_\infty(\alpha)$ is a positive recurrent reversible Markov chain with $p(\theta|Y_{\text{obs}})$ as its unique stationary distribution, see Theorem 2 of Meng and van Dyk (1999).

Liu and Wu (1999) explored the benefits of PX-DA samplers with Haar measure working priors which are accessible when the set of transformations indexed by α, i.e., $\{\mathcal{G}_\alpha^{-1}, \alpha \in \mathcal{A}\}$, forms a locally compact group. We briefly review the basic knowledge of group structure and Haar measure below. See Liu and Wu (1999) and Hobert and Marchev (2008) for more details.

A set $\mathcal{A}$ is a *group* with respect to an operator "$\cdot$" if i) for all $\alpha \in \mathcal{A}$, $\beta \in \mathcal{A}$, $\alpha \cdot \beta \in \mathcal{A}$; ii) there exist an identity element $e \in \mathcal{A}$ so that $\alpha \cdot e = e \cdot \alpha = \alpha$, for all $\alpha \in \mathcal{A}$; iii) for all $\alpha \in \mathcal{A}$, we can find a unique $\alpha^{-1} \in \mathcal{A}$ so that $\alpha \cdot \alpha^{-1} = \alpha^{-1} \cdot \alpha = e$. Following Liu and Wu (1999), we assume that $\{\mathcal{G}_\alpha^{-1}, \alpha \in \mathcal{A}\}$ has the same group structure as $\mathcal{A}$, that is, $\mathcal{G}_e^{-1}$ is an identity mapping, and for all $\alpha \in \mathcal{A}$, $\beta \in \mathcal{A}$, $\mathcal{G}_\alpha^{-1}(\mathcal{G}_\beta^{-1}(Y_{\text{mis}})) = \mathcal{G}_{\alpha\cdot\beta}^{-1}(Y_{\text{mis}})$. Group $\mathcal{A}$ is called a *locally compact group* or *topological group*, if topologically $\mathcal{A}$ is locally compact and the operations $(\alpha, \beta) \to \alpha \cdot \beta$ and $\alpha \to \alpha^{-1}$ are continuous. If Group $\mathcal{A}$ is finite, then it is automatically a locally compact group. If the operations are analytic, $\mathcal{A}$ is called a *Lie group*.

For any measurable subset $A \subset \mathcal{A}$ and any element $\beta \in \mathcal{A}$, $A \cdot \beta$ defines a subset of $\mathcal{A}$ resulting from "operating" on every element of A by β. A *right Haar measure* $H(\mathrm{d}\alpha)$ on Group $\mathcal{A}$ is a measure that is invariant under the group operation on the right side, that is,

$$H(A) = \int_A H(\mathrm{d}\alpha) = \int_{A\cdot\beta} H(\mathrm{d}\alpha) = H(A \cdot \beta), \, \forall \beta \in \mathcal{A}, \tag{1.17}$$

for all measurable subset $A \subset \mathcal{A}$. A *left Haar measure* can be defined in a similar manner. Under mild conditions, the right (or left) Haar measure is unique up to a positive constant, see Rao (1987). When the right Haar measure of $\mathcal{A}$ is also its left Haar measure, $\mathcal{A}$ is called *unimodular*. When $\mathcal{A}$ is compact or abelian (i.e., $\alpha \cdot \beta = \beta \cdot \alpha$; e.g., the translation and scale groups), its right Haar measure is unimodular (Rao, 1987). If $\mathcal{A}$ is a compact group, its unimodular Haar measure is the uniform probability measure. When $\mathcal{A}$ is a translation group (e.g., $\mathcal{G}_\alpha^{-1}(Y_{\text{mis}}) = Y_{\text{mis}} + \alpha$), the unimodular Haar measure for $\mathcal{A}$ is the Lebesgue measure. When $\mathcal{A}$ is a scale group (e.g., $\mathcal{G}_\alpha^{-1}(Y_{\text{mis}}) = \alpha Y_{\text{mis}}$), the unimodular Haar measure for $\mathcal{A}$ is proportional to $|\alpha|^{-1}\mathrm{d}\alpha$. Following Liu and Wu (1999), we assume that a density $H(\alpha)$ exists for the unimodular Haar measure with respect to the Lebesgue or counting measure, i.e., $H(\mathrm{d}\alpha) = H(\alpha)\mathrm{d}\alpha$.

The Haar PX-DA proceeds by

Step 1: $Y_{\text{mis}}^\star \sim p(Y_{\text{mis}}|\theta^{(t)}, Y_{\text{obs}})$, (Sampler 1.21)

Step 2: $\alpha^\star \sim \tilde{p}(\alpha|Y_{\text{mis}}^\star, Y_{\text{obs}})$; set $Y_{\text{mis}}^{(t+1)} = \mathcal{G}_{\alpha^\star}^{-1}(Y_{\text{mis}}^\star)$,

Step 3: $\theta^{(t+1)} \sim p(\theta|Y_{\text{mis}}^{(t+1)}, Y_{\text{obs}})$, where Step 2 imposes a transition from $Y_{\text{mis}}^{\star}$ to $Y_{\text{mis}}^{(t+1)}$. Let $\mathcal{Z}$ be the state space of Y_{mis}. For a fixed Y_{mis}, Liu and Wu (1999) defined the set $\{\mathcal{G}_{\alpha}^{-1}(Y_{\text{mis}}), \alpha \in \mathcal{A}\} \subset \mathcal{Z}$ as an *orbit*; $Y_{\text{mis}}^{(t+1)}$ and $Y_{\text{mis}}^{\star}$ lie on the same orbit if and only if there exists a unique $\alpha \in \mathcal{A}$ so that $Y_{\text{mis}}^{(t+1)} = \mathcal{G}_{\alpha}^{-1}(Y_{\text{mis}}^{\star})$. Different orbits do not intersect, and $\mathcal{Z}$ can be regarded as the union of all of the orbits, each of which has the same structure as $\mathcal{A}$. The set of orbits can be represented by a smooth cross-section $\mathcal{Q}$, which is defined by a subset of $\mathcal{Z}$ that intersects with (almost) every orbit exactly once. See Liu and Wu (1999) for illustrative examples of orbits and cross-sections. With the cross-section $\mathcal{Q}$, any $Y_{\text{mis}} \in \mathcal{Z}$ can be represented together by its orbit $r \in \mathcal{Q}$ and its position $\beta \in \mathcal{A}$ on the orbit. Step 2 of Sampler 1.2l is equivalent to generating a new position of Y_{mis} conditioning on its orbit. To prove the propriety of this step, Liu and Wu (1999) stated that the following condition should hold, that is,

LW-2: The group $\mathcal{A}$ is locally compact and has a unimodular Haar measure $H(\alpha)\mathrm{d}\alpha$. There exists a smooth cross-section $\mathcal{Q} \subset \mathcal{Z}$, and the mapping Z: $Z(\beta, r) = \mathcal{G}_{\beta}^{-1}(r)$ for $\mathcal{A} \times \mathcal{Q} \to \mathcal{Z}$ is one-to-one and continuously differentiable, that is, a diffeomorphism.

With the mapping Z, we can reparameterize each $Y_{\text{mis}} \in \mathcal{Z}$ as that there exists a pair of (β, r), such that $Y_{\text{mis}} = Z(\beta, r)$. For all $\alpha \in \mathcal{A}$ and $Y_{\text{mis}} = Z(\beta, r)$, $\mathcal{G}_{\alpha}^{-1}(Y_{\text{mis}}) = \mathcal{G}_{\alpha}^{-1}(Z(\beta, r)) = Z(\alpha \cdot \beta, r)$. Thus Sampler 1.2l can simply be described by

$$\theta^{(t)} \to (\beta, r) \to (\alpha^{\star} \cdot \beta, r) \to \theta^{(t+1)}. \tag{1.18}$$

Liu and Wu (1999) proved that under Conditions LW-1 and LW-2, Step 2 of Sampler 1.2l keeps $p(Y_{\text{mis}}|Y_{\text{obs}})$ invariant, that is, if $Y_{\text{mis}}^{\star} \sim p(Y_{\text{mis}}|Y_{\text{obs}})$, then $Y_{\text{mis}}^{(t+1)} \sim p(Y_{\text{mis}}|Y_{\text{obs}})$. Thus the Haar PX-DA sampler, i.e., Sampler 1.2l, is proper, that is, it maintains the target stationary distribution $p(Y_{\text{mis}}, \theta|Y_{\text{obs}})$, see Theorem 4 and Corollary 1 of Liu and Wu (1999). As mentioned above, Hobert and Marchev (2008) verified that when the group structure is present, the Haar PX-DA is equivalent to the MDA algorithm under the limiting improper working prior, i.e., Sampler 1.2. In this dissertation, we confine our attention to Haar PX-DA algorithms to obtain optimal computational efficiency.

1.2.2 Ancillarity-Sufficiency Interweaving Strategy

As described above, ASIS utilizes two special DA schemes: a sufficient augmentation, $Y_{\text{mis,S}}$, such that $p(Y_{\text{obs}}|Y_{\text{mis,S}}, \theta)$ is free of the parameter θ, and an ancillary augmentation, $Y_{\text{mis,A}}$, such that $p(Y_{\text{mis,A}}|\theta)$ does not depend on θ. Normally, given θ, $Y_{\text{mis,S}}$ is related to $Y_{\text{mis,A}}$ via a one-to-one mapping, $\mathcal{H}_\theta$, that is, $Y_{\text{mis,A}} = \mathcal{H}_\theta(Y_{\text{mis,S}})$ (but see Yu and Meng (2011) for an exception). We assume $\mathcal{H}_\theta$ is differentiable when $Y_{\text{mis,S}}$ is continuous. ASIS forms a new sampler by interweaving a DA algorithm constructed with a sufficient augmentation into one constructed with an ancillary augmentation:

Step 1: $Y^\star_{\text{mis,S}} \sim p(Y_{\text{mis,S}}|\theta^{(t)}, Y_{\text{obs}})$, (Sampler 1.3)

Step 2: $\theta^\star \sim p(\theta|Y^\star_{\text{mis,S}}, Y_{\text{obs}})$; set $Y^{(t+1)}_{\text{mis,A}} = \mathcal{H}_{\theta^\star}(Y^\star_{\text{mis,S}})$,

Step 3: $\theta^{(t+1)} \sim p(\theta|Y^{(t+1)}_{\text{mis,A}}, Y_{\text{obs}})$; set $Y^{(t+1)}_{\text{mis,S}} = \mathcal{H}^{-1}_{\theta^{(t+1)}}(Y^{(t+1)}_{\text{mis,A}})$.

Here, we assume the joint distribution, $p(Y_{\text{mis,S}}, Y_{\text{mis,A}}, \theta|Y_{\text{obs}})$, is well defined with marginal distributions $p(Y_{\text{mis,S}}, \theta|Y_{\text{obs}})$, $p(Y_{\text{mis,S}}, Y_{\text{mis,A}}|Y_{\text{obs}})$, and $p(Y_{\text{mis,A}}, \theta|Y_{\text{obs}})$. In particular, $p(Y_{\text{mis,S}}, \theta|Y_{\text{obs}})$ and $p(Y_{\text{mis,A}}, \theta|Y_{\text{obs}})$ share the same marginal distribution for θ, i.e., $p(\theta|Y_{\text{obs}})$. (This is a necessary consequence of both augmentation schemes being legitimate DA schemes.) Step 2 of Sampler 1.3 is equivalent to sampling $Y_{\text{mis,A}}$ from $p(Y_{\text{mis,A}}|Y^\star_{\text{mis,S}}, Y_{\text{obs}})$. Thus, the transition kernel of θ under ASIS is

$$\begin{aligned}&\mathcal{K}(\theta^{(t+1)}|\theta^{(t)}) = \\ &\int\int p(\theta^{(t+1)}|Y^{(t+1)}_{\text{mis,A}}, Y_{\text{obs}})p(Y^{(t+1)}_{\text{mis,A}}|Y^\star_{\text{mis,S}}, Y_{\text{obs}})p(Y^\star_{\text{mis,S}}|\theta^{(t)}, Y_{\text{obs}})\text{d}Y^\star_{\text{mis,S}}\text{d}Y^{(t+1)}_{\text{mis,A}}.\end{aligned} \tag{1.19}$$

It is easy to verify that the stationary distribution of $\mathcal{K}(\theta^{(t+1)}|\theta^{(t)})$ is the target marginal, $p(\theta|Y_{\text{obs}})$, see Yu and Meng (2011) for details.

Yu and Meng (2011) found that Steps 1-3 of Sampler 1.3 can be regarded as sampling $(\theta, Y_{\text{mis,S}})$ along different directions. ASIS selects a particular combination of sampling

(a) Gibbs Sampler	(b) Marginalization	(c) Permute	(d) Trim
1. $p(\psi_1 \vert \psi_2', \psi_3', \psi_4')$	1. $p(\psi_1, \psi_3^\star \vert \psi_2', \psi_4')$	1. $p(\psi_2 \vert \psi_1', \psi_3', \psi_4')$	1. $p(\psi_2 \vert \psi_1', \psi_3', \psi_4')$
2. $p(\psi_2 \vert \psi_1, \psi_3', \psi_4')$	2. $p(\psi_2 \vert \psi_1, \psi_3^\star, \psi_4')$	2. $p(\psi_1, \psi_3^\star \vert \psi_2, \psi_4')$	2. $p(\psi_1 \vert \psi_2, \psi_4')$
3. $p(\psi_3, \psi_4 \vert \psi_1, \psi_2)$	3. $p(\psi_3, \psi_4 \vert \psi_1, \psi_2)$	3. $p(\psi_3, \psi_4 \vert \psi_1, \psi_2)$	3. $p(\psi_3, \psi_4 \vert \psi_1, \psi_2)$

Figure 1.1: An example of using three tools to transform a Gibbs sampler into a proper PCG sampler. The parent Gibbs sampler appears in (a). The sampler in (b) updates ψ_3 rather than conditioning on it in Step 1. The steps of this sampler are permuted in (c) to make the draw of $\psi_3^\star$—in Step 2 of (c)—to be redundant. Trimming $\psi_3^\star$, we obtain the PCG sampler in (d).

directions such that the resulting sampler substantially outperforms both parent DA samplers, while the additional computational expense is small, see Yu and Meng (2011).

1.2.3 Partially Collapsed Gibbs Sampling

The PCG sampler replaces some of the complete conditional distributions of an ordinary Gibbs sampler with the corresponding conditionals of marginal distributions of the target joint posterior distribution (van Dyk and Park, 2008). This generally leads to a larger variance of the conditional distribution, bigger average jump sizes, and thus improved convergence relative to the parent Gibbs sampler.

One must be careful, however, to ensure that the target distribution is maintained; van Dyk and Park (2008) provided three tools for transforming a Gibbs sampler into a PCG sampler, while maintaining the target distribution, specifically, *marginalization*, *permutation* and *trimming*. Suppose, for example, $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)$ and we wish to sample $p(\psi)$. The marginalization stage replaces one or more steps of a Gibbs sampler with steps that update rather than condition on some components of ψ, see Step 1 in Figure 1.1(b). This results in the same component being sampled in two or more steps. The steps of the marginalized sampler are then permuted to make some intermediate quantities *redundant*, which means they are neither subsequently conditioned upon nor part of the final output, see ψ_3 in Step 2 of Figure 1.1(c). Finally, we trim the redundant quantities; this results in one or more steps that samples from a conditional distribution of a marginal distribution of $p(\psi)$, such as Step 2 in Figure 1.1(d). We refer to such steps as *reduced steps*. We must guarantee only redundant quantities are trimmed, because trimming intermediate quantities conditioned upon in subsequent steps can alter the stationary distribution. Since all three stages preserve the stationary distribution of the parent Gibbs sampler, the resulting PCG sampler maintains the target stationary distribution. Marginalization can significantly improve the rate of convergence, while permutation typically has a minor effect and trimming has no effect (van Dyk and Park, 2008). PCG samplers generally exhibit better and often much better convergence properties than their parent Gibbs samplers. More details and examples can be found in van Dyk and Park (2008) and Park and van Dyk (2009).

In some cases, a PCG sampler is simply a blocked or collapsed version of its parent Gibbs sampler. PCG is a more general strategy, however, because in some cases, PCG samplers involve *incompatible conditional distributions*, that is, conditionals for which there is no corresponding joint distribution, e.g., the PCG sampler in Figure 1.1(d). Unlike the Gibbs sampler, permuting the steps of such a PCG sampler may alter its stationary distribution. Suppose, for example, we obtain $(\psi_1', \psi_2, \psi_3', \psi_4')$ from $p(\psi_1, \psi_2, \psi_3, \psi_4)$ and update ψ_1 according to Step 2 of the PCG sampler in Figure 1.1(d). The joint distribution of $(\psi_1, \psi_2, \psi_3', \psi_4')$ would be

$$\int p(\psi_1|\psi_2, \psi_4')p(\psi_1', \psi_2, \psi_3', \psi_4')d\psi_1' = p(\psi_1, \psi_2, \psi_4')p(\psi_3'|\psi_2, \psi_4'), \tag{1.20}$$

which is different from the target in that ψ_1 and ψ_3' are conditionally independent. However, the joint distribution of ψ_1 and ψ_2 in (1.20) is the target posterior distribution and Step 3 of the PCG sampler conditions only on (ψ_1, ψ_2). Thus after Step 3, the joint distribution of the parameters is again the target. Thus after a cyclic permutation of steps of the PCG sampler in Figure 1.1(d), the sampler ending with either Step 1 or 3 is proper, whereas ending with Step 2 is improper. With non-cyclic permutation, the stationary distribution is unknown.

1.3 Tools for Comparing Efficiencies of Algorithms

In this manuscript, we typically need to compare the convergence efficiencies of different algorithms. The tools we use to make comparisons are i) visualizing methods: time-series and autocorrelation plots of parameters (the lag-k autocorrelation of a time series $\{\psi^{(t)}, t = 0, 1, \dots\}$ is defined by $\text{corr}(\psi^{(t)}, \psi^{(t-k)}) = \gamma_k/\gamma_0$, where $\gamma_k = \text{cov}(\psi^{(t)}, \psi^{(t-k)})$ and γ_0 is the unconditional variance of the time series), ii) effective sample size per second of parameters, and iii) the cyclic-permutation bound of a sampler.

1.3.1 Effective Sample Size

The effective sample size (ESS) is defined as

$$\text{ESS}(\psi) = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t(\psi)}, \tag{1.21}$$

where T is the total posterior sample size and $\rho_t(\psi)$ is the lag-t autocorrelation of the parameter ψ. The ESS approximates the size of an independent sample with equivalent information in terms of the Monte Carlo variance of the sample mean, and is indicative of how well the chain mixes, see Kass *et al.* (1998) and Liu (2001). We use the function "effectiveSize" in the R package `coda` to estimate the ESS. To account for computational time, we use ESS/sec (i.e., ESS/(CPU time)) of a parameter to compare different algorithms. The larger the ESS/sec, the more efficient is the convergence.

1.3.2 Cyclic-Permutation Bound

We use $L^2(p)$ to denote the set of all functions $h(\psi)$ such that $\int h^2(\psi)p(\psi)\text{d}\psi < \infty$. This set is a Hilbert space with inner product $< h, g >= \text{E}_p[h(\psi)g(\psi)]$. For a general Markov chain, $\{\psi^{(t)}, t = 0, 1, \dots\}$ with transition kernel $\mathcal{K}(\psi|\psi')$, we define the forward operator on $L^2(p)$ of $\mathcal{K}$ by

$$Ph(\psi') = \int h(\psi)\mathcal{K}(\psi|\psi')\mathrm{d}\psi. \tag{1.22}$$

Denote $L_0^2(p) = \{h \in L^2(p), \mathrm{E}_p[h(\psi)] = 0, \mathrm{var}_p[h(\psi)] < \infty\}$, which is also a Hilbert space with the same inner product and forward operator as $L^2(p)$. For $h \in L_0^2(p)$, $||h||^2 = \mathrm{var}_p(h)$. We define P_0 the forward operator on $L_0^2(p)$ induced by P. The norm of this forward operator is $||P_0|| \equiv \sup_{h\in L_0^2(p)} \frac{||P_0 h||}{||h||}$, and it is easy to verify that

$$||P_0|| \equiv \sup_{h\in L_0^2(p), \mathrm{var}_p(h)=1} \left\{\mathrm{var}_p\left[\mathrm{E}(h(\psi^{(1)})|\psi^{(0)})\right]\right\}^{1/2} = \rho(\psi^{(1)}, \psi^{(0)}), \tag{1.23}$$

where $\rho(\psi^{(1)}, \psi^{(0)})$ represents the maximum correlation of $\psi^{(1)}$ and $\psi^{(0)}$. The spectral radius of P_0, $r(P_0)$, which is defined by

$$r(P_0) = \lim_{t\to\infty} ||P_0^t||^{1/t}, \tag{1.24}$$

is called the *convergence rate* of the Markov chain, and $r(P_0) \leq ||P_0||$. Smaller convergence rate implies faster convergence to the stationary distribution of the Markov chain. Although Liu *et al.* (1994) provided the explicit form of $r(P_0)$ for two-step Gibbs samplers, it is typically difficult to derive $r(P_0)$ for Gibbs samplers with more than two steps. Thus van Dyk and Park (2008) introduced an upper-bound of $r(P_0)$, named by *cyclic-permutation bound*. Consider an N-step Gibbs sampler and define a j-step-lagged Gibbs sampler, for $j = 0, \ldots, N-1$, by the Gibbs sampler starting with Step $j+1$ obtained via cyclically permuting the steps of the original sampler. The j-step-lagged Gibbs samplers ($j = 0, \ldots, N-1$) have the same spectral radius, but possibly different norms (or equivalently, maximum correlations). We denote the norm of the forward operator corresponding to the j-step-lagged Gibbs sampler by γ_j, and the cyclic-permutation bound is defined by $\min_{j\in\{0,\ldots,N-1\}}\{\gamma_j\}$. The cyclic-permutation bound is much easier to handle than the convergence rate. Thus we also use this quantity to analyse the convergence properties of Gibbs-type samplers in this dissertation. Smaller cyclic-permutation bounds indicate better convergence properties.

2

Embedding the MH Algorithm into PCG Samplers

In this chapter, we introduce difficulties that may arise when using MH updates within a PCG sampler and develop a general strategy for using such updates while maintaining the target stationary distribution. We begin in Section 2.1 with a motivating example from our applied work in X-ray astronomy to illustrate the complications that arise when MH is introduced into PCG samplers and set the stage for the methodological and theoretical contributions of our strategy. The MH within PCG sampler is introduced in Section 2.2 along with methods for ensuring that its stationary distribution is the target distribution and two strategies for implementing the sampler while maintaining this target. Theoretical arguments are presented in Section 2.2.3 that aim to guide the choice between different implementations of the MH within PCG sampler. The proposed methods and theoretical results are illustrated in Section 2.3 with several examples, including the spectral analysis model in astrophysics, the hierarchical Gaussian model in supernova cosmology, and a factor analysis model. Final discussion appears in Section 2.4, concluding our strategy of embedding MH steps into PCG samplers and shedding light on its connection with the combining strategy, which to be introduced in Chapter 3.

2.1 Motivating Example

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

$$Y_i \overset{\text{ind}}{\sim} \text{Poisson}\left\{\Lambda_i = \alpha({E_i}^{-\beta} + \gamma I\{i = \mu\})e^{-\phi/E_i}\right\}, \text{ for } i = 1, \ldots, n, \tag{2.1}$$

where Y_i is the count in bin i; α, β, γ, μ, and ϕ are model parameters; $I\{\cdot\}$ is the indicator function; and n is the number of energy bins. The $\alpha {E_i}^{-\beta}$ term in (2.1) is a *continuum*—a smooth term that extends over a wide range of energies. The $\alpha\gamma I\{i = \mu\}$ term is an *emission line*—a sharp narrow term that describes a distinct aberration from the continuum. The emission line in (2.1) is very narrow in that it is contained entirely in one energy bin. The parameters of the continuum and emission line describe the composition, temperature, and general physical environment of the source. The factor $e^{-\phi/E_i}$ in (2.1) accounts for absorption—lower energy photons are more likely to be absorbed by inter-stellar material and not be recorded by the detector. A typical spectral model might contain multiple summed continua and emission lines. We use a simple example here to focus attention on computational issues. We assume that α, β, γ, μ, and ϕ are *a priori* independent and that μ is *a priori* uniform on $\{1, \ldots, n\}$, while the

Sampler 2.1

1. $\mathcal{M}_{\mu|Y,\alpha,\beta,\gamma,\phi}(\mu|Y,\alpha',\beta',\gamma',\mu',\phi')$
2. $p(Y_L|Y,\alpha',\beta',\gamma',\mu',\phi')$
3. $p(\alpha|Y,Y_L,\beta',\gamma',\mu,\phi')$
4. $\mathcal{M}_{\beta|Y,Y_L,\alpha,\gamma,\mu,\phi}(\beta|Y,Y_L,\alpha,\beta',\gamma',\mu,\phi')$
5. $p(\gamma|Y,Y_L,\alpha,\beta,\mu,\phi')$
6. $\mathcal{M}_{\phi|Y,Y_L,\alpha,\beta,\gamma,\mu}(\phi|Y,Y_L,\alpha,\beta,\gamma,\mu,\phi')$

Sampler 2.2

1. $\mathcal{M}_{\mu|Y,\beta,\gamma,\phi}(\mu|Y,\beta',\gamma',\mu',\phi')$
2. $\mathcal{M}_{\phi|Y,\beta,\gamma,\mu}(\phi|Y,\beta',\gamma',\mu,\phi')$
3. $\mathcal{M}_{\beta|Y,\gamma,\mu,\phi}(\beta|Y,\beta',\gamma',\mu,\phi)$
4. $p(\alpha|Y,\beta,\gamma',\mu,\phi)$
5. $p(Y_L|Y,\alpha,\beta,\gamma',\mu,\phi)$
6. $p(\gamma|Y,Y_L,\alpha,\beta,\mu,\phi)$

Sampler 2.3

1. $\mathcal{M}_{\mu|Y,\beta,\gamma,\phi}(\mu|Y,\beta',\gamma',\mu',\phi')$
2. $\mathcal{M}_{\beta,\phi|Y,\gamma,\mu}(\beta,\phi|Y,\beta',\gamma',\mu,\phi')$
3. $p(\alpha|Y,\beta,\gamma',\mu,\phi)$
4. $p(Y_L|Y,\alpha,\beta,\gamma',\mu,\phi)$
5. $p(\gamma|Y,Y_L,\alpha,\beta,\mu,\phi)$

Sampler 2.4

1. $\mathcal{M}_{\mu|Y,\beta,\gamma,\phi}(\mu|Y,\beta',\gamma',\mu',\phi')$
2. $\mathcal{M}_{\phi|Y,\beta,\gamma,\mu}(\phi|Y,\beta',\gamma',\mu,\phi')$
3. $\mathcal{M}_{\alpha,\beta|Y,\gamma,\mu,\phi}(\alpha,\beta|Y,\alpha',\beta',\gamma',\mu,\phi)$
4. $p(Y_L|Y,\alpha,\beta,\gamma',\mu,\phi)$
5. $p(\gamma|Y,Y_L,\alpha,\beta,\mu,\phi)$

Figure 2.1: Samplers 2.1–2.4. The four samplers are all MH within PCG samplers for fitting the spectral model in (2.1). Sampler 2.1 (top-left) is the proper sampler with the lowest degree of partial collapsing. Sampler 2.2 (top-right) is another proper sampler but with a higher degree of partial collapsing. Samplers 2.3 (bottom-left), which blocks Steps 2 and 3 of Sampler 2.2 into a single MH step, is the proper sampler with the highest degree of partial collapsing. Sampler 2.4 (bottom-right) blocks Steps 3 and 4 of Sampler 2.2. Unlike Sampler 2.3, however, Sampler 2.4 is improper.

other four parameters are *a priori* uniform on the positive real line $\mathbb{R}^+$. In practice, we do not observe $Y = (Y_1, \ldots, Y_n)$ directly because photon counts are subject to stochastic censoring, misclassification, and background contamination, see Lee *et al.* (2011) and van Dyk and Jiao (2015) for the complete spectral analysis model considering all these factors. In this dissertation, for simplicity, we assume that Y is observed directly and ignore censoring, misclassification, and background contamination.

The model in (2.1) is a finite mixture model and can be fitted via the standard data augmentation scheme that sets $Y_i = Y_{iC} + Y_{iL}$, where $Y_{iC} \overset{\text{ind}}{\sim} \text{Poisson}\left(\alpha E_i^{-\beta} e^{-\phi/E_i}\right)$ and $Y_{iL} \overset{\text{ind}}{\sim} \text{Poisson}\left(\alpha\gamma I\{i=\mu\} e^{-\phi/E_i}\right)$, are the photon counts in bin i generated from the continuum and emission line, respectively. We consider samplers that target $p(Y_L, \alpha, \beta, \gamma, \mu, \phi|Y)$ rather than $p(\alpha, \beta, \gamma, \mu, \phi|Y)$ because introduction of the augmented data,

$Y_C = (Y_{1C}, \ldots, Y_{nC})$ and $Y_L = (Y_{1L}, \ldots, Y_{nL})$, highly simplifies the complete conditional distributions. We begin with the six-step Gibbs sampler which updates each of Y_L, α, β, γ, μ, and ϕ from its complete conditional distribution iteratively, see Figure 2.3(a). Three of its steps need MH updates. Because Y_L completely specifies the line location μ, $\mathrm{var}_p(\mu|Y_L) = 0$. Thus this MH within Gibbs sampler is not irreducible, and the subchain for μ does not move from its starting value $\mu^{(0)}$, for any choice of $\mu^{(0)}$. We solve this problem by updating μ without conditioning on Y_L, and obtain an MH within PCG sampler, i.e., Sampler 2.1, given in the top-left panel of Figure 2.1. Sampler 2.2 in the top-right panel of Figure 2.1 is another MH within PCG sampler but with a higher degree of partial collapsing, i.e., more quantities are marginalized out in steps of Sampler 2.2 than in Sampler 2.1. Sampler 2.2 marginalizes X_L out when updating not only μ, but also α, β, and ϕ, and marginalizes α out of its first three steps, whereas Sampler 2.1 does not remove α from any step. Samplers 2.3 and 2.4 attempt to further improve Sampler 2.2 by blocking two of its steps. Sampler 2.3 blocks the updates of β and ϕ into one single MH step, see the bottom-left panel of Figure 2.1, while Sampler 2.4 blocks the updates of β and α, see the bottom-right panel of Figure 2.1. Details of Samplers 2.1–2.4 and other samplers introduced in this chapter appear in Appendix A.

Unfortunately, Sampler 2.4 is improper, while Samplers 2.1–2.3 are proper MH within PCG samplers with common parent Gibbs sampler. The failure of Sampler 2.4 rests in the MH update of Step 3. The MH transition kernel $\mathcal{M}_{\alpha,\beta|Y,\gamma,\mu,\phi}(\alpha, \beta|Y, \alpha', \beta', \gamma', \mu, \phi)$ depends on (α', β') via its acceptance probability and its output if its proposal is rejected, see Section 1.1.2 of Chapter 1. If the joint distribution of (α', β') and (γ', μ, ϕ) were the target, $\mathcal{M}_{\alpha,\beta|Y,\gamma,\mu,\phi}(\alpha, \beta|Y, \alpha', \beta', \gamma', \mu, \phi)$ would deliver a draw from $p(\alpha, \beta|Y, \gamma', \mu, \phi)$. However, Steps 1 and 2 of Sampler 2.4 update μ and ϕ without conditioning on α'. Thus α' and (μ, ϕ) are conditionally independent and $\mathcal{M}_{\alpha,\beta|Y,\gamma,\mu,\phi}(\alpha, \beta|Y, \alpha', \beta', \gamma', \mu, \phi)$ fails to deliver a draw from $p(\alpha, \beta|Y, \gamma', \mu, \phi)$. If the conditional distribution of (α, β) were available without resorting to an MH update, Sampler 2.4 would be proper because under this scenario, the transition kernel of (α, β) would only depend on (γ', μ, ϕ), whose joint distribution is the target. Thus unlike an ordinary Gibbs sampler, introducing MH into a PCG sampler can destroy its stationary distribution. Examples appear even in simplest two-step MH within PCG samplers, see Section 2.2.1. We must take extra care to guarantee that an MH within PCG sampler is proper.

Like a standard PCG sampler, permuting the steps of a proper MH within PCG sampler may alter its stationary distribution. Suppose, for example, we obtain $(Y_L', \alpha', \beta', \gamma', \mu', \phi')$ from $p(Y_L, \alpha, \beta, \gamma, \mu, \phi|Y)$ and update μ according to Step 1 of Sampler 2.1. The joint distribution of $(Y_L', \alpha', \beta', \gamma', \mu, \phi')$ would be

$$\begin{aligned}&\int p(\mu|Y, \alpha', \beta', \gamma', \phi')p(Y_L', \alpha', \beta', \gamma', \mu', \phi'|Y)d\mu' \\ &= p(\alpha', \beta', \gamma', \mu, \phi'|Y)p(Y_L'|Y, \alpha', \beta', \gamma', \phi').\end{aligned} \tag{2.2}$$

Because the joint distribution of $(\alpha', \beta', \gamma', \phi')$ and μ in (2.2) is the target posterior distribution and Step 2 of Sampler 2.1 conditions only on $(\alpha', \beta', \gamma', \phi')$ and μ, the joint distribution of the parameters after Step 2, that is, of $(Y_L, \alpha', \beta', \gamma', \mu, \phi')$, is again the target. Thus a cyclic permutation of the steps in Sampler 2.1 that ends with each of Steps 2–6 leads to a proper sampler, but ending with Step 1 does not. With non-cyclic permutations, the stationary distribution is unknown. It is the conditional independence of X_L' and μ in (2.2) that makes Sampler 2.1 much faster than its parent MH with Gibbs sampler; recall $\text{var}_p(\mu|X_L) = 0$. In addition, Sampler 2.3, i.e., the proper MH within PCG sampler with the highest degree of partial collapsing, converges much faster than Sampler 2.1, the sampler with the lowest degree of partial collapse. Sampler 2.2 is also more efficient than Sampler 2.1, but worse than Sampler 2.3. See Section 2.3.1 for numerical illustration. Thus proper MH within PCG samplers outperform their parent Gibbs samplers in computational efficiency, and higher degree of partial collapsing can typically lead to better performance in convergence.

Motivated by this example, we consider developing a general strategy for using MH updates within PCG samplers while maintaining the target stationary distribution, and illustrating the computational advantage of proper MH within PCG samplers using examples from our applied work.

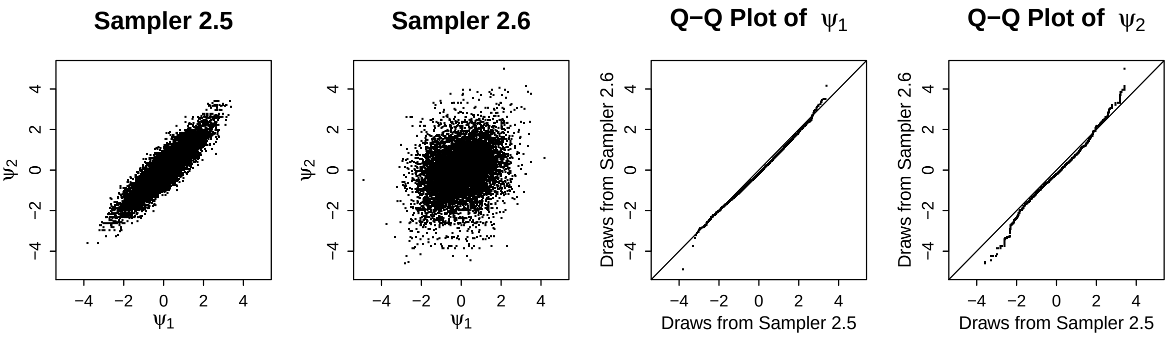

Figure 2.2: Proper and improper samplers, for the bivariate Gaussian target distribution. The first two panels give scatter plots of ψ_1 and ψ_2 for 10,000 draws from Samplers 2.5 and 2.6, respectively. The marginal distributions from the two samplers are compared in the two quantile-quantile plots. The improper Sampler 2.6 severely underestimates the correlation between ψ_1 and ψ_2, and slightly overestimates the variance of ψ_2.

2.2 Using the MH Algorithm within PCG samplers

2.2.1 An error of embedding MH into simplest PCG samplers

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

Step 1: $\psi_1^{(t+1)} \sim p(\psi_1|\psi_2^{(t+1)})$, (Sampler 2.5)

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

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

Step 1: $\psi_1^{(t+1)} \sim p(\psi_1)$, (Sampler 2.6)

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

The problem with Sampler 2.6 can be illustrated using a simulation study. Figure 2.2 compares 10,000 draws generated by Samplers 2.5 and 2.6 with $p(\psi_1, \psi_2)$ given by

$$\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \sim \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right]. \tag{2.3}$$

The MH jumping rule in Step 2 of both samplers is a Gaussian distribution centered at the previous draw of ψ_2 with variance equal to 6. Sampler 2.6 underestimates the correlation of the target distribution and overestimates the marginal variance of ψ_2. Of course, if we repeat Step 2 a sufficient number of times within each iteration of Sampler 2.6, it would deliver a draw (approximately) from its target, $p(\psi_2|\psi_1)$, and Sampler 2.6 would deliver (approximately) independent draws from $p(\psi_1, \psi_2)$. We discuss this strategy for constructing an approximately proper sampler in Section 2.2.2. Similarly, iterating Step 2 of Sampler 2.5 would (approximately) lead to a standard two-step Gibbs sampler.

Like Sampler 2.4 in Section 2.1, the key to understanding the failure of Sampler 2.6 (without iterating Step 2) lies in the MH jumping rule used in Step 2 of both samplers. The kernel $\mathcal{M}_{2|1}(\psi_2|\psi_1^{(t+1)}, \psi_2^{(t)})$ depends on $\psi_2^{(t)}$. Although $\mathcal{M}(\psi_2|\psi_1^{(t+1)}, \psi_2^{(t)})$ delivers a draw from $p(\psi_2|\psi_1)$ if given a sample $(\psi_1^{(t+1)}, \psi_2^{(t)})$ from the target distribution, in Sampler 2.6, $\psi_1^{(t+1)}$ and $\psi_2^{(t)}$ are independent and $\mathcal{M}(\psi_2|\psi_1^{(t+1)}, \psi_2^{(t)})$ does not deliver a draw from the target conditional distribution.

Unfortunately, there are a number of samplers in the literature that have the same structure as the improper Sampler 2.6, for instance, Liu *et al.* (2009), Lunn *et al.* (2009), McCandless *et al.* (2010), and even in the popular WinBUGS package (Spiegelhalter, Thomas, Best and Lunn 2003). These samplers do not generally exhibit the desired stationary distributions.

2.2.2 Embedding the MH algorithm into PCG samplers

Verifying the stationary distribution

As the example in Section 2.2.1 illustrates, introducing MH into a well behaved PCG sampler can alter the sampler's stationary distribution. Here we describe the basic complication that arises when MH is introduced into a PCG sampler and give advice on how to guarantee that the sampler is proper.

When deriving a PCG sampler (without MH), the marginalization stage means some components of ψ are updated in multiple steps. If the same component is updated in consecutive steps, the Markov transition kernel does not depend on the first update. The first update is therefore redundant and can be omitted without affecting the stationary distribution of the chain. The situation becomes more complicated when some of the steps of the PCG sampler require MH updates. Suppose, for example, we wish to sample from $p(\psi)$ with $\psi = (\psi_1, \psi_2, \psi_3)$ using a proper PCG sampler in which ψ_1 and ψ_2 are jointly updated in Step j via a draw from the conditional distribution $p(\psi_1, \psi_2|\psi_3)$. Suppose also that ψ_2 is to be updated according to its full conditional distribution, $p(\psi_2|\psi_1, \psi_3)$ in Step $j+1$, but this cannot be done directly and we wish to use an MH update. The remaining unknowns, ψ_3, are updated in other steps of the sampler, which perhaps involve dividing ψ_3 into multiple subcomponents. That is, Steps j and $j+1$ of the sampler are

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

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

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

Fortunately, the stationary distribution of an MH within PCG sampler can be verified using the same three-stage framework of van Dyk and Park (2008) that are used for an ordinary PCG sampler. The first two stages, marginalization and permutation, apply equally well to MH within Gibbs samplers. Neither updating additional components of ψ in one or more steps nor permuting the order of the steps upsets the stationary distribution of an MH within Gibbs sampler. The final stage involves removing redundant updates. Because an MH step depends on the current draws of not only the components of ψ conditioned upon in the target conditional distribution of that step, but those to be updated in the step also, there are fewer redundant draws when some steps involve MH. Nonetheless, any redundant updates that are identified can safely be removed in the trimming stage—by definition they do not affect the transition kernel. The critical point is that unlike with an ordinary Gibbs sampler, we cannot simply replace some of the component draws of a PCG sampler with MH updates. Rather we must construct an MH within PCG sampler by applying the three-stage framework.

Now suppose we wish to implement the marginalization in an MH step. In Sampler Fragment 1, for example, we aim to sample ψ_3 along with ψ_2 using a single MH update in Step $j+1$. This typically requires replacing Step $j+1$ with $(\psi_2^{(t+1)}, \psi_3) \sim \mathcal{M}_{2,3|1}(\psi_2, \psi_3|\psi_1^{(t+1)}, \psi_2^\star, \psi_3')$. Because the MH transition kernel of (ψ_2, ψ_3) depends on ψ_3', this new Step $j+1$ cannot follow a step that reduces ψ_3 out. If $p(\psi_3|\psi_1, \psi_2)$ is a standard distribution, to update ψ_2 and ψ_3 together, we propose to replace the full MH step with the reduced MH step *followed immediately* by a direct draw from the complete conditional of the reduced quantities. This would entail replacing Step $j+1$ of Sampler Fragment 1 with

Step $j+1$ with marginalization:

$$\psi_2^{(t+1)} \sim \mathcal{M}_{2|1}(\psi_2|\psi_1^{(t+1)}, \psi_2^\star) \text{ and } \psi_3 \sim p(\psi_3|\psi_1^{(t+1)}, \psi_2^{(t+1)}).$$

We denote the transition kernel of the full step (i.e., the reduced MH step followed by the complete conditional of the reduced quantities) by $\mathcal{M}^\star$. In Sampler Fragment 1, we rewrite the step with marginalization as

Step $j+1$ with marginalization: $(\psi_2^{(t+1)}, \psi_3) \sim \mathcal{M}^\star_{2,3|1}(\psi_2, \psi_3|\psi_1^{(t+1)}, \psi_2^\star)$.

(a) Parent MH within Gibbs Sampler	(b) Marginalization
1. $p(Y_L\|Y,\alpha',\beta',\gamma',\mu',\phi')$	1. $p(Y_L^\star\|Y,\alpha',\beta',\gamma',\mu',\phi')$
2. $p(\alpha\|Y,Y_L,\beta',\gamma',\mu',\phi')$	2. $p(\alpha^\star,Y_L^\star\|Y,\beta',\gamma',\mu',\phi')$
3. $\mathcal{M}_{\beta\|Y,Y_L,\alpha,\gamma,\mu,\phi}(\beta\|Y,Y_L,\alpha,\beta',\gamma',\mu',\phi')$	3. $\mathcal{M}^\star_{\beta,Y_L,\alpha\|Y,\gamma,\mu,\phi}(\beta,Y_L^\star,\alpha^\star\|Y,\beta',\gamma',\mu',\phi')$
4. $p(\gamma\|Y,Y_L,\alpha,\beta,\mu',\phi')$	4. $p(\gamma\|Y,Y_L^\star,\alpha^\star,\beta,\mu',\phi')$
5. $\mathcal{M}_{\mu\|Y,Y_L,\alpha,\beta,\gamma,\phi}(\mu\|Y,Y_L,\alpha,\beta,\gamma,\mu',\phi')$	5. $\mathcal{M}^\star_{\mu,Y_L,\alpha\|Y,\beta,\gamma,\phi}(\mu,Y_L^\star,\alpha^\star\|Y,\beta,\gamma,\mu',\phi')$
6. $\mathcal{M}_{\phi\|Y,Y_L,\alpha,\beta,\gamma,\mu}(\phi\|Y,Y_L,\alpha,\beta,\gamma,\mu,\phi')$	6. $\mathcal{M}^\star_{\phi,Y_L,\alpha\|Y,\beta,\gamma,\mu}(\phi,Y_L,\alpha\|Y,\beta,\gamma,\mu,\phi')$

(c) Permute	(d) Trim (Sampler 2.2)
1. $\mathcal{M}^\star_{\mu,Y_L,\alpha\|Y,\beta,\gamma,\phi}(\mu,Y_L^\star,\alpha^\star\|Y,\beta',\gamma',\mu',\phi')$	1. $\mathcal{M}_{\mu\|Y,\beta,\gamma,\phi}(\mu\|Y,\beta',\gamma',\mu',\phi')$
2. $\mathcal{M}^\star_{\phi,Y_L,\alpha\|Y,\beta,\gamma,\mu}(\phi,Y_L^\star,\alpha^\star\|Y,\beta',\gamma',\mu,\phi')$	2. $\mathcal{M}_{\phi\|Y,\beta,\gamma,\mu}(\phi\|Y,\beta',\gamma',\mu,\phi')$
3. $\mathcal{M}^\star_{\beta,Y_L,\alpha\|Y,\gamma,\mu,\phi}(\beta,Y_L^\star,\alpha^\star\|Y,\beta',\gamma',\mu,\phi)$	3. $\mathcal{M}_{\beta\|Y,\gamma,\mu,\phi}(\beta\|Y,\beta',\gamma',\mu,\phi)$
4. $p(\alpha,Y_L^\star\|Y,\beta,\gamma',\mu,\phi)$	4. $p(\alpha\|Y,\beta,\gamma',\mu,\phi)$
5. $p(Y_L\|Y,\alpha,\beta,\gamma',\mu,\phi)$	5. $p(Y_L\|Y,\alpha,\beta,\gamma',\mu,\phi)$
6. $p(\gamma\|Y,Y_L,\alpha,\beta,\mu,\phi)$	6. $p(\gamma\|Y,Y_L,\alpha,\beta,\mu,\phi)$

Figure 2.3: The three-stage framework used to derive Sampler 2.2 from its parent MH within Gibbs sampler. The parent sampler appears in (a) with Steps 3, 5 and 6 requiring MH updates. The conditioning in steps 2, 3, 5, and 6 is reduced in (b). The steps are permuted in (c) to allow redundant draws of $Y_L^\star$ and $\alpha^\star$ to be trimmed in Steps 1–4. The resulting proper MH within PCG sampler, i.e., Sampler 2.2, appears in (d).

Note that this full step is not formally an MH update but has the advantage that it does not depend on the current update of ψ_3. Thus, this step can follow a step that reduces ψ_3 out. This MH update ensures that the target stationary distribution is maintained. Moreover, using this strategy, the updates of the reduced quantities (i.e, ψ_3 here) are expected to be trimmed after the steps are appropriately permuted and that the reduced MH step, i.e., $\psi_2^{(t+1)} \sim \mathcal{M}_{2|1}(\psi_2|\psi_1^{(t+1)}, \psi_2^\star)$, can be employed in the final sampler.

We now illustrate the construction of a proper MH within PCG sampler for the spectral analysis model given in Section 2.1. Figure 2.3(a) gives the six-step Gibbs sampler and three of its steps require MH updates. The conditioning in four steps is reduced in Figure 2.3(b), and the steps are permuted in Figure 2.3(c) to allow the redundant draws of $Y_L^\star$ and $\alpha^\star$ to be trimmed in four steps. The resulting proper MH within PCG sampler, i.e., Sampler 2.2, appears in Figure 2.3(d).

Using MH following a reduced step

Using a full MH step immediately following a reduced step can be problematic. Sampler 2.6 in Section 2.2.1 illustrates this in its simplest form: a draw from a marginal distribution followed by an MH update of the conditional distribution of the remaining unknowns. As noted in Section 2.2.1, this is a particularly common problem in practice, even in its simplest form. In more complicated PCG samplers, the general phenomenon of introducing a full MH step immediately following a reduced step is the typical path by which introducing MH leads to an improper sampler, e.g., the improper Sampler 2.6 in Section 2.1. This is illustrated in Sampler Fragment 1, where we are unable to replace the update in Step j with the reduced step $\psi_1^{(t+1)} \sim p(\psi_1|\psi_3')$. Thus, this case is particularly important and we propose two alternate strategies that maintain the basic structure of the underlying PCG sampler while allowing a form of MH in the step following a reduced step. Both solutions are conceptually straightforward.

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

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

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

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

One straightforward solution to the intractability of $p(\psi_2|\psi_1^{(t+1)}, \psi_3')$ is iterating the MH update within Step $j+1$ to obtain a draw from the target conditional distribution, *Iterated MH Strategy*:

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

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

With sufficiently large L (see the following subsection for methods of determining how large L must be), the Iterative MH Strategy delivers a sample approximately from $p(\psi_2|\psi_1^{(t+1)}, \psi_3')$ and thus the sampler is *approximately proper*. In this special case the Iterated MH Strategy effectively blocks Steps j and $j+1$ to (approximately) deliver an independent draw from $p(\psi_1, \psi_2|\psi_3')$.

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

Joint MH Strategy:

Step $\boldsymbol{j}$: Update (ψ_1, ψ_2) jointly via the MH jumping rule

$$\mathcal{J}_{1,2|3}(\psi_1, \psi_2|\psi_2^{(t)}, \psi_3') = p(\psi_1|\psi_3')\mathcal{J}_{2|1,3}(\psi_2|\psi_1, \psi_2^{(t)}, \psi_3'),$$

Step $\boldsymbol{j+1}$: Omit. (Sampler Fragment 4)

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

The Iterated MH Strategy is in some sense a thinned version of the Joint MH Strategy. This, however, is an over simplification for two reasons. First, the Iterated MH Strategy updates ψ_1 only once for every L updates of ψ_2 whereas the Joint MH Strategy updates both together in each iteration. Second, although the jumping rule in the Joint MH Strategy is the concatenation of Step j and the jumping rule used by the first subiteration in Step $j+1$ of Iterated MH Strategy, the acceptance probabilities differ.

(a) Gibbs Sampler	(b) Marginalization	(c) Permute	(d) Trim
1. $p(\psi_3 \vert \psi_1', \psi_2')$	1. $p(\psi_3 \vert \psi_1', \psi_2')$	1. $p(\psi_3 \vert \psi_1', \psi_2')$	1. $p(\psi_3 \vert \psi_1', \psi_2')$
2. $p(\psi_2 \vert \psi_1', \psi_3)$	2. $p(\psi_2^\star \vert \psi_1', \psi_3)$	2. $\mathcal{M}^\star_{1,2 \vert 3}(\psi_1, \psi_2^\star \vert \psi_1', \psi_3)$	2. $\mathcal{M}_{1 \vert 3}(\psi_1 \vert \psi_1', \psi_3)$
3. $p(\psi_1 \vert \psi_2, \psi_3)$	3. $\mathcal{M}^\star_{1,2 \vert 3}(\psi_1, \psi_2 \vert \psi_1', \psi_3)$	3. $p(\psi_2 \vert \psi_1, \psi_3)$	3. $p(\psi_2 \vert \psi_1, \psi_3)$

1. $p(\psi_2 \vert \psi_1', \psi_3')$	1. $p(\psi_2^\star \vert \psi_1', \psi_3')$	1. $\mathcal{M}^\star_{1,2 \vert 3}(\psi_1, \psi_2^\star \vert \psi_1', \psi_3')$	1. $\mathcal{M}_{1 \vert 3}(\psi_1 \vert \psi_1', \psi_3')$
2. $p(\psi_1 \vert \psi_2, \psi_3')$	2. $\mathcal{M}^\star_{1,2 \vert 3}(\psi_1, \psi_2 \vert \psi_1', \psi_3')$	2. $p(\psi_2 \vert \psi_1, \psi_3')$	2. $p(\psi_2 \vert \psi_1, \psi_3')$
3. $p(\psi_3 \vert \psi_1, \psi_2)$	3. $p(\psi_3 \vert \psi_1, \psi_2)$	3. $p(\psi_3 \vert \psi_1, \psi_2)$	3. $p(\psi_3 \vert \psi_1, \psi_2)$

Figure 2.4: The three-stage framework to derive Sampler Fragment 6 from its parent Gibbs sampler. The first row corresponds to updating ψ_3 before Steps j and $j+1$, while the second row updating ψ_3 after that.

This results in a systematic difference in the performance of the resulting samplers, see Section 2.2.3 for further explanation.

Generalizing Sampler Fragment 2, suppose $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)$ and the parent PCG sampler contains the two steps

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

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

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

To block or not to block

In the first subsection of Section 2.2.2, we discuss the case where Step $j+1$ of Sampler Fragment 2 requires MH. We now consider the case where Step j requires MH, i.e.,

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

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

Sampler Fragment 6 does not lead to convergence problems because the inputs to Step $j+1$, i.e., $\psi_1^{(t+1)}$ and ψ_3', follow the desired distribution. Figure 2.4 illustrates using the three-stage framework to verify the stationary distribution of Sampler Fragment 6, with ψ_3 sampled from its complete conditional distribution either before (first row) or after (second row) Steps K and $K+1$.

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

Step j: Update (ψ_1, ψ_2) jointly via the MH jumping rule

$$\mathcal{J}_{1,2|3}(\psi_1, \psi_2|\psi_1^{(t)}, \psi_3') = \mathcal{J}_{1|3}(\psi_1|\psi_1^{(t)}, \psi_3')p(\psi_2|\psi_1, \psi_3'),$$

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

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

Steps 3 and 4 of Sampler 2.2 (top-right panel of Figure 2.1) are an example of Sampler Fragment 6, with $\psi_1 = \beta$, $\psi_2 = \alpha$, and $\psi_3 = (\gamma, \mu, \phi)$. Blocking Steps 3 and 4 of Sampler 2.2 results in Sampler 2.4 (bottom-right panel of Figure 2.1). As stated in Section 2.1, unfortunately, Sampler 2.4 is an improper sampler, which we verify using a simulation study. We begin by generating an artificial data set consisting of $n = 550$ bins with $\alpha = 37.62$, $\beta = 1$, $\gamma = 40/37.62$, $\mu = 250$, and $\phi = 0.2$, see Figure 2.5. We run two versions of Sampler 2.4 with different jumping rules for updating (α, β). Sampler 2.4(a) uses the concatenated jumping rule given in Sampler Fragment 7, while

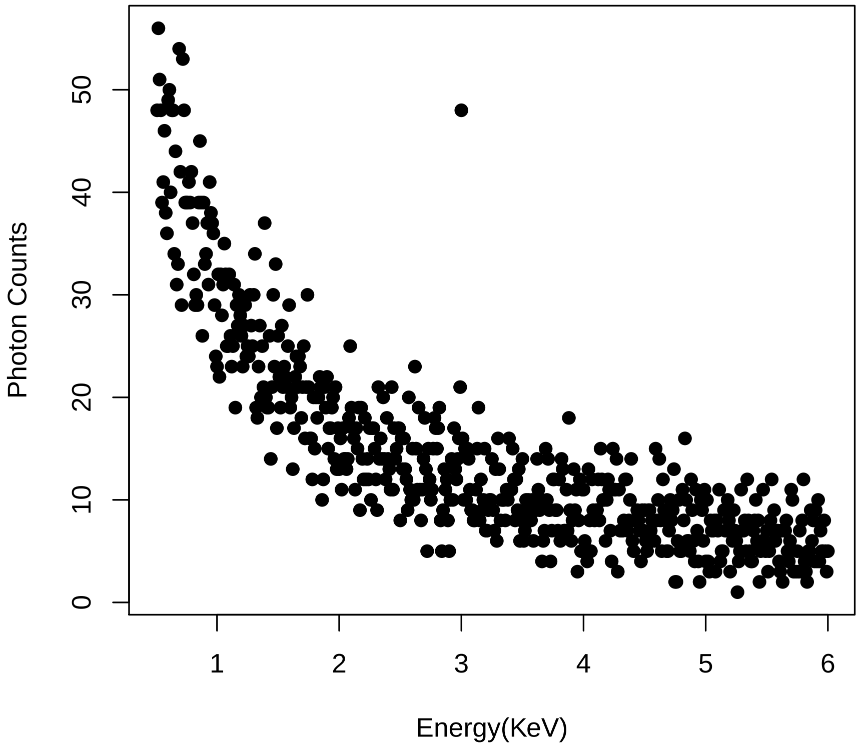

Figure 2.5: A dataset simulated under the spectral model in (2.1) and used in the simulation study in Sections 2.2.2 and 2.3.

Sampler 2.4(b) uses an independent bivariate normal jumping rule centered at the current value of (α, β). We use a uniform prior distribution for each parameter, and run 30,000 iterations of Samplers 2.2, 2.4(a), and 2.4(b) using the same starting values. Scatter plots of (α, β, ϕ) for the last 10,000 draws from the three samplers appear in Figure 2.6, which shows that Samplers 2.4(a) and 2.4(b) underestimate the correlations of the target distribution; this effect is especially significant for Sampler 2.4(b). Figure 2.7 compares the marginal distributions of α, β, and ϕ generated with Samplers 2.2 and 2.4(b), and shows that Sampler 2.4(b) underestimates the marginal variances of all three parameters. (The marginals generated with Sampler 2.4(a) are more similar to those generated with Sampler 2.2.)

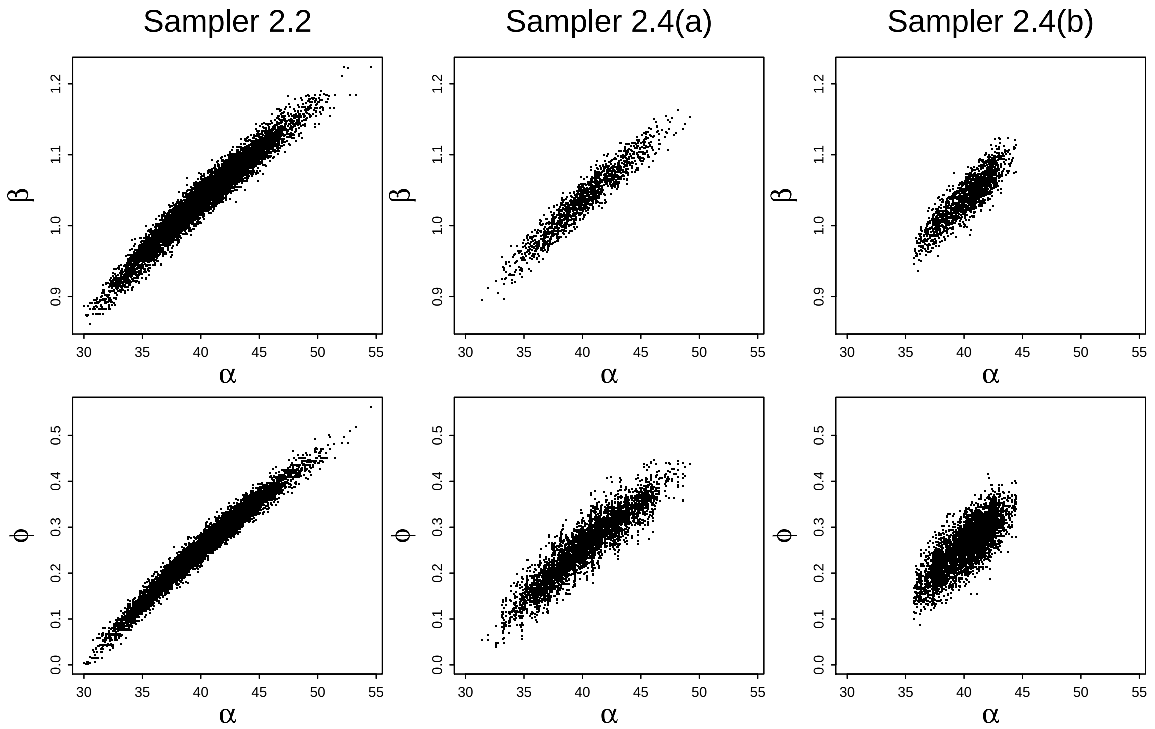

Figure 2.6: Scatter plots of α, β and ϕ for 10,000 draws from Samplers 2.2, 2.4(a), and 2.4(b) respectively. The first row displays scatter plots of α and β, whereas the second row displays that of α and ϕ. The left, central, and right columns correspond to Samplers 2.2, 2.4(a), and 2.4(b) respectively. Sampler 2.4 blocks the two steps of Sampler 2.2 that update α and β. Unfortunately, this results in an improper sampler. When updating (α, β), Sampler 2.4(a) uses the concatenation of Sampler 2.2's jumping rule for β and its step of updating α, while Sampler 2.4(b) uses an independent bivariate normal jumping rule. The impropriety of Sampler 2.4(b) is especially dramatic.

The problem with Sampler 2.4 can be understood in the terms of Section 2.2.2. Blocking the updates for α and β results in an MH step that follows directly after a pair of reduced steps (the updates of μ and ϕ). As it is, the stationary distribution of Sampler 2.4 cannot be verified with the three-phase framework.

Like the comparison of Sampler Fragments 3 and 4, theoretical arguments on the choice between Sampler Fragments 6 and 7 appear in Section 2.2.3.

2.2.3 Theoretical Comparisons

In this section we compare the iterated and joint MH strategies in terms of their acceptance probabilities.

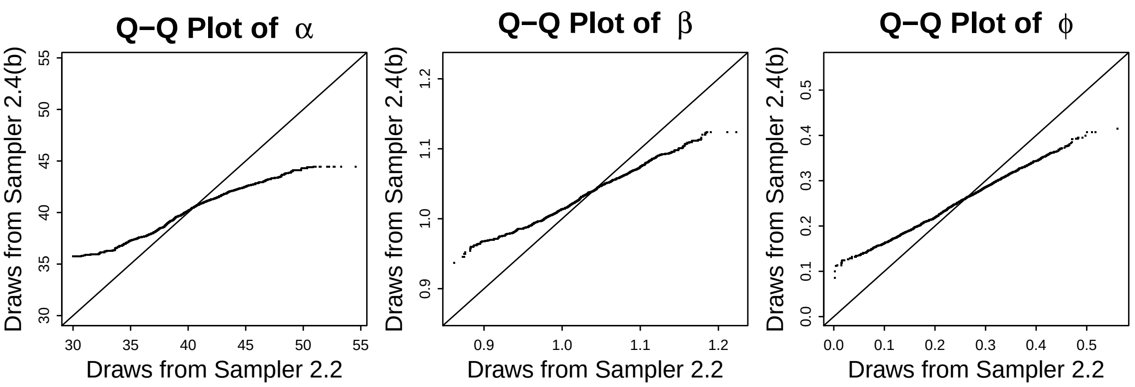

Figure 2.7: Quantile-quantile plots of α, β, and ϕ corresponding to draws generated with Samplers 2.2 and 2.4(b). Sampler 2.4(b) severely underestimates the marginal variances of all three parameters.

Comparing the Iterated and Joint MH Strategies

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

To simplify notation we suppress the conditioning on ψ_3 in Sampler Fragments 3 and 4. This is equivalent to a formal comparison of the iterated and joint MH strategies as alternatives to the improper two-step Sampler 2.6. While the transition kernel used to update ψ_2, i.e., $\mathcal{M}_{2|1}(\psi_2|\psi_1', \psi_2')$ will typically depend on ψ_1', the jumping rule often will not, for example, a symmetric Metropolis jumping rule. Thus we assume that (i) the samplers for the Iterated and Joint MH Strategies are proper and (ii) their jumping rule for updating ψ_2 does not depend on the current draw of ψ_1, i.e., $\mathcal{J}_{2|1}(\psi_2|\psi_1', \psi_2') = \mathcal{J}_{2|1}(\psi_2|\psi_2')$.

