Essays on Bayesian Computation with Ancillarity-Sufficiency Interweaving Strategy

Makoto Nakakita

Centre for Finance, Technology and Economics at Keio

Keio University

Abstract

Researchers have developed more complex models for more realistic data analysis. In general, model complexity tends to increase computational burdens in terms of both computing time and memory/storage usage. As for Bayesian statistics in particular, the model complexity makes statistical inference with the posterior distribution almost intractable and impractical. To tackle this problem, numerous computational methods have been developed since the late 20th century. Among them, the most prominent ones are Markov chain Monte Carlo (MCMC) methods such as the Gibbs sampler (Geman and Geman (1984); Gelfand and Smith (1990)), the Metropolis-Hastings (MH) method (Metropolis *et al.* (1953); Hastings (1970)) and the data augmentation (Tanner and Wong (1987)). Since the adoption of MCMC in late 1980s, computational Bayesian statistics has attracted more attention for its ability to deal with highly complex problems that were previously unsolvable.

Even after the extraordinary progress of computers in the last several decades, however, naive implementation of MCMC methods is still insufficient to handle the increasing complexity of statistical models. For a high-dimensional complex model, random series of model parameters drawn from the posterior distribution with MCMC often exhibits strong positive autocorrelation. Since such high autocorrelation causes slow convergence to the true posterior distribution, acceptable precision of the posterior statistics cannot be achieved in practice. Although it may be possible to solve this problem by generating a huge sample from the posterior distribution through gigantic computer processing power (e.g., supercomputers), this is neither practical nor eco-friendly.

Numerous studies have been conducted to improve the sampling efficiency of MCMC. Among them, the ancillarity-sufficiency interweaving strategy (ASIS) proposed by Yu and Meng (2011) is an easy-to-implement and widely applicable sampling algorithm for

improving the sampling efficiency of MCMC. In principle, ASIS samples random series of parameters and latent variables alternately from the posterior distribution with centralized parametrization (CP) or from the one with non-centralized parametrization (NCP). Whether CP outperforms NCP or not depends on models and data sets and we cannot precisely say which is better in general. ASIS tries to solve this problem by combining two types of parametrization in one sampling cycle of MCMC.

In our doctoral dissertation, we will focus on applications of ASIS to several models used in applied econometric analysis and demonstrate the efficacy of ASIS in Bayesian computation. The organization of our dissertation is as follows.

In Chapter 2, the formal definition of ASIS and an illustrative example of its application will be presented. As the example, we estimated a panel data regression model with wage data used by Vella and Verbeek (1998), which is a balanced panel data set, to check whether ASIS can improve the sampling efficiency of random series from the posterior distribution and the precision of Monte Carlo evaluation of the posterior means and variances. As a result, we found that ASIS could help improve MCMC efficiency in panel data regression analysis.

In Chapter 3, we propose a hierarchical Bayesian model of evaluating horse ability and jockey skills in horse racing and estimate it with ASIS. In the proposed method, we aim to estimate unobservable individual effects of horses and jockeys simultaneously along with regression coefficients for explanatory variables such as horse age, racetrack conditions and others in the regression model. The data used in this study are records on 1800-m races (excluding steeplechases) held by the Japan Racing Association from 2016 to 2018, including 22,183 runs with 4,063 horses and 143 jockeys. Since the number of entries varies by racehorses and jockeys, unlike the example in Chapter 2, it is an unbalanced panel data set. We apply the hierarchical Bayesian model to stably estimate such a large amount of individual effects. Since some racehorses and jockeys have extremely small numbers of runs, it is difficult to make stable estimation of the individual effects with conventional sampling methods. Hence, we use the Gibbs sampling coupled with ASIS for Bayesian estimation of the model and choose the best model with the widely applicable information criterion (WAIC) as a model selection criterion. As a result, we found a large difference in the ability among horses and jockeys. Additionally, we observed a

strong relationship between the individual effects and the race records for both horses and jockeys.

In Chapter 4, we apply ASIS to complex stochastic volatility (SV) models with high-frequency intraday time series data of stock returns. Since intraday financial data tend to have time-dependent characteristics such as volatility clustering and intraday seasonality, it is crucial to properly capture them. Our modeling strategy is two-fold. First, we model the intraday seasonality of return volatility as a Bernstein polynomial and estimate it along with the stochastic volatility simultaneously. Second, we incorporate skewness and excess kurtosis of stock returns into the SV model by assuming that the error term follows a family of generalized hyperbolic distributions including variance-gamma and Student's t distributions. Furthermore, we developed an efficient MCMC sampling algorithm for Bayesian inference of the proposed model. To improve efficiency of MCMC implementation, we apply ASIS and generalized Gibbs sampling. As a demonstration of our new method, we estimated intraday SV models with 1-minute return data of a stock price index (TOPIX), and conducted model selection among various specifications with WAIC. The result shows that the SV model with the skew variance-gamma error is the best among the candidates.

In Chapter 5, we provide the summary of this dissertation and remarks on the future prospects of further research.

Acknowledgements

My most profound appreciation goes to my supervisor, Professor Teruo Nakatsuma, for his time, effort, and understanding in helping me succeed in my studies. His vast knowledge of Bayesian statistics, economics and computation, and wealth of experience have inspired me throughout my studies. He is, needless to say, a top-notch researcher, but his personality is also very respectable and I truly honor him as a person.

In addition, I would like to express my deepest gratitude to my Ph.D. and preliminary exam committee members, Professor Takahiro Hoshino and Professor Daisuke Nagakura for their support. Without their sharp suggestions and tips, none of my studies would have been successful. Also, I would like to extend my sincere thanks to my colleague graduate students and members in Econometric Seminar.

Finally, I am also thankful to my family members for persistent support and their love.

Contents

1 Introduction **1**

2 Ancillarity-Sufficiency Interweaving Strategy (ASIS) **5**

2.1 Introduction 5

2.2 Definition of ASIS 7

2.3 Hierarchical Bayesian Modeling of Panel Data 11

2.4 Application to Wage Panel Data 16

2.5 Conclusion 26

3 Application to Unbalanced Panel Data: Horse Race Results in Japan **29**

3.1 Introduction 29

3.2 Data 34

3.3 Hierarchical Bayesian Modeling 41

3.4 Results and Discussion on Empirical Analysis 48

3.5 Conclusion 62

4 Application to Time Series Data: High Frequency Stock Returns **65**

4.1 Introduction 65

4.2 Stochastic Volatility Model with Intraday Seasonality 68

4.3 Conditional Posterior Distributions 75

4.3.1 NCP Form 75

4.3.2 CP Form 83

4.4 Extension: Skew Heavy-Tailed Distributions 86

4.4.1 Mean-Variance Mixture of the Normal Distribution 86

4.4.2 Conditional Posterior Distributions 88

4.5 Empirical Study 92
4.6 Conclusion 101

5 Summary and Future Prospects **103**

A Derivations of Conditional Posterior Distributions **107**

Chapter 1

Introduction

Researchers in the field of data science have developed more complex and sophisticated models for more realistic data analysis. In general, model complexity tends to increase computational burdens in terms of both computing time and memory/storage usage. As for Bayesian statistics in particular, the model complexity not only raises such burdens prohibitively high but also makes statistical inference with the posterior distribution almost intractable and impractical. To tackle this problem, numerous computational methods have been developed since the late 20th century. Among them, the most prominent ones are Markov chain Monte Carlo (MCMC) methods such as the Gibbs sampler (Geman and Geman (1984); Gelfand and Smith (1990)) and the Metropolis-Hastings (MH) method (Metropolis *et al.* (1953); Hastings (1970)). Since the adoption of MCMC in late 1980s, with the further development of computers, computational Bayesian statistics has attracted more attention of academic researchers as well as practitioners for its ability to deal with highly complex problems that were previously unsolvable.

Even after the extraordinary progress of computers in the last several decades, however, naive implementation of MCMC methods is still insufficient to handle the increasing complexity of statistical models. For a high-dimensional complex model, random series of model parameters drawn from the posterior distribution with MCMC often exhibits strong positive autocorrelation. Since such high autocorrelation causes slow convergence to the true posterior distribution, acceptable precision of the posterior statistics cannot be achieved in a meaningful time even by a high-powered workstation. Although it may be possible to solve this problem by generating a huge sample from the posterior distribu-

tion through gigantic computer processing power (e.g., supercomputers), this is neither practical nor eco-friendly. In addition, in fields such as national security and disaster forecasting, where computing time is a crucial factor, the convergence inefficiency would be a major issue.

Numerous studies have been conducted to improve the sampling efficiency of MCMC. Among them, the ancillarity-sufficiency interweaving strategy (ASIS) proposed by Yu and Meng (2011) is an easy-to-implement and widely applicable sampling algorithm for improving the sampling efficiency of MCMC. ASIS is a strategy that alternately samples parameters and latent variables from the posterior distribution with centralized parametrization (CP) or from the one with non-centralized parametrization (NCP). Whether CP outperforms NCP or not depends on models and data sets and we cannot precisely say which is better in general. ASIS tries to solve this problem by combining two types of parametrization in one sampling cycle of MCMC.

In our doctoral dissertation, we will focus on applications of ASIS to several models used in applied econometric analysis and demonstrate the efficacy of ASIS in Bayesian computation. The organization of our dissertation is as follows.

- Chapter 2, *Ancillarity-Sufficiency Interweaving Strategy (ASIS)*

The purpose of this chapter is to clarify how ASIS is defined and why ASIS works, and to present a simple example of how ASIS works to improve the convergence efficiency of MCMC sampling. As the illustrative example to demonstrate improvement in the convergence rate due to ASIS, we applied ASIS to hierarchical Bayesian modeling of a panel data regression model with wage data used by Vella and Verbeek (1998) and compared estimation results, sample autocorrelation and other aspects of convergence with and without ASIS.

It is barely recognized among researchers that ASIS can be applied to the panel data regression analysis. As far as the author knows, Gelfand *et al.* (1995) first proposed to use the centered parametrization (CP) form of a panel data regression model. They showed that centering the parameters would make a significant difference in convergence rate as well as stability of sampled series. What Gelfand *et al.* (1995) applied in parametrization, however, was not ASIS but mere centralization of the hierarchical prior for regression

coefficients in the panel data regression model. ASIS, on the other hand, combines random number generation from both CP form and NCP form into a unified sampling scheme.

By comparing generated random series with and without ASIS in terms of sample autocorrelation and convergence diagnostic, we demonstrated that ASIS not only improved the convergence rate, but also enabled the convergence itself, which seems impossible for the original parametrization.

- Chapter 3, *Application to Unbalanced Panel Data: Horse Race Results in Japan*

In Chapter 2, we showed that ASIS worked well in panel data analysis with simple and well-prepared balanced panel data datasets. Contrarily, this chapter aims to show that ASIS also works well for panel data analysis with more complex model specifications using unbalanced panel data; horse race running data in Japan. The data used in this study are records on 1800-m races (excluding steeplechases) held by the Japan Racing Association from 2016 to 2018, including 4,063 horses and 143 jockeys.

With the horse race data, we proposed a new method of evaluating horse ability and jockey skills in horse racing. In the proposed method, we aimed to estimate unobservable individual effects of horses and jockeys simultaneously with regression coefficients for explanatory variables such as horse age, racetrack conditions and others in the regression model. We applied a hierarchical Bayesian model to stably estimate such a large amount of the individual effects. We used the MCMC method coupled with ASIS for Bayesian estimation of the model and chose the best model with the widely applicable information criterion (WAIC) by Watanabe (2010) as a model selection criterion. Since some of these racehorses and jockeys have a small number of runs (the minimum number of runs is four), plain-vanilla MCMC was not able to estimate the individual effects stably. However, the application of ASIS made this estimation possible, indicating that ASIS is particularly effective for high-dimensional and small sample-size data. As a result, we found a large difference in the ability among horses and jockeys. Additionally, we observed a strong relationship between the individual effects and the race records for both horses and jockeys.

- Chapter 4, *Application to Time Series Data: High Frequency Stock Returns*

This chapter is based on Nakakita and Nakatsuma (2021). In Chapter 4, we intend to show that ASIS can improve the MCMC efficiency not only in the framework of panel data analysis, but also in the framework of time-series data analysis. As a framework for time series analysis, we employed complex stochastic volatility (SV) models of high frequency intraday stock returns. High frequency intraday data of stock returns exhibit not only typical characteristics (e.g., volatility clustering and leverage effect) but also a cyclical pattern of return volatility that is known as intraday seasonality. In this chapter, we extend the SV model for application with high frequency intraday financial time series data and develop an efficient MCMC sampling algorithm for Bayesian inference of the proposed model. As a demonstration of our new method, we estimated intraday SV models with 1-minute return data of a stock price index (TOPIX).

Our contributions in this chapter may be summarized as follows. First, we successfully captured a feature of high frequency intraday stock returns called intraday seasonality by using Bernstein polynomials. Second, we extended the error term of the SV model to a family of generalized hyperbolic distributions including variance-gamma and Student's t distributions, as well as skew variance gamma and skew t distributions. Needless to say, as the distributions used for the model become more complex, computational efficiency and convergence rate of generated random series become more crucial. Therefore, we applied ASIS to efficiently generate random series and estimated the posterior distribution. As a result, the application of ASIS successfully reduced the sample autocorrelation of (latent) volatility and parameters in the model.

Finally, we provide the summary of this dissertation and comments on the future prospects of further research in Chapter 5.

Chapter 2

Ancillarity-Sufficiency Interweaving Strategy (ASIS)

2.1 Introduction

In the recent development of statistics, data science and other related fields, model specifications used in data analysis have been becoming more and more complex over years. In particular, most of probabilistic models used in Bayesian inference, which seem conceptually sound, are computationally infeasible for application. To deal with such complex models, a simulation-based approach called Markov chain Monte Carlo (MCMC) has been developed since the late 20th century. The MCMC approach includes the Gibbs sampler (Geman and Geman (1984); Gelfand and Smith (1990)), the Metropolis-Hastings (MH) method (Metropolis *et al.* (1953); Hastings (1970)), During the same time period, computing technology made a great leap forward. Thanks to so-called Moore's law, computers are becoming faster and having more memory and storage. A smartphone of today is far more capable than a supercomputer in 1980s.

Unfortunately, however, even the ground-breaking invention of MCMC and the rapid evolution of computing technology are not sufficient to tackle increasing complexity of statistical models. One problem we often encounter in MCMC implementation for a large-scale complex model with many parameters is strong positive autocorrelation in generated random series. By construction of MCMC, generated random series are realizations of a Markov chain whose stationary density is equivalent to the posterior density. Therefore

the generated random series are no longer independent, though it is well known that the law of large numbers still holds for those random series under certain conditions. See Robert and Casella (2004) for more details. In theory, any random series drawn with MCMC will eventually converge to the true posterior distribution and the sample statistics of the series such as the sample mean and the sample variance will also converge to the true ones in the posterior distribution due to the law of large numbers. In practice, however, the convergence rate of the sample statistics depends on mixing of generated random series. If autocorrelation in the random series is strongly positive even for longer lags, the convergence of the sample statistics will become rather slow. Thus the precision of the sample statistics with the MCMC sample as estimates of the true posterior statistics is possibly far worse than those with the i.i.d. sample, and we need a much larger sample of generated parameters.

As a possible solution to the problem of slow convergence, Yu and Meng (2011) proposed the ancillarity-sufficiency interweaving strategy (ASIS). ASIS is a strategy that alternately samples parameters with and without centralization ("interweaving"), thus taking advantages of both convergence and mixing of random series. ASIS is a very simple tool that can be adaptable to a wide variety of models and data, and has shown significant efficiency improvements in many cases.

In this chapter, we first introduce the basic idea behind ASIS and explain why ASIS works. Then, as an illustrative example, we derived a ASIS sampling scheme for hierarchical Bayesian analysis of a panel data regression model and apply it to real-world data (wage panel data used by Vella and Verbeek (1998)) for demonstrating how ASIS improves the sampling efficiency of MCMC for the panel data regression model.

Application of ASIS in the context of panel data regression is little studied in the literature of MCMC. Gelfand *et al.* (1995) first proposed to use an alternative parametrization of a panel data regression model for facilitating mixing of random series and improving the convergence rate. In fact, their reparametrizing approach is related to the centered parametrization form of the panel data regression model, which we will describe later in this section. They showed that the reparametrization would make a significant difference in convergence rate and improve results of the convergence diagnostic (e.g., Gelman and Rubin (1992)) of generated random series.

In our study, on the other hand, we propose to interweave the Gibbs sampling scheme based on the centered parametrization by Gelfand *et al.* (1995) with the widely-applied Gibbs sampling scheme based on the non-centered parametrization, and form an ASIS sampling scheme for the panel data regression model. In our best knowledge, no previous studies have applied ASIS to hierarchical Bayesian analysis of the panel data regression model.

The rest of this chapter is organized as follows. Section 2.2 describes the definition and principle of ASIS by using a two-level normal hierarchical model as an example. Section 2.3 explains the ASIS scheme for the panel data regression model. Section 2.4 presents the results of empirical comparisons between the ASIS scheme and the non-ASIS Gibbs sampling scheme for the panel data regression model with the wage panel data. Finally, conclusion of this chapter is given in Section 2.5.

2.2 Definition of ASIS

To illustrate the definition and procedure of ASIS, we cite a two-level normal hierarchical model from Yu and Meng (2011), which was originally used in Liu and Wu (1999),

$$Y_{obs}|\theta, Y_{mis} \sim \mathcal{N}\left(Y_{mis}, 1\right), \tag{2.1}$$

$$Y_{mis}|\theta \sim \mathcal{N}(\theta, V), \tag{2.2}$$

where Y_{obs} is the observed value while Y_{mis} is the missing value. We suppose the expected value of Y_{mis}, θ, is unknown but the variance of Y_{mis}, V, is a known constant. Thus the hierarchical model (2.1)–(2.2) have two unknown quantities (Y_{mis}, θ).

The standard data augmentation (DA) algorithm for (Y_{mis}, θ) draws them from $Y_{mis}|\theta, Y_{obs}$ and $\theta|Y_{mis}, Y_{obs}$ alternately as

$$Y_{mis}|\theta, Y_{obs} \sim \mathcal{N}\left(\frac{VY_{obs}+\theta}{1+V}, \frac{V}{1+V}\right), \tag{2.3}$$

$$\theta|Y_{mis}, Y_{obs} \sim \mathcal{N}(Y_{mis}, V). \tag{2.4}$$

Because the right-hand side of (2.1) is free of θ, Yu and Meng (2011) call the DA scheme (2.3)–(2.4) *sufficient augmentation* (SA), since Y_{mis} is a sufficient statistic for θ in the

augmented-data model. Note that, in the SA setting, the augmented-data posterior distribution of θ in (2.4) depends on Y_{mis} alone.

Contrarily, if we define

$$\tilde{Y}_{mis} = Y_{mis} - \theta, \tag{2.5}$$

and treat $\tilde{Y}_{mis}$ as the missing value, the model can be rewritten as

$$Y_{obs}|\theta, \tilde{Y}_{mis} \sim \mathcal{N}\ (\tilde{Y}_{mis} + \theta, 1), \tag{2.6}$$

$$\tilde{Y}_{mis} \sim \mathcal{N}(0, V), \tag{2.7}$$

which gives a different DA algorithm:

$$\tilde{Y}_{mis}|\theta, Y_{obs} \sim \mathcal{N}\left(\frac{V(Y_{obs} - \theta)}{1+V}, \frac{V}{1+V}\right), \tag{2.8}$$

$$\theta|\tilde{Y}_{mis}, Y_{obs} \sim \mathcal{N}(Y_{obs} - \tilde{Y}_{mis}, 1). \tag{2.9}$$

Yu and Meng (2011) call the DA scheme (2.8)–(2.9) *ancillary augmentation* (AA), because the distribution of $\tilde{Y}_{mis}$ in (2.7) is free of θ and $\tilde{Y}_{mis}$ is an ancillary statistic for θ. As we can see in (2.6) and (2.7), $\tilde{Y}_{mis}$ and θ are independent a priori in the AA setting.

You may wonder about the fact that the posterior distributions of (2.4) and (2.9) are essentially equivalent in the sense that we can obtain (2.9) by centering (2.4) to zero. Since random series generated from either SA (2.3)–(2.4) or AA (2.8)–(2.9) will eventually converge to the same posterior distribution, it seems that the sampling efficiency is not affected by using SA or AA at first glance. On the contrary to this naive intuition, the convergence rates of two algorithms are usually different.

Previous studies such as Dempster, Laird and Rubin (1977) discussed that the convergence rate of EM-type algorithms depended on "the fraction of missing information." van Dyk and Meng (2001, 2010) extended this idea to the convergence behavior of DA-type algorithms and found that there was a difference between the convergence rates of the centralized form and non-centralized form of DA. In the two-level normal hierarchical model, when the variance V is smaller, Y_{mis} is more informative about θ for SA because the distribution in (2.2) concentrates around the mean θ. For AA, on the other hand, a smaller V implies that we know more about $\tilde{Y}_{mis}$ since the distribution in (2.4) concentrates around zero. In the extreme case of $V \approx 0$, we know $\tilde{Y}_{mis} \approx 0$ for sure. Therefore

we expect SA to be slow in convergence but AA to be fast when V is small, and vice versa when V is large.

To exploit the aforementioned relationship between SA and AA, Yu and Meng (2011) proposed an ASIS algorithm that interweaves SA and AA together, which is called the *global interweaving strategy* or GIS. A general form of GIS is given as follows.

Global Interweaving Strategy (GIS)

Step 1-a: Draw $Y_{mis}^{(t+0.5)}$ given $\theta^{(t)}$ from $p(Y_{mis}|\theta^{(t)})$.

Step 1-b: Draw $\theta^{(t+0.5)}$ given $Y_{mis}^{(t+0.5)}$ from $p(\theta|Y_{mis}^{(t+0.5)})$.

Step 2-a: Draw $\tilde{Y}_{mis}^{(t+1)}$ given $(Y_{mis}^{(t+0.5)}, \theta^{(t+0.5)})$ from $p(\tilde{Y}_{mis}|Y_{mis}^{(t+0.5)}, \theta^{(t+0.5)})$.

Step 2-b: Draw $\theta^{(t+1)}$ given $\tilde{Y}_{mis}^{(t+1)}$ from $p(\theta|\tilde{Y}_{mis}^{(t+1)})$.

where $(Y_{mis}, \tilde{Y}_{mis}, \theta)$ are the missing values in SA (or AA), the missing values in AA (or SA) and the unknown parameters in the model respectively.

If joint sampling of θ from $p(\theta|Y_{mis})$ or $p(\theta|\tilde{Y}_{mis})$ is difficult, we may apply Gibbs sampling to each element or subgroup of θ. Yu and Meng (2011) called this type of ASIS the *component-wise interweaving strategy* or CIS. Basically, CIS breaks down **Step 1-b** and **Step 2-b** in GIS into a series of Gibbs sampling from the conditional distributions of elements/subgroups in θ.

In many applications, the conditional distribution in **Step 2-a** determines $\tilde{Y}_{mis}$ completely given (Y_{mis}, θ). For example, in the two-level normal hierarchical model (2.1)–(2.2), **Step 2-a** is equivalent to $\tilde{Y}_{mis}^{(t+1)} = Y_{mis}^{(t+0.5)} - \theta^{(t+0.5)}$. Thus a GIS algorithm for the two-level normal hierarchical model is given as follows.

GIS for the Two-level Normal Hierarchical Model

Step 1-a: Draw $Y_{mis}^{(t+0.5)}$ given $\theta^{(t)}$ from $p(Y_{mis}|\theta^{(t)})$ in (2.4).

Step 1-b: Draw $\theta^{(t+0.5)}$ given $Y_{mis}^{(t+0.5)}$ from $p(\theta|Y_{mis}^{(t+0.5)})$ in (2.3).

Step 2-a: Let $\tilde{Y}_{mis}^{(t+1)} = Y_{mis}^{(t+0.5)} - \theta^{(t+0.5)}$.

Step 2-b: Draw $\theta^{(t+1)}$ given $\tilde{Y}_{mis}^{(t+1)}$ from $p(\theta|\tilde{Y}_{mis}^{(t+1)})$ in (2.9).

Let us verify that the GIS sampling chain preserves the stationary density $p(\theta)$, i.e.,

$$p(\theta') = \int k(\theta'|\theta)p(\theta)d\theta,$$

where $p(\theta)$ is shared by the original two DA algorithms (SA and AA), $k(\theta'|\theta)$ is the transition kernel of GIS:

$$k(\theta'|\theta) = \int\int p(\theta'|\tilde{Y}_{mis})p(\tilde{Y}_{mis}|Y_{mis})p(Y_{mis}|\theta)dY_{mis}d\tilde{Y}_{mis}, \tag{2.10}$$

and $p(\tilde{Y}_{mis}|Y_{mis}) = \int p(\tilde{Y}_{mis}|Y_{mis},\tilde{\theta})p(\tilde{\theta}|Y_{mis})d\tilde{\theta}$.

Suppose $p(Y_{mis},\tilde{Y}_{mis},\theta)$ is well-defined, and $p(Y_{mis},\theta)$ and $p(\tilde{Y}_{mis},\theta)$ are the stationary densities of DA chains (SA or AA) that share the same $p(\theta)$. By Fubini's theorem, we have

$$\begin{aligned}
\int k(\theta'|\theta)p(\theta)d\theta &= \int\int p(\theta'|\tilde{Y}_{mis})p(\tilde{Y}_{mis}|Y_{mis})\left[\int p(Y_{mis}|\theta)p(\theta)d\theta\right]dY_{mis}d\tilde{Y}_{mis} \\
&= \int p(\theta'|\tilde{Y}_{mis})\left[\int p(\tilde{Y}_{mis}|Y_{mis})p(Y_{mis})dY_{mis}\right]d\tilde{Y}_{mis} \\
&= \int p(\theta'|\tilde{Y}_{mis})p(\tilde{Y}_{mis})d\tilde{Y}_{mis} \\
&= p(\theta').
\end{aligned}$$

Therefore $p(\theta)$ is the stationary density of the transition kernel (2.10).

Finally, we cite two theorems given by Yu and Meng (2011) to show that ASIS is useful for improving MCMC efficiency.

Theorem 1 (Yu and Meng (2011, p.539))**.** *Given a posterior distribution of interest $p(\theta|Y_{obs}), \theta \in \Theta$, suppose we have two augmentation schemes $Y_{mis,1}$ and $Y_{mis,2}$ such that their joint distribution, conditioning on both θ and Y_{obs}, is well defined for $\theta \in \Theta$ (almost surely with respect to $p(\theta|Y_{obs})$). Denote the geometric rate of convergence of the DA algorithm under $Y_{mis,i}$ by $r_i, i = 1, 2$, which are allowed to take value 1 (i.e., being sub-geometric). Then*

$$r_{1\&2} \leq \mathcal{R}_{1,2}\sqrt{r_1 r_2} \tag{2.11}$$

where $r_{1\&2}$ is the geometric rate of the GIS sampler interweaving $Y_{mis,1}$ and $Y_{mis,2}$, and $\mathcal{R}_{1,2}$ is the maximal correlation between $Y_{mis,1}$ and $Y_{mis,2}$ in their joint posterior distribution $p(Y_{mis,1}, Y_{mis,2}|Y_{obs})$.

Theorem 2 (Yu and Meng (2011, p.561))**.** *Given a posterior distribution $p(\theta|Y_{obs})$ of interest, suppose we have two augmentation schemes $Y_{mis,1}$ and $Y_{mis,2}$ such that their joint distribution in well defined conditional on θ and Y_{obs}. Let $\mathcal{N} = \sigma(Y_{mis,1}) \cap \sigma(Y_{mis,2})$, that is, the intersection of the σ-algebras generated by $Y_{mis,1}$ and $Y_{mis,2}$ in the joint posterior of $(\theta, Y_{mis,1}, Y_{mis,2})$. Then $r_{1\&2}$, the geometric rate of convergence of GIS, satisfies*

$$r_{1\&2} \leq \mathcal{R}^2(\theta, \mathcal{N}) + (1 - \mathcal{R}^2(\theta, \mathcal{N}))\mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,1})\mathcal{R}_{\mathcal{N}}(Y_{mis,1}, Y_{mis,2})\mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,2}). \quad (2.12)$$

Theorem 1 and 2 show that ASIS improves the convergence rate. We leave the details of these theorems to Yu and Meng (2011).

2.3 Hierarchical Bayesian Modeling of Panel Data

In the rest of this chapter, we examine the efficacy of ASIS in hierarchical Bayesian modeling of panel data. For this purse, in this section, we first introduce the panel data regression model with individual effects and then derive the ASIS scheme for its hierarchical Bayesian analysis as well as its counterpart in Gibbs sampling as the benchmark.

Suppose we have panel data about N individuals recorded over T periods. Let $i \in \{1, \ldots, N\}$ be the index for each individual and $t \in \{1, \ldots, T\}$ be the index for each period. Furthermore, we suppose the panel data have the following variables:

- y_{it} — dependent variable for individual i in period t
- $\boldsymbol{x}_{it}$ — $K \times 1$ vector of independent variables for individual i in period t

Consider the regression model of panel data:

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad i \in \{1, \ldots, N\}, \quad t \in \{1, \ldots, T\}, \quad (2.13)$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector of regression coefficients. α_i in (2.13) is interpreted as an unobservable variable that only affects the dependent variable of individual i and cannot be explained by other factors. We call α_i the *individual effect*.

Let us define vectors and matrices as follows.

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{iT} \end{bmatrix}, \quad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT} \end{bmatrix}.$$

Then, the regression model in (2.13) can be summarized as

$$\boldsymbol{y}_i = \alpha_i \mathbf{1} + \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}), \tag{2.14}$$

where $\mathbf{1}$ is a vector whose elements are all ones, $\mathbf{0}$ is a a vector whose elements are all zeros, and $\boldsymbol{I}$ is the identity matrix. For simplicity, we will not add any indicator of dimension to $\mathbf{1}$, $\mathbf{0}$ or $\boldsymbol{I}$ as long as the dimension is obvious in the context. By defining

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} \mathbf{1} & & \\ & \ddots & \\ & & \mathbf{1} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix},$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{X} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix},$$

we have

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}\right). \end{aligned} \tag{2.15}$$

To conduct Bayesian inference on the panel data regression model in (2.13), we need to set up the posterior distribution. Note that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{Z}$ in (2.15) is $\mathcal{N}(\boldsymbol{Z}\boldsymbol{\delta}, \sigma_\epsilon^2 \boldsymbol{I})$. Thus the likelihood of the unknown parameters $(\boldsymbol{\delta}, \sigma_\epsilon)$ is

$$p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma_\epsilon) \propto (\sigma_\epsilon^2)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta})\right] \tag{2.16}$$

$$\propto (\sigma_\epsilon^2)^{-\frac{N}{2}} \exp\left[-\frac{\sum_{t=1}^{T} e_{it}^2}{2\sigma_\epsilon^2}\right], \quad e_{it} = y_{it} - \alpha_i - \boldsymbol{x}_{it}'\boldsymbol{\beta}. \tag{2.17}$$

We assume the following prior distributions for the parameters $(\boldsymbol{\delta}, \sigma_\epsilon)$.

$$\boldsymbol{\delta} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_\alpha \mathbf{1} \\ \boldsymbol{\mu}_\beta \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\alpha^2 \boldsymbol{I} & \\ & \boldsymbol{\Sigma}_\beta \end{bmatrix}, \tag{2.18}$$

$$\sigma_\epsilon \sim \mathcal{C}^+(0, s_\epsilon), \tag{2.19}$$

where $\mathcal{C}^+(0, \sigma_\epsilon)$ is the half-Cauchy distribution with the probability density:

$$p(\sigma_\epsilon|s_\epsilon) = \frac{2s_\epsilon}{\pi(\sigma_\epsilon^2 + s_\epsilon^2)}, \quad \sigma_\epsilon > 0, \quad s_\epsilon > 0.$$

Note that the prior distribution in (2.18) is equivalent to assuming

$$\begin{aligned} \alpha_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad i \in \{1, \ldots, N\}, \\ \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \end{aligned} \tag{2.20}$$

If we regard α_i as a random effect in (2.20), it can be interpreted as a latent variable specific to a particular individual that follows $\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$. If α_i is treated as a fixed effect, on the other hand, $\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ can be interpreted as the common prior distribution for all fixed effects that are unknown parameters. Either interpretation does not change the fact that α_i is an unobservable quantity and does not affect the resulting posterior distribution of the individual effect as a parameter.

In the prior distributions of the parameters $(\boldsymbol{\delta}, \sigma_\epsilon)$, $(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ in (2.18) and s_ϵ in (2.19) are fixed to specific values as hyperparameters. For $(\mu_\alpha, \sigma_\alpha)$ in (2.18), however, we set the following hierarchical prior distributions:

$$\mu_\alpha \sim \mathcal{N}(\varphi_\alpha, \tau_\alpha^2), \quad \sigma_\alpha \sim \mathcal{C}^+(0, s_\alpha), \tag{2.21}$$

and attempt to estimate them simultaneously with $(\boldsymbol{\delta}, \sigma_\epsilon)$ in the Bayesian approach. The hyperparameters $(\varphi_\alpha, \tau_\alpha^2, s_\alpha)$ in (2.21) are fixed to preset values. In summary, the parameters to be estimated in the hierarchical Bayesian analysis of the panel data regression model (2.13) are

$$\boldsymbol{\theta} = (\boldsymbol{\delta}, \mu_\alpha, \sigma_\alpha, \sigma_\epsilon) = (\alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_K, \mu_\alpha, \sigma_\alpha, \sigma_\epsilon).$$

By applying Bayes' theorem to the likelihood $p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma_\epsilon)$ in (2.16) and the prior distribution $p(\boldsymbol{\theta})$ in (2.18), (2.19) and (2.21), we obtain the posterior distribution of $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma_\epsilon) p(\boldsymbol{\theta}), \quad \mathcal{D} = (\boldsymbol{y}, \boldsymbol{Z}). \tag{2.22}$$

Because we cannot analytically evaluate either posterior distribution or posterior statistics (e.g., mean, median, variance, quantiles), we will employ the Markov chain Monte Carlo (MCMC) method. In this case, we can derive the conditional posterior distributions for

all parameters (their derivations are given in Appendix):

$$\boldsymbol{\delta}|\mathcal{D},\boldsymbol{\theta}_{-\boldsymbol{\delta}} \sim \mathcal{N}\left(\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right),\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\right), \tag{2.23}$$

$$\mu_\alpha|\mathcal{D},\boldsymbol{\theta}_{-\mu_\alpha} \sim \mathcal{N}\left(\frac{\sigma_\alpha^{-2}\sum_{i=1}^N \alpha_i+\tau_\alpha^{-2}\varphi_\alpha}{\sigma_\alpha^{-2}N+\tau_\alpha^{-2}},\frac{1}{\sigma_\alpha^{-2}N+\tau_\alpha^{-2}}\right), \tag{2.24}$$

$$\sigma_\alpha^2|\mathcal{D},\boldsymbol{\theta}_{-\sigma_\alpha},\xi_\alpha \sim \mathcal{IG}\left(\frac{N+1}{2},\frac{\sum_{i=1}^N(\alpha_i-\mu_\alpha)^2}{2}+\frac{1}{\xi_\alpha}\right),\ \xi_\alpha|\sigma_\alpha \sim \mathcal{IG}\left(1,\frac{1}{\sigma_\alpha^2}+\frac{1}{s_\alpha^2}\right), \tag{2.25}$$

$$\sigma_\epsilon^2|\mathcal{D},\boldsymbol{\theta}_{-\sigma_\epsilon},\xi_\epsilon \sim \mathcal{IG}\left(\frac{T+1}{2},\frac{\sum_{i=1}^N\sum_{t=1}^{T_i} e_{it}^2}{2}+\frac{1}{\xi_\epsilon}\right),\ \xi_\epsilon|\sigma_\epsilon \sim \mathcal{IG}\left(1,\frac{1}{\sigma_\epsilon^2}+\frac{1}{s_\epsilon^2}\right), \tag{2.26}$$

where $\boldsymbol{\theta}_{-a}$ indicates that the parameter a is excluded from $\boldsymbol{\theta}$, and $\mathcal{IG}(a,b)$ is the inverse gamma distribution with the probability density:

$$p(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}.$$

The conditional posterior distributions in (2.23)–(2.26) are standard ones and efficient random number generation algorithms are available for them. Therefore Gibbs sampling can be used to generate the parameters $\boldsymbol{\theta}$ from the posterior distribution (2.22).

If the individual effects $\boldsymbol{\alpha}$ are regarded as latent variables or missing values in the model, the Gibbs sampling scheme based on (2.23)–(2.26) is regarded as a SA scheme for μ_α. This is because the panel data regression model (2.13) does not depend on μ_α given $\boldsymbol{\alpha}$ while the hierarchical prior of $\boldsymbol{\alpha}$, $\mathcal{N}(\mu_\alpha,\sigma_\alpha^2)$ in (2.21), does depend on μ_α. In other words, what (2.1) and (2.2) are to the two-level normal hierarchical model, (2.13) and $\mathcal{N}(\mu_\alpha,\sigma_\alpha^2)$ are to the panel data regression model. Furthermore, since μ_α in $\mathcal{N}(\mu_\alpha,\sigma_\alpha^2)$ is not necessarily equal to zero, the panel data regression model in (2.13) is said to be in the non-centered parametrization (NCP) form. We use this SA-type Gibbs sampling scheme based on the NCP form as the benchmark and compare it with the ASIS scheme for the panel data regression model.

To derive the ASIS scheme for the panel data regression model, we introduce its centered parametrization (CP) form:

$$y_{it}=\mu_\alpha+\boldsymbol{x}_{it}'\boldsymbol{\beta}+\tilde{\alpha}_i+\epsilon_{it},\quad \tilde{\alpha}_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\alpha^2),\quad \epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2), \tag{2.27}$$

where $\tilde{\alpha}_i$ can be interpreted as a random effect for individual i in the sense of the traditional panel data regression analysis. Since the mean of the hierarchical prior $\mathcal{N}(0,\sigma_\alpha^2)$ is

equal to zero, the Gibbs sampling scheme based on the CP form (2.27) is regarded as an AA scheme for μ_α. Therefore, by interweaving it with the SA-type Gibbs sampling based on the NCP form (2.13), we can construct the ASIS scheme for the panel data regression model.

Note that $\tilde{\alpha}_i$ and α_i have the following relationship:

$$\tilde{\alpha}_i = \alpha_i - \mu_\alpha, \quad \alpha_i = \tilde{\alpha}_i + \mu_\alpha, \quad i \in \{1, \ldots, N\}. \tag{2.28}$$

With (2.28) and

$$\tilde{y}_{it} = y_{it} - \tilde{\alpha}_i, \quad i \in \{1, \ldots, N\}, \quad t \in \{1, \ldots, T\}, \tag{2.29}$$

(2.27) is rewritten as

$$\tilde{y}_{it} = \mu_\alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2). \tag{2.30}$$

Furthermore, by introducing the notations:

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{y}_{11} \\ \vdots \\ \tilde{y}_{NT} \end{bmatrix}, \quad \tilde{\boldsymbol{Z}} = \begin{bmatrix} 1 & \boldsymbol{x}_{11}' \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_{NT}' \end{bmatrix}, \quad \tilde{\boldsymbol{\delta}} = \begin{bmatrix} \mu_\alpha \\ \boldsymbol{\beta} \end{bmatrix},$$

(2.30) becomes

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\delta}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}). \tag{2.31}$$

It is straightforward to show that the prior for $\tilde{\delta}$ is

$$\tilde{\boldsymbol{\delta}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \quad \tilde{\boldsymbol{\mu}} = \begin{bmatrix} \varphi_\alpha \\ \boldsymbol{\mu}_\beta \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tau_\alpha^2 & \\ & \boldsymbol{\Sigma}_\beta \end{bmatrix}. \tag{2.32}$$

Then the conditional posterior distribution of $\tilde{\boldsymbol{\delta}}$ is obtained as

$$\begin{aligned} &\tilde{\boldsymbol{\delta}}|\mathcal{D}, \boldsymbol{\theta}_{-\boldsymbol{\delta}} \\ &\sim \mathcal{N}\left(\left(\sigma_\epsilon^{-2}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} + \tilde{\boldsymbol{\Sigma}}^{-1} \right)^{-1} \left(\sigma_\epsilon^{-2}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{y}} + \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} \right), \left(\sigma_\epsilon^{-2}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} + \tilde{\boldsymbol{\Sigma}}^{-1} \right)^{-1} \right), \end{aligned} \tag{2.33}$$

in exactly the same manner as (2.23).

Thus the Gibbs sampling scheme with ASIS can be formulated as follows.

ASIS Scheme for the Panel Data Regression Model

Step 1: Given the s-th generated $\boldsymbol{\theta}^{(s)}$, apply Gibbs sampling based on the conditional posterior distributions of (2.23)–(2.26) to generate $\boldsymbol{\theta}^{(s+0.5)}$ and compute

$$\tilde{\alpha}_i^{(s+0.5)} = \alpha_i^{(s+0.5)} - \mu_\alpha^{(s+0.5)}.$$

Step 2: Given $\boldsymbol{\theta}^{(s+0.5)}$, apply Gibbs sampling based on the conditional posterior distributions of (2.33) and (2.25)–(2.26) to generate $(\boldsymbol{\beta}^{(s)}, \mu_\alpha^{(s)}, \sigma_\alpha^{(s)}, \sigma_\epsilon^{(s)})$, and compute

$$\alpha_i^{(s)} = \tilde{\alpha}_i^{(s+0.5)} + \mu_\alpha^{(s)}.$$

In **Step 1**, we generate the individual effects $\boldsymbol{\alpha}$ along with the rest of the parameters in $\boldsymbol{\theta}$ from the posterior distribution in the NCP form. Since it is impractical to draw $\boldsymbol{\theta}$ from the posterior distribution at once, we apply CIS, i.e., a typical Gibbs sampling scheme to random number generation of $\boldsymbol{\theta}$. At the last stage of **Step 1**, we transform $\{\alpha_i\}_{i=1}^N$ into $\{\tilde{\alpha}_i\}_{i=1}^N$ with (2.28), which is corresponding to **Step 2-a** in GIS. In **Step 2**, we generate $(\boldsymbol{\beta}, \mu_\alpha, \sigma_\alpha, \sigma_\epsilon)$ from the posterior distribution in the CP form where we treat $\{\tilde{\alpha}_i\}_{i=1}^N$ as fixed values. At the last state of **Step 2**, we transform $\{\tilde{\alpha}_i\}_{i=1}^N$ back into $\{\alpha_i\}_{i=1}^N$. This is optional. If we update $\{\alpha_i\}_{i=1}^N$ at the very first stage of **Step 1**, the old $\{\alpha_i\}_{i=1}^N$ computed at the last stage of **Step 2** in the previous cycle of the ASIS scheme will be overwritten with the newly generated ones in **Step 1** before they are used in sampling of the other parameters.

2.4 Application to Wage Panel Data

In this section, we apply the ASIS scheme for the panel data regression model, which we derived in Section 2.3, to real-world data and demonstrate its efficacy and improvement in the convergence rate.

For this purpose, we employ a data set on individual wages and attributes used in the study by Vella and Verbeek (1998), which is publicly available in the data archive of *Journal of Applied Econometrics* (`http://qed.econ.queensu.ca/jae/`) , statistical packages such as R, and popular econometrics textbooks such as Wooldridge (2019).

Table 2.1: Variables in *Wagepan* data set

Variable	Description
lwage	log of wages
educ	years of education
exper	years of experience
expersq	years of experience squared
black	black dummy
hisp	Hispanic dummy
hours	hours worked per year
married	marriage dummy
occ1–occ9	occupation dummies
d81–d87	year dummies
union	labor union membership dummy

Since this date set is often called *Wagepan* (or *wagepan*) in those archives, we use this tag to refer to the data set in this section.

Vella and Verbeek (1998) estimated the premium of joining a labor union among young male workers in the face of declining rates of union membership. In their estimation, they obtained data on years of education, race, and whether or not the worker was married in order to control for effects other than union membership or nonunion membership. Specifically, the variables in Table 2.1 are included in *Wagepan*. These variables are cleaned up so that they have no missing values. As a result, we have a balanced panel data that includes complete records of $N = 545$ persons for $T = 8$ years. Thus the total number of records is 4360.

With *Wagepan*, we estimate a panel data regression model of the log wage ("lwage" in Table 2.1). Independent variables in the regression model are those in Table 2.1 except for the dependent variable "lwage" and a occupation dummy "occ1" to avoid multicollinearity. Thus the number of the independent variables is $K = 23$. The hyperparameters

$(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \varphi_\alpha, \tau_\alpha^2, s_\alpha, s_\epsilon)$ in the prior distribution are set as follows.

$$\boldsymbol{\mu}_\beta = \mathbf{0}_K, \quad \boldsymbol{\Sigma}_\beta = 100\boldsymbol{I}_K, \quad \varphi_\alpha = 0, \quad \tau_\alpha^2 = 100, \quad s_\alpha = s_\epsilon = 1. \tag{2.34}$$

For both Gibbs sampling scheme with and without ASIS, we initiate the sampling chain with 10,000 random number generation as burn-in until the random series of the parameters become stable. Then we iterate 10,000 random number generation to obtain the Monte Carlo sample $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^S (S = 10,000)$ for parameter estimation. The posterior statistics of the regression coefficient $\boldsymbol{\beta}$ and other parameters are summarized in Table 2.2. To visualize the convergence of generate random series, we draw time series plots of random series in Figure 2.1 and their sample autocorrelation plots in Figure 2.2. To save the space, Figure 2.2 shows the plots only for selected parameters.

The p-value of Geweke (1992)'s convergence diagnostic in Table 2.2 is at least 0.07 or higher for 571 random series of the parameters (the number of the parameters is $N + k + 3 = 545 + 23 + 3 = 571$). Thus the convergence diagnostic fails to reject the null hypothesis of the constant mean at 5% significance level for all parameters in the panel data regression model. This result indicates that the generated random series in the ASIS scheme are well stabilized after 10,000 burn-in. We get the same impression from the time series plots in Figure 2.1. All random series from the ASIS scheme seem stable and well mixed around the constant level which is corresponding to the posterior mean. This finding is also consistent with the sample autocorrelation plots in Figure 2.2. All plots for the ASIS scheme go down to zero fairly quickly.

For the non-ASIS scheme, one the other hand, the convergence diagnostic in Table 2.2 rejects the null hypothesis for many variables including all year dummies. Time series plots in Figure 2.1 also show wild and often persistent swings in trajectories of generated random series. These results suggest that the non-ASIS scheme fails to generate stable random series for some parameters.

Let us pick up a few parameters and examine them in detail. First we examine two regression coefficients: β_{educ} for "education" and β_{union} for "union" as representative cases. When we compare the posterior mean and the 95% credible interval in Table 2.2, we do not see any noteworthy difference between the ASIS scheme and the non-ASIS scheme for both β_{educ} and β_{union}, though the posterior mean of β_{educ} may be a little

underestimated without ASIS. As for Geweke (1992)'s convergence diagnostic, β_{union} passes the convergence diagnostic test with or without ASIS since the p-value is above the conventional level of significance for both cases. For β_{educ}, on the other hand, we see a stark difference between the ASIS scheme and the non-ASIS scheme. The random series with ASIS seems sufficiently convergent because the p-value exceeds the conventional level of significance. For the random series without ASIS, however, the p-value is less than 0.0001, which implies that the null hypothesis of the constant mean is rejected. This finding is consistent with the sample autocorrelation plots in Figure 2.2. The sample autocorrelation plot of β_{union} is almost zero for both cases. For β_{educ}, however, the sample autocorrelation plot for the non-ASIS scheme exhibits strongly positive and persistent autocorrelation while the plot for the ASIS scheme goes down to zero quickly. These results indicate that the sampling efficiency of the non-ASIS scheme may heavily rely on the nature of the data we use and it may be a matter of luck for researchers whether the non-ASIS scheme works or not. Given its consistent performance throughout the parameters, the ASIS scheme seems more trustworthy than the non-ASIS scheme in the sense that the former is more likely to produce stable and well mixed random series from the posterior distribution than the latter.

Next, we examine parameters related to the individual effects: α_i ($i \in \{1, \ldots, 545\}$), μ_α and σ_α. Since the panel data regression model has 545 individual effects, we focus only on α_1. In Figure 2.1, we find a noticeable difference in stability for α_i and μ_α. Random series of α_i and μ_α generated with the non-ASIS scheme exhibit persistent up-and-down movements and, as a result, they does not seem stationary at all. Furthermore their movements are more or less synchronized with each other. This pattern of fluctuation we observe in Figure 2.1 virtually disappears when we apply ASIS. This result strongly suggests that ASIS improves the efficiency of random number generation of these parameters. The sample autocorrelation plots in Figure 2.2 also demonstrate that ASIS can dramatically reduce autocorrelation and improve mixing of the generated random series. For σ_α, however, there is no remarkable difference between the ASIS scheme and the non-ASIS scheme in either Figure 2.1 or Figure 2.2.

These findings have important implications on how ASIS should be designed. Recall that the distinction between the NCP form (2.13) and the CP form (2.27) is whether

μ_α appear in the regression equation or not. If we try a sampling scheme based on the NCP form and encounter poor mixing as shown in Figure 2.1 and Figure 2.2, we must suspect that the latent variables $\boldsymbol{\alpha}$ have little information about μ_α. Since the sampling scheme based on the NCP form is a SA, non-informative latent variables lead to a poor performance in sampling μ_α (see the discussion about distinction between SA and AA in Section 2.2). If this is the case, it may be worth trying ASIS on a problematic parameter (μ_α in this example) to boost the mixing. Since μ_α is directly linked to observed dependent variables y_{it} in the CP form (2.27), we can utilize more information about μ_α in sampling with ASIS.

Table 2.2: Posterior statistics

$\boldsymbol{\beta}$	ASIS not applied	ASIS applied
	$1.0903 * 10^{-1}$	$1.1210 * 10^{-1}$
exper	$[\ 7.8653 * 10^{-2},\ 1.3927 * 10^{-1}]$	$[\ 8.1297 * 10^{-2},\ 1.4296 * 10^{-1}]$
	0.0004	0.7896
	$-5.0933 * 10^{-3}$	$-5.0847 * 10^{-3}$
expersq	$[\ -6.4737 * 10^{-3},\ -3.7491 * 10^{-3}]$	$[\ -6.4634 * 10^{-3},\ -3.7328 * 10^{-3}]$
	0.3643	0.6115
	$-1.5128 * 10^{-1}$	$-1.4687 * 10^{-1}$
hours∗	$[\ -1.9534 * 10^{-1},\ -1.0796 * 10^{-1}]$	$[\ -1.9139 * 10^{-1},\ -1.0208 * 10^{-1}]$
	0.0005	0.4967
	$7.9404 * 10^{-1}$	$8.4341 * 10^{-1}$
educ∗	$[\ 5.9500 * 10^{-1},\ 9.8375 * 10^{-1}]$	$[\ 6.2176 * 10^{-1},\ 1.0754]$
	0.0000	0.2806
	$1.0546 * 10^{-1}$	$1.0507 * 10^{-1}$
union	$[\ 7.0247 * 10^{-2},\ 1.4015 * 10^{-1}]$	$[\ 6.9144 * 10^{-2},\ 1.4075 * 10^{-1}]$
	0.4807	0.4985
	$6.5194 * 10^{-2}$	$6.4275 * 10^{-2}$
married	$[\ 3.2305 * 10^{-2},\ 9.7504 * 10^{-2}]$	$[\ 3.1095 * 10^{-2},\ 9.8140 * 10^{-2}]$
	0.6280	0.7096
race dummies		
	$-1.3882 * 10^{-1}$	$-1.3856 * 10^{-1}$
black	$[\ -5.0847 * 10^{-3},\ 0.0000]$	$[\ -2.3338 * 10^{-1},\ -4.3749 * 10^{-2}]$
	0.0000	0.7326
	$2.0601 * 10^{-2}$	$2.3650 * 10^{-2}$
hisp	$[\ -6.5229 * 10^{-2},\ 1.0764 * 10^{-1}]$	$[\ -6.1646 * 10^{-2},\ 1.1086 * 10^{-2}]$
	0.0825	0.3289
occupation dummies		
	$-2.3871 * 10^{-2}$	$-2.3619 * 10^{-2}$
occ2	$[\ -8.4192 * 10^{-2},\ 3.7516 * 10^{-2}]$	$[\ -8.3586 * 10^{-2},\ 3.6566 * 10^{-2}]$
	0.0937	0.6773
	$-7.3565 * 10^{-2}$	$-7.2782 * 10^{-2}$
occ3	$[\ -1.4389 * 10^{-1},\ -4.7938 * 10^{-3}]$	$[\ -1.4305 * 10^{-1},\ -2.9789 * 10^{-3}]$
	0.7707	0.9860
	$-1.0808 * 10^{-1}$	$-1.0672 * 10^{-1}$
occ4	$[\ -1.6519 * 10^{-1},\ -5.0719 * 10^{-2}]$	$[\ -1.6396 * 10^{-1},\ -4.9217 * 10^{-2}]$
	0.4829	0.8015
	$-4.9790 * 10^{-2}$	$-4.8114 * 10^{-2}$
occ5	$[\ -1.0478 * 10^{-1},\ 4.7431 * 10^{-3}]$	$[\ -1.0335 * 10^{-1},\ 6.9549 * 10^{-3}]$
	0.4785	0.4729
	$-5.9567 * 10^{-2}$	$-5.8000 * 10^{-2}$
occ6	$[\ -1.1581 * 10^{-1},\ -3.5336 * 10^{-3}]$	$[\ -1.1499 * 10^{-1},\ -3.6984 * 10^{-4}]$
	0.5421	0.9445
	$-8.7038 * 10^{-2}$	$-8.5258 * 10^{-2}$
occ7	$[\ -1.4979 * 10^{-1},\ -2.3085 * 10^{-2}]$	$[\ -1.4812 * 10^{-1},\ -2.2730 * 10^{-2}]$
	0.4890	0.3196
	$-1.4486 * 10^{-1}$	$-1.4413 * 10^{-1}$
occ8	$[\ -2.6774 * 10^{-1},\ 2.1343 * 10^{-2}]$	$[\ -2.6536 * 10^{-1},\ -2.0586 * 10^{-2}]$
	0.8954	0.7886
	$-1.2512 * 10^{-1}$	$-1.2356 * 10^{-1}$
occ9	$[\ -1.8797 * 10^{-1},\ -6.1441 * 10^{-2}]$	$[\ -1.8602 * 10^{-1},\ -6.0517 * 10^{-2}]$
	0.3724	0.8851

Note: Data marked with ∗ are logarithmic.

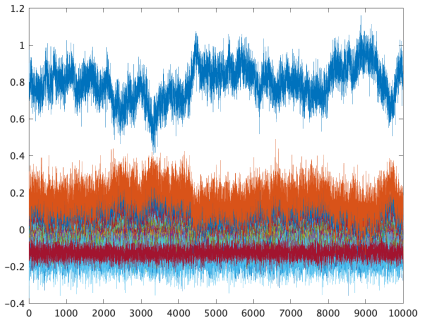
In each cell, the upper number is the posterior mean, the middle [·,·] is the 95% credible interval and the lower number is Geweke's Diagnostics.

Table 2.2: Posterior statistics (continued)

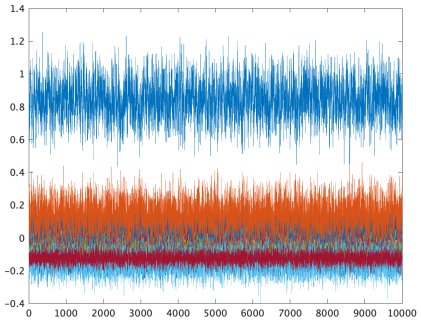
β	ASIS not applied	ASIS applied
year dummies		
	$4.9033 * 10^{-2}$	$4.5404 * 10^{-2}$
d81	$[\ 1.0474 * 10^{-3},\ 9.6422 * 10^{-2}]$	$[\ -3.1301 * 10^{-3},\ 9.4534 * 10^{-2}]$
	0.0007	0.3584
	$4.1197 * 10^{-2}$	$3.4497 * 10^{-2}$
d82	$[\ -2.0760 * 10^{-2},\ 1.0412 * 10^{-1}]$	$[\ -3.0911 * 10^{-2},\ 9.8730 * 10^{-2}]$
	0.0013	0.9079
	$3.9470 * 10^{-2}$	$2.9798 * 10^{-2}$
d83	$[\ -4.1054 * 10^{-2},\ 1.2311 * 10^{-1}]$	$[\ -5.6492 * 10^{-2},\ 1.1424 * 10^{-1}]$
	0.0000	0.9140
	$6.5380 * 10^{-2}$	$5.2504 * 10^{-2}$
d84	$[\ -3.1462 * 10^{-2},\ 1.6717 * 10^{-1}]$	$[\ -5.2801 * 10^{-2},\ 1.5560 * 10^{-1}]$
	0.0001	0.7074
	$8.2989 * 10^{-2}$	$6.6732 * 10^{-2}$
d85	$[\ -3.5523 * 10^{-2},\ 2.0430 * 10^{-1}]$	$[\ -5.8809 * 10^{-2},\ 1.9011 * 10^{-1}]$
	0.0000	0.7244
	$1.1823 * 10^{-1}$	$9.8518 * 10^{-2}$
d86	$[\ -1.8787 * 10^{-2},\ 2.5852 * 10^{-1}]$	$[\ -4.8764 * 10^{-2},\ 2.4322 * 10^{-1}]$
	$0.0003 * 10^{-2}$	0.6906
	$1.6806 * 10^{-1}$	$1.4541 * 10^{-1}$
d87	$[\ 1.1368 * 10^{-2},\ 3.2963 * 10^{-1}]$	$[\ -0.6012 * 10^{-2},\ 3.1164 * 10^{-1}]$
	0.0002 & 0.6128	
	$3.5965 * 10^{-1}$	$1.9439 * 10^{-1}$
μ_α	$[\ -1.8911 * 10^{-1},\ 9.3572 * 10^{-1}]$	$[\ -5.0302 * 10^{-1},\ 8.7775 * 10^{-1}]$
	0.0000	0.2866
	$3.3442 * 10^{-1}$	$3.3458 * 10^{-1}$
σ_α	$[\ 3.1244 * 10^{-1},\ 3.5859 * 10^{-1}]$	$[3.1213 * 10^{-1},\ 3.5871 * 10^{-1}]$
	0.2984	0.3137
	$3.4922 * 10^{-1}$	$3.4917 * 10^{-1}$
σ_ϵ	$[\ 3.4165 * 10^{-1},\ 3.5725 * 10^{-1}]$	$[\ 3.4135 * 10^{-1},\ 3.5700 * 10^{-1}]$
	0.0376 & 0.0861	

Note: Data marked with $*$ are logarithmic.

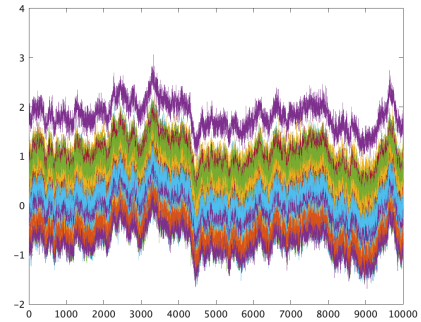
In each cell, the upper number is the posterior mean, the middle $[\cdot,\cdot]$ is the 95% credible interval and the lower number is Geweke's Diagnostics.