The acceptance probability of the first subiteration in Step $j+1$ of the Iterated MH Strategy is

$$r_{\text{iter}} = \frac{p(\psi_2^{\text{prop}}|\psi_1^{(t+1/L)})\mathcal{J}_{2|1}(\psi_2^{(t)}|\psi_2^{\text{prop}})}{p(\psi_2^{(t)}|\psi_1^{(t+1/L)})\mathcal{J}_{2|1}(\psi_2^{\text{prop}}|\psi_2^{(t)})}, \tag{2.4}$$

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

$$r_{\text{joint}} = \frac{p(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})\{p(\psi_1^{(t)})\mathcal{J}_{2|1}(\psi_2^{(t)}|\psi_2^{\text{prop}})\}}{p(\psi_1^{(t)}, \psi_2^{(t)})\{p(\psi_1^{\text{prop}})\mathcal{J}_{2|1}(\psi_2^{\text{prop}}|\psi_2^{(t)})\}} = \frac{p(\psi_2^{\text{prop}}|\psi_1^{\text{prop}})\mathcal{J}_{2|1}(\psi_2^{(t)}|\psi_2^{\text{prop}})}{p(\psi_2^{(t)}|\psi_1^{(t)})\mathcal{J}_{2|1}(\psi_2^{\text{prop}}|\psi_2^{(t)})}, \tag{2.5}$$

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

Lemma 2.1. *In the setting described in the previous paragraph,*

$$\mathrm{E}_p[r_{\text{iter}}/r_{\text{joint}}] \geq 1. \tag{2.6}$$

The expectation in (2.6) is with regard to the common stationary distribution, p, of both chains and is conditional on the random seed used at the start of each iteration. That is, since $(\psi_1^{(t+1/L)}, \psi_2^{\text{prop}})$ sampled under the Iterated MH strategy and $(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})$ sampled under the Joint MH strategy are drawn in exactly the same way, we assume these quantities are numerically equal. Expression (2.6) asserts that while both strategies start with the same proposal $((\psi_1^{(t+1/L)}, \psi_2^{\text{prop}})$ for the Iterated MH Strategy and $(\psi_1^{\text{prop}}, \psi_2^{\text{prop}})$ for the Joint MH Strategy), the iterated MH strategy is on average more likely to accept ψ_2. (The Iterated MH Strategy *always* accepts ψ_1.)

Proof. With the numerical equality of the proposals,

$$\frac{r_{\text{iter}}}{r_{\text{joint}}} = \frac{p(\psi_2^{(t)}|\psi_1^{(t)})}{p(\psi_2^{(t)}|\psi_1^{(t+1/L)})}, \tag{2.7}$$

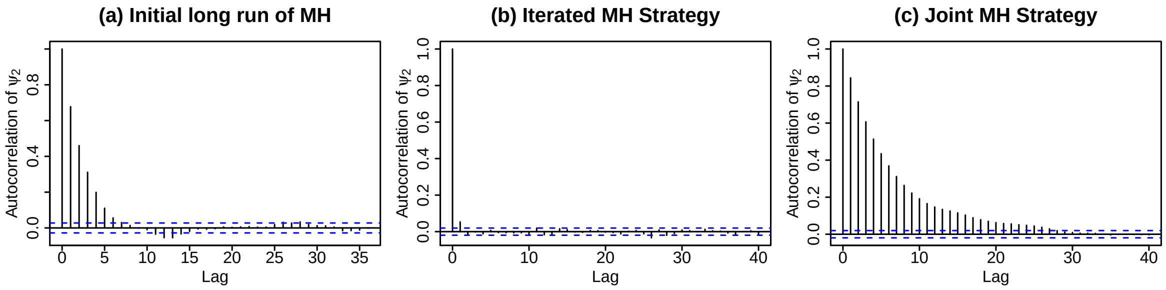

Figure 2.8: Autocorrelation functions of ψ_2 for (a) an initial MH run of Step 2 of Sampler 2.6 with ψ_1 fixed, (b) the Iterated MH Strategy, and (c) the Joint MH Strategy, all under the bivariate normal simulation described in Section 2.2.1. Panel (a) shows that the initial MH runs deliver essentially independent draws after 7 iterations, so that Iterated MH Strategy is run with $L = 7$. Panels (b) and (c) show that the Iterated MH Strategy outperforms the joint one in terms of its computational efficiency.

where $(\psi_1^{(t)}, \psi_2^{(t)}, \psi_1^{(t+1/L)}) \sim p(\psi_1^{(t)}, \psi_2^{(t)})p(\psi_1^{(t+1/L)})$. Because $(\psi_1^{(t)}, \psi_2^{(t)}) \sim p(\psi_1, \psi_2)$, the numerator of (2.7) is the conditional density of ψ_2 evaluated at $\psi_2^{(t)}$. This is not true of the denominator because $\psi_2^{(t)}$ is independent of $\psi_1^{(t+1/L)}$. Thus, we might expect that the numerator of (2.7) is typically larger than the denominator, as claimed in (2.6).

Substituting (2.7) into (2.6), and applying Jensen's inequality, we need only verify that

$$\int \log\left[p(\psi_2|\psi_1)\right] p(\psi_1, \psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2 \geq \int \log\left[p(\psi_2|\psi_1)\right] p(\psi_1)\pi(\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2. \tag{2.8}$$

Expression (2.8) can be verified using a standard property of entropy along with the Kullback-Leiber (KL) divergence. In particular, because KL is nonnegative,

$$\int \log\left[p(\psi_2)\right] p(\psi_1)p(\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2 \geq \int \log\left[p(\psi_2|\psi_1)\right] p(\psi_1)p(\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2. \tag{2.9}$$

But a standard property of entropy (e.g., Ebrahimi *et al.*, 1999) is

$$\int \log\left[p(\psi_2|\psi_1)\right] p(\psi_1, \psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2 \geq \int \log\left[p(\psi_2)\right] p(\psi_1)p(\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2. \tag{2.10}$$

Combining (2.9) and (2.10) gives (2.8) and hence the desired result. □

We now return to the bivariate Gaussian example of Section 2.2.1 to compare the computational performance of the Iterated and Joint MH strategies. As in Sampler 2.6, we sample ψ_1 from its marginal distribution and use the same MH jumping rule to update ψ_2 according to its conditional distribution. The Iterated MH Strategy is run with $L = 7$, in order to return $\psi_2^{(t+1)}$ that is essentially independent of $\psi_2^{(t+1)}$. The value of L was set using an initial MH run of $5,000$ iterations and inspecting the autocorrelation function. The initial MH sampler delivers essentially independent draws after 7 iterations, see Figure 2.8(a). The computational cost per iteration of the Iterated MH Strategy surely depends on L. With $L = 7$, each iteration requires eight univariate normal draws, whereas the Joint MH Strategy requires two. The autocorrelation functions of ψ_2 from both the Iterated and Joint MH Strategies appear in Figure 2.8(b) and (c), and show the clear computational advantage of the Iterated MH Strategy, which returns essentially independent draws, whereas the Joint MH Strategy requires almost thirty full iterations to obtain nearly independent draws.

In practice, it is important to check that the value of L used in Sampler Fragment 3 delivers samples that are essentially independent of the starting value of the Iterated MH Strategy. Fortunately, the autocorrelation function of $\psi_2^{(t)}$ in Sampler Fragment 3 can serve as a simple diagnostic, e.g., Figure 2.8(b). If the lag-1 autocorrelation is not essentially zero, the run should be repeated with a larger value of L. If ψ_2 is updated elsewhere in the sampler, the efficacy of the Iterated MH Strategy can be measured by the correlation between the initial input of ψ_2 and the final output after L iterations of the MH update in Step $j+1$ of Sampler Fragment 3.

Comparing the samplers with and without blocking

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

$$\begin{aligned} r_{\text{blocked}} &= \frac{p(\psi_1^{\text{prop}}, \psi_2^{\text{prop}}) \mathcal{J}_1(\psi_1^{(t)} | \psi_1^{\text{prop}}) p(\psi_2^{(t)} | \psi_1^{(t)})}{p(\psi_1^{(t)}, \psi_2^{(t)}) \mathcal{J}_1(\psi_1^{\text{prop}} | \psi_1^{(t)}) p(\psi_2^{\text{prop}} | \psi_1^{\text{prop}})} \\ &= \frac{p(\psi_1^{\text{prop}}) \mathcal{J}_1(\psi_1^{(t)} | \psi_1^{\text{prop}})}{p(\psi_1^{(t)}) \mathcal{J}_1(\psi_1^{\text{prop}} | \psi_1^{(t)})} = r_{\text{not blocked}}, \end{aligned} \tag{2.11}$$

(a) Parent MH within Gibbs Sampler

1. $p(Y_L|Y, \alpha', \beta', \gamma', \mu', \phi')$
2. $p(\alpha|Y, Y_L, \beta', \gamma', \mu', \phi')$
3. $\mathcal{M}_{\beta|Y,Y_L,\alpha,\gamma,\mu,\phi}(\beta|Y, Y_L, \alpha, \beta', \gamma', \mu', \phi')$
4. $p(\gamma|Y, Y_L, \alpha, \beta, \mu', \phi')$
5. $\mathcal{M}_{\mu|Y,Y_L,\alpha,\beta,\gamma,\phi}(\mu|Y, Y_L, \alpha, \beta, \gamma, \mu', \phi')$
6. $\mathcal{M}_{\phi|Y,Y_L,\alpha,\beta,\gamma,\mu}(\phi|Y, Y_L, \alpha, \beta, \gamma, \mu, \phi')$

(b) Marginalization

1. $p(Y_L^\star|Y, \alpha', \beta', \gamma', \mu', \phi')$
2. $p(\alpha^\star, Y_L^\star|Y, \beta', \gamma', \mu', \phi')$
3. $\mathcal{M}^\star_{\beta,Y_L,\alpha,\phi|Y,\gamma,\mu}(\beta^\star, Y_L^\star, \alpha^\star, \phi^\star|Y, \beta', \gamma', \mu', \phi')$
4. $p(\gamma|Y, Y_L^\star, \alpha^\star, \beta^\star, \mu', \phi^\star)$
5. $\mathcal{M}^\star_{\mu,Y_L,\alpha|Y,\beta,\gamma,\phi}(\mu, Y_L^\star, \alpha^\star|Y, \beta^\star, \gamma, \mu', \phi^\star)$
6. $\mathcal{M}^\star_{\phi,Y_L,\alpha,\beta|Y,\gamma,\mu}(\phi, Y_L, \alpha, \beta|Y, \beta^\star, \gamma, \mu, \phi^\star)$

(c) Permute

1. $\mathcal{M}^\star_{\mu,Y_L,\alpha|Y,\beta,\gamma,\phi}(\mu, Y_L^\star, \alpha^\star|Y, \beta', \gamma', \mu', \phi')$
2. $\mathcal{M}^\star_{\phi,Y_L,\alpha,\beta|Y,\gamma,\mu}(\phi^\star, Y_L^\star, \alpha^\star, \beta^\star|Y, \beta', \gamma', \mu, \phi')$
3. $\mathcal{M}^\star_{\beta,Y_L,\alpha,\phi|Y,\gamma,\mu}(\beta, Y_L^\star, \alpha^\star, \phi|Y, \beta^\star, \gamma', \mu, \phi^\star)$
4. $p(\alpha, Y_L^\star|Y, \beta, \gamma', \mu, \phi)$
5. $p(Y_L|Y, \alpha, \beta, \gamma', \mu, \phi)$
6. $p(\gamma|Y, Y_L, \alpha, \beta, \mu, \phi)$

(d) Trim (Sampler 2.3)

1. $\mathcal{M}_{\mu|Y,\beta,\gamma,\phi}(\mu|Y, \beta', \gamma', \mu', \phi')$
2. $\mathcal{M}_{\beta,\phi|Y,\gamma,\mu}(\beta, \phi|Y, \beta', \gamma', \mu, \phi')$
3. $p(\alpha|Y, \beta, \gamma', \mu, \phi)$
4. $p(Y_L|Y, \alpha, \beta, \gamma', \mu, \phi)$
5. $p(\gamma|Y, Y_L, \alpha, \beta, \mu, \phi)$

Figure 2.9: The three-stage framework used to derive Sampler 2.3 from its parent MH within Gibbs sampler. The parent sampler appears in (a). The conditioning in steps 2, 3, 5, and 6 is reduced in (b). The steps are permuted in (c) to allow redundant draws of $Y_L^\star$, $\alpha^\star$, $\beta^\star$, and $\phi^\star$ to be trimmed in Steps 1–4. The resulting proper MH within PCG sampler, i.e., Sampler 2.3, appears in (d).

where $r_{\text{not blocked}}$ is the acceptance probability of Step j in Sampler Fragment 6, where no blocking occurs. The result in (2.11) means that Sampler Fragments 6 and 7 are identical in terms of their update of ψ_1, but whereas Sampler Fragment 6 updates ψ_2 with a new value at every iteration, blocking of Sampler Fragment 7 causes ψ_2 to only be updated if ψ_1 is updated. Thus, we expect the blocking strategy of Sampler Fragment 7 to reduce the efficiency of the sampler, and contrary to general advice regarding blocking (e.g., Liu *et al.*, 1994), the blocking strategy of Sampler Fragment 7 should be avoided.

Together, the results in Section 2.2.3 discourage the combining of an MH update and a direct draw from a conditional distribution into a single MH update. This advice applies equally to ordinary Gibbs samplers with simple generalization.

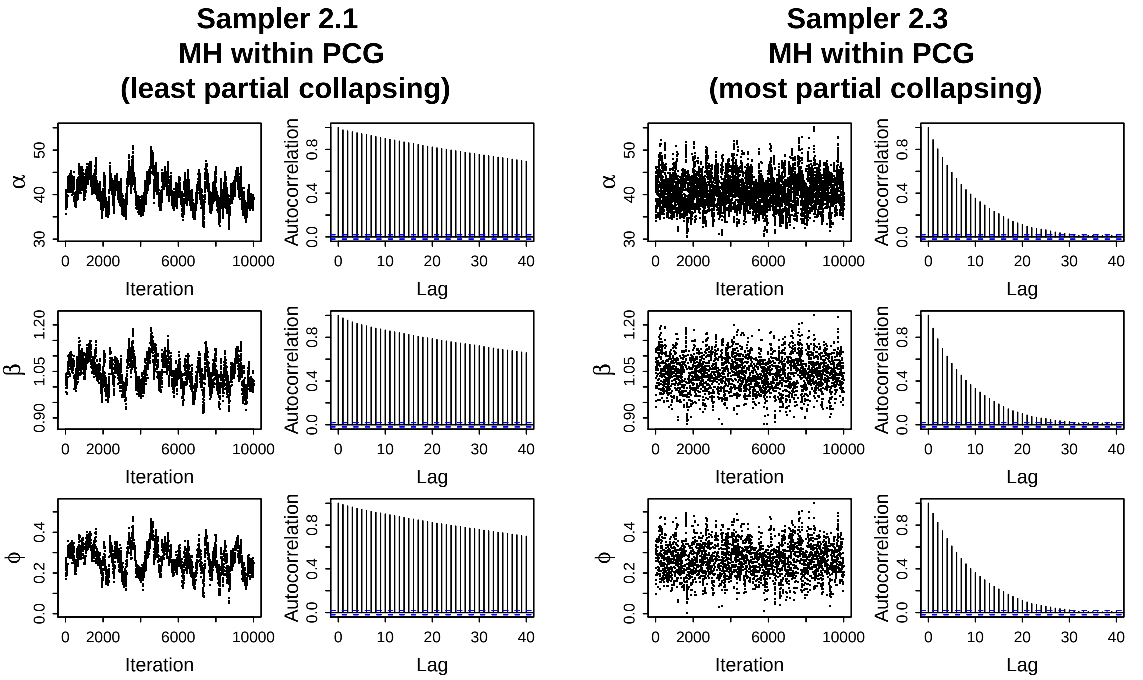

Figure 2.10: The sampling results of Samplers 2.1 and 2.3. The left two columns are the time-series and autocorrelation plots for the posterior draws of α, β, and ϕ respectively from Sampler 2.1, whereas the right two columns are those from Sampler 2.3. Sampler 2.3 performs significantly better than Sampler 2.1.

2.3 Illustration Examples

2.3.1 Spectral analysis in high-energy astrophysics

As pointed out in Section 2.1, the standard Gibbs sampler for fitting the spectral model breaks down. To solve this problem, we construct three proper MH within PCG samplers, i.e., Samplers 2.1–2.3 in Figure 2.1. The three samplers have the common parent Gibbs sampler (see top-left panel of Figure 2.3 or 2.9), but different degrees of partial collapsing. Sampler 2.1 has the least partial collapsing, while Sampler 2.3 has the most. We present using the three-stage framework to derive Sampler 2.3 from its parent sampler in Figure 2.9. (The derivation of Sampler 2.2 appears in Figure 2.3 and that of Sampler 2.1 is omitted to save space.)

Section 2.2.2 uses a simulation study of the spectral model in (2.1) to illustrate a potential problem with the blocking strategy in Sampler Fragment 7. Here we use the same simulation study and starting value settings to illustrate the improved convergence properties of Samplers 2.1–2.3 relative to their parent Gibbs sampler. The only difference is that for each sampler here, a chain of 20,000 iterations is run with a burnin of 10,000 iterations.

The convergence properties of α, β, and ϕ using Samplers 2.1 and 2.3 are compared in Figure 2.10; γ and μ converge well for all three samplers. As noted in Section 2.1, all three MH within PCG samplers outperform the parent MH Gibbs sampler, since the latter does not converge to the target. Sampler 2.3 performs much better than Sampler 2.1 in terms of the mixing and autocorrelations of α, β, and ϕ. The performance of Sampler 2.2 is better than Sampler 2.1, but not as good as Sampler 2.3. (The results of Sampler 2.2 are omitted in Figure 2.10.) To further compare the convergence by accounting for computational time, we estimate the ESS/sec of α, β, and ϕ in Table 2.1. For each parameter, both Samplers 2.2 and 3.3 produce larger ESS/sec than Sampler 2.1, and Sampler 2.3 has the largest value among the three samplers. As stated in Section 1.3 of Chapter 1, the larger the ESS per second, the more efficient is the sampler. By this measurement, we confirm that Sampler 2.3 is most efficient in improving convergence properties. These results show that proper MH within PCG samplers outperform their parent Gibbs sampler in computational efficiency and a higher degree of partial collapsing can improve the convergence even further.

2.3.2 Gaussian hierarchical model in supernova cosmology

We now consider the Gaussian hierarchical model for supernova cosmology mentioned in Chapter 1. We assume the absolute magnitudes of Type Ia SNe follow a Gaussian population distribution, that is,

$$M_i^o \overset{\text{iid}}{\sim} \mathrm{N}(M_0, \sigma_0^2), \text{ for } i = 1, \ldots, n. \tag{2.12}$$

Since the absolute magnitudes are similar, σ_0 is relatively small, but still too large to use Type Ia SNe as distance indicators without further adjustment. This intrinsic variability

	α	β	ϕ
Sampler 2.1	0.066	0.080	0.067
Sampler 2.2	0.096	0.101	0.098
Sampler 2.3	0.535	0.579	0.517

Table 2.1: The ESS per second of α, β, and ϕ for Samplers 2.1–2.3. Sampler 2.3 has the largest values for all the three parameters.

in the absolute magnitudes is due to variations in the properties of the progenitor star (e.g., mass and composition) and/or its environment. Fortunately, we can adjust for two covariates, the stretch parameter, x_i, and the color correction parameter, c_i, to reduce this scatter; these empirical adjustments are known as "Phillips corrections", see Phillips (1993) and Phillips *et al.* (1999) for details. Specifically,

$$M_i^o = M_i - \alpha x_i + \beta c_i, \text{ for } i = 1, \ldots, n, \tag{2.13}$$

with $M_i \overset{\text{iid}}{\sim} \text{N}(M_0, \sigma^2_{\text{res}})$,where M_i is the adjusted absolute magnitude and $\sigma^2_{\text{res}} \ll \sigma^2_0$. (Because of their similar adjusted absolute magnitudes, Type Ia SNe are called "standardizable candles".)

For Type Ia SN i ($i = 1, \ldots, n$), four quantities are observed with error, the apparent magnitude $\hat{m}_{Bi}$, the *observed* stretch and color correction parameters, $\hat{x}_i$ and $\hat{c}_i$, and the redshift z_i[*]. That is,

$$\begin{pmatrix} \hat{c}_i \\ \hat{x}_i \\ \hat{m}_{Bi} \end{pmatrix} \overset{\text{ind}}{\sim} \text{N}\left[\begin{pmatrix} c_i \\ x_i \\ m_{Bi} \end{pmatrix}, \hat{C}_i \right], \text{ for } i = 1, \ldots, n. \tag{2.14}$$

Because its measurement error is very small, in this article, we assume z_i is known. For

[*] The raw data are time-series observations of the evolving SN explosion in each of several color bands. These observations are summarized into the apparent magnitude, stretch parameter and color parameter using the SALT-II method (Guy *et al.*, 2007). The apparent magnitude is the peak magnitude in the B-band.

MH within Gibbs (Sampler 2.7)	MH within PCG (Sampler 2.8)
1. $p(\xi, X \mid Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$	1. $\mathcal{M}_{\mathscr{C} \mid Y,\alpha,\beta,\Sigma_P}(\mathscr{C} \mid Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$
2. $\mathcal{M}_{\mathscr{C} \mid Y,\xi,X,\alpha,\beta,\Sigma_P}(\mathscr{C} \mid Y, \xi, X, \mathscr{C}', \alpha', \beta', \Sigma'_P)$	2. $\mathcal{M}_{\alpha,\beta \mid Y,\mathscr{C},\Sigma_P}(\alpha, \beta \mid Y, \mathscr{C}, \alpha', \beta', \Sigma'_P)$
3. $p(\alpha, \beta \mid Y, \xi, X, \mathscr{C}, \Sigma'_P)$	3. $p(\xi, X \mid Y, \mathscr{C}, \alpha, \beta, \Sigma'_P)$
4. $p(\Sigma_P \mid Y, \xi, X, \mathscr{C}, \alpha, \beta)$	4. $p(\Sigma_P \mid Y, \xi, X, \mathscr{C}, \alpha, \beta)$

Figure 2.11: Samplers 2.7 and 2.8. The left and right panels show the steps of the MH within Gibbs sampler (Sampler 2.7) and the proper MH within PCG sampler (Sampler 2.8) for fitting the cosmological hierarchical model.

illustration, we ignore the small correlations among the observed quantities and take the matrix $\hat{C}_i$ to be diagonal, i.e., $\hat{C}_i = \text{Diag}\left(\hat{\sigma}^2_{ci}, \hat{\sigma}^2_{xi}, \hat{\sigma}^2_{mBi}\right)$. The distance modulus is defined to be $\mu_i = m_{Bi} - M^o_i$, so that (2.13) can be written as

$$m_{Bi} = \mu_i + M_i - \alpha x_i + \beta c_i, \text{ for } i = 1, \ldots, n. \tag{2.15}$$

This forms the first level of our hierarchical model and because of (2.14), it can be viewed as an *errors-in-variables* regression model (Carroll *et al.*, 2006). The second level of the hierarchical model describes the population distribution of the SNe,

$$M_i \overset{\text{iid}}{\sim} \text{N}(M_0, \sigma^2_{\text{res}}),\ x_i \overset{\text{iid}}{\sim} \text{N}(x_0, R^2_x),\ c_i \overset{\text{iid}}{\sim} \text{N}(c_0, R^2_c), \text{ for } i = 1, \ldots, n. \tag{2.16}$$

In a Friedman-Robertson-Walker cosmology, the distance modulus, μ_i, is predicted as a deterministic function of the redshift z and the cosmological parameter $\mathscr{C} = (\Omega_m, \Omega_\Lambda \text{ or } w, \Omega_\kappa, H_0)$, where Ω_m is the total matter density, Ω_Λ is the dark energy density, w is the dark energy equation of state, and H_0 is the Hubble constant. Specifically,

$$\mu_i = \mu_i(z_i, \mathscr{C}) = 25 + 5\log_{10}\left[\frac{c}{H_0} d_L(z_i, \mathscr{C})\right], \tag{2.17}$$

where the speed of light $c = 3 \times 10^5$ km/s, and

$$d_L(z_i, \mathscr{C}) = \frac{(1+z_i)}{\sqrt{|\Omega_\kappa|}} \text{sinn}\left\{\sqrt{|\Omega_\kappa|}\int_0^{z_i}\left[(1+z')^3\Omega_m + \Omega_{\text{DE}}(z') + (1+z')^2\Omega_\kappa\right]^{-1/2} \text{d}z'\right\}, \tag{2.18}$$

where Ω_κ is the curvature parameter; $\text{sinn}(x) = x$, $\text{sinn}(x) = \sin(x)$, or $\text{sinn}(x) = \sinh(x)$

(a) parent MH within Gibbs (Sampler 2.7)	(b) Marginalization
1. $p(\xi, X \vert Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$	1. $p(\xi^\star, X^\star \vert Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$
2. $\mathcal{M}_{\mathscr{C} \vert Y,\xi,X,\alpha,\beta,\Sigma_P}(\mathscr{C} \vert Y, \xi, X, \mathscr{C}', \alpha', \beta', \Sigma'_P)$	2. $\mathcal{M}^\star_{\mathscr{C},\xi,X \vert Y,\alpha,\beta,\Sigma_P}(\mathscr{C}, \xi^\star, X^\star \vert Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$
3. $p(\alpha, \beta \vert Y, \xi, X, \mathscr{C}, \Sigma'_P)$	3. $\mathcal{M}^\star_{\alpha,\beta,\xi,X \vert Y,\mathscr{C},\Sigma_P}(\alpha, \beta, \xi, X \vert Y, \mathscr{C}, \alpha', \beta', \Sigma'_P)$
4. $p(\Sigma_P \vert Y, \xi, X, \mathscr{C}, \alpha, \beta)$	4. $p(\Sigma_P \vert Y, \xi, X, \mathscr{C}, \alpha, \beta)$

(c) Permute	(d) Trim (Sampler 2.8)
1. $\mathcal{M}^\star_{\mathscr{C},\xi,X \vert Y,\alpha,\beta,\Sigma_P}(\mathscr{C}, \xi^\star, X^\star \vert Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$	1. $\mathcal{M}_{\mathscr{C} \vert Y,\alpha,\beta,\Sigma_P}(\mathscr{C} \vert Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$
2. $\mathcal{M}^\star_{\alpha,\beta,\xi,X \vert Y,\mathscr{C},\Sigma_P}(\alpha, \beta, \xi^\star, X^\star \vert Y, \mathscr{C}, \alpha', \beta', \Sigma'_P)$	2. $\mathcal{M}_{\alpha,\beta \vert Y,\mathscr{C},\Sigma_P}(\alpha, \beta \vert Y, \mathscr{C}, \alpha', \beta', \Sigma'_P)$
3. $p(\xi, X \vert Y, \mathscr{C}, \alpha, \beta, \Sigma'_P)$	3. $p(\xi, X \vert Y, \mathscr{C}, \alpha, \beta, \Sigma'_P)$
4. $p(\Sigma_P \vert Y, \xi, X, \mathscr{C}, \alpha, \beta)$	4. $p(\Sigma_P \vert Y, \xi, X, \mathscr{C}, \alpha, \beta)$

Figure 2.12: The three-stage framework used to derive Sampler 2.8 from its parent MH within Gibbs sampler, i.e., Sampler 2.7. The parent sampler appears in (a) with Step 2 requiring MH. Steps 2 and 3 are marginalized in (b). The steps are permuted in (c) to allow redundant draws of (ξ, X) to be trimmed in Steps 1–2. The resulting proper MH within PCG sampler, that is, Sampler 2.8, appears in (d).

for $\Omega_\kappa = 0$, $\Omega_\kappa < 0$, and $\Omega_\kappa > 0$, respectively. For a general dark energy equation of state as a function of redshift, $w(z)$, we express

$$\Omega_{\mathrm{DE}}(z) = \Omega_\Lambda \exp\left[3\int_0^z \frac{1+w(z')}{1+z'}\mathrm{d}z'\right]. \tag{2.19}$$

In our analyses, we either assume a flat Universe with $w(z)$ equal to a constant other than -1 ($\Omega_\kappa = 0$, the wCDM model) or a curved Universe with a cosmological constant $w(z) = -1$ (the ΛCDM model). In either case, $w(z) = w$ becomes a time-independent constant, and thus Ω_κ is completely determined by Ω_Λ and Ω_m via $\Omega_\kappa = 1 - \Omega_m - \Omega_\Lambda$. We assume the ΛCDM model all through this manuscript. In addition, because the Hubble constant H_0 is completely degenerate with M_0, we fix it at the value determined by other measurements.

Finally, we specify weakly informative prior distributions for model parameters,

$$M_0 \sim \mathrm{N}(M_m, \sigma^2_{M_0}),\ x_0 \sim \mathrm{N}(0, \sigma^2_{x_0}),\ c_0 \sim \mathrm{N}(0, \sigma^2_{c_0}), \tag{2.20}$$

where $M_m = -19.3$, $\sigma_{M_0} = 2$, $\sigma_{x_0} = 10$, and $\sigma_{c_0} = 1$. According to March *et al.* (2011), these variances are large enough to make the priors for M_0, x_0 and c_0 sufficiently diffuse. They also find that the choice of mean and variance in the prior distribution of M_0 has little influence on numerical results. Furthermore, we specify $\log(\sigma_{\text{res}}) \sim \text{Unif}(-5, 2)$, $\log(R_x) \sim \text{Unif}(-5, 2)$, $\log(R_c) \sim \text{Unif}(-5, 2)$, $\alpha \sim \text{Unif}(0, 1)$, $\beta \sim \text{Unif}(0, 4)$, $\Omega_m \sim \text{Unif}(0, 1)$, $\Omega_\Lambda \sim \text{Unif}(0, 2)$, and $w \sim \text{Unif}(-2, 0)$. We choose ranges of these uniform priors following March *et al.* (2011), which stated that they generously cover all plausible values of the parameters.

To simplify notation, we let Y denote the $(3n \times 1)$ vector of observed quantities, i.e., $Y = (\hat{c}_1, \hat{x}_1, \hat{m}_{B1}, \ldots, \hat{c}_n, \hat{x}_n, \hat{m}_{Bn})$, ξ denote the (3×1) mean vector of the distribution in the second level of the hierarchical model, i.e., $\xi = (c_0, x_0, M_0)$, X denote the $(3n \times 1)$ vector of latent variables, i.e., $X = (c_1, x_1, M_1, \ldots, c_n, x_n, M_n)$, Σ_C denote the $(3n \times 3n)$ observed variance-covariance matrix of Y, i.e., $\Sigma_C = \text{Diag}(\hat{C}_1, \ldots, \hat{C}_n)$, and Σ_P denote the $(3n \times 3n)$ population variance-covariance matrix of the latent quantities in X, i.e., $\Sigma_P = \text{Diag}(S, \ldots, S)$, where $S = \text{Diag}(R_c^2, R_x^2, \sigma_{\text{res}}^2)$.

To sample from the posterior distribution of the hierarchical model, we start with a standard MH within Gibbs sampler, i.e., Sampler 2.7, where each (sometimes multivariate) component is updated from its complete conditional distribution. We list the steps of Sampler 2.7 in the left panel of Figure 2.11. We sample $\mathscr{C}$ with the help of MH because its conditional is not a standard distribution. (It is evaluated numerically.) Unfortunately, both $\mathscr{C}$ and (α, β) exhibit poor convergence in this sampler. Because these two parameters are highly correlated with (ξ, X) *a posteri*, we derive a proper MH within PCG sampler, i.e., Sampler 2.8, to break this correlation and thus improve convergence. Sampler 2.8 updates both $\mathscr{C}$ and (α, β) without conditioning on (ξ, X) and its steps are listed in the right panel of Figure 2.11. Details of Samplers 2.7 and 2.8 appear in Appendix A. We verify the propriety of Sampler 2.8 using the three-stage framework in Figure 2.12.

To illustrate the relative efficiencies of Samplers 2.7 and 2.8, we use a data set consisting of 288 Type Ia SN observations compiled by Kessler *et al.* (2009). We run each of these

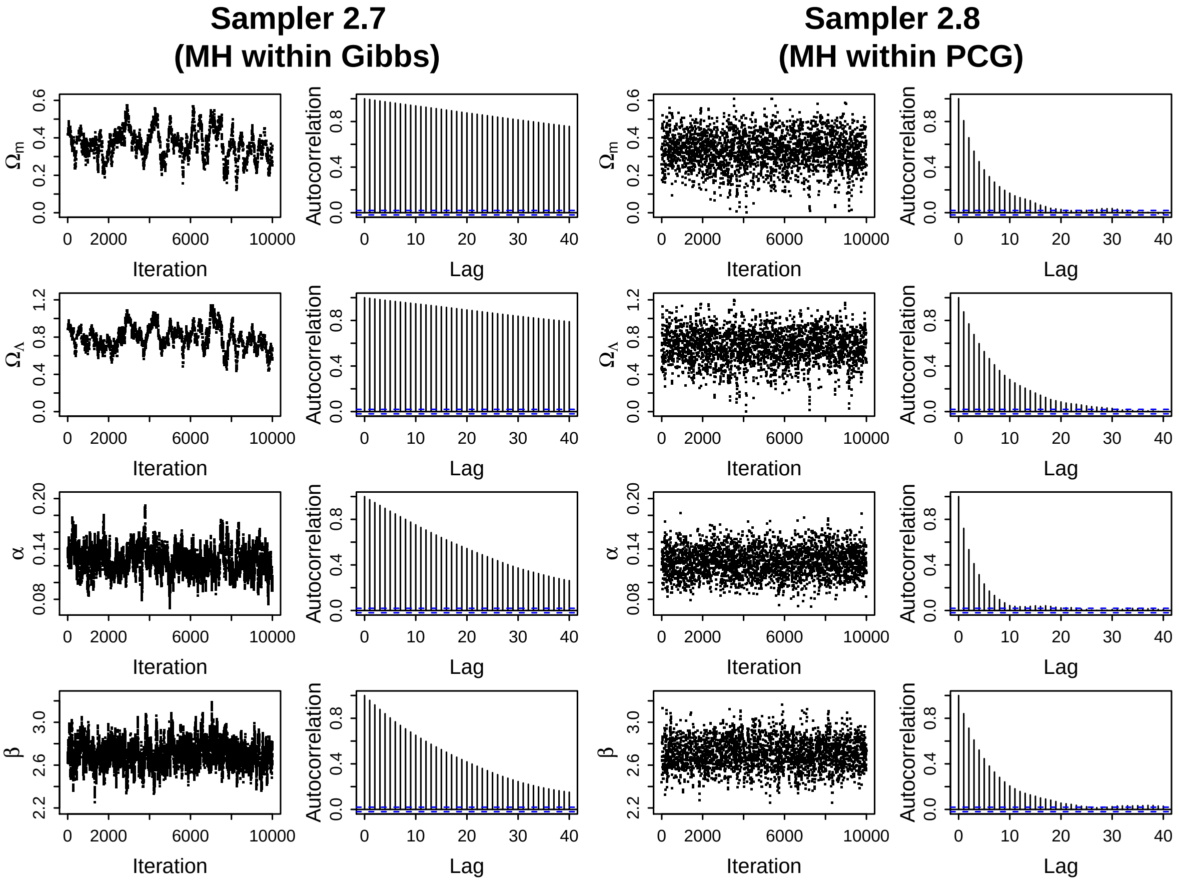

Figure 2.13: The sampling results of Samplers 2.7 and 2.8. The left two columns are the time-series and autocorrelation plots for the posterior draws of Ω_m, Ω_Λ, α and β respectively from Sampler 2.7, while the right two columns are those from Sampler 2.8. Sampler 2.8 converges much faster than Sampler 2.7.

two samplers for a chain of 11,000 iteration with a burn-in of 1,000, starting from the same sets of initial values.

In Figure 2.13, we display the time-series and autocorrelation plots for Ω_m, Ω_Λ, α, and β from both Sampler 2.7 (left two columns) and Sampler 2.8 (right two columns). For all four parameters, Sampler 2.8 produce chains with much faster mixing and lower autocorrelation than Sampler 2.7. We present the ESS/sec of the four parameters in Table 2.2. For each parameter, Samplers 2.8 has larger ESS/sec than Sampler 2.7. We conclude from these results that the MH within PCG sampler is efficient in improving the convergence of its parent Gibbs sampler.

	$\boldsymbol{\Omega_m}$	$\boldsymbol{\Omega_\Lambda}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
Sampler 2.7	0.002	0.001	0.008	0.010
Sampler 2.8	0.037	0.023	0.051	0.029

Table 2.2: The ESS per second of Ω_m, Ω_Λ, α, and β for Samplers 2.7 and 2.8. Sampler 2.8 has larger values for all the four parameters than Sampler 2.7.

2.3.3 Factor analysis model

Consider the following factor-analysis model:

$$Y_i \overset{\text{ind}}{\sim} \text{N}_p\left[\beta Z_i, \Sigma\right], \text{ for } i = 1, \dots, n, \tag{2.21}$$

where Y_i is a $(p \times 1)$ vector of observation; Z_i is a $(q \times 1)$ vector with $Z_i \overset{\text{iid}}{\sim} \text{N}_q(0, I)$; β is the $(p \times q)$ factor-loading matrix; and $\Sigma = \text{Diag}(\sigma_1^2, \dots, \sigma_p^2)$ is the variance-covariance matrix. We set $Y = (Y_1, \dots, Y_n)$ and $Z = (Z_1, \dots, Z_n)$. To identify the model, following Geweke and Zhou (1996), we specify that the first q rows of β as a lower-triangular $(q \times q)$ matrix with positive diagonal elements, that is, $\beta_{kj} = 0$ and $\beta_{kk} > 0$, for $j = (k+1, \dots, q)$ and $k = 1, \dots, q$. The marginal distribution of Y_i is $\text{N}_p(0, S)$, where $S = \beta\beta^T + \Sigma$ is the variance-covariance matrix. There are $p(p+1)/2$ free parameters in a $(p \times p)$ variance-covariance matrix, but a total of $l = p(q+1) - q(q-1)/2$ different parameters in S. Thus we choose q to ensure that $l \leq p(p+1)/2$. We specify conjugate prior distributions for (β, Σ), specifically, $p(\beta) \propto 1$ and $\sigma_j^2 \overset{\text{ind}}{\sim} \text{Inv-Gamma}(0.01, 0.01)$, for $j = 1, \dots, p$. The parameters Z, β, and Σ are unknown and we wish to sample from their joint posterior distribution.

We simulate a dataset for this factor analysis model. Particularly, we set $p = 6$, $q = 2$, and $n = 100$; σ_j^2 $(j = 1, \dots, 6)$ are generated from Inv-Gamma$(1, 0.5)$, and β_{jk} $(j = 1, \dots, 6,\ k = 1, 2, \text{ and } j \geq k)$ from $\text{N}(0, 3^2)$, subject to the constraint that β_{11} and β_{22} are positive.

To sample from the posterior distribution of factor analysis model (2.21) based on the simulated dataset above, we start with a three-step Gibbs sampler, which updates

Gibbs (Sampler 2.9)	**MH within PCG (Sampler 2.10)**
1. $p(Z_i\|Y,\beta',\Sigma')$, for $i=1,\dots,100$ 2. $p(\sigma_j^2\|Y,Z,\beta')$, for $j=1,\dots,6$ 3. $p(\beta_j\|Y,Z,\Sigma)$, for $j=1,\dots,6$, where β_j is the jth row of β	j. $\mathcal{M}_{\sigma_j^2\|Y,\sigma_{-j}^2,\beta}(\sigma_j^2\|Y,\Sigma',\beta')$, for $j=1,\dots,4$ 5. $p(Z_i\|Y,\beta',\Sigma')$, for $i=1,\dots,100$ 6. $p(\sigma_j^2\|Y,Z,\beta')$, for $j=5,6$ 7. $p(\beta_j\|Y,Z,\Sigma)$, for $j=1,\dots,6$ Note that when updating σ_j^2 ($j=1,\dots,4$), $\sigma_{-j}^2=(\sigma_1^2,\dots,\sigma_{j-1}^2,\sigma_{j+1}^2,\dots,\sigma_6^2)$ and $\Sigma'=\text{Diag}(\sigma_1^2,\dots,\sigma_{j-1}^2,\sigma_j^{2'},\sigma_{j+1}^{2'},\dots,\sigma_6^{2'})$; when updating Z, $\Sigma'=\text{Diag}(\sigma_1^2,\dots,\sigma_4^2,\sigma_5^{2'},\sigma_6^{2'})$.

Figure 2.14: Samplers 2.9 and 2.10. The left and right panels show the steps of the Gibbs sampler (Sampler 2.9) and the proper MH within PCG sampler (Sampler 2.10) for fitting the factor analysis model.

each of Z, Σ, and β from its complete conditional distribution, i.e., Sampler 2.9 in the left panel of Figure 2.14. Unfortunately, both β and Σ exhibit poor convergence with this sampler. Because Σ and Z are highly correlated *a posteri*, we design a PCG sampler, i.e., Sampler 2.10 in the right panel of Figure 2.14, to break the correlation and boost the efficiency of convergence. In particular, Sampler 2.10 updates σ_1^2–σ_4^2 without conditioning on Z; these reduced updates require MH steps. Because σ_5^2 and σ_6^2 converge well with Sampler 2.9, we do not alter their updates in Sampler 2.10. We derive Sampler 2.10 from its parent Gibbs sampler, i.e., Sampler 2.9, strictly following the three-stage framework, see Figure 2.15. Thus Sampler 2.10 is a proper MH within PCG sampler.

We run 50,000 iterations for each of Samplers 2.9 and 2.10 with a burnin of 10,000 using the same starting values. Figure 2.16 compares Samplers 2.9 and 2.10 in terms of mixing and autocorrelation of $\log(\sigma_1^2)$–$\log(\sigma_4^2)$; the left two columns correspond to Sampler 2.9 and the right two columns correspond to Sampler 2.10. The computational advantage of Sampler 2.10 is evident. Furthermore, we display the ESS/sec of $\log(\sigma_1^2)$–$\log(\sigma_4^2)$ in Table 2.3. For each parameter, Samplers 2.10 has larger ESS/sec than Sampler 2.9. Thus the proper MH within PCG sampler outperforms its parent Gibbs sampler in computational efficiency. While highly effective for Σ, PCG samplers do not appreciably

(a) Parent MH within Gibbs Sampler

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

2. $p(\sigma_j^2|Y, Z, \beta')$, for $j = 1, \ldots, 6$

3. $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 6$, where β_j is the jth row of β

(b) Marginalization

1. $p(Z_i^\star|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$

$j+1$. $\mathcal{M}^\star_{Z,\sigma_j^2|Y,\sigma_{-j}^2,\beta}(Z^\star, \sigma_j^2|Y, \Sigma', \beta')$, for $j = 1, 2, 3$

5. $\mathcal{M}^\star_{Z,\sigma_4^2|Y,\sigma_{-4}^2,\beta}(Z, \sigma_4^2|Y, \Sigma', \beta')$

6. $p(\sigma_j^2|Y, Z, \beta')$, for $j = 5, 6$

7. $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 6$

(c) Permute

j. $\mathcal{M}_{Z,\sigma_j^2|Y,\sigma_{-j}^2,\beta}(Z^\star, \sigma_j^2|Y, \Sigma', \beta')$, for $j = 1, \ldots, 4$

5. $p(Z_i|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$

6. $p(\sigma_j^2|Y, Z, \beta')$, for $j = 5, 6$

7. $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 6$

(d) Trim

j. $\mathcal{M}_{\sigma_j^2|Y,\sigma_{-j}^2,\beta}(\sigma_j^2|Y, \Sigma', \beta')$, for $j = 1, \ldots, 4$

5. $p(Z_i|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$

6. $p(\sigma_j^2|Y, Z, \beta')$, for $j = 5, 6$

7. $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 6$

Figure 2.15: The three-phase framework to derive Sampler 2.10 from its parent Gibbs sampler, i.e., Sampler 2.9. The parent Gibbs sampler is in (a); Steps 2–5 are marginalized in (b); and the steps are permuted in (c) to allow redundant draws of $Z^\star$ to be trimmed in Steps 1–4 The resulting Sampler 2.10 appears in (d).

improve the convergence of β. We will address this problem in Chapter 3.

2.4 Discussion

Since introduced in 2008, the PCG sampler has been deployed to improve the convergence properties of numerous Gibbs-type samplers in a variety of applied settings. As with ordinary Gibbs samplers, MH updates are sometimes required within PCG samplers. However, unlike an ordinary Gibbs samplers, embedding MH steps into a PCG sampler may upset its stationary distribution. This has led to a number of improper samplers in the literature. This chapter illustrates the subtleties of introducing MH updates into PCG samplers, offers a strategy for guaranteeing the propriety of such samplers, and provides advice on the choice between alternative implementations of MH within PCG samplers. Some of the advice is contrary to what is commonly un-

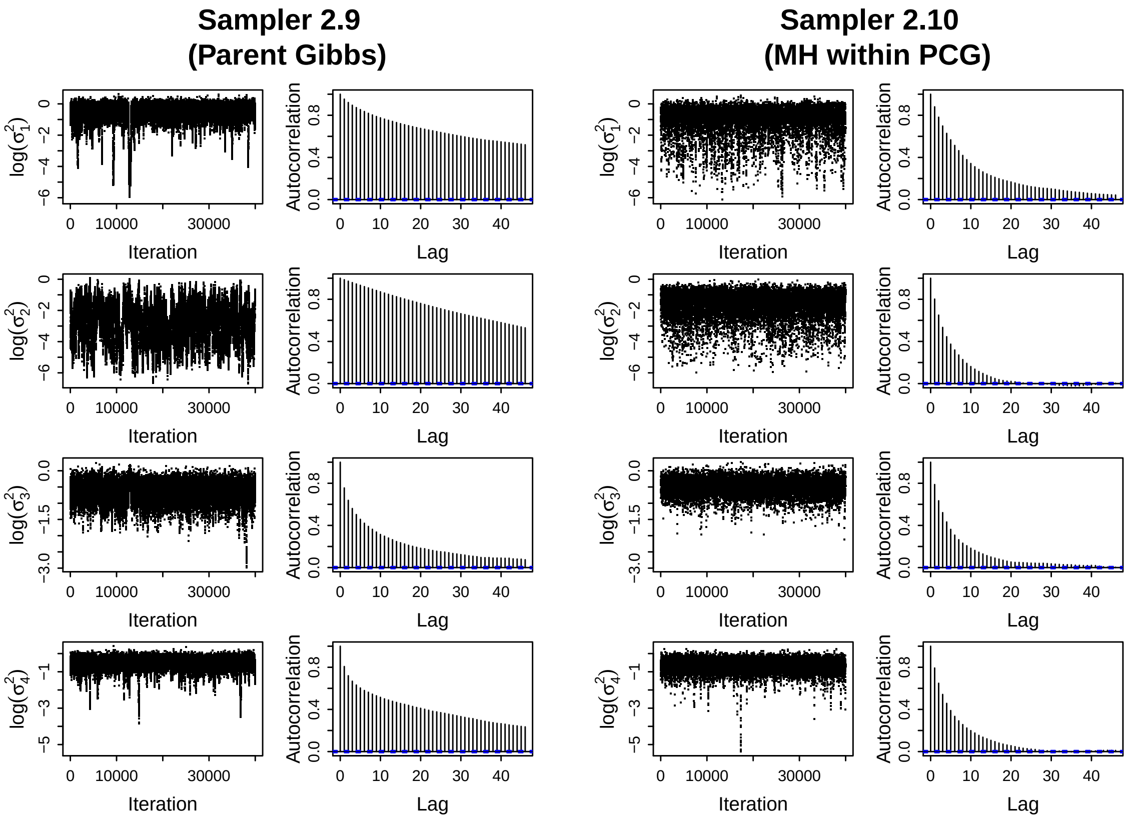

Figure 2.16: The sampling results of Samplers 2.9 and 2.10. The left two columns are the time-series and autocorrelation plots for the posterior draws of σ_1^2–σ_4^2 from Sampler 2.9, while the right two columns are those from Sampler 2.10. Sampler 2.10 is efficient in improving the convergence of Σ.

derstood. For example, blocking steps within a Gibbs-type sampler should improve its convergence. We find, however, that this may not be true if MH is involved.

Marginalization in one or more steps of a Gibbs sampler can only improve the convergence (van Dyk and Park, 2008). When MH is required to implement the reduced steps, however, the overall performance of the PCG sampler may decay, especially if the choice for the MH jumping rule is poor. Thus, there is a natural trade-off between the computational complexity of MH and the improved efficiency offered by partial collapsing. Generally speaking, some trial and error may be needed to negotiate this trade-off. In practice, it is ideal if we can start with an MH within Gibbs sampler, which already involves MH and can be improved by partial collapsing without any added complexity. As implied by numerical examples in this chapter, however, there also exist a variety of

	$\log(\sigma_1^2)$	$\log(\sigma_2^2)$	$\log(\sigma_3^2)$	$\log(\sigma_4^2)$
Sampler 2.9	0.129	0.142	0.958	0.372
Sampler 2.10	0.931	1.777	1.686	1.578

Table 2.3: The ESS per second of $\log(\sigma_1^2)$–$\log(\sigma_4^2)$ for Samplers 2.9 and 2.10. Sampler 2.10 has larger values for all the four parameters than Sampler 2.9.

cases where the reduced correlation afforded by partial collapsing makes up the lost efficiency caused by more MH updates. The expectation is that our strategies can extend the application of PCG samplers in practice and provide researchers with additional tools to improve the convergence of Gibbs-type samplers.

Embedding the MH algorithm into PCG samplers can be regarded as a specific example of the combining strategy, which encompasses more flexibility and power to facilitate the implementation and boost the efficiency of Gibbs-type samplers by integrating two or more strategies into one coherent proper sampler. We will elucidate the combining strategy in the next chapter, i.e., Chapter 3.

3

Combining Strategies for Improving the Convergence of Gibbs-type Samplers

To simplify the implementation and improve the efficiency of Gibbs-type samplers, we combine different strategies into one sampler. For example, in Chapter 1, we combine PCG and MH. In this chapter, we construct a general framework to combine different algorithms into a coherent sampler. We use both theoretical arguments and numerical examples to illustrate that the combined samplers maintain their target stationary distributions and can only improve the convergence properties of their parent Gibbs-type samplers. In Section 3.1, we use the factor analysis model described in Section 2.3.3 of Chapter 2 as the motivating example to show the benefit of combining different acceleration strategies into a single sampler. In Section 3.2, we construct the framework of combining acceleration strategies and provide theoretical arguments to prove that under our construction, the combined sampler i) maintains the target stationary distribution and ii) is more efficient than any of its component algorithms. In Section 3.3, we use the factor analysis model, the hierarchical model in supernova cosmology, a hierarchical t model, and a hierarchical probit model to illustrate the construction of combined samplers and show their computational advantages. In Section 3.4, we summarize the combining strategy and point out that combining different acceleration strategies into a single sampler is a specific application of the surrogate distribution framework, which is the topic of Chapter 4.

3.1 Motivating Example

Here we outline our computational strategy for sampling from the posterior distribution of the factor analysis model introduced in Section 2.3.3 of Chapter 2, in order to motivate the advantage of combining several strategies in a single sampler. Recall that in Section 2.3.3 of Chapter 2, to sample from the posterior distribution of the factor analysis model, we first implement a three-step Gibbs sampler, by updating each of Z, Σ and β from their complete conditional distributions. Unfortunately, both β and Σ exhibit poor convergence in this sampler. Thus we design a PCG sampler which updates some (diagonal) components of Σ without conditioning on Z. Although this strategy is highly effective for Σ, PCG does not appreciably improve the convergence of β. To address this, we consider implementing ASIS. Because the distribution $Z_i \overset{\text{iid}}{\sim} \text{N}_q(0, I)$ does not depend on β or Σ, Z is an ancillary augmentation for both β and Σ. It is more difficult to derive a sufficient augmentation for β and Σ. However, conditioning on Σ, we can easily find a sufficient augmentation for β; letting $W = \beta Z$, $Y_i|W_i \overset{\text{ind}}{\sim} \text{N}(W_i, \Sigma)$ is free of β and W is a (conditional) sufficient augmentation. Thus we implement an ASIS algorithm for β conditioning on Σ, see Sampler 3.1 in the left panel of Figure 3.1. (Details of Sampler 3.1 and other samplers in this chapter appear in Appendix B.) This involves i) updating Z and Σ from their complete conditional distributions; ii) sampling β conditioning on Z and Σ, and transforming from Z to W; iii) updating β again but conditioning on W and Σ, and finally transforming W back to Z. Numerically we find

```
ASIS (Sampler 3.1)
1. $p(Z_i^\star|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$
2. $p(\sigma_j^2|Y, Z^\star, \beta')$, for $j = 1, \ldots, 6$
3. $p(\beta_j^\star|Y, Z^\star, \Sigma)$, for $j = 1, \ldots, 6$;
   set $W_i = \beta^\star Z_i^\star$, for $i = 1, \ldots, 100$;
   denote the $(2 \times 2)$ sub-matrix of $\beta^\star$ consisting of the first two rows of $\beta^\star$ by $\tilde{\beta}$
4. $p(\beta|Y, W, \Sigma)$, consisting of
   • $p(\tilde{\beta}^{\mathrm{T}}\tilde{\beta}|Y, W, \Sigma)$;
     set $\tilde{\beta}$ to the Cholesky factor of $\tilde{\beta}^{\mathrm{T}}\tilde{\beta}$,
   • $p(\beta_j|Y, W, \tilde{\beta}, \Sigma)$, for $j = 3, \ldots, 6$;
   set $Z_i = \tilde{\beta}^{-1}\tilde{W}_i$, for $i = 1, \ldots, 100$, where $\tilde{W}_i$ is a $(2 \times 1)$ vector consisting of the first two components of $W_i$
```

```
MH within PCG+ASIS (Sampler 2.10)
j. $\mathcal{M}_{\sigma_j^2|Y,\sigma_{-j}^2,\beta}(\sigma_j^2|Y, \Sigma', \beta')$, for $j = 1, \ldots, 4$
5. $p(Z_i|Y, \beta', \Sigma')$, for $i = 1, \ldots, 100$
6. $p(\sigma_j^2|Y, Z, \beta')$, for $j = 5, 6$
7. $p(\beta_j|Y, Z, \Sigma)$, for $j = 1, \ldots, 6$;
   set $W_i = \beta^\star Z_i^\star$, for $i = 1, \ldots, 100$
8. $p(\beta|Y, W, \Sigma)$, consisting of
   • $p(\tilde{\beta}^{\mathrm{T}}\tilde{\beta}|Y, W, \Sigma)$;
     set $\tilde{\beta}$ to the Cholesky factor of $\tilde{\beta}^{\mathrm{T}}\tilde{\beta}$,
   • $p(\beta_j|Y, W, \tilde{\beta}, \Sigma)$, for $j = 3, \ldots, 6$;
   set $Z_i = \tilde{\beta}^{-1}\tilde{W}_i$, for $i = 1, \ldots, 100$
```

Figure 3.1: Samplers 3.1 and 3.2. The left and right panels show the steps of the ASIS sampler (Sampler 3.1) and the sampler combining PCG and ASIS (Sampler 3.2) for fitting the factor analysis model.

that this conditional ASIS algorithm is remarkable in improving the convergence of β, but has little effect on Σ. (This kind of conditional ASIS algorithm is called as *partial component-wise interweaving strategy* in Yu and Meng (2011).)

Since PCG and ASIS have none-overlapping effects in improving the convergence of β and Σ, we naturally consider combining these two algorithms into one single sampler to improve the convergence of both parameters simultaneously. That is, conditional on β, we update Σ parameters with PCG, and then conditional on Σ, we sample β via ASIS. This results in Sampler 3.2 in the right panel of Figure 3.1. We find that this combined strategy indeed simultaneously improves the convergence of β and Σ with a fairly small extra computational cost.

Motivated by this example, we consider obtaining more capacity to improve the convergence of Gibbs-type samplers by combining several acceleration strategies into a single coherent sampler. We construct a general framework to achieve this goal and guarantee that under our framework, the combined sampler maintains the target stationary

Gibbs Sampler	**PCG Sampler**	**MDA Sampler**	**Haar PX-DA**	**ASIS Sampler**
1. $p(\psi_1\|\psi_2', \psi_3')$ 2. $p(\psi_2\|\psi_1, \psi_3')$ 3. $p(\psi_3\|\psi_1, \psi_2)$	1. $p(\psi_1\|\psi_2')$ 2. $p(\psi_3\|\psi_1, \psi_2')$ 3. $p(\psi_2\|\psi_1, \psi_3)$	1. $\tilde{p}(\alpha^\star, \tilde{\psi}_1\|\psi_2', \psi_3')$ 2. $\tilde{p}(\alpha, \psi_2\|\tilde{\psi}_1, \psi_3')$; set $\psi_1 = \mathcal{G}_\alpha^{-1}(\tilde{\psi}_1)$ 3. $p(\psi_3\|\psi_1, \psi_2)$	1. $p(\psi_1^\star\|\psi_2', \psi_3')$ 2. $\tilde{p}(\alpha^\star\|\psi_1^\star, \psi_3')$; set $\psi_1 = \mathcal{G}_{\alpha^\star}^{-1}(\psi_1^\star)$ 3. $p(\psi_2\|\psi_1, \psi_3')$ 4. $p(\psi_3\|\psi_1, \psi_2)$	1. $p(\psi_1^\star\|\psi_2', \psi_3')$ 2. $p(\psi_2^\star\|\psi_1^\star, \psi_3')$; set $\tilde{\psi}_1 = \mathcal{H}_{\psi_2^\star}(\psi_1^\star)$ 3. $p(\psi_2\|\tilde{\psi}_1, \psi_3')$; set $\psi_1 = \mathcal{H}_{\psi_2}^{-1}(\tilde{\psi}_1)$ 4. $p(\psi_3\|\psi_1, \psi_2)$

Figure 3.2: Five samplers for updating from the distribution $p(\psi_1, \psi_2, \psi_3)$. They are the parent Gibbs sampler (first panel), PCG sampler (second panel), MDA sampler (third panel) with a proper working prior, Haar PX-DA sampler (fourth panel), and ASIS sampler (last panel), respectively.

distribution, and can only improve the convergence of the parent Gibbs sampler.

3.2 Combining Two or More Acceleration Strategies into One Sampler

3.2.1 Transition kernels for PCG, MDA (Haar PX-DA), and ASIS

Suppose we wish to sample from $p(\psi)$ with $\psi = \{\psi_1, \ldots, \psi_N\}$. As introduced in Section 1.1.2 of Chapter 1, the transition kernel of the standard Gibbs sampler for updating $p(\psi)$ is a product of N component transition kernels, that is, $\mathcal{K}(\psi|\psi') = \prod_{j=1}^N \mathcal{K}_j[\psi_j|; \mathcal{F}_j(\psi, \psi')]$, where $\mathcal{F}_j(\psi, \psi') = (\psi_1, \ldots, \psi_{j-1}, \psi_{j+1}', \ldots, \psi_N')$. The stationary distribution of each $\mathcal{K}_j$ is the corresponding complete conditional of the target distribution, i.e., $p(\psi_j|\psi_{-j})$. By implementing acceleration strategies on the Gibbs sampler, we equivalently replace $\mathcal{K}_j[\psi_j|; \mathcal{F}_j(\psi, \psi')]$ with a new kernel, $\mathcal{K}_j'[\psi_j|; \mathcal{F}_j'(\psi, \psi')]$, for some j. The stationary distribution of $\mathcal{K}_j'$ can be different from $p(\psi_j|\psi_{-j})$ so long as the overall stationary distribution of the sampler is maintained. We focus on three acceleration algorithms, that is, PCG, MDA (Haar PX-DA), and ASIS, and provide the explicit forms of their transition kernels.

Without loss of generality, we set $N = 3$, that is, $\psi = (\psi_1, \psi_2, \psi_3)$. The steps of the standard Gibbs sampler for updating $p(\psi_1, \psi_2, \psi_3)$ are presented in the first panel of

Figure 3.2. The transition kernel of the Gibbs sampler is

$$\mathcal{K}(\psi|\psi') = p(\psi_1|\psi_2', \psi_3')p(\psi_2|\psi_1, \psi_3')p(\psi_3|\psi_1, \psi_2). \tag{3.1}$$

To improve convergence properties of the parent Gibbs sampler, we consider another four samplers, which are a PCG sampler, an MDA sampler with a proper working prior, a Haar PX-DA sampler, and an ASIS sampler, respectively.

Transition kernel of PCG

Implementing PCG is equivalent to replacing some component transition kernels in (3.1) by kernels with reduced conditioning. For example, the PCG sampler in the second panel of Figure 3.2, which updates ψ_1 without conditioning on ψ_3, in fact replaces the component kernel $\mathcal{K}_1[\psi_1|; (\psi_2', \psi_3')] = p(\psi_1|\psi_2', \psi_3')$ in (3.1) with a new kernel, $\mathcal{K}_1'(\psi_1|; \psi_2') = p(\psi_1|\psi_2')$. The stationary distribution of $\mathcal{K}_1'(\psi_1|; \psi_2')$ is $p(\psi_1|\psi_2')$, whereas that of $\mathcal{K}_1[\psi_1|; (\psi_2', \psi_3')]$ is $p(\psi_1|\psi_2', \psi_3')$.

Transition kernel of MDA and Haar PX-DA

Suppose conditioning on ψ_3', ψ_1 is the augmented data for ψ_2. Thus Steps 1 and 2 of the Gibbs sampler in Figure 3.2 can be regarded as a DA algorithm conditioning on ψ_3'. We introduce a working parameter α into the model via a one-to-one mapping $\tilde{\psi}_1 = \mathcal{G}_\alpha(\psi_1)$.

First, we implement MDA on Steps 1 and 2 of the Gibbs sampler specifying a proper prior $p(\alpha)$ on α, see the third panel of Figure 3.2. With a proper working prior, implementing the MDA algorithm is equivalent to replacing first two component kernels in (3.1), $\mathcal{K}_1[\psi_1|; (\psi_2', \psi_3')] = p(\psi_1|\psi_2', \psi_3')$ and $\mathcal{K}_2[\psi_2|; (\psi_1, \psi_3')] = p(\psi_2|\psi_1, \psi_3')$, with

$$
\begin{aligned}
&\mathcal{K}'_{1,2}(\psi_1,\psi_2|\psi'_2;\psi'_3)\\
&=\int\int\int\tilde{p}(\alpha^\star,\tilde{\psi}_1|\psi'_2,\psi'_3)\tilde{p}(\alpha,\psi_1,\psi_2|\tilde{\psi}_1,\psi'_3)\mathrm{d}\alpha\mathrm{d}\alpha^\star\mathrm{d}\tilde{\psi}_1\\
&=\int\int\tilde{p}(\alpha^\star,\mathcal{G}_\alpha(\psi_1)|\psi'_2,\psi'_3)\tilde{p}(\alpha,\psi_2|\mathcal{G}_\alpha(\psi_1),\psi'_3)|\mathrm{J}(\psi_1|\alpha)|\mathrm{d}\alpha\mathrm{d}\alpha^\star\\
&=\int p(\alpha^\star)\int\tilde{p}(\mathcal{G}_\alpha(\psi_1)|\alpha^\star,\psi'_2,\psi'_3)\tilde{p}(\alpha,\psi_2|\mathcal{G}_\alpha(\psi_1),\psi'_3)|\mathrm{J}(\psi_1|\alpha)|\mathrm{d}\alpha\mathrm{d}\alpha^\star,
\end{aligned}
\tag{3.2}
$$

where $\mathrm{J}(\psi_1|\alpha)$ is the Jacobian matrix of the transformation $\psi_1=\mathcal{G}_\alpha^{-1}(\tilde{\psi}_1)$ conditioning on α. Updating (α,ψ_1,ψ_2) from $\tilde{p}(\alpha,\psi_1,\psi_2|\tilde{\psi}_1,\psi'_3)$, shown in the second row of (3.2), is equivalent to Step 2 of the MDA sampler in Figure 3.2, because sampling (α,ψ_1,ψ_2) from $\tilde{p}(\alpha,\psi_1,\psi_2|\tilde{\psi}_1,\psi'_3)$ consists of sampling (α,ψ_2) from $\tilde{p}(\alpha,\psi_2|\tilde{\psi}_1,\psi'_3)$ and then updating ψ_1 from $\tilde{p}(\psi_1|\alpha,\psi_2,\tilde{\psi}_1,\psi'_3)$, which is simply the transformation from $\tilde{\psi}_1$ to ψ_1. The stationary distribution of the new kernel $\mathcal{K}'_{1,2}(\psi_1,\psi_2|\psi'_2;\psi'_3)$ is $p(\psi_1,\psi_2|\psi'_3)$ since

$$
\int\mathcal{K}'_{1,2}(\psi_1,\psi_2|\psi'_2;\psi'_3)p(\psi'_1,\psi'_2|\psi'_3)\mathrm{d}\psi'_1\mathrm{d}\psi'_2=p(\psi_1,\psi_2|\psi'_3). \tag{3.3}
$$

Then suppose Conditions LW-1 and LW-2 described in Section 1.2.1 of Chapter 1 hold. (For all of the numerical examples related with MDA in this dissertation, the validity of these two conditions can be easily verified.) We specify the Haar piror to α and use Haar PX-DA to update (ψ_1,ψ_2) conditioning on ψ'_3, see the fourth panel of Figure 3.2. The Haar PX-DA algorithm replaces the component kernel $\mathcal{K}_1[\psi_1|;(\psi'_2,\psi'_3)]=p(\psi_1|\psi'_2,\psi'_3)$ in (3.1) with the new kernel,

$$
\begin{aligned}
\mathcal{K}'_1[\psi_1|;(\psi'_2,\psi'_3)]\;&=\int\int p(\psi_1^\star|\psi'_2,\psi'_3)\tilde{p}(\alpha^\star,\psi_1|\psi_1^\star,\psi'_3)\mathrm{d}\alpha^\star\mathrm{d}\psi_1^\star\\
&=\int p(\mathcal{G}_{\alpha^\star}(\psi_1)|\psi'_2,\psi'_3)\tilde{p}(\alpha^\star|\mathcal{G}_{\alpha^\star}(\psi_1),\psi'_3)|\mathrm{J}(\psi_1|\alpha^\star)|\mathrm{d}\alpha^\star.
\end{aligned}
\tag{3.4}
$$

Like the equivalence of updating (α,ψ_1,ψ_2) from $\tilde{p}(\alpha,\psi_1,\psi_2|\tilde{\psi}_1,\psi'_3)$to Step 2 of the MDA sampler, updating (α,ψ_1) from $\tilde{p}(\alpha,\psi_1|\psi_1^\star,\psi'_3)$, shown in the first row of (3.4), is equivalent to Step 2 of the Haar PX-DA sampler in Figure 3.2. As stated in Section 1.2.1, Liu and Wu (1999) verified that Step 2 of the standard Haar PX-DA sampler, i.e., Sampler 1.2l, is proper. The proof can be applied directly to show the propriety of Step 2 of the conditional Haar PX-DA sampler in Figure 3.2. Thus, $\mathcal{K}'_1[\psi_1|;(\psi'_2,\psi'_3)]$

in (3.4) maintains the stationary distribution of $\mathcal{K}_1(\psi_1|;\psi_2',\psi_3')$, that is, $p(\psi_1|\psi_2',\psi_3')$.