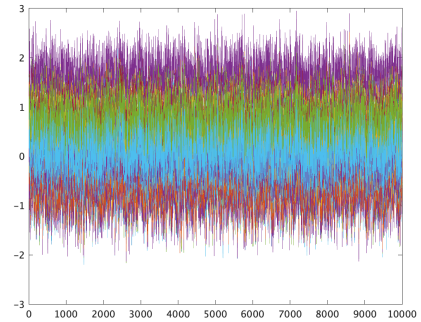
(a) $\boldsymbol{\beta}$ (ASIS not applied)



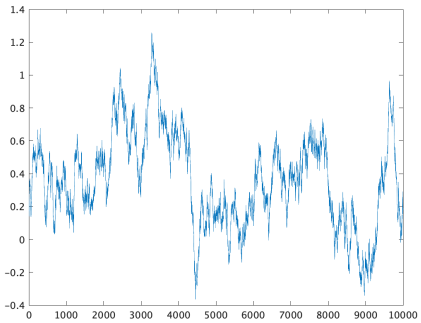
(b) $\boldsymbol{\beta}$ (ASIS applied)



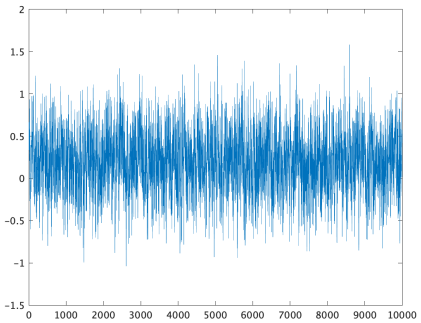
(c) $\alpha_1 - \alpha_{545}$ (ASIS not applied)



(d) $\alpha_1 - \alpha_{545}$ (ASIS applied)

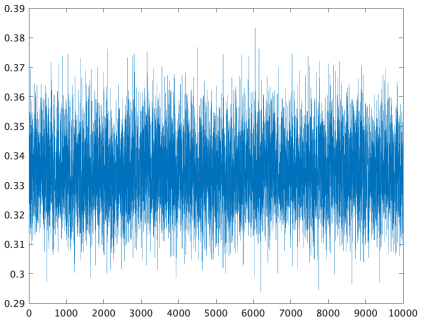


(e) μ_α (ASIS not applied)

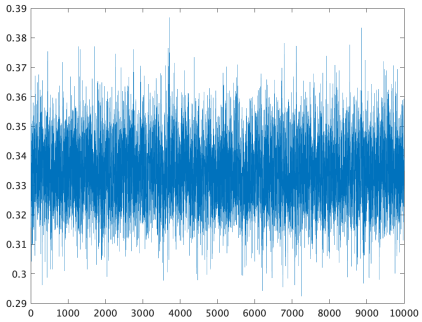


(f) μ_α (ASIS applied)

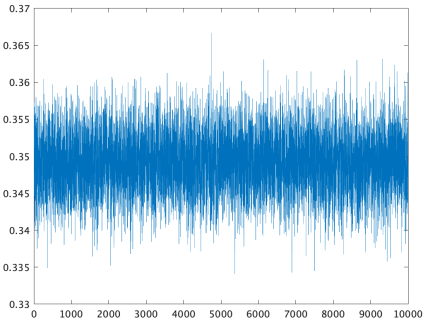
Figure 2.1: 10,000 sampling after 10,000 burn-ins



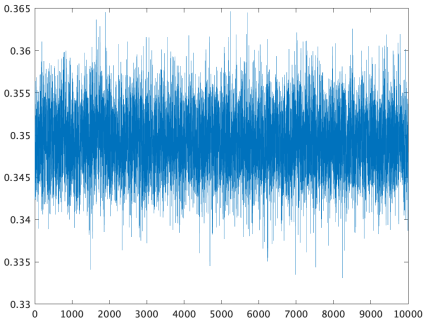
(g) σ_α (ASIS not applied)



(h) σ_α (ASIS applied)

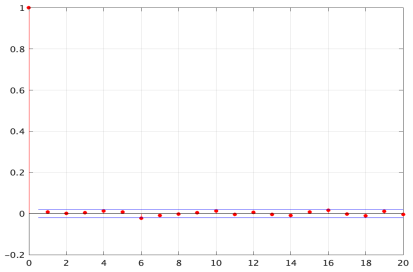


(i) σ_ϵ (ASIS not applied)

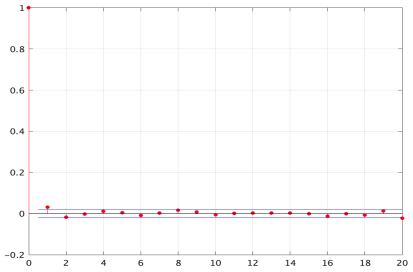


(j) σ_ϵ (ASIS applied)

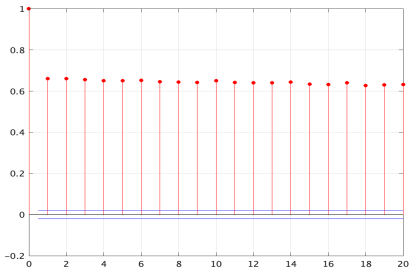
Figure 2.1: 10,000 sampling after 10,000 burn-ins (continued)



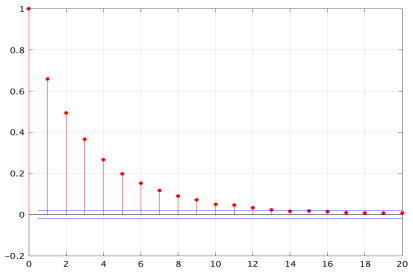
(a) β_{union} (ASIS not applied)



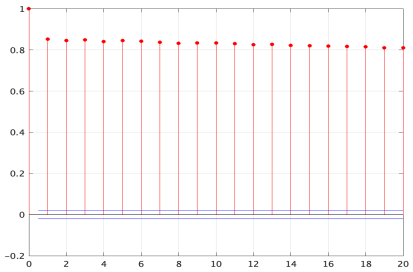
(b) β_{union} (ASIS applied)



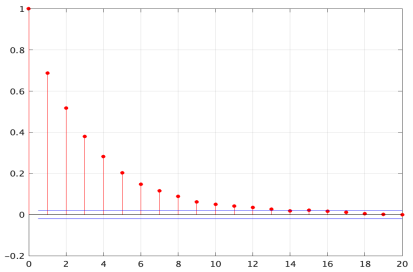
(c) β_{educ} (ASIS not applied)



(d) β_{educ} (ASIS applied)

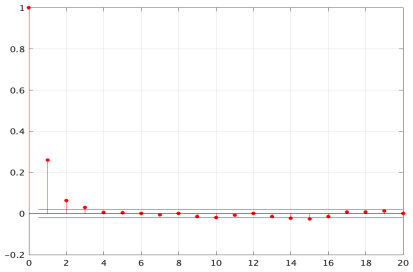


(e) α_1 (ASIS not applied)

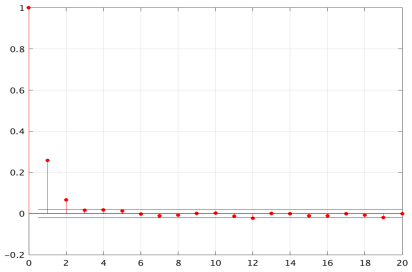


(f) α_1 (ASIS applied)

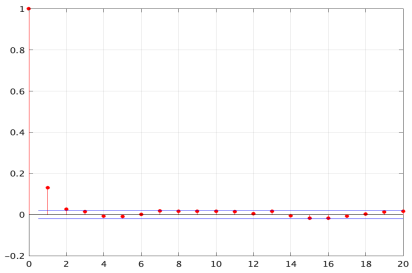
Figure 2.2: Autocorrelation of 10,000 sampling of hierarchical parameters



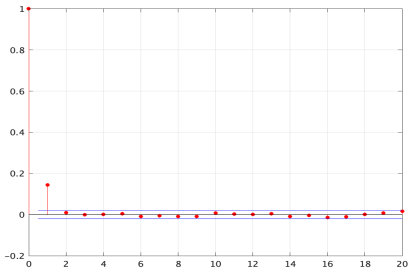
(g) σ_α (ASIS not applied)



(h) σ_α (ASIS applied)



(i) σ_ϵ (ASIS not applied)



(j) σ_ϵ (ASIS applied)

Figure 2.2: Autocorrelation of 10,000 sampling of hierarchical parameters (continued)

2.5 Conclusion

In this chapter, we provided the definition and principle of ASIS and examined its efficacy in hierarchical Bayesian modeling of panel data. For this purpose, we derived a new ASIS scheme for the panel data regression model with individual effects and applied it to wage panel data used in the study by Vella and Verbeek (1998).

In the empirical study, we showed that ASIS could improve the efficiency of MCMC sampling in terms of stability and mixing of generated random series. Without ASIS, due to strongly positive and persistent autocorrelation in random series, many parameters in the panel data regression model failed to pass Geweke (1992)'s convergence diagnostic test even after the burn-in of 10,000 runs. With ASIS, on the other hand, generated random series are far more stable and their sample autocorrelation tends to decline much faster. Moreover, although the model consists of near six hundred parameters, all of them successfully passed the convergence diagnostic test.

In overall assessment, the ASIS scheme archived efficient sampling for all parameters while the performance of the non-ASIS scheme was a mixed bag to say the least. Although our study is limited in terms of the data set we used and the model specification we examined, the superb performance of ASIS is rather noteworthy, especially given the fact that the number of parameters (571) is about 13% of the number of observations (4360).

The flexibility is another advantage of ASIS. As we will explore in Chapter 3, the panel data regression model may have two-way effects (e.g., individual effect and time effects) or more. It is straightforward to extend the ASIS for such models. See Chapter 3 for more details. Moreover, as Gelfand *et al.* (1995) proposed, we may extend the ASIS scheme to mixed effects models in which individual-specific regression coefficients, not necessarily the constant term, follow a common prior distribution. This extension is also straightforward.

In addition to the panel data regression model, we may apply ASIS in time series analysis. For example, Kastner and Frühwirth-Schnatter (2014) proposed an ASIS scheme to facilitate efficient sampling of the latent log volatility in a stochastic volatility (SV) model. In Chapter 4, we will develop a new efficient ASIS-based sampling algorithm for the SV model with heavy-tailed and possibly skewed error and leverage effect. Since this type of SV model is a nonlinear non-Gaussian state space model, it is hard to generate stable and well mixed random series of parameters and latent variables from the posterior distribution. Our new algorithm can handle complexity and scalability of such a complex model in an efficient manner. See Chapter 4 for more details.

Chapter 3

Application to Unbalanced Panel Data: Horse Race Results in Japan

3.1 Introduction

Recently, there has been a worldwide movement to measure the performance of athletes and improve the training methods by analyzing data in sports. Similar movements are being promoted by the public and private sectors in Japan. For example, the Japan Sports Agency is promoting the formation of the Sports Open Innovation Platform (SOIP), a testing ground for the growth and industrialization of sports, and supports the use of data for the development of sports businesses. Additionally, organizations, such as the Japan Sports Analysts Association and Japan Statistical Society, hold regular competitions for students to analyze data on various sports such as baseball, soccer, and rugby, with the aim of promoting academic research on sports statistics and increasing its recognition in society.

Horse racing is not exempted in the trend to utilize these data. Horse racing in Japan is attracting a great deal of attention, with ticket sales for races held by the Japan Racing Association (JRA) exceeding 2.7 trillion yen in 2018[1]. Horse racing's popularity has led to the appearance of websites that provide horse racing data and software that analyzes the data available on these websites to help users purchase the most appropriate horse racing tickets. DWANGO Co., Ltd holds a horse racing algorithm competition called

[1] `http://company.jra.jp/0000/gaiyo/g_22/g_22_01.pdf`.

"Dennou Sho (Spring)" since 2017, and JRA-VAN, a data provision service operated by JRA, has been provided to participants of this competition, which has begun to encourage quantitative analysis of horse racing using data.

However, due to the characteristics of horse racing as public gambling, it is undeniable that data analysis of horse racing tends to be biased toward the viewpoint of "which horse to bet on to make a profit." Therefore, to the best of our knowledge, research on horse racing conducted from the viewpoint of sports statistics in Japan is lacking, and only a few related studies have been conducted in other countries as well. Therefore, this study presents a quantitative analysis of the running ability of racehorses in horse racing in Japan with the aim of pioneering the study of sports statistics in horse racing in Japan.

To measure a racehorse's running ability, in addition to race time as a measure of leg speed, various quantitative indicators such as lifetime or annual winning percentage and prize money can be considered. However, since the winning percentage and prize money are interpreted as a cumulative index of the relative time difference from other racehorses in each race, they are indices that depend on the opponents in the race in which the horse runs. Furthermore, since the prize money of a race depends on the rating of the racehorses that are eligible to participate (the higher the rating of a race, such as G1, the higher the prize money tends to be), the prize money earned also depends on the race type the horse has entered. Contrarily, race times can be regarded as a numerical value focusing only on individual racehorses, and thus are considered to be appropriate as a quantitative indicator of a racehorse's running ability in a race. Therefore, in this study, "speed" (in meters per second), calculated by dividing the racehorse's time in each race by the distance of the race, is used as a quantitative index of the racehorse's running ability, and a regression model with this as the explained variable is used to examine the determinants of the racehorse's running ability.

The factors affecting the speed of a racehorse as a quantitative index can be broadly divided into the following three categories: racehorse's ability; race environment; and jockey's ability. Regarding a racehorse's ability, which can be further divided into the horse's observable (i.e., horse, such as the horse's weight, age, and sex) and unobservable factors (horse's own natural constitution and temperament, and its physical condition before the race.) Contrarily, the race environment includes the type of track surface and

the condition of the track surface on the day of the race. The observable variables in the first and second groups can be used as explanatory variables in a regression model with the horse's speed as the explained variable. These data are detailed in Section 3.2.

However, unobservable factors that determine a racehorse performance cannot be included in the regression model as explanatory variables. Therefore, these factors are added to the regression model as racehorse-specific parameters (hereafter referred to as racehorse individual effects). This racehorse individual effect corresponds to a random or fixed effect that appears in the panel data analysis. In a textbook panel data analysis, it is customary to assume a balanced panel where the length of the sample period (number of races in the case of horse racing) for each individual is uniform. In reality, however, some horses continue to run in many races, whereas others retire after only a few races. Therefore, the number of races run during the sample period varies greatly from horse to horse; thus, the run data of racehorses result in an unbalanced panel. Of course, it would be possible to construct a model that explicitly incorporates the mechanism of missing data due to cessation of running[2], but this would require estimating the racehorse survival time model as well, making the model considerably more complex. Combined with the limitations of the available data, this approach would be impractical. Therefore, in this study, we assume that the "cessation of running" associated with the racehorse retirement occurs randomly and independently, and we aim to avoid problems associated with panel data incompleteness by applying the method of hierarchical Bayesian analysis to the estimation of individual effects of racehorses.

There is no disagreement that the third factor, the jockey's ability, is extremely important in determining the winner of a horse race. Horse racing is a competition in which jockeys compete for position on a racehorse. In other words, horse racing is a sport similar to motor sports, except that the jockeys ride horses, which are living animals. The jockey controls the racehorse to run at optimal pace, taking into account the horse's temperament, horse's physical condition on the day of the race, and condition of the course where the race is being held. This is the jockey's role during a race; it is the reason why horse racing is not just a simple game between horses, but a sophisticated sport in which

[2]The "cessation of running" here does not mean a temporary cessation of running, but rather a complete cessation of the racehorse's activity.

humans engage in intense games of strategy and skill while controlling a racehorse as a vehicle.

However, unlike motor sports such as four-wheeled vehicle, two-wheeled vehicle, boat, and auto racing that are also public gambling, racehorse "vehicles" are not man-made objects but living organisms, and there is a large variation in running performance among individuals[3]. Moreover, racehorses that are currently considered competitive tend to be assigned to jockeys with excellent past records and who are considered to be excellent jockeys. Therefore, if we use winning percentage and prize money as a measure of a jockey's ability, it becomes difficult to distinguish whether a high winning percentage and high prize money are achieved due to the individual jockey's superior ability, or whether the jockey is simply aided by the ability of the racehorse he/she rides. Therefore, this study aimed to measure the jockey's ability to control a racehorse by estimating the jockey's contribution to the speed recorded by the ridden racehorse in the race. Specifically, an unobserved parameter unique to the jockey (hereafter referred to as the jockey's individual effect) is added as the jockey's contribution to the racehorse's speed in the incomplete panel data regression model with the racehorse's speed as the explained variable, and this individual effect is estimated simultaneously with the racehorse's individual effect using a hierarchical Bayesian analysis. The jockey's individual effect is estimated simultaneously with the racehorse's individual effect using a hierarchical Bayesian analysis. In other words, the jockey's individual effect in this study is interpreted as the increase or decrease in the speed of a racehorse caused by a particular jockey's ride provided that all other factors, including racehorse ability and track condition, remain the same (ceteris paribus)[4]. However, since the individual effects of racehorses are already included in the regression model, including all jockeys' individual effects in the same model would lead to complete multicollinearity, making the model estimation impossible. To avoid this problem, in this study, the jockey's individual effect of Yutaka Take[5] is omitted from the

[3]In motorsports, it is customary to establish detailed rules and restrictions to equalize the performance of vehicles among riders and teams as much as possible.

[4]In this study, for the sake of simplicity, the interaction between the jockey's individual effects and racehorse's individual effects (i.e., the compatibility between the jockey and racehorse) is ignored.

[5]Yutaka Take is, at the time of writing this study, the record holder for the most wins and most rides in JRA's history, and is onamong the best jockeys who can bring out the best in racehorses in the world

model. Then, the jockey's individual effect estimated based on Yutaka Take is interpreted as the difference in ability of the jockeys in question compared to Yutaka Take.

In summary, the model underlying the empirical analysis in this study is an unbalanced panel data regression model with the racehorse speed in each race as the explained variable and the following items as the determinants of the racehorse speed: racehorse's individual effect; jockey's individual effect (relative ability difference between the jockey and Yutaka Take); observed racehorse characteristics; and observed environmental characteristics of the race. The estimation of this regression model applies the hierarchical Bayesian analysis method. As explained in Section 3, the model itself is standard; thus it is relatively easy to generate unknown parameters such as all individual effects and coefficients of explanatory variables from the posterior distribution by means of Gibbs sampling. Although the model itself is standard, the large number of racehorses and jockeys in the analysis makes stable estimation difficult with conventional methods. Therefore, this study incorporates a sampling technique called Ancillarity-Sufficiency Interweaving Strategy (ASIS) to improve sampling efficiency, which enables efficient estimation despite the fact that the data itself is difficult to estimate using conventional methods. Since a large number of racehorse and race environmental characteristics are candidates for explanatory variables to be included in the regression model, the model selection criterion Widely Applicable Information Criterion (WAIC) proposed by Watanabe (2010) was used to select a combination of these variables.

Silverman (2012) is an example of a regression model that uses hierarchical Bayesian analysis to explain racehorse speed. The fundamental difference between Silverman (2012) and this study, with the exception of the data used in this study, is that it focuses its empirical analysis on measuring racehorses and jockeys the performances, whereas Silverman (2012) focuses on measuring the performance of race winners. Silverman (2012) claims that the predicted success rate of the first-place finisher in the estimated model is 21.63%, which is better than the predicted success rate when the horses are chosen at random (7.14%–12.5%). This result is superior to that of random selection (7.14%–12.5%). Outside of Japan, more research is being conducted on race winners rather than racehorse speed. For example, Sung and Johnson (2007) use conditional logit models

of horse racing in Japan.

to analyze race winners in British horse racing. Meanwhile, Edelman (2007), Lessmann et al. (2007), Lessman et al. (2009), Chung et al. (2017), and others have attempted to predict racehorse winners using support vector machines (SVM). Benter (2008) proposed using conditional logistic regression models for horse racing analysis, and Silverman and Suchard (2013) developed their model by including parameters inspired by the Cox proportional hazards model. In addition, Lessmann et al. (2010) used a random forest algorithm to predict the winner of a race. However, to the best of our knowledge, no academic research has been conducted using scientific statistical methods on data from horse racing in Japan. Additionally, only Silverman (2012) had employed the hierarchical Bayesian analysis to examine the determinants of racehorse's speed, and ours is the first study to take into account jockey's individual effect. Thus, the contribution of this study is not insignificant. Furthermore, it should be emphasized that the data used in this study were obtained by the authors by scraping from "netkeiba.com," a website that provides horse racing information. Considering that similar data are not available in a form that can be used for statistical analysis, the findings of the empirical study in this study are valuable.

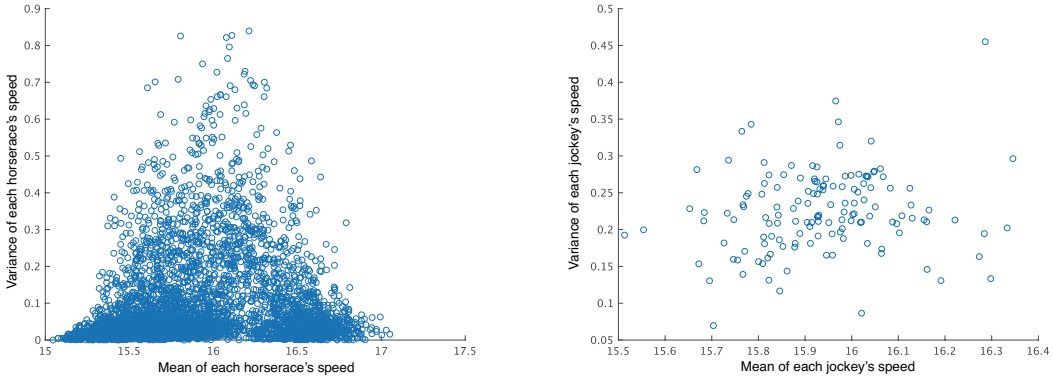
Finally, this chapter is organized as follows. Section 3.2 describes the horse racing data used in this chapter and presents descriptive statistics. Section 3.3 describes the panel data regression model used to explain racehorse speed and the hierarchical Bayesian analysis method used to estimate it. Section 3.4 presents the detailed results of the analysis, followed by a summary and discussion of the findings from the empirical analysis in Section 3.5.

3.2 Data

First, this section describes the details of the horse race data used in this study. We analyzed 1800-m races held by JRA between 2016 and 2018, excluding steeplechase races. The 1800-m racefs were chosen because they had the largest number of race records among the races held by JRA during the same period. These records were obtained by scraping from netkeiba.com (`https://db.netkeiba.com/`), a horse racing information site operated by Net Dreamers, Inc. Before screening, the total number of entries was

32,667, with 11,198 racehorses and 224 jockeys. However, if racehorses and jockeys who ran extremely infrequently, or if race records that seemed to have had some kind of trouble (e.g., extremely long running times) were included, they would become outliers, and it would be difficult to conduct an accurate analysis. Therefore, we screened the data by removing records that fulfilled at least one of the following three conditions: Condition 1, records of entries for jockeys who have ridden <10 times in 1800-m races; Condition 2, records of racehorses that has raced <3 times in 1800-m races; and Condition 3, records of runs by horses whose speed is in the bottom 5% of all runs. As a result, the number of racehorses to be eligible for analysis was 4,063, excluding 7,135 racehorses that met conditions 1–3 above from 11,198. Similarly, the remaining 140 jockeys were selected for analysis, excluding 83 of the 224 jockeys who met conditions 1–3. Finally, 22,183 race records were used for analysis. In the screening process, the number of rides by jockeys was selected before the number of rides by racehorses, resulting in 10 jockeys having fewer than 10 rides in the data selected.

For a bird's-eye view of the data, Figure 3.3 shows scatter plots of the mean and variance of the speeds of 4,063 racehorses and 141 jockeys. Figure 3.3(a) shows the racehorse speed, which can be roughly categorized into two based on the average of 16.2. The left cluster is for racehorses that mainly run on dirt courses, whereas the right cluster is for racehorses that mainly run on turf courses. The racehorses with the largest variance are those with an average of 16.2, but these racehorses run on both turf and dirt courses, and we can assume that the variance is greatly influenced by the course and the condition of the racehorses and jockeys. Next, we examine the relationship between the mean and variance of speed in terms of the jockeys. Figure 3.3(b) shows a scatter plot of the mean and speed of 141 jockeys, and it can be seen that none of them have an extreme small variance compared to the racehorses. The scatterplot of the jockeys shows the left–right divergence observed in the scatterplot of the racehorses, indicating that the jockeys ride on both turf and dirt courses equally.



(a) Mean and variance of speed of 4,063 racehorses

(b) Mean and variance of speed of 141 jockeys

Figure 3.1: Mean and variance of racehorse and jockey speed

The explained variable in the regression model, the racehorse speed, is the speed per second (the distance of the race (1800 m) divided by the time (seconds)). Since it is inappropriate for speed to be non-negative, the speed as the explained variable is log-transformed to impose a non-negative constraint when estimating, and the regression coefficient of each explanatory variable is interpreted as the ratio of its effect on speed. As explanatory variables for the regression model, we use the following features from the netkeiba.com database that are expected to affect racehorse speed in the race. Table 3.1 summarizes the descriptive statistics (mean, standard deviation, median, maximum, and minimum) for each feature.

1. Features of a racehorse

   - Horse weight (kg)
   - Gain or loss in horse weight (kg)
   - Horse age (years)
   - Horse age squared
   - Mare dummy (mare = 1)
   - Gelding dummy (castrated stallion = 1)
   - Trainer's evaluation (4 levels: A, B, C, D)

2. Environmental features of the run

   - weight to carry (weight to be borne by jockey, saddle, etc., in kg)
   - outer track dummy (run on outer track = 1)
   - turf course dummy (run on turf course = 1)
   - Racecourse dummies (10 locations: Tokyo, Chukyo, Nakayama, Kyoto, Hakodate, Kokura, Niigata, Sapporo, Fukushima, and Hanshin)
   - track index[6]
   - race index (average time excluding racehorse i record in race it)

In preparation for a hierarchical Bayesian analysis, let us consider the relationship between racehorse speed and each of the explanatory variables as a preliminary analysis. For this purpose, we created a correspondence table (Table 3.2) between the frequency distribution of each explanatory variable and the average racehorse speed within the class. Figure 3.2 shows the average speed for each quintile of the quantitative explanatory variables. Table 3.2 shows the average speed within each category for the qualitative explanatory variables. As shown in Table 3.2, we can guess in advance what effect the characteristics related to the racehorse and race can have on racehorse speed as an explanatory variable in the regression model.

Table 3.1: Descriptive statistics for dependent and explanatory variables

	Mean	Standard Deviation	Median	Max	Min
speed (m/s)	15.9634	0.4913	15.8730	17.2249	15.0250
horse weight	475.6023	28.9928	476	598	378
gain or loss in horse weight	0.3999	6.1309	0	40	−38
Horse age	3.6850	1.2565	3	10	2
Horse age squared	15.1577	11.0182	9	100	4
Weight to carry	54.9787	1.5463	55	60	49
Track index	−1.8314	10.7431	−3	71	−30
Race index	15.9210	0.5000	15.8195	17.1227	14.6559

[6]Track index is a numerical measure of the track condition. The hard-to-run condition is indicated by a positive value and the easy-to-run condition is indicated by a negative value.

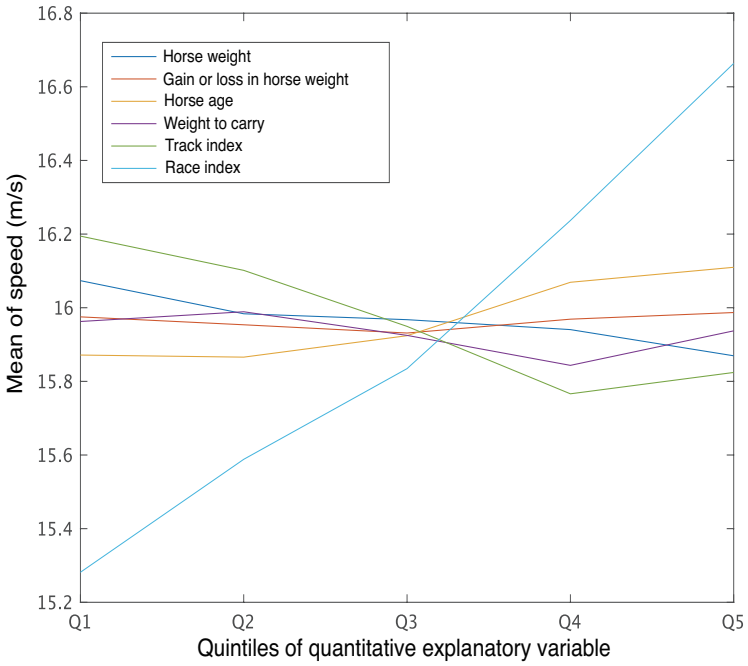


Figure 3.2: Frequency distribution of explanatory variables and within-class means of racehorse speed (quantitative explanatory variables)

Table 3.2: Frequency distribution of explanatory variables and within-class means of racehorse speed (qualitative explanatory variables)

	Frequency		Mean	
Dummy variable	**1**	**0**	**1**	**0**
Mare	7686	14497	16.0091	15.9391
Gelding	839	21344	16.0208	15.9611
Outer course	3068	19115	16.5636	15.8670
Turf course	8174	14009	16.4876	15.6575

Trainer evaluatoin	A	B	C	D
Frequency	378	12789	8281	735
Mean	16.3583	16.0288	15.8604	15.7824

Racetrack	Tokyo	Chukyo	Nakayama	Kyoto	Hakodate	Kokura	Niigata	Sapporo	Fukushima	Hanshin
Frequency	1450	1775	4622	4469	490	970	2734	415	833	4425
Mean	16.5103	15.6112	15.6975	15.9365	16.4354	16.5154	15.9254	16.2837	16.4202	15.9644

Horse weight Figure 3.2 shows that lighter horses tend to run faster. Therefore, the regression coefficient of horse weight is expected to be negative in the regression model.

Gain or loss in horse weight There is no clear relationship between horse weight and speed in Figure fig:q5-quantitative-predictor. However, the gain or loss in horse weight is thought to be related to the training and diet menu of the racehorse. It is natural to assume that racehorses that gained weight strengthened their instantaneous power, whereas those that lost weight strengthened their endurance, and that this affected the unobserved abilities of the racehorses, thus we add gain or loss in horse weight as an explanatory variable.

Horse age In Figure 3.2, we found that speed generally tends to increase as horse age increases, but it decreases at later ages. At a closer look, we can observe that speed tends to increase until the horse reaches 8 years of age. It is natural to assume that racehorses grow up to a certain age, but after reaching a certain age, their ability as racehorses diminishes due to physical strength decline caused by aging. Gramm and Marksteiner (2010) report that the average speed of racehorses increases until they reach the age of four, but after that, the average speed gradually slows down. Their study used data from U.S. horse races, and it is thought that the absolute running ability of racehorses begins to decline at that age because the load weight is smaller after 4 years and 8 months of age due to the load weight regulations of U.S. horse races. Therefore, the horse age squared is also added as an explanatory variable in the analysis. This allows us to estimate the age at which a racehorse's ability begins to decline.

Weight to carry, mare dummy, gelding dummy The weight to carry in Figure 3.2 must be considered together with the mare and gelding dummies in Table 3.2. In JRA races, the weight to be borne by the jockey, saddle, etc. is regulated, and although this regulated weight varies depending on the type of race, mares and gelding generally bear a lighter weight than stallions. Therefore, contrary to our intuition, the higher the weight to carry in Figure 3.2, the higher the speed of the runners. In Table 3.2, the average speed of the stallions and gelding is lower than

that of the mares and gelding combined. This result may be due to the influence of the individual factors on each other. Hence, it would be significant to add these variables to the regression model to elucidate the influence of the individual factors.

track index This is an index published by netkeiba.com. The smaller this value is, the better the race time is expected to be. Figure 3.2 shows that the smaller the value of the track index, the faster the horse actually runs. Since we expect horses to run more easily when track conditions are better, and consequently to run faster, we would expect the regression coefficient to be negative when the track index is added as an explanatory variable.

Race index The race index is an index original to this study and is the average time of race t excluding racehorse i. The race index is an explanatory variable for the hypothesis that the racehorse speed in a given race is influenced by the speed of other runners. The regression coefficient of the race index is expected to be positive because the racehorse speed is not only affected by the speed of other runners, but also because the race with the highest average speed is likely to be run by the best runners. Figure 3.2 shows that the average speed clearly increases as the race index increases.

Outer course dummy This dummy variable indicates that the racehorse raced on the outer course of the racetrack. Generally, each racetrack has two courses, with one course encompassing the other course in a double circle-like structure. The outer course has a gentler curve than the inner course, and the deceleration at curves is relatively small, thus it is considered easier to run faster on the outer course. In fact, Table 3.2 shows that the horses that ran the outer course were able to run faster than those that ran the inner course. Therefore, the regression coefficient for the outside track dummy is expected to be positive.

Turf course dummy This dummy variable indicates the race was held on a turf course. There are two types of courses: turf and dirt courses. Horses tend to run faster on turf courses than on dirt courses. In fact, Table 3.2 shows that horses that ran on the turf course had a higher average speed than those that ran on the dirt course.

Therefore, the regression coefficient for the turf course dummy is expected to be positive.

Trainer evaluation This variable is the trainer's evaluation of the condition of the racehorses in each race on a four-point scale of A, B, C, and D. The data obtained include some cases without trainer evaluation, but the number of such cases was only 14, so the corresponding race data are excluded. In Table 3.2, the average speeds for the four levels, A, B, C, and D, are tabulated. Although the trainer evaluation is a subjective evaluation, Table 3.2 shows that the speed decreases as the trainer's evaluation decreases from A, B, C, to D. This trend suggests that the trainer evaluation may be capturing an aspect of the horse's ability that is not observed as an external numerical value. Therefore, we add dummy variables corresponding to the three levels of evaluation, A, B, and C, to the regression model. In this way, it would be possible to test the validity of the trainer evaluation based on D.

The track index is standardized to have a mean of 0 and a variance of 1 according to Silverman (2012). The positive quantitative variables, including horse weight, weight to carry, and race index, are logarithmized.

3.3 Hierarchical Bayesian Modeling

In this section, we introduce the panel data regression model of racehorse speed and explain the procedure for its hierarchical Bayesian analysis. Suppose that N racehorses ran and M jockeys rode in all the horse races held during the sample period, in the data used in this study, $N = 4,063$ and $M = 140$ (excluding Yutaka Take). Let $i \in \{1, \ldots, N\}$ denote the index of the racehorse and $j \in \{1, \ldots, M\}$ denote the index of the jockey. Then, let i be a racehorse that ran T_i times in the sample period, and let $t \in \{1, \ldots, T_i\}$ be the index of each race.

Additionally, we introduce the following variables.

- y_{it} — speed of racehorse i in race t (m/s)

- d_{ijt} — jockey dummy

$$d_{ijt} = \begin{cases} 1, & \text{(jockey } j \text{ rides racehorse } i \text{ in race } t.\text{);} \\ 0, & \text{(otherwise).} \end{cases}$$

- $\boldsymbol{x}_{it}$ — A $K \times 1$ vector of features for racehorse i and horse condition in race t (this contains the explanatory variables introduced in Section 3.2)

Then, consider the regression model in which the error term ϵ_{it} follows a normal distribution $\mathcal{N}(0, \sigma_\epsilon^2)$ for these variables.

$$y_{it} = \alpha_i + \sum_{j=1}^{M} \gamma_j d_{ijt} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad t \in \{1, \ldots, T_i\}. \tag{3.1}$$

The α_i in (3.1) is interpreted as the racehorse's individual effect, which cannot be explained by other factors. Contrarily, γ_j is the part of the racehorse's speed that is increased when the jockey j exceeds the performance of the reference jockey (Yutaka Take in this study) (when $\gamma_j > 0$) or decreased when the jockey's performance is below (when $\gamma_j < 0$). Since the right-hand side of (3.1) controls for racehorse's individual effect and racehorse and track conditions, γ_j is interpreted as the ability difference between the jockey j and reference jockey. Finally, $\boldsymbol{\beta}$ is a vector of regression coefficients for $K \times 1$, each of which is interpreted as the marginal effect of the variables in $\boldsymbol{x}_{it}$ on the racehorse speed.

Here, let us define vectors and matrixes as follows.

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT_i} \end{bmatrix}, \quad \boldsymbol{D}_i = \begin{bmatrix} d_{i11} & \cdots & d_{iM1} \\ \vdots & \ddots & \vdots \\ d_{i1T_i} & \cdots & d_{iMT_i} \end{bmatrix}, \quad \boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{iT_i} \end{bmatrix},$$

$$\boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT_i} \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_M \end{bmatrix}.$$

Then, the regression model of (3.1) corresponding to the run record of racehorse i can be summarized as

$$\boldsymbol{y}_i = \alpha_i \mathbf{1}_{T_i} + \boldsymbol{D}_i\boldsymbol{\gamma} + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}_{T_i}, \sigma_\epsilon^2 \boldsymbol{I}_{T_i}). \tag{3.2}$$

Here $\mathbf{1}_{T_i}$ is a $T_i \times 1$-vector whose elements are all ones, $\mathbf{0}_{T_i}$ is a $T_i \times 1$-vector whose elements are all zeros, and $\boldsymbol{I}_{T_i}$ is a T_i-dimensional unit matrix.

Lastly, we define

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} \mathbf{1}_{T_1} & & \\ & \ddots & \\ & & \mathbf{1}_{T_N} \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 \\ \vdots \\ \boldsymbol{D}_N \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix},$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{D} & \boldsymbol{X} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{bmatrix},$$

then the regression model for all N racehorses can be summarized as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{D}\boldsymbol{\gamma} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}_T, \sigma_\epsilon^2 \boldsymbol{I}_T\right), \quad T = \sum_{i=1}^{N} T_i. \end{aligned} \tag{3.3}$$

Next, we set up the posterior distribution to conduct a hierarchical Bayesian analysis on the regression model in (3.1). Since the conditional distribution of $\boldsymbol{y}$ under a given $\boldsymbol{Z}$ in (3.3) is $\mathcal{N}(\boldsymbol{Z}\boldsymbol{\delta}, \sigma_\epsilon^2 \boldsymbol{I})$, the likelihood of the unknown parameter $(\boldsymbol{\delta}, \sigma_\epsilon)$ is

$$p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma_\epsilon) \propto (\sigma_\epsilon^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta})\right] \tag{3.4}$$

$$\propto (\sigma_\epsilon^2)^{-\frac{T}{2}} \exp\left[-\frac{\sum_{i=1}^{T_i} e_{it}^2}{2\sigma_\epsilon^2}\right], \tag{3.5}$$

$$e_{it} = y_{it} - \alpha_i - \sum_{j=1}^{M} \gamma_j d_{jit} - \boldsymbol{x}_{it}'\boldsymbol{\beta}.$$

Then, the following prior distribution is assumed for the parameter $(\boldsymbol{\delta}, \sigma_\epsilon)$.

$$\boldsymbol{\delta} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_\alpha \mathbf{1}_N \\ \mu_\gamma \mathbf{1}_{M-1} \\ \boldsymbol{\mu}_\beta \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\alpha^2 \boldsymbol{I}_N & & \\ & \sigma_\gamma^2 \boldsymbol{I}_{M-1} & \\ & & \boldsymbol{\Sigma}_\beta \end{bmatrix}, \tag{3.6}$$

$$\sigma_\epsilon \sim \mathcal{C}^+(0, s_\epsilon). \tag{3.7}$$

Here $\mathcal{C}^+(0, \sigma_\epsilon)$ is the half-Cauchy distribution with the probability density:

$$p(\sigma_\epsilon|s_\epsilon) = \frac{2s_\epsilon}{\pi(\sigma_\epsilon^2 + s_\epsilon^2)}, \quad \sigma_\epsilon > 0, \quad s_\epsilon > 0.$$

Since this model is a hierarchical Bayesian model, we used the half-Cauchy distribution recommended by Gelman (2006) for hierarchical models. Note that the prior distribution in (3.6) is equivalent to assuming

$$\begin{aligned} &\alpha_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad i \in \{1, \ldots, N\}, \\ &\gamma_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\gamma, \sigma_\gamma^2), \quad j \in \{1, \ldots, M\}, \\ &\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \end{aligned} \tag{3.8}$$

If we regard racehorse i's individual effect α_i and jockey j's individual effect γ_j as random effects in (3.8), they can be interpreted as randomly generated from a normal distribution of $\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ and $\mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$. Contrarily, if α_i and γ_j are treated as fixed effects, $\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ and $\mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$ can be interpreted as a common prior distribution of the fixed effects as unknown parameters. Either interpretation does not change the fact that neither α_i nor γ_j are unobservable parameters in the Bayesian approach and does not affect the resulting posterior distribution of individual effects as parameters.

In the prior distribution of the parameter $(\boldsymbol{\delta}, \sigma_\epsilon)$, $(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ in (3.6) and s_ϵ in (3.7) are fixed to a specific value as hyperparameters. However, for $(\mu_\alpha, \mu_\gamma, \sigma_\alpha, \sigma_\gamma)$ in (3.6), we set the following hierarchical prior distribution

$$\begin{aligned} &\mu_\alpha \sim \mathcal{N}(\varphi_\alpha, \tau_\alpha^2), \quad \mu_\gamma \sim \mathcal{N}(\varphi_\gamma, \tau_\gamma^2), \\ &\sigma_\alpha \sim \mathcal{C}^+(0, s_\alpha), \quad \sigma_\gamma \sim \mathcal{C}^+(0, s_\gamma), \end{aligned} \tag{3.9}$$

and attempt to estimate it simultaneously with $(\boldsymbol{\delta}, \sigma_\epsilon)$ using a Bayesian approach.

The name hierarchical Bayesian analysis is derived from the use of this hierarchical prior distribution. One advantage of hierarchical Bayesian analysis is that $(\mu_\alpha, \mu_\gamma, \sigma_\alpha, \sigma_\gamma)$ in the prior distribution (3.8) are not fixed as hyperparameters, but can be estimated simultaneously with other parameters. Another advantage of this analysis is that the shrinkage method can be used to stabilize estimation of the individual effects. $(\varphi_\alpha, \varphi_\gamma, \tau_\alpha^2, \tau_\gamma^2, s_\alpha, s_\gamma)$ in (3.9) is fixed to a specific value as a hyperparameter.

From the above explanation, we can conclude that the parameters to be estimated in the hierarchical Bayesian analysis of the regression model (3.1) for the ability of the

racehorse and jockey are summarized as

$$\begin{aligned}\boldsymbol{\theta} &= (\boldsymbol{\delta}, \mu_\alpha, \mu_\gamma, \sigma_\alpha, \sigma_\gamma, \sigma_\epsilon) \\ &= (\alpha_1, \ldots, \alpha_N, \gamma_1, \ldots, \gamma_M, \beta_1, \ldots, \beta_K, \mu_\alpha, \mu_\gamma, \sigma_\alpha, \sigma_\gamma, \sigma_\epsilon).\end{aligned}$$

However, since the posterior distribution

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma_\epsilon)p(\boldsymbol{\theta}), \quad \mathcal{D} = (\boldsymbol{y}, \boldsymbol{Z}), \tag{3.10}$$

of parameter $\boldsymbol{\theta}$, consisting of the likelihood $p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma_\epsilon)$ of (3.4) and the prior distribution $p(\boldsymbol{\theta})$ of (3.6), (3.7) and (3.9), cannot be evaluated analytically, we will proceed with the hierarchical Bayesian analysis by using the Markov chain Monte Carlo (MCMC) method. The conditional posterior distribution of each parameter is derived as follows.

$$\boldsymbol{\delta}|\mathcal{D}, \boldsymbol{\theta}_{-\boldsymbol{\delta}} \sim \mathcal{N}\left(\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right), \tag{3.11}$$

$$\mu_\alpha|\mathcal{D}, \boldsymbol{\theta}_{-\mu_\alpha} \sim \mathcal{N}\left(\frac{\sigma_\alpha^{-2}\sum_{i=1}^N \alpha_i + \tau_\alpha^{-2}\varphi_\alpha}{\sigma_\alpha^{-2}N + \tau_\alpha^{-2}}, \frac{1}{\sigma_\alpha^{-2}N + \tau_\alpha^{-2}}\right), \tag{3.12}$$

$$\mu_\gamma|\mathcal{D}, \boldsymbol{\theta}_{-\mu_\gamma} \sim \mathcal{N}\left(\frac{\sigma_\gamma^{-2}\sum_{j=1}^M \gamma_i + \tau_\gamma^{-2}\varphi_\gamma}{\sigma_\gamma^{-2}M + \tau_\gamma^{-2}}, \frac{1}{\sigma_\gamma^{-2}M + \tau_\gamma^{-2}}\right), \tag{3.13}$$

$$\sigma_\alpha^2|\mathcal{D}, \boldsymbol{\theta}_{-\sigma_\alpha}, \xi_\alpha \sim \mathcal{IG}\left(\frac{N+1}{2}, \frac{\sum_{i=1}^N(\alpha_i - \mu_\alpha)^2}{2} + \frac{1}{\xi_\alpha}\right), \ \xi_\alpha|\sigma_\alpha \sim \mathcal{IG}\left(1, \frac{1}{\sigma_\alpha^2} + \frac{1}{s_\alpha^2}\right), \tag{3.14}$$

$$\sigma_\gamma^2|\mathcal{D}, \boldsymbol{\theta}_{-\sigma_\gamma}, \xi_\gamma \sim \mathcal{IG}\left(\frac{M+1}{2}, \frac{\sum_{j=1}^M(\gamma_j - \mu_\gamma)^2}{2} + \frac{1}{\xi_\gamma}\right), \ \xi_\gamma|\sigma_\gamma \sim \mathcal{IG}\left(1, \frac{1}{\sigma_\gamma^2} + \frac{1}{s_\gamma^2}\right), \tag{3.15}$$

$$\sigma_\epsilon^2|\mathcal{D}, \boldsymbol{\theta}_{-\sigma_\epsilon}, \xi_\epsilon \sim \mathcal{IG}\left(\frac{T+1}{2}, \frac{\sum_{i=1}^N\sum_{t=1}^{T_i} e_{it}^2}{2} + \frac{1}{\xi_\epsilon}\right), \ \xi_\epsilon|\sigma_\epsilon \sim \mathcal{IG}\left(1, \frac{1}{\sigma_\epsilon^2} + \frac{1}{s_\epsilon^2}\right), \tag{3.16}$$

where $\boldsymbol{\theta}_{-a}$ indicates that the parameter a is excluded from $\boldsymbol{\theta}$, and $\mathcal{IG}(a, b)$ is the inverse gamma distribution

$$p(x|a, b) = \frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}.$$

The derivations of these conditional posterior distributions are explained in Appendix.

In the conditional posterior distribution of (3.14)–(3.16), new latent variables $(\xi_\alpha, \xi_\gamma, \xi_\epsilon)$ are introduced. This is because $x \sim \mathcal{C}^+(0, a)$ is expressed as

$$x^2|z \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{z}\right), \quad z \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{a^2}\right), \tag{3.17}$$

and is used to derive (3.14). See Appendix for more details.

The conditional posterior distributions in (3.11)–(3.16) have all known efficient random number generation algorithms, such as the normal and inverse gamma distributions. Therefore, Gibbs sampling can be used to generate parameter $\boldsymbol{\theta}$ from the posterior distribution. However, when we applied simple Gibbs sampling from the conditional posterior distribution of (3.11)–(3.16) to the running record data used in this study, the convergence of the random series tended to take a long time. Therefore, we consider the racehorse's individual effect $\boldsymbol{\alpha}$ and the jockey's individual effect $\boldsymbol{\gamma}$ as parameters, and apply the ASIS proposed by Yu and Meng (2011) to improve the efficiency of random number generation. To demonstrate the ASIS algorithm used in this study, we first assume that $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \mu_\alpha, \mu_\gamma)$ are generated by Gibbs sampling, and consider the following transformation.

$$\begin{aligned} \tilde{\alpha}_i &= \alpha_i - \mu_\alpha, \quad i \in \{1, \ldots, N\}, \\ \tilde{\gamma}_j &= \gamma_j - \mu_\gamma, \quad j \in \{1, \ldots, M\}, \\ \tilde{y}_{it} &= y_{it} - \tilde{\alpha}_i - \sum_{j=1}^{M} \tilde{\gamma}_j d_{jit}, \quad t \in \{1, \ldots, T_i\}. \end{aligned} \tag{3.18}$$

Then, we use these to rewrite (3.1) as

$$\tilde{y}_{it} = \mu_\alpha + \mu_\gamma \tilde{d}_{it} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad \tilde{d}_{it} = \sum_{j=1}^{M} d_{jit}, \tag{3.19}$$

where $\tilde{d}_{it}$ is a dummy variable interpreted as

$$\tilde{d}_{it} = \begin{cases} 0, & \text{(Yutaka Take rides racehorse } i \text{ in race } t\text{);} \\ 1, & \text{(otherwise).} \end{cases}$$

Furthermore, by introducing the notation

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{y}_{11} \\ \vdots \\ \tilde{y}_{NT_N} \end{bmatrix}, \quad \tilde{\boldsymbol{Z}} = \begin{bmatrix} 1 & \tilde{d}_{11} & \boldsymbol{x}'_{11} \\ \vdots & \vdots & \vdots \\ 1 & \tilde{d}_{NT_N} & \boldsymbol{x}'_{NT_N} \end{bmatrix}, \quad \tilde{\boldsymbol{\delta}} = \begin{bmatrix} \mu_\alpha \\ \mu_\gamma \\ \boldsymbol{\beta} \end{bmatrix},$$

(3.19) becomes

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\delta}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}_T, \sigma_\epsilon^2 \boldsymbol{I}_N). \tag{3.20}$$

Then the conditional posterior distribution of $\tilde{\boldsymbol{\delta}}$ is obtained as

$$\tilde{\boldsymbol{\delta}}|\mathcal{D}, \boldsymbol{\theta}_{-\boldsymbol{\delta}} \sim \mathcal{N}\left(\left(\sigma_\epsilon^{-2}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{y}} + \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right), \left(\sigma_\epsilon^{-2}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}\right), \quad (3.21)$$

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \varphi_\alpha \\ \varphi_\gamma \\ \boldsymbol{\mu}_\beta \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tau_\alpha^2 & & \\ & \tau_\gamma^2 & \\ & & \boldsymbol{\Sigma}_\beta \end{bmatrix},$$

by exactly the same procedure with (3.11): thus, the Gibbs sampling with the addition of ASIS can be summarized as follows[7].

Step 1 Given the sth generated $\boldsymbol{\theta}^{(s)}$, apply Gibbs sampling based on the conditional posterior distribution of (3.11)–(3.16) to generate $\boldsymbol{\theta}^{(s+0.5)}$ and denote

$$\tilde{\alpha}_i^{(s+0.5)} = \alpha_i^{(s+0.5)} - \mu_\alpha^{(s+0.5)}, \quad \tilde{\gamma}_j^{(s+0.5)} = \gamma_j^{(s+0.5)} - \mu_\gamma^{(s+0.5)}.$$

Step 2 Given $\boldsymbol{\theta}^{(s+0.5)}$, apply Gibbs sampling based on the conditional posterior distribution of (3.21) and (3.14)–(3.16) to generate $(\boldsymbol{\beta}^{(s)}, \mu_\alpha^{(s)}, \mu_\gamma^{(s)}, \sigma_\alpha^{(s)}, \sigma_\gamma^{(s)}, \sigma_\epsilon^{(s)})$, and denote

$$\alpha_i^{(s)} = \tilde{\alpha}_i^{(s+0.5)} + \mu_\alpha^{(s)}, \quad \gamma_j^{(s)} = \tilde{\gamma}_j^{(s+0.5)} + \mu_\gamma^{(s)},$$

In this study, we use this Gibbs sampling to generate a Monte Carlo sample $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^S$ of $\boldsymbol{\theta}$ from the posterior distribution, and analyze the hierarchical Bayesian analysis of a racehorse's running ability and a jockey's ability.

However, as shown in Section 3.2, there are multiple candidate explanatory variables that should be included in the regression model of (3.1), and it is necessary to determine which variables should be included in the model in some way. In this study, the WAIC proposed by Watanabe (2010) is used as a model selection criterion for variable selection. According to Gelman *et al.* (2014), the WAIC for (3.1) is calculated as follows

$$\begin{aligned} \mathrm{WAIC} &= -2(\mathrm{lppd} - p_{\mathrm{WAIC}}), \\ \mathrm{lppd} &= \sum_{i=1}^N \sum_{t=1}^{T_i} \log\left(\frac{1}{S}\sum_{s=1}^S p(y_{it}|\boldsymbol{\theta}^{(s)})\right), \\ p_{\mathrm{WAIC}} &= \sum_{i=1}^N \sum_{t=1}^{T_i} \frac{1}{S-1}\sum_{s=1}^S \left(\log p(y_{it}|\boldsymbol{\theta}^{(s)}) - \frac{1}{S}\sum_{s=1}^S \log p(y_{it}|\boldsymbol{\theta}^{(s)})\right)^2, \end{aligned} \quad (3.22)$$

[7] In short, the prior distribution of $\tilde{\boldsymbol{\delta}}$ is $\mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

where $p(y_{it}|\boldsymbol{\theta}^{(s)})$ is calculated as the probability density of y_{it} (speed of racehorse i in race t) in (3.1) using $\boldsymbol{\theta}$ generated by the sth (after burn-in) Gibbs sampling. Among the characteristics introduced in Section 3.2 regarding the racehorse and the environment in the race, the four variables "horse age," "horse age squared," "weight to carry," "turf course dummy" and "racetrack dummy" are always included in the regression model, whereas "gain or loss in horse weight," "track index," "race index," "outside track dummy," "gelding dummy," and "trainer evaluation dummy." In other words, Gibbs sampling is performed for $2^6 = 64$ regression models, and WAIC is obtained from the resulting 64 sets of Monte Carlo samples $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$ using (3.22), then the combination of explanatory variables with the smallest WAIC value is adopted as the optimal regression model.

3.4 Results and Discussion on Empirical Analysis

To present the results of the hierarchical Bayesian analysis of racehorse running ability and jockey ability conducted with the data presented in Section 3.2, we first describe the detailed settings of the prior distribution and the Gibbs sampling used in this study. The hyperparameters $(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \varphi_\alpha, \varphi_\gamma, \tau_\alpha^2, \tau_\gamma^2, s_\alpha, s_\gamma, s_\epsilon)$ of the prior distribution are set as follows.

$$\begin{aligned} &\boldsymbol{\mu}_\beta = \mathbf{0}_K, \quad \boldsymbol{\Sigma}_\beta = 100\boldsymbol{I}_K, \quad \varphi_\alpha = \varphi_\gamma = 0, \quad \tau_\alpha^2 = \tau_\gamma^2 = 100, \\ &s_\alpha = s_\gamma = s_\epsilon = 1. \end{aligned} \tag{3.23}$$

The Gibbs sampling procedure begins with 2,000 random number generation as burn-in until the random series of parameters become stable, followed by 10,000 random number generation to obtain the Monte Carlo sample $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$ used for parameter estimation and WAIC calculations. That is, $S = 10,000$. As a result of the analysis, the inefficiency factors are in the single digits for all of the generated random number series (mostly < 2), and the results of Geweke's convergence diagnostic proposed by Geweke (1992) are also good, exceeding 0.2 for all variables.

The WAICs are calculated using this Monte Carlo sample, and as shown in Table 3.3, the results show that the regression model with all explanatory variables is the

best. Although the regression coefficients for gelding dummies and weight to carry are not significant because they include 0 in the 95% credible interval, the model including these variables is selected as the best. The second and third models selected are regression models that excluded the gelding dummy and gain or loss in horse weight from the explanatory variables, respectively. Surely, the exclusion of the gelding dummy is understandable, given that a gelding is a castrated stallion, and its inborn physical abilities do not change. Moreover, since gain or loss in horse weight has nothing to do with the racehorse's intrinsic ability, it is reasonable to exclude this variable.

The posterior statistics of the regression coefficient $\boldsymbol{\beta}$ and other parameters for the top three models are summarized in Table 3.3. In Table 3.3, there is no remarkable difference in the posterior statistics for each parameter among the models. Therefore, Model 1, which has the lowest WAIC, is used to analyze the racehorse's and jockey's performance.

Our first interest is in regression coefficient $\boldsymbol{\beta}$: thus, in turn, we will look at the posterior statistics of $\boldsymbol{\beta}$ summarized in Table 3.3. Note that in the following discussion, the significance of a regression coefficient means that the 95% credible interval does not contain zero, and a significantly positive (negative) regression coefficient means that the 95% credible interval contains only positive (negative) values.

First, we discuss the explanatory variables for the physical characteristics of racehorses. The regression coefficient of horse weight is significantly positive. This indicates that heavy racehorses, which are thick and muscular, are more suitable for 1800-m races than light racehorses, which are thin and lean. This result is contrary to the prediction presented in Section 3.2, but this may be interpreted as a result of the fact that racehorses competing in races with speedier conditions tend to be lighter in weight. For example, as described below, the regression coefficients for the turf and outer courses are significant and positive, but the racehorses that participated in the race conducted in the turf and outer courses are on average 15.85 and 11.86 kg lighter than those that did not participate in these courses, respectively. Therefore, the opposite prediction may have been made when only horse weight is taken into account. The gain or loss in horse weight is significantly negative, but this may be due to the fact that horses had lost weight because of adjusting their condition.

Next, horse age must be interpreted in conjunction with horse age squared. From the

formula for the quadratic function $ax^2 + bx = a(x + b/2a)^2 + c$, noting that the quadratic term is negative, the descending junction point is 4.18 years old by substituting the second-order coefficient of horse age, $a = -0.2359 \times 10^{-3}$, and the first-order coefficient, $b = 1.9701 \times 10^{-3}$, for $x = -b/2a$ respectively. This indicates that the ability of a racehorse gradually declines from 4 years of age. The average speeds in the speed classes shown in Figure 3.2 become larger as the horse age increases, which probably reflects a survival bias. The fastest and most talented racehorses survive in the racing world, whereas the slowest racehorses are retired at a young age because they are judged to have no chance of winning even if they run. Therefore, at first glance, it would appear that racehorses with larger ages are faster than those with smaller ages, and many of the older racehorses are superior individuals who have achieved success throughout their racing careers. However, when we control for individual ability, we find that the ability declines after 4 years of age. The regression coefficient for the mare dummy is significantly negative, but the gelding dummy is not significant. The negative coefficients are probably appropriate given that mares are on average less physically fit than stallions. However, since the sign of the regression coefficient for the gelding dummy is unclear, we can say that the stallion's castration does not affect his performance as a racehorse.

Then, we examined the explanatory variables related to the race environment. First, the regression coefficient for the weight to carry is negative for the posterior mean, but the 95% credible interval includes 0. The sign of the regression coefficient is understandable since the jockey's weight is a burden for the racehorse, but it is not a factor that significantly affects speed. The regression coefficient for the track index is significantly negative, but this is a reasonable result because a smaller track index indicates a better track condition. For the race index, the regression coefficient is significantly positive, indicating that the faster the other racehorses run in the same race, the faster a racehorse runs. The regression coefficients for the outer and turf course dummies are also significantly positive as expected. These results are reasonable considering the fact that the outer course has more gentle curves and is less likely to slow down, and the turf course is easier to run than the dirt course. For the trainer evaluation dummies, the regression coefficients for the posterior means are all significantly positive. Additionally, the values become smaller as the ratings decrease from A to C, indicating that racehorses

with higher trainer evaluations in advance are able to perform as expected in the races. Finally, the racetrack dummies are significantly positive for Kokura and Fukushima, and significantly negative for Chukyo, Nakayama, Hakodate, Sapporo, and Hanshin. Since the turf quality, track dirt, and elevation difference are naturally different between Tokyo Racecourse, which is the criterion for the racetrack dummies, and other racetracks, it is reasonable to assume that the speed is significantly higher or lower at these tracks than at the other tracks.