Transition kernel of ASIS

Suppose conditioning on ψ_3', ψ_1 is a sufficient augmentation for ψ_2, and let $\tilde{\psi}_1 = \mathcal{H}_{\psi_2}(\psi_1)$ be the corresponding ancillary augmentation. Then we replace Steps 1 and 2 of the parent Gibbs sampler with ASIS updates (last panel in Figure 3.2), which is equivalent to replacing first two component kernels in (3.1), $\mathcal{K}_1[\psi_1|;(\psi_2',\psi_3')] = p(\psi_1|\psi_2',\psi_3')$ and $\mathcal{K}_2[\psi_2|;(\psi_1,\psi_3')] = p(\psi_2|\psi_1,\psi_3')$, with the following new kernel,

$$\begin{aligned}
&\mathcal{K}_{1,2}'(\psi_1,\psi_2|\psi_2';\psi_3') \\
&= \int\int\int p(\psi_1^\star|\psi_2',\psi_3')p(\psi_2^\star,\tilde{\psi}_1|\psi_1^\star,\psi_3')p(\psi_1,\psi_2|\tilde{\psi}_1,\psi_3')\mathrm{d}\psi_1^\star\mathrm{d}\psi_2^\star\mathrm{d}\tilde{\psi}_1 \qquad (3.5)\\
&= \int\int p(\psi_1^\star|\psi_2',\psi_3')p(\psi_2^\star,\mathcal{H}_{\psi_2}(\psi_1)|\psi_1^\star,\psi_3')p(\psi_2|\mathcal{H}_{\psi_2}(\psi_1),\psi_3')|\mathrm{J}(\psi_1|\psi_2)\mathrm{d}\psi_1^\star\mathrm{d}\psi_2^\star,
\end{aligned}$$

where $\mathrm{J}(\psi_1|\psi_2)$ is the Jacobian matrix of the transformation $\psi_1 = \mathcal{H}_{\psi_2}^{-1}(\tilde{\psi}_1)$ conditioning on ψ_2. Sampling (ψ_1,ψ_2) from $p(\psi_1,\psi_2|\tilde{\psi}_1,\psi_3')$ shown in the second row of (3.5) is equivalent to Step 3 of the ASIS sampler in Figure 3.2 because updating ψ_1 from $p(\psi_1|\tilde{\psi}_1,\psi_2,\psi_3')$ is simply a transformation from $\tilde{\psi}_1$ to ψ_1. The stationary distribution of the ASIS kernel $\mathcal{K}_{1,2}'(\psi_1,\psi_2|\psi_2';\psi_3')$ is $p(\psi_1,\psi_2|\psi_3')$ since it is easy to verify that

$$\int \mathcal{K}_{1,2}'(\psi_1,\psi_2|\psi_2';\psi_3')p(\psi_1',\psi_2'|\psi_3')\mathrm{d}\psi_1'\mathrm{d}\psi_2' = p(\psi_1,\psi_2|\psi_3'). \qquad (3.6)$$

Note that when one step of a sampler requires MH, the corresponding component kernel also depends on the current iteration of the parameter to be updated. For example, if Step 1 of the Gibbs sampler in Figure 3.2 requires MH update, its transition kernel $\mathcal{K}_1$ has the form $\mathcal{K}_1[\psi_1|\psi_1';(\psi_2',\psi_3')]$, see Section 1.1.2 of Chapter 1.

3.2.2 Identifying stationary distributions of combined samplers

To construct a sampler combining two or more acceleration strategies, we replace some component transition kernels of the parent Gibbs sampler with the kernels for the strategies we tend to use. As shown in Section 3.2.1, new component kernels may not have the target stationary distribution. Thus we must take care to guarantee that the stationary distribution of the overall combined sampler is the target.

Generally, we can directly verify the stationary distribution of a sampler by finding $\pi(\psi)$ such that

$$\pi(\psi) = \int \mathcal{K}(\psi|\psi')\pi(\psi')\mathrm{d}\psi. \tag{3.7}$$

In addition, we provide two conditions which are sufficient to ensure that the combined sampler maintains the target stationary distribution, that is,

i) if $\psi' \sim p(\psi)$, then the input to each component kernel of the sampler follows the target distribution. For example, suppose the current iteration of the parameter ψ' is a draw from the target distribution $p(\psi)$. If the j^{th} component kernel is $\mathcal{K}'_j(\psi_j|;\mathcal{F}'_j(\psi,\psi'))$, then $\mathcal{F}'_j(\psi,\psi')$ must follow the target distribution $p(\mathcal{F}'_j(\psi,\psi'))$. If $\mathcal{K}'_j$ has the form $\mathcal{K}'_j(\psi_j|\psi'_j;\mathcal{F}'_j(\psi,\psi'))$, the joint distribution of ψ'_j and $\mathcal{F}'_j(\psi,\psi')$ should be the target;

ii) the last step of the sampler is a draw from the complete conditional of the target joint distribution.

These two conditions ensure that the last component kernel produces a draw from $p(\psi)$ if the input to the first component kernel follows the target distribution. Using Conditions i) and ii) to verify the stationary distribution is simpler than directly applying (3.7).

Sometimes we need to permute the component kernels to guarantee that the two conditions hold. For instance, after replacing the component kernel $\mathcal{K}_1[\psi_1|;(\psi'_2,\psi'_3)] = p(\psi_1|\psi'_2,\psi'_3)$ of the Gibbs sampler in Figure 3.2 with the PCG kernel, $\mathcal{K}'_1(\psi_1|;\psi'_2) = p(\psi_1|\psi'_2)$, Condition i) above does not hold, because sampling ψ_1 from $p(\psi_1|\psi'_2)$ leads to the conditional independence of ψ'_3 and ψ_1 in Step 2 and the input to Step 2 does not follow the target distribution. Thus we change the order of Steps 2 and 3, that is, sampling ψ_3 immediately after ψ_1, and obtain the PCG sampler in Figure 3.2. Now both Conditions i) and ii) hold and the PCG sampler maintains the target stationary distribution.

3.2.3 Convergence rate of combined samplers

In this section, we establish the computational advantage of the samplers combining acceleration strategies. The acceleration strategies we consider are PCG, MDA (Haar PX-DA), and ASIS. Thus to prove the combined sampler outperforms the parent Gibbs sampler in efficiency, we simply need to show that replacing a component kernel of the parent Gibbs sampler with one of the PCG, MDA (Haar PX-DA), and ASIS transition kernels can only improve the rate of convergence. Without loss of generality, we just verify that the PCG, MDA, Haar PX-DA, and ASIS samplers in Figure 3.2 all have better convergence properties than the parent Gibbs sampler in Figure 3.2. We use the cyclic-permutation bound introduced in Section 1.3 of Chapter 1 to compare computational efficiency of different samplers, since it is easier to handle than the spectral radius of a sampler. Recall that smaller cyclic-permutation bound indicates faster convergence.

Sampling more components of ψ in any set of steps of a Gibbs sampler can only reduce the cyclic-permutation bound, see Theorem 1 of van Dyk and Park (2008). Thus the PCG sampler in Figure 3.2 reduces the cyclic-permutation bound of its parent Gibbs sampler.

To compare the relative efficiencies of the Gibbs sampler and the MDA sampler with a proper working prior in Figure 3.2, we consider another sampler, which proceeds by

1. $\tilde{p}(\alpha, \tilde{\psi}_1 | \psi_2', \psi_3')$
2. $\tilde{p}(\psi_2 | \alpha, \tilde{\psi}_1, \psi_3')$; set $\psi_1 = \mathcal{G}_\alpha^{-1}(\tilde{\psi}_1)$
3. $p(\psi_3 | \psi_1, \psi_2)$.

We name this sampler by Gibbs Sampler 2. The stationary distribution of Gibbs Sampler-2 is $p(\alpha)p(\psi_1, \psi_2, \psi_3)$. More interestingly, the cyclic-permutation bound of Gibbs Sampler 2 equals to that of the Gibbs sampler in Figure 3.2. We present the corresponding proof in Appendix B. The MDA sampler in Figure 3.2 differs from Gibbs Sampler 2 only in that the MDA sampler updates α in Step 2, whereas Gibbs Sampler 2 conditions on it. Then the MDA sampler has smaller cyclic-permutation bound than Gibbs Sampler 2, because sampling more components in one step of a Gibbs-type sampler can only reduce the cyclic-permutation bound. Thus the MDA sampler reduces the cyclic-permutation bound of its parent Gibbs sampler, which owns the same cyclic-permutation bound as Gibbs Sampler 2.

Under Conditions LW-1 and LW-2, Liu and Wu (1999) verified the optimality of a standard Haar PX-DA sampler, i.e., Sampler 1.2l in Section 1.2.1 of Chapter 1, among a class of such algorithms. Specifically, with these two condition, each $Y_{\text{mis}} \in \mathcal{Z}$ can be represented by its orbit $r \in \mathcal{Q}$ and its position $\beta \in \mathcal{A}$ on the orbit. Thus we can parameterize the original augmented model as

$$p(Y_{\text{mis}}, \theta | Y_{\text{obs}}) \mathrm{d}Y_{\text{mis}} \mathrm{d}\theta = p(\beta, r, \theta | Y_{\text{obs}}) \mathrm{d}r \mathrm{d}\theta H(\mathrm{d}\beta), \tag{3.8}$$

and the expanded augmented model as

$$\tilde{p}(\alpha, Y_{\text{mis}}, \theta | Y_{\text{obs}}) \mathrm{d}Y_{\text{mis}} \mathrm{d}\theta H(\mathrm{d}\alpha) = p(\alpha \cdot \beta, r, \theta | Y_{\text{obs}}) p(\alpha) \mathrm{d}r \mathrm{d}\theta H(\mathrm{d}\alpha) H(\mathrm{d}\beta). \tag{3.9}$$

Liu and Wu (1999) showed that the DA sampler is equivalent to iterating between $\beta, r|\theta$ and $\theta|\beta, r$ based on (3.8), the MDA sampler with the working prior $p(\alpha)$, i.e., Sampler 1.1 in Section 1.2.1, also induces the iteration between $\beta, r|\theta$ and $\theta|\beta, r$, but based on (3.9), whereas the Haar PX-DA sampler, i.e., Sampler 1.2l in Section 1.2.1, iterates between $r|\theta$ and $\theta|r$, based on (3.9) with $p(\alpha) = 1$. Henceforth, updating θ without conditioning on β, the Haar PX-DA sampler outperforms the parent DA sampler and any MDA sampler with a proper working prior under Conditions LW-1 and LW-2. The arguments above can be directly applied to prove that the Haar PX-DA sampler in Figure 3.2 has smaller cyclic-permutation bound than both its parent Gibbs sampler and the MDA sampler in Figure 3.2.

Yu and Meng (2011) proved that under Conditions YM-1 and YM-2 below, the ASIS sampler, i.e., Sampler 1.3 in Section 1.2.2 of Chapter 1 is identical to the optimal Haar

PX-DA sampler for the expanded model $\tilde{p}(\alpha, \tilde{Y}_{\text{mis,S}}, \theta|Y_{\text{obs}})$, where α is introduced into the model via the transformation $\tilde{Y}_{\text{mis,S}} = \mathcal{H}_\alpha(Y_{\text{mis,S}})$ (recall that $Y_{\text{mis,A}} = \mathcal{H}_\theta(Y_{\text{mis,S}})$), see Theorem 4 of Yu and Meng (2011).

YM-1: The state space of θ, Θ, forms a group (induced by $\mathcal{H}_\theta$) with a unimodular Haar measure;

YM-2: The prior distribution for θ, $p_\infty(\theta)$, with respect to the Haar measure satisfies that $p_\infty(\theta \cdot \theta') \propto p_\infty(\theta)p_\infty(\theta')$.

Considering the ASIS sampler in Figure 3.2, we replace θ with ψ_2, and $Y_{\text{mis,S}}$ with ψ_1. Then suppose that Conditions YM-1 and YM-2 hold. We can directly use the proof for Theorem 4 of Yu and Meng (2011) to show that the ASIS update induced by Steps 1 and 2 of the sampler in the last panel Figure 3.2 is identical to a Haar PX-DA algorithm conditioning on ψ_3'. Thus we conclude that under Conditions YM-1 and YM-2, the ASIS sampler reduces the cyclic-permutation bound of its parent Gibbs sampler. Although Conditions YM-1 and YM-2 seem relatively restrictive, they are satisfied in all of the numerical examples related with ASIS in this Chapter.

In general, replacing a component kernel of the parent Gibbs sampler with one of the PCG, MDA (Haar PX-DA), and ASIS transition kernels can only reduce the cyclic-permutation bound. Thus the samplers combining two or more of these acceleration strategies improve the convergence rate of their parent Gibbs samplers.

3.3 Illustration Examples

3.3.1 Factor analysis model

In Section 3.1, we use the factor analysis model described in Section 2.3.3 of Chapter 2 to motivate combining several acceleration strategies into one sampler. In this section, we use the same simulation study as in Section 2.3.3 to show numerically the efficiency of the combined sampler relative to the samplers that use only one strategy. Specifically, we start with the parent Gibbs sampler, i.e., Sampler 2.9, design an MH within PCG

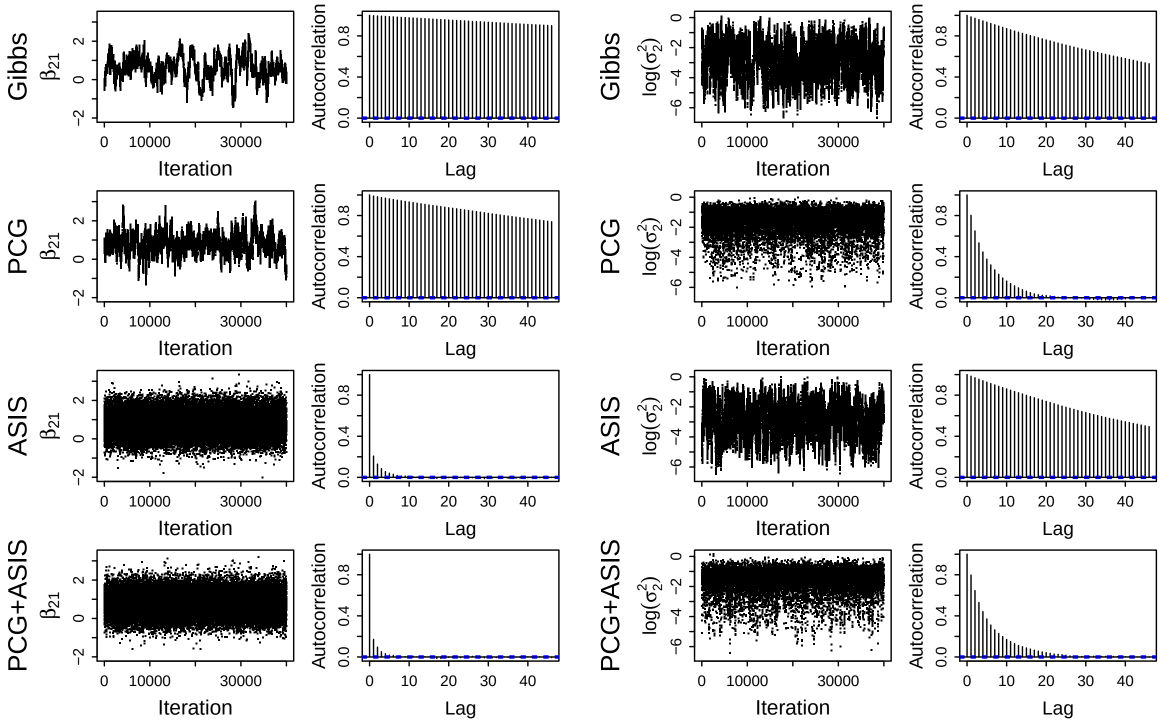

Figure 3.3: Comparing four samplers for the factor analysis model. The left two columns are the time-series and autocorrelation plots for the posterior draws of β_{21}, while the right two columns are those for $\log(\sigma_2^2)$. The four rows from top to bottom correspond to the Gibbs, PCG, ASIS, and combined samplers, respectively. The combined sampler outperforms other three in efficiency.

sampler (Sampler 2.10) and an ASIS sampler (Sampler 3.1), and demonstrate how they can be combined to a proper MH within PCG + ASIS sampler.

Recall that the PCG sampler updates σ_1^2–σ_4^2 without conditioning on Z, which requires the help of MH. Since we derive Sampler 2.10 from its parent Gibbs sampler by following the three-stage framework, see Figure 2.15, Sampler 2.10 maintains the target stationary distribution. It is easy to verify that the ASIS sampler also has the correct stationary distribution. Suppose (Z', Σ', β') is a sample from $p(Z, \Sigma, \beta|Y)$. Because the first three steps of Sampler 3.2 all update the complete conditionals of the target, they maintain the target distribution. After the transformation at the end of Step 3, we have $(W, \Sigma, \beta^\star) \sim p(W, \Sigma, \beta|Y)$, which is equivalent to $p(Z, \Sigma, \beta|Y)$. After updating β from $p(\beta|Y, W, \Sigma)$ in Step 4, (W, Σ, β) follows the distribution $p(W, \Sigma, \beta|Y)$. Finally, the transformation at the end of Step 4 makes (Z, Σ, β) a new sample from the target distribution. By

	Gibbs	**MH within PCG**	**ASIS**	**MH within PCG+ASIS**
$\boldsymbol{\beta_{21}}$	0.023	0.062	8.165	9.388
$\boldsymbol{\log(\sigma_2^2)}$	0.142	1.777	0.137	1.502

Table 3.1: The ESS per second of β_{21} and σ_2^2 for Samplers 2.9, 2.10, 3.1, and 3.2. The combined sampler, i.e., Sampler 3.2, outperforms the other three.

combining the arguments for the MH within PCG and ASIS samplers, we ensure the MH within PCG + ASIS sampler also has the target stationary distribution.

We run 50,000 iterations with a burnin of 10,000 and use the same starting values for each sampler. Figure 3.3 compares the four samplers in terms of the mixing and autocorrelation of β_{21} and $\log(\sigma_2^2)$; the left two columns correspond to results for β_{21}, and the right two columns for $\log(\sigma_2^2)$. The other β components behave similarly to β_{21}, and $\log(\sigma_1^2)$, $\log(\sigma_3^2)$, and $\log(\sigma_4^2)$ behave similarly to $\log(\sigma_2^2)$, while $\log(\sigma_5^2)$ and $\log(\sigma_6^2)$ converge well for all four samplers. We find the MH within PCG sampler is efficient in improving the convergence of Σ, but has little effect on β. ASIS has the opposite effect. By combining MH within PCG and ASIS, we improve the convergence of both Σ and β simultaneously.

Because it has more steps, the combined sampler is computationally more expensive than the other three. To check whether its improved efficiency compensates for the additional computational cost, we estimate the ESS per second. We present the ESS per second of β_{21} and $\log(\sigma_2^2)$ for the four samplers in Table 3.1. By this measure, we find the sampler combining MH within ASIS and PCG substantially outperforms the other three in efficiency with a fairly small extra computational cost. (Although the MH within PCG sampler is slightly better than the combined sampler for $\log(\sigma_2^2)$, it takes around 150 times longer to obtain the same ESS for β_{21}.)

3.3.2 Cosmological hierarchical model

Recall the cosmological hierarchical model described in Section 2.3.2 of Chapter 2. In Section 2.3.2, to sample from the posterior distribution of this model, we start with

```
ASIS (Sampler 3.3)
1. $p(\xi, X^\star|Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$
2. $\mathcal{M}_{\mathscr{C}|Y,\xi,X,\alpha,\beta,\Sigma_P}(\mathscr{C}^\star|Y, \xi, X^\star, \mathscr{C}', \alpha', \beta', \Sigma'_P)$;
   use $\mathscr{C}^\star$ to construct $L^\star$
3. $p(\alpha^\star, \beta^\star|Y, \xi, X^\star, \mathscr{C}^\star, \Sigma'_P)$;
   use $(\alpha^\star, \beta^\star)$ to construct $A^\star$;
   set $\bar{X} = A^\star X^\star + L^\star$
4. $\mathcal{M}_{\mathscr{C}|Y,\xi,\bar{X},\alpha,\beta,\Sigma_P}(\mathscr{C}|Y, \xi, \bar{X}, \mathscr{C}^\star, \alpha^\star, \beta^\star, \Sigma'_P)$;
   use $\mathscr{C}$ to construct $L$
5. $p(\alpha, \beta|Y, \xi, \bar{X}, \mathscr{C}, \Sigma'_P)$;
   use $(\alpha, \beta)$ to construct $A$;
   set $X = A^{-1}(\bar{X} - L)$
6. $p(\Sigma_P|Y, \xi, X, \mathscr{C}, \alpha, \beta)$
```

```
MH within PCG+ASIS (Sampler 3.4)
1. $\mathcal{M}_{\mathscr{C}|Y,\alpha,\beta,\Sigma_P}(\mathscr{C}|Y, \mathscr{C}', \alpha', \beta', \Sigma'_P)$;
   use $\mathscr{C}$ to construct $L$
2. $p(\xi, X^\star|Y, \mathscr{C}, \alpha', \beta', \Sigma'_P)$
3. $p(\alpha^\star, \beta^\star|Y, \xi, X^\star, \mathscr{C}, \Sigma'_P)$;
   use $(\alpha^\star, \beta^\star)$ to construct $A^\star$;
   set $\bar{X} = A^\star X^\star + L$
4. $p(\alpha, \beta|Y, \xi, \bar{X}, \mathscr{C}, \Sigma'_P)$;
   use $(\alpha, \beta)$ to construct $A$;
   set $X = A^{-1}(\bar{X} - L)$
5. $p(\Sigma_P|Y, \xi, X, \mathscr{C}, \alpha, \beta)$
```

Figure 3.4: Samplers 3.3 and 3.4. The left and right panels show the steps of the ASIS sampler (Sampler 3.3) and the sampler combining MH within PCG and ASIS (Sampler 3.4) for fitting the cosmological hierarchical model.

the parent MH within Gibbs sampler, i.e., Sampler 2.7 in Figure 2.11, where both $\mathscr{C} = (\Omega_m, \Omega_\Lambda)$ and (α, β) exhibit poor convergence. In order to improve the convergence of these two parameters, we use three other samplers. First we consider the MH within PCG sampler, Sampler 2.8, introduced in Section 2.3.2, which uses MH to update $\mathscr{C}$ and (α, β) without conditioning on ξ or X. We derive Sampler 2.8 from Sampler 2.7 using the three-stage framework described in Section 2.2.2 of Chapter 2, see Figure 2.12. This guarantees that the MH within PCG sampler is proper. Next we construct an ASIS sampler. We derive the sufficient and ancillary augmentations for $\mathscr{C}$ and (α, β) conditioning on the other parameters ξ and Σ_P. The distribution of X conditioning on $\mathscr{C}$ and (α, β) is

$$X|\mathscr{C}, \alpha, \beta \sim \mathrm{N}(J\xi, \Sigma_P), \tag{3.10}$$

where $J_{(3n\times 3)} = (I, \ldots, I)^T$ with $I = \mathrm{Diag}(1, 1, 1)$. Because this distribution is free of $\mathscr{C}$ and (α, β), X is an ancillary augmentation for both of them. To derive a sufficient augmentation, we set $\bar{X} = AX + L$, where $A_{(3n\times 3n)} = \mathrm{Diag}(T, \ldots, T)$ with $T_{(3\times 3)} =$

$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta & -\alpha & 1 \end{bmatrix}$, and let L denote the $(3n \times 1)$ vector $(0, 0, \mu_1, \ldots, 0, 0, \mu_n)$, which is a deterministic function of $\mathscr{C}$. The distribution of observed quantities Y conditioning on $\bar{X}$, $\mathscr{C}$ and (α, β) is

$$Y|\bar{X}, \mathscr{C}, \alpha, \beta \sim \mathrm{N}(\tilde{X}, \Sigma_C), \tag{3.11}$$

where $\Sigma_{C(3n \times 3n)} = \mathrm{Diag}(\hat{C}_1, \ldots, \hat{C}_n)$. Because this distribution is free of $\mathscr{C}$ and (α, β), $\bar{X}$ is a sufficient augmentation for both parameters. Sampler 3.3 in the left panel of Figure 3.4 is the ASIS sampler corresponding to this pair of sufficient and ancillary augmentations. Specifically, in Sampler 3.3, we implement ASIS conditioning on ξ and Σ_P: (i) (ξ, X) is sampled from its complete conditional distribution; (ii) $\mathscr{C}$ and (α, β) are updated conditioning on X, and X is transformed to $\bar{X}$ conditioning on $\mathscr{C}$ and (α, β); and (iii) $\mathscr{C}$ and (α, β) are updated again but conditioning on $\bar{X}$. Both of the updates of $\mathscr{C}$ require MH. Sampler 3.4 combines MH within PCG and ASIS. In particular, conditioning on (α, β), we update $\mathscr{C}$ with MH within PCG, and then conditioning on $\mathscr{C}$, we update (α, β) with ASIS, see the right panel of Figure 3.4 for its steps.

It is easy to verify that Sampler 3.4 is proper. Suppose $(\xi', X', \mathscr{C}', \alpha', \beta', \Sigma'_P)$ is a draw from the target distribution. Although in Step 1 of Sampler 3.4, we sample $\mathscr{C}$ without conditioning on (ξ, X), the marginal distribution of $(\mathscr{C}, \alpha', \beta', \Sigma'_P)$ is still that of the target. After updating (ξ, X) from its complete conditional distribution in Step 2, $(\xi, X, \mathscr{C}, \alpha', \beta', \mathscr{C}')$ follows the target distribution. Since the distribution of (X, α, β) is equivalent to that of $(\bar{X}, \alpha, \beta)$ conditioning on the other parameters, when we transform back to (X, α, β) at the end of Step 4, $(\xi, X, \mathscr{C}, \alpha, \beta, \Sigma'_P)$ follows the target distribution. Finally, the step that updates Σ_P is a standard Gibbs step which preserves the target stationary distribution. Thus, Sampler 3.4 is proper.

We run each of the MH within Gibbs, MH within PCG, ASIS, and MH within PCG + ASIS samplers for 11,000 iterations with a burn-in of 1,000 using the same data set and initial values as in Section 2.3.2.

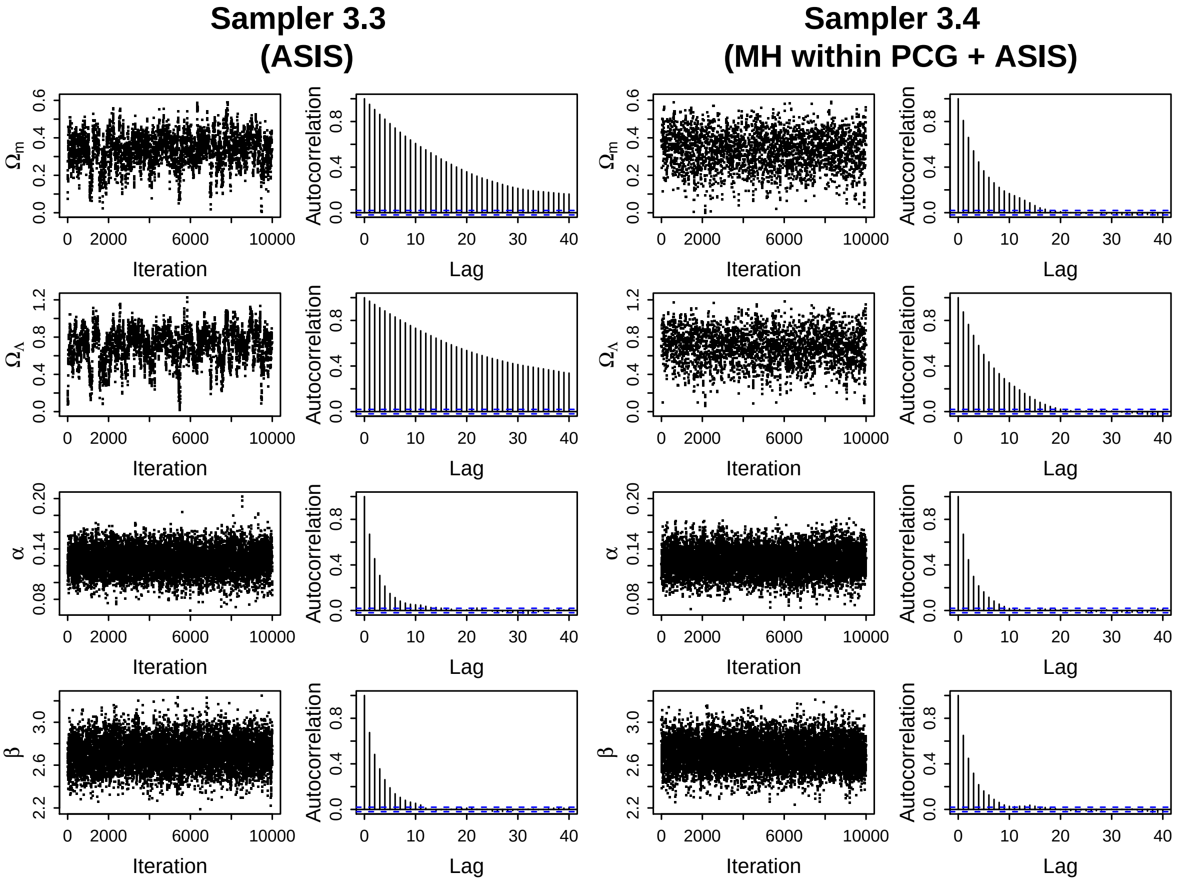

Figure 3.5: The sampling results of Samplers 3.3 and 3.4. The left two columns are the time-series and autocorrelation plots for the posterior draws of Ω_m, Ω_Λ, α, and β respectively from Sampler 3.3, while the right two columns are those from Sampler 3.4. The combined sampler, Sampler 3.4, converges better than the other three.

The sampling results of the MH within Gibbs and MH within PCG samplers are shown in Figure 2.13 of Section 2.3.2. Figure 3.5 shows the convergence properties of the ASIS and combined samplers. For each sampler, we display the time-series (left column) and autocorrelation plots (right column) for Ω_m, Ω_Λ, α, and β. For all of the four parameters, the MH within PCG, ASIS, and MH within PCG + ASIS samplers produce chains with much faster mixing and lower autocorrelation than the parent MH within Gibbs sampler. We display the ESS per second of Ω_m, Ω_Λ, α, and β in Table 3.2 and conclude that the MH within PCG, ASIS, and MH within PCG + ASIS samplers all substantially improve the convergence properties of the MH within Gibbs sampler. Thus we confirm

	MH within Gibbs	**MH within PCG**	**ASIS**	**MH within PCG + ASIS**
$\boldsymbol{\Omega_m}$	0.002	0.037	0.010	0.041
$\boldsymbol{\Omega_\Lambda}$	0.001	0.023	0.006	0.027
$\boldsymbol{\alpha}$	0.008	0.051	0.076	0.073
$\boldsymbol{\beta}$	0.010	0.029	0.064	0.074

Table 3.2: The ESS per second of Ω_m, Ω_Λ, α, and β for Samplers 2.7, 2.8, 3.3, and 3.4. The combined sampler, i.e., Sampler 3.4, outperforms the other three samplers.

that both PCG and ASIS are efficient in improving convergence. More interestingly, ASIS is less efficient in improving the convergence of $\mathscr{C}$ than MH within PCG, while better in improving the convergence of (α, β). When we combine these two strategies into Sampler 3.4, the result outperforms both the MH within PCG and ASIS samplers in terms of ESS per second. (Although the ASIS sampler is slightly better than the combined sampler for α, it takes around 15% longer to obtain the same ESS for β, and more than four times longer for $\mathscr{C}$.)

3.3.3 Hierarchical t model

In this section, we use a hierarchical t model to illustrate the further efficiency obtained by combining Haar PX-DA and ASIS algorithms into one sampler. Specifically, the observations, Y_i $(i = 1, \ldots, n)$, follow Gaussian distributions independently conditioning on all the other parameters, that is,

$$Y_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\beta_i X_i, \frac{\sigma^2}{Z_i}\right), \text{ for } i = 1, \ldots, n, \tag{3.12}$$

where $Y = (Y_1, \ldots, Y_n)$ are the observations and $X = (X_1, \ldots, X_n)$ are known covariates. We specify Gaussian and chi-square distributions to the regression coefficients, $\beta = (\beta_1, \ldots, \beta_n)$, and variance parameters, $Z = (Z_1, \ldots, Z_n)$, respectively, that is,

$$\beta_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^2) \text{ and } Z_i \overset{\text{iid}}{\sim} \chi^2_\nu/\nu. \tag{3.13}$$

Gibbs (Sampler 3.5)

1. $p(Z|Y,(\sigma^2)',\beta',\tau',\mu')$
2. $p(\sigma^2|Y,Z,\beta',\tau',\mu')$
3. $p(\beta|Y,Z,\sigma^2,\tau',\mu')$
4. $p(\tau,\mu|Y,Z,\sigma^2,\beta)$

Haar PX-DA (Sampler 3.6)

1. $p(Z^\star|Y,(\sigma^2)',\beta',\tau',\mu')$
2. $\tilde{p}(\alpha|Y,Z^\star,\beta',\tau',\mu')$; set $Z=\tilde{Z}/\alpha$
3. $p(\sigma^2|Y,Z,\beta',\tau',\mu')$
4. $p(\beta|Y,Z,\sigma^2,\tau',\mu')$
5. $p(\tau,\mu|Y,Z,\sigma^2,\beta)$

ASIS (Sampler 3.7)

1. $p(Z|Y,(\sigma^2)',\beta',\tau',\mu')$
2. $p(\sigma^2|Y,Z,\beta',\tau',\mu')$
3. $p(\beta^\star|Y,Z,\sigma^2,\tau',\mu')$
4. $p(\tau^\star,\mu^\star|Y,Z,\sigma^2,\beta^\star)$; set $\bar{\beta}=(\beta^\star-\mu^\star)/\tau^\star$
5. $p(\tau,\mu|Y,Z,\sigma^2,\bar{\beta})$; set $\beta=\tau\bar{\beta}+\mu$

Haar PX-DA+ASIS (Sampler 3.8)

1. $p(Z^\star|Y,(\sigma^2)',\beta',\tau',\mu')$
2. $\tilde{p}(\alpha|Y,\tilde{Z},\beta',\tau',\mu')$; set $Z=\tilde{Z}/\alpha$
3. $p(\sigma^2|Y,Z,\beta',\tau',\mu')$
4. $p(\beta^\star|Y,Z,\sigma^2,\tau',\mu')$
5. $p(\tau^\star,\mu^\star|Y,Z,\sigma^2,\beta^\star)$; set $\bar{\beta}=(\beta^\star-\mu^\star)/\tau^\star$
6. $p(\tau,\mu|Y,Z,\sigma^2,\bar{\beta})$; set $\beta=\tau\bar{\beta}+\mu$

Figure 3.6: Four samplers for fitting the hierarchical t model. The top-left and top-right panels show the steps of the parent Gibbs sampler (Sampler 3.5) and the Haar PX-DA sampler (Sampler 3.6). The bottom-left and bottom-right panels display the steps of the ASIS sampler (Sampler 3.7) and the sampler combining Haar PX-DA and ASIS (Sampler 3.8).

Combining (3.12) and (3.13), Y_i is marginally t-distributed. We specify non-informative prior distributions to unknown parameters (τ,μ,σ) as $p(\tau,\mu,\sigma)\propto 1$. Here we set Y_i, β_i, and Z_i as univariate variables, but see van Dyk (2000) for a multivariate version of the hierarchical t model. We wish to sample from the posterior distribution, $p(Z,\sigma^2,\beta,\tau,\mu|Y)$.

To update $p(Z,\sigma^2,\beta,\tau,\mu|Y)$, we start with the standard Gibbs sampler, which updates Z, σ^2, β, and (τ,μ) iteratively from their complete conditional distributions, see Sampler 3.5 in the top-left panel of Figure 3.6. With the Gibbs sampler, both (τ,μ) and σ^2 exhibit poor convergence. Thus we consider another three samplers to improve the efficiency of (τ,μ) and σ^2. First, we construct an MDA sampler. Specifically, we introduce the working parameter α into the model via $\tilde{Z}=(\tilde{Z}_1,\ldots,\tilde{Z}_n)$, that is,

$$Y_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\beta_i X_i, \frac{\alpha\sigma^2}{\tilde{Z}_i}\right), \tag{3.14}$$

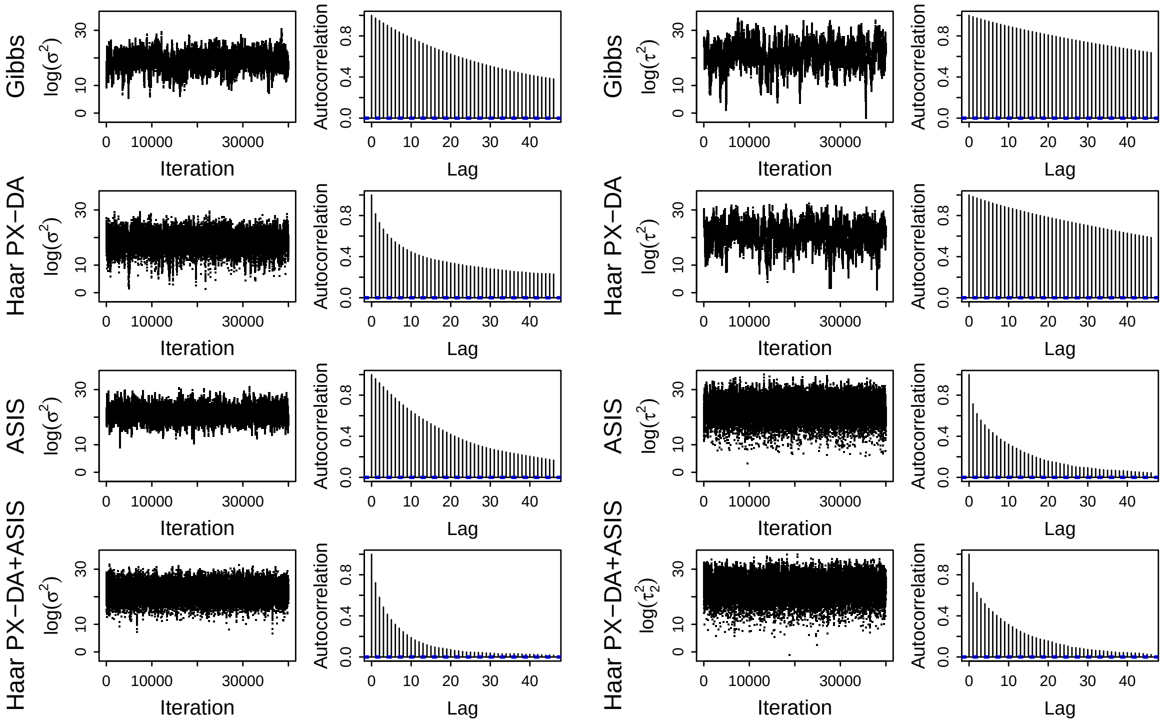

Figure 3.7: Comparing four samplers for the hierarchical t model. The left two columns are the mixing and autocorrelation plots for the posterior draws of $\log(\sigma^2)$, while the right two columns are those for $\log(\tau^2)$. The four rows from top to bottom correspond to the Gibbs, Haar PX-DA, ASIS, and combined samplers, respectively. The combined sampler outperforms other three in efficiency.

where $\tilde{Z} = \alpha Z$ and thus $\tilde{Z}_i | \alpha \overset{\text{iid}}{\sim} \alpha\chi^2_\nu/\nu$. In this example, Conditions LW-1 and LW-2 both hold. Thus we specify the Haar measure prior to α, that is, $p_\infty(\alpha) \propto 1/\alpha$, and use Haar PX-DA to obtain optimal computational efficiency. Although it is difficult to derive the Haar PX-DA algorithm for both (τ, μ) and σ^2, constructing a Haar PX-DA sampler for σ^2 conditioning on (τ, μ) is simple. Thus we implement Haar PX-DA conditioning on (τ, μ) and obtain a conditional Haar PX-DA sampler, i.e., Sampler 3.6. Specifically, Sampler 3.6 proceeds by i) updating σ^2 and Z with the Haar PX-DA algorithm conditioning on β and (τ, μ), and ii) sampling β and (τ, μ) from their complete conditional distributions; see the top-right panel of Figure 3.6 for steps of Sampler 3.6. Next we use ASIS to improve the convergence. We derive the sufficient and ancillary augmentations for (τ, μ) conditioning on Z and σ^2. Because the distribution of observed

	Gibbs	**Haar PX-DA**	**ASIS**	**Haar PX-DA+ASIS**
$\log(\boldsymbol{\sigma^2})$	46.028	64.875	66.972	265.627
$\log(\boldsymbol{\tau^2})$	19.978	21.865	162.177	167.040

Table 3.3: The ESS per second of $\log(\sigma^2)$ and $\log(\tau^2)$ for Samplers 3.5–3.8. The combined sampler, i.e., Sampler 3.8, outperforms the other three samplers.

quantities Y conditioning on Z, β, and σ^2 is

$$Y_i|Z, \beta_i, \sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(\beta_i X_i, \frac{\sigma^2}{Z_i}\right), \tag{3.15}$$

free of (τ, μ), β is the sufficient augmentation for (τ, μ). $\bar{\beta} = (\beta - \mu)/\tau$ is the corresponding ancillary augmentation because its distribution, $\bar{\beta}_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, is free of (τ, μ). We construct the ASIS sampler by i) sampling all the parameters from their complete conditionals, and ii) sampling (τ, μ) again conditioning on $\bar{\beta}$, see Sampler 3.7 in the bottom-left panel of Figure 3.6. Finally, we combine Haar PX-DA and ASIS into one coherent sampler by updating σ^2 with Haar PX-DA conditioning on (τ, μ) and sampling (τ, μ) with ASIS conditioning on σ^2, and obtain Sampler 3.8 in the bottom-right panel of Figure 3.6.

Now we verify that the combined sampler, i.e., Sampler 3.8, is proper. Suppose $(Z', (\sigma^2)', \beta', \tau', \mu')$ is a draw from the target $p(Z, \sigma^2, \beta, \tau, \mu|Y)$. With $\psi_1 = Z$, $\psi_2 = \sigma^2$, and $\psi_3 = (\beta, \tau, \mu)$, Sampler 3.8 as a specific example of the Haar PX-DA sampler in Figure 3.2. As shown in Section 3.2.1, the transition kernel induced by Steps 1 and 2 of Sampler 3.8 maintains the target stationary distribution. Thus after sampling Z in Steps 1 and 2, $(Z, (\sigma^2)', \beta', \tau', \mu')$ follows the target distribution. After updating σ^2 from its complete conditional distribution in Step 3, the distribution of $(Z, \sigma^2, \beta', \tau', \mu')$ is $p(Z, \sigma^2, \beta, \tau, \mu|Y)$. Since the distribution of (β, τ, μ) is equivalent to that of $(\bar{\beta}, \tau, \mu)$ conditioning on the other parameters, when we transform back to (β, τ, μ) at the end of Step 6, $(Z, \sigma^2, \beta, \tau, \mu)$ follows the target distribution. Thus, Sampler 3.8 is proper.

We use a simulation study to compare Samplers 3.5–3.8. We set $\mu = 2$, $\tau = 3$, $\sigma = 0.1$, $\nu = 0.1$, $n = 10$, and draw X from a uniform distribution on $[-1, 1]$. For each sampler, we use the same starting values and run a chain of 50,000 iterations with a burn-in of 10,000. Figure 3.7 compares the four samplers with regard to the mixing and autocorrelation of $\log(\sigma^2)$ and $\log(\tau^2)$, the left two columns correspond to results for $\log(\sigma^2)$, and the right two columns for $\log(\tau^2)$; μ behaves similarly to $\log(\tau^2)$. The Haar PX-DA sampler is effective in improving the convergence of σ^2, but has little effect on τ. The ASIS sampler is useful for improving the convergence of both parameters, especially for τ. By combining Haar PX-DA and ASIS, Sampler 3.8 significantly improves the convergence of both σ^2 and τ^2, and outperforms the samplers using either of the two methods alone. Furthermore, we account for the computational time and display the ESS per second of $\log(\sigma^2)$ and $\log(\tau^2)$ from all of the four samplers in Table 3.3. The combined sampler produces the largest ESS/sec for both parameters, which helps us confirm that combining Haar PX-DA and ASIS into one sampler is more efficient in improving convergence than using either strategy alone.

3.3.4 Hierarchical probit model

We use a hierarchical probit model to illustrate the efficiency we can gain by combining either Haar PX-DA and ASIS algorithms or two ASIS algorithms into one sampler. In this model, the observation $Y = (Y_1, \ldots, Y_n)$, are the indicators of the values of the latent variables $Z = (Z_1, \ldots, Z_n)$, that is,

$$Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i = 1, \ldots, n, \tag{3.16}$$

where $Z_i \overset{\text{ind}}{\sim} \mathrm{N}(X_i\beta_i, 1)$. Whereas $X = (X_1, \ldots, X_n)$ are known covariates, $\beta = (\beta_1, \ldots, \beta_n)$ are unknown parameters and $\beta_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^2)$. We specify conjugate priors to μ and τ as $\mu \sim \mathrm{N}(0, 10)$ and $p(\tau) \propto 1$. See van Dyk (2000) for a multivariate version of this model. We wish to sample from the posterior distribution $p(Z, \beta, \mu, \tau|Y)$.

We compare the performance of six samplers for fitting the hierarchical probit model. We start with the Gibbs sampler, which samples each of Z, (μ, β), and τ from its

Gibbs (Sampler 3.9)

1. $p(Z|Y, \beta', \mu', \tau')$
2. $p(\mu, \beta|Y, Z, \tau')$
3. $p(\tau|Y, Z, \beta, \mu)$

Haar PX-DA (Sampler 3.10)

1. $p(Z^\star|Y, \beta', \mu', \tau')$
2. $\tilde{p}(\alpha|Y, Z^\star, \tau')$; set $Z = Z^\star/\alpha$
3. $p(\mu, \beta|Y, Z, \tau')$
4. $p(\tau|Y, Z, \beta, \mu)$

ASIS I (Sampler 3.11)

1. $p(Z^\star|Y, \mu', \tau')$
2. $p(\mu^\star|Y, Z, \tau')$; set $\bar{Z} = Z^\star - X\mu^\star$
3. $p(\mu|Y, \bar{Z}, \tau')$; set $Z = \bar{Z} + X\mu$
4. $p(\beta|Y, Z, \mu, \tau')$
5. $p(\tau|Y, Z, \beta, \mu)$

ASIS II (Sampler 3.12)

1. $p(Z|Y, \beta', \mu', \tau')$
2. $p(\mu^\star, \beta^\star|Y, Z, \tau')$
3. $p(\tau^\star|Y, Z, \beta^\star, \mu^\star)$; set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$
4. $p(\mu|Y, Z, \bar{\beta}, \tau^\star)$
5. $p(\tau|Y, Z, \bar{\beta}, \mu)$; set $\beta = \tau\bar{\beta} + \mu$

Haar PX-DA+ASIS II (Sampler 3.13)

1. $p(Z^\star|Y, \beta', \mu', \tau')$
2. $\tilde{p}(\alpha|Y, Z^\star, \tau')$; set $Z = Z^\star/\alpha$
3. $p(\mu^\star, \beta^\star|Y, Z, \tau')$
4. $p(\tau^\star|Y, Z, \beta^\star, \mu^\star)$; set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$
5. $p(\mu|Y, Z, \bar{\beta}, \tau^\star)$
6. $p(\tau|Y, Z, \bar{\beta}, \mu)$; set $\beta = \tau\bar{\beta} + \mu$

ASIS I+ASIS II (Sampler 3.14)

1. $p(Z^\star|Y, \mu', \tau')$
2. $p(\mu^\star|Y, Z, \tau')$; set $\bar{Z} = Z^\star - X\mu^\star$
3. $p(\mu|Y, \bar{Z}, \tau')$; set $Z = \bar{Z} + X\mu$
4. $p(\beta^\star|Y, Z, \mu, \tau')$
5. $p(\tau^\star|Y, Z, \beta^\star, \mu)$; set $\bar{\beta} = (\beta^\star - \mu)/\tau^\star$
6. $p(\tau|Y, Z, \bar{\beta}, \mu)$; set $\beta = \tau\bar{\beta} + \mu$

Figure 3.8: Six samplers for fitting the hierarchical probit model. From top to bottom, the left panels show the steps of the parent Gibbs sampler, the sampler with ASIS I, and the sampler combining Haar PX-DA and ASIS II, whereas the right panels present the steps of the Haar PX-DA sampler, the sampler with ASIS II, and the sampler combining ASIS I and ASIS II.

complete conditional distribution, see Sampler 3.9 in the top-left panel of Figure 3.8. To improve the convergence of the Gibbs sampler, we consider using MDA and ASIS. First, we construct an MDA sampler. The working parameter α is introduced into the model by setting $\tilde{Z} = \alpha Z$. Then

$$Y_i = \begin{cases} 1 & \text{if } \tilde{Z}_i > 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i = 1, \ldots, n, \tag{3.17}$$

where $\tilde{Z}_i \overset{\text{ind}}{\sim} \mathrm{N}(X_i\alpha\beta_i, \alpha^2)$. To facilitate the update of α, we also set $\tilde{\beta} = \alpha\beta$ and

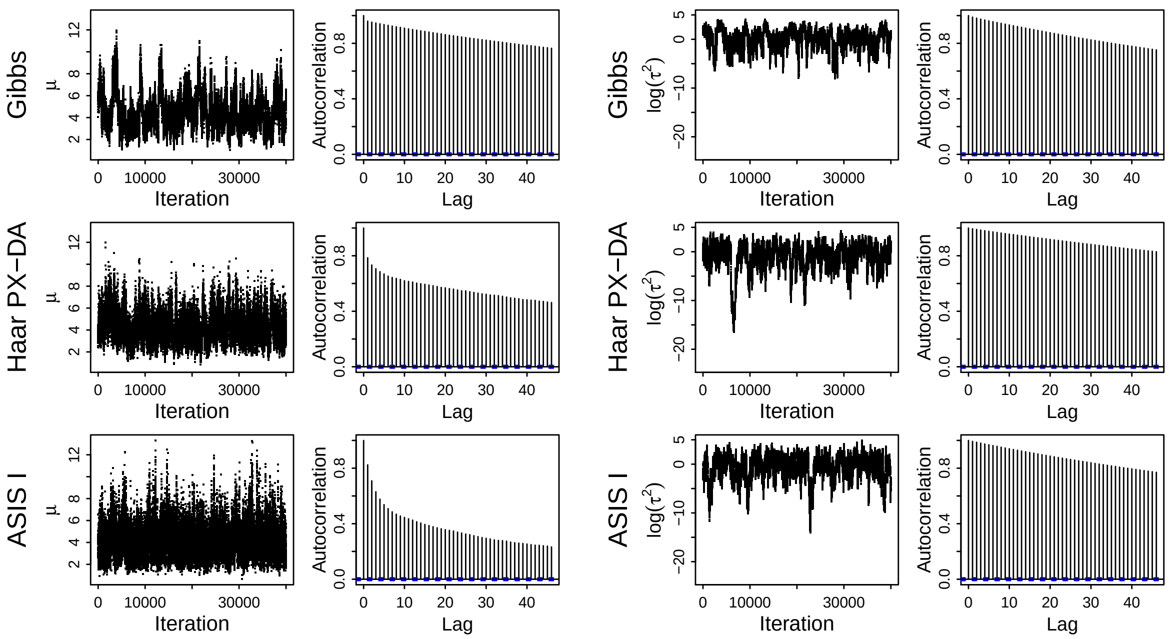

Figure 3.9: The sampling results of three samplers for the hierarchical probit model. The left two columns are the mixing and autocorrelation plots for the posterior draws of μ, while the right two columns are those for $\log(\tau^2)$. The three rows from top to bottom correspond to the Gibbs, Haar PX-DA, and ASIS I samplers. The Haar PX-DA and ASIS I samplers improve the convergence of μ, but have little effect on $\log(\tau^2)$.

$\tilde{\mu} = \alpha\mu$. Thus equivalently, $\tilde{Z}_i \overset{\text{ind}}{\sim} \mathrm{N}(X_i\tilde{\beta}_i, \alpha^2)$ and $\tilde{\beta}_i \overset{\text{ind}}{\sim} \mathrm{N}(\tilde{\mu}, \alpha^2\tau^2)$, and the prior of $\tilde{\mu}$ is $\tilde{\mu} \sim \mathrm{N}(0, \alpha^2 10)$. Since Conditions LW-1 and LW-2 hold for this example, we use the Haar PX-DA algorithm to achieve optimality. We specify the Haar measure prior to α, that is, $p_\infty(\alpha) \propto 1$, and obtain the Haar PX-DA sampler, i.e., Sampler 3.10, by i) updating Z and (μ, β) with the Haar PX-DA algorithm conditioning on τ, and ii) sampling τ from its complete conditional distribution, see the top-right panel of Figure 3.8. Next, we construct two samplers by using different ASIS schemes. First, we implement ASIS for μ without conditioning on β. After marginalizing β out, Z is a sufficient augmentation for μ conditioning on τ, and $\bar{Z} = Z - X\mu$ is the corresponding ancillary augmentation. We name this scheme by ASIS I and construct the sampler using ASIS I, i.e., Sampler 3.11, by i) sampling Z and μ via ASIS conditioning on τ but not on β, ii) updating β from its complete conditional distribution, and iii) sampling τ from their complete conditional distributions, see the middle-left panel of Figure 3.8. Note that this ASIS sampler also utilize PCG. This is equivalent to replacing some

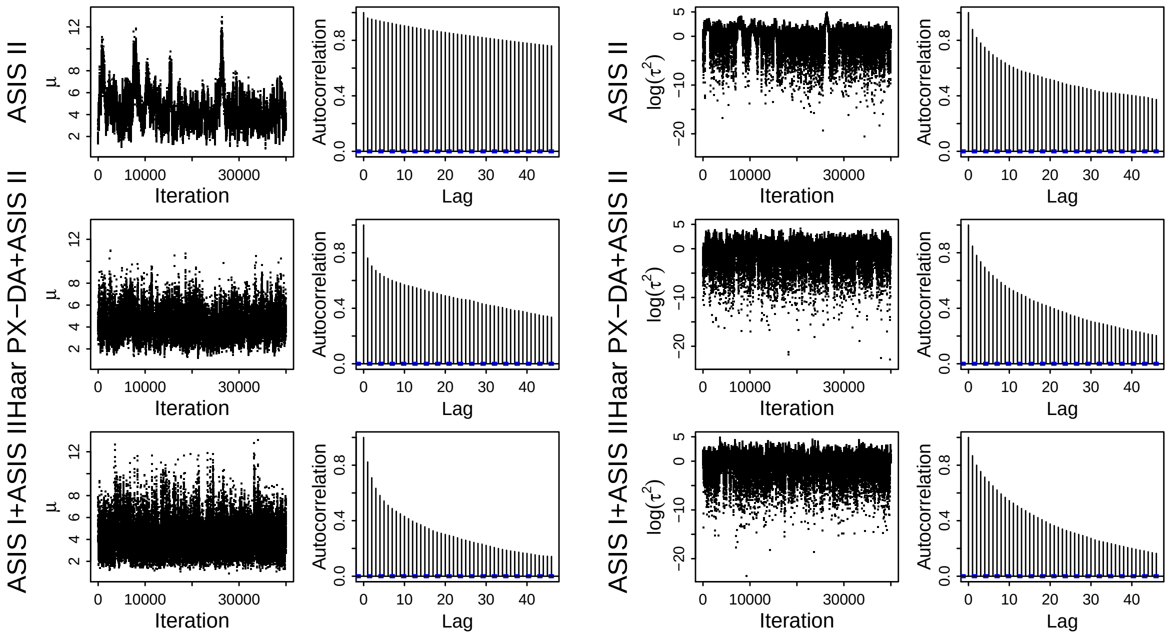

Figure 3.10: The sampling results of three samplers for the hierarchical probit model. The left two columns are the mixing and autocorrelation plots for the posterior draws of μ, while the right two columns are those for $\log(\tau^2)$. The three rows from top to bottom correspond to the ASIS II, the sampler combining Haar PX-DA and ASIS II, and the sampler combining ASIS I and ASIS II. The sampler combining ASIS I and ASIS II exhibits the best convergence among all of the six samplers for fitting the hierarchical probit model.

component kernels of the Gibbs sampler with PCG kernels and then swapping the PCG kernels with an ASIS kernel. The overlapping replacement is applicable as long as the overall stationary distribution of the sampler is maintained. Sampler 3.12 in the middle-right panel of Figure 3.8 is another ASIS sampler. Similar as the hierarchical t model in Section 3.3.3, conditioning on Z, β is a sufficient augmentation for (τ, μ), and $\bar{\beta} = (\beta - \mu)/\tau$ is the corresponding ancillary augmentation. We name this scheme by ASIS II. Sampler 3.12 implements ASIS II by first updating all the parameters from their complete conditional distributions and sampling (τ, μ) again conditioning on $\bar{\beta}$. Sampler 3.13 combines Haar PX-DA and ASIS II by sampling (μ, β) conditioning on τ with Haar PX-DA, and updating (τ, μ) with ASIS II conditioning on Z, see the bottom left panel of Figure 3.8. Finally, Sampler 3.14 combines two ASIS algorithms by sampling Z and μ via ASIS I, and updating β and τ via ASIS II. (Sampler 3.14 does not update μ with ASIS II).

	Gibbs	**Haar PX-DA**	**ASIS I**	**ASIS II**
μ	1.636	3.632	8.807	1.770
$\log(\tau^2)$	2.198	1.150	1.477	6.162
	Haar PX-DA+ASIS II	**ASIS I+ASIS II**		
μ	5.514	12.849		
$\log(\tau^2)$	9.748	11.092		

Table 3.4: The ESS per second of μ and $\log(\tau^2)$ for Samplers 3.9–3.14. The sampler combining ASIS I and ASIS II, i.e., Sampler 3.14, outperforms the other five samplers.

Now we verify the propriety of the two combined samplers, i.e., Samplers 3.13 and 3.14. Suppose $(Z', \beta', \mu', \tau')$ is a draw from the target distribution $p(Z, \beta, \mu, \tau|Y)$. For Sampler 3.13, since the kernel of Haar PX-DA maintains the target stationary distribution, after Step 2, the distribution of $(Z, \beta^\star, \mu^\star, \tau')$ is the target. Because the distribution of $(Z, \bar{\beta}, \mu, \tau)$ is equivalent to that of (Z, β, μ, τ). After transforming $\bar{\beta}$ to β at the end of Step 5, the distribution of (Z, β, μ, τ) is $p(Z, \beta, \mu, \tau|Y)$. For Sampler 3.14, after Step 3, (Z, μ) and β' are conditionally independent because Z and μ are updated without conditioning on β'. Fortunately, the marginal distribution of (Z, μ, τ') is still the target and Step 4 only depends on these three quantities. Thus after Step 4, the distribution of $(Z, \beta^\star, \mu, \tau')$ is again the target. As stated above, after transforming $\bar{\beta}$ to β at the end of Step 6, the distribution of (Z, β, μ, τ) is $p(Z, \beta, \mu, \tau|Y)$. Thus Samplers 3.13 and 3.14 are both proper.

We use a simulation study to compare the relative efficiencies of Samplers 3.9–3.14. Specifically, we set $\mu = 2$, $\tau = 0.05$, $n = 50$ and sample X from the uniform distribution on the interval $[-2, 2]$. For each sampler, we run a chain of 50,000 iterations with a burn-in of 10,000 using the same starting values. Figures 3.9 and 3.10 compare Samplers 3.9–3.14 in terms of the mixing and autocorrelation of μ and $\log(\tau^2)$. For either of Figures 3.9 and 3.10, the left two columns are the time-series and autocorrelation plots of μ, and the right two columns are those of $\log(\tau)$. The Haar PX-DA and ASIS I samplers improve the convergence of μ, but have little effect on τ. ASIS I behaves slightly better than Haar PX-DA. ASIS II has the opposite effect to Haar PX-DA and ASIS I. The samplers combining ASIS II and Haar PX-DA/ASIS I outperform the samplers using any of these algorithms alone. The sampler combining ASIS II and ASIS I exhibits the best convergence properties. Furthermore, we account for the computational time and compare the ESS per second of μ and $\log(\tau^2)$ from all of the six samplers in Table 3.4. The sampler combining ASIS II and ASIS I produces the largest ESS/sec for both parameters.

3.4 Discussion

Although acceleration strategies like MDA (Haar PX-DA), ASIS, and PCG are efficient in improving convergence, if there are more than one parameter exhibiting poor convergence, we may not be able to improve the convergence of all the parameters simultaneously by using one strategy alone. Numerical examples in this Chapter show that combining different algorithms into a coherent sampler can effectively break the plight and further improve computational efficiency. However, we need to implement combined samplers carefully to guarantee that they maintain the target stationary distribution. In fact, combining strategies into one sampler is an examples of using surrogate distributions, which is the topic of the next chapter, i.e., Chapter 4.

4

Surrogate Distribution Strategy

In Section 3.4 of Chapter 3, we mention that a sampler combining several acceleration strategies is an example of samplers using surrogate distributions. We have provided the definition of a surrogate distribution in Chapter 1, that is, a joint distribution of all unknown parameters in the model which shares certain marginal distributions with the target, but has lower correlations among its components. In this chapter, we propose a general framework to construct more efficient Gibbs-type samplers by replacing some of the conditional distributions of the target distribution with conditionals of a surrogate distribution. Like combining acceleration strategies into one sampler, using surrogate distributions may lead to incompatible conditional distributions. Thus, we must be cautious to guarantee that the desired stationary distribution is retained. Both theoretical arguments and numerical examples are deployed to illustrate the implementation of samplers using surrogate distributions and show the obtained computational efficiency. In Section 4.1, we use Gaussian models as motivating examples to shed light on the construction and computational advantage of samplers using surrogate distributions.

In Section 4.2, we derive the forms of surrogate distributions derived from PCG, Haar PX-DA, and ASIS algorithms, provide sufficient conditions to guarantee that samplers using surrogate conditionals maintain the desired stationary distributions, and then discuss the computational efficiency of the surrogate distribution strategy. Finally, in Section 4.3, we use univariate and multivariate t models, the spectral analysis model, and a simple hierarchical Gaussian model to illustrate the derivation of surrogate distributions from acceleration strategies, demonstrate the construction of samplers using surrogate distributions, and show their efficiency in improving convergence.

4.1 Motivating Examples

4.1.1 Two-step samplers using surrogate distributions

Suppose we wish to sample from $p(\psi_1, \psi_2)$. To achieve this goal, we start with a two-step Gibbs sampler:

Step 1: $\psi_1^{(t+1)} \sim p(\psi_1|\psi_2^{(t)})$, (Sampler 4.1)

Step 2: $\psi_2^{(t+1)} \sim p(\psi_2|\psi_1^{(t+1)})$.

Besides $p(\psi_1, \psi_2)$, We have a surrogate distribution of (ψ_1, ψ_2), $p_s(\psi_1, \psi_2)$, which shares the same marginal distributions with $p(\psi_1, \psi_2)$, but has lower correlation between ψ_1 and ψ_2. we construct Sampler 4.2 by replacing the conditional distribution in Step 1 of Sampler 4.1 with the conditional of $p_s(\psi_1, \psi_2)$, i.e.,

Step 1: $\psi_1^{(t+1)} \sim p_s(\psi_1|\psi_2^{(t)})$, (Sampler 4.2)

Step 2: $\psi_2^{(t+1)} \sim p(\psi_2|\psi_1^{(t+1)})$.

Furthermore, we replace both conditional distributions of Sampler 4.1 with conditionals of $p_s(\psi_1, \psi_2)$, and obtain Sampler 4.3,

Step 1: $\psi_1^{(t+1)} \sim p_s(\psi_1|\psi_2^{(t)})$, (Sampler 4.3)

Step 2: $\psi_2^{(t+1)} \sim p_s(\psi_2|\psi_1^{(t+1)})$,

which is in fact the ordinary Gibbs sampler updating $p_s(\psi_1, \psi_2)$. From Sampler 4.1 to Sampler 4.3, we increase the degree of using conditionals of the surrogate distribution.

We use a simulation study to illustrate the relative efficiencies of Samplers 4.1, 4.2, and 4.3 by setting $p(\psi_1, \psi_2)$ to

$$(\psi_1, \psi_2) \sim \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix} \right] \tag{4.1}$$

and $p_s(\psi_1, \psi_2)$ to

$$(\psi_1, \psi_2) \sim \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \right]. \tag{4.2}$$

Figure 4.1 compares 10,000 draws of ψ_1 generated by the three samplers in terms of mixing (first column) and autocorrelation (second column). (The performance of ψ_2 is similar to ψ_1.) Sampler 4.2 significantly outperforms Sampler 4.1. Although slightly, Sampler 4.3 exhibits even better convergence than Sampler 4.2. The results imply that for a two-step sampler, we can gain more computational efficiency by replacing more steps with draws from conditionals of a surrogate distribution. We discuss more on this point in Section 4.2.3.

While the stationary distributions of both Samplers 4.1 and 4.2 are $p(\psi_1, \psi_2)$, the stationary distribution of Sampler 4.3 is $p_s(\psi_1, \psi_2)$, which has lower correlation between ψ_1 and ψ_2 than $p(\psi_1, \psi_2)$, see the last column of Figure 4.1. If $p(\psi_1, \psi_2)$ is aimed to be maintained, Sampler 4.3 is improper and we can only use Samplers 4.1 and 4.2. If only marginal distributions of $p(\psi_1, \psi_2)$ are of interest, Sampler 4.3 can be manipulated to further improve the convergence. Moreover, permuting steps of Sampler 4.2 changes its stationary distribution, because the stationary distribution of the sampler ending with Step 1 after a cyclic permutation of steps of Sampler 4.2 is $p_s(\psi_1, \psi_2)$.

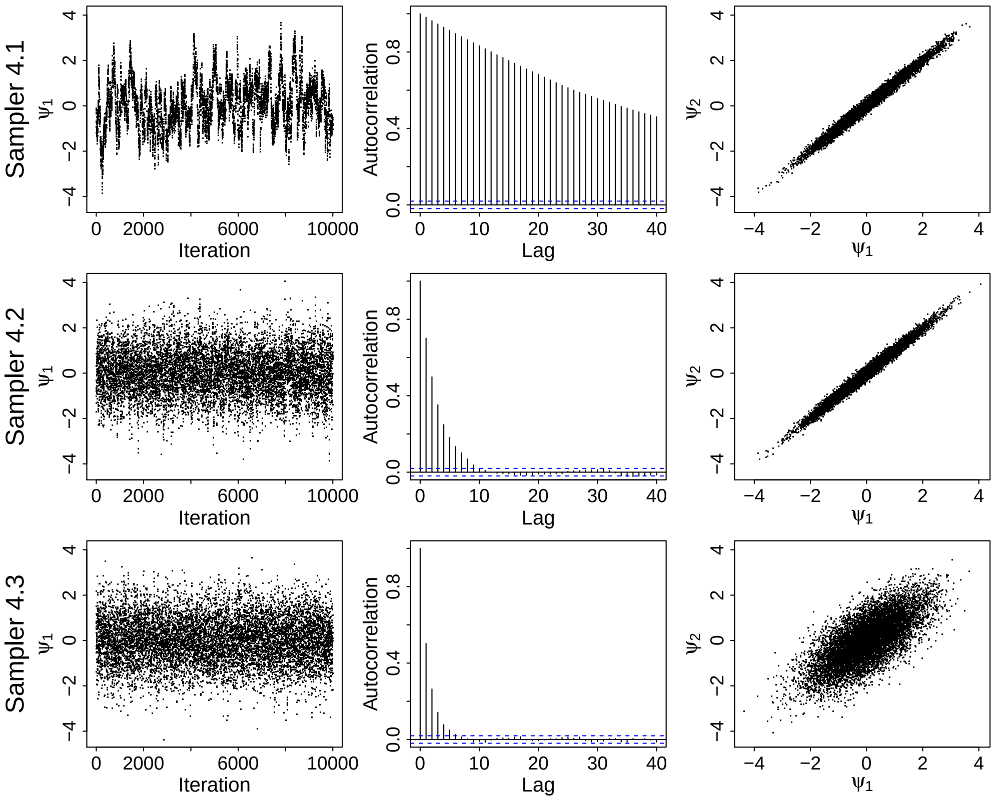

Figure 4.1: Samplers with different degrees of using the surrogate distribution for bivariate Gaussian distributions. The left two columns are the time-series and autocorrelation plots for 10,000 draws of ψ_1 from Samplers 4.1–4.3. Both Samplers 4.2 and 4.3 significantly outperform Sampler 4.1, and Sampler 4.3 behaves slightly better than Sampler 4.2. The last column displays scatter plots of ψ_1 and ψ_2 from the three samplers. The stationary distributions of both Samplers 4.1 and 4.2 are $p(\psi_1, \psi_2)$, whereas that of Sampler 4.3 is $p_s(\psi_1, \psi_2)$, with lower correlation between ψ_1 and ψ_2 than $p(\psi_1, \psi_2)$.

4.1.2 Three-step samplers using surrogate distributions

In this section we consider using surrogate distributions in N-step ($N > 2$) samplers. Without loss of generality, we set $N = 3$ and the target distribution to $p(\psi_1, \psi_2, \psi_3)$. The ordinary Gibbs sampler for updating $p(\psi_1, \psi_2, \psi_3)$ is

Step 1: $\psi_1^{(t+1)} \sim p(\psi_1 | \psi_2^{(t)}, \psi_3^{(t)})$, (Sampler 4.4)

Step 2: $\psi_2^{(t+1)} \sim p(\psi_2 | \psi_1^{(t+1)}, \psi_3^{(t)})$,

Step 3: $\psi_3^{(t+1)} \sim p(\psi_3 | \psi_1^{(t+1)}, \psi_2^{(t+1)})$.

When there are more than two unknown parameters, it can be practically difficult to derive a joint surrogate distribution for all the variables. Nevertheless, it can be relatively easy to obtain a surrogate distribution for a subset of parameters conditioning on others. Thus we consider a surrogate distribution, $p_s(\psi_1, \psi_2, \psi_3) = p_s(\psi_1, \psi_2|\psi_3)p(\psi_3)$, which has the same marginal of ψ_3 as the target, but different distribution of (ψ_1, ψ_2) conditioning on ψ_3. Specifically, $p_s(\psi_1|\psi_3) = p(\psi_1|\psi_3)$ and $p_s(\psi_2|\psi_3) = p(\psi_2|\psi_3)$ (equivalently, $p_s(\psi_1, \psi_3) = p(\psi_1, \psi_3)$ and $p_s(\psi_2, \psi_3) = p(\psi_2, \psi_3)$), but the correlation between ψ_1 and ψ_2 in $p_s(\psi_1, \psi_2|\psi_3)$ is lower than that in $p(\psi_1, \psi_2|\psi_3)$. First, we construct Sampler 4.5 by replacing the first step of Sampler 4.4 with a draw from the conditional distribution of the surrogate, $p(\psi_1, \psi_2|\psi_3)$, that is,

Step 1: $\psi_1^{(t+1)} \sim p_s(\psi_1|\psi_2^{(t)}, \psi_3^{(t)})$, (Sampler 4.5)

Step 2: $\psi_2^{(t+1)} \sim p(\psi_2|\psi_1^{(t+1)}, \psi_3^{(t)})$,

Step 3: $\psi_3^{(t+1)} \sim p(\psi_3|\psi_1^{(t+1)}, \psi_2^{(t+1)})$.

Next we replace Steps 1 and 3 of Sampler 4.4 with updates from conditionals of the surrogate, change the order of Step 2 and 3, and obtain Sampler 4.6,

Step 1: $\psi_1^{(t+1)} \sim p_s(\psi_1|\psi_2^{(t)}, \psi_3^{(t)})$, (Sampler 4.6)

Step 2: $\psi_3^{(t+1)} \sim p_s(\psi_3|\psi_1^{(t+1)}, \psi_2^{(t)})$,

Step 3: $\psi_2^{(t+1)} \sim p(\psi_2|\psi_1^{(t+1)}, \psi_3^{(t+1)})$.

Finally, we obtain Sampler 4.7 by replacing all the steps of Sampler 4.4 with draws from conditionals of $p_s(\psi_1, \psi_2, \psi_3)$, that is,

Step 1: $\psi_1^{(t+1)} \sim p_s(\psi_1|\psi_2^{(t)}, \psi_3^{(t)})$, (Sampler 4.7)

Step 2: $\psi_2^{(t+1)} \sim p_s(\psi_2|\psi_1^{(t+1)}, \psi_3^{(t)})$,

Step 3: $\psi_3^{(t+1)} \sim p_s(\psi_3|\psi_1^{(t+1)}, \psi_2^{(t+1)})$.

Sampler 4.7 is simply the standard three-step Gibbs sampler for updating $p_s(\psi_1, \psi_2, \psi_3)$. From Sampler 4.4 to Sampler 4.7, we increase the degree of using surrogate conditionals.

We also use a simulation study to illustrate the relative efficiencies of Samplers 4.4–4.7. We set $p(\psi_1, \psi_2, \psi_3)$ to

$$(\psi_1, \psi_2, \psi_3) \sim \mathrm{N}_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.95 & 0.81 \\ 0.95 & 1 & 0.95 \\ 0.81 & 0.95 & 1 \end{pmatrix} \right], \tag{4.3}$$

and $p_s(\psi_1, \psi_2, \psi_3)$ to

$$(\psi_1, \psi_2, \psi_3) \sim \mathrm{N}_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.70 & 0.81 \\ 0.70 & 1 & 0.95 \\ 0.81 & 0.95 & 1 \end{pmatrix} \right], \tag{4.4}$$

which is different from $p(\psi_1, \psi_2, \psi_3)$ only in the correlation between ψ_1 and ψ_2. We run a chain of 10,000 iterations using each of Samplers 4.4–4.7. In Figure 4.2, we compare the four samplers in terms of mixing (first column) and autocorrelation (second column) of ψ_3 (ψ_1 and ψ_2 behave similarly to ψ_3). Samplers 4.5–4.7 all improve the convergence of Sampler 4.4, Sampler 4.6 is better than Sampler 4.5, and Sampler 4.7 performs the best among the three. These results also indicate that replacing more steps with draws from conditionals of a surrogate distribution can lead to better convergence properties.

Note that the stationary distribution of Samplers 4.7 is the surrogate, $p_s(\psi_1, \psi_2, \psi_3)$, whereas the stationary distributions of other three samplers are all the target, $p(\psi_1, \psi_2, \psi_3)$. The surrogate distribution has lower correlation between ψ_1 and ψ_2 (see the last column of Figure 4.2), but has the same marginal distributions as the target for ψ_1, ψ_2, ψ_3, (ψ_1, ψ_3), and (ψ_2, ψ_3). If $p(\psi_1, \psi_2, \psi_3)$ is necessary to be maintained, Sampler 4.7 is improper and we can only consider using Samplers 4.5 and 4.6 to improve the convergence of Sampler 4.4. Or else the correlation between ψ_1 and ψ_2 is not of main interest, Sampler 4.7 can also be considered to further improve the convergence. Like two-step

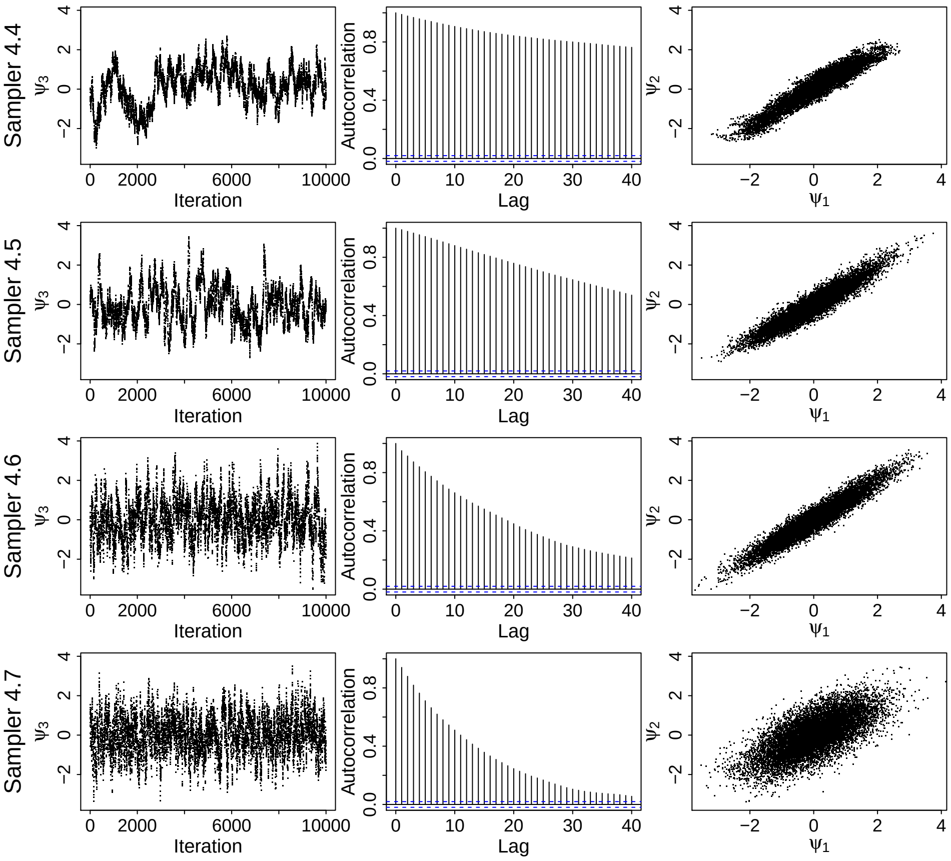

Figure 4.2: Samplers with different degrees of using the surrogate distribution for multivariate Gaussian distributions. The left two columns are the time-series and autocorrelation plots for 10,000 draws of ψ_3 from Samplers 4.4–4.7. Samplers 4.5–4.7 all outperform Sampler 4.4, Sampler 4.6 behaves better than Sampler 4.5, and Sampler 4.7 is the best. The last column displays scatter plots of ψ_1 and ψ_2 from the four samplers. The stationary distributions of Samplers 4.4–4.6 are $p(\psi_1, \psi_2, \psi_3)$, whereas that of Sampler 4.7 is $p_s(\psi_1, \psi_2, \psi_3)$, with lower correlation between ψ_1 and ψ_2 than $p(\psi_1, \psi_2, \psi_3)$.

samplers, permuting steps of a three-step sampler using conditionals of surrogate distributions may alter its stationary distribution. For example, the stationary distribution of the sampler ending with Step 1 after a cyclic permutation of steps of Sampler 4.5 is $p_s(\psi_1, \psi_2, \psi_3)$. For Sampler 4.6, a cyclic permutation of the steps ending with either Step 1 or 2 leads to a sampler that has the surrogate as its stationary distribution. With non-cyclic permutations, the stationary distributions of Samplers 4.5 and 4.6 can be unpredictable. One more complication of N-step ($N > 2$) samplers is that some combinations of conditionals of the target and surrogate distributions cannot produce samplers with known stationary distributions. For example, the sampler that updates

ψ_1 and ψ_2 from their conditional distributions of the target, and updates ψ_3 from its conditional of the surrogate does not have a known stationary distribution, neither does the sampler updating ψ_1 and ψ_2 from their conditional distributions of the surrogate, and ψ_3 from its conditional of the target, no matter how we permute the steps.

Motivated by the examples in this section, we believe that it is promising to improve the convergence of a Gibbs sampler by replacing some of its steps with draws from conditionals of a surrogate distribution. Moreover, samplers using surrogate distributions to a higher degree may have better computational efficiency. However, extra care must be taken to guarantee that the samplers using surrogate distributions maintain the desired stationary distributions. Thus, we provide general conditions for the samplers using surrogate distributions to preserve the desired stationary distributions. Furthermore, we use both theoretical arguments and numerical examples to illustrate the computational advantage of using surrogate distributions in a Gibbs sampler.

4.2 Using Surrogate Distributions in Gibbs-type Samplers

4.2.1 Deriving surrogate distributions via PCG, limiting MDA, and ASIS

In this section, we introduce how to derive surrogate distributions from existing acceleration strategies. Specifically, we focus on PCG, Haar PX-DA, and ASIS algorithms. First, we define the j^{th} $(j = 1, \ldots, N)$ *intermediate stationary distribution* of an N-step Gibbs-type sampler by the stationary distribution of the sampler ending with Step j after a cyclic permutation. Intermediate stationary distributions can be either the same as or different from the target distribution, and those which are different from the target are crucial for deriving surrogate distributions.

Deriving surrogate distributions via PCG

Recall the PCG sampler in Figure 1.1(d), which samples ψ_1 without conditioning on ψ_3. In Section 1.2.3 of Chapter 1, we derive the second intermediate stationary distribution of this sampler, that is, $p(\psi_1, \psi_2, \psi_4)p(\psi_3|\psi_2, \psi_4)$, which is different from the target distribution $p(\psi_1, \psi_2, \psi_3, \psi_4)$ only in that conditioning on (ψ_2, ψ_4), ψ_1 and ψ_3 are independent. This intermediate distribution is the surrogate that we can derive from the PCG sampler, and we denote it by $p_s(\psi_1, \psi_2, \psi_3, \psi_4)$. Thus we can consider the PCG sampler in Figure 1.1(d) as a sampler constructed by replacing Step 2 of its parent Gibbs sampler in Figure 1.1(a) with a draw from the conditional of the surrogate distribution $p_s(\psi_1, \psi_2, \psi_3, \psi_4)$ and permuting the steps. Analogously, Step 1 of Sampler 2.1, that is, the MH within PCG sampler with least partial collapsing for fitting the spectral model in (2.1), can be regarded as a draw from the conditional of its first intermediate stationary distribution, which is displayed in (2.2). Generally, the surrogate distribution we obtain from a PCG sampler is its intermediate stationary distribution after a reduced step.

Deriving surrogate distributions via Haar PX-DA

For MDA, we can only derive surrogate distributions from either the limiting MDA algorithm or Haar PX-DA algorithm. We focus on Haar PX-DA to obtain optimal efficiency. Suppose Conditions LW-1 and LW-2 described in Section 1.2.1 of Chapter 1 hold. Steps 2 and 3 of the Haar PX-DA sampler, i.e., Sampler 1.2l, are equivalent to first sampling $(\alpha^\star, \theta)$ from $\tilde{p}(\alpha, \theta|Y^\star_{\text{mis}}, Y_{\text{obs}})$ and then setting $Y_{\text{mis}} = \mathcal{G}^{-1}_{\alpha^\star}(Y^\star_{\text{mis}})$. Thus we specify the surrogate distribution from Haar PX-DA as the intermediate stationary distribution after updating $\tilde{p}(\alpha, \theta|Y^\star_{\text{mis}}, Y_{\text{obs}})$, that is,

$$p_s(Y_{\text{mis}}, \theta|Y_{\text{obs}}) = \left[\int \tilde{p}(\alpha, \theta|Y_{\text{mis}}, Y_{\text{obs}})\mathrm{d}\alpha\right] p(Y_{\text{mis}}|Y_{\text{obs}}). \tag{4.5}$$

This surrogate distribution has the same marginal distribution of Y_{mis} as the target. As stated in Section 1.2.1, Sampler 1.2l is proper. Thus the surrogate distribution also shares the same marginal distribution of θ as the target. (Note that the surrogate distribution derived from the limiting MDA algorithm has the form as (4.5).) With the surrogate distribution $p_s(Y_{\text{mis}}, \theta|Y_{\text{obs}})$, the sub-chain $\{\theta^{(t)}, t = 0, 1, \dots\}$ induced by

Sampler 1.2l is equivalent to that induced by the following sampler using the surrogate distribution, that is,

Step 1: $Y^{\star}_{\text{mis}} \sim p(Y_{\text{mis}}|\theta^{(t)}, Y_{\text{obs}})$, (Sampler 4.8)

Step 2: $\theta^{(t+1)} \sim p_s(\theta|Y^{\star}_{\text{mis}}, Y_{\text{obs}})$,

which is resulted from replacing the step updating θ of the parent DA sampler with a draw from the conditional of the surrogate distribution.

Recall that under Conditions LW-1 and LW-2, Y_{mis} can be represented by (β, r), where r is the orbit and β is the position on r. As stated in Section 3.2.3 of Chapter 3, sampling θ from $p(\theta|Y_{\text{mis}}, Y_{\text{obs}})$ is equivalent to updating θ conditioning on (β, r), whereas sampling θ from $\tilde{p}(\theta|Y_{\text{mis}}, Y_{\text{obs}})$ is equivalent to updating θ conditioning on r alone. Because the correlation between θ and r is lower than the correlation between θ and (β, r), the correlation between θ and Y_{mis} in $p_s(Y_{\text{mis}}, \theta|Y_{\text{obs}})$ is lower than that in $p(Y_{\text{mis}}, \theta|Y_{\text{obs}})$.

Deriving surrogate distributions via ASIS

We can also use ASIS to derive surrogate distributions. Suppose the target distribution is $p(\theta, Y_{\text{mis,S}}|Y_{\text{obs}})$. The surrogate distribution we obtain from the ASIS sampler, that is, Sampler 1.3 in Section 1.2.1 of Chapter 1 is the intermediate stationary distribution after sampling θ from $p(\theta|Y^{(t+1)}_{\text{mis,A}}, Y_{\text{obs}})$ in Step 3. As stated in Section 1.2.2 of Chapter 1, Step 2 of the ASIS sampler, i.e., Sampler 1.3, is equivalent to sampling $Y_{\text{mis,A}}$ from $p(Y_{\text{mis,A}}|Y^{\star}_{\text{mis,S}}, Y_{\text{obs}})$. Thus the surrogate distribution is

$$p_s(Y_{\text{mis,S}}, \theta|Y_{\text{obs}}) = \left[\int p(\theta|Y_{\text{mis,A}}, Y_{\text{obs}})p(Y_{\text{mis,A}}|Y_{\text{mis,S}}, Y_{\text{obs}})\mathrm{d}Y_{\text{mis,A}}\right] p(Y_{\text{mis,S}}|Y_{\text{obs}}). \quad (4.6)$$

Apparently, this surrogate distribution has the same marginal distribution of $Y_{\text{mis,S}}$ as the target. We verify here that the surrogate distribution shares the same marginal distribution of θ as the target, that is, $p_s(\theta|Y_{\text{obs}}) = p(\theta|Y_{\text{obs}})$:

Proof.

$$\begin{aligned}
p_s(\theta|Y_{\text{obs}}) &= \int p_s(\theta, Y_{\text{mis,S}}|Y_{\text{obs}})\mathrm{d}Y_{\text{mis,S}} = \int p_s(\theta|Y_{\text{mis,S}}, Y_{\text{obs}})p(Y_{\text{mis,S}}|Y_{\text{obs}})\mathrm{d}Y_{\text{mis,S}} \\
&= \int \left[\int p(\theta|Y_{\text{mis,A}}, Y_{\text{obs}})p(Y_{\text{mis,A}}|Y_{\text{mis,S}}, Y_{\text{obs}})\mathrm{d}Y_{\text{mis,A}}\right] p(Y_{\text{mis,S}}|Y_{\text{obs}})\mathrm{d}Y_{\text{mis,S}} \\
&= \int \left[\int p(\theta|Y_{\text{mis,A}}, Y_{\text{obs}}) \left(\int p(Y_{\text{mis,A}}|\theta^\star, Y_{\text{mis,S}}, Y_{\text{obs}})\right.\right. \\
&\quad \left.\left. p(\theta^\star|Y_{\text{mis,S}}, Y_{\text{obs}})\mathrm{d}\theta^\star\right) \mathrm{d}Y_{\text{mis,A}}\right] p(Y_{\text{mis,S}}|Y_{\text{obs}})\mathrm{d}Y_{\text{mis,S}} \\
&= \int\int\int p(\theta|Y_{\text{mis,A}}, Y_{\text{obs}})p(Y_{\text{mis,A}}|\theta^\star, Y_{\text{mis,S}}, Y_{\text{obs}}) \\
&\quad p(\theta^\star|Y_{\text{mis,S}}, Y_{\text{obs}})p(Y_{\text{mis,S}}|Y_{\text{obs}})\mathrm{d}\theta^\star\mathrm{d}Y_{\text{mis,A}}\mathrm{d}Y_{\text{mis,S}} \\
&= p(\theta|Y_{\text{obs}}).
\end{aligned} \tag{4.7}$$

□

The fourth equality holds due to the mathematical equivalence between sampling $Y_{\text{mis,A}}$ from $p(Y_{\text{mis,A}}|Y^\star_{\text{mis,S}}, Y_{\text{obs}})$ and Step 2 of Sampler 1.3. If the target distribution is $p(\theta, Y_{\text{mis,A}}|Y_{\text{obs}})$, we can derive the corresponding surrogate distribution, $p_s(\theta, Y_{\text{mis,A}}|Y_{\text{obs}})$, in the similar manner. Similar as the Haar PX-DA sampler, i.e., Sampler 1.2l, the subchain of θ induced by Sampler 1.3 is equivalent to that induced by the following sampler using the surrogate distribution $p_s(\theta, Y_{\text{mis,S}}|Y_{\text{obs}})$, that is,

Step 1: $Y^\star_{\text{mis,S}} \sim p(Y_{\text{mis,S}}|\theta^{(t)}, Y_{\text{obs}})$, (Sampler 4.9)

Step 2: $\theta^{(t+1)} \sim p_s(\theta|Y^\star_{\text{mis,S}}, Y_{\text{obs}})$,

constructed by replacing the step updating θ in the parent DA algorithm for sampling the target distribution $p(\theta, Y_{\text{mis,S}}|Y_{\text{obs}})$ with a draw from the conditional of $p_s(\theta, Y_{\text{mis,S}}|Y_{\text{obs}})$.

Yu and Meng (2011) verified that under Conditions YM-1 and YM-2 described in Section 3.2.3 of Chapter 3, the ASIS sampler is identical to the Haar PX-DA algorithm for the expanded model $\tilde{p}(\alpha, \tilde{Y}_{\text{mis,S}}, \theta|Y_{\text{obs}})$, where $\tilde{Y}_{\text{mis,S}} = \mathcal{H}_\alpha(Y_{\text{mis,S}})$. Thus, as for the surrogate distribution in (4.5) derived from the Haar PX-DA, the correlation between θ and $Y_{\text{mis,S}}$ in $p_s(Y_{\text{mis,S}}, \theta|Y_{\text{obs}})$ is lower than that in $p(Y_{\text{mis,S}}, \theta|Y_{\text{obs}})$.

4.2.2 Identifying Stationary Distributions of Samplers Using Surrogate Distributions

We construct a sampler using surrogate conditional distributions by replacing some conditionals of the parent Gibbs sampler with the conditionals of surrogate distributions. The intermediate stationary distributions may be different from the target. Thus, we must take care to guarantee that the stationary distribution of the overall sampler is the desired distribution.

Recall Sampler 4.2 in Section 4.1.1, which is a two-step sampler with one step replaced by a draw from the conditional distribution of the surrogate. One sufficient condition for Sampler 4.2 to maintain $p(\psi_1, \psi_2)$ as its stationary distribution is that $p_s(\psi_1) = p(\psi_1)$ and $p_s(\psi_2) = p(\psi_2)$. Suppose this condition holds and $(\psi_1^{(t)}, \psi_2^{(t)})$ is a sample from $p(\psi_1, \psi_2)$. Thus then

$$\begin{aligned} & \int p(\psi_1^{(t)}, \psi_2^{(t)}) p_s(\psi_1^{(t+1)}|\psi_2^{(t)}) p(\psi_2^{(t+1)}|\psi_1^{(t+1)}) \mathrm{d}\psi_1^{(t)} \mathrm{d}\psi_2^{(t)} \\ = & \int p(\psi_2^{(t)}) p_s(\psi_1^{(t+1)}|\psi_2^{(t)}) p(\psi_2^{(t+1)}|\psi_1^{(t+1)}) \mathrm{d}\psi_2^{(t)} \\ = & \; p_s(\psi_1^{(t+1)}) p(\psi_2^{(t+1)}|\psi_1^{(t+1)}) \\ = & \; p(\psi_1^{(t+1)}, \psi_2^{(t+1)}). \end{aligned} \tag{4.8}$$

Samplers 4.5 and 4.6 in Section 4.1.2 are three-step samplers using conditional distributions of the surrogate. One sufficient condition for both samplers to maintain the target stationary distribution is that $p_s(\psi_2, \psi_3) = p(\psi_2, \psi_3)$ and $p_s(\psi_1, \psi_3) = p(\psi_1, \psi_3)$. We verify its sufficiency for Sampler 4.6 and omit the proof for Sampler 4.5. Suppose $p_s(\psi_2, \psi_3) = p(\psi_2, \psi_3)$ and $p_s(\psi_1, \psi_3) = p(\psi_1, \psi_3)$ hold, and $(\psi_1^{(t)}, \psi_2^{(t)}, \psi_3^{(t)})$ is a draw from $p(\psi_1^{(t)}, \psi_2^{(t)}, \psi_3^{(t)})$, then

$$\begin{aligned} & \int p(\psi_1^{(t)}, \psi_2^{(t)}, \psi_3^{(t)}) p_s(\psi_1^{(t+1)}|\psi_2^{(t)}, \psi_3^{(t)}) p_s(\psi_3^{(t+1)}|\psi_1^{(t+1)}, \psi_2^{(t)}) \\ & p(\psi_2^{(t+1)}|\psi_1^{(t+1)}, \psi_3^{(t+1)}) \mathrm{d}\psi_1^{(t)} \mathrm{d}\psi_2^{(t)} \mathrm{d}\psi_3^{(t)} \\ = & \int p_s(\psi_1^{(t+1)}, \psi_2^{(t)}, \psi_3^{(t)}) p_s(\psi_3^{(t+1)}|\psi_1^{(t+1)}, \psi_2^{(t)}) p(\psi_2^{(t+1)}|\psi_1^{(t+1)}, \psi_3^{(t+1)}) \mathrm{d}\psi_2^{(t)} \mathrm{d}\psi_3^{(t)} \\ = & \; p(\psi_1^{(t+1)}, \psi_2^{(t+1)}, \psi_3^{(t+1)}). \end{aligned} \tag{4.9}$$

Generally, the following two conditions are sufficient to ensure that the overall sampler maintains the target stationary distribution, that is,

i) The input to each conditional distribution of the sampler follows the correct distribution. Specifically, consider a generic N-step sampler using surrogate distributions, which updates ψ_j in Step j ($j = 1, \ldots, N$). If Step j is a draw from the conditional of the target distribution, that is, $\psi_j \sim p(\psi_j | \mathcal{F}_j(\psi, \psi'))$, we must guarantee $\mathcal{F}_j(\psi, \psi') \sim p(\mathcal{F}_j(\psi, \psi'))$. If $\psi_j \sim p_s(\psi_j | \mathcal{F}_j(\psi, \psi'))$, $\mathcal{F}_j(\psi, \psi')$ must follow the corresponding surrogate distribution, $p_s(\mathcal{F}_j(\psi, \psi'))$.

ii) the last step of the sampler is a draw from the complete conditional of the desired stationary distribution.

These two conditions ensure that the last steps produces a draw from the target if the input to the first step follows the target distribution.

In order to guarantee the two conditions holding, we sometimes need to permute the steps of the sampler after replacing some of its updates with draws from conditional distributions of the surrogate. For example, after replacing Steps 1 and 3 of Sampler 4.4 with updates from conditionals of the surrogate distribution, we change the order of Steps 2 and 3 to guarantee the two conditions above holding for Sampler 4.6.

If the correlations among parameters are not of main interest, we allow the stationary distribution of the sampler to be the surrogate, which is different from the target, so that it is possible to replace more steps with draws from conditionals of the surrogate distribution and thus obtain more computational efficiency (e.g., Samplers 4.3 and 4.7). Under this scenario, Conditions i) and ii) are also sufficient to ensure that the stationary distribution of the overall sampler is the surrogate.

We can construct a proper sampler by replacing conditional distributions of the parent Gibbs sampler with conditionals from MORE THAN ONE surrogate distribution, so long as Conditions i) and ii) above hold. The combining strategy introduced in Chapter 3 is an example for this case, because combining different acceleration strategies is equivalent to replacing some conditionals of the parent sampler with conditionals of the surrogate distributions derived from these strategies.

4.2.3 Computational Efficiency Gained by Using Surrogate Distributions

Comparing two-step samplers using the surrogate distribution to different degrees

In Section 4.1, we investigate three two-step samplers, i.e., Sampler 4.1–4.3. Recall that Sampler 4.1 is the parent Gibbs sampler for updating the target distribution $p(\psi_1, \psi_2)$, Sampler 4.2 replaces Step 1 of Sampler 4.1 with an update from the conditional of the surrogate distribution $p_s(\psi_1, \psi_2)$, and Sampler 4.3 replaces both steps of Sampler 4.1 with conditionals of the surrogate, resulting in a Gibbs sampler for updating the surrogate distribution $p_s(\psi_1, \psi_2)$.

To compare the convergence rates of the three samplers using the surrogate distribution to different degrees, we start with the special case that both the target and surrogate distributions are bivariate Gaussian. Specifically, suppose

$$p(\psi_1, \psi_2) = \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r_p \\ r_p & 1 \end{pmatrix} \right] \text{ and } p_s(\psi_1, \psi_2) = \mathrm{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r_{p_s} \\ r_{p_s} & 1 \end{pmatrix} \right]. \tag{4.10}$$

Under this scenario, we can derive the convergence rate, that is, the spectral radius of the forward operator induced by each of Samplers 4.1–4.3 analytically. We denote the forward operators corresponding to Samplers 4.1–4.3 by P_1, P_2, and P_3 respectively. By Theorem 3.2 of Liu *et al.* (1994), the spectral radius of P_1, $r(P_1)$, is ρ_p^2, and that of P_3, $r(P_3)$, is $\rho_{p_s}^2$, where ρ_p and ρ_{p_s} are the maximum correlation between ψ_1 and ψ_2 for $p(\psi_1, \psi_2)$ and $p_s(\psi_1, \psi_2)$ respectively. For bivariate Gaussian distributions, the maximal correlation between ψ_1 and ψ_2 is the absolute value of their correlation (Lancaster, 1958). Thus $r(P_1) = r_p^2$ and $r(P_3) = r_{p_s}^2$. The spectral radius of P_2, $r(P_2)$, is equal to the maximum eigenvalue of the transition matrix M such that $\mathrm{E}(\psi^{(t+1)}|\psi^{(t)}) = M\psi^{(t)} + v$,

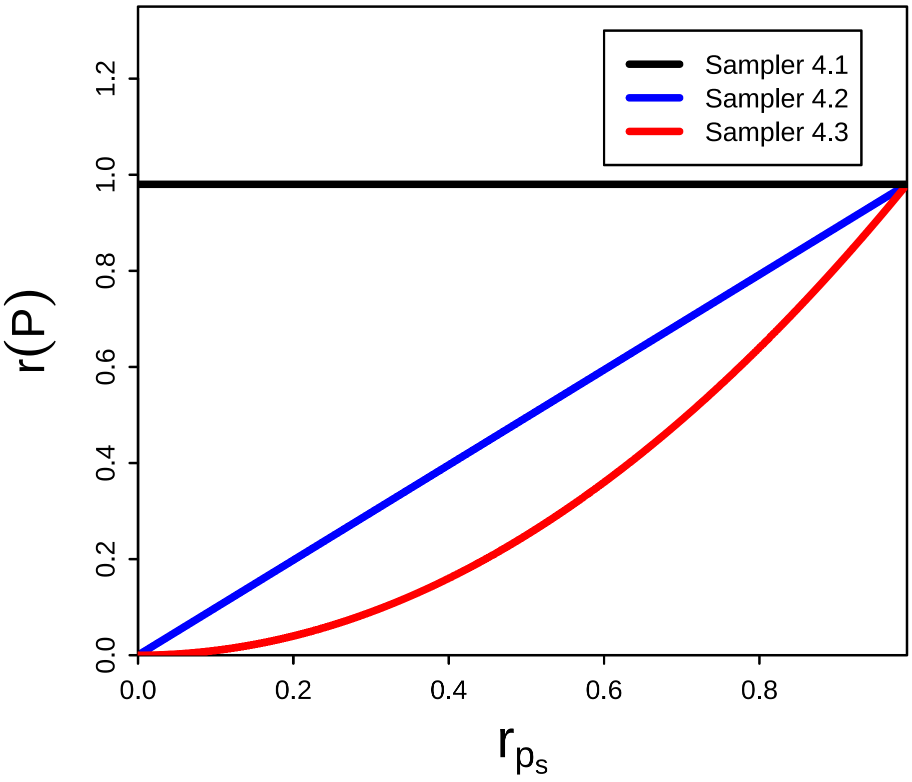

Figure 4.3: Convergence rates of Samplers 4.1–4.3 as functions of the correlation between ψ_1 and ψ_2 in the surrogate distribution $p_s(\psi_1, \psi_2)$ for bivariate Gaussian target and surrogate. The two samplers using conditionals of the surrogate distribution, Samplers 4.2 and 4.3, are both efficient in improving the convergence of the parent Gibbs sampler, i.e., Sampler 4.1. Moreover, Sampler 4.3, the sampler using the surrogate distribution to the highest degree, performs even better than Sampler 4.2.

where v is a vector of constants. With (4.10), $r(P_2) = |r_p r_{p_s}|$. We fix r_p at 0.99, and let r_{p_s} vary in the range $[0, r_p]$. Figure 4.3 shows the convergence rates of Samplers 4.1–4.3 as a function of r_{p_s}. Because smaller convergence rate corresponds to faster convergence (see Section 1.3 of Chapter 1), and $r(P_1) \geq r(P_2) \geq r(P_3)$ holds uniformly for $r_{p_s} \in [0, r_p]$, we conclude that with bivariate Gaussian target and surrogate distributions, the two samplers using conditional distributions of the surrogate, Samplers 4.2 and 4.3, both converge faster than the parent Gibbs sampler, i.e., Sampler 4.1, and Sampler 4.3, the sampler using the surrogate distribution to the highest degree, has faster convergence than Sampler 4.2.

For general bivariate target and surrogate distributions, although we can obtain the convergence rates of the two Gibbs samplers, i.e., Samplers 4.1 and 4.3, via Theorem 3.2 of Liu *et al.* (1994), we can hardly derive the convergence rate of Sampler 4.2. Thus we instead compare the three samplers in terms of their cyclic-permutation bounds. For a standard two-step Gibbs sampler, the norms of the forward operators induced by the corresponding 0-step-lagged and 1-step-lagged Gibbs samplers are equal. Thus by definition, the cyclic-permutation bound of a two-step Gibbs sampler is equal to the norm of the corresponding forward operator, equivalently the maximum correlation between the two components updated in two steps of the Gibbs sampler (see Theorem 3.2 of Liu *et al.* (1994)). Henceforth the cyclic-permutation bound of Sampler 4.1 is $||P_1|| = \rho_p$ and that of Sampler 4.3 is $||P_3|| = \rho_{p_s}$. The cyclic-permutation bound of Sampler 4.2 is $\min\{\rho_p, \rho_{p_s}\}$. (The proof is displayed in Appendix C.) As long as $\rho_{p_s} < \rho_p$, the two samplers using the surrogate distribution, i.e., Samplers 4.2 and 4.3 have the same cyclic-permutation bound, which is smaller than that of the parent Gibbs sampler, i.e., Sampler 4.1. This implies that replacing steps of a two-step Gibbs sampler with updates from conditional distributions of the surrogate has effects on improving the convergence properties of the sampler. Although the exact relationship of the convergence rate of Sampler 4.2 with those of Samplers 4.1 and 4.3 remains unclear, the special case of bivariate Gaussian models indicates that for a two-step Gibbs sampler, using the surrogate distribution to higher degree leads to better convergence.

Comparing three-step samplers using the surrogate distribution to different degrees

Recall Samplers 4.4–4.7 in Section 4.1. Sampler 4.4 is the parent three-step Gibbs sampler for updating the target distribution $p(\psi_1, \psi_2, \psi_3)$. Samplers 4.5, 4.6, and 4.7 replace one, two, and three steps of Sampler 4.4 with updates from the conditionals of the surrogate distribution $p_s(\psi_1, \psi_2, \psi_3)$, respectively.

If the target and surrogate distributions are both Gaussian, we can derive the spectral radius of the forward operator induced by each of Samplers 4.4–4.7 analytically via computing the maximum eigenvalue of the transition matrix M such that $\mathrm{E}(\psi^{(t+1)}|\psi^{(t)}) =$

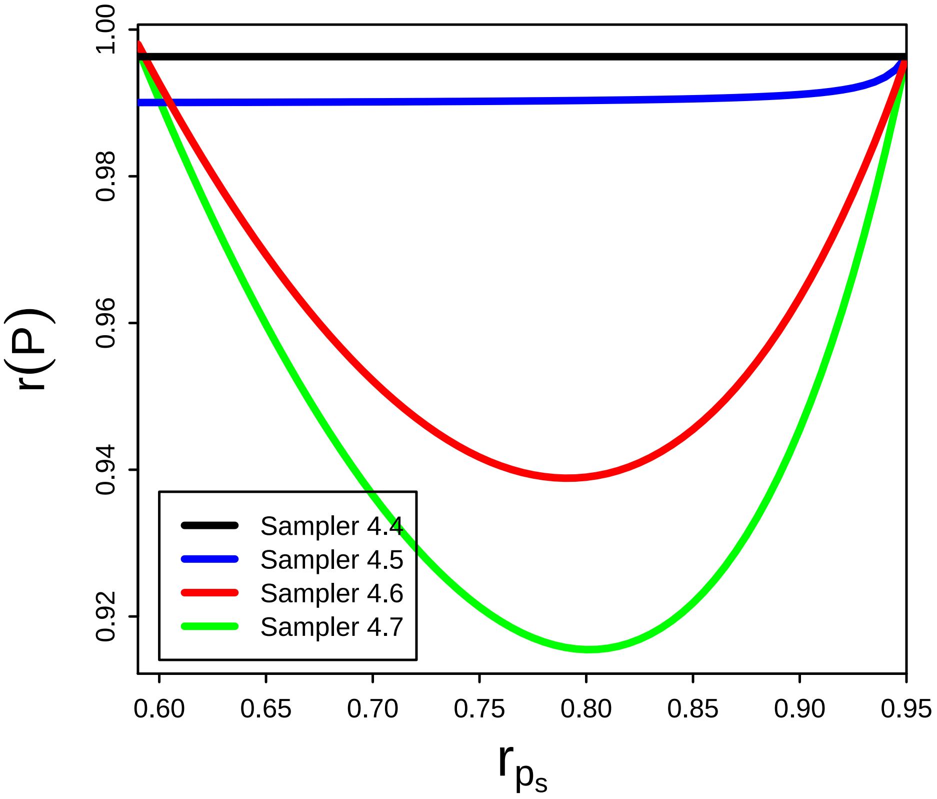

Figure 4.4: Convergence rates of Samplers 4.4–4.7 as functions of the correlation between ψ_1 and ψ_2 in the surrogate distribution $p_s(\psi_1, \psi_2, \psi_3)$ for Gaussian target and surrogate. When $r_{p_s} > 0.62$, the three samplers using the surrogate distribution, i.e., Samplers 4.5–4.7, all have smaller convergence rates than the parent Gibbs sampler, i.e., Sampler 4.4, and using the surrogate distribution to higher degree produces faster convergence. When r_{p_s} gets smaller, the three samplers keep converging faster than the parent Gibbs sampler. However, the samplers using more than one surrogate conditional, i.e., Samplers 4.6 and 4.7, can perform worse than the sampler using just one surrogate conditional, i.e., Sampler 4.5. When r_{p_s} gets close to 0.59, the sampler using surrogate conditionals in two steps even converges slower than the parent sampler.

$M\psi^{(t)} + v$. Thus we first compare Samplers 4.4–4.7 for the case that both $p(\psi_1, \psi_2, \psi_3)$ and $p_s(\psi_1, \psi_2, \psi_3)$ are Gaussian distributions. We specify $p(\psi_1, \psi_2, \psi_3)$ as in (4.3), and set $p_s(\psi_1, \psi_2, \psi_3)$ to

$$(\psi_1, \psi_2, \psi_3) \sim \mathrm{N}_3\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r_{p_s} & 0.81 \\ r_{p_s} & 1 & 0.95 \\ 0.81 & 0.95 & 1 \end{pmatrix}\right]. \tag{4.11}$$

Suppose $r_{p_s} \geq 0$, and let r_{p_s} vary in the range $[0.59, 0.95]$, where 0.59 is the smallest value of $r_{p_s} \geq 0$ to make the variance-covariance matrix in (4.11) positive-definite. Figure 4.4 shows the convergence rates of Samplers 4.4–4.7 as a function of r_{p_s}. With Gaussian target and surrogate distributions in (4.3) and (4.11) respectively, if $r_{p_s} > 0.62$, then the three samplers using the surrogate distribution all have smaller convergence rates than the parent Gibbs sampler updating the target distribution, and using the surrogate distribution to higher degree leads to faster convergence. When r_{p_s} gets smaller, the three samplers using the surrogate distributions keep converging faster than the parent Gibbs sampler. However, the samplers using more than one surrogate conditional can converge slower than the sampler using just one surrogate conditional. When r_{p_s} gets close to 0.59, the sampler using surrogate conditionals in two steps even converges slower than the parent sampler.

With general target and surrogate distributions, comparing the convergence rates, or simply the cyclic-permutation bounds of Samplers 4.4–4.7 becomes more complicated. In practice, we typically derive surrogate distributions from the PCG, Haar PX-DA, and ASIS algorithms. The efficiency of these acceleration algorithms in improving convergence properties has been confirmed. Henceforth, if the surrogate distribution $p_s(\psi_1, \psi_2, \psi_3)$ is derived from one of PCG, Haar PX-DA, and ASIS, it is easy to verify that the sampler using the surrogate conditional in one step, i.e., Sampler 4.5, has smaller cyclic-permutation bound than the parent Gibbs sampler, i.e., Sampler 4.4. However, using the surrogate distribution to higher degree does not necessarily leads to faster convergence. Thus in practice, finding the optimal algorithm from Samplers 4.5–4.7 to improve the convergence of the parent sampler is on a case-by-case basis. The results for three-step samplers can be easily generalized to N-step($N \geq 3$) samplers.

In all, replacing steps with draws from conditionals of the surrogate distribution generally improves the convergence of the parent sampler. For two-step samplers, using surrogate conditional distributions in more steps typically leads to more efficient samplers, whereas for N-step($N \geq 3$) samplers, this trend does not often hold. Moreover, we can consider more than one surrogate distribution when constructing a sampler using surrogate conditionals, as the combining strategy does. We have proved the computational advantage of the combining strategy in Section 3.2.3 of Chapter 3. Thus the samplers constructed by replacing some steps of the parent Gibbs sampler with updates from conditionals of the surrogate distributions derived from different acceleration strategies are expected to be efficient in improving the convergence of the parent sampler.

4.3 Illustration Example

In this section, we use three examples, t-distribution models, the spectral analysis model, and a simple hierarchical Gaussian model to demonstrate deriving surrogate distributions from Haar PX-DA, PCG, and ASIS algorithms, and show the computational efficiency of the samplers using surrogate distributions.

4.3.1 t-distribution model

We first use t-distribution model to illustrate the deriving of a surrogate distribution from the Haar PX-DA algorithm, and compare the convergence properties of three two-step samplers, which use the Haar PX-DA surrogate distribution to different degrees. We consider both univariate and multivariate t-distributions.

Univariate t-distribution

We consider the following univariate t-distribution model:

$$Y_i \overset{\text{iid}}{\sim} t_\nu(\mu, \sigma^2), \text{ for } i = 1, \ldots, n, \tag{4.12}$$

where $Y = (Y_1, \ldots, Y_n)$ are the observations; ν is the known degrees of freedom of the t-distribution; μ and σ are the unknown location and scale parameters respectively and we specify non-informative prior distributions to them, that is, $p(\mu, \sigma^2) \propto 1/\sigma^2$. It is difficult to sample $p(\mu, \sigma^2|Y)$ directly. To simplify the implementation, we extend the model in (4.12) with the latent variable $q = (q_1, \ldots, q_n)$ as in Meng and van Dyk (1999),

DA
(Sampler 4.10)

1. $p(q|Y, \mu', \sigma^{2'})$
2. $p(\mu, \sigma^2|q, Y)$

Haar PX-DA
(Sampler 4.11)

1. $p(q^\star|Y, \mu', \sigma^{2'})$
2. $\tilde{p}(\alpha|Y, q^\star)$; set $q = q^\star/\alpha$
3. $p(\mu, \sigma^2|Y, q)$

Surrogate
(Sampler 4.12)

1. $p_s(q|Y, \mu', \sigma^{2'})$,
2. $p_s(\mu, \sigma^2|q, Y)$.

Figure 4.5: Three samplers for fitting the univariate t-distribution model in (4.13). The left, middle, and right panels show the steps of the DA sampler (Sampler 4.10), the Haar PX-DA sampler (Sampler 4.11), and the two-step Gibbs sampler for updating the surrogate distribution derived from the Haar PX-DA algorithm (Sampler 4.12).

that is,

$$Y_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{q_i}\right), \text{ for } i = 1, \ldots, n, \tag{4.13}$$

where $q_i \overset{\text{iid}}{\sim} \chi^2_\nu/\nu$.

We compare the efficiencies of three samplers for fitting the univariate t-distribution model. We start with the standard DA sampler which iteratively updates q and (μ, σ^2) from their conditional distributions of $p(q, \mu, \sigma^2|Y)$, see Sampler 4.10 in the left panel of Figure 4.5. To improve the convergence of the DA sampler, we construct an MDA sampler. The working parameter α is introduced into the model by setting

$$Y_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu, \frac{\alpha\sigma^2}{\tilde{q}_i}\right), \tag{4.14}$$

where $\tilde{q}_i|\alpha \sim \alpha\chi^2_\nu/\nu$. Since Conditions LW-1 and LW-2 introduced in Section 1.2.1 of Chapter 1 hold for this example, we use the Haar PX-DA algorithm to achieve optimality. We specify the Haar measure prior to α, that is, $p_\infty(\alpha) \propto 1$, and obtain the Haar PX-DA sampler, i.e., Sampler 4.11 in the middle panel of Figure 4.5. We derive a surrogate distribution via the method introduced in Section 4.2.1 from the Haar PX-DA sampler, that is,

$$p_s(\mu, \sigma^2, q|Y) = \left[\int \tilde{p}(\alpha, \mu, \sigma^2|q, Y)\mathrm{d}\alpha\right] p(q|Y). \tag{4.15}$$

As stated in Section 4.2.1, the sub-chain of (μ, σ^2) induced by the Haar PX-DA sampler, i.e., Sampler 4.11, is equivalent to that induced by the sampler which updates from

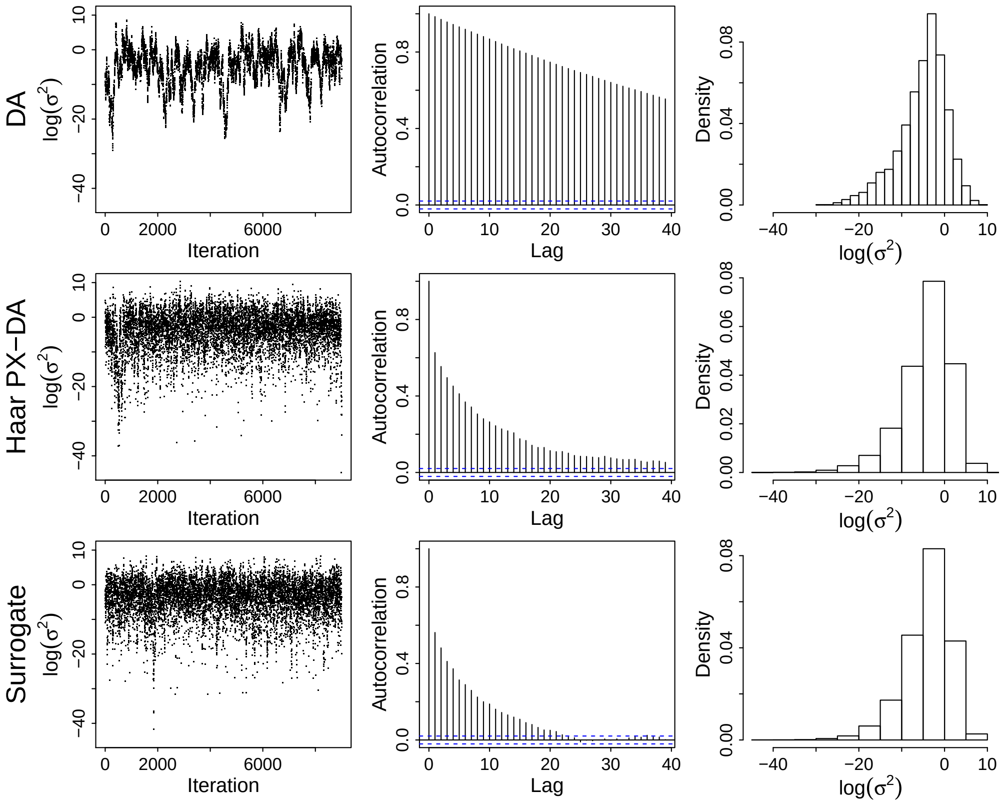

Figure 4.6: The sampling results of σ^2 for Samplers 4.10–4.12. The first, second, and third columns correspond to the trace plots, autocorrelation plots, and histograms of $\log(\sigma^2)$. Both samplers using the surrogate distribution, i.e., Samplers 4.11 and 4.12, are efficient in improving the convergence of σ^2, and the sampler using the surrogate distribution to a higher degree, i.e., Sampler 4.12, performs slightly better.

$p(q|Y, \mu', \sigma^{2\prime})$ and $p_s(\mu, \sigma^2|q, Y)$ iteratively. Thus the Haar PX-DA sampler can be considered as a two-step sampler which replaces one step of the parent DA sampler with a draw from the conditional of the surrogate distribution $p_s(\mu, \sigma^2, q|Y)$. Because in this example, our main interest focuses on the marginal distribution of (μ, σ^2), we also consider the two-step Gibbs sampler for updating $p_s(\mu, \sigma^2, q|Y)$, that is, Sampler 4.12 in the right panel of Figure 4.5. The construction of Sampler 4.12 can be considered as replacing both steps of the parent DA sampler with draws from the conditionals of the surrogate distribution. Henceforth, Samplers 4.10–4.12 are an example for Samplers 4.1–4.3, which are three two-step samplers using the surrogate distribution to different degrees. Details of Samplers 4.10–4.12 and other samplers in this chapter appear in Appendix C.

We use a simulation study to compare the relative efficiencies of Samplers 4.10–4.12. We set $n = 6$, $\nu = 0.1$, $\mu = 0$, and $\sigma = 1$. For each sampler, we run a chain of 10,000 with a burn-in of 1,000, starting from the same initial values. Figure 4.6 shows the mixing, autocorrelation, and posterior density estimation of $\log(\sigma^2)$ for Samplers 4.10–4.12. The three samplers behave similarly on μ and we omit the corresponding results here. Both samplers using the surrogate distribution, i.e., Samplers 4.11 and 4.12 are efficient in improving the convergence of σ^2, and the sampler using the surrogate distribution to a higher degree, i.e., Sampler 4.12, performs slightly better.

Multivariate t-distribution

Generalizing the data augmentation scheme in (4.13) to the multivariate version is straightforward. We write

$$Y_i \overset{\text{ind}}{\sim} \mathrm{N}_d\left(\mu, \frac{\Sigma}{q_i}\right), \text{ for } i = 1, \dots, n, \tag{4.16}$$

where $Y = (Y_1, \dots, Y_n)$ are the observation and each Y_i is a $(d \times d)$ vector, marginally following a multivariate location-scale t-distribution with ν degrees of freedom (ν is known); $q_i \overset{\text{iid}}{\sim} \chi^2_\nu/\nu$; μ and Σ are the unknown $(d \times 1)$ location vector and $(d \times d)$ scale matrix. We also give the non-informative prior distribution to (μ, Σ), that is, $p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$. Generalizing the marginal augmentation scheme in (4.14) to the multivariate case, we obtain

$$Y_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu, \frac{\alpha\Sigma}{\tilde{q}_i}\right), \tag{4.17}$$

where $\tilde{q}_i|\alpha \sim \alpha\chi^2_\nu/\nu$. We also specify the Haar measure prior to α, that is, $p_\infty(\alpha) \propto 1$.

The algorithms we consider for fitting the multivariate are also the standard DA sampler, Haar PX-DA sampler, and the two-step Gibbs sampler for updating the surrogate distribution derived from the PX-DA sampler. The steps of these samplers are almost the same as those corresponding to the univariate t-distribution, which are listed in Figure 4.5. The only modification is that σ^2 is replaced by Σ. We omit the plot display-

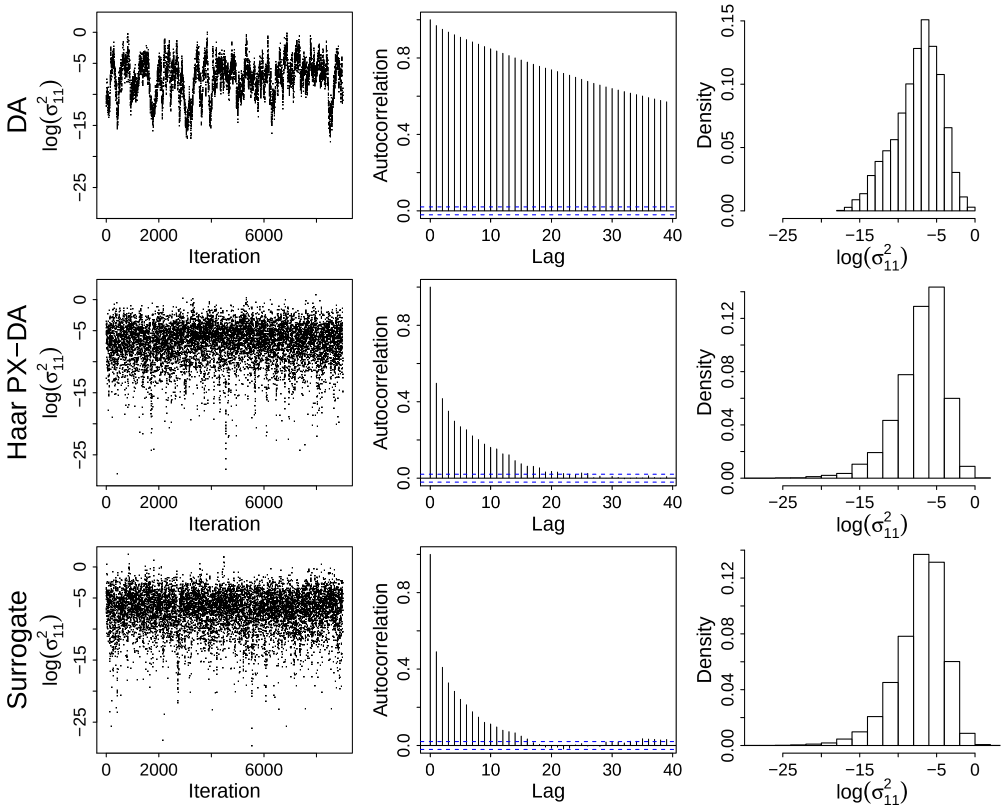

Figure 4.7: The sampling results of σ^2_{11} for the three samplers for fitting the multivariate t-distribution model. The first, second, and third columns correspond to the trace plots, autocorrelation plots, and histograms of $\log(\sigma^2_{11})$. Both samplers using the surrogate distribution, i.e., the Haar PX-DA sampler and the Gibbs sampler for updating the surrogate distribution, are efficient in improving the convergence of σ^2_{11}, and the sampler using the surrogate distribution to a higher degree, i.e., the Gibbs sampler for updating the surrogate distribution, performs slightly better.

ing the samplers for fitting the multivariate t-distribution model. The deriving of the surrogate distribution from the Haar PX-DA sampler for the multivariate t-distribution model is also a copy of that for the univariate t-distribution model as in (4.15), except replacing σ^2 with Σ.

We use a simulation study to compare the three samplers for fitting the multivariate t-distribution model. We set $n = 10$, $d = 2$, $\nu = 0.1$, $\mu = (0, 0)$, and $\Sigma = \begin{pmatrix} 0.1 & 0.7 \times \sqrt{0.1} \times \sqrt{5} \\ 0.7 \times \sqrt{0.1} \times \sqrt{5} & 5 \end{pmatrix}$. For each sampler, we run a chain of 10,000 with a burn-in of 1,000, starting from the same initial values. Figure 4.7 presents the mixing, autocorrelation, and posterior density estimation of the logarithm of the first diagonal component of the scale matrix Σ, that is, $\log(\sigma^2_{11})$ for the three samplers. Both samplers using the surrogate distribution, i.e., the Haar PX-DA sampler and the Gibbs sampler for updating the surrogate distribution, are efficient in improving the convergence of σ_{11}^2, and the sampler using the surrogate distribution to a higher degree, i.e., the Gibbs sampler for updating the surrogate, performs slightly better. The three samplers perform similarly on μ and the correlation parameter of the scale matrix r_{12}, with the Gibbs sampler for updating the surrogate slightly more efficient, $\log(\sigma_{22}^2)$ behaves similarly to $\log(\sigma_{11}^2)$, and we omit the results for μ, r, and $\log(\sigma_{22}^2)$ here.

This results of the univariate and multivariate t-distribution model confirm our on two-step samplers using the surrogate distribution, that is, replacing more steps with draws from conditionals of the surrogate distribution leads to more efficient samplers.

4.3.2 Spectral Analysis in X-ray Astrophysics

Recall the spectral analysis model introduced in Section 2.1 of Chapter 2. In this section we use an extension of this model to illustrate the deriving of a surrogate distribution from the PCG algorithm, and compare the convergence properties of two three-step samplers, which use the PCG surrogate distribution to different degrees.

The model we use here is similar to that in (2.1). The only difference is that we specify the emission line part as a sum of K lines, instead of just one line, that is,

$$Y_i \overset{\text{ind}}{\sim} \text{Pois}\left(\alpha(E_i^{-\beta} + \sum_{k=1}^{K} \gamma_k I\{i = \mu_k\})e^{-\phi/E_i}\right), \text{ for } i = 1, \ldots, n, \tag{4.18}$$

where $Y = (Y_1, \ldots, Y_n)$ are the recorded photon counts and $E = (E_1, \ldots, E_n)$ are the known energies in n bins. For simplicity, we fix the value of $\gamma = (\gamma_1, \ldots, \gamma_K)$ in this example. Thus α, β, μ_1, …, μ_K, and ϕ are the unknown parameters. We assume these parameters are *a priori* independent, each of $\mu = (\mu_1, \ldots, \mu_K)$ is *a priori* uniform on $\{1, \ldots, n\}$, and α, β, and ϕ are *a priori* uniform on the positive real line $\mathbb{R}^+$.