Next, let us examine the racehorse's individual effect for racehorses and jockeys. First, consider the racehorse's individual effect. Tables 3.4 and 3.5 show the results of the analysis for the top 10 and bottom 10 racehorses, respectively, and Figure 3.4 shows the posterior means and 95% credible intervals of the individual effects for the 4,063 horses. The horizontal axis in Figure 3.4(a) is the rank of the racehorse's individual effect α_i ($i = 1, \ldots, 4,063$) sorted by the posterior mean. Contrarily, the vertical axis of Figure 3.4(a) shows the individual effect α_i of each horse as a contribution to speed. The results of the top racehorse in Table 3.4 show that the racehorses with a high individual effect have all achieved excellent total results. Contrarily, although the names of individual racehorses are withheld to protect the honor of the racehorses and their related jockeys, the results of the racehorse's individual effects are not as good as those of the racehorses with low individual effects in Table 3.5. Furthermore, since the difference in effects between the highest and lowest ranked racehorses is 2.65%, the maximum effect on running distance due to differences in racehorse ability at the finish line has the potential to be 47.7 m. All the above mentioned data suggest that the estimated racehorse's individual effect well captures the racehorse's running ability, which is not observed, although no significant difference can be confirmed.

As for the jockey's individual effect, Tables 3.6 and 3.7 show the results of the analysis for the top 10 and bottom 10 jockeys, respectively, and Figure 3.4 shows the posterior mean and 95% confidence interval of the individual effect for 140 jockeys with Yutaka Take as the benchmark jockey. The horizontal axis in Figure 3.4(b) shows the rank of each jockey's individual effect γ_j ($j = 1, \ldots, 140$) sorted by the posterior mean. The vertical axis of Figure 3.4(b) shows the jockey's individual effect γ_j as a contribution to speed. The results for the high-ranking jockeys in Table 3.6 show that the posterior means of

the top 9 jockeys are higher than 0. Considering that we adopted Yutaka Take, who is well known as a great jockey, as the benchmark, it is indicative that these 9 jockeys are very capable. However, the 95% credible intervals of the jockey's individual effects for eight of these nine jockeys all contain 0, indicating that there is no significant difference between them and Yutaka Take. These results indicate that Yutaka Take is as good a jockey as his reputation, and he is almost as good as most of the jockeys we compared in this study. Although the names of the jockeys in the lower ranks of Table 3.7 are withheld for the sake of honor, a comparison of Tables 3.6 and 3.7 shows that the jockeys in the higher ranks have raced more frequently and have more experience than those in the lower ranks[8]. The effect of the difference in ability between the best and the worst jockeys is 0.698%, and the effect of the difference in jockey ability at the finish line on the distance can cause a difference of up to 12.6 m. In horse racing, the horses' ability and condition of the horses tend to be the focus of attention. However, the fact that a jockey can make a difference of approximately 12.6 m that can significantly overturn a ranking indicates that horse racing is not simply a mere game between horses, but a highly developed sport between human beings.

To check the robustness of the estimation results, the prior distributions of $(\sigma_\epsilon, \sigma_\alpha, \sigma_\gamma)$ in (3.7) and (3.9) are replaced half-Cauchy distributions with inverse gamma distributions

$$\sigma_\epsilon^2 \sim \mathcal{IG}(0.0001, 0.0001), \quad \sigma_\alpha^2 \sim \mathcal{IG}(0.0001, 0.0001), \quad \sigma_\gamma^2 \sim \mathcal{IG}(0.0001, 0.0001),$$

and Bayesian estimation is performed by Gibbs sampling with ASIS applied as well. Although the details are omitted due to the limited number of papers, the best model by WAIC is the same as that of the half-Cauchy distribution from the first to the fourth ranks even when the prior distributions are replaced with the inverse gamma distributions, and there is almost no difference in the posterior statistics. The same analysis is conducted for the data excluding racehorses that had raced < 4 times in 1800-m races instead of three. Applying this exclusion criterion, the total number of runs is 19,014, consisting of 2,922 racehorses and 140 jockeys. In this case, the results of model selection by WAIC are the same for all of the top five models as when the criterion of the three runs is applied.

[8]Since we could not confirm the exact data for the number of races run outside Japan, we compared the number of races run in Japan.

Next, we show the improvement in sampling efficiency by applying ASIS to the simulation. Figures 3.6 and 2.2 show respectively the MCMC sampling process and correlogram of $(\mu_\alpha, \mu_\gamma, \sigma_\alpha, \sigma_\gamma)$ in (3.9). Figure 3.6 clearly reveals that the sampling efficiency of the three parameters except for σ_γ has improved, especially for μ_α, which clearly does not converge when ASIS is not applied (a), although it converges when ASIS is applied (b). Additionally, for μ_γ, the sampling bias can be seen in (e) where ASIS is not applied and the autocorrelation is large, whereas in (f) where ASIS is applied, the bias is solved and convergence is attained. In fact, using Geweke's convergence diagnostic, the results are good for all four parameters.

The correlograms in Figure 2.2 display that, for autocorrelations (a) and (c) for which ASIS is not applied for μ_α, μ_γ for which sampling is inefficient and does not converge in Figure 3.6, the values are at inappropriate levels to use the MCMC method for all lags from the first to the 20th order. Especially for μ_α, considering that the value of Figure 2.2(a) sticks to approximately 1 even at lag 20, the number of samplings required to determine that the model is statistically converged without the application of ASIS is enormous, In the conventional approach, the model would have to be modified because the analysis would have been impossible. However, the autocorrelation (b), (d) of μ_α, μ_γ with ASIS applied shows that the autocorrelation becomes smaller and closer to zero as the lag order increases, indicating that a dramatic improvement in sampling efficiency has been realized. This improvement in efficiency makes the model of this study analyzable.

Finally, to evaluate the validity of the selected model, we demonstrated the accuracy with which the model can predict speed. We split 90% (19,965 runs) of the data randomly selected from the data into training data and the remaining 10% (2,218 runs) of the data into validation data, and compared the speed of the actual run with that of the model trained from the training data only. Figure 3.3 shows the actual speed of the actual run on the horizontal axis and the predicted speed of the run on the vertical axis. The solid red line drawn diagonally represents the area where the actual and predicted values coincide, and the observations are generally distributed around the solid line, indicating that the accuracy of the model's predictions is high and that it is a reasonable model.

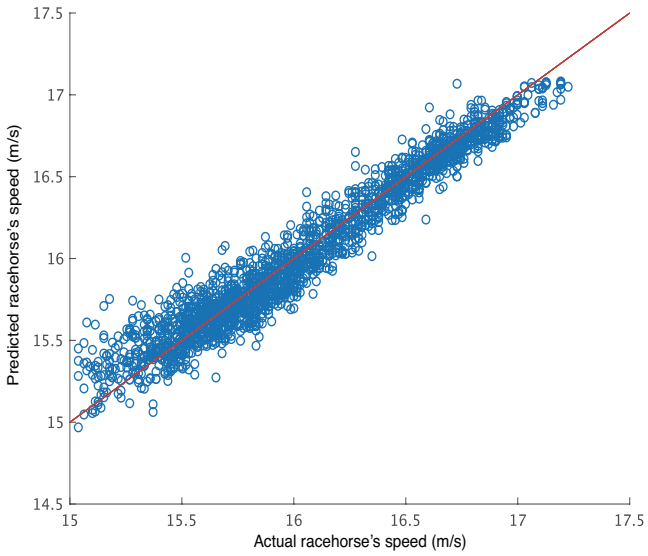
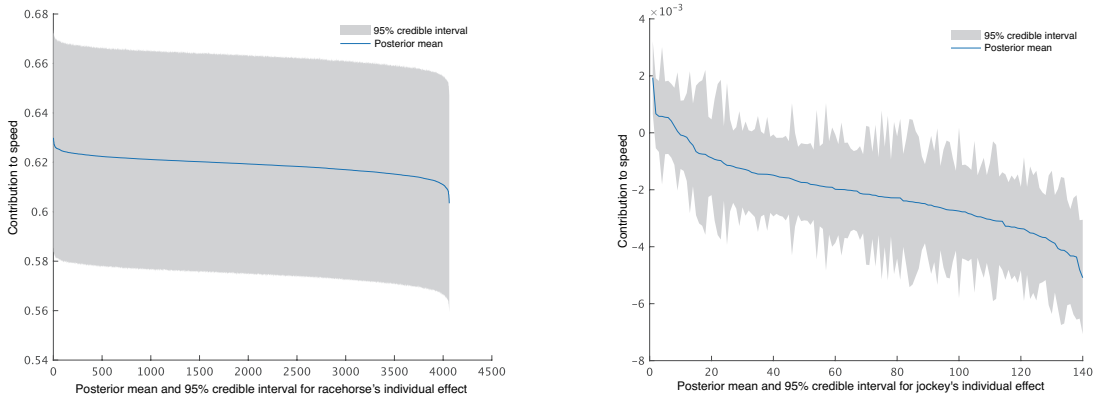


Figure 3.3: Actual and predicted speeds from the best model



(a) 4,063 racehorse's individual effect

(b) 140 jockey's individual effect with Yutaka Take as the benchmark

Figure 3.4: Racehorse's and jockey's individual effect

Table 3.3: Posterior statistics for the top three models by WAIC

$\boldsymbol{\beta}$	Model 1 (best model)	Model 2	Model 3
Horse weight*	4.3321×10^{-3}	4.3007×10^{-3}	4.3009×10^{-3}
	$[1.2864 \times 10^{-3}, 7.3972 \times 10^{-3}]$	$[1.2126 \times 10^{-3}, 7.3114 \times 10^{-3}]$	$[1.2171 \times 10^{-3}, 7.3051 \times 10^{-3}]$
Horse age	1.9701×10^{-3}	1.9834×10^{-3}	1.9760×10^{-3}
	$[1.3412 \times 10^{-3}, 2.6020 \times 10^{-3}]$	$[1.3590 \times 10^{-3}, 2.6365 \times 10^{-3}]$	$[1.3279 \times 10^{-3}, 2.6140 \times 10^{-3}]$
Horse age squared	-0.2359×10^{-3}	-0.2369×10^{-3}	-0.2359×10^{-3}
	$[-0.3041 \times 10^{-3}, -0.1671 \times 10^{-3}]$	$[-0.3080 \times 10^{-3}, -0.1693 \times 10^{-3}]$	$[-0.3053 \times 10^{-3}, -0.1664 \times 10^{-3}]$
Gain or loss in horse weight	-0.0248×10^{-3}	-0.0248×10^{-3}	
	$[-0.0423 \times 10^{-3}, -0.0072 \times 10^{-3}]$	$[-0.0427 \times 10^{-3}, -0.0070 \times 10^{-3}]$	
Mare dummy	-0.7326×10^{-3}	-0.7435×10^{-3}	-0.7576×10^{-3}
	$[-1.1832 \times 10^{-3}, -0.2778 \times 10^{-3}]$	$[-1.1921 \times 10^{-3}, -0.2831 \times 10^{-3}]$	$[-1.2104 \times 10^{-3}, -0.2908 \times 10^{-3}]$
Gelding dummy	0.1793×10^{-3}		0.2052×10^{-3}
	$[-0.6128 \times 10^{-3}, 0.9834 \times 10^{-3}]$		$[-0.5946 \times 10^{-3}, 0.9993 \times 10^{-3}]$
Weight to carry*	-2.0011×10^{-3}	-1.9986×10^{-3}	-2.0028×10^{-3}
	$[-8.7095 \times 10^{-3}, 4.7337 \times 10^{-3}]$	$[-8.7194 \times 10^{-3}, 4.8627 \times 10^{-3}]$	$[-8.8484 \times 10^{-3}, 4.8023 \times 10^{-3}]$
Trainer evaluation dummies			
A	5.3830×10^{-3}	5.3678×10^{-3}	5.3692×10^{-3}
	$[4.3416 \times 10^{-3}, 6.4635 \times 10^{-3}]$	$[4.2889 \times 10^{-3}, 6.4601 \times 10^{-3}]$	$[4.2883 \times 10^{-3}, 6.4570 \times 10^{-3}]$
B	3.4841×10^{-3}	3.4718×10^{-3}	3.4721×10^{-3}
	$[2.8561 \times 10^{-3}, 4.1062 \times 10^{-3}]$	$[2.8554 \times 10^{-3}, 4.1209 \times 10^{-3}]$	$[2.8535 \times 10^{-3}, 4.1204 \times 10^{-3}]$
C	1.3586×10^{-3}	1.3446×10^{-3}	1.3394×10^{-3}
	$[0.7414 \times 10^{-3}, 1.9850 \times 10^{-3}]$	$[0.7205 \times 10^{-3}, 1.9774 \times 10^{-3}]$	$[0.7146 \times 10^{-3}, 1.9715 \times 10^{-3}]$

Note: "horse age", "horse age squared", "weight to carry", "turf course dummy", and "racetrack dummy" are included in all models.

Data marked with $*$ are logarithmic.

The $**$ mark indicates that the data are standardized.

In each cell, the upper number is the posterior mean and the lower $[\cdot,\cdot]$ is the 95% credible interval.

Table 3.3: Posterior statistics for the top three models by WAIC (continued)

$\boldsymbol{\beta}$	Model 1 (best model)	Model 2	Model 3
Racetrack dummies			
Chukyo	-2.0140×10^{-3}	-1.9986×10^{-3}	-1.9628×10^{-3}
	$[-2.8948 \times 10^{-3}$, -1.1491×10^{-3}]	$[-2.8928 \times 10^{-3}$, $-1.1156 \times 10^{-3}]$	$[-2.8549 \times 10^{-3}$, $-1.0793 \times 10^{-3}]$
Nakayama	-2.4091×10^{-3}	-2.4021×10^{-3}	-2.3977×10^{-3}
	$[-3.1030 \times 10^{-3}$, -1.7310×10^{-3}]	$[-3.0892 \times 10^{-3}$, $-1.7093 \times 10^{-3}]$	$[-3.0840 \times 10^{-3}$, $-1.7049 \times 10^{-3}]$
Kyoto	-0.7926×10^{-3}	-0.7792×10^{-3}	-0.7684×10^{-3}
	$[-1.6417 \times 10^{-3}$, 0.0461×10^{-3}]	$[-1.6339 \times 10^{-3}$, $0.0752 \times 10^{-3}]$	$[-1.6218 \times 10^{-3}$, $0.0836 \times 10^{-3}]$
Hakodate	-1.1168×10^{-3}	-1.1058×10^{-3}	-1.1380×10^{-3}
	$[-2.0202 \times 10^{-3}$, -0.2124×10^{-3}]	$[-1.9844 \times 10^{-3}$, $-0.1865 \times 10^{-3}]$	$[-2.0179 \times 10^{-3}$, $-0.2211 \times 10^{-3}]$
Kokura	0.9387×10^{-3}	0.9487×10^{-3}	0.9644×10^{-3}
	$[0.1677 \times 10^{-3}$, 1.6937×10^{-3}]	$[0.1957 \times 10^{-3}$, $1.6939 \times 10^{-3}]$	$[0.2098 \times 10^{-3}$, $1.7983 \times 10^{-3}]$
Niigata	0.0719×10^{-3}	0.0800×10^{-3}	0.0973×10^{-3}
	$[-0.7418 \times 10^{-3}$, 0.8871×10^{-3}]	$[-0.7591 \times 10^{-3}$, $0.8837 \times 10^{-3}]$	$[-0.7429 \times 10^{-3}$, $0.9030 \times 10^{-3}]$
Sapporo	-3.9396×10^{-3}	-3.9450×10^{-3}	-3.9314×10^{-3}
	$[-4.8776 \times 10^{-3}$, -2.9831×10^{-3}]	$[-4.9286 \times 10^{-3}$, $-2.9553 \times 10^{-3}]$	$[-2.9197 \times 10^{-3}$, $-2.9411 \times 10^{-3}]$
Fukushima	0.9086×10^{-3}	0.9216×10^{-3}	0.9523×10^{-3}
	$[0.1603 \times 10^{-3}$, 1.6228×10^{-3}]	$[0.1941 \times 10^{-3}$, $1.6355 \times 10^{-3}]$	$[0.2273 \times 10^{-3}$, $1.6662 \times 10^{-3}]$
Hanshin	-0.9096×10^{-3}	-0.8962×10^{-3}	-0.8598×10^{-3}
	$[-1.7695 \times 10^{-3}$, -0.0689×10^{-3}]	$[-1.7489 \times 10^{-3}$, $-0.0556 \times 10^{-3}]$	$[-1.7128 \times 10^{-3}$, $-0.0191 \times 10^{-3}]$
μ_α	2.7439	2.7441	2.7475
	[2.7113 , 2.7770]	[2.7112, 2.7769]	[2.7154, 2.7811]
μ_γ	-2.1057×10^{-3}	-2.1145×10^{-3}	-2.1107×10^{-3}
	$[-3.1258 \times 10^{-3}$, 1.0776×10^{-3}]	$[-3.1256 \times 10^{-3}$, $-1.1307 \times 10^{-3}]$	$[-3.1282 \times 10^{-3}$, $-1.1285 \times 10^{-3}]$
σ_α	4.2679×10^{-3}	4.2642×10^{-3}	4.2587×10^{-3}
	$[4.0790 \times 10^{-3}$, 4.4498×10^{-3}]	$[4.0806 \times 10^{-3}$, $4.4502 \times 10^{-3}]$	$[4.0756 \times 10^{-3}$, $4.4449 \times 10^{-3}]$
σ_γ	1.5005×10^{-3}	1.5092×10^{-3}	1.5079×10^{-3}
	$[1.2501 \times 10^{-3}$, 1.7948×10^{-3}]	$[1.2537 \times 10^{-3}$, $1.7971 \times 10^{-3}]$	$[1.2552 \times 10^{-3}$, $1.7975 \times 10^{-3}]$
σ_ϵ	7.6718×10^{-3}	7.6735×10^{-3}	7.6761×10^{-3}
	$[7.5904 \times 10^{-3}$, 7.7517×10^{-3}]	$[7.5940 \times 10^{-3}$, $7.7540 \times 10^{-3}]$	$[7.5964 \times 10^{-3}$, $7.7573 \times 10^{-3}]$
WAIC	-150420.6	-150418.2	-150406.4

Note: "horse age", "horse age squared", "weight to carry", "turf course dummy", and "racetrack dummy" are included in all models.

Data marked with $*$ are logarithmic.

The $**$ mark indicates that the data are standardized.

In each cell, the upper number is the posterior mean and the lower $[\cdot,\cdot]$ is the 95% credible interval.

Table 3.4: Top 10 racehorse information

			Racehorse's individual effect	
Racehorse name	Date of birth	Win-loss records*	Posterior mean	95% credible interval
T M Jinsoku	04/21/2012	9 wins in 28 races [9-6-4-9]	0.6299	[0.5854, 0.6737]
Cloudscape	02/13/2016	0 wins in 6 races [0-3-1-2]	0.6295	[0.5854, 0.6735]
Jordan King	05/26/2013	2 wins in 8 races [2-5-0-1]	0.6286	[0.5838, 0.6729]
Inti	04/08/2014	5 wins in 6 races [5-0-0-1]	0.6284	[0.5838, 0.6719]
Highland Peak	05/16/2014	6 wins in 14 races [6-4-1-3]	0.6283	[0.5837, 0.6723]
Omega Perfume	04/06/2015	4 wins in 7 races [4-1-1-1]	0.6280	[0.5835, 0.6724]
Suzuka Burg	02/24/2014	2 wins in 8 races [2-4-0-2]	0.6273	[0.5825, 0.6714]
Pegasus Shine	04/24/2014	3 wins in 6 races [3-2-1-0]	0.6272	[0.5826, 0.6714]
Le Vent Se Leve	01/26/2015	3 wins in 4 races[3-1-0-0]	0.6271	[0.5828, 0.6715]
Meiner Brocken	03/27/2014	3 wins in 27 races [3-7-6-11]	0.6272	[0.5830, 0.6706]

∗ The table shows the total number of races in Central Japan horse races up to 2018.

Table 3.5: Bottom 10 racehorse information

			Racehorse's individual effect	
Racehorse name	Date of birth	Win-loss records*	Posterior mean	95% credible interval
Racehorse A	03/–/2014	0 wins in 4 races [0-0-0-4]	0.6034	[0.5593, 0.6471]
Racehorse B	04/–/2013	1 wins in 10 races [1-0-0-9]	0.6043	[0.5601, 0.6478]
Racehorse C	03/–/2013	0 wins in 5 races [0-0-0-5]	0.6053	[0.5608, 0.6490]
Racehorse D	04/–/2016	0 wins in 3 races [0-0-0-3]	0.6056	[0.5614, 0.6493]
Racehorse E	02/–/2014	0 wins in 5 races [0-0-0-5]	0.6062	[0.5617, 0.6498]
Racehorse F	05/–/2011	3 wins in 34 races [3-6-1-24]	0.6069	[0.5626, 0.6508]
Racehorse G	03/–/2014	0 wins in 4 races [0-0-0-4]	0.6070	[0.5626, 0.6510]
Racehorse H	04/–/2012	1 wins in 18 races [1-1-2-14]	0.6071	[0.5629, 0.6510]
Racehorse I	04/–/2014	0 wins in 7 races [0-0-0-7]	0.6072	[0.5628, 0.6510]
Racehorse J	02/–/2014	0 wins in 4 races [0-0-0-4]	0.6081	[0.5639, 0.6520]

∗ The table shows the total number of races in Central Japan horse races up to 2018.

Table 3.6: Top 10 jockey information

			jockey's individual effect	
Jockey name	Year of first riding*	# of races run**	Posterior mean	95% credible interval
Ryuji Wada	1996	16706	1.9184×10^{-3}	$[0.6794 \times 10^{-3}, 3.1833 \times 10^{-3}]$
Shane Foley	2007	256	0.6390×10^{-3}	$[-0.5694 \times 10^{-3}, 1.8776 \times 10^{-3}]$
Hironobu Tanabe	2002	9992	0.5748×10^{-3}	$[-0.6819 \times 10^{-3}, 1.8081 \times 10^{-3}]$
Suguru Hamanaka	2007	8723	0.5574×10^{-3}	$[-1.8409 \times 10^{-3}, 2.9858 \times 10^{-3}]$
Yuichi Fukunaga	1996	16684	0.5554×10^{-3}	$[-0.7420 \times 10^{-3}, 1.8188 \times 10^{-3}]$
Yuga Kawada	2004	9405	0.5287×10^{-3}	$[-0.7333 \times 10^{-3}, 1.8212 \times 10^{-3}]$
Joao Moreira	2014	347	0.3547×10^{-3}	$[-0.9309 \times 10^{-3}, 1.6491 \times 10^{-3}]$
Mirco Demuro	2004	5063	0.1942×10^{-3}	$[-1.1732 \times 10^{-3}, 1.5333 \times 10^{-3}]$
Keita Tosaki	2005	6538	0.0497×10^{-3}	$[-1.9985 \times 10^{-3}, 2.0965 \times 10^{-3}]$
Christophe Lemaire	1999	4904	-0.0910×10^{-3}	$[-1.2939 \times 10^{-3}, 1.1212 \times 10^{-3}]$

∗ This mark indicates the year of the first ride whether in a domestic or international race.

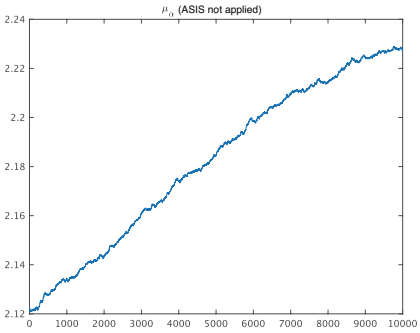
∗∗ This mark indicates the number of domestic races only.

Table 3.7: Bottom 10 jockey information

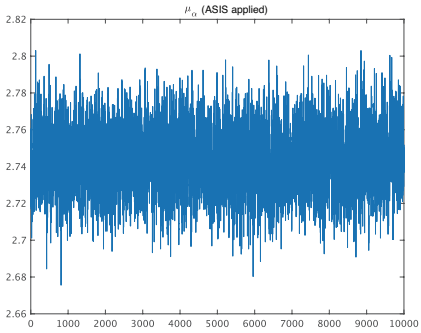
			jockey's individual effect	
Jockey name	Year of first riding*	# of races run†	Posterior mean	95% credible interval
Jockey A	2013	1837	-5.0592×10^{-3}	$[-7.0296 \times 10^{-3}, -3.0272 \times 10^{-3}]$
Jockey B	2015	869	-4.7377×10^{-3}	$[-6.4222 \times 10^{-3}, -2.9959 \times 10^{-3}]$
Jockey C	1993	7192	-4.3762×10^{-3}	$[-6.5525 \times 10^{-3}, -2.2102 \times 10^{-3}]$
Jockey D	2010	1727	-4.3306×10^{-3}	$[-6.3710 \times 10^{-3}, -2.3061 \times 10^{-3}]$
Jockey E	2016	1281	-4.2910×10^{-3}	$[-5.9396 \times 10^{-3}, -2.6827 \times 10^{-3}]$
Jockey F	1995	6411	-4.2002×10^{-3}	$[-6.0308 \times 10^{-3}, -2.3493 \times 10^{-3}]$
Jockey G	2001	3954	-4.1596×10^{-3}	$[-6.7135 \times 10^{-3}, -1.6041 \times 10^{-3}]$
Jockey H	2006	3703	-4.1369×10^{-3}	$[-6.7727 \times 10^{-3}, -1.5874 \times 10^{-3}]$
Jockey I	2008	3403	-4.0557×10^{-3}	$[-6.1679 \times 10^{-3}, -1.9684 \times 10^{-3}]$
Jockey J	2015	1554	-3.8123×10^{-3}	$[-5.5383 \times 10^{-3}, -2.1056 \times 10^{-3}]$

∗ This mark indicates the year of the first ride whether in a domestic or international race.

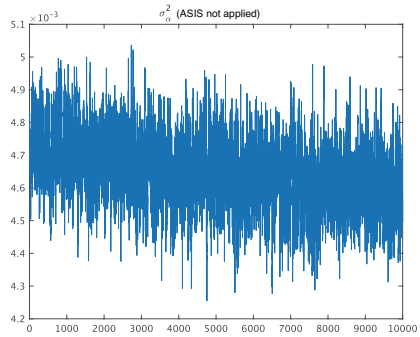
† The jockeys in the bottom 10 are all Japanese and they ride only in domestic races.



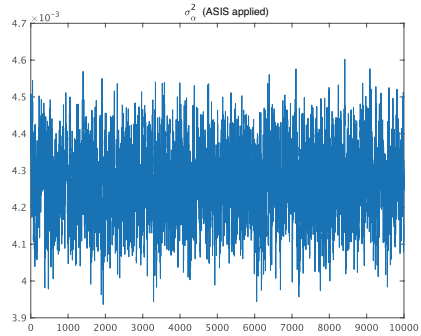
(a) μ_α (ASIS not applied)



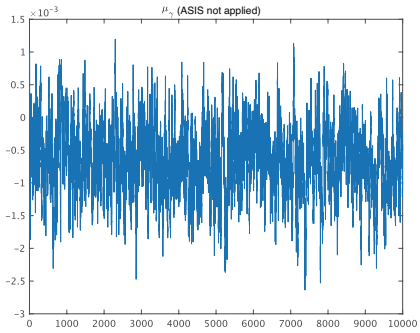
(b) μ_α (ASIS applied)



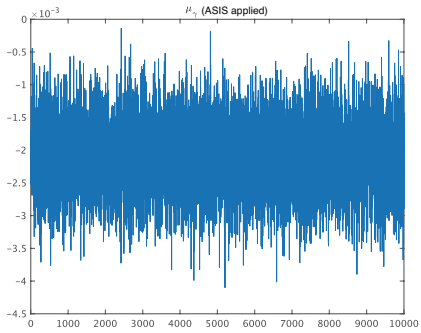
(c) σ_α (ASIS not applied)



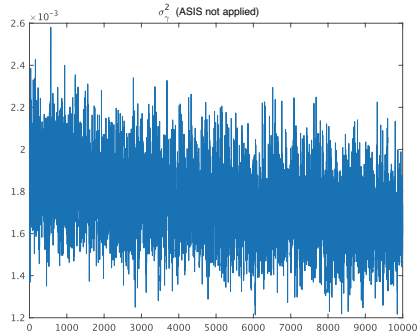
(d) σ_α (ASIS applied)



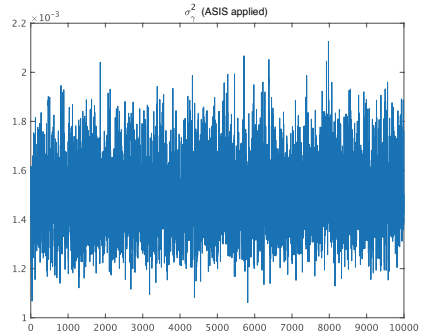
(e) μ_γ (ASIS not applied)



(f) μ_γ (ASIS applied)

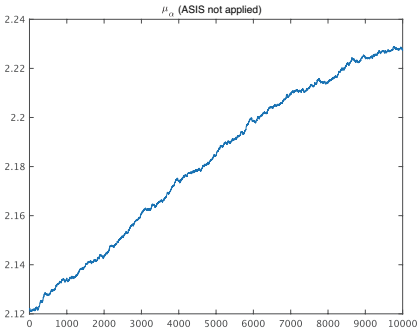


(g) σ_γ (ASIS not applied)

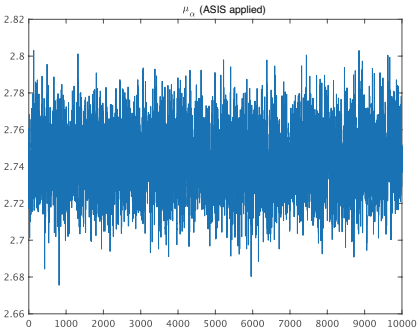


(h) σ_γ (ASIS applied)

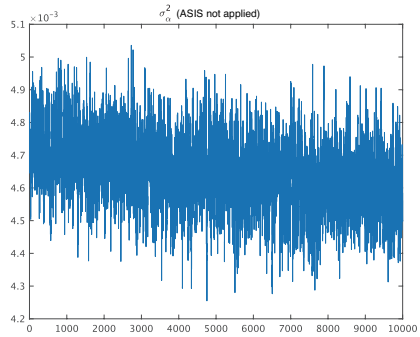
Figure 3.5: 10,000 sampling after 2,000 burn-ins



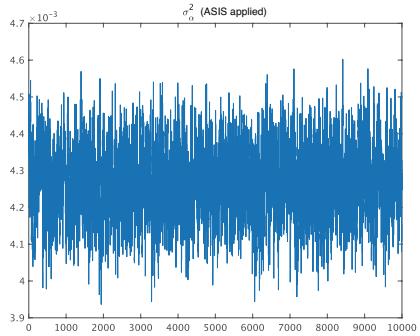
(a) μ_α (ASIS not applied)



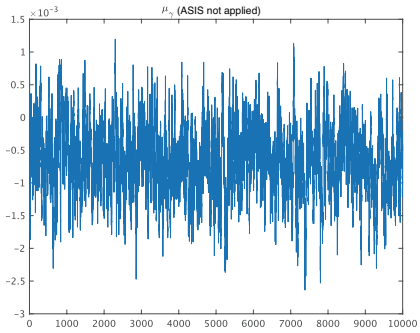
(b) μ_α (ASIS applied)



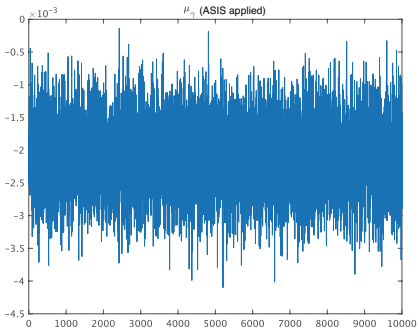
(c) σ_α (ASIS not applied)



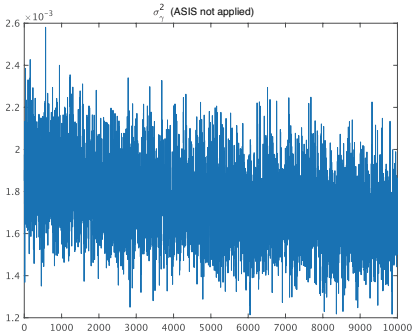
(d) σ_α (ASIS applied)



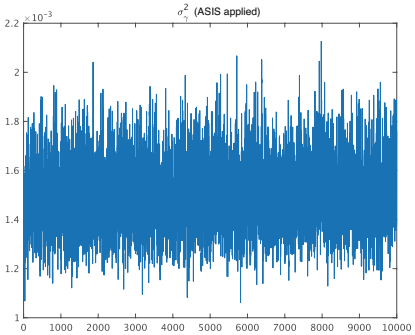
(e) μ_γ (ASIS not applied)



(f) μ_γ (ASIS applied)

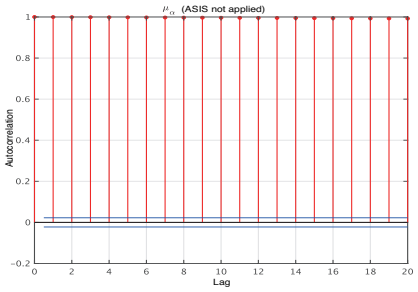


(g) σ_γ (ASIS not applied)

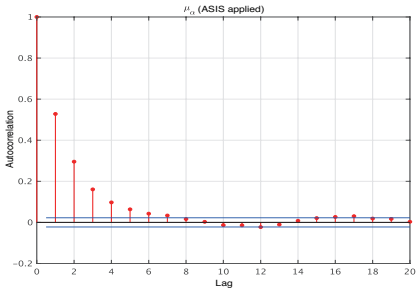


(h) σ_γ (ASIS applied)

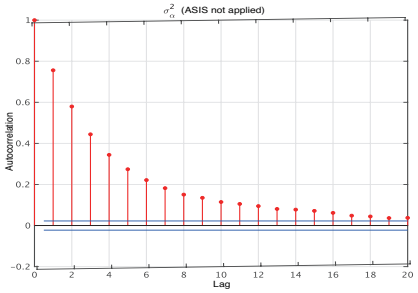
Figure 3.6: 10,000 sampling after 2,000 burn-ins



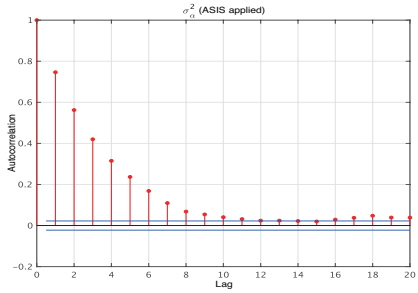
(a) μ_α (ASIS not applied)



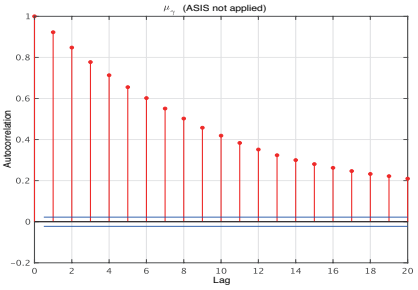
(b) μ_α (ASIS applied)



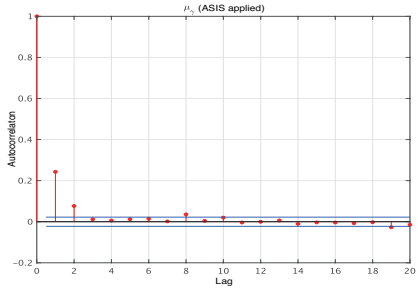
(c) σ_α (ASIS not applied)



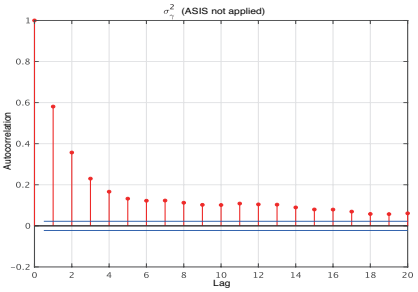
(d) σ_α (ASIS applied)



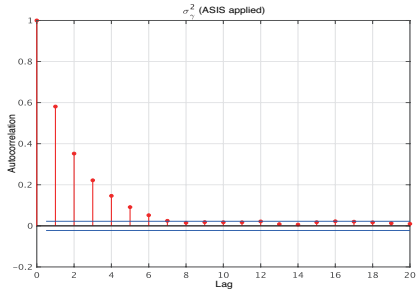
(e) μ_γ (ASIS not applied)



(f) μ_γ (ASIS applied)



(g) σ_γ (ASIS not applied)



(h) σ_γ (ASIS applied)

Figure 3.7: Autocorrelation of 10,000 sampling of hierarchical parameters

3.5 Conclusion

In this chapter, we conducted a hierarchical Bayesian analysis of racehorse running ability and jockey skills by using the running records of all 1800-m races held by JRA from 2016 to 2018 excluding obstacle courses. Average speed, calculated by dividing the distance of the course by the time of the race, is used as a quantitative measure of the racehorse's running ability. Three factors are considered to affect the racehorse speed during a race, which are as follows: the individual racehorse's ability, the environment in which the race is run, and the jockey's ability to control the horse.

In addition to the characteristics of the model that includes a large number of parameters as racehorse's and jockey's individual effects, the racehorse race data used in this study are unbalanced panel data in which the number of times each racehorse has run is not uniform; thus, we apply a hierarchical Bayesian analysis method following Silverman (2012). Conveniently, the conditional posterior distributions of all parameters in the model are typical (normal and inverse gamma distributions); thus, it is possible to generate random numbers from the posterior distributions using only Gibbs sampling. Furthermore, ASIS is applied to improve the efficiency of convergence of the random series. Moreover, WAIC is used to select combinations of explanatory variables in the regression model.

As a result of the analysis, the effects of horse weight, mare dummy, weight to carry, track index, race index, outside track dummy and turf track dummy are consistent with the expectation. The results for horse age and horse age squared indicate that the marginal effect of horse age on racehorse speed is positive until the racehorse reaches 4.18 years of age.

The application of ASIS also leads to a lower autocorrelation of the MCMC sample, which results in a more efficient sampling. The improvement in efficiency is so great that although the analysis of this study's model is impractical without ASIS, it is now possible to perform the analysis at a level that can be considered efficient in general.

Next, we consider the racehorse's individual effect, and find that the racehorses with the highest ability have the best total results, whereas the racehorses with the lowest ability do not have good total results. The difference in the effect between the best and

the worst racehorses is 2.65%, and considering that this data is obtained from an 1800-m race, the effect of the difference in the ability of the racehorses at the finish line causes a maximum difference of approximately 47.7 m in the running distance.

Finally, as for the jockey's individual effect, the posterior means of the top nine jockeys exceeded 0. Considering that the jockey used as the benchmark is Yutaka Take, well known as a distinguished jockey, all of these nine jockeys have extremely high ability. However, since the 95% credible interval of the jockey's individual effect for eight of the nine jockeys includes 0, there is no decisive difference in the ability between them and Yutaka Take. Therefore, we can conclude that Yutaka Take is as good as almost all the jockeys we compared in this study. Moreover, a comparison of the top and bottom jockeys revealed that the top jockeys were more experienced and had run more often than the bottom jockeys. The difference in ability between the highest and lowest ranked jockeys is 0.698%, and the difference in jockey ability at the finish line has a maximum effect of approximately 12.6 m on the distance run. In horse racing, the ability and condition of the horses tend to be the focus of attention, but the fact that the jockey's ability can overturn a large difference of approximately 12.6 m suggests that horse racing is a sophisticated sport in which human jockeys control a racehorse based on strategy and skill as a vehicle.

Chapter 4

Application to Time Series Data: High Frequency Stock Returns

4.1 Introduction

It is well documented that (a) probability distributions of stock returns are heavy-tailed (both tails of the probability density function go down to zero much slower than in the case of the normal distribution and as a result the kurtosis of the distribution exceeds three), (b) they are often asymmetric around the mean (the skewness of the distribution is either positive or negative), (c) they exhibit volatility clustering (positive autocorrelation among the day-to-day variance of returns) and (d) the leverage effect (the current volatility and the previous return are negatively correlated so that downturns in the stock market tend to predate sharper spikes in the volatility). In practice of financial risk management, it is imperative to develop a statistical model that can capture these characteristics of stock returns because they are thought to be related to steep drops and rebounds in stock prices during the periods of financial turmoil. Without factoring them in risk management, financial institutions might unintentionally take a higher risk and as a result would be faced with grave consequences, which we already observed during the Global Financial Crisis.

As a time series model with the aforementioned characteristics, a family of time series model called the stochastic volatility (SV) model has been developed in the field of financial econometrics. The standard SV model is a simple state-space model in which

the measurement equation is a mere distribution of stock returns with the time-varying variance (volatility) and the system equation is an AR(1) process of the latent log volatility. In the standard setting, both measurement and system errors are supposed to be Gaussian and negatively correlated in order to incorporate the leverage effect into the model. The standard SV model can explain three stylized facts: heavy-tailed distribution, volatility clustering and leverage effect, but it cannot make the distribution of stock returns asymmetric. Furthermore, although in theory the standard SV model incorporates the heavy-tail behavior of stock returns, many empirical studies demonstrated that it was insufficient to explain extreme fluctuations of stock prices that were caused by large shocks in financial markets.

Based on this plain-vanilla SV model, researchers has developed numerous variants that are designed to capture all aspects of stock returns sufficiently well. A direct way to introduce a more heavy-tailed distribution to the SV model is to assume that the error term of the measurement equation follows a distribution with much heavier tails than the normal distribution. The Student's t distribution is a popular choice (Berg *et al.* (2004), Omori *et al.* (2007), Nakajima and Omori (2009), Nakajima (2012) among others). In the literature, the asymmetry in stock returns can be handled by assuming that the error term follows an asymmetric distribution (Nakajima and Omori (2012), Tsiotas (2012), Abanto-Valle *et al.* (2015) among others). In particular, the generalized hyperbolic (GH) distribution proposed by Barndorff-Nielsen (1977) has recently drawn increasing attention among researchers (e.g., Nakajima and Omori (2012)) since it is regarded as a broad family of heavy-tailed distributions such as variance-gamma and Student's t as well as their skewed variants such as skew variance-gamma and skew Student's t.

Traditionally, empirical studies with the SV model used time series data of daily stock returns. However, the availability of high frequency tick data and the advent of high-frequency trading (HFT), which is a general term for algorithmic trading in full use of high performance computing and high speed communication technology, has shifted the focus of research on volatility from closing-to-closing daily volatility to intraday volatility in a very short interval (e.g., 5 minutes or shorter). This shift paved the way for a new type of SV model. In addition to the traditional stylized facts on daily volatility, intraday volatility is known to exhibit a cyclical pattern during the trading hours. On a

typical trading day, the volatility tends to be high immediately after the market opens, but it gradually declines in the middle of the trading hours. In the late trading hours, the volatility again becomes higher as it nears the closing time. This U-shaped trend in volatility is called intraday seasonality in the literature (see Chan *et al.* (1991) among others). Although it is crucial to take the intraday seasonality into consideration in estimation of any intraday volatility models, only a few studies (e.g., Stroud and Johannes (2014), Fičula and Witzany (2015a, 2015b)) explicitly incorporate it into their volatility models.

In this chapter, we propose to directly embed intraday seasonality into the SV model by approximating the U-shaped seasonality pattern with a linear combination of Bernstein polynomials. In order to capture skewness and excess kurtosis in high frequency stock returns, we employ two distributions (variance-gamma and Student's t) and their skewed variants (skew variance-gamma and skew Student's t) in the family of GH distributions as the distribution of stock returns in the SV model. Since the proposed SV model is intractably complicated, we develop an efficient Markov chain Monte Carlo (MCMC) sampling algorithm for full Bayesian estimation of all parameters and state variables (latent log volatilities in our case) in the model.

The rest of this chapter is organized as follows. In Section 4.2, we introduce a reparameterized Gaussian SV model with leverage and intraday seasonality and derive an efficient MCMC sampling algorithm for its Bayesian estimation. In Section 4.3, we show the conditional posterior distributions and prepare for application of ancillarity-sufficiency interweaving strategy (ASIS) proposed by Yu and Meng (2011). In Section 4.4, we extend the Gaussian SV model to the case of variance gamma and Student's t error as well as their skewed variants. In Section 4.5, we report the estimation results of our proposed SV models with 1-minute return data of TOPIX. Finally, conclusion is given in Section 4.6.

4.2 Stochastic Volatility Model with Intraday Seasonality

Consider the log difference of a stock price in a short interval (say, 1 or 5 minutes). We divide trading hours evenly into T periods and normalize them so that the length of the trading hours is equal to 1, that is, the length of each period is $\frac{1}{T}$ and the time stamp of the t-th period is $\frac{t}{T}$ $(t = 1, \ldots, T)$. Note that the market opens at time 0 and closes at time 1 in our setup. Let y_t $(t = 1, \ldots, T)$ denote the stock return in the t-th period (at time $\frac{t}{T}$ in the trading hours) and consider the following stochastic volatility (SV) model of y_t with intraday seasonality:

$$\begin{cases} y_t = \exp(x_t'\beta + h_t)\epsilon_t, \\ h_{t+1} = \phi h_t + \eta_t, \end{cases} \qquad \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\tau \\ \rho\tau & \tau^2 \end{bmatrix} \right), \quad |\rho| < 1, \quad \tau > 0, \quad (4.1)$$

and

$$h_1 \sim \mathcal{N}\left(0, \frac{\tau^2}{1-\phi^2}\right), \quad |\phi| < 1.$$

It is well known that the estimate of the correlation coefficient ρ is negative in most stock markets. This negative correlation is often referred to as the leverage effect. Note that the stock volatility in the t-th period (the natural logarithm of the conditional standard deviation of y_t) is

$$\log \sqrt{\mathrm{Var}[y_t|\mathcal{F}_{t-1}]} = x_t'\beta + h_t,$$

where $\mathcal{F}_{t-1}$ is the filtration that represents all available information at time $\frac{t-1}{T}$. Hence the stock volatility in the SV model (4.1) is decomposed into two parts: a linear combination of covariates $x_t'\beta$ and the unobserved AR(1) process h_t. In this study, we regard $x_t'\beta$ as the intraday seasonal component of the stock volatility, though it can be interpreted as any function of covariates x_t in a different situation. On the other hand, h_t is supposed to capture volatility clustering. We call h_t the *latent log volatility* since it is unobservable.

Although the intraday seasonal component $x_t'\beta$ is likely to be a U-shaped function of time stamps (the stock volatility is higher right after the opening or near the closing, but it is lower in the middle of the trading hours), we have no information about the exact

functional form of the intraday seasonality. To make it in a flexible functional form for the intraday seasonality, we assume that $x_t'\beta$ is a Bernstein polynomial:

$$x_t'\beta = \sum_{k=0}^{n} \beta_k x_{k,t} = \sum_{k=0}^{n} \beta_k b_{k,n}\left(\frac{t}{T}\right), \tag{4.2}$$

where $b_{k,n}(\cdot)$ is called a Bernstein basis polynomial of degree n:

$$b_{k,n}(v) = {}_nC_k v^k (1-v)^{n-k}, \quad k = 0, \ldots, n, \quad v \in [0,1].$$

According to the Weierstrass approximation theorem, the Bernstein polynomial (4.2) can approximate any continuous function on $[0,1]$ as n goes to infinity. In practice, however, the number of observations T is finite. Thus we need to choose a finite n via a model selection procedure. We will discuss this issue in Section 4.4.

Although the parametrization of the SV model in (4.1) is widely applied in the literature, we propose an alternative parametrization that facilitates MCMC implementation in non-Gaussian SV models. By replacing the covariance matrix in (4.1) with

$$\begin{bmatrix} \mathrm{Var}[\epsilon_t] & \mathrm{Cov}[\eta_t, \epsilon_t] \\ \mathrm{Cov}[\epsilon_t, \eta_t] & \mathrm{Var}[\eta_t] \end{bmatrix} = \begin{bmatrix} 1+\gamma^2\tau^2 & \gamma\tau^2 \\ \gamma\tau^2 & \tau^2 \end{bmatrix}, \quad \gamma \in \mathbb{R}, \tag{4.3}$$

we obtain an alternative formulation of the SV model:

$$\begin{cases} y_t = \exp(x_t'\beta + h_t)\epsilon_t, & \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1+\gamma^2\tau^2 & \gamma\tau^2 \\ \gamma\tau^2 & \tau^2 \end{bmatrix} \right). \\ h_{t+1} = \phi h_t + \eta_t, & \end{cases} \tag{4.4}$$

Since in (4.4) the variance of ϵ_t is no longer equal to one, the interpretation of β and h_t in (4.4) is slightly different from the original one in (4.1). Nonetheless the SV model (4.4) has essentially the same characteristics as (4.1). Since the correlation coefficient in (4.3) is

$$\mathrm{Corr}[\epsilon_t, \eta_t] = \frac{\gamma\tau}{\sqrt{1+\gamma^2\tau^2}},$$

the sign of γ always coincides with the correlation coefficient and the leverage effect exists if $\gamma < 0$. To distinguish γ in (4.4) from the correlation parameter ρ in (4.1), we call γ the *leverage parameter* in this study.

Note that the inverse of (4.3) is

$$\begin{bmatrix} \mathrm{Var}[\epsilon_t] & \mathrm{Cov}[\eta_t, \epsilon_t] \\ \mathrm{Cov}[\epsilon_t, \eta_t] & \mathrm{Var}[\eta_t] \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -\gamma \\ -\gamma & \gamma^2 + \tau^{-2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\gamma & \tau^{-1} \end{bmatrix} \begin{bmatrix} 1 & -\gamma \\ 0 & \tau^{-1} \end{bmatrix},$$

and the determinant of (4.3) is τ^2. Using

$$\begin{bmatrix}\epsilon_t & \eta_t\end{bmatrix}\begin{bmatrix}1 & -\gamma \\ -\gamma & \gamma^2+\tau^{-2}\end{bmatrix}\begin{bmatrix}\epsilon_t \\ \eta_t\end{bmatrix} = \begin{bmatrix}\epsilon_t & \eta_t\end{bmatrix}\begin{bmatrix}1 & 0 \\ -\gamma & \tau^{-1}\end{bmatrix}\begin{bmatrix}1 & -\gamma \\ 0 & \tau^{-1}\end{bmatrix}\begin{bmatrix}\epsilon_t \\ \eta_t\end{bmatrix}$$
$$= (\epsilon_t - \gamma\eta_t)^2 + \frac{\eta_t^2}{\tau^2},$$

we can easily show that the SV model (4.4) is equivalent to

$$\begin{cases} y_t = \exp(x_t'\beta + h_t)(z_t + \gamma\eta_t), \\ h_{t+1} = \phi h_t + \eta_t, \end{cases} \tag{4.5}$$

where

$$z_t \sim \mathcal{N}(0,1),\ \eta_t \sim \mathcal{N}(0,\tau^2),\ z_t \perp \eta_t.$$

In the alternative formulation of the SV model (4.5), we can interpret η_t as a common shock that affects both the stock return y_t and the log volatility h_{t+1} while z_t as an idiosyncratic shock that affects y_t only.

The likelihood for the SV model (4.5) given the observations $y_{1:T} = [y_1; \ldots; y_T]$ and the latent log volatility $h_{1:T+1} = [h_1; \ldots; h_{T+1}]$ is

$$p(y_{1:T}, h_{1:T+1}|\theta) = \underbrace{\prod_{t=1}^{T} p(y_t|h_t, h_{t+1}, \theta)}_{p(y_{1:T}|h_{1:T+1},\theta)} \cdot \underbrace{p(h_1|\theta)\prod_{t=1}^{T} p(h_{t+1}|h_t,\theta)}_{p(h_{1:T+1}|\theta)}, \tag{4.6}$$

where

$$p(y_t|h_t, h_{t+1}, \theta) = \frac{1}{\sqrt{2\pi}}\exp\left[-x_t'\beta - h_t - \frac{\{y_t\exp(-x_t'\beta - h_t) - \gamma(h_{t+1} - \phi h_t)\}^2}{2}\right], \tag{4.7}$$

$$p(h_{t+1}|h_t,\theta) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left[-\frac{(h_{t+1}-\phi h_t)^2}{2\tau^2}\right], \quad t = 1,\ldots,T,$$
$$p(h_1|\theta) = \sqrt{\frac{1-\phi^2}{2\pi\tau^2}}\exp\left[-\frac{(1-\phi^2)h_1^2}{2\tau^2}\right],$$

and $\theta = (\beta, \gamma, \tau^2, \phi)$. Since h_t follows a stationary AR(1) process, the joint probability

distribution of $h_{1:T+1}$ is $\mathcal{N}(0, \tau^2 V^{-1})$ where

$$V = \begin{bmatrix} 1 & -\phi & & & & & \\ -\phi & 1+\phi^2 & -\phi & & & & \\ & -\phi & 1+\phi^2 & -\phi & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -\phi & 1+\phi^2 & -\phi & \\ & & & & -\phi & 1+\phi^2 & -\phi \\ & & & & & -\phi & 1 \end{bmatrix}, \tag{4.8}$$

is a tridiagonal matrix, and it is positive definite as long as $|\phi| < 1$. Thus the joint p.d.f. of $h_{1:T+1}$ is

$$p(h_{1:T+1}|\theta) = (2\pi\tau^2)^{-\frac{T+1}{2}} |V|^{\frac{1}{2}} \exp\left[-\frac{1}{2\tau^2} h'_{1:T+1} V h_{1:T+1}\right], \quad |V| = 1 - \phi^2. \tag{4.9}$$

The prior distributions for $(\beta, \gamma, \tau^2, \phi)$ in our study are

$$\begin{aligned} &\beta \sim \mathcal{N}(\bar{\mu}_\beta, \bar{\Omega}_\beta^{-1}), \quad \gamma \sim \mathcal{N}(\bar{\mu}_\gamma, \bar{\omega}_\gamma^{-1}), \\ &\tau^2 \sim \mathcal{IG}(a_\tau, b_\tau), \quad \frac{\phi+1}{2} \sim \mathcal{B}(a_\phi, b_\phi), \end{aligned} \tag{4.10}$$

where $\mathcal{B}(a, b)$ is the beta distribution with the probability density:

$$p(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \quad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx.$$

Then the joint posterior density of $(h_{1:T+1}, \theta)$ for the SV model (4.5) is

$$p(h_{1:T+1}, \theta|y_{1:T}) \propto \prod_{t=1}^{T} p(y_t|h_t, h_{t+1}, \theta) \cdot p(h_{1:T+1}|\theta) \cdot p(\theta), \tag{4.11}$$

where $p(\theta)$ is the prior density of the parameters in (4.10).

Since analytical evaluation of the joint posterior distribution (4.11) is impractical, we apply an MCMC method to generating a random sample $\{(h_{1:T+1}^{(r)}, \beta^{(r)}, \gamma^{(r)}, \tau^{2(r)}, \phi^{(r)})\}_{r=1}^{R}$ from the joint posterior distribution (4.11), and numerically evaluate the posterior statistics necessary for Bayesian inference with Monte Carlo integration. The outline of the standard MCMC sampling scheme for the posterior distribution (4.11) is given as follows:

Outline of the MCMC sampling for the SV model

Step 0: Initialize $(h_{1:T+1}^{(0)}, \beta^{(0)}, \gamma^{(0)}, \tau^{2(0)}, \phi^{(0)})$ and set the counter $r = 0$.

Step 1: Generate $(h_{1:T+1}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \tau^{2(r+1)}, \phi^{(r+1)})$ with the following scheme:

Step 1-1: Generate $h_{1:T+1}^{(r+1)}$ from $p(h_{1:T+1}|\beta^{(r)}, \gamma^{(r)}, \tau^{2(r)}, \phi^{(r)}, y_{1:T})$.

Step 1-2: Generate $\beta^{(r+1)}$ from $p(\beta|h_{1:T+1}^{(r+1)}, \gamma^{(r)}, \tau^{2(r)}, \phi^{(r)}, y_{1:T})$.

Step 1-3: Generate $\gamma^{(r+1)}$ from $p(\gamma|h_{1:T+1}^{(r+1)}, \beta^{(r+1)}, \tau^{2(r)}, \phi^{(r)}, y_{1:T})$

Step 1-4: Generate $\tau^{2(r+1)}$ from $p(\tau|h_{1:T+1}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \phi^{(r)}, y_{1:T})$

Step 1-5: Generate $\phi^{(r+1)}$ from $p(\phi|h_{1:T+1}^{(r+1)}, \beta^{(r+1)}, \gamma^{(r+1)}, \tau^{2(r+1)}, y_{1:T})$.

Step 2: Let $r = r + 1$ and go to **Step 1** until the burn-in iterations are completed.

Step 3: Reset the counter $r = 0$ and repeat **Step 1**–**2** R times in order to obtain the Monte Carlo sample $\{(h_{1:T+1}^{(r)}, \beta^{(r)}, \gamma^{(r)}, \tau^{2(r)}, \phi^{(r)})\}_{r=1}^{R}$.

Although the above MCMC sampling scheme is ubiquitous in the literature of the SV model, the generated Monte Carlo sample $\{(h_{1:T+1}^{(r)}, \beta^{(r)}, \gamma^{(r)}, \tau^{2(r)}, \phi^{(r)})\}_{r=1}^{R}$ tends to exhibit strongly positive autocorrelation. To improve efficiency of MCMC implementation, Yu and Meng (2011) proposed an ancillarity-sufficiency interweaving strategy (ASIS). In the literature of the SV model, Kastner and Frühwirth-Schnatter (2014) applied ASIS to the SV model of daily US-dollar/Euro exchange rate data with the Gaussian error. Their SV model did not include either intraday seasonality or leverage effect since they applied it to daily exchange rate data that exhibited no leverage effect in most cases. We extend the algorithm developed by Kastner and Frühwirth-Schnatter (2014) to facilitate the converge of the sample path in the SV model (4.5). The basic principle of ASIS is to construct MCMC sampling schemes for two different but equivalent parametrizations of a model with missing/latent variables ($h_{1:T+1}$ in our case) and generate the parameters alternately with each of them.

According to Kastner and Frühwirth-Schnatter (2014), the SV model (4.5) is in a non-centered parametrization (NCP). On the other hand, we may transform h_t as

$$\tilde{h}_t = x_t'\beta + h_t, \tag{4.12}$$

and rearrange the SV model (4.5) as

$$\begin{cases} y_t = \exp(\tilde{h}_t)(z_t + \gamma\eta_t), \\ \tilde{h}_{t+1} - x'_{t+1}\beta = \phi(\tilde{h}_t - x'_t\beta) + \eta_t. \end{cases} \tag{4.13}$$

The above SV model (4.13) is in a centered parametrization (CP).