We compare the performance of two three-step samplers for fitting the model in (4.18). The first sampler we consider is an MH within PCG sampler, which updates μ and (β, ϕ)

MH within PCG (Sampler 4.13)	**Surrogate (Sampler 4.14)**
1. $p(\mu \vert \beta', \phi', Y)$	1. $p(\mu \vert \beta', \phi', Y)$
2. $p(\beta, \phi \vert \mu, Y)$	2. $p(\beta, \phi \vert \mu, Y)$
3. $p(\alpha \vert \beta, \mu, \phi, Y)$	3. $p(\alpha \vert \beta, \phi, Y)$

Figure 4.8: Two samplers for fitting the spectral analysis model in (4.18). The left and right panels show the steps of the MH within PCG sampler (Sampler 4.13), which uses the conditional of the surrogate distribution $p_s(\alpha, \beta, \mu, \phi|Y)$ in one step, and the three-step sampler using the surrogate conditionals in two steps (Sampler 4.14).

without conditioning on α, see Sampler 4.13 in the first panel of Figure 4.8. Steps 1 and 2 require MH updates. We derive a surrogate distribution via the method introduced in Section 4.2.1 from the PCG sampler. Specifically, the surrogate distribution is the intermediate stationary distribution after the reduced Step 1 of Sampler 4.13, that is,

$$\begin{aligned} p_s(\alpha, \beta, \mu, \phi|Y) &= \textstyle\int p(\mu|\beta, \phi, Y) p(\alpha, \beta, \mu', \phi|Y) \mathrm{d}\mu' \\ &= p(\beta, \mu, \phi, Y) p(\alpha|\beta, \phi, Y), \end{aligned} \tag{4.19}$$

which breaks the correlation between α and μ conditioning on (β, ϕ). Then Sampler 4.13 can be considered as a sampler using the conditional of the surrogate distribution $p_s(\alpha, \beta, \mu, \phi|Y)$ in Step 1. We replace Step 3 of Sampler 4.13 with a draw from the conditional of the surrogate distribution, i.e., $p_s(\alpha|\beta, \mu, \phi, Y) = p(\alpha|\beta, \phi, Y)$, and obtain Sampler 4.14 in the second panel of Figure 4.8. Steps 1 and 2 of Samplers 4.13 and 4.14 also require MH. Thus both Samplers 4.13 and 4.14 use the surrogate distribution in (4.19), with Sampler 4.14 to a higher degree. Note that the stationary distribution of Sampler 4.14 is $p_s(\alpha, \beta, \mu, \phi|Y)$, not $p(\alpha, \beta, \mu, \phi|Y)$. In fact, both samplers use conditionals from two surrogate distributions, because $p(\beta, \phi|\mu, Y)$ is the conditional distribution of $p(\beta, \phi, \mu|Y)p(\alpha|\mu, Y)$, which is a surrogate distribution different from $p_s(\alpha, \beta, \mu, \phi|Y)$.

We use a simulation study to compare Samplers 4.13 and 4.14. We specify $n = 100$, $E_1 = 0.5$ and the bin width as 0.03, $\alpha = 30$, $\beta = 1$, $K = 2$, $\gamma_1 = 1$, $\gamma_2 = 0.3$, $\mu_1 = 50$,

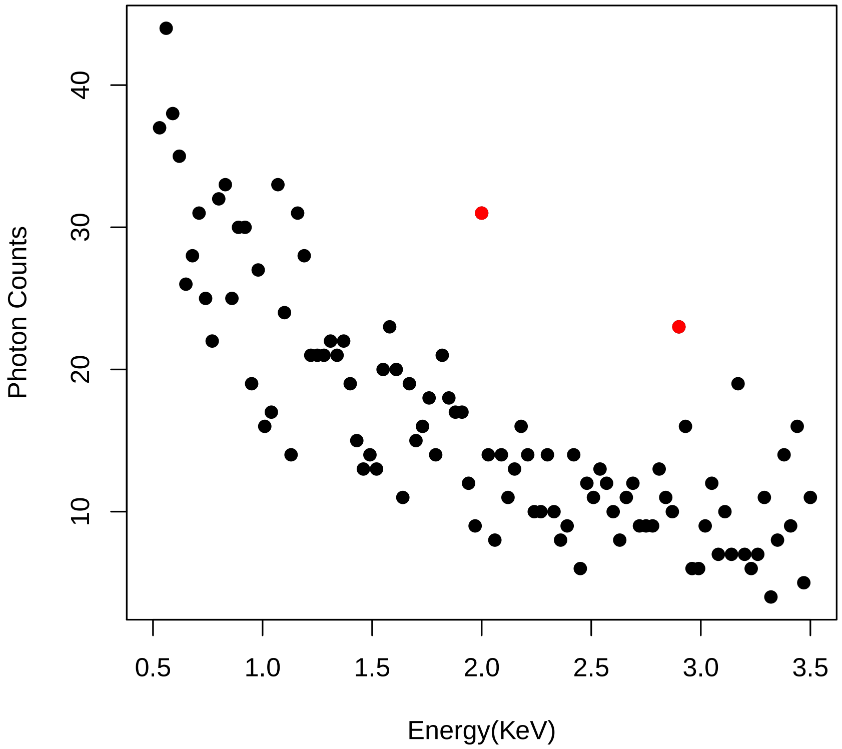

Figure 4.9: A dataset simulated under the spectral model in (4.18) and used in the simulation study for comparing Samplers 4.13 and 4.14.

$\mu_2 = 80$, and $\phi = 0.2$, see Figure 4.9. For either of the two samplers, we run one chain of 35,000 with the same starting values. Figure 4.10 compares Samplers 4.13 and 4.14 in terms of the mixing and autocorrelation of α, β, and ϕ. Sampler 4.14 performs better than Sampler 4.13 in convergence, which is a signal that using the surrogate conditionals in more steps of a sampler leads to better convergence properties. However, one plight we encounter in this example is that Step 3 of Sampler 4.14 (updating a mixture of n^K Gamma distributions, see Appendix C) is computationally demanding, which diminishes the power of this example for verifying the efficiency of using surrogate distributions to higher degrees.

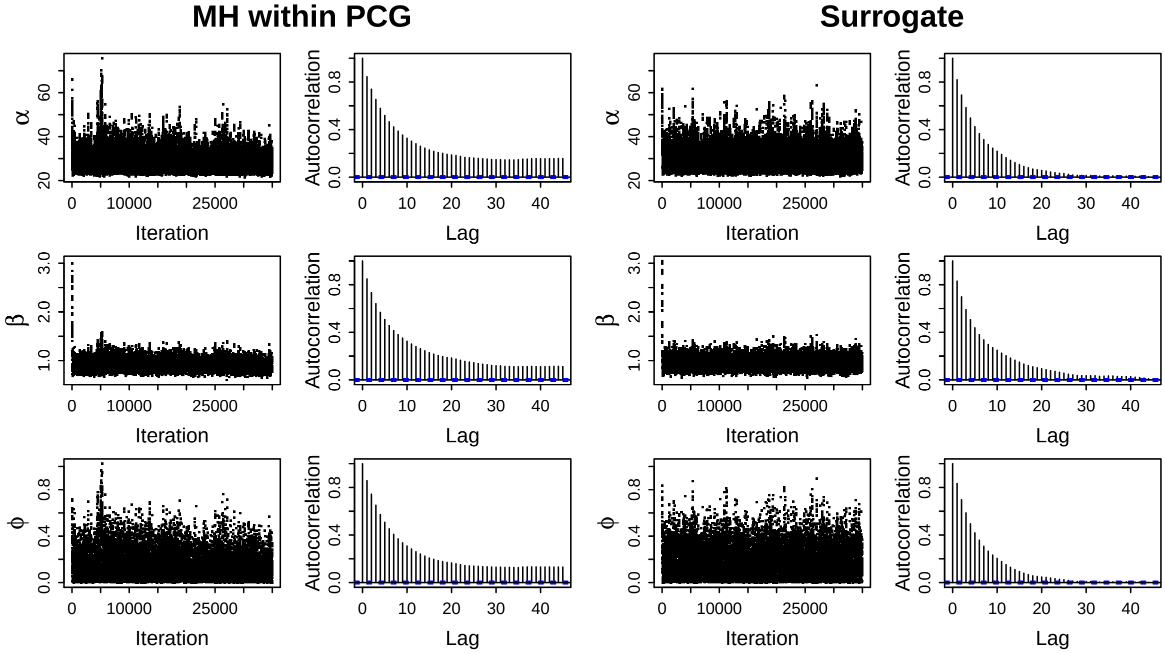

Figure 4.10: The sampling results of Samplers 4.13 and 4.14. The left two columns are the time-series and autocorrelation plots for the posterior draws of α, β, and ϕ respectively from Sampler 4.13, whereas the right two columns are those from Sampler 4.14. Sampler 4.14 performs better than Sampler 4.13.

4.3.3 Hierarchical Gaussian Model

Finally, we use a simple Gaussian hierarchical model to demonstrate deriving surrogate distributions from ASIS. This example is not convincing enough to prove the computational advantage of samplers using the surrogate distribution to a higher degree. However, it sheds light on the possible equivalence of PCG, MDA, and ASIS under the framework of surrogate distributions.

Suppose there is a single observation Y, and

$$Y \sim \mathrm{N}(X, 1) \text{ and } X \sim \mathrm{N}(\psi, V), \tag{4.20}$$

where V is the known latent variance; X and ψ are unknown parameters. We specify the non-informative prior distribution to ψ, that is, $p(\psi) \propto 1$.

DA (Sampler 4.15)	**ASIS (Sampler 4.16)**	**Surrogate (Sampler 4.17)**
1. $p(X \mid \psi', Y)$	1. $p(X^{\star} \mid \psi', Y)$	1. $p_s(\psi \mid X', Y)$
2. $p(\psi \mid X, Y)$	2. $p(\psi^{\star} \mid X^{\star}, Y)$; set $\bar{X} = X^{\star} - \psi^{\star}$	2. $p_s(X \mid \psi, Y)$
	3. $p(\psi \mid \bar{X}, Y)$; set $X = \bar{X} + \psi$	

Figure 4.11: Three samplers for fitting the simple hierarchical Gaussian model in (4.20). The left, middle, and right panels show the steps of the DA sampler (Sampler 4.15), the ASIS sampler (Sampler 4.16), and the two-step Gibbs sampler for updating the surrogate distribution derived from the ASIS algorithm (Sampler 4.17).

We consider three samplers for fitting the hierarchical Gaussian model. The first is the standard DA sampler which updates X and ψ iteratively from their conditional distributions of $p(\psi, X|Y)$, see Sampler 4.15 in the left panel of Figure 4.11. To improve the convergence of Sampler 4.15, we then consider using ASIS. The latent X is the sufficient augmentation for ψ and $\bar{X} = X - \psi$ is the corresponding ancillary augmentation. We construct the ASIS sampler, i.e., Sampler 4.16, based on this pair of augmentation schemes, see the middle panel of Figure 4.11. We derive a surrogate distribution via the method introduced in Section 4.2.1 from the ASIS sampler, that is,

$$\begin{aligned} p_s(\psi, X|Y) &= \left[\int \left(\int p(\psi^{\star}|X, Y) p(\bar{X}|\psi^{\star}, X, Y) \mathrm{d}\psi^{\star}\right) p(\psi|\bar{X}, Y) \mathrm{d}\tilde{X}\right] p(X|Y) \\ &= p(\psi|Y) p(X|Y), \end{aligned} \tag{4.21}$$

where sampling $\bar{X}$ from $p(\bar{X}|\psi, X, Y)$ is equivalent to the transformation at the end of Step 2 of the ASIS sampler. As stated in Section 4.2.1, the sub-chain of ψ induced by the ASIS sampler, i.e., Sampler 4.16, is equivalent to that induced by the sampler which updates from $p(X|\psi', Y)$ and $p_s(\psi|X, Y)$ iteratively. Thus the ASIS sampler can be considered as a two-step sampler which replaces one step of the parent DA sampler with a draw from the conditional of the surrogate distribution $p_s(\psi, X|Y)$. Since for this example, ψ is the parameter of main interest, we also consider the two-step Gibbs sampler for updating $p_s(\psi, X|Y)$, that is, Sampler 4.17 in the right panel

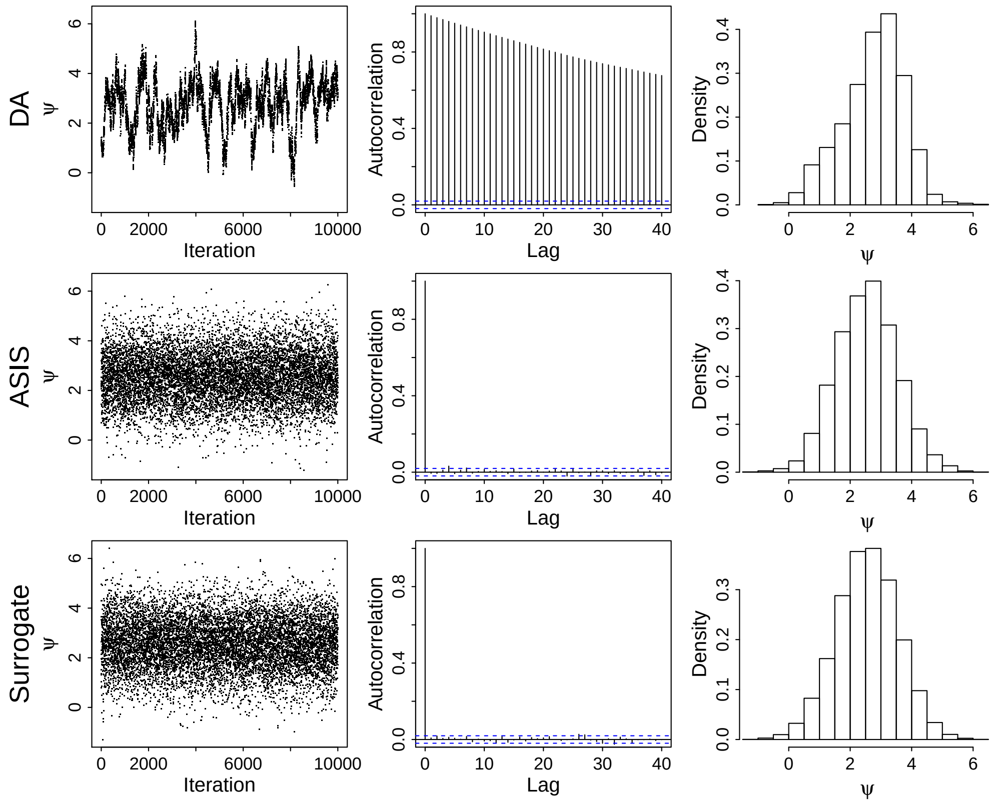

Figure 4.12: The sampling results of ψ for Samplers 4.15–4.17. The first, second, and third columns correspond to the trace plots, autocorrelation plots, and histograms of ψ. Both samplers using the surrogate distribution, i.e., Samplers 4.16 and 4.17, are efficient in improving the convergence of ψ, and the two samplers perform similarly.

of Figure 4.11. The construction of Sampler 4.17 can be considered as replacing both steps of the parent DA sampler with draws from the conditionals of the surrogate distribution. Henceforth, Samplers 4.15–4.17 are another example for Samplers 4.1–4.3, the three two-step samplers using the surrogate distribution to different degrees.

We use a simulation study to compare Samplers 4.15–4.17. We set $V = 0.01$ and $\psi = 1$. For each sampler, we run a chain of 10,000 with the same initial values. Figure 4.12 presents the mixing, autocorrelation, and posterior density estimation of ψ for Samplers 4.15–4.17. Both samplers using the surrogate distribution, i.e., Samplers 4.16 and 4.17 are efficient in improving the convergence of ψ, and the two samplers perform similarly, although they use the surrogate distribution to different degrees.

One interesting result of this example is that, for this simple hierarchical Gaussian model, the surrogate distributions derived from three different samplers are all the same. The three samplers are the ASIS sampler, i.e., Sampler 4.16, the PCG sampler

1. $p(\psi|Y)$
2. $p(X|\psi, Y)$,

and the Haar PX-DA sampler, where the working parameter α is introduced into the model via $\tilde{X} = X - \alpha$ and the working prior is the Haar measure prior $p(\alpha) \propto 1$. Thus the surrogate distribution strategy unveils the possible equivalence of the PCG, Haar PX-DA, and ASIS algorithms, and we unify different acceleration strategies under the general framework of surrogate distributions.

4.4 Conclusion

With surrogate distributions, we are able to unify different acceleration algorithms and also the combining strategy introduced in Chapter 3 under one general framework, because all the algorithms can be constructed by replacing some steps of the parent Gibbs sampler with draws from the conditionals of surrogate distributions. We use both theoretical arguments and numerical examples to illustrate the flexibility and strength of the surrogate distribution strategy in improving convergence properties of Gibbs-type samplers. Especially, for an N-step($N \geq 3$) sampler, there typically exist numerous choices of using surrogate distributions. However, the sampler with the optimal efficiency often need to be selected on a case-by-case basis.

5

Applied Work in Supernova Cosmology

Recall the Gaussian hierarchical model (2.14)–(2.16) in Section 2.3.2 of Chapter 2, which is designed for analysing supernova cosmology. In this section, we discuss the advantage of the hierarchical structure in this model, and introduce several extensions of the model that are of physical interests. We begin in Section 5.1 with describing the shrinkage effects of the Gaussian hierarchical model. In Section 5.2, we explore the influence of adding systematical errors onto the variance-covariance matrix in the first level of the model, i.e., (2.14). In Section 5.3, we specify the residual distance modulus as a cubic spline function to check the propriety of the hierarchical model. Then we investigate two generalizations of the Phillips corrections to further reduce the residual scatter around the Hubble diagram. In Section 5.5, we let the color correction parameter β vary with redshift and In Section 5.4, we add the host galaxy mass as a covariate in the regression model. Final discussion appears in Section 5.6.

5.1 Shrinkage Effects

We illustrate the shrinkage effects of the hierarchical model by plotting the posterior expectation and standard deviation of the absolute magnitude M_i conditioning on the other unknown parameters (but integrating out c_i and x_i), as the function of the population standard deviation σ_{res}, for thirty Type Ia SNe, in Figures 5.1 and 5.2. The thirty SNe are selected for having large difference between the likelihood and posterior estimates of M_i relative to the other SNe. As the population standard deviation σ_{res} gets close to zero, the conditional expectations of the absolute magnitudes for different SNe are shunk to the population mean M_0, and the conditional standard deviations are shrunk to zero. The larger σ_{res} becomes, the more variation appears in the population of absolute magnitudes. The shaded areas in the plots correspond to the 95% posterior credible interval of σ_{res}. The posterior estimate of σ_{res} under the hierarchical model is so small that the conditional expectations of the absolute magnitudes for different SNe are similar and the conditional standard deviations are small. In fact, the hierarchical model accounts for the uncertainty in the observed values of c_i, x_i, and m_{Bi}, and adjusts their estimates (i.e., their posterior distributions) by "shrinking" them towards their estimated population means and the fitted regression line. By "borrowing strength" across the SNe, the hierarchical regression structure of the model reduces the residual scatter around the regression plane. Recall that Type Ia SNe are supposed to be "standardizable candles" with similar absolute magnitudes (see Chapter 1 and Section 2.3.2 of Chapter 2). The hierarchical model reflects this assumption properly with shrinkage.

In the remainder of the chapter, we consider the sampling results of applying the hierarchical model to another data set, which consists of 740 spectroscopically confirmed Type Ia SNe obtained by the SDSS-II and SNLS collaboration from the "Joint Light-curve Analysis" (JLA) (Betoule *et al.*, 2014). This data set contains Type Ia SNe observed with SNLS (Canada-France-Hawaii Telescope), Hubble Space Telescope (HST), SDSS, and several other telescopes for low-z SNe. For the JLA data set, we specify an

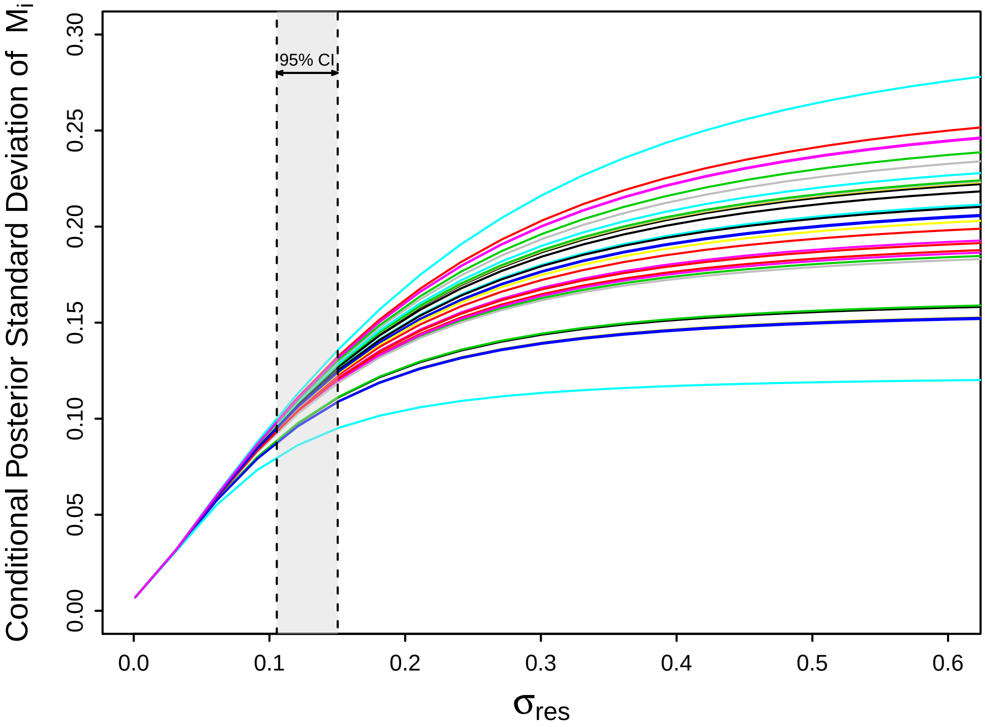

Figure 5.1: The posterior standard deviation of the absolute magnitude M_i conditioning on the other unknown parameters (but integrating out c_i and x_i), as the function of the population standard deviation $\sigma_{\rm res}$, for thirty Type Ia SNe. The thirty SNe are selected for having larger difference between the likelihood and posterior estimates of M_i. The shaded area corresponds to the 95% posterior credible interval of $\sigma_{\rm res}$.

inverse-Gamma prior distribution to $\sigma^2_{\rm res}$, that is, $\sigma^2_{\rm res} \sim \text{Inv-Gamma}(u, v)$, instead of a log-uniform prior, which might lead to difficulties in interpreting the posterior distribution. We perform a sensitivity analysis for the choice of scale for this inverse-Gamma distribution, to test the robustness of our posterior inference on $\sigma^2_{\rm res}$ with respect to its prior specification. Specifically, we compare the posterior distributions of $\sigma^2_{\rm res}$ obtained under three different inverse-Gamma priors, with parameters $u = v = 0.003$, 0.03, and 0.1 respectively. The resulting posterior distributions (along with the priors) are shown in Figure 5.3. Despite the widely differing priors, the posterior distributions are nearly identical, verifying the prior-independence of our results. Posterior distributions of all the other parameters are similarly insensitive to the choice of prior for $\sigma^2_{\rm res}$. In the rest of this section, we use $\sigma^2_{\rm res} \sim \text{Inv-Gamma}(0.003, 0.003)$.

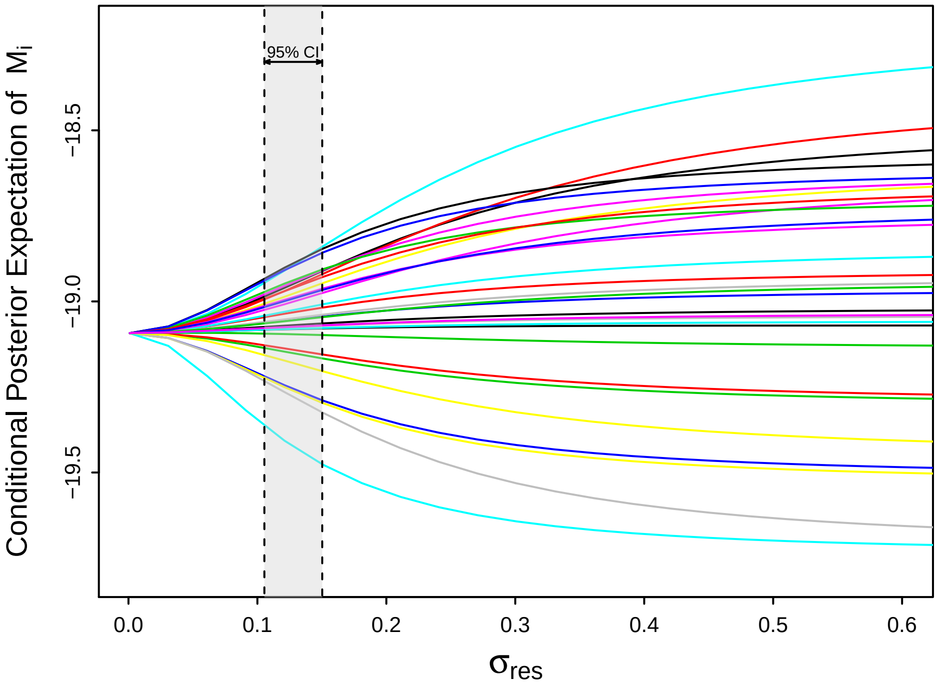

Figure 5.2: the posterior expectation of the absolute magnitude M_i conditioning on the other unknown parameters (but integrating out c_i and x_i), as the function of the population standard deviation $\sigma_{\rm res}$, for thirty Type Ia SNe. The thirty SNe are selected for having larger difference between the likelihood and posterior estimates of M_i. The shaded area in the plots corresponds to the 95% posterior credible interval of $\sigma_{\rm res}$.

5.2 Influence of the Systematical Covariance Matrix

In Section 2.3.2 of Chapter 2, we assume that the observations of c, x, and m_B for each SN are conditionally independent (given their means and variances, see (2.14)), i.e., the $(3n \times 3n)$ variance-covariance matrix $\Sigma_C = C_{\rm stat} \equiv {\rm Diag}(\hat{C}_1, \ldots, \hat{C}_n)$ is block diagonal. Betoule *et al.* (2014) derived a systematical variance-covariance matrix, $C_{\rm syst}$, with correlations among the SNe, see Figure 5.4 for the systematical correlation matrix, where the blocks correspond to different surveys, that is, SNLS, HST, SDSS, and low-z. The systematical covariance matrix includes contributions from calibration, model uncertainty, bias correction, host, dust, peculiar velocities, and contamination. We account for these systematical error by replacing the matrix $\Sigma_C = C_{\rm stat}$ with $\Sigma_C = C_{\rm stat} + C_{\rm syst}$ in the full posterior distribution.

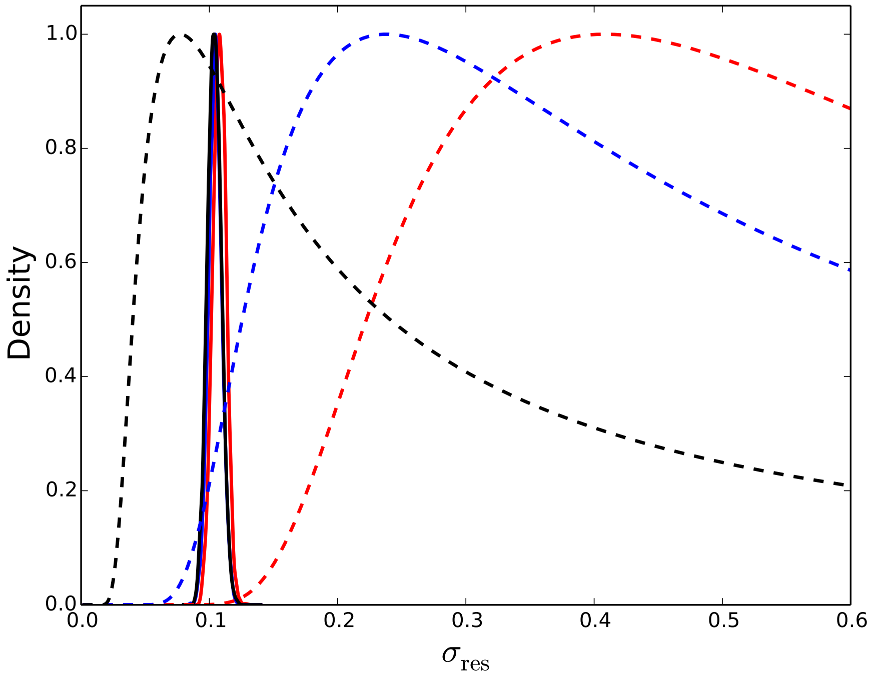

Figure 5.3: Robustness of the posterior distribution for σ^2_{res} (solid lines) with respect to three different prior specifications (dashed lines). The black, blue, and red correspond to the prior $\sigma^2_{\text{res}} \sim$ Inv-Gamma$(0.003, 0.003)$, $\sigma^2_{\text{res}} \sim$ Inv-Gamma$(0.03, 0.03)$, and $\sigma^2_{\text{res}} \sim$ Inv-Gamma$(0.1, 0.1)$. Since the three posterior distributions are similar, we conclude that the posterior distribution of σ_{res} is largely insensitive to its prior specification.

To assess the relative importance of the statistical and systematical variance-covariance matrices, we use an MH within PCG sampler similar to Sampler 2.8 in Section 2.3.2 of Chapter 2 to fit the hierarchical models with $\Sigma_C = C_{\text{stat}}$ and $\Sigma_C = C_{\text{stat}} + C_{\text{syst}}$ respectively. The only difference between this MH within PCG sampler and Sampler 2.8 rests in the sampling of σ^2_{res}, since we change the prior of σ^2_{res} for the JLA data set. (See Appendix D for details.)

Figure 5.5 shows the 68% and 95% contours of the joint posterior distributions of $(\Omega_m, \Omega_\Lambda)$, where blue represents including only statistical errors in the Level 1 variance-covariance matrix, and black including both statistical and systematical errors. Adding the systematical covariance matrix not only enlarges the size of the contours—as expected—but also significantly shifts the mean value of the posterior distribution of Ω_m to larger values, which leads to a smaller Ω_Λ. Thus we conclude that the shift in cosmology is

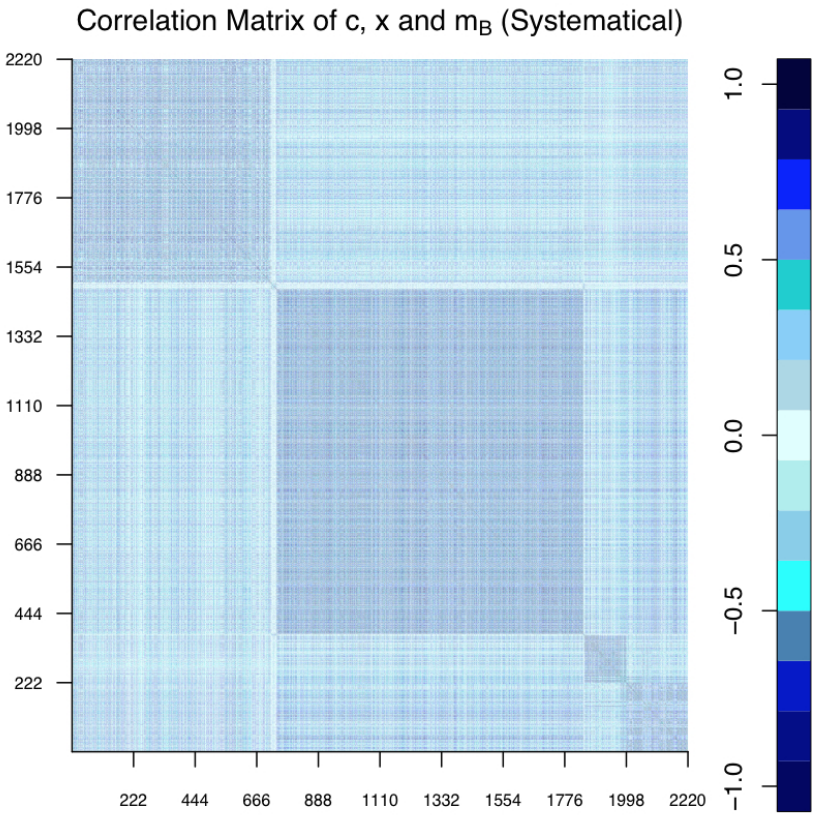

Figure 5.4: the systematical correlation matrix for the JLA data set. The blocks correspond to different surveys, that is, SNLS, HST, SDSS, and low-z.

driven by some aspect of the systematical errors in JLA.

As stated above, the systematical covariance matrix contains contributions from different sources. Analysing each source individually, we conclude that the main driver shifting the posterior estimate of Ω_m to larger values is the calibration uncertainty.

In order to further investigate the origin of the observed shift in the fitted cosmological parameters obtained by add the systematical errors, we compute the percent increase in the standard deviation of $\hat{c}$, $\hat{x}$, and $\hat{m}_B$ when adding the systematical covariance matrix onto the statistical covariance matrix, i.e.,

$$F_y = \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma_{y,i}^{2,\mathrm{syst}}}{\sigma_{y,i}^{2,\mathrm{stat}}} \tag{5.1}$$

where $y = \hat{c}$, $\hat{x}$, or $\hat{m}_B$. The quantity $F_y^{1/2}$ is the average percent increase in the standard

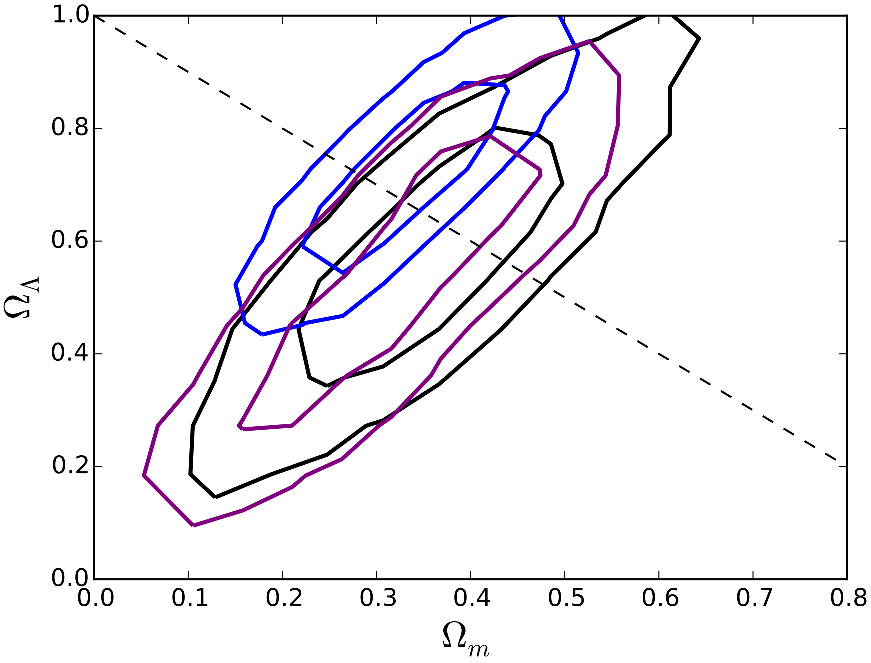

Figure 5.5: Comparing the 68% and 95% contours of the joint posterior distributions of $(\Omega_m, \Omega_\Lambda)$ for including only statistical errors in the Level 1 variance-covariance matrix (blue) to those for including both statistical and systematical errors (black). Purple contours correspond to statistics covariance matrix with diagonal errors on m_B inflated by the average m_B variance resulted from the systematical covariance matrix.

deviation for SN i when the systematics covariance matrix is added to the statistical covariance matrix (considering diagonal elements only). We find $F_{\hat{c}}^{1/2} = 0.36$, $F_{\hat{x}}^{1/2} = 0.16$, and $F_{\hat{m}_B}^{1/2} = 2.66$, which shows that the increased error on $\hat{m}_B$ is by far the dominant contribution from the systematical covariance matrix. This is because the dominant source of systematical error in the JLA data is the flux calibration (Betoule *et al.*, 2014). To check whether the increase in the $\hat{m}_B$ variance is responsible for the shift in the cosmological parameter estimates, we multiply the variance of $\hat{m}_B$ in the statistical covariance matrix by $(1 + F_{\hat{m}_B})$, and refit (without adding the systematical covariance matrix) the Gaussian hierarchical model. The resulting cosmological constraints are shown as purple contours in Figure 5.5. Comparing with the results obtained from including both statistical and systematical errors in the variance-covariance matrix (black contours), it is clear that most of the shift in the fitted cosmological parameter is due to the large systematic variance of $\hat{m}_B$. If the model were Gaussian and linear, inflating the errors would only enlarge the uncertainty on the parameters, but would not shift the mean of the posterior distribution. Hence we conclude that the cosmology shift is a reflection of the non-Gaussian and non-linear nature of our model.

In the remainder of this chapter, we always set the observed variance-covariance matrix to $\Sigma_C = C_{\text{stat}} + C_{\text{syst}}$, and name the Gaussian hierarchical model (2.14)–(2.16) with $\Sigma_C = C_{\text{stat}} + C_{\text{syst}}$ by the "Baseline Model". The following models are all extensions of the Baseline Model.

5.3 Model checking

To check whether the Baseline Model reflects the underlying physical truth of supernova cosmology properly, we quantify the residual scatter around the Hubble diagram as,

$$t(z_i) = \tilde{\mu}_i - \mu_i, \tag{5.2}$$

where $t(z)$ is a function of the red shift z, $\tilde{\mu}_i$ is the distance modulus specified by the Baseline Model, i.e., $\tilde{\mu}_i = m_{Bi} - M_i + \alpha x_i - \beta c_i$, and μ_i is the theoretical distance modulus, i.e., $\mu_i = \mu_i(z_i, \mathscr{C})$ as displayed in (2.17). Here we specify $t(z_i)$ as a cubic spline function, that is,

$$t(z_i) = b_1 z_i + b_2 z_i^2 + b_3 z_i^3 + \sum_{k=1}^{K} b_{k+3}(z_i - z_k^0)_+^3, \tag{5.3}$$

where $(z_1^0, \ldots, z_K^0)$ are known knots and $b = (b_1, \ldots, b_{K+3})$ are unknown parameters. Then we replace the regression model in (2.15) with

$$m_{Bi} = \mu_i + M_i - \alpha x_i + \beta c_i + t(z_i), \text{ for } i = 1, \ldots, n, \tag{5.4}$$

specify the non-informative flat prior to b, that is, $p(b) \sim 1$, and obtain a new hierarchical model, which is named by the "Cubic Residual Model".

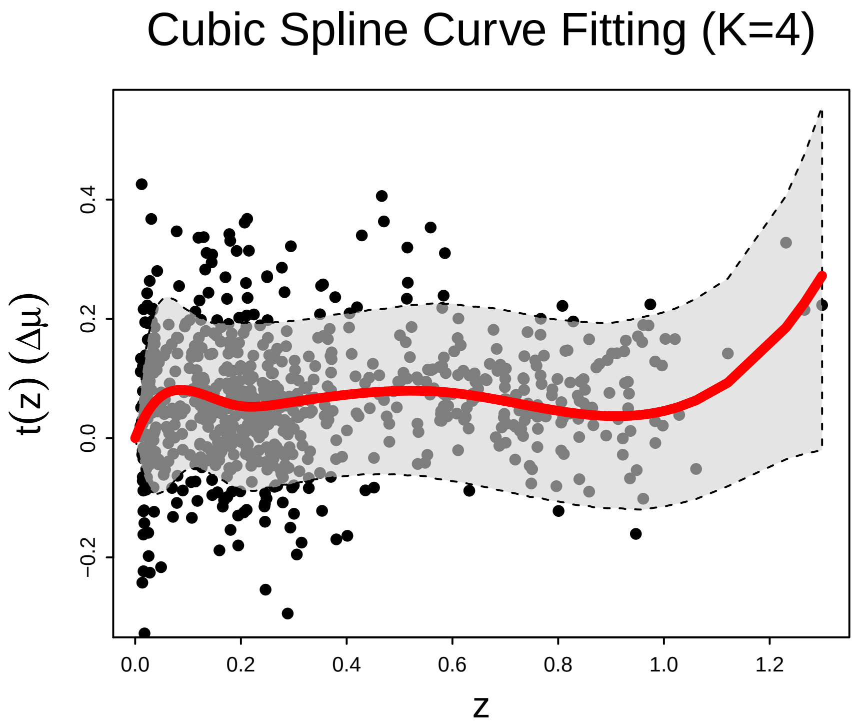

Figure 5.6: The posterior estimate of the cubic spline function $t(z)$ in (5.4). The red line is its posterior mean and the gray band is its point-wise 95% credible region. The black dots are estimates of the residual distance moduli based on the observations, that is, $\Delta\mu = \hat{m}_{Bi} - M_0 + \alpha\hat{x}_i - \beta\hat{c}_i - \mu_i$, where M_0, α, and β are imputed with their posterior means obtained from fitting the hierarchical model (2.14)–(2.16).

We fix the values of $\mathscr{C} = (\Omega_m, \Omega_\Lambda)$ at their posterior means obtained from fitting the Baseline Model, and then use an MH within PCG sampler to fit the Cubic Residual Model. Specifically, we sample (α, β) without conditioning on (X, ξ) and b. Details of this sampler and other samplers in this Chapter are given in Appendix D.

We plot the posterior estimate of the cubic spline function $t(z)$ in Figure 5.6. The red line represents its posterior mean and the gray band represents its point-wise 95% credible region. The black dots are estimates of the residual distance moduli based on the observations, that is,

$$\Delta\mu_i = \hat{m}_{Bi} - M_0 + \alpha\hat{x}_i - \beta\hat{c}_i - \mu_i, \tag{5.5}$$

where M_0, α, and β are replaced by their posterior means obtained from fitting the Baseline Model. Since the posterior estimate of the residual distance modulus $t(z)$

is close to zero, we conclude that the Baseline Model performs well in reflecting the underlying physical truth of supernova cosmology.

The Phillips corrections are expected to reduce the residual variance in (2.15) and thus increase the precision in the estimates of $\mathscr{C}$, see Section 2.3.2 of Chapter 2. Introducing additional covariates may further improve precision. Thus in the following two sections, we generalize the Phillips corrections by adding more covariates into the regression model.

5.4 Dependency on Host Galaxy Mass

There is strong evidence that the adjusted absolute magnitude of Type Ia SN correlates with host galaxy mass (e.g., Sullivan *et al.*, 2006). Current results indicate that more massive galaxies ($\log_{10}(M/M_\odot) > 10$) host brighter SNe, with their average absolute magnitude being of order ~ 0.1mag smaller than that in less massive hosts (e.g., Sullivan *et al.*, 2010). This could be a reflection of dust, age, and/or metallicity of the progenitor systems (Childress *et al.*, 2013).

5.4.1 Models incorporating host galaxy mass

We investigate three models that incorporate host galaxy mass as a covariate in (2.15) and study how they affect the inference of $\mathscr{C}$. In particular, we consider the models that (i) divide the SNe into two populations using a hard host galaxy mass threshold ("Hard Classification Model"), (ii) divide the SNe into two populations using soft probabilistic classification ("Soft Classification Model"), and (iii) adjust for host galaxy mass as a covariate in the regression, analogously to the stretch and color corrections ("Covariate Adjustment Model"). Specifically, we model the observed host galaxy masses (on the $\log_{10}$ scale) as

$$\widehat{M}_{\mathrm{g}\,i} \overset{\text{ind}}{\sim} \mathrm{N}\left(M_{\mathrm{g}\,i}, \sigma^2_{\mathrm{g}\,i}\right), \text{ for } i = 1, \ldots, n, \tag{5.6}$$

where $M_{\mathrm{g}\,i}$ is the (true) host galaxy mass of SN i (in $\log_{10}$ solar mass) and $\sigma_{\mathrm{g}\,i}$ is the observed standard deviation of $\widehat{M}_{\mathrm{g}\,i}$.

In the "Hard Classification Model", we divide the SNe into two classes using the observed mass: high host galaxy mass class if $\widehat{M}_{\mathrm{g}\,i} \geq 10$ and low host galaxy mass class if $\widehat{M}_{\mathrm{g}\,i} < 10$. (Thus in this model, we ignore measurement errors in $\widehat{M}_{\mathrm{g}\,i}$.) We fix the host galaxy mass classification at 10^{10} solar masses, analogous to the location of the step function used for the host galaxy mass by Betoule *et al.* (2014). The two classes are allowed to have their own population-level values for the mean absolute SN magnitude and residual standard deviation, i.e., $(M_0^{\mathrm{hi}}, \sigma_{\mathrm{res}}^{\mathrm{hi}})$ for high mass hosts and $(M_0^{\mathrm{lo}}, \sigma_{\mathrm{res}}^{\mathrm{lo}})$ for low mass hosts. Common values are used for α and β (and of course for $\mathscr{C}$) for both classes. We do not assume a redshift dependency for the color correction parameter. The prior distributions we specify to the new population-level parameters are $M_0^{\mathrm{hi}} \sim \mathrm{N}(-19.3, 2^2)$, $M_0^{\mathrm{lo}} \sim \mathrm{N}(-19.3, 2^2)$, $\left(\sigma_{\mathrm{res}}^{\mathrm{hi}}\right)^2 \sim \text{Inv-Gamma}(0.003, 0.003)$, and $\left(\sigma_{\mathrm{res}}^{\mathrm{lo}}\right)^2 \sim \text{Inv-Gamma}(0.003, 0.003)$.

The "Soft Classification Model" is identical to the Hard Classification Model except that measurement errors in the observed masses are accounted for by probabilistically classifying each SNe; these errors can be quite significant. Specifically, we let Z_i be an indicator variable that equals one for an SN with high host galaxy mass and equals zero for an SN with low host galaxy mass, that is,

$$Z_i = \begin{cases} 0, & \text{if } M_{\mathrm{g}\,i} < 10 \\ 1, & \text{if } M_{\mathrm{g}\,i} \geq 10. \end{cases} \tag{5.7}$$

We treat $Z = (Z_1, \ldots, Z_n)$ as a vector of unknown latent variables that are estimated along with the other model parameters and latent variables via Bayesian model fitting. This requires specification of a prior distribution on each $M_{\mathrm{g}\,i}$. We choose a flat prior so that $M_{\mathrm{g}\,i} | \widehat{M}_{\mathrm{g}\,i} \overset{\mathrm{ind}}{\sim} \mathrm{N}(\widehat{M}_{\mathrm{g}\,i}, \sigma_{\mathrm{g}\,i}^2)$, see Appendix D for details.

The "Covariate Adjustment Model" introduces $M_{\mathrm{g}\,i}$ as a covariate in the regression model (2.15) rather than classifying the SNe by host galaxy masses. Specifically, we replace (2.15) with

$$m_{Bi} = \mu_i + M_i - \alpha x_i + \beta c_i + \gamma M_{\mathrm{g}\,i}, \tag{5.8}$$

expand the observed quantities $\hat{c}_i$, $\hat{x}_i$, and $\hat{m}_{Bi}$ in (2.14) to include the observed host galaxy masses in (5.6), and also expand the population model for the latent variables, c_i, x_i, and M_i, given in (2.16) to include host galaxy mass, that is,

$$M_{\mathrm{g}\,i} \overset{\mathrm{iid}}{\sim} \mathrm{N}(M_{\mathrm{g}\star}, R_{\mathrm{g}}^2), \tag{5.9}$$

where $M_{\mathrm{g}\star}$ and R_{g} are hyperparameters analogous e.g., to x_0 and R_x. The prior distributions we specify to the added parameters are $\gamma\mathrm{Unif}(-4, 4)$, $M_{\mathrm{g}\star} \sim \mathrm{N}(10, 100^2)$, and $\log(R_{\mathrm{g}}) \sim \mathrm{Unif}(-5, 2)$.

5.4.2 Results under the models including host galaxy mass

We fit all of the Hard Classification Soft Classification and Covariate Adjustment Models using MH within PCG samplers, which update $\mathscr{C}$ and (α, β) (for the Hard Classification and Soft Classification Models, or (α, β, γ) for the Covariate Adjustment Model) without conditioning on (X, ξ). We compare the sampling results under the three models with those under the Baseline Model.

W detect significant difference (with 95% probability) between the mean absolute magnitudes of SNe in low-mass and high-mass classes. Specifically, we define

$$\Delta M_0 = M_0^{\mathrm{hi}} - M_0^{\mathrm{lo}} \tag{5.10}$$

as the difference of mean absolute magnitude between the two classes. The 95% equal-tail posterior credible interval for ΔM_0 is

$$-0.10 < \Delta \mathcal{F}_0 < 0.00, \tag{5.11}$$

with $\Delta M_0 = 0$ excluded. Figure 5.7 shows the posterior distribution of ΔM_0, where the

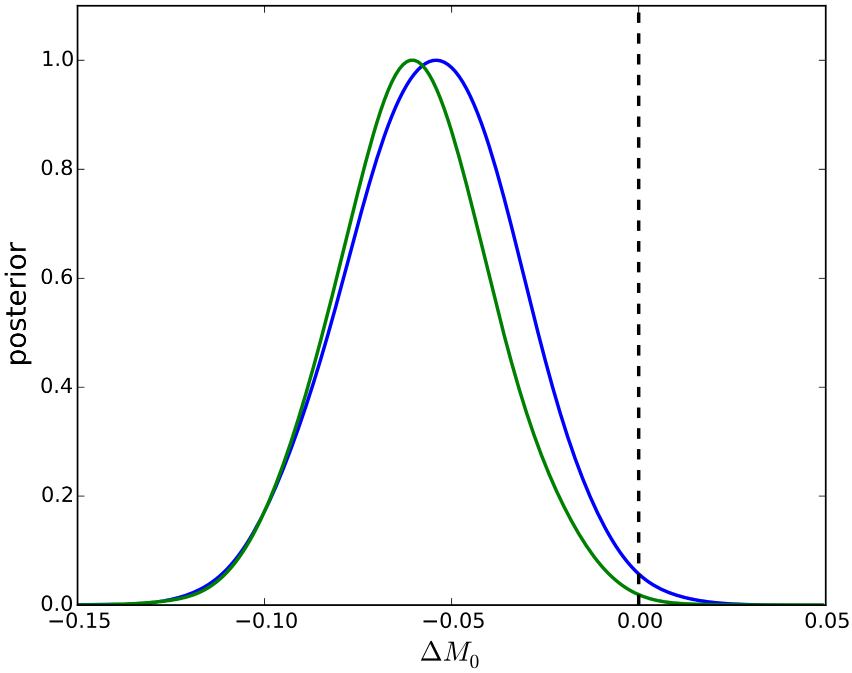

Figure 5.7: Posterior distributions of ΔM_0, the difference of mean absolute magnitudes for SNe in high-mass ($M_{g\,i} > 10$) and low-mass classes ($M_{g\,i} < 10$). Blue and green correspond to the Hard Classification and Soft Classification Models, respectively. Under both models, the posterior probability that $\Delta M_0 < 0$ is greater than 95%, implying that SNe in more massive hosts are most intrinsically brighter.

result for the Hard Classification Model is compared with that the for Soft Classification Model. There is not an appreciable difference in ΔM_0 between the Hard Classification Model and the Soft Classification Model. We find that SNe in more massive host galaxies are intrinsically brighter, with our posterior estimate of the magnitude difference under the as $\Delta M_0 = -0.055 \pm 0.022$. The SNe residing in more massive galaxies have smaller residual standard deviation, for $\sigma_{\rm res}^{\rm hi} = 0.097 \pm 0.007$ and $\sigma_{\rm res}^{\rm lo} = 0.110 \pm 0.009$.

Figure 5.8 shows the posterior estimates of the empirically corrected SNe's absolute magnitudes, M_i, as a function of the measured host galaxy mass. Histograms on either side of the plot show the posterior distributions of the mean absolute magnitudes for the two classes. The average measurement error of the host galaxy mass is fairly large, especially for low-mass host galaxies. Therefore, the SNe whose host galaxy masses are close to the cut-off, i.e., $M_{g\,i} = 10$, are of uncertain classification, once the measurement errors are taken into account. This could influence the estimate of ΔM_0 and the ensuing

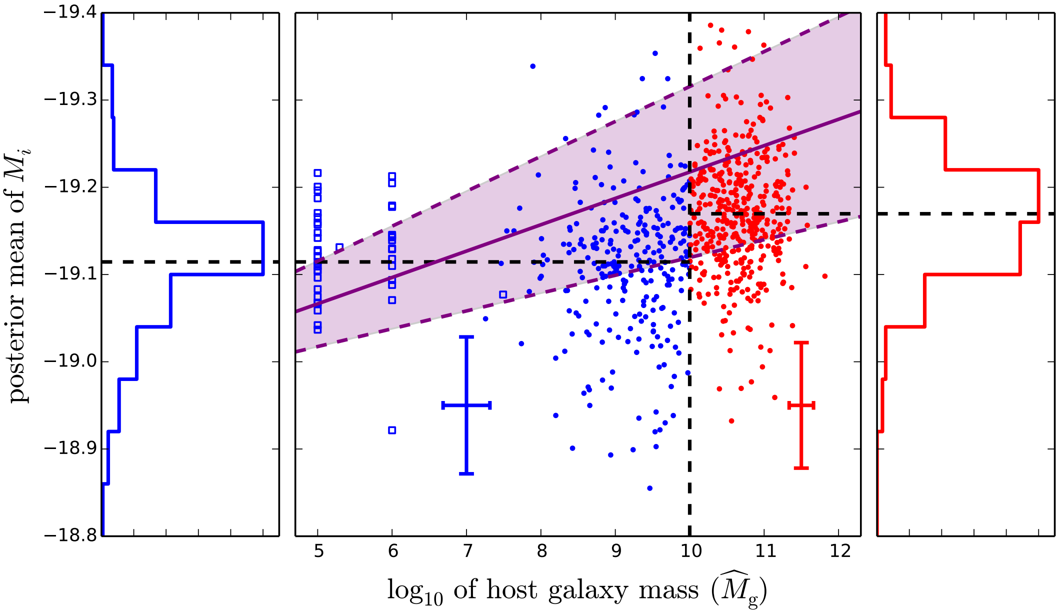

Figure 5.8: Posterior means and standard deviations for the empirically corrected absolute magnitudes versus measured host galaxy mass. The SNe are divided into two populations, with $M_{g\,i}$ smaller (larger) than 10, colored in blue (red). A hollow square represents the SN whose measurement error of $M_{g\,i}$ is equal to or larger than 5. The population means of the absolute magnitudes are $M_0^{\text{lo}} = -19.114 \pm 0.023$ and $M_0^{\text{hi}} = -19.169 \pm 0.022$ (horizontal dashed lines) respectively for the low-mass and high-mass classes. The blue and red vertical errorbars represent the average posterior standard deviations of the absolute magnitudes in the low-mass and high-mass classes, respectively. The horizontal errorbars represent the average measurement errors of $M_{g\,i}$ in the two classes. These average values exclude the SNe represented by hollow squares. The slope of the purple regression line is the posterior mean of γ under the Covariate Adjustment Model, and the purple shaded area represents the 1σ credible region for γ.

cosmological constraints.

To investigate the importance of the measurement errors in host galaxy masses, we fit the Soft Classification Model which includes an indicator variable Z_i for each SN; recall that Z_i is one if SN i belongs to the high-mass class and zero if it does not. Treating Z_i as an unknown variable allows us to compute the posterior probability that each SN belongs to the high-mass class. Figure 5.9 presents the posterior mean and standard deviation for each Z_i. The posterior mean of Z_i is the posterior probability that SN i belongs to the high-mass class. Although the measurement errors in host galaxy masses are suppressed for clarity in Figure 5.9, they are fully accounted by the Soft Classification Model.

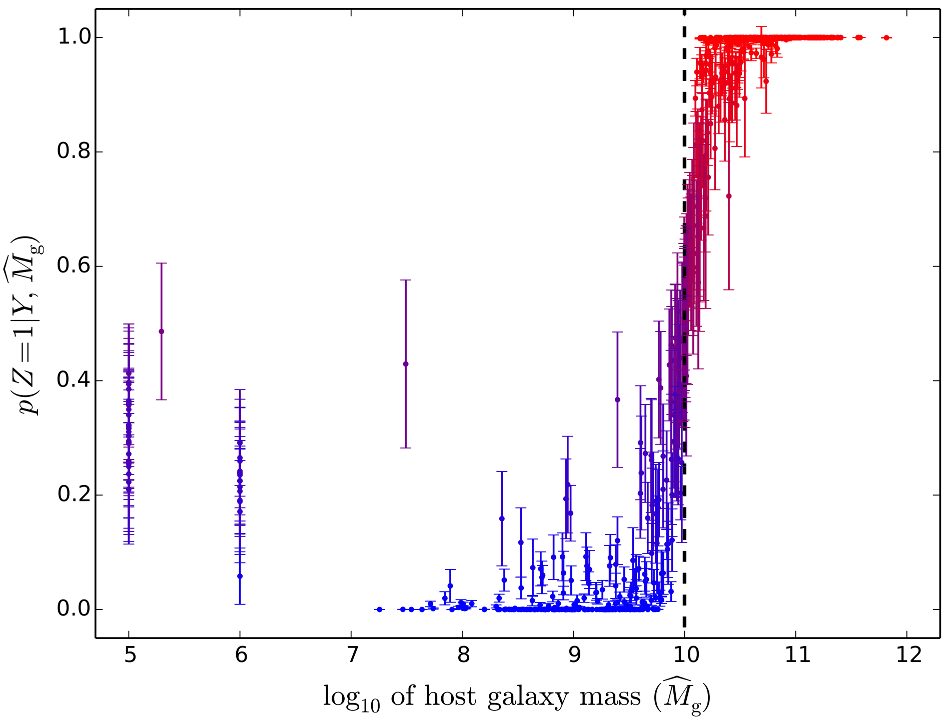

Figure 5.9: The posterior mean and standard deviation of each Z_i, the indicator variable for SN i belonging to the high-mass class ($Z_i = 1$, red) versus measured host galaxy mass. If $Z_i = 0$ (blue), SN i belongs to the low-mass class. The posterior mean of Z_i is the posterior probability that SN i belongs to the high-mass class. Although the errorbars are suppressed for clarity, the Soft Classification Model fully accounts for the measurement errors in host galaxy masses.

Under either the Hard Classification or Soft Classification Model, the posterior distributions of the cosmological parameters, $\mathscr{C}$, hardly change compared with the Baseline Model.

Finally, under the Covariate Adjustment Model, we express the fitted regression line with respect to $M_{g\,i}$ as $\hat{m}_{Bi} - \mu_i = \text{intercept} + \bar{\gamma}\ \widehat{M}_{g\,i}$, where $\bar{\gamma}$ is the posterior mean of γ and the intercept is $(M_0 - \alpha x + \beta c)$ with M_0, α, and β replaced by their posterior means, i.e., $\bar{M}_0$, $\bar{\alpha}$, and $\bar{\beta}$, x replaced by $\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i$, and c replaced by $\frac{1}{n}\sum_{i=1}^{n}\hat{c}_i$. The regression line is plotted as a solid purple line in Figure 5.8. The shaded purple area corresponds to the 68% posterior credible interval of γ (with the intercept fixed as described above). Figure 5.10 shows the posterior distribution of the regression coefficient γ. The posterior probability that $\gamma < 0$ is 99%. The 68% posterior credible interval for γ is -0.030 ± 0.010.

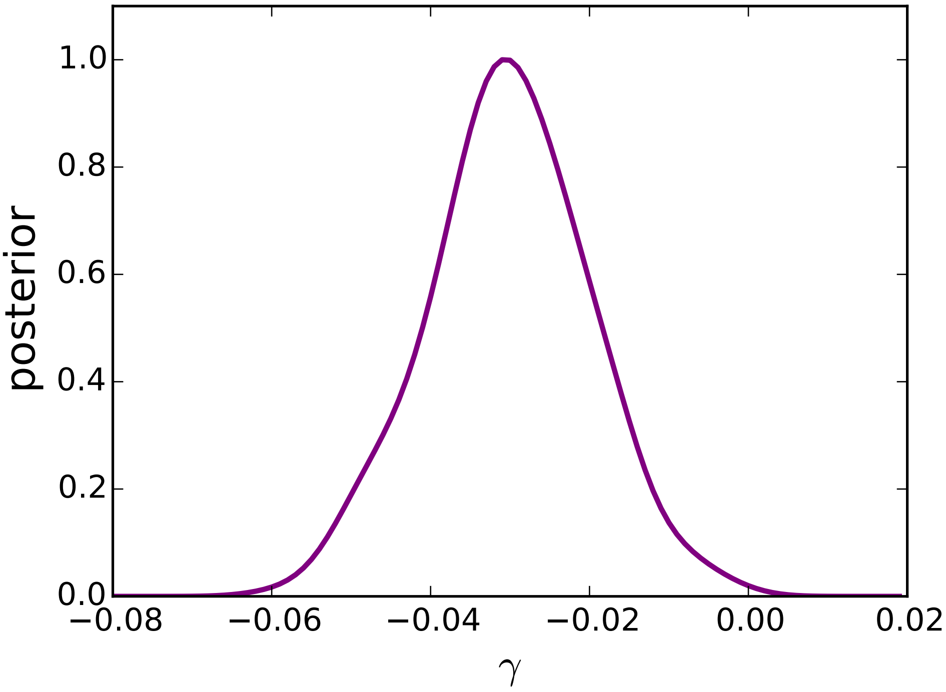

Figure 5.10: The marginal posterior distribution of γ, the regression coefficient for $M_{g\,i}$ in the Covariate Adjustment Model. The posterior probability that γ is less than zero is 99%.

Despite the fact that the posterior probability that $\gamma < 0$ is 99%, there is not a significant shift in the cosmological parameters, $\mathscr{C}$, or the residual standard deviation, $\sigma_{\rm res}$, compared with the Baseline Model. Although the intuition stemming from the standard linear regression suggests that adding a significant covariate should reduce residual variance, the situation is more complicated in (5.8) owing to the measurement errors in both the independent and dependent variables. While the variances of the left and right sides of (5.8) must be equal, there are numerous random quantities whose variances and covariances can be altered because of adding a covariate to the model.

5.5 Redshift Evolution of the Color Correction Parameter

It is possible that the color correction varies with redshift, as a consequence of evolution of the progenitor and/or changes in the environment, for example, variation in the dust composition with galactic evolution (Childress *et al.*, 2013). This is not captured by the

SALT-II fits, since they use a training sample that is distributed over a large redshift range ($0.002 \leq z \lesssim 1$) (Guy *et al.*, 2007) and thus the training sample color correction is averaged across redshift. It is therefore important to check for a redshift dependence in the color correction by allowing β, which controls the amplitude of the linear correction to the absolute magnitude, to vary with z.

5.5.1 Models with z-dependent color correction parameter

We consider two phenomenological models that allow the color correction to depend on z. In the first model, the dependence is linear. Specifically, we replace the constant β in (2.15) with the z-dependent $\beta_0 + \beta_1 \hat{z}_i$, which leads to

$$m_{Bi} = \mu_i + M_i - \alpha x_i + \beta_0 c_i + \beta_1 z_i c_i. \tag{5.12}$$

We specify uniform prior distributions to β_0 and β_1, that is, $\beta_0 \sim \text{Unif}(0, 4)$ and $\beta_1 \sim \text{Unif}(-4, 4)$. We refer to this model as the "z-Linear color Correction Model".

The second model allows for a sharp transition from a high-z to a low-z regime. Specifically, we replace the constant β in (2.15) with $\beta_0 + \Delta\beta \left[\frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{\hat{z}_i - z_t}{0.01}\right)\right]$, where β_0, $\Delta\beta$, and z_t are unknown parameters. This can be viewed as a smoothed step function which approaches β_0 as $z \to 0$ and approaches $\beta_0 + \Delta\beta$ as $z \to \infty$, with a smooth monotone local transition centered at $z = z_t$. Substituting this into (2.15), we obtain

$$m_{Bi} = \mu_i + M_i - \alpha x_i + \beta_0 c_i + \Delta\beta \left[\frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{z_i - z_t}{0.01}\right)\right] c_i, \tag{5.13}$$

where the covariate associated with $\Delta\beta$ depends nonlinearly on z_t. The priors we specify to β_0, $\Delta\beta$, and z_t are uniform distributions, that is, $\beta_0 \sim \text{Unif}(0, 4)$, $\Delta\beta \sim \text{Unif}(-1.5, 1.5)$, and $z_t \sim \text{Unif}(0.2, 1)$. We refer to this model as the "z-Jump color Correction Model".