The posterior distribution in the CP form (4.13) is equivalent to the one in the NCP form (4.5) in the sense that they give us the same posterior distribution of θ. Let us verify this claim. The likelihood for the SV model (4.13) given the observations $y_{1:T}$ and the latent log volatility $\tilde{h}_{1:T+1} = [\tilde{h}_1; \ldots; \tilde{h}_{T+1}]$ is

$$p(y_{1:T}, \tilde{h}_{1:T+1}|\theta) = \underbrace{\prod_{t=1}^{T} p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \theta)}_{p(y_{1:T}|\tilde{h}_{1:T+1},\theta)} \cdot \underbrace{p(\tilde{h}_1|\theta) \prod_{t=1}^{T} p(\tilde{h}_{t+1}|\tilde{h}_t, \theta)}_{p(\tilde{h}_{1:T+1}|\theta)}, \tag{4.14}$$

where

$$p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \theta) = \frac{1}{\sqrt{2\pi}} \exp\left[-\tilde{h}_t - \frac{\left\{ y_t e^{-\tilde{h}_t} - \gamma((\tilde{h}_{t+1} - x'_{t+1}\beta) - \phi(\tilde{h}_t - x'_t\beta)) \right\}^2}{2} \right], \tag{4.15}$$

$$p(\tilde{h}_{t+1}|\tilde{h}_t, \theta) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{\{(\tilde{h}_{t+1} - x'_{t+1}\beta) - \phi(\tilde{h}_t - x'_t\beta)\}^2}{2\tau^2} \right], \quad t = 1, \ldots, T,$$

$$p(\tilde{h}_1|\theta) = \sqrt{\frac{1-\phi^2}{2\pi\tau^2}} \exp\left[-\frac{(1-\phi^2)(\tilde{h}_1 - x'_1\beta)^2}{2\tau^2} \right].$$

Note that the joint p.d.f. of $\tilde{h}_{1:T+1}$ is

$$p(\tilde{h}_{1:T+1}|\theta) = (2\pi\tau^2)^{-\frac{T+1}{2}} |V|^{\frac{1}{2}} \exp\left[-\frac{1}{2\tau^2} (\tilde{h}_{1:T+1} - X\beta)' V (\tilde{h}_{1:T+1} - X\beta) \right], \tag{4.16}$$

where $X = [x'_1; \ldots; x'_{T+1}]$. With the prior of θ in (4.10), the joint posterior density of $(\tilde{h}_{1:T+1}, \theta)$ for the SV model (4.13) is obtained as

$$p(\tilde{h}_{1:T+1}, \theta|y_{1:T}) \propto \prod_{t=1}^{T} p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \theta) \cdot p(\tilde{h}_{1:T+1}|\theta) \cdot p(\theta). \tag{4.17}$$

Note that θ is unchanged between the NCP form (4.11) and the CP form (4.17). Although the latent variables are transformed with (4.12), the "marginal" posterior p.d.f. of θ is unchanged because

$$\int p(\tilde{h}_{1:T+1}, \theta|y_{1:T+1}) d\tilde{h}_{1:T+1} = \int p(h_{1:T+1}, \theta|y_{1:T+1}) |J| dh_{1:T+1},$$

where the Jacobian $|J| = 1$.

With this fact in mind, we can incorporate ASIS into the MCMC sampling scheme by replacing **Step 1** with

NCP-based ASIS step

Step 1: Generate $(h_{1:T+1}^{(r+0.5)}, \beta^{(r+0.5)}, \gamma^{(r+0.5)}, \tau^{2(r+0.5)}, \phi^{(r+0.5)})$ with the sampling scheme based on the NCP form (4.5) and compute

$$\tilde{h}_t^{(r+0.5)} = h_t^{(r+0.5)} + x_t'\beta^{(r+0.5)}, \quad t = 1, \ldots, T+1.$$

Step 1.5: Generate $(\beta^{(r+1)}, \gamma^{(r+1)}, \tau^{2(r+1)}, \phi^{(r+1)})$ with the sampling scheme based on the CP form (4.13) and compute

$$h_t^{(r+1)} = \tilde{h}_t^{(r+0.5)} - x_t'\beta^{(r+1)}, \quad t = 1, \ldots, T+1.$$

Note that we generate a new latent log volatility $h_{1:T+1}$ from its conditional posterior distribution in the NCP form (4.11) only once at the beginning of **Step 1**. This is the reason why we call it the NCP-based ASIS step. After this update, we merely shift the location of $h_{1:T+1}$ by $x_t'\beta^{(r+0.5)}$ (**Step 1**) or by $-x_t'\beta^{(r+1)}$ (**Step 1.5**). In ASIS, these shifts are applied with probability 1 even if all elements in $h_{1:T+1}$ are not updated at the beginning of **Step 1**, which is highly probable in practice because we need to use the MH algorithm to generate $h_{1:T+1}$. Although we also utilize the MH algorithm to generate β as explained later, the acceptance rate of β in the MH step is much higher than that of $h_{1:T+1}$ in our experience. Thus we expect that both $x_t'\beta^{(r+0.5)}$ and $-x_t'\beta^{(r+1)}$ will be updated more often than $h_{1:T+1}$ itself. As a result, the above ASIS step may improve mixing of the sample sequence of $h_{1:T+1}$. Conversely, we may apply the following CP-based ASIS step:

CP-based ASIS step

Step 1: Generate $(\tilde{h}_{1:T+1}^{(r+0.5)}, \beta^{(r+0.5)}, \gamma^{(r+0.5)}, \tau^{2(r+0.5)}, \phi^{(r+0.5)})$ with the sampling scheme based on the CP form (4.13) and compute

$$h_t^{(r+0.5)} = \tilde{h}_t^{(r+0.5)} - x_t'\beta^{(r+0.5)}, \quad t = 1, \ldots, T+1.$$

Step 1.5: Generate $(\beta^{(r+1)}, \gamma^{(r+1)}, \tau^{2(r+1)}, \phi^{(r+1)})$ with the sampling scheme based on the NCP form (4.5) and compute

$$\tilde{h}_t^{(r+1)} = h_t^{(r+0.5)} + x_t'\beta^{(r+1)}, \quad t = 1, \ldots, T+1.$$

In the CP-based ASIS step, we generate $\tilde{h}_{1:T+1}$ from its conditional posterior distribution in the CP form (4.17) once. The rest is the same as in the NCP-based ASIS step except that the order of sampling is reversed.

4.3 Conditional Posterior Distributions

In this section, we derive the conditional posterior distribution of the latent log volatility and that of each parameter in the SV model for both NCP and CP.

4.3.1 NCP Form

Latent Log Volatility $h_{1:T+1}$

The conditional posterior density of the latent log volatility $h_{1:T+1}$ is

$$p(h_{1:T+1}|\theta, y_{1:T}) \propto \prod_{t=1}^{T} p(y_t|h_t, h_{t+1}, \theta) \cdot p(h_{1:T+1}|\theta). \tag{4.18}$$

We apply the Metropolis- Hastings (MH) algorithm to generate $h_{1:T+1}$ from (4.18). To derive a suitable proposal distribution for the MH algorithm, we first consider consider the second-order Taylor approximation of $\ell(h_{1:T+1}) = \log p(y_{1:T}|h_{1:T+1}, \theta)$ in the neighborhood of $h_{1:T+1}^*$:

$$\begin{aligned}\ell(h_{1:T+1}) \approx \ell(h_{1:T+1}^*) &+ g(h_{1:T+1}^*)'(h_{1:T+1} - h_{1:T+1}^*) \\ &- \frac{1}{2}(h_{1:T+1} - h_{1:T+1}^*)'Q(h_{1:T+1}^*)(h_{1:T+1} - h_{1:T+1}^*),\end{aligned} \tag{4.19}$$

where $g(h_{1:T+1})$ is the gradient vector of $\ell(h_{1:T+1})$:

$$g(h_{1:T+1}) = \begin{bmatrix} g_1(h_{1:T+1}) \\ \vdots \\ g_t(h_{1:T+1}) \\ \vdots \\ g_T(h_{1:T+1}) \\ g_{T+1}(h_{1:T+1}) \end{bmatrix} = \begin{bmatrix} \nabla_1 \log p(y_{1:T}|h_{1:T+1}, \theta) \\ \vdots \\ \nabla_t \log p(y_{1:T}|h_{1:T+1}, \theta) \\ \vdots \\ \nabla_T \log p(y_{1:T}|h_{1:T+1}, \theta) \\ \nabla_{T+1} \log p(y_{1:T}|h_{1:T+1}, \theta) \end{bmatrix},$$

and $Q(h_{1:T+1})$ is the Hessian matrix of $\log p(y_{1:T}|h_{1:T+1}, \theta)$ times -1:

$$Q(h_{1:T+1}) = \begin{bmatrix} q_{11}(h_{1:T+1}) & q_{12}(h_{1:T+1}) & & \cdots & 0 \\ q_{21}(h_{1:T+1}) & q_{22}(h_{1:T+1}) & q_{23}(h_{1:T+1}) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & q_{T,T-1}(h_{1:T+1}) & q_{T,T}(h_{1:T+1}) & q_{T,T+1}(h_{1:T+1}) \\ 0 & \cdots & & q_{T+1,T}(h_{1:T+1}) & q_{T+1,T+1}(h_{1:T+1}) \end{bmatrix}.$$

which is a $(T+1) \times (T+1)$ band matrix.

Let us derive the explicit form of each element in $g(h_{1:T+1})$ and $Q(h_{1:T+1})$. By defining

$$\epsilon_t = y_t \exp(-x_t'\beta - h_t), \quad \eta_t = h_{t+1} - \phi h_t, \tag{4.20}$$

the log density of y_t (4.7) is rewritten as

$$\log p(y_t|h_t, h_{t+1}, \theta) = -x_t'\beta - h_t - \frac{1}{2}(\epsilon_t - \gamma\eta_t)^2 + \text{constant}.$$

Note that

$$\nabla_t \epsilon_t = -\epsilon_t, \quad \nabla_t \eta_t = -\phi, \quad \nabla_t \eta_{t-1} = 1.$$

where $\nabla_t = \frac{\partial}{\partial h_t}$. Each element of $g(h_{1:T+1})$ is derived as

$$\begin{aligned} g_t(h_{1:T+1}) &= \nabla_t \log p(y_{1:T}|h_{1:T+1}, \theta) = \nabla_t \log p(y_t|h_t, h_{t+1}, \theta) + \nabla_t \log p(y_{t-1}|h_{t-1}, h_t, \theta) \\ &= -1 - (\epsilon_t - \gamma\eta_t)(-\epsilon_t - \gamma(-\phi)) - (\epsilon_{t-1} - \gamma\eta_{t-1})(-\gamma) \\ &= -1 + (\epsilon_t - \gamma\eta_t)(\epsilon_t - \gamma\phi) + \gamma(\epsilon_{t-1} - \gamma\eta_{t-1}), \end{aligned}$$

for $t = 2, \ldots, T$,

$$\begin{aligned} g_1(h_{1:T+1}) &= \nabla_1 \log p(y_{1:T}|h_{1:T+1}, \theta) = \nabla_1 \log p(y_1|h_1, h_2, \theta) \\ &= -1 - (\epsilon_1 - \gamma\eta_1)(-\epsilon_1 - \gamma(-\phi)) \\ &= -1 + (\epsilon_1 - \gamma\eta_1)(\epsilon_1 - \gamma\phi), \end{aligned}$$

for $t = 1$, and

$$\begin{aligned} g_{T+1}(h_{1:T+1}) &= \nabla_{T+1} \log p(y_{1:T}|h_{1:T+1}, \theta) = \nabla_{T+1} \log p(y_T|h_T, h_{T+1}, \theta) \\ &= -(\epsilon_T - \gamma\eta_T)(-\gamma) = \gamma(\epsilon_T - \gamma\eta_T), \end{aligned}$$

for $t = T + 1$. The diagonal element in $Q(h_{1:T+1})$ is given as

$$\begin{aligned} q_{t,t}(h_{1:T+1}) &= (-1) \times \nabla_t^2 \log p(y_{1:T}|h_{1:T+1}, \theta) \\ &= -(-\epsilon_t - \gamma(-\phi))(\epsilon_t - \gamma\phi) - (\epsilon_t - \gamma\eta_t)(-\epsilon_t) - \gamma(-\gamma) \\ &= (\epsilon_t - \gamma\phi)^2 + \epsilon_t(\epsilon_t - \gamma\eta_t) + \gamma^2, \end{aligned}$$

for $t = 2, \ldots, T$,

$$\begin{aligned} q_{11}(h_{1:T+1}) &= (-1) \times \nabla_1^2 \log p(y_{1:T}|h_{1:T+1}, \theta) \\ &= -(-\epsilon_1 - \gamma(-\phi))(\epsilon_1 - \gamma\phi) - (\epsilon_1 - \gamma\eta_1)(-\epsilon_1) \\ &= (\epsilon_1 - \gamma\phi)^2 + \epsilon_1(\epsilon_1 - \gamma\eta_1), \end{aligned}$$

for $t = 1$, and

$$\begin{aligned} q_{T+1,T+1}(h_{1:T+1}) &= (-1) \times \nabla_{T+1}^2 \log p(y_{1:T}|h_{1:T+1}, \theta) \\ &= -\gamma(-\gamma) = \gamma^2, \end{aligned}$$

for $t = T + 1$. Furhtermore the first off-diagonal element of $Q(h_{1:T+1})$ is derived as

$$\begin{aligned} q_{t,t+1}(h_{1:T+1}) &= (-1) \times \nabla_{t,t+1} \log p(y_{1:T}|h_{1:T+1}, \theta) \\ &= -(-\gamma)(\epsilon_t - \gamma\phi) \\ &= \gamma(\epsilon_t - \gamma\phi), \end{aligned}$$

for $t = 1, \ldots, T$. In summary,

$$g_t(h_{1:T+1}) = \{-1 + (\epsilon_t - \gamma\eta_t)(\epsilon_t - \gamma\phi)\}\mathbf{1}(t \leqq T) + \gamma(\epsilon_{t-1} - \gamma\eta_{t-1})\mathbf{1}(t \geqq 2), \quad (4.21)$$

$$q_t(h_{1:T+1}) = \{(\epsilon_t - \gamma\phi)^2 + \epsilon_t(\epsilon_t - \gamma\eta_t)\}\mathbf{1}(t \leqq T) + \gamma^2\mathbf{1}(t \geqq 2), \quad (4.22)$$

$$q_{t,t+1}(h_{1:T+1}) = \gamma(\epsilon_t - \gamma\phi). \quad (4.23)$$

Since the log prior density of $h_{1:T+1}$ is

$$\bar{p}(h_{1:T+1}) = -\frac{T+1}{2}\log(2\pi\tau^2) + \frac{1}{2}\log|V| - \frac{1}{2\tau^2}h'_{1:T+1}Vh_{1:T+1}, \quad (4.24)$$

the conditional posterior density of $h_{1:T+1}$ (4.18) can be approximated by

$$\begin{aligned}
&p(h_{1:T+1}|\theta, y_{1:T}) \\
&= C \exp\left[\ell(h_{1:T+1}) + \bar{p}(h_{1:T+1})\right] \\
&\approx C \exp\Bigg[\ell(h^*_{1:T+1}) + g(h^*_{1:T+1})'(h_{1:T+1} - h^*_{1:T+1}) \\
&\qquad - \frac{1}{2}(h_{1:T+1} - h^*_{1:T+1})' Q(h^*_{1:T+1})(h_{1:T+1} - h^*_{1:T+1}) + \bar{p}(h_{1:T+1})\Bigg] \\
&= C \exp\left[\ell(h^*_{1:T+1}) - \frac{T+1}{2}\log(2\pi\tau^2) + \frac{1}{2}\log|V| + f(h_{1:T+1})\right], \qquad (4.25)
\end{aligned}$$

where C is the normalizing constant of the conditional posterior density and

$$\begin{aligned}
f(h_{1:T+1}) = {} & g(h^*_{1:T+1})'(h_{1:T+1} - h^*_{1:T+1}) \\
& - \frac{1}{2}(h_{1:T+1} - h^*_{1:T+1})' Q(h^*_{1:T+1})(h_{1:T+1} - h^*_{1:T+1}) \\
& - \frac{1}{2\tau^2} h'_{1:T+1} V h_{1:T+1}, \qquad (4.26)
\end{aligned}$$

By completing the square in (4.26), we have

$$\begin{aligned}
f(h_{1:T+1}) = {} & -\frac{1}{2}\left(h_{1:T+1} - \mu_h(h^*_{1:T+1})\right)' \Sigma_h(h^*_{1:T+1})^{-1}\left(h_{1:T+1} - \mu_h(h^*_{1:T+1})\right) \qquad (4.27) \\
& + \text{constant},
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_h(h^*_{1:T+1}) &= \left(Q(h^*_{1:T+1}) + \frac{1}{\tau^2}V\right)^{-1}, \\
\mu_h(h^*_{1:T+1}) &= \Sigma_h(h^*_{1:T+1})\left(g(h^*_{1:T+1}) + Q(h^*_{1:T+1})h^*_{1:T+1}\right).
\end{aligned}$$

Therefore the right-hand side of (4.25) is approximately proportional to the pdf of the following normal distribution:

$$h_{1:T+1} \sim \mathcal{N}\left(\mu_h(h^*_{1:T+1}), \Sigma_h(h^*_{1:T+1})\right). \qquad (4.28)$$

Recall that both $Q(h^*_{1:T+1})$ and V are tridiagonal matrices. Thus $\Sigma_h(h^*_{1:T+1})^{-1} = Q(h^*_{1:T+1}) + \frac{1}{\tau^2}V$ is also tridiagonal. Since the Cholesky decomposition of a tridiagonal matrix and the inverse of a triangular matrix can be efficiently computed if they exist, $h_{1:T+1}$ is readily generated from (4.28) with

$$h_{1:T+1} = (L')^{-1}\left(L^{-1}\left(g(h^*_{1:T+1}) + Q(h^*_{1:T+1})h^*_{1:T+1}\right) + \tilde{z}\right), \quad \tilde{z} \sim \mathcal{N}(0, \mathrm{I}),$$

where L is a lower triangular matrix obtained by the Cholesky decomposition as

$$L'L = Q(h^*_{1:T+1}) + \frac{1}{\tau^2}V.$$

The above algorithm, which is called the *all without a loop* (AWOL) in Kastner and Frühwirth-Schnatter (2014), has been applied to Gaussian Markov random fields (e.g., Rue (2001)) and state-space models (e.g., Chan and Jeliazkov (2009), McCausland *et al.* (2011)).

Hoping that the approximation (4.25) is sufficiently accurate, we use (4.28) as the proposal distribution in the MH algorithm. In practice, however, we need to address two issues:

1. the choice of $h^*_{1:T+1}$ is crucial to make the approximation (4.25) workable.

2. the acceptance rate of the MH algorithm tends to be too low when $h_{1:T+1}$ is a high-dimensional vector.

We address the former issue by using the mode of the conditional posterior density as $h^*_{1:T+1}$. The search of the mode is performed by the following recursion:

Step 1: Initialize $h^{*(0)}_{1:T+1}$ and set the counter $r = 1$.

Step 2: Update $h^{*(r)}_{1:T+1}$ by $h^{*(r)}_{1:T+1} = \mu_h(h^{*(r-1)}_{1:T+1})$.

Step 3: Let $r = r + 1$ and go to **Step 2** unless $\max_{t=1,\ldots,T+1} |h^{*(r)}_t - h^{*(r-1)}_t|$ is less than the preset tolerance level.

In our experience, it mostly attains convergence in a few iterations.

We address the latter issue by apply a so-called block sampler. In the block sampler, we randomly partition $h_{1:T+1}$ into several sub-vectors (blocks), generate each block from its conditional distribution given the rest of the blocks and apply the MH algorithm to each generated block. Without loss of generality, suppose the proposal distribution (4.28) is partitioned as

$$\underbrace{\begin{bmatrix} h_1 \\ h_2 \end{bmatrix}}_{h_{1:T}} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} \mu_{h1} \\ \mu_{h2} \end{bmatrix}}_{\mu_h}, \underbrace{\begin{bmatrix} \Sigma_{h11} & \Sigma_{h12} \\ \Sigma_{h21} & \Sigma_{h11} \end{bmatrix}}_{\Sigma_h} \right), \tag{4.29}$$

where $\mu_h = \mu_h\left(h^*_{1:T+1}\right)$ and $\Sigma_h = \Sigma_h\left(h^*_{1:T+1}\right)$ (we ignore the dependence on $h^*_{1:T+1}$ for brevity) and h_1 is the block to be updated in the current MH step while h_2 contains either elements that were already updated in the previous MH steps or those to be updated in the following MH steps. It is well known that the conditional distribution of h_1 given h_2 is given by

$$h_1|h_2 \sim \mathcal{N}\left(\mu_{h1} + \Sigma_{h11}\Sigma_{12}^{-1}\left(h_2 - \mu_{h2}\right), \Sigma_{h11} - \Sigma_{h12}\Sigma_{h22}^{-1}\Sigma_{h21}\right). \tag{4.30}$$

Note that the inverse of the covariance matrix Σ_h in (4.29) is

$$\Sigma_h^{-1} = \begin{bmatrix} \Sigma_{h11} & \Sigma_{h12} \\ \Sigma_{h21} & \Sigma_{h22} \end{bmatrix}^{-1} = \begin{bmatrix} \Omega_{h11} & -\Omega_{h11}\Sigma_{h12}\Sigma_{h22}^{-1} \\ -\Sigma_{h22}^{-1}\Sigma_{h21}\Omega_{h11} & \Sigma_{h22}^{-1} + \Sigma_{h22}^{-1}\Sigma_{h21}\Omega_{h11}\Sigma_{h12}\Sigma_{h22}^{-1} \end{bmatrix},$$
$$\Omega_{h11} = \left(\Sigma_{h11} - \Sigma_{h12}\Sigma_{h22}^{-1}\Sigma_{h21}\right)^{-1}.$$

Furthermore, if we let Ω_{h12} denote the upper-right block of Σ_h^{-1}, we have

$$\Omega_{h12} = -\Omega_{h11}\Sigma_{h12}\Sigma_{h22}^{-1}.$$

Therefore the conditional distribution of h_1 given h_2 in (4.30) is rearranged as

$$h_1|h_2 \sim \mathcal{N}\left(\mu_{h1} - \Omega_{h11}^{-1}\Omega_{h12}\left(h_2 - \mu_{h2}\right), \Omega_{h11}^{-1}\right). \tag{4.31}$$

Recall that Σ_h^{-1} is tridiagonal and so is Ω_{h11} by construction. Thus we can apply the AWOL algorithm:

$$h_1 = \mu_{h1} - \left(L_1'\right)^{-1}\left(L_1^{-1}\Omega_{h12}(h_2 - \mu_{h2}) - \tilde{z}_1\right), \quad \tilde{z}_1 \sim \mathcal{N}(0, \mathrm{I}), \quad L_1 L_1' = \Omega_{h11},$$

to generate h_1 from (4.31). In essence, our approach is an AWOL variant of the block sampler proposed by Omori and Watanabe (2008).

Regression Coefficients β

The sampling scheme for the regression coefficients β is almost identical to the one for the log volatility $h_{1:T+1}$. Let $\ell(\beta)$ denote $\log p(y_{1:T}|h_{1:T+1}, \theta)$ given $y_{1:T}$ and the parameters other than β. In the same manner as (4.19), consider the second-order Taylor approximation of $\ell(\beta)$ in the neighborhood of β^*:

$$\ell(\beta) \approx \ell(\beta^*) + g(\beta^*)'(\beta - \beta^*) - \frac{1}{2}(\beta - \beta^*)'Q(\beta^*)(\beta - \beta^*), \tag{4.32}$$

where $g(\beta)$ is the gradient vector of $\ell(\beta)$ and $Q(\beta)$ is the Hessian matrix of $\ell(\beta)$ times -1. Since $\nabla_\beta \epsilon_t = -\epsilon_t x_t$, we have

$$\nabla_\beta \log p(y_t|h_t, h_{t+1}, \theta) = -x_t + (\epsilon_t^2 - \gamma\eta_t\epsilon_t)x_t,$$
$$\nabla'_\beta \nabla_\beta \log p(y_t|h_t, h_{t+1}, \theta) = (-2\epsilon_t^2 + \gamma\eta_t\epsilon_t)x_t x'_t.$$

Therefore $g(\beta)$ and $Q(\beta)$ are obtained as

$$g(\beta) = \sum_{t=1}^{T} (\epsilon_t(\epsilon_t - \gamma\eta_t) - 1)x_t,$$
$$Q(\beta) = \sum_{t=1}^{T} \epsilon_t(2\epsilon_t - \gamma\eta_t)x_t x'_t.$$

With the prior $\beta \sim \mathcal{N}(\bar{\mu}_\beta, \bar{\Omega}_\beta^{-1})$, the conditional posterior density of β can be approximated by

$$\begin{aligned}
&p(\beta|h_{1:T+1}, \theta_{-\beta}, y_{1:T}) \\
&= C \exp\left[\ell(\beta) + \log p(\beta)\right] \\
&\approx C \exp\left[\ell(\beta^*) - \frac{1}{2}\log(2\pi) + \frac{1}{2}\log|\bar{\Omega}_\beta|\right] \\
&\quad \times \exp\left[g(\beta^*)'(\beta - \beta^*) - \frac{1}{2}(\beta - \beta^*)'Q(\beta^*)(\beta - \beta^*) - \frac{1}{2}(\beta - \bar{\mu}_\beta)'\bar{\Omega}_\beta(\beta - \bar{\mu}_\beta)\right]. \quad (4.33)
\end{aligned}$$

By completing the square as in (4.33), the proposal distribution for the MH algorithm is derived as

$$\beta \sim \mathcal{N}\left(\mu_\beta(\beta^*), \Sigma_\beta(\beta^*)\right), \quad (4.34)$$

where

$$\Sigma_\beta(\beta^*) = \left(Q(\beta^*) + \bar{\Omega}_\beta\right)^{-1}, \quad \mu_\beta(\beta^*) = \Sigma_\beta(\beta^*)\left(g(\beta^*) + Q(\beta^*)\beta^* + \bar{\Omega}_\beta\bar{\mu}_\beta\right).$$

The search algorithm for β^* is the same as $h^*_{1:T+1}$.

Since the dimension of β is considerably smaller than $h_{1:T+1}$, it is not necessary to apply the block sampler in our experience.

Leverage Parameter γ

Since we use the standard conditionally conjugate prior distributions for γ, the conditional posterior distribution is given by

$$\gamma|h_{1:T+1}, \theta_{-\gamma}, y_{1:T} \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T}\eta_t\epsilon_t + \bar{\omega}_\gamma\bar{\mu}_\gamma}{\sum_{t=1}^{T}\eta_t^2 + \bar{\omega}_\gamma}, \frac{1}{\sum_{t=1}^{T}\eta_t^2 + \bar{\omega}_\gamma}\right). \tag{4.35}$$

Variance τ^2

Since we use the standard conditionally conjugate prior distribution for τ^2, the conditional posterior distribution is given by

$$\tau^2|h_{1:T+1}, \theta_{-\tau^2}, y_{1:T} \sim \mathcal{IG}\left(\frac{T+1}{2} + a_\tau, \frac{1}{2}h_{1:T+1}'Vh_{1:T+1} + b_\tau\right). \tag{4.36}$$

AR(1) Coefficient ϕ

Once the state variables $h_{1:T+1}$ are generated, the conditional posterior density of ϕ is given by

$$\begin{aligned} p(\phi|h_{1:T+1}, \theta_{-\phi}, y_{1:T}) \propto \sqrt{1-\phi^2}\exp\left[-\frac{(1-\phi^2)h_1^2 + \sum_{t=1}^{T}(h_{t+1}-\phi h_t)^2}{2\tau^2}\right] \\ \times (1+\phi)^{a_\phi - 1}(1-\phi)^{b_\phi - 1}\mathbf{1}_{(-1,1)}(\phi). \end{aligned} \tag{4.37}$$

By completing the square, we have

$$\begin{aligned} &(1-\phi^2)h_1^2 + \sum_{t=1}^{T}(h_{t+1}-\phi h_t)^2 \\ &= (1-\phi^2)h_1^2 + \sum_{t=1}^{T}h_{t+1}^2 - 2\phi\sum_{t=1}^{T}h_{t+1}h_t + \phi^2\sum_{t=1}^{T}h_t^2 \\ &= \sum_{t=1}^{T+1}h_t^2 - 2\phi\sum_{t=1}^{T}h_{t+1}h_t + \phi^2\sum_{t=2}^{T}h_t^2 \\ &= \sum_{t=2}^{T}h_t^2\left(\phi - \frac{\sum_{t=1}^{T}h_{t+1}h_t}{\sum_{t=2}^{T}h_t^2}\right)^2 + \sum_{t=1}^{T+1}h_t^2 - \frac{\left(\sum_{t=1}^{T}h_{t+1}h_t\right)^2}{\sum_{t=2}^{T}h_t^2}. \end{aligned}$$

With the above expression in mind, we use the following truncated normal distribution:

$$\phi \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T}h_{t+1}h_t}{\sum_{t=2}^{T}h_t^2}, \frac{\tau^2}{\sum_{t=2}^{T}h_t^2}\middle| -1 < \phi < 1\right), \tag{4.38}$$

as the proposal distribution for ϕ in the MH algorithm.

4.3.2 CP Form

Latent Log Volatility $\tilde{h}_{1:T+1}$

The sampling scheme from the conditional posterior distribution of $\tilde{h}_{1:T+1}$:

$$p(\tilde{h}_{1:T+1}|\theta, y_{1:T}) \propto \prod_{t=1}^{T} p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \theta) \cdot p(\tilde{h}_{1:T+1}|\theta), \tag{4.39}$$

is based on the MH algorithm, which is similar to the case of the NCP form. To construct the proposal distribution of $\tilde{h}_{1:T+1}$, we consider the second-order Taylor approximation of $\ell(\tilde{h}_{1:T+1}) = \log p(\tilde{h}_{1:T+1}|\theta, y_{1:T})$ in the neighborhood of $\tilde{h}^*_{1:T+1}$ as in (4.19). We first derive the explicit form of each element in $g(\tilde{h}_{1:T+1})$ and $Q(\tilde{h}_{1:T+1})$. By defining

$$\tilde{\epsilon}_t = y_t \exp(-\tilde{h}_t), \quad \tilde{\eta}_t = \tilde{h}_{t+1} - \phi\tilde{h}_t - (x_{t+1} - \phi x_t)'\beta, \tag{4.40}$$

the log density of y_t in (4.15) is rewritten as

$$\log p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \theta) = -\tilde{h}_t - \frac{1}{2}(\tilde{\epsilon}_t - \gamma\tilde{\eta}_t)^2 + \text{constant}.$$

Since

$$\nabla_t\tilde{\epsilon}_t = -\tilde{\epsilon}_t, \quad \nabla_t\tilde{\eta}_t = -\phi, \quad \nabla_t\tilde{\eta}_{t-1} = 1,$$

$g_t(\tilde{h}_{1:T+1})$, $q_t(\tilde{h}_{1:T+1})$ and $q_{t,t+1}(\tilde{h}_{1:T+1})$ are identical to (4.21), (4.22) and (4.23) except that ϵ_t and η_t are replaced with $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ respectively.