5.5.2 Results under the models with z-dependent color correction

We fit both the z-Linear color Correction and z-Jump color Correction Models with MH within PCG + ASIS samplers. Under the z-Linear color Correction Model, we sample $\mathscr{C}$ without conditioning on (X, ξ), which requires the MH update, and sample $(\alpha, \beta_0, \beta_1)$ via ASIS conditioning on other parameters. Under the z-Jump color Correction Model, we sample z_t and $\mathscr{C}$ without conditioning on (X, ξ), which require the MH updates, and sample $(\alpha, \beta_0, \Delta\beta)$ via ASIS conditioning on other parameters. We compare the sampling results under these two models with those under the Baseline Model.

When evolution with the redshift is linear (as in the z-Linear color Correction Model), a non-zero, negative linear term β_1 is preferred with $\sim 95\%$ probability, $\beta_1 = -0.622 \pm 0.342$. Because the standard deviation of $\hat{c}_i$ is of order ~ 0.1, high-z SNe (at $z \sim 1$) are typically ~ 0.06 mag brighter than those nearby. When the evolution is a sharp transition with redshift (as in the z-Jump color Correction Model), there is strong evidence for a significant drop in β at $z_t = 0.66 \pm 0.06$. At z_t, β drops from its low-z value, $\beta_0 = 3.14 \pm 0.09$ by $\Delta\beta = -1.12 \pm 0.24$, with a nominal significance of approximately 4.6σ. This represents a correction of typically ~ 0.11mag for SNe at $z > z_t$. The posterior mean and 1σ credible region of the sharply transited $\beta(z)$ are shown in Figure 5.11.

Despite significant evidence for redshift evolution of the color correction parameter, the cosmological parameters estimated from the z-Linear color Correctionand z-Jump color Correction Models are only mildly affected compared with the Baseline Model. The posterior distribution of the residual intrinsic scatter also remains unchanged.

To quantify the residual scatter around the Hubble diagram, we use $\Delta\mu_i$ in (5.5) of Section 5.3, that is, the difference between the estimated distance modulus based on the observations and the theoretical distance modulus. The sample variance of $\Delta\mu_i$ is

$$\sigma^2_{\Delta\mu} = \frac{1}{n-1} \sum_{i=1}^{n} (\Delta\mu_i - \Delta\bar{\mu})^2, \tag{5.14}$$

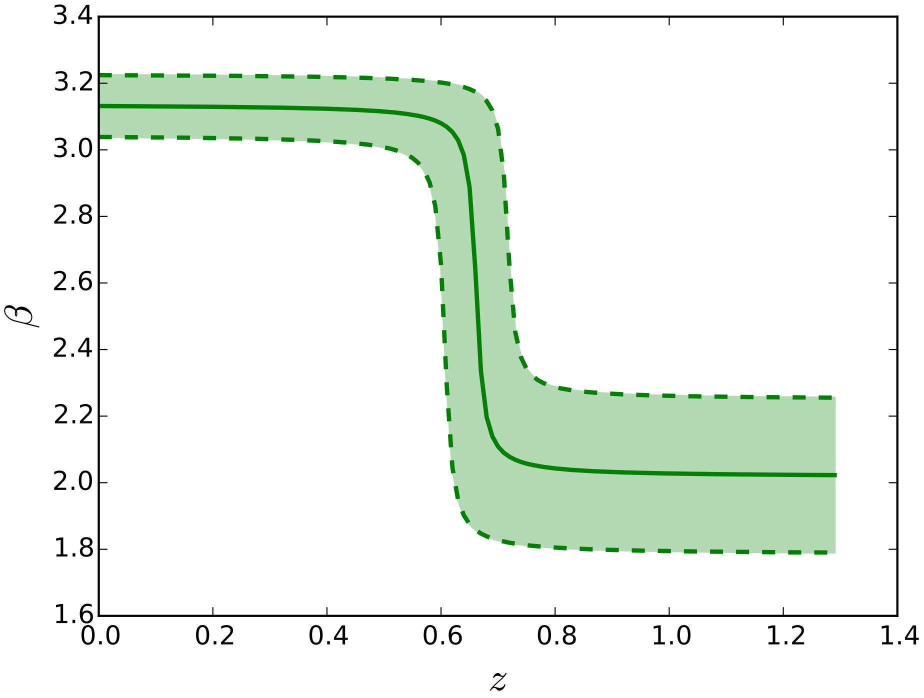

Figure 5.11: Redshift evolution of the color correction parameter β, assuming the z-Jump color Correction Model. The green line is the posterior mean, and the shaded area represents the 1σ credible region.

where $\Delta\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\Delta\mu_i$. Note that both $\Delta\mu_i$ and $\sigma^2_{\Delta\mu}$ are functions of the unknown parameters, and thus have posterior distributions.

We compare the Hubble diagram residuals, $\Delta\mu_i$, for the Baseline Model, with those for the z-Jump color Correction Model in Figure 5.12. The unknown parameters in $\Delta\mu_i$ are replaced with their posterior means. We only present SNe with $z > 0.6$, because the β values from the two models are similar for low-z SNe and thus the residual distance moduli are similar for $z > 0.6$. The left panel of Figure 5.12 shows the Hubble residuals under the Baseline Model; the central panel shows those under the z-Jump color Correction Model; the right panel compares the two models by plotting residuals under the Baseline Model versus those under the z-Jump color Correction Model. The scatter under the z-Jump color Correction Model is closer to zero than that under the Baseline Model. This indicates that allowing for a sharp transition in $\beta(z)$ improves the standardization of Type Ia SNe.

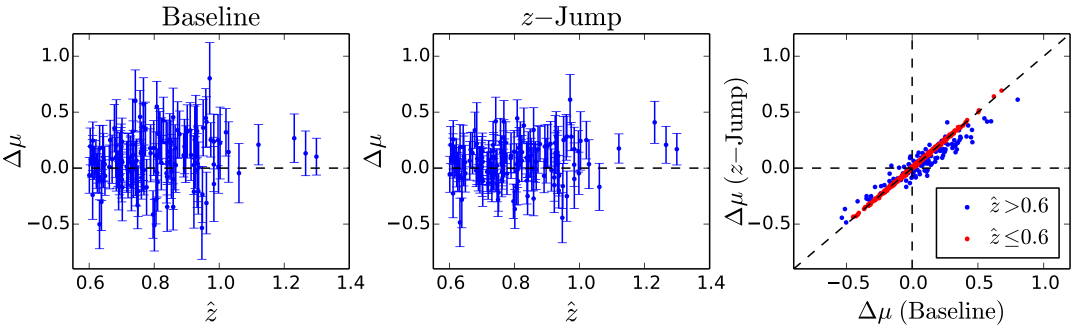

Figure 5.12: Hubble residuals of the Baseline Model (left, $\beta =$ constant with z), z-Jump color Correction model (centre), and comparison of the two models (right). In the left and central panels, only SNe with $z > 0.6$ are plotted to highlight the difference between the two models. Errorbars represent the posterior standard deviations of $\Delta\mu_i$. In the right panel, SNe with $z \leq 0.6$ are plotted in red. This panel shows that the z-Jump color Correction Model reduces the scatter around the Hubble diagram noticeably for $z > 0.6$, while its Hubble residuals are similar to the Baseline Model for $z \leq 0.6$.

We define the cumulative (i.e., summed over z) Hubble residual as

$$s_i = \sum_{z_j \leq z_i} |\Delta\mu_j| \quad (1 \leq i \leq n). \tag{5.15}$$

In Figure 5.13, we use the cumulative residual to highlight the difference between the fits under the Baseline, z-Linear color Correction and z-Jump color Correction Models. Figure 5.13 shows the cumulative residual as a function of redshift, where at each redshift the Baseline Model residual has been subtracted to facilitate comparison. For $z \lesssim 0.7$, the Baseline Model offers a slightly better fit than either of the $\beta(z)$ models. But for $z \geq 0.8$ both the z-Linear color Correction and especially the z-Jump color Correction Model provide improved residuals with respect to the Baseline Model. This is shown by the negative values of their relative cumulative residuals with respect to the Baseline Model. In other words, Figure 5.13 shows that both $\beta(z)$ models improve the fit for high-z SNe. Formal model comparison should be deployed to weigh the evidence for the evolving color correction models relative to the Baseline model.

It is conceivable that the evidence for a step in the evolution of $\beta(z)$ is a spurious consequence of the mass-step correction, which is not included in the above analysis. Since more massive ($M_{\mathrm{g}\,i} > 10$) host galaxies are preferentially found at low redshift,

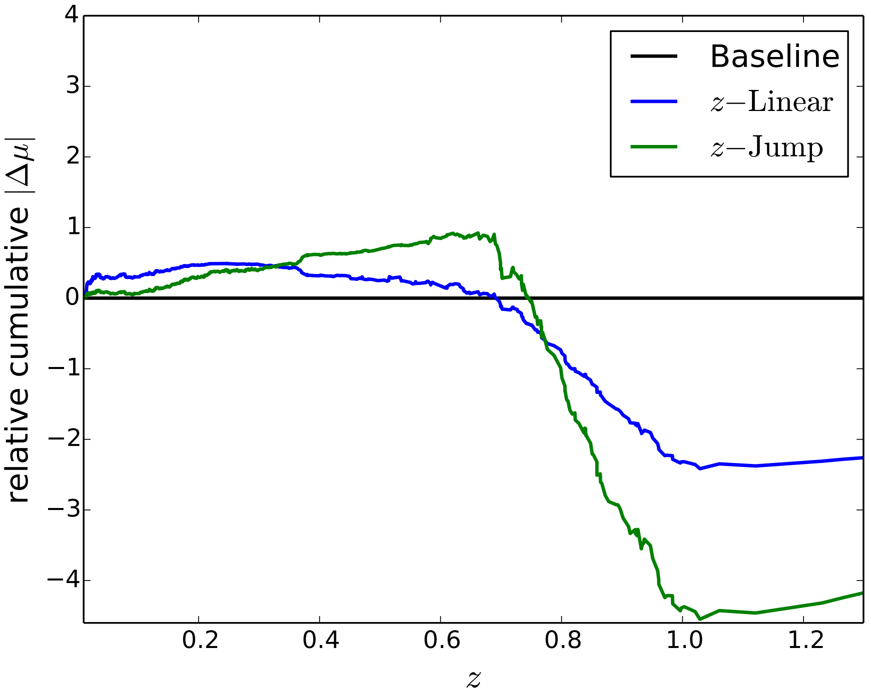

Figure 5.13: Cumulative Hubble residuals for the two $\beta(z)$ models relative to the Baseline Model. For $z \gtrsim 0.8$, both z-dependent models improve the fit with respect to the Baseline Model, which has $\beta = \text{constant}$. The z-Jump color Correction Model shows the largest improvement.

and SNe in those galaxies are brighter (see Section 5.4), it is possible that such galaxies require on average a smaller color correction than SNe in galaxies at high redshift (which are on average less luminous). However, if such a color-mass-redshift interaction were to exist, it could be identified by fitting a model that allows for both a host galaxy mass correction and evolution in the color correction. To investigate this possibility, we fit a model that includes both a mass-step correction (as in the Hard Classification Model) and a sharp transition in $\beta(z)$ (as in the z-Jump color Correction Model). The posterior distributions of all the model parameters from this fit change negligibly compared with those from the fit of the z-Jump color Correction Model without mass-step correction.

The top panel in Figure 11 of Betoule *et al.* (2014) suggested that un-modelled selection effects on the color correction at $z \gtrsim 0.6$ might lead to our detection of a drop in the value of $\beta(z)$ in that range. To test this possibility, we artificially correct the trend to negative colors (as shown in Figure 11 of Betoule *et al.* (2014)) for $z > z_t$, and re-fit the

z-Jump color Correction Model. This correction alters the posterior distributions of the cosmological parameters significantly, while leaving the strong detection of a jump in the value of $\beta(z)$ largely unchanged. This argues against the existence of un-modelled color correction selection effects causing the observed jump in $\beta(z)$ in the z-Jump color Correction Model. By the same token, it is unlikely that our result is driven by the redshift evolution of the color (or stretch) correction, as a consequence of selection effects.

In all of our models above, the population mean and variance of the color correction and stretch parameters are assumed to be redshift-independent. However, the observed color corrections drift towards the blue near the magnitude limit of a survey (i.e., to larger z). This happens because intrinsically brighter SNe (which are more likely to be observed) are bluer in color. This selection effect leads to a z-dependency of the observed color correction, even if the underlying color does not change with redshift. We allow the population mean and variance of the color corrections to differ for low-z ($z < 0.66$) and high-z ($z \geq 0.66$) SNe. (The threshold of $z = 0.66$ is chosen as the posterior mean of the jump location in the z-Jump color Correction Model.) The we re-fit both the Baseline Model and the z-Jump color Correction Model. The joint posterior distribution of $(\Omega_m, \Omega_\Lambda)$ shifts appreciably towards lower matter and lower cosmological constant values, while the evidence for a drop in β remains. This shows that the results of the hierarchical model are sensitive to the detailed modelling of a potential redshift-dependency (induced by selection effects, or otherwise) of the color correction parameter. However, the model for the redshift dependence of color is not what is driving the shift in the posterior distribution of Ω_m toward higher values.

5.6 Conclusion

In this section, we focus on the Gaussian hierarchical model in supernova cosmology and investigate several extensions of the model. We find that the hierarchical regression structure of the model reduces the residual scatter around the regression plane by "borrowing strength" across SNe, and this hierarchical model properly reflects the underlying physical assumptions. Moreover, for the JLA data set, the systematical errors can significantly shift the cosmological constraint estimates and the contribution from calibration is the main driver of the shift. To further reduce the residual scatter around the Hubble diagram, we generalize the Philip correction to explore the influence of host galaxy mass and redshift-dependent color correction parameter. We confirm (at the 95% probability level) the existence of two sub-populations segregated by host galaxy mass, at $\log_{10}(M/M_{\odot}) = 10$, differing in mean absolute magnitude by 0.055 ± 0.022mag. Cosmological parameter constraints are however unaffected by inclusion of host galaxy mass corrections. We also find $\sim 4\sigma$ evidence for a sharp drop in the value of the color correction parameter, $\beta(z)$, at a redshift $z_t = 0.662 \pm 0.055$. We rule out some possible explanations for this behaviour, which remains unexplained.

6

Corrections to Algorithms for Fitting the MNP Models in Imai and van Dyk (2005)

The multinomial probit (MNP) model is a useful tool for describing discrete-choice data in social sciences and transportation studies. There are a variety of methods for fitting the model, see Chapter 1. Among them, the algorithms provided by Imai and van Dyk (2005), based on Marginal Data Augmentation, are widely used, because they are efficient in convergence and allow the prior distribution to be specified directly on identifiable parameters. Burgette and Nordheim (2012) modified a model and algorithm of Imai and van Dyk (2005) to avoid an arbitrary choice that is often made to establish identifiability. However, there are errors in the algorithms of Imai and van Dyk (2005), which affect both their algorithms and that of Burgette and Nordheim (2012). These errors can alter the stationary distribution and the resulting fitted parameters as well as the efficiency of these algorithms. We propose corrections and use both a simulation study and real-data analyses to illustrate the difference between the original and corrected algorithms, both in terms of their estimated posterior distributions and their convergence properties. In some cases, the effect on the stationary distribution can be substantial.

6.1 Multinomial Probit Model

We consider a $(p+1)$-class multinomial model. Each observation is a binary $(p+1)$-vector, $d_i = (d_{i1}, \ldots, d_{i,(p+1)})$. We model d_i by conditioning on a latent multivariate normal variable, $U_i = \big(U_{i1}, \ldots, U_{i,(p+1)}\big)$; d_{ij} is one if U_{ij} is larger than all the other components of U_i. Specifically,

$$U_i \sim \mathrm{N}_{p+1}\left(X_i^0\beta, \Sigma^0\right) \text{ and } d_{ij} = \begin{cases} 1 & \text{if } U_{ij} = \max\{U_{i1}, \ldots, U_{i,(p+1)}\} \\ 0 & \text{otherwise} \end{cases}, \text{ for } i = 1, \ldots, n, \tag{6.1}$$

where X_i^0 is a $((p+1) \times q)$ matrix of known covariates, β is a q-vector of unknown parameters, and Σ^0 is a $((p+1) \times (p+1))$ unknown variance-covariance matrix.

Model (6.1) is unidentifiable because shifting U_i by any constant or rescaling U_i by any positive constant, does not alter the distribution of d_i. To avoid this, IvD and BN both followed McCulloch and Rossi (1994), by expressing each U_{ij} relative to a base category (e.g., $U_{i,(p+1)}$), and obtained the new latent variable, $W_i = (W_{i1}, \ldots, W_{ip})$, where $W_{ij} = U_{ij} - U_{i,(p+1)}$. The distribution of W_i is still multivariate normal, i.e.,

$$W_i \sim \mathrm{N}_p(X_i\beta, \Sigma), \tag{6.2}$$

where $X_i = PX_i^0$ and $\Sigma = P\Sigma^0P^{\mathrm{T}}$ with $P = [I_p, -J]$, with I_p a $(p \times p)$ identity matrix and J a column p-vector of ones. For simplicity, we collapse d_i into Y_i, which is an integer in $\{0, \ldots, p\}$, defined as

$$Y_i = \begin{cases} 0 & \text{if } \max\{W_{i1}, \dots, W_{ip}\} < 0 \\ k & \text{if } W_{ik} = \max\{0, W_{i1}, \dots, W_{ip}\} \end{cases}, \quad \text{for } i = 1, \dots, n. \tag{6.3}$$

To ensure identifiability, we must also set the scale. IvD adopted the standard solution of McCulloch and Rossi (1994); they set the first diagonal element of Σ to one, i.e., $\sigma_{11}^2 = 1$. BN proposed a different solution; they fixed the trace of the variance-covariance matrix, i.e., $\text{trace}(\Sigma) = p$. BN argued that the trace restriction is better because the resulting posterior distributions do not depend on the choice of category with unit variance and are easier to interpret.

To overcome difficulties stemming from the constraint, $\sigma_{11}^2 = 1$, on the variance-covariance matrix, motivated by McCulloch and Rossi (1994), IvD set $\tilde{W}_i = \alpha W_i$, for $i = 1, \dots, n$, where $\alpha > 0$. Then $\tilde{W}_i \sim \text{N}_p(X_i \tilde{\beta}, \tilde{\Sigma})$, where $\tilde{\beta} = \alpha\beta$ and $\tilde{\Sigma} = \alpha^2 \Sigma$. Because $\tilde{\Sigma}$ can be any positive-definite matrix, IvD specified an inverse-Wishart prior distribution, $\tilde{\Sigma} \sim \text{Inv-Wishart}(\nu, \tilde{S})$. After transforming to $\alpha^2 = \tilde{\sigma}_{11}^2$ and $\Sigma = \tilde{\Sigma}/\tilde{\sigma}_{11}^2$, the implied prior distribution on (α^2, Σ) is

$$\alpha^2 | \Sigma \sim \alpha_0^2 \text{trace}(S\Sigma^{-1})/\chi_{\nu p}^2, \text{ and } p(\Sigma) \propto |\Sigma|^{-(\nu+p+1)/2} [\text{trace}(S\Sigma^{-1})]^{-\nu p/2} I\{\sigma_{11}^2 = 1\}, \tag{6.4}$$

where α_0 is a positive constant, $S = \tilde{S}/\alpha_0^2$, and the first diagonal element of S is one; I is an indicator function which equals one when the condition in the brackets is satisfied, and zero otherwise. They also specified a normal prior distribution for β, $\beta \sim \text{N}_q(\beta_0, A)$. For simplicity, we set $\beta_0 = 0$. BN adopted the same strategy for setting their prior distribution in the context of the constraint, $\text{trace}(\Sigma) = p$. In particular, their implied prior distribution for (α^2, Σ) is almost the same as IvD except that

$$p(\Sigma) \propto |\Sigma|^{-(\nu+p+1)/2} [\text{trace}(S\Sigma^{-1})]^{-\nu p/2} I\{\text{trace}(\Sigma) = p\}, \tag{6.5}$$

where $\text{trace}(S) = p$. They use the same prior distribution as IvD for β, i.e., $\beta \sim \text{N}_q(0, A)$. As IvD stated, this choice of prior distribution allows both informative and diffuse priors for the unknown parameters while maintaining simplicity and efficiency of the algorithms.

6.2 Errors in Algorithms and the Corrections

We refer to Algorithms 1 and 2 of IvD as Algorithms IvD-1 and IvD-2, and to the algorithm of BN as Algorithm BN. We denote the corrected versions of these algorithms as Algorithms IvD-1c, IvD-2c, and BNc, respectively. Algorithm IvD-1c is displayed here and Algorithms IvD-2c and BNc in Appendix E. Except for the two boxed expressions, Algorithm IvD-1 is identical to Algorithm IvD-1c.[*]

To obtain posterior samples under the MNP model, Algorithm IvD-1 proceeds by sampling iteratively from $p(\alpha^2, \tilde{W}|Y, \beta, \Sigma)$, $p(\alpha^2, \beta|Y, \tilde{W}, \Sigma)$ and $p(\alpha^2, \Sigma|Y, \tilde{Z}, \beta)$, where $\tilde{Z} = \tilde{W} - \alpha X\beta$.[†] In the first of these draws, $\tilde{W}$ is obtained via a sequence of conditional draws; this is identical to Step 1(b) of Algorithm IvD-1c. Note that this algorithm marginalizes α out in each step. Algorithm IvD-2 proceeds by sampling iteratively from $p(\alpha^2, \tilde{W}|Y, \beta, \Sigma)$, $p(\alpha^2, \Sigma|Y, \tilde{Z}, \beta)$ and $p(\beta|Y, W, \Sigma)$, again using a sequence of conditional draws for updating $\tilde{W}$. Algorithm IvD-2 does not marginalize α out when sampling β. Algorithm BN is an adaptation of Algorithm IvD-1. The only difference occurs in Step 3 when sampling (α^2, Σ). In Algorithm IvD-1, α^2 is set to the first element of $\tilde{\Sigma}$ in Step 3(b), while it is set to trace$(\tilde{\Sigma})/p$ in Step 3(b) of Algorithm BN. Details of Algorithms IvD-2 and BN appear in Appendix E.

Unfortunately, there are two errors in these algorithms, which may severely alter their stationary distributions, fitted values, and convergence properties. In Algorithm IvD-1, both errors occur in Step 3. The corrections to these errors are the boxed expressions in Algorithm IvD-1c. Correction 1 is rather simple. The transformation from $(\tilde{Z}, \beta^{(t+1)}, \alpha^{(t+1)}, \tilde{\Sigma}^{\star})$ to the original variables $(W^{(t+1)}, \beta^{(t+1)}, \alpha^{(t+1)}, \Sigma^{(t+1)})$ should

[*] In Algorithm IvD-1, the constraint in the first box of Algorithm IvD-1c is ignored and the expression in the second box is replaced by $W^{(t+1)} = \tilde{W}^{\star}/\alpha^{(t+1)}$.

[†] In the original version of the paper, we denoted $(\tilde{W} - \alpha X\beta)$ by Z. However, there exist two transformed variables, one in the original model, i.e., $(W - X\beta)$, and the other in the expanded model, i.e., $(\tilde{W} - X\tilde{\beta})$. Thus here we denote $(W - X\beta)$ by Z and $(\tilde{W} - X\tilde{\beta})$ by $\tilde{Z}$, with $\tilde{Z} = \alpha Z$.

Algorithm IvD-1c (with corrections in boxes)

```
Step 0: Initialize parameters $t = 0$, $\beta^{(0)}$, $\alpha^{(0)}$, $\Sigma^{(0)}$, and $W^{(0)}$.
while $t < T$ do
    Step 1: Update $\left((\alpha^2)^\star, \tilde{W}^\star\right)$ via $p(\alpha^2, \tilde{W}|Y, \beta^{(t)}, \Sigma^{(t)})$ by
    (a) sampling $(\alpha^2)^\star$ from $p(\alpha^2|\Sigma^{(t)})$: $(\alpha^2)^\star \sim \alpha_0^2 \text{trace}\left(S{\Sigma^{(t)}}^{-1}\right)/\chi^2_{\nu p}$; setting $\alpha^\star = \sqrt{(\alpha^2)^\star}$;
    (b) sampling $\tilde{W}^\star$ from $p(\tilde{W}|Y, \alpha^\star, \beta^{(t)}, \Sigma^{(t)})$:
    for $i := 1, \ldots, n$ do
        for $k := 1, \ldots, p$ do
            sampling $W^\star_{ik}$ from $p(W_{ik}|Y_i, W^\star_{i,-k}, \beta^{(t)}, \Sigma^{(t)})$: $W^\star_{ik} \sim \text{TN}(\mu_{ik}, \tau^2_{ik})$, see Appendix E.3 for details;
        end for
        setting $\tilde{W}^\star_i = \alpha^\star W^\star_i$.
    end for
    Step 2: Update $\left((\alpha^2)^\star, \beta^{(t+1)}\right)$ via $p(\alpha^2, \beta|Y, \tilde{W}^\star, \Sigma^{(t)})$ by
      (a) sampling $(\alpha^2)^\star$ from $p(\alpha^2|Y, \tilde{W}^\star, \Sigma^{(t)})$:

          $$(\alpha^2)^\star \sim \frac{\sum_{i=1}^n (\tilde{W}^\star_i - X_i\hat{\beta})^{\mathrm{T}} {\Sigma^{(t)}}^{-1} (\tilde{W}^\star_i - X_i\hat{\beta}) + \hat{\beta}^{\mathrm{T}} A^{-1} \hat{\beta} + \text{trace}\left(\tilde{S}{\Sigma^{(t)}}^{-1}\right)}{\chi^2_{(n+\nu)p}},$$

    where $\hat{\beta} = \left(\sum_{i=1}^n X_i^{\mathrm{T}} {\Sigma^{(t)}}^{-1} X_i + A^{-1}\right)^{-1} \left(\sum_{i=1}^n X_i^{\mathrm{T}} {\Sigma^{(t)}}^{-1} \tilde{W}^\star_i\right)$;
      (b) sampling $\tilde{\beta}^\star$ from $p(\tilde{\beta}|Y, \tilde{W}^\star, (\alpha^2)^\star, \Sigma^{(t)})$:

          $$\tilde{\beta}^\star \sim \mathrm{N}_q \left[\hat{\beta}, (\alpha^2)^\star \left(\sum_{i=1}^n X_i^{\mathrm{T}} {\Sigma^{(t)}}^{-1} X_i + A^{-1}\right)^{-1}\right];$$

    setting $\alpha^\star = \sqrt{(\alpha^2)^\star}$ and $\beta^{(t+1)} = \tilde{\beta}^\star/\alpha^\star$.
    Step 3: Update $\left((\alpha^2)^{(t+1)}, \Sigma^{(t+1)}\right)$ via $p(\alpha^2, \Sigma|Y, \tilde{W}^\star, \beta^{(t+1)})$ by
      (a) sampling $\tilde{\Sigma}^\star$ from $p(\tilde{\Sigma}|Y, \tilde{Z}, \beta^{(t+1)})$:

          $$\tilde{\Sigma}^\star \sim \text{Inv-Wishart}\left(n + \nu, \sum_{i=1}^n \tilde{Z}_i \tilde{Z}_i^{\mathrm{T}} + \tilde{S}\right), \boxed{\text{subject to the constraint in (6.7)}},$$

    where $\tilde{Z}_i = \tilde{W}^\star_i - \alpha^\star X_i \beta^{(t+1)}$;
      (b) setting $\alpha^{(t+1)} = \tilde{\sigma}^\star_{11}$, $\Sigma^{(t+1)} = \tilde{\Sigma}^\star/\left(\alpha^{(t+1)}\right)^2$, and $\boxed{W_i^{(t+1)} = (\tilde{Z}_i + \alpha^{(t+1)} X_i \beta^{(t+1)})/\alpha^{(t+1)}}$.

    return $\beta^{(t+1)}$, $\Sigma^{(t+1)}$ and $W^{(t+1)}$
    $t + 1 \leftarrow t$
end while
```

involve setting

$$W_i^{(t+1)} = \left(\tilde{Z}_i + \alpha^{(t+1)} X_i \beta^{(t+1)}\right) / \alpha^{(t+1)}, \text{ for } i = 1, \ldots, n, \tag{6.6}$$

see Step 3(b) of Algorithm IvD-1c, instead of $W_i^{(t+1)} = \tilde{W}_i^\star / \alpha^{(t+1)}$, as in Algorithm IvD-1. The correct inverse transformation is necessary to guarantee that the joint stationary distribution of $(W^{(t+1)}, \beta^{(t+1)}, \alpha^{(t+1)}, \Sigma^{(t+1)})$ is the target posterior distribution.

Correction 2 is more subtle. When sampling $\tilde{\Sigma}^\star$ while conditioning on Y, $\tilde{Z}$, and $\beta^{(t+1)}$, Algorithm IvD-1 uses Inv-Wishart$(n + \nu, \sum_{i=1}^n \tilde{Z}_i \tilde{Z}_i^{\mathrm{T}} + \tilde{S})$. IvD, however, ignored a constraint on $\tilde{\Sigma}^\star$ imposed by Y and the current values of $\tilde{Z}$ and β. This constraint is on the first diagonal element of $\tilde{\Sigma}^\star$, i.e., $(\tilde{\sigma}_{11}^\star)^2$. In particular, if we set $\tilde{\xi}_i\left(\tilde{\sigma}_{11}^\star\right) = \tilde{Z}_i + (\tilde{\sigma}_{11}^\star) X_i \beta^{(t+1)}$, for $i = 1, \ldots, n$, the updated value of $\tilde{\sigma}_{11}^\star$ must satisfy

$$\left\{ \begin{array}{ll} \max\left\{\tilde{\xi}_{i1}\left(\tilde{\sigma}_{11}^\star\right), \ldots, \tilde{\xi}_{ip}\left(\tilde{\sigma}_{11}^\star\right)\right\} < 0 & \text{if } Y_i = 0 \\ \max\left\{0, \tilde{\xi}_{i1}\left(\tilde{\sigma}_{11}^\star\right), \ldots, \tilde{\xi}_{ip}\left(\tilde{\sigma}_{11}^\star\right)\right\} = \tilde{\xi}_{ik}\left(\tilde{\sigma}_{11}^\star\right) & \text{if } Y_i = k \end{array} \right., \quad \text{for } \; i = 1, \ldots, n. \tag{6.7}$$

Thus, the conditional distribution of $\tilde{\Sigma}^\star$ given Y, $\tilde{Z}$, and $\beta^{(t+1)}$ in Step 3(a) of Algorithm IvD-1c is a constrained inverse-Wishart distribution, whereas that of Algorithm IvD-1 is unconstrained. The constrained sample in Algorithm IvD-1c is obtained via rejection sampling; we iteratively sample from the unconstrained inverse-Wishart distribution until (6.7) is satisfied.

Algorithms IvD-2 and BN are adaptations of Algorithm IvD-1. Thus, both corrections affect these algorithms as well. See Appendix E for details.

6.3 Simulation Study

We use a simulation study to illustrate the differences of Algorithms IvD-1 and IvD-1c, Algorithms IvD-2 and IvD-2c, and Algorithms BN and BNc. We set $n = 50$, $p = 2$, $q = 2$, $\beta = (-\sqrt{2}, 1)$, and $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. For $X_i = \begin{pmatrix} X_{i1,1} & X_{i1,2} \\ X_{i2,1} & X_{i2,2} \end{pmatrix}$, we sample

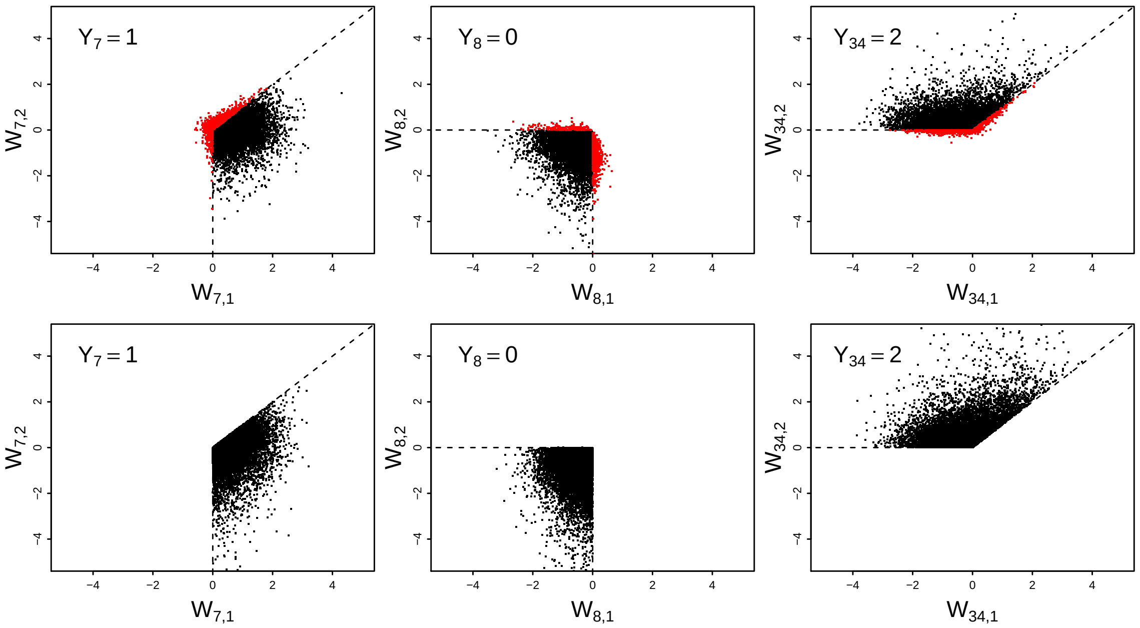

Figure 6.1: The posterior samples of W_7, W_8, and W_{34} from Algorithm IvD-1 with only Correction 1 implemented and Algorithm IvD-1 with both corrections implemented (i.e., Algorithm IvD-1c) appear in the first and second rows, respectively. The draws from the sampler without Correction 2 not adhering to the constraint (6.7) are in red.

$X_{ij,1}$ $(j = 1, 2)$ from a uniform distribution on $(-0.5, 0.5)$ for $i = 1, \ldots, 25$, on $(0.4, 1.5)$ for $i = 26, \ldots, 50$, and sample $X_{ij,2}$ $(j = 1, 2)$ from a uniform distribution on $(-1, 1)$ for $i = 1, \ldots, 25$, on $(0.8, 3)$ for $i = 26, \ldots, 50$. We specify the prior distribution for Σ and α^2 as in Section 6.1, with $\nu = p$, $\alpha_0^2 = \nu$, and $S = \mathrm{Diag}(1, 1)$, and for β, as $\beta \sim \mathrm{N}_q[0, \mathrm{Diag}(100, 100)]$. For each algorithm, we run a chain of 15,000 and discard the first 5,000 draws.

Figure 6.1 presents the posterior samples of W_7, W_8, and W_{34} obtained with Algorithm IvD-1 with Correction 1 implemented, but not Correction 2 (row 1), and with Algorithm IvD-1c, i.e., with both corrections (row 2). The draws from the sampler without Correction 2 that do not adhere to the constraint (6.7) are colored red, which illustrates the second problem of Algorithm IvD-1 described in Section 6.2. These draws are rejected in Step 3(a) of Algorithm IvD-1c.

Unfortunately, Algorithms IvD-1, IvD-2, and BN do not return draws from the target

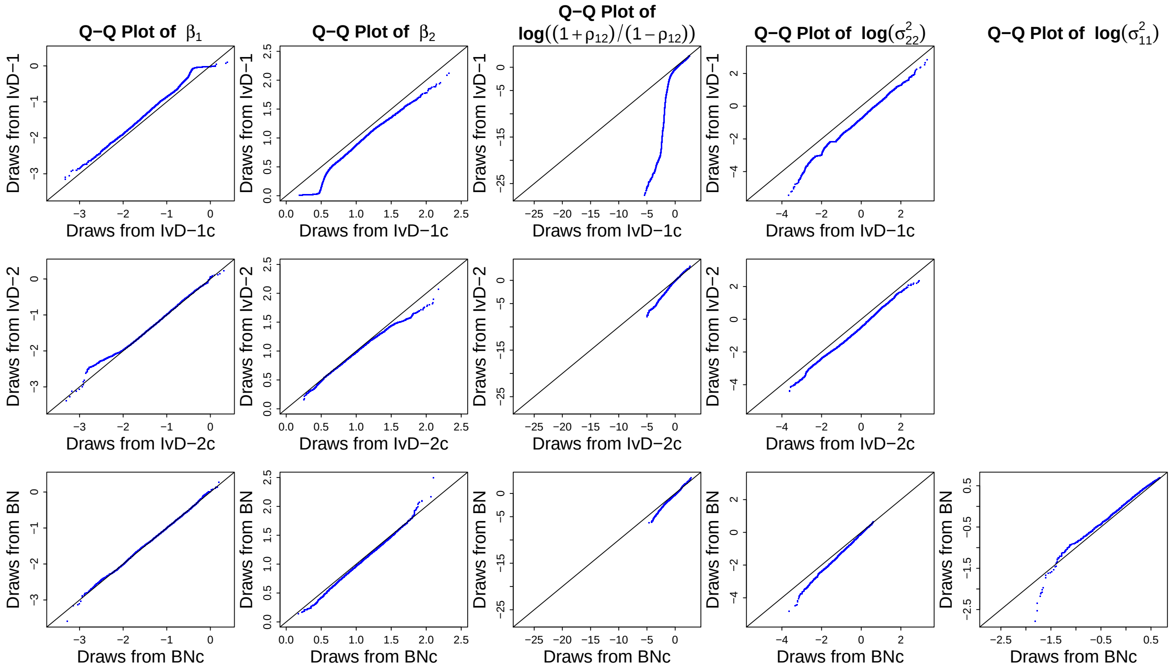

Figure 6.2: Quantile-quantile plots comparing the posterior draws from different algorithms in the simulation study. The columns correspond to five parameters, i.e., β_1, β_2, $\log\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right)$, $\log(\sigma_{22}^2)$, and $\log(\sigma_{11}^2)$. The first row compares draws from Algorithms IvD-1 and IvD-1c, the second row compares draws from Algorithms IvD-2 and IvD-2c, and the last row compares draws from Algorithms BN and BNc. The algorithms of IvD fix $\sigma_{11}^2 = 1$ and thus no comparisons are included for these algorithms in Column 5.

posterior distribution. Figure 6.2 shows quantile-quantile plots that compare the stationary distributions of original and corrected algorithms. The first row of Figure 6.2 compares Algorithms IvD-1 and IvD-1c. The distributions of β differ slightly for the two algorithms, while the distributions of Σ differ significantly, especially the correlation parameter, $\rho_{12} = \sigma_{12}/(\sigma_{11}\sigma_{22})$. The second row shows the quantile-quantile plots that compare Algorithms IvD-2 and IvD-2c. The distributions of both β and Σ differ slightly for the two algorithms. The last row of Figure 6.2 compares Algorithms BN and BNc. The distributions of β are similar for the two algorithms, while the distributions of Σ are different, particularly ρ_{12} and σ_{22}^2.

Trace plots of the parameters obtained with Algorithms IvD-1c, IvD-2c, and BNc exhibit better convergence than the corresponding plots obtained with Algorithms IvD-1, IvD-2, and BN. These plots are omitted to save space.

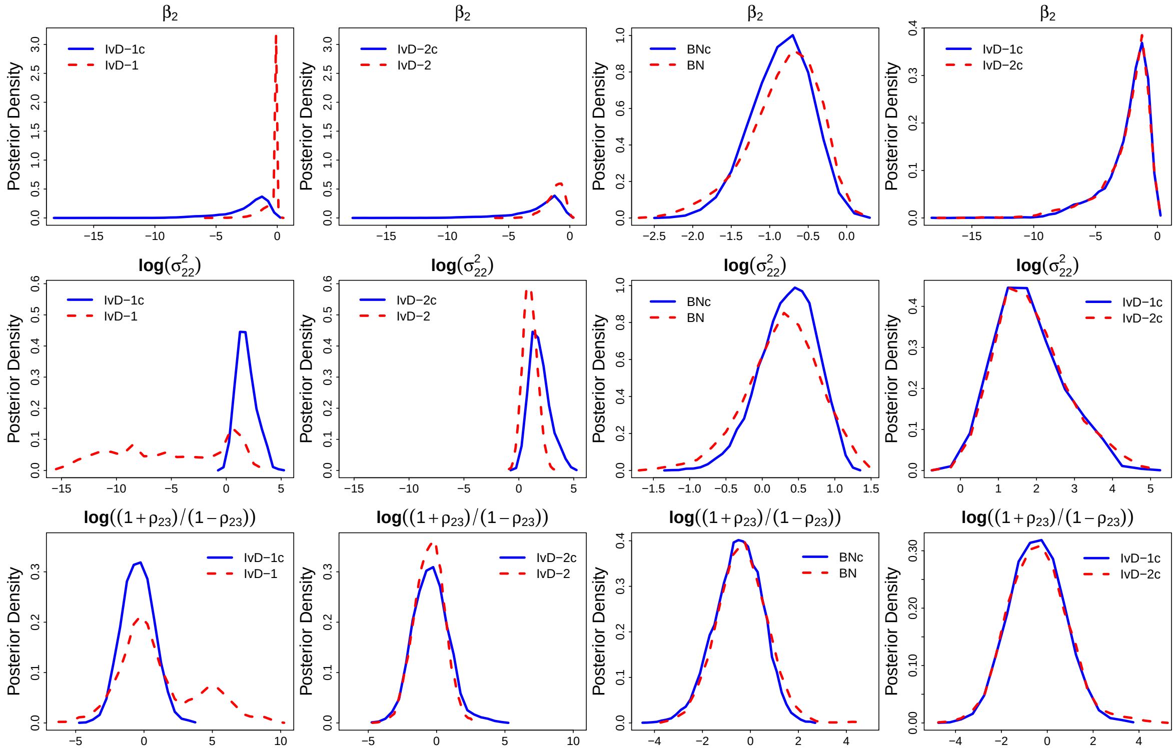

Figure 6.3: Comparison of algorithms in the margarine-purchase data analysis. The first three columns compare posterior distributions of β_2, $\log(\sigma^2_{22})$, and $\log\left(\frac{1+\rho_{23}}{1-\rho_{23}}\right)$ obtained with Algorithms IvD-1c and IvD-1, Algorithms IvD-2c and IvD-2, and Algorithms BNc and BN respectively. Blue solid line represents the corrected algorithms, and red dashed line represents the original algorithms. The last column compares posterior distributions of the three parameters from two corrected algorithms, Algorithms IvD-1c (blue solid line) and IvD-2c (red dashed line).

6.4 Data Analysis

For a further comparison of the algorithms, we consider three real-data analyses. First, we fit the MNP model to a data set describing margarine purchases which is available in the `bayesm` R package. Following BN, we limit analysis to purchases of six brands: “Parkay stick”, “Blue Bonnet stick”, “Fleischmanns stick”, “House brand stick”, “Generic stick”, and “Shedd Spread tub”, and only consider the first purchase of one of these brands for each household. This results in a data set consisting of $n = 507$ observations.

We set “Parkay stick” as the base category, and $p = 5$. Again following BN, we set up

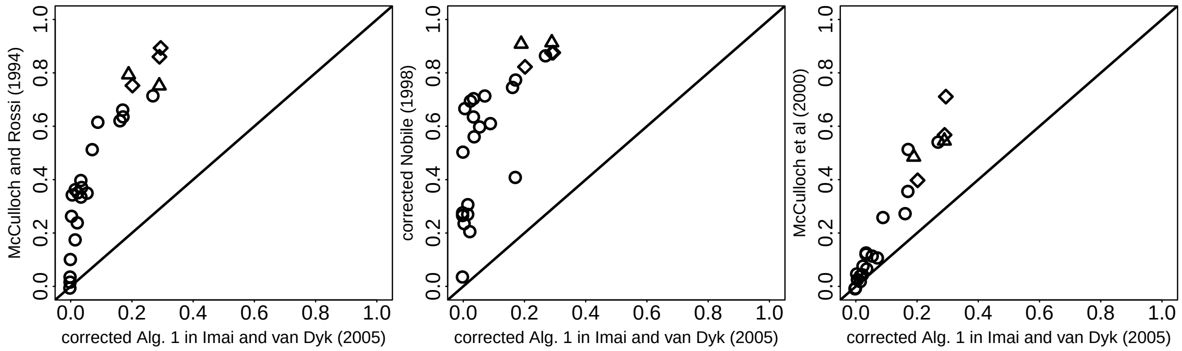

Figure 6.4: Comparison of lag-100 autocorrelations of all the model parameters sampled with Algorithm IvD-1c and the samplers of McCulloch and Rossi (1994), Nobile (1998), and McCulloch *et al.* (2000) in the Dutch election example. Circles, triangles, and rectangles represent coefficient, variance, and correlation parameters respectively.

a model that only includes intercept terms for the other five categories and a coefficient for log prices. Thus $q = 6$, and $X_i = [I_p, g_i]$, where I_p is an identity matrix and g_i is the p-vector of differences in log prices between each category and the base. We again specify the prior distribution for Σ and α^2 as in Section 6.1, with $\nu = p$, $\alpha_0^2 = \nu$, and $S = \text{Diag}(1, \ldots, 1)$, and for β, as $\beta \sim \text{N}_q[0, \text{Diag}(100, \ldots, 100)]$. When implementing Algorithms IvD-1, IvD-1c, IvD-2, and IvD-2c, we set the variance corresponding to "Blue Bonnet stick" to one. For each algorithm, we run a chain of length 300,000, discard the first 100,000 draws, and thin the remaining draws by 10. In this way we obtain 20,000 draws from each algorithm.

The first three columns of Figure 6.3 compare the posterior distributions of β_2, $\log(\sigma_{22}^2)$, and $\log\left(\frac{1+\rho_{23}}{1-\rho_{23}}\right)$ sampled with Algorithms IvD-1 and IvD-1c (column 1), Algorithms IvD-2 and IvD-2c (column 2), and Algorithms BN and BNc (column 3). The three parameters are selected because their stationary distributions show relatively obvious difference for all three pairs of original and corrected algorithms. Algorithms IvD-1, IvD-2, and BN all fail to deliver draws from the target posterior distributions. The situation is most substantial for Algorithm IvD-1. The last column of Figure 6.3 compares two corrected algorithms, Algorithms IvD-1c and IvD-2c, and shows that the posterior distributions of β_2, $\log(\sigma_{22}^2)$, and $\log\left(\frac{1+\rho_{23}}{1-\rho_{23}}\right)$ obtained with Algorithms IvD-1c and IvD-2c are identical. (The corresponding distributions of the other parameters are also identical.) We do not compare Algorithm BNc with Algorithms IvD-1c or IvD-2c because the difference in its model specification means its posterior distribution will differ.

IvD compared the convergence properties of Algorithm IvD-1c with those of the samplers of McCulloch and Rossi (1994), Nobile (1998), and McCulloch *et al.* (2000), using a series of numerical examples. Here we replicate these comparisons in order to verify that Algorithm IvD-1c maintains the advantages of Algorithm IvD relative to the other samplers. Thus we redo both real-data analyses in IvD, replacing Algorithm IvD-1 with Algorithm IvD-1c. The first example considers a data set describing the voter choice in Dutch parliamentary elections. Using the same model setting, prior specification, and starting values as IvD, we run a chain of length 50,000 with Algorithm IvD-1c. Following IvD, we compare Algorithm IvD-1c with the samplers of McCulloch and Rossi (1994), Nobile (1998), and McCulloch *et al.* (2000) in terms of lag-100 autocorrelations of all the model parameters, see Figure 6.4. With smaller autocorrelations for all the parameters, Algorithm IvD-1c exhibits better convergence than the other three algorithms.

The second data set in IvD is on purchases of liquid laundry detergent. We specify the same model and prior distributions as IvD, and run Algorithm IvD-1c to obtain three chains of length 10,000 with the same three sets of starting values used by IvD. Trace plots of the price coefficient, $\log(\sigma^2_{55})$, and $\log\left(\frac{1+\rho_{45}}{1-\rho_{45}}\right)$ obtained with Algorithm IvD-1c are similar to the corresponding plots from Algorithm IvD-1 (presented in the top left panel of Fig. 8 in IvD). Thus Algorithm IvD-1c is less sensitive to the choice of starting values than are the other samplers. (The trace plots from these three samplers are also displayed in Fig. 8 of IvD.) Unfortunately, in this example, however, the rejection sampling of $\tilde{\Sigma}$ is rather time-consuming for the two chains which are initialized far from regions of high posterior density. To solve this problem, we start running the chains without Correction 2, that is, by sampling $\tilde{\Sigma}$ from the unconstrained inverse-Wishart distribution. After an initial run (1,000 iterations in this example), when the chain approaches the high-density region of the target distribution, we deploy Correction 2 and discard the initial run.

6.5 Discussion

The algorithms of IvD and BN are implemented in the popular R package `MNP` and are widely used for fitting MNP models. We point out errors in these algorithms and propose corrections. Using both a simulation study and real-data analyses, we illustrate the differences between the original and corrected algorithms. From these analyses, we find that the errors can significantly affect the final results, especially in that they alter the stationary distributions and hence the fitted parameters. Considering the popularity of these algorithms, it is important that these errors are corrected. We have done so here and also in the `MNP` package. The corrected algorithms require somewhat more computational time due to the additional rejection sampling steps. Like other MCMC samplers for the MNP model, computational time increases with both the sample size n and the number of alternatives p, and like other samplers, is roughly proportional to np. For most of the examples considered in the paper, the extra computational time is fairly small and at least in some cases it is made up by the improved autocorrelation of the corrected algorithms. When the initial values are far from the high-density region of the posterior distribution, the rejection sampling of $\tilde{\Sigma}$ can be computationally expensive. In this case, we propose doing an initial run without Correction 2 and initializing Correction 2 when the chain reaches high-density regions.

Moreover, to further improve the convergence of Algorithm IvD-1c, we have tried combining MDA and ASIS. The ASIS algorithm is constructed conditioning on β. Because the distribution of observed quantities Y conditioning on $\tilde{W}$, β, and $\tilde{\Sigma}$ is free of $\tilde{\Sigma}$; $\tilde{W}$ is the sufficient augmentation for $\tilde{\Sigma}$; $\bar{W} = \Sigma^{-1/2}(\tilde{W} - \alpha X\beta)$ is the corresponding ancillary augmentation because its distribution, $\bar{W}_i \overset{\text{iid}}{\sim} \mathrm{N}_p(0, I)$, is free of $\tilde{\Sigma}$. The first three steps of the combined sampler are the same as the steps of Algorithm IvD-1c. After Step 3, the combined sampler transforms $\tilde{W}$ to $\bar{W}$, samples some components of $\tilde{\Sigma}$ conditioning on $\bar{W}$, β, and the rest $\tilde{\Sigma}$ components, and finally transforms $(\bar{W}, \tilde{\Sigma})$ back to (W, Σ, α). However, this combined sampler does not show significantly more efficiency than Algorithm IvD-1c, either in simulation studies or real-data analyses. One important reason can be that it is difficult to sample $\tilde{\Sigma}$ conditioning on $\bar{W}$. We solve this problem by only updating some components of $\tilde{\Sigma}$ conditioning on others. However, this solution is not sufficiently effective. We are exploring more useful methods to circumvent the intractability of sampling $\tilde{\Sigma}$ conditioning on $\bar{W}$, which may help the combined sampler work better.

7

Conclusion and Future Work

7.1 Conclusion

Gibbs-type samplers are widely used tools for obtaining Monte Carlo samples from posterior distributions under complicated Bayesian models. Standard Gibbs samplers update component quantities of the parameter by sequentially sampling their conditional distributions from the target joint distribution. However, this strategy can be slow to converge if the components are highly correlated. In this manuscript, we make efforts to formalize a general strategy to construct more efficient samplers by using surrogate distributions, which are designed to share certain marginal distributions with the target, but with lower correlations among its components. Specifically, we replace some of the conditional distributions in a Gibbs sampler with conditionals of a surrogate distribution. Although not necessarily recognized when they were introduced, a number of existing strategies for improving Gibbs can be formulated in this way (e.g., Marginal Data Augmentation, Partially Collapsed Gibbs sampling, Ancillarity-Sufficiency Interweaving Strategy, etc.). Under particular settings, these existing strategies are even equivalent in terms of surrogate distributions. However, the use of surrogate distributions in Gibbs-type samplers may lead to incompatible conditional distributions and thus sensitivity to the order of the component draws. Thus we propose a framework to combine different strategies involving surrogate distributions into a single coherent sampler that maintains the target stationary distribution and outperforms any of its component algorithms in terms of convergence. This combining strategy, as a special case of implementing the surrogate distribution strategy, greatly amplifies our power to improve the convergence of Gibbs-type samplers. A problem in supernova cosmology has motivated our work and serves as a realistic testing ground for our methods. Our new algorithms are efficient in fitting the Gaussian hierarchical model in supernova cosmology and several extensions of this model.

7.2 Future Work

In the future, our main task is to refine the theory of our framework and generalize our algorithms for applications in wider field. At this stage, we are exploring the following three problems. First, the combining strategy developed in Chapter 3 can be extended by considering more algorithms. That is, the algorithms to be combined do not need to be restricted within PCG, MDA, ASIS, and MH strategies. Now we are exploring the possibility of combining pseudo marginal Monte Carlo methods and PCG to facilitate implementation and boost efficiency. Second, so far, we have not found convincing numerical examples to illustrate the computational advantage of using surrogate conditionals in a Gibbs-type sampler, especially a sampler with more than two steps. Thus we need to keep searching for such examples that we can verify the power of our surrogate distribution strategy numerically. Furthermore, we should improve the theory for the surrogate distribution strategy, because so far we cannot provide clear guidance on the usage of surrogate conditional distributions in a Gibbs-type (particularly multi-step) sampler, which guarantees an improvement of the convergence properties.

References

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

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

Betoule, M., Kessler, R., Guy, J., Mosher, J., Hardin, D., Biswas, R., and et al. (2014). Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. *Astronomy & Astrophysics* **568**, id.A22–32.

Burgette, L. F. and Nordheim, E. V. (2012). The trace restriction: an alternative identification strategy for the Bayesian multinomial probit model. *Journal of Business and Economic Statistics* **30**, 404–410.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability, London, second edition edn.

Childress, M., Aldering, G., Antilogus, P., Aragon, C., Bailey, S., and Baltay, C. (2013). Host galaxy properties and Hubble residuals of Type Ia supernovae from the nearby supernova factory. *The Astrophysical Journal* **770**, 108–125.

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

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

Filippone, M., Marquand, A. F., Blain, C. R. V., Williams, S. C. R., Mourao-Miranda, J., and Girolami, M. (2012). Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. *The Annals of Applied Statistics* **6**, 1883–1905.

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.

Geweke, J. and Zhou, G. (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. *Review of Financial Studies* **9**, 557–587.

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

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society* **44**, 455–472.

Guy, J., Astier, P., Baumont, S., Hardin, D., Pain, R., Regnault, N., and et al. (2007). Salt2: using distant supernovae to improve the use of type Ia supernovae as distance indicators. *Astronomy & Astrophysics* **466**, 11–21.

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

Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation, and PX-DA algorithms. *The Annals of Statistics* **36**, 532–554.

Imai, K. and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics* **124**, 311–334.

Jiao, X. Y. and van Dyk, D. A. (2016). A corrected and more efficient suite of MCMC samplers for the multinomial probit model. *Journal of Econometrics,* in press.

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

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician* **52**, 93–100.

Kastner, G. and Fruhwirth-Schnatter, S. (2014). Ancillarity-Sufficiency Interweaving Strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics and Data Analysis* **76**, 408–423.

Kelly, B. C., Shetty, R., Stutz, A. M., Kauffmann, J., Goodman, A. A., and Launhardt, R. (2012). Dust spectral energy distributions in the era of Herschel and Planck: a hierarchical Bayesian-fitting technique. *The Astrophysical Journal* **752**, 55–71.

Kessler, R., Becker, A. C., Cinabro, D., Vanderplas, J., and et al. (2009). First-year Sloan Digital Sky Survey-II supernova results: Hubble diagram and cosmological parameters. *The Astrophysical Journal Supplement* **185**, 32–84.

Lancaster, H. O. (1958). The structure of bivariate distributions. *The Annals of Mathematical Statistics* **29**, 719–736.

Lavaux, G. (2016). Bayesian 3D velocity field reconstruction with VIRBIUS. *Monthly Notices of the Royal Astronomical Society* **457**, 172–197.

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

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

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer, 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., Wong, W. H., and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scan. *Journal of the Royal Statistical Society, Series B, Methodological* **57**, 157–169.

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

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

March, M. C., Trotta, R., Berkes, P., Starkman, G. D., and Vaudrevange, P. M. (2011). Improved constraints on cosmological parameters from snia data. *Monthly Notices of the Royal Astronomical Society* **418**, 2308–2329.

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

McCulloch, R., Polson, N., and Rossi, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173–193.

McCulloch, R. E. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **64**, 207–240.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag, London.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association* **108**, 656–665.

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing* **8**, 229–242.

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2007). A general framework for the parameterization of hierarchical models. *Statistical Science* **22**, 59–73.

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

Phillips, M. M. (1993). The absolute magnitudes of type Ia supernovae. *The Astrophysical Journal* **413**, L105–L108.

Phillips, M. M., Lira, P., Suntzeff, N. B., Schommer, R. A., Hamuy, M., and Maza, J. (1999). The reddening-free decline rate versus luminosity relationship for type Ia supernovae. *The Astrophysical Journal* **118**, 1766–1776.

Rao, M. M. (1987). *Measure Theory and Integration.* Wiley, New York.

Reed, C. and Yu, K. (2009). A partially collapsed Gibbs sampler for Bayesian quantile regression. In *Mathematics School of Information Systems, Computing and Mathematics Research Papers.* Brunel University Research Archive, London.

Robert, C. (1995). Simulation of Truncated Normal Variables. *Statistics and Computing* **5**, 121–125.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer Texts in Statistics, New York, second edition edn.

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

Royle, J. A. and Dorazio, R. M. (2012). Parameter-expanded data augmentation for Bayesian analysis of capture-recapture models. *Journal of Ornithology* **152**, 521–537.

Schliep, E. M. and Hoeting, J. A. (2015). Data augmentation and parameter expansion for independent or spatially correlated ordinal data. *Computational Statistics and Data Analysis* **90**, 1–14.

Shariff, H., Jiao, X. Y., Trotta, R., and van Dyk, D. A. (2016). BAHAMAS: new analysis of type Ia supernovae reveals inconsistencies with standard cosmology. *The Astrophysical Journal,* to appear.

Smith, A. and Roberts, G. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **55**, 3–24.

Sullivan, M., Conley, A., Howell, D. A., Neill, J. D., Astier, P., and Balland, C. (2010). The dependence of Type Ia supernovae luminosities on their host galaxies. *Monthly Notices of the Royal Astronomical Society* **406**, 782–802.

Sullivan, M., Le Borgne, D., Pritchet, C. J., Hodsman, A., Neill, J. D., and Howell, D. A. (2006). Rates and properties of Type Ia supernovae as a function of mass and star formation in their host galaxies. *The Astrophysical Journal* **648**, 868–883.

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.

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

van Dyk, D. A. (2010). Marginal Markov chain Monte Carlo methods. *Statistica Sinica* **20**, 1423–1454.

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 Jiao, X. Y. (2015). Metropolis-Hastings within partially collapsed Gibbs samplers. *Journal of Computational and Graphical Statistics* **24**, 301–327.

van Dyk, D. A. and Meng, X. L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.

van Dyk, D. A. and Meng, X. L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: a graphical guide book. *Statistical Science* **25**, 429–449.

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

Woodard, D. B., Crainiceanu, C., and Ruppert, D. (2013). Hierarchical adaptive regression kernels for regression with functional predictors. *Journal of Computational and Graphical Statistics* **22**, 777–800.

Yu, Y. and Meng, X. L. (2011). To center or not to center: that is not the question—an Ancillarity-Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency (with discussion). *Journal of Computational and Graphical Statistics* **20**, 531–570.

A

Details of Steps of the Gibbs-type Samplers in Chapter 2

A.1 Details of the MH within Gibbs and MH within PCG samplers for fitting the spectral model in (2.1)

With noninformative uniform prior distributions for all of the parameters, the posterior distribution of α, β, γ, μ, and ϕ under the spectral model (2.1) is

$$\begin{aligned} p(\alpha, \beta, \gamma, \mu, \phi | Y) \propto \; & \prod_{i=1}^{n} \left[\alpha({E_i}^{-\beta} + \gamma I\{i = \mu\}) e^{-\phi/E_i}\right]^{Y_i} \\ & \exp\left\{-\alpha \sum_{i=1}^{n} ({E_i}^{-\beta} + \gamma I\{i = \mu\}) e^{-\phi/E_i}\right\}. \end{aligned} \tag{A.1}$$

The joint posterior distribution of α, β, γ, μ, ϕ, and the augmented data Y_L is

$$\begin{aligned} p(Y_L, \alpha, \beta, \gamma, \mu, \phi | Y) \; & \propto \alpha^{\sum_{i=1}^{n} Y_i} e^{-\phi \sum_{i=1}^{n} (Y_i/E_i)} \left[\prod_{i=1}^{n} \frac{E_i^{-\beta(Y_i - Y_{iL})}}{(Y_i - Y_{iL})! Y_{iL}!}\right] \gamma^{\sum_{i=1}^{n} Y_{iL}} \\ & \times \left\{\prod_{i=1}^{n} [I\{i = \mu\}]^{Y_{iL}}\right\} \exp\left\{-\alpha \sum_{i=1}^{n} ({E_i}^{-\beta} + \gamma I\{i = \mu\}) e^{-\phi/E_i}\right\}. \end{aligned} \tag{A.2}$$

Thus the steps of the parent MH within Gibbs sampler in Figure 2.3(a) or 2.9(a) are

Step 1: Sample Y_{iL} from Binomial$\left\{Y_i, \frac{\gamma' I\{i=\mu'\}}{{E_i}^{-\beta'}+\gamma' I\{i=\mu'\}}\right\}$, for $i = 1, \dots, n$,

Step 2: Sample α from Gamma$\left\{\sum_{i=1}^n Y_i + 1,\ \sum_{i=1}^n ({E_i}^{-\beta'} + \gamma' I\{i = \mu'\})e^{-\phi'/E_i}\right\}$,

Step 3: Use MH to sample β from $p(\beta|Y, Y_L, \alpha, \gamma', \mu', \phi') \propto p(Y_L, \alpha, \beta, \gamma', \mu', \phi'|Y)$,

Step 4: Sample γ from Gamma$\left\{\sum_{i=1}^n Y_{iL} + 1,\ \alpha \sum_{i=1}^n I\{i = \mu'\}e^{-\phi'/E_i}\right\}$,

Step 5: Use MH to sample μ from $p(\mu|Y, Y_L, \alpha, \beta, \gamma, \phi') \propto p(Y_L, \alpha, \beta, \gamma, \mu, \phi'|Y)$,

Step 6: Use MH to sample ϕ from $p(\phi|Y, Y_L, \alpha, \beta, \gamma, \mu) \propto p(Y_L, \alpha, \beta, \gamma, \mu, \phi|Y)$.

The steps of the MH within PCG sampler with lowest degree of partial collapsing, i.e., Sampler 2.1, are

Step 1: Use MH to sample μ from $p(\mu|Y, \alpha', \beta', \gamma', \phi') \propto p(\alpha', \beta', \gamma', \mu, \phi'|Y)$,

Step 2: Sample Y_{iL} from Binomial$\left\{Y_i, \frac{\gamma' I\{i=\mu\}}{{E_i}^{-\beta'}+\gamma' I\{i=\mu\}}\right\}$, for $i = 1, \dots, n$,

Step 3: Sample α from Gamma$\left\{\sum_{i=1}^n Y_i + 1,\ \sum_{i=1}^n ({E_i}^{-\beta'} + \gamma' I\{i = \mu\})e^{-\phi'/E_i}\right\}$,

Step 4: Use MH to sample β from $p(\beta|Y, Y_L, \alpha, \gamma', \mu, \phi') \propto p(Y_L, \alpha, \beta, \gamma', \mu, \phi'|Y)$,

Step 5: Sample γ from Gamma$\left\{\sum_{i=1}^n Y_{iL} + 1,\ \alpha \sum_{i=1}^n I\{i = \mu\}e^{-\phi'/E_i}\right\}$,

Step 6: Use MH to sample ϕ from $p(\phi|Y, Y_L, \alpha, \beta, \gamma, \mu) \propto p(Y_L, \alpha, \beta, \gamma, \mu, \phi|Y)$.

Integrating (A.1) over α, we have,

$$\begin{aligned} p(\beta, \gamma, \mu, \phi|Y) \propto & \prod_{i=1}^n \left[({E_i}^{-\beta} + \gamma I\{i = \mu\})e^{-\phi/E_i}\right]^{Y_i} \times \\ & \left[\sum_{i=1}^n ({E_i}^{-\beta} + \gamma I\{i = \mu\})e^{-\phi/E_i}\right]^{-(\sum_{i=1}^n Y_i + 1)}. \end{aligned} \tag{A.3}$$

Hence, the steps of the MH within PCG sampler with medium degree of partial collapsing, i.e., Sampler 2.2, are

Step 1: Use MH to sample μ from $p(\mu|Y, \beta', \gamma', \phi') \propto p(\beta', \gamma', \mu, \phi'|Y)$,

Step 2: Use MH to sample ϕ from $p(\phi|Y, \beta', \gamma', \mu) \propto p(\beta', \gamma', \mu, \phi|Y)$,

Step 3: Use MH to sample β from $p(\beta|Y, \gamma', \mu, \phi) \propto p(\beta, \gamma', \mu, \phi|Y)$,

Step 4: Sample α from Gamma$\left\{\sum_{i=1}^n Y_i + 1,\ \sum_{i=1}^n ({E_i}^{-\beta} + \gamma' I\{i = \mu\})e^{-\phi/E_i}\right\}$,

Step 5: Sample Y_{iL} from Binomial$\left\{Y_i, \frac{\gamma' I\{i=\mu\}}{{E_i}^{-\beta}+\gamma' I\{i=\mu\}}\right\}$, for $i = 1, \dots, n$,

Step 6: Sample γ from Gamma$\left\{\sum_{i=1}^n Y_{iL} + 1,\ \alpha \sum_{i=1}^n I\{i = \mu\}e^{-\phi/E_i}\right\}$.

The steps of the MH within PCG sampler with highest degree of partial collapsing, i.e., Sampler 2.3, are almost the same as Sampler 2.2, except that Steps 2 and 3 are combined into one single step. That is, we use MH to sample (β, ϕ) from $p(\beta, \phi|Y, \gamma', \mu) \propto p(\beta, \gamma', \mu, \phi|Y)$.

We use a uniform distribution on $\{1, \dots, n\}$ as the jumping rule when updating μ. When updating either β or ϕ via MH, we specify a Gaussian distribution centered at the current draw of the parameter as the jumping rule; the variance of the jumping rule is adjusted to obtain an acceptance rate of around 40%. Analogously, when sampling β and ϕ jointly via MH, the jumping rule is a bivariate Gaussian distribution centered at the current draw with variance-covariance matrix adjusted to obtain an acceptance rate of around 20%.

A.2 Details of the MH within Gibbs and MH within PCG samplers for fitting the cosmological hierarchical model in (2.14)–(2.16)

The posterior distribution of $(\xi, X, \mathscr{C}, \alpha, \beta, \Sigma_P)$ is

$$\begin{aligned} p(\xi, X, \mathscr{C}, \alpha, \beta, \Sigma_P|Y) \;&\propto |\Sigma_C|^{-1/2}|\Sigma_P|^{-1/2}|\Sigma_0|^{-1/2}\frac{1}{R_c^2}\frac{1}{R_x^2}\frac{1}{\sigma_{\text{res}}^2} \\ &\exp\bigg\{-\frac{1}{2}\Big[(Y - AX - L)^T\Sigma_C^{-1}(Y - AX - L) \\ &+ (X - J\xi)^T\Sigma_P^{-1}(X - J\xi) + (\xi - \xi_m)^T\Sigma_0^{-1}(\xi - \xi_m)\Big]\bigg\}, \end{aligned} \tag{A.4}$$

where $J_{(3n\times 3)} = (I, \dots, I)$ with $I = \text{Diag}(1, 1, 1)$, $A_{(3n\times 3n)} = \text{Diag}(T, \dots, T)$ with $T_{(3\times 3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta & -\alpha & 1 \end{bmatrix}$, L denotes the $(3n \times 1)$ vector $(0, 0, \mu_1, \dots, 0, 0, \mu_n)$, which is a deterministic function of $\mathscr{C}$, $\xi_m = (0, 0, M_m)$, Σ_C and Σ_P are defined in Section 2.3.2, and $\Sigma_0 = \text{Diag}(\sigma_{c_0}^2, \sigma_{x_0}^2, \sigma_{M_0}^2)$.

The steps of the MH within Gibbs sampler, i.e., Sampler 2.7, are

Step 1: Sample (ξ, X), which consists of two sub-steps:

- Sample ξ from $\text{N}(k_0, K)$;
- Sample X from $\text{N}(\mu_A, \Sigma_A)$.

where $\Sigma_A^{-1} = A^T\Sigma_C^{-1}A + \Sigma_P^{-1}$, $K^{-1} = -J^T\Sigma_P^{-1}\Sigma_A\Sigma_P^{-1}J + J^T\Sigma_P^{-1}J + \Sigma_0^{-1}$, $\Delta = A^T\Sigma_C^{-1}(Y - L)$, $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1}J\xi)$, and $k_0 = K(J^T\Sigma_P^{-1}\Sigma_A\Delta + \Sigma_0^{-1}\xi_m)$.

Step 2: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y,\xi,X,\alpha,\beta,\Sigma_P)$, which is proportional to $p(\xi,X,\mathscr{C},\alpha,\beta,\Sigma_P|Y)$, under the constraint $\mathscr{C}=(\Omega_m,\Omega_\Lambda)\in[0,1]\times[0,2]$.

Step 3: Sample (α,β) from $\mathrm{N}(\mu_B,\Sigma_B)$ with constraint $(\alpha,\beta)\in[0,1]\times[0,4]$, where

$$\Sigma_B^{-1}=\begin{bmatrix}\sum\limits_{i=1}^{n}\frac{x_i^2}{\hat{\sigma}_{mBi}^2} & \sum\limits_{i=1}^{n}\frac{-x_ic_i}{\hat{\sigma}_{mBi}^2}\\ \sum\limits_{i=1}^{n}\frac{-x_ic_i}{\hat{\sigma}_{mBi}^2} & \sum\limits_{i=1}^{n}\frac{c_i^2}{\hat{\sigma}_{mBi}^2}\end{bmatrix}\text{ and }\mu_B=\Sigma_B\begin{bmatrix}\sum_{i=1}^{n}\frac{x_i(M_i-\hat{m}_{Bi}+\mu_i)}{\hat{\sigma}_{mBi}^2}\\ \sum_{i=1}^{n}\frac{-c_i(M_i-\hat{m}_{Bi}+\mu_i)}{\hat{\sigma}_{mBi}^2}\end{bmatrix}.$$

Step 4: Sample Σ_P, which consists of three sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2},\frac{\sum_{i=1}^{n}(c_i-c_0)^2}{2}\right]$ with $\log(R_c)\in[-5,2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2},\frac{\sum_{i=1}^{n}(x_i-x_0)^2}{2}\right]$ with $\log(R_x)\in[-5,2]$.
- Sample σ_{res}^2 from Inv-Gamma $\left[\frac{n}{2},\frac{\sum_{i=1}^{n}(M_i-M_0)^2}{2}\right]$ with $\log(\sigma_{\text{res}})\in[-5,2]$.

Moreover, integrating (ξ,X) out of (A.4), the marginal distribution of $(\mathscr{C},\alpha,\beta,\Sigma_P)$ is

$$\begin{aligned}p(\mathscr{C},\alpha,\beta,\Sigma_P|Y)\ \propto\ & |\Sigma_C|^{-1/2}|\Sigma_P|^{-1/2}|\Sigma_A|^{1/2}|K|^{1/2}|\Sigma_0|^{-1/2}\frac{1}{R_c^2}\frac{1}{R_x^2}\frac{1}{\sigma_{\text{res}}^2}\\ \times\ & \exp\Big\{-\frac{1}{2}\big[(Y-L)^T\Sigma_C^{-1}(Y-L)-\Delta^T\Sigma_A\Delta\\ & -k_0^TK^{-1}k_0+\xi_m^T\Sigma_0^{-1}\xi_m\big]\Big\}.\end{aligned}\tag{A.5}$$

The steps of the MH within PCG sampler, i.e., Sampler 2.8, are

Step 1: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y,\alpha,\beta,\Sigma_P)$, which is proportional to $p(\mathscr{C},\alpha,\beta,\Sigma_P|Y)$, with $\mathscr{C}=(\Omega_m,\Omega_\Lambda)\in[0,1]\times[0,2]$.

Step 2: Use MH to sample (α,β) from $p(\alpha,\beta|Y,\mathscr{C},\Sigma_P)$, which is proportional to $p(\mathscr{C},\alpha,\beta,\Sigma_P|Y)$, with $(\alpha,\beta)\in[0,1]\times[0,4]$.

Step 3: Sample (ξ,X), which consists of two sub-steps:

- Sample ξ from $\mathrm{N}(k_0,K)$;
- Sample X from $\mathrm{N}(\mu_A,\Sigma_A)$, where $\mu_A=\Sigma_A(\Delta+\Sigma_P^{-1}J\xi)$.

Step 4: Sample Σ_P, which consists of three sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2},\frac{\sum_{i=1}^{n}(c_i-c_0)^2}{2}\right]$ with $\log(R_c)\in[-5,2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2},\frac{\sum_{i=1}^{n}(x_i-x_0)^2}{2}\right]$ with $\log(R_x)\in[-5,2]$.

- Sample σ^2_{res} from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^{n}(M_i - M_0)^2}{2}\right]$ with $\log(\sigma_{\text{res}}) \in [-5, 2]$.

When MH updates are required in any of the samplers, we use truncated Gaussian distributions as the proposal distributions. These distributions are centered at the current draws with variance-covariance matrices adjusted to obtain an acceptance rate of around 25%. The truncation enforces prior constraints and in all cases the MH updates are bivariate.

When generating parameters from a truncated distribution, we repeat drawing from the corresponding unconstrained distribution until the truncation condition is satisfied. In the cosmological example, rejection sampling is not computationally demanding, since the ranges of the prior distributions are fairly large.

A.3 Details of the Gibbs and MH within PCG samplers for fitting the factor analysis model in (2.21)

With priors $p(\sigma_j^2) = $ Inv-Gamma(a, b) $(j = 1, \dots, 6)$ and $p(\beta) \propto 1$, the joint posterior distribution of Z, β, and Σ under the factor analysis model (2.21) is

$$\begin{aligned} p(Z, \beta, \Sigma | Y) \quad \propto \quad & \exp\left\{ -\frac{1}{2} \sum_{i=1}^{100} \left[(Y_i - \beta Z_i)^T \Sigma^{-1} (Y_i - \beta Z_i) + Z_i{}^T Z_i \right] \right\} \\ & |\Sigma|^{-100/2} \left(\prod_{j=1}^{6} \sigma_j^{-2(a+1)} \right) \exp\left\{ -b \sum_{j=1}^{6} \sigma_j^{-2} \right\}. \end{aligned} \tag{A.6}$$

Thus the steps of the Gibbs sampler, i.e., Sampler 2.9, are

Step 1: Sample Z_i from $\text{N}_2 \left[(I_2 + \beta'^T \Sigma'^{-1} \beta')^{-1} \beta'^T \Sigma'^{-1} Y_i, (I_2 + \beta'^T \Sigma'^{-1} \beta')^{-1} \right]$, for $i = 1, \dots, 100$.

Step 2: Sample σ_j^2 from Inv-Gamma $\left[a + \frac{100}{2}, b + \frac{1}{2} \sum_{i=1}^{100} (Y_{ij} - \beta_j' Z_i)^2 \right]$, for $j = 1, \dots, 6$, where β_j' denotes the jth row of β'.

Step 3: Because of the constraint that $\beta_{11} > 0$, $\beta_{22} > 0$ and $\beta_{12} = 0$, we sample β from $p(\beta | Y, Z, \Sigma)$ by

- sampling β_{11} from TN $\left[\frac{\sum_{i=1}^{100} Y_{i1} Z_{i1}}{\sum_{i=1}^{100} Z_{i1}^2}, \frac{\sigma_1^2}{\sum_{i=1}^{100} Z_{i1}^2} \right] \Bigg|_{\beta_{11} > 0}$,

  where $\text{TN}(\mu_0, \sigma_0^2)\big|_F$ denotes a normal distribution $\text{N}(\mu_0, \sigma_0^2)$ truncated by the constraint F,

- sampling β_2 from $\text{TN}_2\left[\left(ZZ^T\right)^{-1}ZY_{.2}, \left(ZZ^T/\sigma_2^2\right)^{-1}\right]\Bigg|_{\beta_{22}>0}$,

  where $Y_{.2}$ denotes the vector $(Y_{12}\ldots, Y_{n2})$,
- sampling β_j from $\text{N}_2\left[\left(ZZ^T\right)^{-1}ZY_{.j}, \left(ZZ^T/\sigma_j^2\right)^{-1}\right]$, for $j = 3, \ldots, 6$.

Integrating (A.6) out of Z, we obtain the marginal posterior distribution of β and Σ, that is,

$$\begin{aligned} p(\beta, \Sigma|Y) \quad \propto \quad & \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[Y_i^T(\Sigma^{-1} - \Sigma^{-1}\beta\left(I_2 + \beta^T\Sigma^{-1}\beta\right)^{-1}\beta^T\Sigma^{-1})Y_i^T\right]\right\} \\ & \left|I_2 + \beta^T\Sigma^{-1}\beta\right|^{-100/2}|\Sigma|^{-100/2}\left(\prod_{j=1}^{6}\sigma_j^{-2(a+1)}\right)\exp\left\{-b\sum_{j=1}^{6}\sigma_j^{-2}\right\}. \end{aligned} \tag{A.7}$$

Henceforth, the steps of the MH within PCG sampler, i.e., Sampler 2.10, are

Step j**:** Use MH to sample σ_j^2 from $p(\sigma_j^2|Y, \sigma_{-j}^{2'}, \beta') \propto p(\beta, \Sigma|Y)$, for $j = 1, \ldots, 4$,

where $\sigma_{-j}^{2'}$ denotes the (5×1) vector $(\sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_{j+1}^{2'}, \ldots, \sigma_6^{2'})$.

Step 5**:** Sample Z_i from $\text{N}_2\left\{\left[I_2 + \beta'^T(\Sigma')^{-1}\beta'\right]^{-1}\beta'^T(\Sigma')^{-1}Y_i, \left[I_2 + \beta'^T(\Sigma')^{-1}\beta'\right]^{-1}\right\}$, for $i = 1, \ldots, 100$,

where $\Sigma' = \text{Diag}(\sigma_1^2, \ldots, \sigma_4^2, \sigma_5^{2'}, \sigma_6^{2'})$.

Step 6**:** Sample σ_j^2 from Inv-Gamma $\left[a + \frac{100}{2}, b + \frac{1}{2}\sum_{i=1}^{100}\left(Y_{ij} - \beta_j' Z_i\right)^2\right]$, for $j = 5, 6$.

Step 7**:** Sample β from $p(\beta|Y, Z, \Sigma)$ by

- sampling β_{11} from $\text{TN}\left[\frac{\sum_{i=1}^{100} Y_{i1}Z_{i1}}{\sum_{i=1}^{100} Z_{i1}^2}, \frac{\sigma_1^2}{\sum_{i=1}^{100} Z_{i1}^2}\right]\Bigg|_{\beta_{11}>0}$,
- sampling β_2 from $\text{TN}_2\left[\left(ZZ^T\right)^{-1}ZY_{.2}, \left(ZZ^T/\sigma_2^2\right)^{-1}\right]\Bigg|_{\beta_{22}>0}$,
- sampling β_j from $\text{N}_2\left[\left(ZZ^T\right)^{-1}ZY_{.j}, \left(ZZ^T/\sigma_j^2\right)^{-1}\right]$, for $j = 3, \ldots, 6$.

To sample σ_j^2 ($j = 1, \ldots, 4$) without Z in Sampler 2.10, we first update $\log(\sigma_j^2)$ via MH. The proposal distribution is specified as a Gaussian distribution centered at the logarithm of the current draw of σ_j^2; the variance is adjusted to obtain an acceptance rate of around 40%. Then the new iteration of σ_j^2 is set to the exponential of the updated $\log(\sigma_j^2)$.

B

The Proof of One Statement and Details of the Gibbs-type Samplers in Chapter 3

B.1 Proof of the equality of the cyclic-permutation bound of the Gibbs sampler in Figure 3.2 and that of Gibbs Sampler 2

Proof. Recall that the cyclic-permutation bound of an N-step Gibbs-type sampler is defined by $\min_{j\in\{0,\dots,N-1\}}\{\gamma_j\}$, where γ_j is the norm of the forward operator corresponding to the j-step-lagged sampler, see Section 1.3 of Chapter 1. Henceforth, to prove that the cyclic-permutation bound of the Gibbs sampler in Figure 3.2 equals to that of Gibbs Sampler 2, we just need to show that these two samplers have equal values for all of $\{\gamma_j, j = 0, 1, 2\}$. We use γ_j^1 to denote the norm of the forward operator corresponding to the j-step-lagged sampler of the Gibbs sampler in Figure 3.2, and γ_j^2 to denote that of Gibbs Sampler 2. We present the proof of $\gamma_1^1 = \gamma_1^2$ here; $\gamma_0^1 = \gamma_0^2$ and $\gamma_2^1 = \gamma_2^2$ can be shown in the similar way.

The 1-step-lagged sampler of Gibbs Sampler 2 proceeds by

1. $\tilde{p}(\psi_2|\alpha', \tilde{\psi}_1', \psi_3')$
2. $p(\psi_3|\psi_1', \psi_2)$ (where $\psi_1' = \mathcal{G}_{\alpha'}^{-1}(\tilde{\psi}_1')$)
3. $\tilde{p}(\alpha, \tilde{\psi}_1|\psi_2, \psi_3)$; set $\psi_1 = \mathcal{G}_\alpha^{-1}(\tilde{\psi}_1)$,

where Step 1 is equivalent to sampling ψ_2 from $p(\psi_2|\psi_1', \psi_3')$ and Step 3 is equivalent to sampling (α, ψ_1) from $p(\alpha)p(\psi_1|\psi_2, \psi_3)$. The stationary distribution of this 1-step-lagged sampler is $\tilde{p}(\alpha, \psi_1, \psi_2, \psi_3)p(\alpha)p(\psi_1, \psi_2, \psi_3)$.

The 1-step-lagged sampler of the Gibbs sampler in Figure 3.2 proceeds by

1. $p(\psi_2|\psi_1', \psi_3')$
2. $p(\psi_3|\psi_1', \psi_2)$
3. $p(\psi_1|\psi_2, \psi_3)$,

and its stationary distribution is $p(\psi_1, \psi_2, \psi_3)$.

Then $\forall h \in L_0^2(\tilde{p})$, with $\mathrm{var}_{\tilde{p}}(h) = 1$, we have

$$\begin{aligned}
& \mathrm{var}_{\tilde{p}}[\mathrm{E}(h(\alpha, \psi_1, \psi_2, \psi_3)|\psi_1', \psi_3')] = \mathrm{E}_{\tilde{p}}[\mathrm{E}^2(h(\alpha, \psi_1, \psi_2, \psi_3)|\psi_1', \psi_3')] \\
= \;& \int \left[\int h(\alpha, \psi_1, \psi_2, \psi_3)p(\psi_2|\psi_1', \psi_3')p(\psi_3|\psi_1', \psi_2)p(\alpha)p(\psi_1|\psi_2, \psi_3)\mathrm{d}\alpha\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3\right]^2 \\
& p(\alpha')p(\psi_1', \psi_2', \psi_3')\mathrm{d}\alpha'\mathrm{d}\psi_1'\mathrm{d}\psi_2'\mathrm{d}\psi_3' \\
= \;& \mathrm{var}_p(h^\star) \int \left[\int \frac{h^\star(\psi_1, \psi_2, \psi_3)}{\sqrt{\mathrm{var}_p(h^\star)}} p(\psi_2|\psi_1', \psi_3')p(\psi_3|\psi_1', \psi_2)p(\psi_1|\psi_2, \psi_3)\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3\right]^2 \\
& p(\psi_1', \psi_2', \psi_3')\mathrm{d}\psi_1'\mathrm{d}\psi_2'\mathrm{d}\psi_3' \\
\leq \;& (\gamma_1^1)^2 \mathrm{var}_p(h^\star) \leq (\gamma_1^1)^2,
\end{aligned} \tag{B.1}$$

where $h^\star(\psi_1, \psi_2, \psi_3) = \int h(\alpha, \psi_1, \psi_2, \psi_3)p(\alpha)\mathrm{d}\alpha$; $h^\star \in L_0^2(p)$, and $\mathrm{var}_p(h^\star) \leq 1$ because

i) $\mathrm{E}_p(h^\star) = \int [\int h(\alpha, \psi_1, \psi_2, \psi_3)p(\alpha)\mathrm{d}\alpha]p(\psi_1, \psi_2, \psi_3)\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3 = \mathrm{E}_{\tilde{p}}(h) = 0$;

ii) $\mathrm{var}_p(h^\star) = \int [\int h(\alpha, \psi_1, \psi_2, \psi_3)p(\alpha)\mathrm{d}\alpha]^2 p(\psi_1, \psi_2, \psi_3)\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3 \leq \mathrm{var}_{\tilde{p}}(h) = 1$.

From (B.1), we conclude that $\gamma_1^2 \leq \gamma_1^1$.

Next, $\forall g \in L_0^2(p)$ with $\mathrm{var}_p(g) = 1$, it also holds that $g \in L_0^2(\tilde{p})$ with $\mathrm{var}_{\tilde{p}}(g) = 1$ since

i) $\mathrm{E}_{\tilde{p}}(g) = \int g(\psi_1, \psi_2, \psi_3)p(\alpha)p(\psi_1, \psi_2, \psi_3)\mathrm{d}\alpha\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3 = \mathrm{E}_p(g) = 0$;

ii) $\mathrm{var}_{\tilde{p}}(g) = \int g^2(\psi_1, \psi_2, \psi_3)p(\alpha)p(\psi_1, \psi_2, \psi_3)\mathrm{d}\alpha\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3 = \mathrm{var}_p(g) = 1$.

Thus, in addition, we have

$$\begin{aligned}
& \mathrm{var}_p[\mathrm{E}(g(\psi_1, \psi_2, \psi_3)|\psi_1', \psi_3')] = \mathrm{E}_p[\mathrm{E}^2(g(\psi_1, \psi_2, \psi_3)|\psi_1', \psi_3')] \\
= \;& \int \left[\int g(\psi_1, \psi_2, \psi_3)p(\psi_2|\psi_1', \psi_3')p(\psi_3|\psi_1', \psi_2)p(\psi_1|\psi_2, \psi_3)\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3\right]^2 \\
& p(\psi_1', \psi_2', \psi_3')\mathrm{d}\psi_1'\mathrm{d}\psi_2'\mathrm{d}\psi_3' \\
= \;& \int \left[\int g(\psi_1, \psi_2, \psi_3)p(\psi_2|\psi_1', \psi_3')p(\psi_3|\psi_1', \psi_2)p(\alpha)p(\psi_1|\psi_2, \psi_3)\mathrm{d}\alpha\mathrm{d}\psi_1\mathrm{d}\psi_2\mathrm{d}\psi_3\right]^2 \\
& p(\alpha')p(\psi_1', \psi_2', \psi_3')\mathrm{d}\alpha'\mathrm{d}\psi_1'\mathrm{d}\psi_2'\mathrm{d}\psi_3' \\
\leq \;& (\gamma_1^2)^2.
\end{aligned} \tag{B.2}$$

From (B.2), we conclude that $\gamma_1^1 \leq \gamma_1^2$.

Combining (B.1) and (B.2), $\gamma_1^1 = \gamma_1^2$. □

B.2 Details of the ASIS and combined samplers for fitting the factor analysis model (2.21)

Setting $W_i = \beta Z_i$ ($i = 1, \ldots, 100$), the joint posterior distribution of W, β, and Σ is,

$$p(W, \beta, \Sigma | Y) \;\propto\; \exp\left\{ -\frac{1}{2} \sum_{i=1}^{100} \left[(Y_i - W_i)^T \Sigma^{-1} (Y_i - W_i) + \tilde{W}_i^T \left(\tilde{\beta}\tilde{\beta}^T \right)^{-1} \tilde{W}_i \right] \right\}$$
$$\left| \tilde{\beta}\tilde{\beta}^T \right|^{-100/2} |\Sigma|^{-100/2} \left(\prod_{j=1}^{6} \sigma_j^{-2(a+1)} \right) \exp\left\{ -b \sum_{j=1}^{6} \sigma_j^{-2} \right\}, \tag{B.3}$$

where $\tilde{W}_i$ ($i = 1, \ldots, 100$) denotes the sub-vector (W_{i1}, W_{i2}), and $\tilde{\beta}$ is defined in Section 3.1.

The steps of the ASIS sampler, i.e., Sampler 3.1, are

Step 1: Sample $Z_i^\star$ from $\mathrm{N}_2 \left[\left(I_2 + \beta'^T \Sigma'^{-1} \beta' \right)^{-1} \beta'^T \Sigma'^{-1} Y_i, \left(I_2 + \beta'^T \Sigma'^{-1} \beta' \right)^{-1} \right]$, for $i = 1, \ldots, 100$.

Step 2: Sample σ_j^2 from Inv-Gamma $\left[a + \frac{100}{2}, b + \frac{1}{2} \sum_{i=1}^{100} \left(Y_{ij} - \beta_j' Z_i^\star \right)^2 \right]$, for $j = 1, \ldots, 6$.

Step 3: Sample $\beta^\star$ from $p(\beta | Y, Z^\star, \Sigma)$ by

- sampling $\beta_{11}^\star$ from TN $\left[\frac{\sum_{i=1}^{100} Y_{i1} Z_{i1}^\star}{\sum_{i=1}^{100} Z_{i1}^{\star\,2}}, \frac{\sigma_1^2}{\sum_{i=1}^{100} Z_{i1}^{\star\,2}} \right] \Bigg|_{\beta_{11}^\star > 0}$,
- sampling $\beta_2^\star$ from $\mathrm{TN}_2 \left[\left(Z^\star Z^{\star T} \right)^{-1} Z^\star Y_{.2}, \left(Z^\star Z^{\star T} / \sigma_2^2 \right)^{-1} \right] \Bigg|_{\beta_{22}^\star > 0}$,
- sampling $\beta_j^\star$ from $\mathrm{N}_2 \left[\left(Z^\star Z^{\star T} \right)^{-1} Z^\star Y_{.j}, \left(Z^\star Z^{\star T} / \sigma_j^2 \right)^{-1} \right]$, for $j = 3, \ldots, 6$;

Then set $W_i = \beta^\star Z_i^\star$, for $i = 1, \ldots, 100$.

Step 4: Sample β from $p(\beta | Y, W, \Sigma)$ by

- sampling $\tilde{\beta}$ from $p(\tilde{\beta} | Y, W, \Sigma)$:
  first, sampling S from Inv-wishart(ν, S_0), where the degrees of freedom $\nu = 100 - 6 - 1$ and the (2×2) scale matrix $S_0 = \sum_{i=1}^{100} \tilde{W}_i \tilde{W}_i^T$; Then set $\tilde{\beta}$ as the Cholesky factor of S, i.e., $\tilde{\beta}$ is the unique (2×2) lower triangle matrix with positive diagonal elements satisfying $S = \tilde{\beta}\tilde{\beta}^T$,

- sampling β_j from $\mathrm{N}_2[m_j, V_j]$, for $j = 3, \ldots, 6$,

  where $V_j = \left[\frac{\tilde{\beta}^{-1}\left(\sum_{i=1}^{100} \tilde{W}_i \tilde{W}_i^T\right)(\tilde{\beta}^T)^{-1}}{\sigma_j^2}\right]^{-1}$, and $m_j = V_j \tilde{\beta}^{-1} \left[\begin{array}{c} \frac{\sum_{i=1}^{100} W_{i1} Y_{ij}}{\sigma_j^2} \\ \frac{\sum_{i=1}^{100} W_{i2} Y_{ij}}{\sigma_j^2} \end{array}\right]$;

  Finally set $Z_i = \tilde{\beta}^{-1} \tilde{W}_i$, for $i = 1, \ldots, 100$.

The steps of the sampler combining MH within PCG and ASIS, i.e., Sampler 3.2, are

Step j**:** Use MH to sample σ_j^2 from $p(\sigma_j^2 | Y, \sigma_{-j}^{2'}, \beta') \propto p(\beta, \Sigma | Y)$, for $j = 1, \ldots, 4$, where $\sigma_{-j}^{2'}$ denotes the (5×1) vector $(\sigma_1^2, \ldots, \sigma_{j-1}^2, \sigma_{j+1}^{2'}, \ldots, \sigma_6^{2'})$.

Step 5: Sample $Z_i^\star$ from $\mathrm{N}_2 \left\{ \left[I_2 + \beta'^T \left(\Sigma^{(4)}\right)^{-1} \beta'\right]^{-1} \beta'^T \left(\Sigma^{(4)}\right)^{-1} Y_i, \left[I_2 + \beta'^T \left(\Sigma^{(4)}\right)^{-1} \beta'\right]^{-1} \right\}$, for $i = 1, \ldots, 100$.

Step 6: Sample σ_j^2 from Inv-Gamma $\left[a + \frac{100}{2}, b + \frac{1}{2} \sum_{i=1}^{100} \left(Y_{ij} - \beta_j' Z_i^\star\right)^2\right]$, for $j = 5, 6$.

Step 7: Sample $\beta^\star$ from $p(\beta | Y, Z^\star, \Sigma)$ by

- sampling $\beta_{11}^\star$ from TN $\left.\left[\frac{\sum_{i=1}^{100} Y_{i1} Z_{i1}^\star}{\sum_{i=1}^{100} Z_{i1}^{\star\,2}}, \frac{\sigma_1^2}{\sum_{i=1}^{100} Z_{i1}^{\star\,2}}\right]\right|_{\beta_{11}^\star > 0}$,
- sampling $\beta_2^\star$ from $\mathrm{TN}_2 \left.\left[\left(Z^\star Z^{\star T}\right)^{-1} Z^\star Y_{.2}, \left(Z^\star Z^{\star T} / \sigma_2^2\right)^{-1}\right]\right|_{\beta_{22}^\star > 0}$,
- sampling $\beta_j^\star$ from $\mathrm{N}_2 \left[\left(Z^\star Z^{\star T}\right)^{-1} Z^\star Y_{.j}, \left(Z^\star Z^{\star T} / \sigma_j^2\right)^{-1}\right]$, for $j = 3, \ldots, 6$;

Then set $W_i = \beta^\star Z_i^\star$, for $i = 1, \ldots, 100$.

Step 8: Sample β from $p(\beta | Y, W, \Sigma)$ by

- sampling $\tilde{\beta}$ from $p(\tilde{\beta} | Y, W, \Sigma)$:

  first, sampling S from Inv-wishart(ν, S_0); Then set $\tilde{\beta}$ as the Cholesky factor of S,
- sampling β_j from $\mathrm{N}_2[m_j, V_j]$, for $j = 3, \ldots, 6$;

  Finally set $Z_i = \tilde{\beta}^{-1} \tilde{W}_i$, for $i = 1, \ldots, 100$.

B.3 Details of the ASIS and combined samplers for fitting the cosmological hierarchical model in (2.14)–(2.16)

Setting $\bar{X} = AX + L$, the joint distribution of $(\xi, \bar{X}, \mathscr{C}, \alpha, \beta, \Sigma_P^2)$ is

$$\begin{aligned}p(\xi,\bar{X},\Omega_m,\Omega_\Lambda,\alpha,\beta,\Sigma_P^2|Y) \;&\propto |\Sigma_C|^{-1/2}|\Sigma_P|^{-1/2}|\Sigma_0|^{-1/2}\frac{1}{R_c^2}\frac{1}{R_x^2}\frac{1}{\sigma^2_{\text{res}}}\\ &\exp\biggl\{-\frac{1}{2}\Bigl[(Y-\bar{X})^T\Sigma_C^{-1}(Y-\bar{X})\\ &+(A^{-1}\bar{X}-A^{-1}L-J\xi)^T\Sigma_P^{-1}(A^{-1}\bar{X}-A^{-1}L-J\xi)\\ &+(\xi-\xi_m)^T\Sigma_0^{-1}(\xi-\xi_m)\Bigr]\biggr\}.\end{aligned} \tag{B.4}$$

The steps of the ASIS sampler, i.e., Sampler 3.3, are

Step 1: Sample $(\xi, X^\star)$, which consists of two sub-steps:

- Sample ξ from $\mathrm{N}(k_0, K)$;
- Sample $X^\star$ from $\mathrm{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1}J\xi)$.

Step 2: Use MH to sample $\mathscr{C}^\star$ from $p(\mathscr{C}|Y,\xi,X,\alpha,\beta,\Sigma_P)$, which is proportional to $p(\xi,X,\mathscr{C},\alpha,\beta,\Sigma_P|Y)$, under the constraint $\mathscr{C}^\star = (\Omega_m^\star,\Omega_\Lambda^\star) \in [0,1]\times[0,2]$;

Use $\mathscr{C}^\star$ to construct $L^\star$.

Step 3: Sample $(\alpha^\star,\beta^\star)$ from $\mathrm{N}(\mu_B,\Sigma_B)$ with constraint $(\alpha^\star,\beta^\star)\in[0,1]\times[0,4]$;

Use $(\alpha^\star,\beta^\star)$ to construct $A^\star$. Then set $\bar{X} = A^\star X^\star + L^\star$.

Step 4: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y,\xi,\bar{X},\alpha^\star,\beta^\star,\Sigma_P)$, which is proportional to $p(\xi,\bar{X},\mathscr{C},\alpha^\star,\beta^\star,\Sigma_P|Y)$ under the constraint $\mathscr{C} = (\Omega_m,\Omega_\Lambda)\in[0,1]\times[0,2]$;

Use $\mathscr{C}$ to construct L.

Step 5: Sample (α,β) from $\mathrm{N}(\mu_D,\Sigma_D)$ with constraint $(\alpha,\beta)\in[0,1]\times[0,4]$, where

$$\Sigma_D^{-1} = \begin{bmatrix} \sum\limits_{i=1}^n \frac{\bar{x}_i^2}{\sigma^2_{\text{res}}} & \sum\limits_{i=1}^n \frac{-\bar{x}_i\bar{c}_i}{\sigma^2_{\text{res}}} \\ \sum\limits_{i=1}^n \frac{-\bar{x}_i\bar{c}_i}{\sigma^2_{\text{res}}} & \sum\limits_{i=1}^n \frac{\bar{c}_i^2}{\sigma^2_{\text{res}}} \end{bmatrix} \text{ and } \mu_D = \Sigma_D \begin{bmatrix} \sum_{i=1}^n \frac{\bar{x}_i(M_0-\bar{M}_i)}{\sigma^2_{\text{res}}} \\ \sum_{i=1}^n \frac{-\bar{c}_i(M_0-\bar{M}_i)}{\sigma^2_{\text{res}}} \end{bmatrix},$$

where $\bar{c}_i$, $\bar{x}_i$, and $\bar{M}_i$ are the $(3i-2)^{\text{th}}$, $(3i-1)^{\text{th}}$, and $(3i)^{\text{th}}$ components of $(\bar{X}-L)$;

Use (α,β) to construct A. Then set $X = A^{-1}(\bar{X}-L)$.

Step 6: Sample Σ_P, which consists of three sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (c_i-c_0)^2}{2}\right]$ with $\log(R_c)\in[-5,2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (x_i-x_0)^2}{2}\right]$ with $\log(R_x)\in[-5,2]$.

- Sample σ^2_{res} from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (M_i - M_0)^2}{2}\right]$ with $\log(\sigma_{\text{res}}) \in [-5, 2]$.

The steps of the sampler combining MH within PCG and ASIS, i.e., Sampler 3.4, are

Step 1: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y, \alpha, \beta, \Sigma_P)$, which is proportional to $p(\mathscr{C}|\alpha, \beta, \Sigma_P|Y)$ with $\mathscr{C} = (\Omega_m, \Omega_\Lambda) \in [0,1] \times [0,2]$;

Use $\mathscr{C}$ to construct L.

Step 2: Sample $(\xi, X^\star)$, which consists of two sub-steps:

- Sample ξ from $\text{N}(k_0, K)$;
- Sample $X^\star$ from $\text{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1} J\xi)$.

Step 3: Sample $(\alpha^\star, \beta^\star)$ from $\text{N}(\mu_B, \Sigma_B)$ with constraint $(\alpha^\star, \beta^\star) \in [0,1] \times [0,4]$;

Use $(\alpha^\star, \beta^\star)$ to construct $A^\star$. Then set $\bar{X} = A^\star X^\star + L$.

Step 4: Sample (α, β) from $\text{N}(\mu_D, \Sigma_D)$ with constraint $(\alpha, \beta) \in [0,1] \times [0,4]$;

Use (α, β) to construct A. Then set $X = A^{-1}(\bar{X} - L)$.

Step 5: Sample Σ_P, which consists of three sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (c_i - c_0)^2}{2}\right]$ with $\log(R_c) \in [-5, 2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (x_i - x_0)^2}{2}\right]$ with $\log(R_x) \in [-5, 2]$.
- Sample σ^2_{res} from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (M_i - M_0)^2}{2}\right]$ with $\log(\sigma_{\text{res}}) \in [-5, 2]$.

B.4 Details of the Gibbs-type samplers for fitting the hierarchical t model in (3.12) and (3.13)

With the prior $p(\mu, \sigma, \tau) \propto 1$, the joint posterior distribution of Z, σ^2, β, τ, and μ under the hierarchical t model in (3.12) and (3.13) is

$$\begin{aligned} p(Z, \sigma^2, \beta, \tau, \mu|Y) \;\propto\; & \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^n Z_i(Y_i - \beta_i X_i)^2}{\sigma^2} + \nu \sum_{i=1}^n Z_i + \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{\tau^2}\right]\right\} \\ & (\sigma^2)^{-(n+1)/2} \tau^{-n} \left(\prod_{i=1}^n Z_i^{(\nu-1)/2}\right). \end{aligned} \tag{B.5}$$

Thus the steps of the Gibbs sampler, i.e., Sampler 3.5, are

Step 1: Sample Z_i from Gamma $\left[\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(Y_i - \beta_i' X_i)^2}{2(\sigma^2)'}\right]$, for $i = 1, \ldots, n$.

Step 2: Sample σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n Z_i(Y_i-\beta_i' X_i)^2}{2}\right]$.

Step 3: Sample β_i from N $\left[\left(\frac{Z_i Y_i X_i}{\sigma^2}+\frac{\mu'}{(\tau')^2}\right)/\left(\frac{Z_i X_i^2}{\sigma^2}+\frac{1}{(\tau')^2}\right), 1/\left(\frac{Z_i X_i^2}{\sigma^2}+\frac{1}{(\tau')^2}\right)\right]$, for $i = 1, \ldots, n$.

Step 4: Sample (τ, μ) from $p(\tau, \mu | Y, Z, \beta, \sigma^2)$ by

- sampling τ^2 from Inv-Gamma $\left\{\frac{n}{2}-1, \left[\sum_{i=1}^n \beta_i^2 - \frac{(\sum_{i=1}^n \beta_i)^2}{n}\right]/2\right\}$; $\tau = \sqrt{\tau^2}$,
- sampling μ from N $\left(\frac{\sum_{i=1}^n \beta_i}{n}, \frac{\tau^2}{n}\right)$.

Setting $\tilde{Z} = \alpha Z$, and specifying the Haar measure prior to α as $p_\infty(\alpha) \propto 1/\alpha$, we obtain the joint posterior distribution of $\tilde{Z}$, α, σ^2, β, τ and μ, that is,

$$\begin{aligned} \tilde{p}(\tilde{Z}, \alpha, \sigma^2, \beta, \tau, \mu | Y) \quad \propto \quad & \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^n \tilde{Z}_i (Y_i - \beta_i X_i)^2}{\alpha\sigma^2} + \frac{\nu \sum_{i=1}^n \tilde{Z}_i}{\alpha} + \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{\tau^2}\right]\right\} \\ & (\sigma^2)^{-(n+1)/2} \tau^{-n} \alpha^{-(n\nu+n)/2-1} \left(\prod_{i=1}^n \tilde{Z}_i^{(\nu-1)/2}\right). \end{aligned} \tag{B.6}$$

The steps of the Haar PX-DA sampler, i.e., Sampler 3.6, are

Step 1: Sample $Z_i^\star$ from Gamma $\left[\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(Y_i - \beta_i' X_i)^2}{2(\sigma^2)'}\right]$, for $i = 1, \ldots, n$.

Step 2: Sample α from Inv-Gamma $\left(\frac{n\nu+1}{2}, \frac{\nu \sum_{i=1}^n Z_i^\star}{2}\right)$; Set $Z = Z^\star/\alpha$.

Step 3: Sample σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n Z_i(Y_i-\beta_i' X_i)^2}{2}\right]$.

Step 4: Sample β_i from N $\left[\left(\frac{Z_i Y_i X_i}{\sigma^2}+\frac{\mu'}{(\tau')^2}\right)/\left(\frac{Z_i X_i^2}{\sigma^2}+\frac{1}{(\tau')^2}\right), 1/\left(\frac{Z_i X_i^2}{\sigma^2}+\frac{1}{(\tau')^2}\right)\right]$, for $i = 1, \ldots, n$.

Step 5: Sample (τ, μ) from $p(\tau, \mu | Y, Z, \sigma^2, \beta)$ by

- sampling τ^2 from Inv-Gamma $\left\{\frac{n}{2}-1, \left[\sum_{i=1}^n \beta_i^2 - \frac{(\sum_{i=1}^n \beta_i)^2}{n}\right]/2\right\}$; $\tau = \sqrt{\tau^2}$,
- sampling μ from N $\left(\frac{\sum_{i=1}^n \beta_i}{n}, \frac{\tau^2}{n}\right)$.

Furthermore, setting $\bar{\beta} = (\beta - \mu)/\tau$, we obtain the joint posterior distribution of Z, σ^2, $\bar{\beta}$, τ, and μ, that is,

$$\begin{aligned} p(Z, \sigma^2, \bar{\beta}, \tau, \mu | Y) \quad \propto \quad & \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^n Z_i (Y_i - \tau\bar{\beta}_i X_i - \mu X_i)^2}{\sigma^2} + \nu \sum_{i=1}^n Z_i + \sum_{i=1}^n \bar{\beta}_i^2\right]\right\} \\ & (\sigma^2)^{-(\mathrm{n}+1)/2} \left(\prod_{i=1}^n Z_i^{(\nu-1)/2}\right). \end{aligned} \tag{B.7}$$

The steps of the ASIS sampler, i.e., Sampler 3.7, are

Step 1: Sample Z_i from Gamma $\left[\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(Y_i-\beta_i' X_i)^2}{2(\sigma^2)'}\right]$, for $i = 1, \ldots, n$.

Step 2: Sample σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n Z_i(Y_i-\beta_i' X_i)^2}{2}\right]$.

Step 3: Sample $\beta_i^\star$ from N $\left[\left(\frac{Z_i Y_i X_i}{\sigma^2} + \frac{\mu'}{(\tau')^2}\right) / \left(\frac{Z_i X_i^2}{\sigma^2} + \frac{1}{(\tau')^2}\right), 1/\left(\frac{Z_i X_i^2}{\sigma^2} + \frac{1}{(\tau')^2}\right)\right]$, for $i = 1, \ldots, n$.

Step 4: Sample $(\tau^\star, \mu^\star)$ from $p(\tau, \mu | Y, Z, \beta^\star, \sigma^2)$ by

- sampling $(\tau^\star)^2$ from Inv-Gamma $\left\{\frac{n}{2} - 1, \left[\sum_{i=1}^n (\beta_i^\star)^2 - \frac{(\sum_{i=1}^n \beta_i^\star)^2}{n}\right]/2\right\}$; $\tau^\star = \sqrt{(\tau^\star)^2}$,
- sampling $\mu^\star$ from N $\left[\frac{\sum_{i=1}^n \beta_i^\star}{n}, \frac{(\tau^\star)^2}{n}\right]$;

Set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$.

Step 5: Sample (τ, μ) from $p(\tau, \mu | Y, Z, \bar{\beta}, \sigma^2)$ by

- sampling τ from $\mathrm{TN}(m_\tau, V_\tau)\big|_{\tau>0}$, where

$$m_\tau = \frac{\sum_{i=1}^n Z_i Y_i \bar{\beta}_i X_i - (\sum_{i=1}^n Z_i Y_i X_i)(\sum_{i=1}^n Z_i \bar{\beta}_i X_i^2)/(\sum_{i=1}^n Z_i X_i^2)}{\sum_{i=1}^n Z_i \bar{\beta}_i^2 X_i^2 - (\sum_{i=1}^n Z_i \bar{\beta}_i X_i^2)^2/(\sum_{i=1}^n Z_i X_i^2)}$$

and

$$V_\tau = \frac{\sigma^2}{\sum_{i=1}^n Z_i \bar{\beta}_i^2 X_i^2 - (\sum_{i=1}^n Z_i \bar{\beta}_i X_i^2)^2/(\sum_{i=1}^n Z_i X_i^2)},$$

- sampling μ from N $\left[\frac{\sum_{i=1}^n Z_i(Y_i - \tau\bar{\beta}_i)X_i}{\sum_{i=1}^n Z_i X_i^2}, \frac{\sigma^2}{\sum_{i=1}^n Z_i X_i^2}\right]$;

Set $\beta = \tau\tilde{\beta} + \mu$.

Finally, the steps of the sampler combining Haar PX-DA and ASIS, Sampler 3.8, are

Step 1: Sample $Z_i^\star$ from Gamma $\left[\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(Y_i-\beta_i' X_i)^2}{2(\sigma^2)'}\right]$, for $i = 1, \ldots, n$.

Step 2: Sample α from Inv-Gamma $\left(\frac{n\nu+1}{2}, \frac{\nu\sum_{i=1}^n Z_i^\star}{2}\right)$; Set $Z = Z^\star/\alpha$.

Step 3: Sample σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n Z_i(Y_i-\beta_i' X_i)^2}{2}\right]$.

Step 4: Sample $\beta_i^\star$ from N $\left[\left(\frac{Z_i Y_i X_i}{\sigma^2} + \frac{\mu'}{(\tau')^2}\right) / \left(\frac{Z_i X_i^2}{\sigma^2} + \frac{1}{(\tau')^2}\right), 1/\left(\frac{Z_i X_i^2}{\sigma^2} + \frac{1}{(\tau')^2}\right)\right]$, for $i = 1, \ldots, n$.

Step 5: Sample $(\tau^\star, \mu^\star)$ from $p(\tau, \mu | Y, Z, \beta^\star, \sigma^2)$ by

- sampling $(\tau^\star)^2$ from Inv-Gamma $\left\{\frac{n}{2} - 1, \left[\sum_{i=1}^n (\beta_i^\star)^2 - \frac{(\sum_{i=1}^n \beta_i^\star)^2}{n}\right]/2\right\}$; $\tau^\star = \sqrt{(\tau^\star)^2}$,
- sampling $\mu^\star$ from N $\left[\frac{\sum_{i=1}^n \beta_i^\star}{n}, \frac{(\tau^\star)^2}{n}\right]$;

Set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$.

Step 6: Sample (τ, μ) from $p(\tau, \mu | Y, Z, \bar{\beta}, \sigma^2)$ by

- sampling τ from $\mathrm{TN}(m_\tau, V_\tau)\big|_{\tau>0}$,
- sampling μ from N $\left[\frac{\sum_{i=1}^n Z_i(Y_i - \tau\bar{\beta}_i)X_i}{\sum_{i=1}^n Z_i X_i^2}, \frac{\sigma^2}{\sum_{i=1}^n Z_i X_i^2}\right]$;

Set $\beta = \tau\tilde{\beta} + \mu$.

B.5 Details of the steps of the Gibbs-type samplers for fitting the hierarchical probit model in (3.16)

With the prior distributions $\mu \sim \mathrm{N}(0, V)$ and $p(\tau) \propto 1$, the joint posterior distribution of Z, β, μ and τ under the hierarchical probit model (3.16) is

$$p(Z, \beta, \tau, \mu | Y) \propto \frac{1}{\tau^n}\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n (Z_i - \beta_i X_i)^2 + \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{\tau^2} + \frac{\mu^2}{V}\right]\right\}\left[\prod_{i=1}^n F_i(\beta_i; Y_i)\right], \tag{B.8}$$

where $F_i(\beta_i; Y_i) = \begin{cases} 1/\Phi(\beta_i X_i), & \text{if } Y_i = 1, \\ 1/\Phi(-\beta_i X_i), & \text{otherwise,} \end{cases}$ with $\Phi(\cdot)$ denoting the cumulative distribution function of the standard normal distribution.

Thus the steps of the Gibbs sampler, i.e., Sampler 3.9, are

Step 1: Sample Z_i from $\mathrm{TN}(\beta_i' X_i, 1)|_{Y_i}$, for $i = 1, \ldots, n$,

where Z_i is constrained by $Z_i > 0$ if $Y_i = 1$, and $Z_i \leq 0$ otherwise.

Step 2: Sample (μ, β) from $p(\mu, \beta | Y, Z, \tau')$ by

- sampling μ from N $\left\{\frac{\sum_{i=1}^n Z_i X_i/[(\tau')^2 X_i^2 + 1]}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2 + 1] + 1/V}, \frac{1}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2 + 1] + 1/V}\right\}$,
- sampling β_i from N $\left[\frac{Z_i X_i + \mu/(\tau')^2}{X_i^2 + 1/(\tau')^2}, \frac{1}{X_i^2 + 1/(\tau')^2}\right]$, for $i = 1, \ldots, n$.

Step 3: Sample τ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{2}\right]$; $\tau = \sqrt{\tau^2}$.

Setting $\tilde{Z} = \alpha Z$, $\tilde{\beta} = \alpha\beta$, and $\tilde{\mu} = \alpha\mu$, and specifying the Haar measure prior to α as $p_\infty(\alpha) \propto 1/\alpha$, we obtain the joint posterior distribution of $\tilde{Z}$, α, $\tilde{\beta}$, $\tilde{\mu}$, and τ, that is,

$$\begin{aligned} \tilde{p}(\tilde{Z}, \alpha, \tilde{\beta}, \tilde{\mu}, \tau|Y) \;\; \propto \;\; & \exp\left\{-\tfrac{1}{2}\left[\tfrac{\sum_{i=1}^n (\tilde{Z}_i - \tilde{\beta}_i X_i)^2}{\alpha^2} + \tfrac{\sum_{i=1}^n (\tilde{\beta}_i - \tilde{\mu})^2}{\alpha^2\tau^2} + \tfrac{\tilde{\mu}^2}{\alpha^2 V}\right]\right\} \\ & \tau^{-n}\alpha^{-2(n+1)}\left[\textstyle\prod_{i=1}^n \tilde{F}_i(\tilde{\beta}_i, \alpha; Y_i)\right], \end{aligned} \tag{B.9}$$

where $\tilde{F}_i(\tilde{\beta}_i, \alpha; Y_i) = \begin{cases} 1/\Phi\left(\frac{\tilde{\beta}_i}{\alpha}X_i\right), & \text{if } Y_i = 1, \\ 1/\Phi\left(-\frac{\tilde{\beta}_i}{\alpha}X_i\right), & \text{otherwise.} \end{cases}$

The steps of the Haar PX-DA sampler, i.e., Sampler 3.10, are

Step 1: Sample $Z_i^\star$ from $\text{TN}(\beta_i' X_i, 1)|_{Y_i}$, for $i = 1, \ldots, n$.

Step 2: Sample α^2 from Inv-Gamma $\left\{\frac{n}{2}, \left[\sum_{i=1}^n \frac{(Z_i^\star)^2}{(\tau')^2 X_i^2 + 1} - \frac{\left(\sum_{i=1}^n \frac{Z_i^\star X_i}{(\tau')^2 X_i^2+1}\right)^2}{\sum_{i=1}^n \frac{X_i^2}{(\tau')^2 X_i^2+1} + \frac{1}{V}}\right] \Big/ 2\right\}$;

Set $\alpha = \sqrt{\alpha^2}$ and $Z = Z^\star/\alpha$.

Step 3: Sample (μ, β) from $p(\mu, \beta|Y, Z, \tau')$ by

- sampling μ from $\text{N}\left\{\frac{\sum_{i=1}^n Z_i X_i/[(\tau')^2 X_i^2+1]}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}, \frac{1}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}\right\}$,
- sampling β_i from $\text{N}\left[\frac{Z_i X_i + \mu/(\tau')^2}{X_i^2 + 1/(\tau')^2}, \frac{1}{X_i^2 + 1/(\tau')^2}\right]$, for $i = 1, \ldots, n$.

Step 4: Sample τ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{2}\right]$; $\tau = \sqrt{\tau^2}$.

Integrating (B.8) over β, we obtain the marginal distribution of Z, μ, and τ, that is,

$$p(Z, \mu, \tau|Y) \propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(Z_i - \mu X_i)^2}{\tau^2 X_i^2 + 1} + \frac{\mu^2}{V}\right]\right\} \prod_{i=1}^n \left[(\tau^2 X_i^2 + 1)^{-1/2} F_i(\tau, \mu; Y_i)\right], \tag{B.10}$$

where $F_i(\mu, \tau; Y_i) = \begin{cases} 1/\Phi\left(\mu X_i/\sqrt{\tau^2 X_i^2 + 1}\right), & \text{if } Y_i = 1, \\ 1/\Phi\left(-\mu X_i/\sqrt{\tau^2 X_i^2 + 1}\right), & \text{otherwise.} \end{cases}$

Furthermore, setting $\bar{Z} = Z - \mu X$, we obtain the joint distribution of $\bar{Z}$, μ, and τ, i.e.,

$$p(\bar{Z}, \tau, \mu|Y) \propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{\bar{Z}_i^2}{\tau^2 X_i^2 + 1} + \frac{\mu^2}{V}\right]\right\} \prod_{i=1}^n \left[(\tau^2 X_i^2 + 1)^{-1/2} F_i(\tau, \mu; Y_i)\right]. \tag{B.11}$$

Thus the steps of the sampler using ASIS I, i.e., Sampler 3.11, are

Step 1: Sample $Z_i^\star$ from TN$\left\{\mu' X_i, \left[(\tau')^2 X_i^2 + 1\right]\right\}|_{Y_i}$, for $i = 1, \dots, n$.

Step 2: Sample $\mu^\star$ from N$\left\{\frac{\sum_{i=1}^n Z_i^\star X_i/[(\tau')^2 X_i^2+1]}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}, \frac{1}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}\right\}$;

Set $\bar{Z} = Z^\star - \mu^\star X$.

Step 3: Sample μ from TN$(0, V)|_{F_\mu}$, where F_μ is the constraint $\{\mu : \ \bar{Z}_i + \mu X_i > 0 \text{ if } Y_i = 1;\ \bar{Z}_i + \mu X_i \leq 0 \text{ otherwise, for } i = 1, \dots, n\}$;

Set $Z = \bar{Z} + \mu X$.

Step 4: Sample β_i from N$\left[\frac{Z_i X_i + \mu/(\tau')^2}{X_i^2 + 1/(\tau')^2}, \frac{1}{X_i^2+1/(\tau')^2}\right]$, for $i = 1, \dots, n$.

Step 5: Sample τ^2 from Inv-Gamma$\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{2}\right]$; $\tau = \sqrt{\tau^2}$.

Furthermore, setting $\bar{\beta} = (\beta - \mu)/\tau$, we get the joint posterior distribution of Z, $\bar{\beta}$, μ, and τ, that is,

$$p(Z, \bar{\beta}, \mu, \tau | Y) \propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n (Z_i - \tau\bar{\beta}_i X_i - \mu X_i)^2 + \sum_{i=1}^n \bar{\beta}_i^2 + \frac{\mu^2}{V}\right]\right\}\left[\prod_{i=1}^n \bar{F}_i(\bar{\beta}_i, \mu, \tau; Y_i)\right], \tag{B.12}$$

where $\bar{F}_i(\bar{\beta}_i, \mu, \tau; Y_i) = \begin{cases} 1/\Phi\left[(\tau\bar{\beta}_i + \mu)X_i\right], & \text{if } Y_i = 1, \\ 1/\Phi\left[-(\tau\bar{\beta}_i + \mu)X_i\right], & \text{otherwise.} \end{cases}$

Thus the steps of the sampler using ASIS II, i.e., Sampler 3.12, are

Step 1: Sample Z_i from TN$(\beta_i' X_i, 1)|_{Y_i}$, for $i = 1, \dots, n$.

Step 2: Sample $(\mu^\star, \beta^\star)$ from $p(\mu, \beta | Y, Z, \tau')$ by

- sampling $\mu^\star$ from N$\left\{\frac{\sum_{i=1}^n Z_i X_i/[(\tau')^2 X_i^2+1]}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}, \frac{1}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}\right\}$,
- sampling $\beta_i^\star$ from N$\left[\frac{Z_i X_i + \mu^\star/(\tau')^2}{X_i^2+1/(\tau')^2}, \frac{1}{X_i^2+1/(\tau')^2}\right]$, for $i = 1, \dots, n$.

Step 3: Sample $(\tau^\star)^2$ from Inv-Gamma$\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n (\beta_i - \mu)^2}{2}\right]$; $\tau^\star = \sqrt{(\tau^\star)^2}$;

Set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$.

Step 4: Sample μ from N$\left\{\left[\sum_{i=1}^n X_i(Z_i - \tau^\star \bar{\beta}_i X_i)\right] / \left(\sum_{i=1}^n X_i^2 + \frac{1}{V}\right), 1/\left(\sum_{i=1}^n X_i^2 + \frac{1}{V}\right)\right\}$.

Step 5: Sample τ from TN$\left\{\left[\sum_{i=1}^n (Z_i - X_i\mu) X_i \bar{\beta}_i\right] / \left(\sum_{i=1}^n X_i^2 \bar{\beta}_i^2\right), 1/\left(\sum_{i=1}^n X_i^2 \bar{\beta}_i^2\right)\right\}|_{\tau>0}$;

Set $\beta = \tau\tilde{\beta} + \mu$.

The steps of the sampler combining Haar PX-DA and ASIS II, Sampler 3.13, are

Step 1: Sample $Z_i^\star$ from TN$(\beta_i' X_i, 1)|_{Y_i}$, for $i = 1, \dots, n$.

Step 2: Sample α^2 from Inv-Gamma $\left\{\frac{n}{2}, \left[\sum_{i=1}^n \frac{(Z_i^\star)^2}{(\tau')^2 X_i^2+1} - \frac{\left(\sum_{i=1}^n \frac{Z_i^\star X_i}{(\tau')^2 X_i^2+1}\right)^2}{\sum_{i=1}^n \frac{X_i^2}{(\tau')^2 X_i^2+1}+\frac{1}{V}}\right] \Big/ 2\right\}$;

Set $\alpha = \sqrt{\alpha^2}$ and $Z = Z^\star/\alpha$.

Step 3: Sample $(\mu^\star, \beta^\star)$ from $p(\mu, \beta|Y, Z, \tau')$ by

- sampling $\mu^\star$ from N $\left\{\frac{\sum_{i=1}^n Z_i X_i/[(\tau')^2 X_i^2+1]}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}, \frac{1}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}\right\}$,
- sampling $\beta_i^\star$ from N $\left[\frac{Z_i X_i+\mu^\star/(\tau')^2}{X_i^2+1/(\tau')^2}, \frac{1}{X_i^2+1/(\tau')^2}\right]$, for $i = 1, \ldots, n$.

Step 4: Sample $(\tau^\star)^2$ from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n (\beta_i-\mu)^2}{2}\right]$; $\tau^\star = \sqrt{(\tau^\star)^2}$;

Set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$.

Step 5: Sample μ from N $\left\{\left[\sum_{i=1}^n X_i(Z_i - \tau^\star \bar{\beta}_i X_i)\right] / \left(\sum_{i=1}^n X_i^2 + \frac{1}{V}\right), 1/\left(\sum_{i=1}^n X_i^2 + \frac{1}{V}\right)\right\}$.

Step 6: Sample τ from TN $\left\{\left[\sum_{i=1}^n (Z_i - X_i\mu) X_i \bar{\beta}_i\right] / \left(\sum_{i=1}^n X_i^2 \bar{\beta}_i^2\right), 1/\left(\sum_{i=1}^n X_i^2 \bar{\beta}_i^2\right)\right\}\Big|_{\tau>0}$;

Set $\beta = \tau\tilde{\beta} + \mu$.

Finally, the steps of the sampler combining ASIS I and ASIS II, Sampler 3.14, are

Step 1: Sample $Z_i^\star$ from TN $\left\{\mu' X_i, \left[(\tau')^2 X_i^2 + 1\right]\right\}\big|_{Y_i}$, for $i = 1, \ldots, n$.

Step 2: Sample $\mu^\star$ from N $\left\{\frac{\sum_{i=1}^n Z_i^\star X_i/[(\tau')^2 X_i^2+1]}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}, \frac{1}{\sum_{i=1}^n X_i^2/[(\tau')^2 X_i^2+1]+1/V}\right\}$;

Set $\bar{Z} = Z^\star - \mu^\star X$.

Step 3: Sample μ from $\text{TN}(0, V)|_{F_\mu}$, where F_μ is the constraint $\{\mu : \bar{Z}_i + \mu X_i > 0 \text{ if } Y_i = 1;\ \bar{Z}_i + \mu X_i \leq 0 \text{ otherwise, for } i = 1, \ldots, n\}$;

Set $Z = \bar{Z} + \mu X$.

Step 4: Sample $\beta_i^\star$ from N $\left[\frac{Z_i X_i+\mu/(\tau')^2}{X_i^2+1/(\tau')^2}, \frac{1}{X_i^2+1/(\tau')^2}\right]$, for $i = 1, \ldots, n$.

Step 5: Sample $(\tau^\star)^2$ from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n (\beta_i-\mu)^2}{2}\right]$; $\tau^\star = \sqrt{(\tau^\star)^2}$;

Set $\bar{\beta} = (\beta^\star - \mu^\star)/\tau^\star$.

Step 6: Sample τ from TN $\left\{\left[\sum_{i=1}^n (Z_i - X_i\mu) X_i \bar{\beta}_i\right] / \left(\sum_{i=1}^n X_i^2 \bar{\beta}_i^2\right), 1/\left(\sum_{i=1}^n X_i^2 \bar{\beta}_i^2\right)\right\}\Big|_{\tau>0}$;

Set $\beta = \tau\tilde{\beta} + \mu$.

C

Derivation of the cyclic-permutation bound of Sampler 4.2 and Details of the Gibbs-type Samplers in Chapter 4

C.1 Deriving the cyclic-permutation bound of Sampler 4.2 with general target and surrogate distribution

Recall that the cyclic-permutation bound of an N-step Gibbs-type sampler is defined by $\min_{j\in\{0,\ldots,N-1\}}\{\gamma_j\}$, where γ_j is the norm of the forward operator corresponding to the j-step-lagged sampler, see Section 1.3 of Chapter 1. Henceforth, to derive the cyclic-permutation bound of Sampler 4.2 in Section 4.1, we first compute the norms of the forward operators corresponding to its 0-step-lagged and 1-step-lagged samplers, i.e., γ_0 and γ_1.

The 0-step-lagged sampler is Sampler 4.2 itself, and its stationary distribution is the target, i.e., $p(\psi_1, \psi_2)$.

Then $\forall h \in L_0^2(p)$ with $\text{var}_p(h) = 1$, since $p(\psi_1, \psi_2)$ and $p_s(\psi_1, \psi_2)$ have the same marginal distributions, we have

$$\begin{aligned}
&\mathrm{var}_p\left[\mathrm{E}(h(\psi_1^{(t+1)},\psi_2^{(t+1)})|\psi_2^{(t)})\right] = \mathrm{E}_p\left[\mathrm{E}^2(h(\psi_1^{(t+1)},\psi_2^{(t+1)})|\psi_2^{(t)})\right]\\
= \;& \int\left[\int h(\psi_1^{(t+1)},\psi_2^{(t+1)})p_s(\psi_1^{(t+1)}|\psi_2^{(t)})p(\psi_2^{(t+1)}|\psi_1^{(t+1)})\mathrm{d}\psi_1^{(t+1)}\mathrm{d}\psi_2^{(t+1)}\right]^2 p(\psi_2^{(t)})\mathrm{d}\psi_2^{(t)}\\
= \;& \mathrm{var}_{p_s}(h^\star)\int\left[\int \frac{h^\star(\psi_1^{(t+1)})}{\sqrt{\mathrm{var}_{p_s}(h^\star)}}p_s(\psi_1^{(t+1)}|\psi_2^{(t)})\mathrm{d}\psi_1^{(t+1)}\right]^2 p_s(\psi_2^{(t)})\mathrm{d}\psi_2^{(t)}\\
\leq \;& \rho_{p_s}^2\mathrm{var}_{p_s}(h^\star) \leq \rho_{p_s}^2.
\end{aligned} \tag{C.1}$$

where ρ_{p_s} is the maximum correlation between ψ_1 and ψ_2 for $p_s(\psi_1,\psi_2)$; $h^\star(\psi_1) = \int h(\psi_1,\psi_2)p(\psi_2|\psi_1)\mathrm{d}\psi_2$; $h^\star \in L_0^2(p_{s,1})$ with $p_{s,1}$ the marginal distribution of ψ_1 for $p_s(\psi_1,\psi_2)$, and $\mathrm{var}_{p_{s,1}}(h^\star) \leq 1$, because

i) $\mathrm{E}_{p_{s,1}}(h^\star) = \int[\int h(\psi_1,\psi_2)p(\psi_2|\psi_1)\mathrm{d}\psi_2]p_s(\psi_1)\mathrm{d}\psi_1 = \mathrm{E}_p(h) = 0$;

ii) $\mathrm{var}_{p_{s,1}}(h^\star) = \int[\int h(\psi_1,\psi_2)p(\psi_2|\psi_1)\mathrm{d}\psi_2]^2 p_s(\psi_1)\mathrm{d}\psi_1 \leq \mathrm{var}_p(h) = 1$.

From (C.1), we conclude that $\gamma_0 \leq \rho_{p_s}$.

Furthermore, $\forall g \in L_0^2(p_{s,1})$ with $\mathrm{var}_{p_{s,1}}(g) = 1$, it also holds that $g \in L_0^2(p)$ with $\mathrm{var}_p(g) = 1$, because

i) $\mathrm{E}_p(g) = \int g(\psi_1)p(\psi_1,\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2 = \int g(\psi_1)p_s(\psi_1)\mathrm{d}\psi_1 = \mathrm{E}_{p_{s,1}}(g) = 0$;

ii) $\mathrm{var}_p(g) = \int g^2(\psi_1)p(\psi_1,\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2 = \int g^2(\psi_1)p_s(\psi_1)\mathrm{d}\psi_1 = \mathrm{var}_{p_{s,1}}(g) = 1$.

Thus, in addition, we have

$$\begin{aligned}
\gamma_0^2 \;\geq\;& \mathrm{var}_p\left[\mathrm{E}(g(\psi_1^{(t+1)})|\psi_2^{(t)})\right] = \mathrm{E}_p\left[\mathrm{E}^2(g(\psi_1^{(t+1)})|\psi_2^{(t)})\right]\\
=\;& \int\left[\int g(\psi_1^{(t+1)})p_s(\psi_1^{(t+1)}|\psi_2^{(t)})p(\psi_2^{(t+1)}|\psi_1^{(t+1)})\mathrm{d}\psi_1^{(t+1)}\mathrm{d}\psi_2^{(t+1)}\right]^2 p(\psi_2^{(t)})\mathrm{d}\psi_2^{(t)}\\
=\;& \int\left[\int g(\psi_1^{(t+1)})p_s(\psi_1^{(t+1)}|\psi_2^{(t)})\mathrm{d}\psi_1^{(t+1)}\right]^2 p_s(\psi_2^{(t)})\mathrm{d}\psi_2^{(t)}\\
=\;& \mathrm{var}_{p_s}(\mathrm{E}_{p_s}(g(\psi_1^{(t+1)})|\psi_2^{(t)})).
\end{aligned} \tag{C.2}$$

From (C.2), we conclude that $\gamma_0 \geq \rho_{p_s}$.

Combining (C.1) and (C.2), $\gamma_0 = \rho_{p_s}$

In the similar manner, we can obtain that $\gamma_1 = \rho_p$. Since $\rho_{p_s} \leq \rho_p$ by the definition of the surrogate distribution, the cyclic-permutation bound of Sampler 4.2 is $\gamma_0 = \rho_{p_s}$.

C.2 Details of the Gibbs-type samplers for fitting the univariate t-distribution model (4.13)

The target posterior distribution of q and (μ, σ^2) is,

$$p(q, \mu, \sigma^2|Y) \propto \left(\prod_{i=1}^{n} q_i\right)^{\frac{\nu+1}{2}-1} (\sigma^2)^{-\frac{n}{2}-1} \exp\left\{-\left[\frac{\sum_{i=1}^{n} q_i(Y_i-\mu)^2}{2\sigma^2} + \frac{\nu\sum_{i=1}^{n} q_i}{2}\right]\right\}. \tag{C.3}$$

The steps of the standard DA sampler, i.e., Sampler 4.10, are

Step 1: Sample q_i from Gamma $\left[\frac{\nu+1}{2}, \frac{(Y_i-\mu')^2}{2\sigma^{2\prime}} + \frac{\nu}{2}\right]$, for $i = 1, \ldots, n$.

Step 2: Sample (μ, σ^2) from $p(\mu, \sigma^2|q, Y)$ by

- sampling σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^{n} q_i(Y_i-\hat{\mu})^2}{2}\right]$, where $\hat{\mu} = \frac{\sum_{i=1}^{n} q_i Y_i}{\sum_{i=1}^{n} q_i}$,
- sampling μ from N $\left(\hat{\mu}, \frac{\sigma^2}{\sum_{i=1}^{n} q_i}\right)$.

Setting $\tilde{q} = \alpha q$, and specifying the Haar measure prior to α as $p_\infty(\alpha) \propto 1/\alpha$, we obtain the joint posterior distribution of $\tilde{q}$, α, and (μ, σ^2), that is,

$$\begin{array}{ll} \tilde{p}(\tilde{q}, \alpha, \mu, \sigma^2|Y) & \propto \left(\prod_{i=1}^{n} \tilde{q}_i\right)^{\frac{\nu+1}{2}-1} (\sigma^2)^{-\frac{n}{2}-1} \alpha^{-\frac{n(\nu+1)}{2}-1} \\ & \exp\left\{-\left[\frac{\sum_{i=1}^{n} \tilde{q}_i(Y_i-\mu)^2}{2\alpha\sigma^2} + \frac{\nu\sum_{i=1}^{n} \tilde{q}_i}{2\alpha}\right]\right\}. \end{array} \tag{C.4}$$

The steps of the Haar PX-DA sampler, i.e., Sampler 4.11, are

Step 1: Sample $q_i^\star$ from Gamma $\left[\frac{\nu+1}{2}, \frac{(Y_i-\mu')^2}{2\sigma^{2\prime}} + \frac{\nu}{2}\right]$, for $i = 1, \ldots, n$.

Step 2: Sample α from Inv-Gamma $\left(\frac{n\nu}{2}, \frac{\nu\sum_{i=1}^{n} q_i^\star}{2}\right)$; Set $q = q^\star/\alpha$.

Step 3: Sample (μ, σ^2) from $p(\mu, \sigma^2|q, Y)$ by

- sampling σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^{n} q_i(Y_i-\hat{\mu})^2}{2}\right]$,
- sampling μ from N $\left(\hat{\mu}, \frac{\sigma^2}{\sum_{i=1}^{n} q_i}\right)$.

We derive the surrogate distribution of q and (μ, σ^2) as in (4.15). We first work out $p(q|Y)$ and $p_s(\mu, \sigma^2|q, Y)$ as

$$p(q|Y) \propto \left(\prod_{i=1}^{n} q_i\right)^{\frac{\nu+1}{2}-1} \left(\sum_{i=1}^{n} q_i\right)^{-\frac{1}{2}} \left[\sum_{i=1}^{n} q_i(Y_i-\hat{\mu})^2\right]^{-\frac{n-1}{2}} e^{-\frac{\nu\sum_{i=1}^{n} q_i}{2}}; \tag{C.5}$$

$$\begin{array}{rcl} p_s(\mu,\sigma^2|q,Y) & = & \int \tilde{p}(\alpha,\mu,\sigma^2|q,Y)\mathrm{d}\alpha \\ & = & \frac{\Gamma[n(\nu+1)/2]}{\Gamma(n\nu/2)\Gamma[(n-1)/2]}\frac{\nu^{n\nu/2}}{\sqrt{\pi}}(\sigma^2)^{-\frac{n}{2}-1}(\sum_{i=1}^n q_i)^{\frac{n\nu+1}{2}}\left[\sum_{i=1}^n q_i(Y_i-\hat{\mu})^2\right]^{\frac{n-1}{2}} \\ & & \left[\nu\sum_{i=1}^n q_i + \frac{\sum_{i=1}^n q_i(Y_i-\mu)^2}{\sigma^2}\right]^{-n(\nu+1)/2}. \end{array} \tag{C.6}$$

Thus the surrogate distribution is

$$\begin{array}{rcl} p_s(q,\mu,\sigma^2|Y) & = & p_s(\mu,\sigma^2|q,Y)p(q|Y) \\ & \propto & (\sigma^2)^{-\frac{n}{2}-1}(\prod_{i=1}^n q_i)^{\frac{\nu+1}{2}-1}(\sum_{i=1}^n q_i)^{\frac{n\nu}{2}}e^{-\frac{\nu\sum_{i=1}^n q_i}{2}} \\ & & \left[\nu\sum_{i=1}^n q_i + \frac{\sum_{i=1}^n q_i(Y_i-\mu)^2}{\sigma^2}\right]^{-n(\nu+1)/2}. \end{array} \tag{C.7}$$

The steps of the Gibbs sampler for updating the surrogate distribution, i.e., Sampler 4.12, are

Step 1: Sample q from $p_s(q|Y,\mu',\sigma^{2\prime})$ by,

- sampling $w_i^\star$ from Gamma $\left[\frac{\nu+1}{2}, \frac{(Y_i-\mu')^2}{2\sigma^{2\prime}}+\frac{\nu}{2}\right]$, for $i=1,\ldots,n$,
- sampling $\beta^\star$ from Gamma $\left(\frac{n\nu}{2}, \frac{\nu\sum_{i=1}^n w_i^\star}{2}\right)$; setting $q=\beta^\star w^\star$, where $w^\star=(w_1^\star,\ldots,w_n^\star)$;

Discard $\beta^\star$ and w.

Step 2: Sample (μ,σ^2) from $p_s(\mu,\sigma^2|q,Y)$ by

- Sample $\alpha^\star$ from Inv-Gamma $\left(\frac{n\nu}{2}, \frac{\nu\sum_{i=1}^n q_i}{2}\right)$,
- sampling σ^2 from Inv-Gamma $\left[\frac{n-1}{2}, \frac{\sum_{i=1}^n q_i(Y_i-\hat{\mu})^2}{2\alpha^\star}\right]$,
- sampling μ from N $\left(\hat{\mu}, \frac{\alpha^\star\sigma^2}{\sum_{i=1}^n q_i}\right)$;

Discard $\alpha^\star$.

C.3 Details of the Gibbs-type samplers for fitting the multivariate t-distribution model (4.16)

The target posterior distribution of q and (μ,σ^2) is,

$$p(q,\mu,\Sigma|Y) \propto \left(\prod_{i=1}^n q_i\right)^{\frac{\nu+d}{2}-1}|\Sigma|^{-\frac{n+d+1}{2}}\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n q_i(Y_i-\mu)^T\Sigma^{-1}(Y_i-\mu)+\nu\sum_{i=1}^n q_i\right]\right\}. \tag{C.8}$$

The steps of the standard DA sampler for updating the target distribution are

Step 1: Sample q_i from Gamma $\left[\frac{\nu+d}{2}, \frac{(Y_i-\mu')^T\Sigma'^{-1}(Y_i-\mu')+\nu}{2}\right]$, for $i = 1, \dots, n$.

Step 2: Sample (μ, Σ) from $p(\mu, \Sigma|q, Y)$ by

- sampling Σ from Inv-Wishart $\left[n-1, \sum_{i=1}^n q_i(Y_i - \hat{\mu})(Y_i - \hat{\mu})^T\right]$, where $\hat{\mu} = \frac{\sum_{i=1}^n q_i Y_i}{\sum_{i=1}^n q_i}$,
- sampling μ from $\mathrm{N}_d\left(\hat{\mu}, \frac{\Sigma}{\sum_{i=1}^n q_i}\right)$.

Setting $\tilde{q} = \alpha q$, and specifying the Haar measure prior to α as $p_\infty(\alpha) \propto 1/\alpha$, we obtain the joint posterior distribution of $\tilde{q}$, α, and (μ, Σ), that is,

$$\begin{aligned} \tilde{p}(\tilde{q}, \alpha, \mu, \Sigma|Y) \;&\propto (\textstyle\prod_{i=1}^n \tilde{q}_i)^{\frac{\nu+d}{2}-1}|\Sigma|^{-\frac{n+d+1}{2}}\alpha^{-\frac{n(\nu+d)}{2}-1} \\ &\exp\left\{-\frac{1}{2\alpha}\left[\textstyle\sum_{i=1}^n \tilde{q}_i(Y_i-\mu)^T\Sigma^{-1}(Y_i-\mu) + \nu\sum_{i=1}^n \tilde{q}_i\right]\right\}. \end{aligned} \tag{C.9}$$

The steps of the Haar PX-DA sampler are

Step 1: Sample $q_i^\star$ from Gamma $\left[\frac{\nu+d}{2}, \frac{(Y_i-\mu')^T\Sigma'^{-1}(Y_i-\mu')+\nu}{2}\right]$, for $i = 1, \dots, n$.

Step 2: Sample α from Inv-Gamma $\left(\frac{n\nu}{2}, \frac{\nu\sum_{i=1}^n q_i^\star}{2}\right)$; Set $q = q^\star/\alpha$.

Step 3: Sample (μ, Σ) from $p(\mu, \Sigma|q, Y)$ by

- sampling Σ from Inv-Wishart $\left[n-1, \sum_{i=1}^n q_i(Y_i - \hat{\mu})(Y_i - \hat{\mu})^T\right]$,
- sampling μ from $\mathrm{N}_d\left(\hat{\mu}, \frac{\Sigma}{\sum_{i=1}^n q_i}\right)$.

We derive the surrogate distribution of q and (μ, Σ) as in (4.15). We first work out $p(q|Y)$ and $p_s(\mu, \Sigma|q, Y)$ as

$$p(q|Y) \propto \left(\prod_{i=1}^n q_i\right)^{\frac{\nu+d}{2}-1}\left(\sum_{i=1}^n q_i\right)^{-\frac{d}{2}}\left|\sum_{i=1}^n q_i(Y_i-\hat{\mu})(Y_i-\hat{\mu})^T\right|^{-\frac{n-1}{2}} e^{-\frac{\nu\sum_{i=1}^n q_i}{2}}; \tag{C.10}$$

$$\begin{aligned} p_s(\mu, \Sigma|q, Y) \;&= \textstyle\int \tilde{p}(\alpha, \mu, \Sigma|q, Y)\mathrm{d}\alpha \\ &= \frac{\Gamma[n(\nu+d)/2]}{\Gamma(n\nu/2)\Gamma_d[(n-1)/2]}\frac{\nu^{n\nu/2}}{\pi^{d/2}}|\Sigma|^{-\frac{n+d+1}{2}} \\ &(\textstyle\sum_{i=1}^n q_i)^{\frac{n\nu+d}{2}}\left|\textstyle\sum_{i=1}^n q_i(Y_i-\hat{\mu})(Y_i-\hat{\mu})^T\right|^{\frac{n-1}{2}} \\ &\left[\nu\textstyle\sum_{i=1}^n q_i + \sum_{i=1}^n \tilde{q}_i(Y_i-\mu)^T\Sigma^{-1}(Y_i-\mu)\right]^{-\frac{n(\nu+d)}{2}}. \end{aligned} \tag{C.11}$$

Thus the surrogate distribution is

$$\begin{aligned} p_s(q,\mu,\Sigma|Y) &= p_s(\mu,\Sigma|q,Y)p(q|Y) \\ &\propto |\Sigma|^{-\frac{n+d+1}{2}} (\textstyle\prod_{i=1}^n \tilde{q}_i)^{\frac{\nu+d}{2}-1} (\textstyle\sum_{i=1}^n q_i)^{\frac{n\nu}{2}} e^{-\frac{\nu \sum_{i=1}^n q_i}{2}} \\ &\quad \left[\nu \textstyle\sum_{i=1}^n q_i + \sum_{i=1}^n \tilde{q}_i (Y_i-\mu)^T \Sigma^{-1}(Y_i-\mu)\right]^{-\frac{n(\nu+d)}{2}}. \end{aligned} \tag{C.12}$$

The steps of the Gibbs sampler for updating the surrogate distribution are

Step 1: Sample q from $p_s(q|Y,\mu',\Sigma')$ by,

- sampling $w_i^\star$ from Gamma $\left[\frac{\nu+d}{2}, \frac{(Y_i-\mu')^T\Sigma'^{-1}(Y_i-\mu')+\nu}{2}\right]$, for $i=1,\ldots,n$,
- sampling $\beta^\star$ from Gamma $\left(\frac{n\nu}{2}, \frac{\nu\sum_{i=1}^n w_i^\star}{2}\right)$; setting $q=\beta^\star w^\star$, where $w^\star=(w_1^\star,\ldots,w_n^\star)$;

Discard $\beta^\star$ and w.

Step 2: Sample (μ,Σ) from $p_s(\mu,\Sigma|q,Y)$ by

- Sample $\alpha^\star$ from Inv-Gamma $\left(\frac{n\nu}{2}, \frac{\nu\sum_{i=1}^n q_i}{2}\right)$,
- sampling Σ from Inv-Wishart $\left[n-1, \frac{\sum_{i=1}^n q_i(Y_i-\hat{\mu})(Y_i-\hat{\mu})^T}{\alpha^\star}\right]$,
- sampling μ from $\mathrm{N}_d\left(\hat{\mu}, \frac{\alpha^\star\Sigma}{\sum_{i=1}^n q_i}\right)$;

Discard $\alpha^\star$.

In fact, the univariate t-distribution model is a special case of the multivariate t-distribution model.

C.4 Details of the Gibbs-type samplers for fitting the spectral analysis model (4.18)

The target posterior distribution of α, β, μ, and ϕ under the spectral model (4.18) is

$$\begin{aligned} p(\alpha,\beta,\mu,\phi|Y) \propto\ & \textstyle\prod_{i=1}^n \left[\alpha(E_i^{-\beta} + \sum_{k=1}^K \gamma_k I\{i=\mu_k\})e^{-\phi/E_i}\right]^{Y_i} \\ & \exp\left\{-\alpha \textstyle\sum_{i=1}^n (E_i^{-\beta} + \sum_{k=1}^K \gamma_k I\{i=\mu_k\})e^{-\phi/E_i}\right\}. \end{aligned} \tag{C.13}$$

Integrating (C.13) over α, we have,

$$\begin{aligned} p(\beta, \mu, \phi|Y) \propto \quad & \textstyle\prod_{i=1}^n \left[(E_i{}^{-\beta} + \sum_{k=1}^K \gamma_k I\{i = \mu_k\}) e^{-\phi/E_i} \right]^{Y_i} \times \\ & \left[\textstyle\sum_{i=1}^n (E_i{}^{-\beta} + \sum_{k=1}^K \gamma_k I\{i = \mu_k\}) e^{-\phi/E_i} \right]^{-(\sum_{i=1}^n Y_i + 1)}. \end{aligned} \tag{C.14}$$

Hence, the steps of the MH within PCG sampler updating μ and (β, ϕ) without conditioning on α, i.e., Sampler 4.13, are

Step 1: Use MH to sample μ from $p(\mu|\beta', \phi', Y) \propto p(\beta', \mu, \phi|Y)$.

Step 2: Use MH to sample (β, ϕ) from $p(\beta, \phi|\mu, Y) \propto p(\beta, \mu, \phi|Y)$.

Step 3: Sample α from Gamma $\left[\sum_{i=1}^n Y_i + 1,\ \sum_{i=1}^n (E_i{}^{-\beta} + \sum_{k=1}^K \gamma_k I\{i = \mu_k\}) e^{-\phi/E_i} \right]$.

The steps of the three-step sampler using the conditionals of the surrogate distribution $p_s(\alpha, \beta, \mu, \phi|Y)$ in Steps 1 and 3, i.e., Sampler 4.14, are

Step 1: Use MH to sample μ from $p(\mu|\beta', \phi', Y) \propto p(\beta', \mu, \phi|Y)$.

Step 2: Use MH to sample (β, ϕ) from $p(\beta, \phi|\mu, Y) \propto p(\beta, \mu, \phi|Y)$.

Step 3: For $K = 2$, as for the simulation study in Section 4.3.2, sample α from $\sum_{l=1}^n \sum_{j=1}^n p_{jl} \text{Gamma}(a, b_{jl})$,

where $a = \sum_{i=1}^n Y_i + 1$; $p_{jl} = \frac{\omega_{jl}}{\sum_{l=1}^n \sum_{j=1}^n \omega_{jl}}$ and $\sum_{l=1}^n \sum_{j=1}^n p_{jl} = 1$; for $j = 1$ and $l = 1$, $b_{jl} = \sum_{i=1}^n E_i^{-\beta} e^{-\phi/E_i} + \gamma_1 e^{-\phi/E_j} + \gamma_2 e^{-\phi/E_l}$, and

- if $j = l$, $\omega_{jl} = \omega_{ll} = (E_l^{-\beta} + \gamma_1 + \gamma_2)^{Y_l} b_{ll}^{-a} (\prod_{i=1, i \neq l}^n E_i^{-\beta Y_i})$,
- otherwise, $\omega_{jl} = (E_j^{-\beta} + \gamma_1)^{Y_j} (E_l^{-\beta} + \gamma_2)^{Y_l} b_{jl}^{-a} (\prod_{i=1, i \neq j \neq l}^n E_i^{-\beta Y_i})$.

We use a uniform distribution on $\{1, \ldots, n\}^K$ as the jumping rule when updating μ. When sampling β and ϕ jointly via MH, the jumping rule is a bivariate Gaussian distribution centered at the current draw with variance-covariance matrix adjusted to obtain an acceptance rate of around 20%.

C.5 Details of the Gibbs-type samplers for fitting the simple hierarchical Gaussian model (4.20)

The target posterior distribution of X and ψ is,

$$p(X, \psi|Y) \propto \exp\left\{ -\frac{1}{2} \left[(Y - X)^2 + \frac{(X - \psi)^2}{V} \right] \right\}. \tag{C.15}$$

The steps of the standard DA sampler, i.e., Sampler 4.15, are

Step 1: Sample X from $\mathrm{N}\left(\frac{VY+\psi'}{V+1}, \frac{V}{V+1}\right)$.

Step 2: Sample ψ from $\mathrm{N}(X, V)$.

Setting $\bar{X} = X - \psi$, the joint posterior distribution of $\bar{X}$ and ψ is,

$$p(X, \psi|Y) \propto \exp\left\{-\frac{1}{2}\left[(Y - \bar{X} - \psi)^2 + \frac{\bar{X}^2}{V}\right]\right\}. \tag{C.16}$$

The steps of the ASIS sampler, i.e., Sampler 4.16, are

Step 1: Sample $X^\star$ from $\mathrm{N}\left(\frac{VY+\psi'}{V+1}, \frac{V}{V+1}\right)$.

Step 2: Sample $\psi^\star$ from $\mathrm{N}(X^\star, V)$; Set $\bar{X} = X^\star - \psi^\star$.

Step 3: Sample ψ from $\mathrm{N}(Y - \bar{X}, 1)$; Set $X = \bar{X} + \psi$.

We derive the surrogate distribution of X and ψ as in (4.21). We first obtain $p(X|Y)$, which is $\mathrm{N}(Y, 1)$, and $p_s(\psi|X, Y)$, which is $\mathrm{N}[Y, (V+1)]$. Because $p_s(\psi|X, Y) = p(\psi|Y) = p_s(\psi|Y)$, $p_s(X|\psi, Y) = p(X|Y)$. Thus it is not necessary to express the joint surrogate distribution $p_s(\psi, X|Y)$ explicitly.

The steps of the Gibbs sampler for updating the surrogate distribution, i.e., Sampler 4.17, are

Step 1: Sample X from $\mathrm{N}(Y, 1)$.

Step 2: Sample ψ from $\mathrm{N}(Y, V + 1)$.

D

Details of the Gibbs-type Samplers in Chapter 5

In this appendix, we provide the details of the samplers for fitting the generalized versions of the Baseline Model in supernova cosmology, which are introduced in Chapter 5.

D.1 The Baseline Model

The joint and marginal posterior distributions under the Baseline Model have almost the same forms as (A.4) and (A.5) in Section A.2 of Appendix A, except that we replace $\frac{1}{\sigma^2_{\text{res}}}$ by the density of Inv-Gamma$(0.003, 0.003)$.

The MH within PCG sampler for the Baseline Model is almost the same as Sampler 2.8 in Section 2.3.2 of Chapter 2, except that we sample σ^2_{res} from Inv-Gamma$\left[\,\frac{n}{2}+0.003,\ \frac{\sum_{i=1}^{n}(M_i-M_0)^2}{2}+0.003\,\right]$.

The posterior distributions for the hierarchical model with $\Sigma_C = C_{\text{stat}}$ have the same forms as those for the Baseline Model. The steps of the sampler for fitting this model are also the same as those for fitting the Baseline Model.

D.2 The Hard Classification Model

In this model, we divide the SN population into two classes according to the observed host galaxy mass. The joint and marginal posterior distributions under the Hard Classification Model have the identical forms as those in (A.4) and (A.5) except that we replace $\frac{1}{\sigma^2_{\text{res}}}$ with the product of two Inv-Gamma(0.003, 0.003) densities. The specification of Y, X, L, Σ_C, and A is identical to that in the Baseline Model. However, the specification of ξ, ξ_m, Σ_P, Σ_0, and J is modified to reflect the existence of two host galaxy mass populations. Under this model, $\xi = (c_0, x_0, M_0^{\text{lo}}, M_0^{\text{hi}})$; $\xi_m = (0, 0, -19.3, -19.3)$; $\Sigma_P = \text{Diag}(S_1, \ldots, S_n)$, where $S_i = \text{Diag}[R_c^2, R_x^2, (1 - Z_i)(\sigma_{\text{res}}^{\text{lo}})^2 + Z_i(\sigma_{\text{res}}^{\text{hi}})^2]$; $\Sigma_0 = \text{Diag}(1^2, 10^2, 2^2, 2^2)$; $J_{(3n\times4)} = \begin{bmatrix} J_1 \\ \vdots \\ J_n \end{bmatrix}$, where $J_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 - Z_i & Z_i \end{bmatrix}$. Under this model, $Z = (Z_1, \ldots, Z_n)$ is known with,

$$Z_i = \begin{cases} 1 & \text{if } \widehat{M}_{\text{g}\,i} \geq 10 \\ 0 & \text{otherwise.} \end{cases} \tag{D.1}$$

The steps of the MH within PCG sampler for fitting the Hard Classification Model are,

Step 1: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y, \alpha, \beta, \Sigma_P)$, which is proportional to $p(\mathscr{C}, \alpha, \beta, \Sigma_P|Y)$, with $\mathscr{C} = (\Omega_m, \Omega_\Lambda) \in [0, 1] \times [0, 2]$.

Step 2: Use MH to sample (α, β) from $p(\alpha, \beta|Y, \mathscr{C}, \Sigma_P)$, which is proportional to $p(\mathscr{C}, \alpha, \beta, \Sigma_P|Y)$, with $(\alpha, \beta) \in [0, 1] \times [0, 4]$.

Step 3: Sample (ξ, X), which consists of two sub-steps:

- Sample ξ from $\text{N}(k_0, K)$;
- Sample X from $\text{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1} J\xi)$.

Step 4: Sample Σ_P, which consists of four sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (c_i - c_0)^2}{2}\right]$ with $\log(R_c) \in [-5, 2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (x_i - x_0)^2}{2}\right]$ with $\log(R_x) \in [-5, 2]$.
- Sample $(\sigma_{\text{res}}^{\text{lo}})^2$ from Inv-Gamma$[\frac{\sum_{i=1}^n (1-Z_i)}{2} + 0.003, \frac{\sum_{i=1}^n (1-Z_i)(M_i - M_0^{\text{lo}})^2}{2} + 0.003]$.
- Sample $(\sigma_{\text{res}}^{\text{hi}})^2$ from Inv-Gamma $\left[\frac{\sum_{i=1}^n Z_i}{2} + 0.003, \frac{\sum_{i=1}^n Z_i(M_i - M_0^{\text{hi}})^2}{2} + 0.003\right]$.

D.3 The Soft Classification Model

For the Soft Classification Model, the SNe are classified by their true (latent) host galaxy masses (rather than by their observed masses as in the Hard Classification Model), and the indicator variables, Z, are treated as unknown. Thus, the joint and marginal posterior distributions under this model should be written as $p(X, \xi, \mathscr{C}, \alpha, \beta, \Sigma_P, Z|Y)$ and $p(\mathscr{C}, \alpha, \beta, \Sigma_P, Z|Y)$. The distributions have the identical forms as those in (A.4) and (A.5) respectively except that we replace $\frac{1}{\sigma^2_{\text{res}}}$ with

$$p\left({\sigma^{\text{lo}}_{\text{res}}}^2\right) p\left({\sigma^{\text{hi}}_{\text{res}}}^2\right) \prod_{i=1}^{n} p_i^{Z_i}(1-p_i)^{1-Z_i}, \tag{D.2}$$

where

$$\begin{aligned} p_i &= \Pr(Z_i = 1 \mid \widehat{M}_{\text{g}\,i}) = \Pr(M_{\text{g}\,i} \geq 10 \mid \widehat{M}_{\text{g}\,i}) \\ &= \textstyle\int_{10}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_{\text{g}\,i}} \exp\left[-(M_{\text{g}\,i} - \widehat{M}_{\text{g}\,i})^2/(2\sigma^2_{\text{g}\,i})\right] \mathrm{d}M_{\text{g}\,i}, \end{aligned} \tag{D.3}$$

for $i = 1, \ldots, n$. The specification of Y, X, ξ, ξ_m, L, Σ_C, Σ_P, Σ_0, A, and J is identical to that in the Hard Classification Model.

The steps of the MH within PCG sampler for fitting the Soft Classification Model are,

Step 1: For each i, sample Z_i from Bernoulli($\tilde{p}_i$), where $\tilde{p}_i = \frac{p_{i,\text{high}}}{p_{i,\text{low}} + p_{i,\text{high}}}$, with

$$p_{i,\text{low}} = \frac{1}{\sigma^{\text{lo}}_{\text{res}}} \exp\left\{-\frac{(M^{\epsilon}_i - M^{\text{lo}}_0)^2}{2(\sigma^{\text{lo}}_{\text{res}})^2}\right\}(1-p_i), \tag{D.4}$$

and

$$p_{i,\text{high}} = \frac{1}{\sigma^{\text{lo}}_{\text{res}}} \exp\left\{-\frac{(M^{\epsilon}_i - M^{\text{hi}}_0)^2}{2(\sigma^{\text{hi}}_{\text{res}})^2}\right\}(1-)p_i); \tag{D.5}$$

Use Z to construct J.

Step 2: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y, \alpha, \beta, \Sigma_P, Z)$, which is proportional to $p(\mathscr{C}, \alpha, \beta, \Sigma_P, Z|Y)$, with $\mathscr{C} = (\Omega_m, \Omega_\Lambda) \in [0, 1] \times [0, 2]$.

Step 3: Use MH to sample (α, β) from $p(\alpha, \beta|Y, \mathscr{C}, \Sigma_P, Z)$, which is proportional to $p(\mathscr{C}, \alpha, \beta, \Sigma_P, Z|Y)$, with $(\alpha, \beta) \in [0, 1] \times [0, 4]$.

Step 4: Sample (ξ, X), which consists of two sub-steps:

- Sample ξ from $\text{N}(k_0, K)$;
- Sample X from $\text{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1}J\xi)$.

Step 5: Sample Σ_P, which consists of four sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^{n}(c_i - c_0)^2}{2}\right]$ with $\log(R_c) \in [-5, 2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i - x_0)^2}{2}\right]$ with $\log(R_x) \in [-5, 2]$.
- Sample $(\sigma_{\text{res}}^{\text{lo}})^2$ from Inv-Gamma$[\frac{\sum_{i=1}^{n}(1-Z_i)}{2}+0.003, \frac{\sum_{i=1}^{n}(1-Z_i)(M_i - M_0^{\text{lo}})^2}{2}+0.003]$.
- Sample $(\sigma_{\text{res}}^{\text{hi}})^2$ from Inv-Gamma $\left[\frac{\sum_{i=1}^{n} Z_i}{2} + 0.003, \frac{\sum_{i=1}^{n} Z_i(M_i - M_0^{\text{hi}})^2}{2} + 0.003\right]$.

D.4 The Covariate Adjustment Model

In this model, since we include $M_{\text{g}\,i}$ as an additional covariate, the specification of quantities in the posterior distribution is different from the Baseline Model. First, for $Y = (Y_1, \ldots, Y_n)$, $Y_i = (\hat{c}_i, \hat{x}_i, \widehat{M}_{\text{g}\,i}, \hat{m}_{Bi})$. Moreover, $X = (c_1, x_1, M_{\text{g},1}, M_1, \ldots, c_n, x_n, M_{\text{g},n}, M_n)$, $\xi = (c_0, x_0, M_{\text{g}\star}, M_0)$, and $\xi_m = (0, 0, 10, -19.3)$. For the variance-covariance matrices, Σ_C now has the dimension of $(4n \times 4n)$. The $(3n \times 3n)$ submatrix of Σ_C, after deleting the $(4i)^{\text{th}}$ $(i = 1, \ldots, n)$ rows and columns, is $(C_{\text{stat}} + C_{\text{syst}})$. The $(4i, 4i)^{\text{th}}$ element of Σ_C is $\sigma_{\text{g}\,i}^2$, while the other elements in the $(4i)^{\text{th}}$ rows and columns are all zero, because we ignore correlations between $\widehat{M}_{\text{g}\,i}$ and other observed quantities; $\Sigma_P = \text{Diag}(S_1, \ldots, S_n)$, where each $S_i = \text{Diag}(R_c^2, R_x^2, R_{\text{g}}^2, \sigma_{\text{res}}^2)$; $\Sigma_0 = \text{Diag}(1^2, 10^2, 100^2, 2^2)$. In addition, $J_{(4n\times 4)} = \begin{bmatrix} J_1 \\ \vdots \\ J_n \end{bmatrix}$, where each J_i is a (4×4) identity matrix; $A_{(4n\times 4n)} =$ $\text{Diag}(T_1, \ldots, T_n)$, where each $T_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \beta & -\alpha & \gamma & 1 \end{bmatrix}$. Under this model, the joint and marginal posterior distributions should be written as $p(X, \xi, \mathscr{C}, \alpha, \beta, \gamma, \Sigma_P | Y, \widehat{\mathscr{D}}_{\text{g}})$ and $p(\mathscr{C}, \alpha, \beta, \gamma, \Sigma_P | Y, \widehat{\mathscr{D}}_{\text{g}})$ respectively. But they are formally identical to (A.4) and (A.5), respectively, except that we replace $\frac{1}{\sigma_{\text{res}}^2}$ with

$$\frac{1}{R_{\text{g}}^2} p(\sigma_{\text{res}}^2). \tag{D.6}$$

The steps of the MH within PCG sampler for fitting the Covariate Adjustment Model are,

Step 1: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C} | Y, \widehat{\mathscr{D}}_{\text{g}}, \alpha, \beta, \gamma, \Sigma_P)$, which is proportional to

$p(\mathscr{C}, \alpha, \beta, \gamma, \Sigma_P | Y, \widehat{\mathscr{D}}_{\mathrm{g}})$, with $\mathscr{C} = (\Omega_m, \Omega_\Lambda) \in [0,1] \times [0,2]$.

Step 2: Use MH to sample (α, β, γ) from $p(\alpha, \beta, \gamma | Y, \widehat{\mathscr{D}}_{\mathrm{g}}, \mathscr{C}, \Sigma_P)$, which is proportional to $p(\mathscr{C}, \alpha, \beta, \gamma, \Sigma_P | Y, \widehat{\mathscr{D}}_{\mathrm{g}})$, with $(\alpha, \beta, \gamma) \in [0,1] \times [0,4] \times [-4,4]$.

Step 3: Sample (ξ, X), which consists of two sub-steps:

- Sample ξ from $\mathrm{N}(k_0, K)$;
- Sample X from $\mathrm{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1} J\xi)$.

Step 4: Sample Σ_P, which consists of four sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (c_i - c_0)^2}{2}\right]$ with $\log(R_c) \in [-5,2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (x_i - x_0)^2}{2}\right]$ with $\log(R_x) \in [-5,2]$.
- Sample R_{g}^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (M_{\mathrm{g}\,i} - M_{\mathrm{g}\star})^2}{2}\right]$ with $\log(R_{\mathrm{g}}) \in [-5,2]$.
- Sample σ_{res}^2 from Inv-Gamma $\left[\frac{n}{2} + 0.003, \frac{\sum_{i=1}^n (M_i - M_0)^2}{2} + 0.003\right]$.

D.5 The z-Linear color Correction Model

In the z-Linear color Correction model, the specification of Y, ξ, ξ_m, L, Σ_C, Σ_P, Σ_0, and J is identical to that in the Baseline model. But $X = (c_1, z_1c_1, x_1, M_1, \ldots, c_n, z_nc_n, x_n, M_n)$, and $A = \mathrm{Diag}(T_1, \ldots, T_n)$, where $T_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta_0 + \beta_1 z_i & -\alpha & 1 \end{bmatrix}$.

Under this model, the joint and marginal posterior distributions should be written as $p(X, \xi, \mathscr{C}, \alpha, \beta_0, \beta_1, \Sigma_P | Y)$ and $p(\mathscr{C}, \alpha, \beta_0, \beta_1, \Sigma_P | Y)$ respectively. But they are formally identical to (A.4) and (A.5), respectively, except that we replace $\frac{1}{\sigma_{\mathrm{res}}^2}$ with the density of Inv-Gamma$(0.003, 0.003)$.

For fitting this model, we combine MH within PCG and ASIS algorithms. We integrate (X, ξ) out when updating $\mathscr{C}$, and use the ASIS algorithm to update $(-\alpha, \beta_0, \beta_1)$. The distribution of X conditioning on $(-\alpha, \beta_0, \beta_1)$ and other parameters is

$$X | \xi, \Sigma_P, \alpha, \beta_0, \beta_1, \mathscr{C} \sim \mathrm{N}_{3n}(J\xi, \Sigma_P). \tag{D.7}$$

Because this distribution is free of $(-\alpha, \beta_0, \beta_1)$, X is an ancillary augmentation for $(-\alpha, \beta_0, \beta_1)$ conditioning on other parameters. To derive a sufficient augmentation, we set $\tilde{X} = AX$. The distribution of Y conditioning on $\tilde{X}$, $(-\alpha, \beta_0, \beta_1)$, and other parameters is

$$Y|\tilde{X}, \xi, \Sigma_P, \alpha, \beta_0, \beta_1, \mathscr{C} \sim \mathrm{N}_{3n}(\tilde{X}\Sigma_C). \tag{D.8}$$

Because this distribution is free of $(-\alpha, \beta_0, \beta_1)$, $\tilde{X}$ is the corresponding sufficient augmentation for $(-\alpha, \beta_0, \beta_1)$.

The steps of the MH within PCG + ASIS sampler for fitting the z-Linear color Correction Model are,

Step 1: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y, \alpha, \beta_0, \beta_1, \Sigma_P)$, which is proportional to $p(\mathscr{C}, \alpha, \beta_0, \beta_1, \Sigma_P|Y)$, with $\mathscr{C} = (\Omega_m, \Omega_\Lambda) \in [0, 1] \times [0, 2]$.

Step 2: Sample $(\xi, X^\star)$, which consists of two sub-steps:

- Sample ξ from $\mathrm{N}(k_0, K)$;
- Sample $X^\star$ from $\mathrm{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1} J\xi)$.

Step 3: Sample $(-\alpha^\star, \beta_0^\star, \beta_1^\star)$ from $\mathrm{N}_3(\zeta_B, \Sigma_B)$ (details about this distribution are given below) with constraint $(-\alpha^\star, \beta_0^\star, \beta_1^\star) \in [-1, 0] \times [0, 4] \times [-4, 4]$;

Use $(-\alpha^\star, \beta_0^\star, \beta_1^\star)$ to construct $A^\star$; Then set $\tilde{X} = A^\star X^\star$.

Step 4: Sample $(-\alpha, \beta_0, \beta_1)$ from $\mathrm{N}_3(\tilde{\zeta}_B, \tilde{\Sigma}_B)$ (details about this distribution are given below) with constraint $(-\alpha, \beta_0, \beta_1) \in [-1, 0] \times [0, 4] \times [-4, 4]$;

Use $(-\alpha, \beta_0, \beta_1)$ to construct A; Then set $X = A^{-1}\tilde{X}$.

Step 5: Sample Σ_P, which consists of four sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (c_i - c_0)^2}{2}\right]$ with $\log(R_c) \in [-5, 2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (x_i - x_0)^2}{2}\right]$ with $\log(R_x) \in [-5, 2]$.
- Sample σ_{res}^2 from Inv-Gamma$[\frac{n}{2} + 0.003, \frac{\sum_{i=1}^n (M_i - M_0)^2}{2} + 0.003]$.

In Step 3, $\Sigma_B^{-1} = E^T V_m^{-1} E$, where V_m is the $(n \times n)$ submatrix of Σ_C after deleting the $(3i-2)^{\text{th}}$ $(i = 1, \ldots, n)$ and $(3i-1)^{\text{th}}$ $(i = 1, \ldots, n)$ rows and columns, and $E_{(n\times 3)} = \begin{bmatrix} c_1 & z_1c_1 & x_1 \\ \vdots & \vdots & \vdots \\ c_n & z_nc_n & x_n \end{bmatrix}$. Furthermore, $\zeta_B = \Sigma_B E^T V_m^{-1}(\hat{\eta}_m - \eta_m - \Delta\eta)$, where $\hat{\eta}_m = (\hat{m}_{B1}^\star - \mu_1, \ldots, \hat{m}_{Bn}^\star - \mu_n)$, $\eta_m = (M_1, \ldots, M_n)$, and $\Delta\eta = V_{m,-m}V_{-m}^{-1}(\hat{\eta}_{-m} - \eta_{-m})$; V_{-m} is the $(2n \times 2n)$ submatrix of Σ_C after deleting the $(3i)^{\text{th}}$ $(i = 1, \ldots, n)$ rows and columns; $V_{m,-m}$ is the $(n \times 2n)$ submatrix of Σ_C after deleting the $(3i-2)^{\text{th}}$ $(i = 1, \ldots, n)$ and $(3i-1)^{\text{th}}$ $(i = 1, \ldots, n)$ rows and the $(3i)^{\text{th}}$ $(i = 1, \ldots, n)$ columns; $\hat{\eta}_{-m} = (\hat{c}_1, \hat{x}_{11}, \ldots, \hat{c}_n, \hat{x}_{1n})$; $\eta_{-m} = (c_1, x_{11}, \ldots, c_n, x_{1n})$.

In Step 4, $\tilde{\Sigma}_B^{-1} = (\tilde{E}^T\tilde{E})/\sigma^2_{\text{res}}$, where $\tilde{E}_{(n\times 3)} = \begin{bmatrix} \tilde{E}_1^T \\ \vdots \\ \tilde{E}_n^T \end{bmatrix}$ with $\tilde{E}_i = (-\tilde{c}_i, -z_i\tilde{c}_i, -\tilde{x}_{1i})$; $\tilde{c}_i$ and $\tilde{x}_{1i}$ are the $(3i-2)^{\text{th}}$ and $(3i-1)^{\text{th}}$ components of $\tilde{X}$ respectively. Furthermore, $\tilde{\zeta}_B = \tilde{\Sigma}_B[\tilde{E}^T(\eta_{M_0} - \tilde{\eta}_m)/\sigma^2_{\text{res}}]$, where $\eta_{M_0} = (\underbrace{M_0, \ldots, M_0}_{n})$ and $\tilde{\eta}_m = (\tilde{M}_1, \ldots, \tilde{M}_n)$; $\tilde{M}_i$ is the $(3i-2)^{\text{th}}$ component of $\tilde{X}$.

D.6 The z-Jump color Correction Model

In the z-Jump color Correction model, the specification of Y, ξ, ξ_m, L, Σ_C, Σ_P, Σ_0, and J is identical to that in the Baseline model. But under this model, $X = \big(c_1, \left(\frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{z_1 - z_t}{0.01}\right)\right) c_1, x_1, M_1, \ldots, c_n, \left(\frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{z_n - z_t}{0.01}\right)\right) c_n, x_n, M_n\big)$, and $A = \text{Diag}(T_1, \ldots, T_n)$, where $T_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta_0 + \Delta\beta\left(\frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{z_i - z_t}{0.01}\right)\right) & -\alpha & 1 \end{bmatrix}$.

Under this model, the joint and marginal posterior distributions should be written as $p(X, \xi, \mathscr{C}, \alpha, \beta_0, \Delta\beta, z_t, \Sigma_P|Y)$ and $p(\mathscr{C}, \alpha, \beta_0, \Delta\beta, z_t, \Sigma_P|Y)$ respectively. But they are formally identical to (A.4) and (A.5), respectively, except that we replace $\frac{1}{\sigma^2_{\text{res}}}$ with the density of Inv-Gamma$(0.003, 0.003)$.

For fitting this model, as in the sampler for the z-Linear color Correction Model, we also combine MH within PCG and ASIS algorithms in this sampler. We integrate (X, ξ) out when updating both $\mathscr{C}$ and z_t, and use the ASIS algorithm to update $(-\alpha, \beta_0, \Delta\beta)$. When implementing ASIS, we also regard X as the ancillary augmentation, and $\tilde{X} = AX$ as the corresponding sufficient augmentation for $(-\alpha, \beta_0, \Delta\beta)$, conditioning on other parameters.

The steps of the MH within PCG + ASIS sampler for fitting the z-Jump color Correction Model are,

Step 1: Use MH to sample $\mathscr{C}$ from $p(\mathscr{C}|Y, \alpha, \beta_0, \Delta\beta, z_t, \Sigma_P)$, which is proportional to $p(\mathscr{C}, \alpha, \beta_0, \Delta\beta, z_t, \Sigma_P|Y)$, with $\mathscr{C} = (\Omega_m, \Omega_\Lambda) \in [0, 1] \times [0, 2]$.

Step 2: Use MH to sample z_t from $p(z_t|Y, \mathscr{C}, \alpha, \beta_0, \Delta\beta, \Sigma_P)$, which is proportional to $p(z_t, \mathscr{C}, \alpha, \beta_0, \Delta\beta, \Sigma_P|Y)$, under the constraint $z_t \in [0.2, 1]$.

Step 3: Sample $(\xi, X^\star)$, which consists of two sub-steps:

- Sample ξ from $\text{N}(k_0, K)$;

- Sample $X^\star$ from $\mathrm{N}(\mu_A, \Sigma_A)$, where $\mu_A = \Sigma_A(\Delta + \Sigma_P^{-1} J\xi)$.

Step 4: Sample $(-\alpha^\star, \beta_0^\star, \Delta\beta^\star)$ from $\mathrm{N}_3(\zeta_B, \Sigma_B)$ with constraint $(-\alpha^\star, \beta_0^\star, \beta_1^\star) \in [-1, 0] \times [0, 4] \times [-1.5, 1.5]$;

The construction of ζ_B and Σ_B is identical to that in the z-Linear color Correction sampler;

Use $(-\alpha^\star, \beta_0^\star, \Delta\beta^\star)$ to construct $A^\star$; Then set $\tilde{X} = A^\star X^\star$.

Step 5: Sample $(-\alpha, \beta_0, \Delta\beta)$ from $\mathrm{N}_3(\tilde{\zeta}_B, \tilde{\Sigma}_B)$ with constraint $(-\alpha, \beta_0, \Delta\beta) \in [-1, 0] \times [0, 4] \times [-1.5, 1.5]$;

The construction of $\tilde{\zeta}_B$ and $\tilde{\Sigma}_B$ is identical to that in the z-Linear color Correction sampler, except that under this model, $\tilde{E}_i = \left(-\tilde{c}_i, -\left(\frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{z_i - z_t}{0.01}\right)\right) \tilde{c}_i, -\tilde{x}_{1i}, \right)$;

Use $(-\alpha, \beta_0, \Delta\beta)$ to construct A; Then set $X = A^{-1}\tilde{X}$.

Step 6: Sample Σ_P, which consists of four sub-steps:

- Sample R_c^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (c_i - c_0)^2}{2}\right]$ with $\log(R_c) \in [-5, 2]$.
- Sample R_x^2 from Inv-Gamma $\left[\frac{n}{2}, \frac{\sum_{i=1}^n (x_i - x_0)^2}{2}\right]$ with $\log(R_x) \in [-5, 2]$.
- Sample σ_{res}^2 from Inv-Gamma$[\frac{n}{2} + 0.003, \frac{\sum_{i=1}^n (M_i - M_0)^2}{2} + 0.003]$.

When MH updates are required in the samplers above, we use truncated normal distributions centered at the current draw with variance-covariance matrix adjusted to obtain an acceptance rate of around 40% (univariate) or 25% (multivariate). Truncations are applied according to prior constraints.

E

Algorithms IvD-2, IvD-2c, BN, and BNc in Chapter 6

E.1 Algorithms IvD-2 and IvD-2c

Algorithm IvD-2 does not marginalize α out when updating β. Thus Algorithm IvD-2 can be used when the prior mean of β, β_0, is not equal to zero, while Algorithm IvD-1 can not.

The error in Algorithm IvD-2 arises in Step 2(a), same as the error in Step 3(a) of Algorithm IvD-1. Thus Steps 0, 1, and 3 of Algorithms IvD-2 and IvD-2c are the same. In Step 2(a), however, Algorithm IvD-2c updates $\tilde{\Sigma}^\star$ by sampling from Inv-Wishart $\left(n+\nu, \sum_{i=1}^{n} \tilde{Z}_i \tilde{Z}_i^{\mathrm{T}} + \tilde{S}\right)$ subject to the constraint in (6.7), whereas Algorithm IvD-2 ignores the constraint. Note that $\beta^{(t+1)}$ in $\tilde{\xi}_i\left(\tilde{\sigma}_{11}^\star\right)$ of the constraint (6.7) should be replaced by $\beta^{(t)}$ in Algorithm IvD-2c. We display Algorithm IvD-2c here. Except for the boxed expression, Algorithm IvD-2 is identical to Algorithm IvD-2c.*

*In Algorithm IvD-2, the constraint in the box of Algorithm IvD-2c is ignored.

Algorithm IvD-2c (with the correction in the box)

```
Step 0: Initialize parameters $t = 0$, $\beta^{(0)}$, $\alpha^{(0)}$, $\Sigma^{(0)}$, and $W^{(0)}$.
while $t < T$ do
    Step 1: Update $\left((\alpha^2)^\star, \tilde{Z}\right)$ from $p(\alpha^2, \tilde{Z}|Y, \beta^{(t)}, \Sigma^{(t)})$ by
    (a) sampling $(\alpha^2)^\star$ from $p(\alpha^2|\Sigma^{(t)})$: $(\alpha^2)^\star \sim \alpha_0^2 \mathrm{trace}\left(S{\Sigma^{(t)}}^{-1}\right)/\chi^2_{\nu p}$; setting $\alpha^\star = \sqrt{(\alpha^2)^\star}$;
    (b) sampling $\tilde{Z}$ from $p(\tilde{Z}|Y, \alpha^\star, \beta^{(t)}, \Sigma^{(t)})$:
    for $i := 1, \ldots, n$ do
        for $k := 1, \ldots, p$ do
            sampling $W^\star_{ik}$ via $p(W_{ik}|Y_i, W^\star_{i,-k}, \beta^{(t)}, \Sigma^{(t)})$: $W^\star_{ik} \sim \mathrm{TN}(\mu_{ik}, \tau^2_{ik})$, see Appendix E.3 for details;
        end for
        Set $\tilde{Z}_i = \alpha^\star(W^\star_i - X_i\beta^{(t)})$.
    end for
    Step 2: Update $\left((\alpha^2)^{(t+1)}, \Sigma^{(t+1)}\right)$ via $p(\alpha^2, \Sigma|Y, \tilde{Z}, \beta^{(t)})$ by
     (a) sampling $\tilde{\Sigma}^\star$ from $p(\tilde{\Sigma}|Y, \tilde{Z}, \beta^{(t)})$:

        $\tilde{\Sigma}^\star \sim \text{Inv-Wishart}\left[n + \nu, \sum_{i=1}^{n} \tilde{Z}_i \tilde{Z}_i^{\mathrm{T}} + \tilde{S}\right], \boxed{\text{subject to the constraint in (6.7)}};$

     (b) setting $\alpha^{(t+1)} = \tilde{\sigma}^\star_{11}$, $\Sigma^{(t+1)} = \tilde{\Sigma}^\star/\left(\alpha^{(t+1)}\right)^2$, and $W_i^{(t+1)} = (\tilde{Z}_i + \alpha^{(t+1)}X_i\beta^{(t)})/\alpha^{(t+1)}$.
    Step 3: Update $\beta^{(t+1)}$ via $p(\beta|Y, W^{(t+1)}, \Sigma^{(t+1)})$:

        $\beta^{(t+1)} \sim \mathrm{N}_q\left[\hat{\beta}, \left(\sum_{i=1}^{n} X_i^{\mathrm{T}} {\Sigma^{(t+1)}}^{-1} X_i + A^{-1}\right)^{-1}\right],$

    where $\hat{\beta} = \left(\sum_{i=1}^{n} X_i^{\mathrm{T}} {\Sigma^{(t+1)}}^{-1} X_i + A^{-1}\right)^{-1} \left(\sum_{i=1}^{n} X_i^{\mathrm{T}} {\Sigma^{(t+1)}}^{-1} W_i^{(t+1)}\right)$.
    return $\beta^{(t+1)}$, $\Sigma^{(t+1)}$, and $W^{(t+1)}$
    $t + 1 \leftarrow t$
end while
```

E.2 Algorithms BN and BNc

Algorithm BN is almost the same as Algorithm IvD-1. The only difference is Step 3(b). Specifically, first, in Algorithm BN, α^2 in this step is set to trace$(\tilde{\Sigma})/p$, while in Algorithm IvD-1, α^2 is set to the first element of $\tilde{\Sigma}$; second, Algorithm BN sets $\beta = \tilde{\beta}/\alpha$ in Step 3(b), while Algorithm IvD-1 does not.

There are three errors in Algorithm BN and all of the errors appear in Step 3. Besides the same two errors as in Algorithm IvD-1, "$\beta^{(t+1)} = \tilde{\beta}^\star/\alpha^{(t+1)}$" in Step 3(b) of Algorithm BN should be removed, because we update $\tilde{\Sigma}^\star$ conditioning on $(Y, \tilde{Z}, \beta^{(t+1)})$, not on $(Y, \tilde{W}^\star, \tilde{\beta}^\star)$. Thus Steps 0, 1, and 2 of Algorithms BN and BNc are the same. However, Step 3(a) of Algorithm BNc updates $\tilde{\Sigma}^\star$ by sampling from a constrained inverse-

Algorithm BNc (with corrections in boxes)

```
Step 0: Initialize parameters $t = 0$, $\beta^{(0)}$, $\alpha^{(0)}$, $\Sigma^{(0)}$, and $W^{(0)}$.
while $t < T$ do
    Step 1: Update $\left((\alpha^2)^\star, \tilde{W}^\star\right)$ via $p(\alpha^2, \tilde{W}|Y, \beta^{(t)}, \Sigma^{(t)})$ by
    (a) sampling $(\alpha^2)^\star$ from $p(\alpha^2|\Sigma^{(t)})$: $(\alpha^2)^\star \sim \alpha_0^2 \mathrm{trace}\left(S{\Sigma^{(t)}}^{-1}\right)/\chi^2_{\nu p}$; setting $\alpha^\star = \sqrt{(\alpha^2)^\star}$;
    (b) sampling $\tilde{W}^\star$ from $p(\tilde{W}|Y, \alpha^\star, \beta^{(t)}, \Sigma^{(t)})$:
    for $i := 1, \ldots, n$ do
        for $k := 1, \ldots, p$ do
            sampling $W_{ik}^\star$ from $p(W_{ik}|Y_i, W_{i,-k}^\star, \beta^{(t)}, \Sigma^{(t)})$: $W_{ik}^\star \sim \mathrm{TN}(\mu_{ik}, \tau_{ik}^2)$, see Appendix E.3 for details;
        end for
        Set $\tilde{W}_i^\star = \alpha^\star W_i^\star$.
    end for
    Step 2: Update $\left((\alpha^2)^\star, \beta^{(t+1)}\right)$ via $p(\alpha^2, \beta|Y, \tilde{W}^\star, \Sigma^{(t)})$ by
     (a) sampling $(\alpha^2)^\star$ from $p(\alpha^2|Y, \tilde{W}^\star, \Sigma^{(t)})$:

         $(\alpha^2)^\star \sim \dfrac{\sum_{i=1}^n (\tilde{W}_i^\star - X_i\hat{\beta})^{\mathrm{T}} {\Sigma^{(t)}}^{-1} (\tilde{W}_i^\star - X_i\hat{\beta}) + \hat{\beta}^{\mathrm{T}} A^{-1}\hat{\beta} + \mathrm{trace}\left(\tilde{S}{\Sigma^{(t)}}^{-1}\right)}{\chi^2_{(n+\nu)p}}$,

    where $\hat{\beta} = \left(\sum_{i=1}^n X_i^{\mathrm{T}} {\Sigma^{(t)}}^{-1} X_i + A^{-1}\right)^{-1} \left(\sum_{i=1}^n X_i^{\mathrm{T}} {\Sigma^{(t)}}^{-1} \tilde{W}_i^\star\right)$;
     (b) sampling $\tilde{\beta}^\star$ from $p(\tilde{\beta}|Y, \tilde{W}^\star, (\alpha^2)^\star, \Sigma^{(t)})$:

         $\tilde{\beta}^\star \sim \mathrm{N}_q \left[\hat{\beta}, (\alpha^2)^\star \left(\sum_{i=1}^n X_i^{\mathrm{T}} {\Sigma^{(t)}}^{-1} X_i + A^{-1}\right)^{-1}\right]$;

    setting $\alpha^\star = \sqrt{(\alpha^2)^\star}$ and $\beta^{(t+1)} = \tilde{\beta}^\star/\alpha^\star$.
    Step 3: Update $\left((\alpha^2)^{(t+1)}, \Sigma^{(t+1)}\right)$ via $p(\alpha^2, \Sigma|Y, \tilde{W}^\star, \beta^{(t+1)})$ by
     (a) sampling $\tilde{\Sigma}^\star$ from $p(\tilde{\Sigma}|Y, \tilde{Z}, \beta^{(t+1)})$:

         $\tilde{\Sigma}^\star \sim \text{Inv-Wishart}\left(n + \nu, \sum_{i=1}^n \tilde{Z}_i \tilde{Z}_i^{\mathrm{T}} + \tilde{S}\right), \boxed{\text{subject to the constraint in (6.7}^\star\text{)}}$,

    where $\tilde{Z}_i = \tilde{W}_i^\star - \alpha^\star X_i \beta^{(t+1)}$;
     (b) setting $\alpha^{(t+1)} = \sqrt{\mathrm{trace}(\tilde{\Sigma}^\star/p)}$, $\Sigma^{(t+1)} = \tilde{\Sigma}^\star/\left(\alpha^{(t+1)}\right)^2$, and $\boxed{W_i^{(t+1)} = (\tilde{Z}_i + \alpha^{(t+1)} X_i \beta^{(t+1)})/\alpha^{(t+1)}}$.
    return $\beta^{(t+1)}$, $\Sigma^{(t+1)}$, and $W^{(t+1)}$
    $t + 1 \leftarrow t$
end while
```

Wishart distribution, that is,

$$\tilde{\Sigma}^{\star} \sim \text{Inv-Wishart}\left(n+\nu, \sum_{i=1}^{n} \tilde{Z}_i \tilde{Z}_i^{\mathrm{T}} + \tilde{S}\right) \text{ subject to the constraint } (6.7^{\star}),$$

whereas Algorithm BN ignores the constraint. The constraint $(6.7^{\star})$ is an adaptation of (6.7) by replacing $\tilde{\sigma}_{11}^{\star}$ with $r = \sqrt{\text{trace}(\tilde{\Sigma}^{\star}/p)}$. Specifically, $\tilde{\xi}_i(r) = \tilde{Z}_i + rX_i\beta^{(t+1)}$, for $i = 1, \ldots, n$. The updated value of r must satisfy

$$\begin{cases} \max\left\{\tilde{\xi}_{i1}(r), \ldots, \tilde{\xi}_{ip}(r)\right\} < 0 & \text{if } Y_i = 0 \\ \max\left\{0, \tilde{\xi}_{i1}(r), \ldots, \tilde{\xi}_{ip}(r)\right\} = \tilde{\xi}_{ik}(r) & \text{if } Y_i = k \end{cases}, \quad \text{for} \quad i = 1, \ldots, n. \qquad (6.7^{\star})$$

Finally, in Step 3(b), Algorithm BNc sets $\alpha^{(t+1)} = \sqrt{\text{trace}(\tilde{\Sigma}^{\star}/p)}$, $\Sigma^{(t+1)} = \tilde{\Sigma}^{\star}/\big(\alpha^{(t+1)}\big)^2$, and $W_i^{(t+1)} = (\tilde{Z}_i + \alpha^{(t+1)} X_i \beta^{(t+1)})/\alpha^{(t+1)}$, while Algorithm BN sets $\alpha^{(t+1)} = \sqrt{\text{trace}(\tilde{\Sigma}^{\star}/p)}$, $\Sigma^{(t+1)} = \tilde{\Sigma}^{\star}/\big(\alpha^{(t+1)}\big)^2$, $W^{(t+1)} = \tilde{W}^{\star}/\alpha^{(t+1)}$, and additionally, $\beta^{(t+1)} = \tilde{\beta}^{\star}/\alpha^{(t+1)}$. We display Algorithm BNc here. Except for Step 3, Algorithm BN is identical to Algorithm BNc.[†]

E.3 Details of Sampling W in Step 1(b) of Algorithms IvD-1, IvD-1c, IvD-2, IvD-2c, BN, and BNc

Updating W in Step 1(b) of Algorithms IvD-1, IvD-1c, IvD-2, IvD-2c, BN, and BNc consists of sampling from a series of univariate truncated normal distributions, that is, for $i = 1, \ldots, n$ and $k = 1, \ldots, p$,

$$W_{ik}^{\star} \sim \text{TN}(\mu_{ik}, \tau_{ik}^2),$$

where $\mu_{ik} = X_{ik}\beta^{(t)} + \Sigma_{k,-k}^{(t)} \Sigma_{-k,-k}^{(t)-1} (W_{i,-k}^{\star} - X_{i,-k}\beta^{(t)})$ with $W_{i,-k}^{\star} = (\, W_{i1}^{\star}, \ldots, W_{i,(k-1)}^{\star}, W_{i,(k+1)}^{(t)}, \ldots, W_{ip}^{(t)}\,)$, and $\tau_{ik}^2 = \left(\sigma_{kk}^{(t)}\right)^2 - \Sigma_{k,-k}^{(t)} \Sigma_{-k,-k}^{(t)-1} \Sigma_{-k,k}^{(t)}$; X_{ik} is the kth row of X_i, and $X_{i,-k}$ is the sub-matrix of X_i with X_{ik} removed. The constraint on $W_{ik}^{\star}$ is, $W_{ik}^{\star} \geq \max\{0, W_{i,-k}^{\star}\}$, if $Y_i = k$; $W_{ik}^{\star} < \max\{0, W_{i,-k}^{\star}\}$, if $Y_i \neq k$.

If the constraint on $W_{ik}^{\star}$ has the form, $W_{ik}^{\star} \geq w$, and $w \leq 0$, we update $W_{ik}^{\star}$ with simple

[†] In Algorithm BN, the constraint in the first box of Algorithm BNc is ignored, the expression in the second box is replaced by $W^{(t+1)} = \tilde{W}^{\star}/\alpha^{(t+1)}$, and $\beta^{(t+1)} = \tilde{\beta}^{\star}/\alpha^{(t+1)}$ is added at the end of Step 3(b).

rejection sampling: we iteratively sample from the unconstrained normal distribution until $W_{ik}^{\star} \geq w$ is satisfied. If $W_{ik}^{\star} \geq w$, but $w > 0$, we update $W_{ik}^{\star}$ with the exponential rejection sampling proposed by Robert (1995). Otherwise, if the constraint on $W_{ik}^{\star}$ has the form $W_{ik}^{\star} \leq w$, we can apply the above sampling scheme with slight adaptation, since $-W_{ik}^{\star} \geq -w$.