Since the log prior density of $\tilde{h}_{1:T+1}$ is

$$\bar{p}(\tilde{h}_{1:T+1}) = -\frac{T+1}{2}\log(2\pi\tau^2) + \frac{1}{2}\log|V| - \frac{1}{2\tau^2}(\tilde{h}_{1:T+1} - X\beta)'V(\tilde{h}_{1:T+1} - X\beta), \tag{4.41}$$

the conditional posterior density of $\tilde{h}_{1:T+1}$ (4.39) can be approximated by

$$\begin{aligned}
&p(\tilde{h}_{1:T+1}|\theta, y_{1:T}) \\
&= C\exp\left[\ell(h_{1:T+1}) + \bar{p}(h_{1:T+1})\right] \\
&\approx C\exp\left[\ell(\tilde{h}^*_{1:T+1}) - \frac{T+1}{2}\log(2\pi\tau^2) + \frac{1}{2}\log|V| + f(\tilde{h}_{1:T+1})\right],
\end{aligned} \tag{4.42}$$

where C is the normalizing constant of the conditional posterior density and

$$\begin{aligned}
f(\tilde{h}_{1:T+1}) = {} & g(\tilde{h}^*_{1:T+1})'(\tilde{h}_{1:T+1} - \tilde{h}^*_{1:T+1}) \\
& - \frac{1}{2}(\tilde{h}_{1:T+1} - \tilde{h}^*_{1:T+1})'Q(\tilde{h}^*_{1:T+1})(\tilde{h}_{1:T+1} - \tilde{h}^*_{1:T+1}) \\
& - \frac{1}{2\tau^2}(\tilde{h}_{1:T+1} - X\beta)'V(\tilde{h}_{1:T+1} - X\beta),
\end{aligned} \tag{4.43}$$

By completing the square in (4.43), we have

$$f(\tilde{h}_{1:T+1}) = -\frac{1}{2}\left(\tilde{h}_{1:T+1} - \mu_{\tilde{h}}(\tilde{h}^*_{1:T+1})\right)' \Sigma_{\tilde{h}}(\tilde{h}^*_{1:T+1})^{-1}\left(\tilde{h}_{1:T+1} - \mu_{\tilde{h}}(\tilde{h}^*_{1:T+1})\right) \qquad (4.44)$$
$$+ \text{constant},$$

where

$$\Sigma_{\tilde{h}}(\tilde{h}^*_{1:T+1}) = \left(Q(\tilde{h}^*_{1:T+1}) + \frac{1}{\tau^2}V\right)^{-1},$$
$$\mu_{\tilde{h}}(\tilde{h}^*_{1:T+1}) = \Sigma_{\tilde{h}}(\tilde{h}^*_{1:T+1})\left(g(\tilde{h}^*_{1:T+1}) + Q(\tilde{h}^*_{1:T+1})\tilde{h}^*_{1:T+1} + \frac{1}{\tau^2}VX\beta\right).$$

Therefore the right-hand side of (4.42) is approximately proportional to the pdf of the following normal distribution:

$$\tilde{h}_{1:T+1} \sim \mathcal{N}\left(\mu_{\tilde{h}}(\tilde{h}^*_{1:T+1}), \Sigma_{\tilde{h}}(\tilde{h}^*_{1:T+1})\right), \qquad (4.45)$$

which we use as the proposal distribution in the MH algorithm. We obtain $\tilde{h}^*_{1:T+1}$ in (4.45) with the same search algorithm as in the case of the NCP form and apply the block sampler to improve the acceptance rate in the MH algorithm.

Regression Coefficients β

By ignoring the terms that do not depend on β, we can rearrange the density of Y_t in (4.15) as

$$p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \beta, \theta_{-\beta}) \propto \exp\left[-\frac{1}{2}(\tilde{y}_t - \tilde{x}_t'\beta)^2\right],$$

where

$$\tilde{y}_t = y_t \exp(-\tilde{h}_t) - \gamma(\tilde{h}_{t+1} - \phi\tilde{h}_t), \quad \tilde{x}_t = -\gamma(x_{t+1} - \phi x_t).$$

By defining $\tilde{y} = [\tilde{y}_1; \ldots; \tilde{y}_T]$ and $\tilde{X} = [\tilde{x}_1'; \ldots; \tilde{x}_T]$, we have

$$p(y_{1:T}|\tilde{h}_{1:T+1}, \beta, \theta_{-\beta}) = \prod_{t=1}^{T} p(y_t|\tilde{h}_t, \tilde{h}_{t+1}, \beta, \theta_{-\beta})$$
$$\propto \exp\left[-\frac{1}{2}(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta)\right]. \qquad (4.46)$$

Then the conditional posterior distribution of β is given by

$$\begin{aligned}
&p(\beta|\tilde{h}_{1:T+1}, \theta_{-\beta}, y_{1:T}) \\
&\propto p(y_{1:T}|\tilde{h}_{1:T+1}, \beta, \theta_{-\beta}) p(\tilde{h}_{1:T+1}|\beta, \theta_{-\beta}) p(\beta) \\
&\propto \exp\Bigg[-\frac{1}{2}(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) - \frac{1}{2\tau^2}(\tilde{h}_{1:T+1} - X\beta)'V(\tilde{h}_{1:T+1} - X\beta) \\
&\qquad - \frac{1}{2}(\beta - \bar{\mu}_\beta)'\bar{\Omega}_\beta(\beta - \bar{\mu}_\beta)\Bigg]. \qquad (4.47)
\end{aligned}$$

By completing the square, we have the conditional posterior distribution of β as

$$\beta \sim \mathcal{N}\left(\tilde{\mu}_\beta, \tilde{\Sigma}_\beta\right), \qquad (4.48)$$

where

$$\begin{aligned}
\tilde{\Sigma}_\beta &= \left(\tilde{X}'\tilde{X} + \frac{1}{\tau^2}X'VX + \bar{\Omega}_\beta\right)^{-1}, \\
\tilde{\mu}_\beta &= \tilde{\Sigma}_\beta\left(\tilde{X}'\tilde{y} + \frac{1}{\tau^2}X'V\tilde{h}_{1:T+1} + \bar{\Omega}_\beta\bar{\mu}_\beta\right).
\end{aligned}$$

Since $\tilde{X}'\tilde{X} = \gamma^2 X'VX$,

$$\begin{aligned}
\tilde{\Sigma}_\beta &= \left(\left(\gamma^2 + \frac{1}{\tau^2}\right)X'VX + \bar{\Omega}_\beta\right)^{-1}, \\
\tilde{\mu}_\beta &= \tilde{\Sigma}_\beta\left(\tilde{X}'\tilde{\epsilon} + \left(\gamma^2 + \frac{1}{\tau^2}\right)X'V\tilde{h}_{1:T+1} + \bar{\Omega}_\beta\bar{\mu}_\beta\right),
\end{aligned}$$

where $\tilde{\epsilon} = [\tilde{\epsilon}_1; \ldots; \tilde{\epsilon}_T]$.

Leverage Parameter γ

Replacing ϵ_t and η_t in (4.35) with $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ respectively, we have

$$\gamma|\tilde{h}_{1:T+1}, \theta_{-\gamma}, y_{1:T} \sim \mathcal{N}\left(\frac{\sum_{t=1}^T \tilde{\eta}_t\tilde{\epsilon}_t + \bar{\omega}_\gamma\bar{\mu}_\gamma}{\sum_{t=1}^T \tilde{\eta}_t^2 + \bar{\omega}_\gamma}, \frac{1}{\sum_{t=1}^T \tilde{\eta}_t^2 + \bar{\omega}_\gamma}\right). \qquad (4.49)$$

Variance τ^2

It is straightforward to show that the conditional posterior distribution of τ^2 is

$$\tau^2|\tilde{h}_{1:T+1}, \theta_{-\tau^2}, y_{1:T} \sim \mathcal{IG}\left(\frac{T+1}{2}a_\tau, \frac{1}{2}(\tilde{h}_{1:T+1} - X\beta)'V(\tilde{h}_{1:T+1} - X\beta) + b_\tau\right). \qquad (4.50)$$

AR(1) Coefficient ϕ

Replacing h_t in derivation of (4.38) with $\tilde{h}_t - x_t'\beta$, we have

$$\phi \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T}(\tilde{h}_{t+1} - x_{t+1}'\beta)(\tilde{h}_t - x_t'\beta)}{\sum_{t=2}^{T}(\tilde{h}_t - x_t'\beta)^2}, \frac{\tau^2}{\sum_{t=2}^{T}(\tilde{h}_t - x_t'\beta)^2} \middle| -1 < \phi < 1 \right). \tag{4.51}$$

We use (4.51) as the proposal distribution for ϕ in the MH algorithm.

4.4 Extension: Skew Heavy-Tailed Distributions

4.4.1 Mean-Variance Mixture of the Normal Distribution

It is a well-known stylized fact that probability distributions of stock returns are almost definitely heavy-tailed (the probability density goes down to zero much slower than the normal distribution) and often have non-zero skewness (they are not symmetric around the mean). Although introduction of stochastic volatility and leverage makes the distribution of y_t skew and heavy-tailed, it may not be sufficient to capture those characteristics of real data. For this reason, instead of the normal distribution, we introduce a skew heavy-tailed distribution to the SV model.

In our study, we suppose that z_t in (4.5) is expressed as a mean-variance mixture of the standard normal distribution:

$$z_t = \alpha\delta_t + \sqrt{\delta_t}u_t, \quad u_t \sim \mathcal{N}(0,1), \quad \delta_t \sim \mathcal{GIG}(\lambda, \psi, \xi), \tag{4.52}$$

where $\mathcal{GIG}(\lambda, \psi, \xi)$ stands for the generalized inverse Gaussian distribution with the probability density:

$$p(\delta_t) = \frac{(\psi/\xi)^{\lambda/2}}{2K_\lambda(\sqrt{\psi\xi})}\delta_t^{\lambda-1}\exp\left[-\frac{1}{2}\left(\psi\delta_t + \frac{\xi}{\delta_t}\right)\right], \tag{4.53}$$

where

$$\lambda \in \mathbb{R}, \quad (\psi, \xi) \in \begin{cases} \{(\psi,\xi) : \psi > 0, \xi \geq 0\} & \text{if } \lambda > 0, \\ \{(\psi,\xi) : \psi > 0, \xi > 0\} & \text{if } \lambda = 0, \\ \{(\psi,\xi) : \psi \geq 0, \xi > 0\} & \text{if } \lambda < 0, \end{cases}$$

and $K_\lambda(\cdot)$ is the modified Bessel function of the second kind. The family of generalized inverse Gaussian distributions includes

- exponential distribution ($\lambda = 1, \xi = 0$),
- gamma distribution ($\lambda > 0, \xi = 0$),
- inverse gamma distribution ($\lambda < 0, \psi = 0$),
- inverse Gaussian distribution ($\lambda = -\frac{1}{2}$)

Under the assumption (4.52), the distribution of z_t belongs to the family of generalized hyperbolic distributions proposed by Barndorff-Nielsen (1977) which includes many well-known skew heavy-tailed distributions such as

- skew variance gamma (VG) distribution[1] ($\lambda = \frac{\nu}{2}, \psi = \nu, \xi = 0$),
- skew t distribution ($\lambda = -\frac{\nu}{2}, \psi = 0, \xi = \nu$),

where $\nu > 0$. Thus we have two additional parameters (α, ν) in the SV model. Since α determines whether the distribution of y_t is symmetric or not while ν determines how heavy-tailed the distribution is, we call α the *asymmetry parameter* and ν the *tail parameter* respectively. In our study, we use the above three skew heavy-tailed distributions as alternatives to the normal distribution. To distinguish each model specification, we use the following abbreviations:

SV-N:	stochastic volatility model with the normal error,
SV-G:	stochastic volatility model with the VG error,
SV-SG:	stochastic volatility model with the skew VG error,
SV-T:	stochastic volatility model with the Student-t error,
SV-ST:	stochastic volatility model with the skew t error.

In this setup, the SV model with heavy-tailed error is formulated as

$$\begin{cases} y_t = \exp(x_t'\beta + h_t)\epsilon_t, \\ h_{t+1} = \phi h_t + \eta_t, \end{cases} \qquad \left.\begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix} \right| \delta_t \sim \mathcal{N}\left(\begin{bmatrix} \alpha\delta_t \\ 0 \end{bmatrix}, \begin{bmatrix} \delta_t + \gamma^2\tau^2 & \gamma\tau^2 \\ \gamma\tau^2 & \tau^2 \end{bmatrix} \right). \tag{4.54}$$

[1] In general, the skew VG distribution is a mean-variance mixture of the standard normal distribution with $\mathcal{GIG}(\lambda, \psi, 0)$. To make the estimation easier, we set $\lambda = \frac{\nu}{2}$ and $\psi = \nu$ so that the skew VG distribution has only two free parameters (α, ν).

It is straightforward to show that the conditional probability density of y_t given (h_t, h_{t+1}) is given by

$$p(y_t|h_t, h_{t+1}, \theta) = \int_0^\infty p(y_t|h_t, h_{t+1}, \delta_t, \theta)p(\delta_t|\nu)\delta_t, \tag{4.55}$$

where $\theta = (\beta, \gamma, \tau^2, \phi, \alpha, \nu)$,

$$\begin{aligned} &p(y_t|h_t, h_{t+1}, \delta_t, \theta) \\ &= \frac{1}{\sqrt{2\pi\delta_t}} \exp\left[-x_t'\beta - h_t - \frac{\{y_t \exp\left(-x_t'\beta - h_t\right) - \alpha\delta_t - \gamma(h_{t+1} - \phi h_t)\}^2}{2\delta_t}\right], \end{aligned} \tag{4.56}$$

and

$$p(\delta_t|\nu) = \begin{cases} \dfrac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\delta_t^{\frac{\nu}{2}-1}\exp\left(-\dfrac{\nu}{2}\delta_t\right) & \text{(SV-SG)}, \\ \dfrac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\delta_t^{-\frac{\nu}{2}-1}\exp\left(-\dfrac{\nu}{2\delta_t}\right) & \text{(SV-ST)}. \end{cases} \tag{4.57}$$

Since it is impractical to evaluate the multiple integral in (4.55), we generate $\delta_{1:T} = (\delta_1, \ldots, \delta_T)$ along with $h_{1:T+1}$ and θ form their joint posterior distribution. In this setup, the likelihood used in the posterior simulation is

$$\begin{aligned} p(y_{1:T}, h_{1:T+1}, \delta_{1:T}|\theta) &= p(y_{1:T}|h_{1:T+1}, \delta_{1:T}, \theta)p(h_{1:T+1}|\theta) \\ &= \prod_{t=1}^{T} p(y_t|h_t, h_{t+1}, \theta) \cdot p(h_{1:T+1}|\theta). \end{aligned} \tag{4.58}$$

We suppose that the prior distributions for α and ν are

$$\alpha \sim \mathcal{N}(\bar{\mu}_\alpha, \bar{\omega}_\alpha^{-1}), \quad \nu \sim \mathcal{G}(a_\nu, b_\nu). \tag{4.59}$$

where $\mathcal{G}(a, b)$ is the gamma distribution with the probability density:

$$p(x|a, b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}.$$

As for the other parameters, we keep the same ones in (4.10).

4.4.2 Conditional Posterior Distributions

Latent Log Volatility $h_{1:T+1}$

Our sampling scheme for $h_{1:T+1}$ is basically the same as before. We first approximate the log likelihood with the second-order Taylor expansion around the mode and construct

a proposal distribution of $h_{1:T+1}$ with the approximated log likelihood. Then we apply a multi-move MH sampler for generating $h_{1:T+1}$ from the conditional posterior distribution. The sole differences are the functional form of $g(h_{1:T+1})$ and $Q(h_{1:T+1})$.

$$g_t(h_{1:T+1}) = \left\{-1 + \frac{1}{\delta_t}\left(\epsilon_t - \alpha\delta_t - \gamma\eta_t\right)\left(\epsilon_t - \gamma\phi\right)\right\}\mathbf{1}(t \leqq T) + \frac{\gamma}{\delta_{t-1}}\left(\epsilon_{t-1} - \alpha\delta_{t-1} - \gamma\eta_{t-1}\right)\mathbf{1}(t \geqq 2), \quad (t = 1, \ldots, T+1), \tag{4.60}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Each diagonal element of $Q(h_{1:T+1})$ is

$$q_{t,t}(h_{1:T+1}) = \frac{1}{\delta_t}\left\{\epsilon_t(\epsilon_t - \alpha\delta_t - \gamma\eta_t) + \left(\epsilon_t - \gamma\phi\right)^2\right\}\mathbf{1}(t \leqq T) + \frac{\gamma^2}{\delta_{t-1}}\mathbf{1}(t \geqq 2), \quad (t = 1, \ldots, T+1), \tag{4.61}$$

and the off-diagonal element is

$$q_{t,t+1}(h_{1:T+1}) = \frac{\gamma}{\delta_t}(\epsilon_t - \gamma\phi), \quad (t = 1, \ldots, T). \tag{4.62}$$

For the NCP form, we use ϵ_t and η_t in (4.20). For the CP form, we replace them with $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ in (4.40).

Regression Coefficients β

The sampling scheme for β is the same as before. For the NCP form, $g(\beta)$ and $Q(\beta)$ are given by

$$g(\beta) = \sum_{t=1}^{T}\left(\frac{\epsilon_t}{\delta_t}\left(\epsilon_t - \alpha\delta_t - \gamma\eta_t\right) - 1\right)x_t, \tag{4.63}$$

$$Q(\beta) = \sum_{t=1}^{T}\frac{\epsilon_t}{\delta_t}\left(2\epsilon_t - \alpha\delta_t - \gamma\eta_t\right)x_t x_t', \tag{4.64}$$

respectively. For the CP form, the conditional posterior distribution of β are given by

$$\beta \sim \mathcal{N}\left(\tilde{\mu}_\beta, \tilde{\Sigma}_\beta\right), \tag{4.65}$$

where

$$\tilde{\Sigma}_\beta = \left(\tilde{X}'D^{-1}\tilde{X} + \frac{1}{\tau^2}X'VX + \bar{\Omega}_\beta\right)^{-1},$$

$$\tilde{\mu}_\beta = \tilde{\Sigma}_\beta\left(\tilde{X}'D^{-1}\tilde{y} + \frac{1}{\tau^2}X'V\tilde{h}_{1:T+1} + \bar{\Omega}_\beta\bar{\mu}_\beta\right),$$

$$\tilde{y}_t = \tilde{\epsilon}_t - \alpha\delta_t - \gamma(\tilde{h}_{t+1} - \phi\tilde{h}_t),\ \tilde{y} = [\tilde{y}_1; \ldots; \tilde{y}_T],\ D = \mathrm{diag}\{\delta_1, \ldots, \delta_T\}.$$

Leverage Parameter γ

Their conditional posterior distribution of γ is given by

$$\gamma|h_{1:T+1}, \delta_{1:T}, \theta_{-\gamma}, y_{1:T} \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T}\eta_t(\epsilon_t/\delta_t - \alpha) + \bar{\omega}_\gamma\bar{\mu}_\gamma}{\sum_{t=1}^{T}\eta_t^2/\delta_t + \bar{\omega}_\gamma}, \frac{1}{\sum_{t=1}^{T}\eta_t^2/\delta_t + \bar{\omega}_\gamma}\right). \quad (4.66)$$

For the NCP form, we use ϵ_t and η_t in (4.20). For the CP form, we replace them with $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ in (4.40).

Random Scale $\delta_{1:T}$

Using the Bayes theorem, we obtain the conditional posterior distribution of δ_t as

$$\delta_t|h_{1:T+1}, \theta, y_{1:T} \sim \mathcal{GIG}(\lambda_t, \psi_t, \xi_t), \quad t = 1, \ldots, T, \quad (4.67)$$

where

$$(\lambda_t, \psi_t, \xi_t) = \begin{cases} \left(\frac{\nu-1}{2}, \alpha^2 + \nu, (\epsilon_t - \gamma\eta_t)^2\right), & \text{(SV-SG)}, \\ \left(-\frac{\nu+1}{2}, \alpha^2, (\epsilon_t - \gamma\eta_t)^2 + \nu\right), & \text{(SV-ST)}. \end{cases}$$

For the NCP form, we use ϵ_t and η_t in (4.20). For the CP form, we replace them with $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ in (4.40).

To improve the performance of the MCMC algorithm, we apply a generalized Gibbs sampler by Liu and Sabatti (2000) to $\{\delta_t\}_{t=1}^T$ after we generate them from the conditional posterior distribution (4.67). This is rather simple. All we need to do is to multiply each of $\{\delta_t\}_{t=1}^T$ by a random number c that is generated from

$$c \sim \begin{cases} \mathcal{GIG}\left(\frac{(\nu-1)T}{2}, (\alpha^2+\nu)\sum_{t=1}^{T}\delta_t, \sum_{t=1}^{T}\frac{(\epsilon_t - \gamma\eta_t)^2}{\delta_t}\right), & \text{(SV-SG)}, \\ \mathcal{GIG}\left(-\frac{(\nu+1)T}{2}, \alpha^2\sum_{t=1}^{T}\delta_t, \sum_{t=1}^{T}\frac{(\epsilon_t - \gamma\eta_t)^2 + \nu}{\delta_t}\right), & \text{(SV-ST)}. \end{cases} \quad (4.68)$$

Asymmetry Parameter α

Using the Bayes theorem, we obtain the conditional posterior distribution of α as

$$\alpha|h_{1:T+1}, \delta_{1:T}, \theta_{-\alpha}, y_{1:T} \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T}(\epsilon_t - \gamma\eta_t) + \bar{\omega}_\alpha\bar{\mu}_\alpha}{\sum_{t=1}^{T}\delta_t + \bar{\omega}_\alpha}, \frac{1}{\sum_{t=1}^{T}\delta_t + \bar{\omega}_\alpha}\right). \quad (4.69)$$

For the NCP form, we use ϵ_t and η_t in (4.20). For the CP form, we replace them with $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ in (4.40).

Tail Parameter ν

The explicit form of the conditional posterior density of ν is not available. Therefore we apply the MH algorithm for generating ν. Note that the gamma density for SV-SG in (4.57) is identical to the inverse gamma density for SV-ST in (4.57) as a function of ν if we exchange δ_t with δ_t^{-1}. Since we use the same gamma prior for ν in either case, the resultant conditional posterior density should be the same in both SV-SG and SV-ST. Therefore it suffices to derive the MH algorithm for SV-ST.

The sampling strategy for ν is basically the same as β, which was originally proposed by Watanabe (2001). We first consider the second-order Taylor expansion of the log conditional posterior density of ν:

$$f(\nu) = \sum_{t=1}^{T} \log p(\delta_t|\nu) + \log p(\nu) + \text{constant} \tag{4.70}$$

$$= \frac{\nu T}{2}\log\frac{\nu}{2} - T\log\Gamma\left(\frac{\nu}{2}\right) - \nu\left\{\frac{1}{2}\sum_{t=1}^{T}\left(\log\delta_t + \frac{1}{\delta_t}\right) + b_\nu\right\} + (a_\nu - 1)\log\nu \tag{4.71}$$

$$+ \text{constant},$$

with respect to ν in the neighborhood of $\nu^* > 0$, i.e.,

$$f(\nu) \approx f(\nu^*) + g(\nu^*)(\nu - \nu^*) - \frac{1}{2}q(\nu^*)(\nu - \nu^*)^2, \tag{4.72}$$

where

$$\begin{aligned} g(\nu^*) &\equiv \nabla_\nu f(\nu^*) \\ &= \frac{T}{2} + \frac{T}{2}\log\frac{\nu^*}{2} - \frac{T}{2}\psi^{(0)}\left(\frac{\nu^*}{2}\right) - \frac{1}{2}\sum_{t=1}^{T}\left(\log\delta_t + \frac{1}{\delta_t}\right) - b_\nu + \frac{a_\nu - 1}{\nu^*}, \\ q(\nu^*) &\equiv -\nabla_\nu^2 f(\nu^*) \\ &= -\frac{T}{2\nu^*} + \frac{T}{4}\psi^{(1)}\left(\frac{\nu^*}{2}\right) + \frac{a_\nu - 1}{\nu^{*2}}, \end{aligned}$$

and $\psi^{(s)}$ is the polygamma function of order s. Note that $q(\nu^*) > 0$ if $T + 2a_\nu > 2$. See Theorem 1 in Watanabe (2001) for the proof. By applying the completing-the-square

technique to (4.72), we obtain the proposal distribution for the MH algorithm:

$$\nu \sim \mathcal{N}\left(\mu_\nu(\nu^*), \sigma_\nu^2(\nu^*)\right), \tag{4.73}$$

where

$$\sigma_\nu^2(\nu^*) = \frac{1}{q(\nu^*)}, \quad \mu_\nu(\nu^*) = \nu^* + \frac{g(\nu^*)}{q(\nu^*)}.$$

If we use the mode of $f(\nu)$ as ν^*, $g(\nu^*) = 0$ always holds due to the global concavity of $f(\nu)$. Thus $\mu_\nu(\nu^*)$ is effectively identical to ν^*.

4.5 Empirical Study

As an application of our proposed models to real data, we analyze high frequency data of the Tokyo Stock Price Index (TOPIX), a market-cap-weighted stock index based on all domestic common stocks listed in the Tokyo Stock Exchange (TSE) First Section, which is provided by Nikkei Media Marketing. We use the data in June 2016, when the referendum for the UK's withdrawal from the EU (Brexit) was held on the 23rd of the month. The Brexit referendum is arguably one of the biggest financial events in the resent years. So we can analyze the effect of the Brexit referendum on the volatility of the Japanese stock market. Another reason for this choice is that Japan has no holiday in June so that all weekdays are trading days. There are five weeks in June 2016. Since the first week of June 2016 includes May 30th and 31st and the last week includes July 1st, we also include them in the sample period.

The morning session of TSE starts at 9:00 am and ends at 11:30 am while the afternoon session of TSE starts at 12:30 am and ends at 15:00 pm. So both sessions last for 150 minutes. We treat the morning session and the afternoon session as if they are separated trading hours, and normalize the time stamps so that they takes value within $[0, 1]$. As a result, $t = 0$ is corresponding to 9:00 am for the morning session while it is corresponding to 12:30 am for the afternoon session. In the same manner, $t = 1$ is corresponding to 11:30 am for the morning session while it is corresponding to 15:00 pm for the afternoon session. In this empirical study, we estimate the Bernstein polynomial of the intraday seasonality in each session by allowing β in (4.2) to differ from session to session.

Table 4.1: Descriptive statistics of standardized TOPIX 1-minute returns in June 2016

Date	Skewness	Kurtosis	Min.	Max
Week1	−0.1081	7.2296	−7.2569	5.4520
Week2	0.2494	7.7468	−5.9886	5.7911
Week3	0.3534	7.4500	−6.4415	5.8413
Week4	−0.0125	7.1031	−6.4212	5.6074
Week5	0.0346	4.9146	−4.6433	5.4437

We pick prices at every 1 minute and compute 1-minute log difference of prices as 1-minute stock returns. Thus the number of observations per session is 150. Furthermore we put together all series of 1-minute returns in each week. As a result, the total number of observations per week is $150 \times 2 \times 5 = 1,500$. In addition, to simplify the interpretation of the estimation results, we standardize each week-long series of 1-minute returns so that the sample mean is 0 and the sample variance is 1. Table 1 shows the descriptive statistics of the standard 1-minute returns of TOPIX in each week while Figure 1–5 show the time series plots of the standardized 1-minute returns for each week.

We consider 5 candidates (SV-N, SV-G, SV-SG, SV-T, SV-ST) in the SV model (4.54) and set the prior distributions as follows:

$$\beta \sim \mathcal{N}(0, 100\mathrm{I}), \quad \gamma \sim \mathcal{N}(0, 100), \quad \tau^2 \sim \mathcal{IG}(1, 0.04),$$
$$\frac{\phi + 1}{2} \sim \mathcal{B}(1, 1), \quad \alpha \sim \mathcal{N}(0, 100), \quad \nu \sim \mathcal{G}(0, 0.1).$$

We vary the order of the Bernstein polynomial from 5 to 10. In sum, we try 30 different model specifications for the SV model (4.54). In the MCMC implementation, we generate 10,000 draws after the first 5,000 draws are discarded as the burn-in periods. To select the best model among the candidates, we employ the widely applicable information criterion (WAIC, Watanabe (2010)). We compute the WAIC of each model specification with the formula by Gelman *et al.* (2014). The results are reported in Table 2–6. According to these tables, SV-G or SV-SG is the best model in all months. It may be a notable finding since the SV model with the variance-gamma error has hardly been applied in the previous studies.

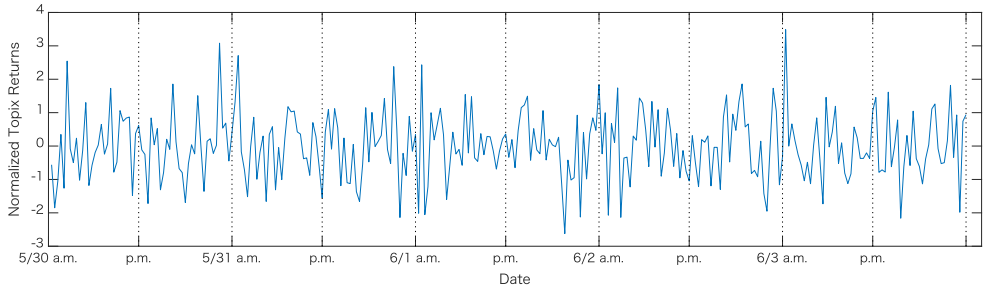


Figure 4.1: Standardized returns of the TOPIX data in the first week of June 2016

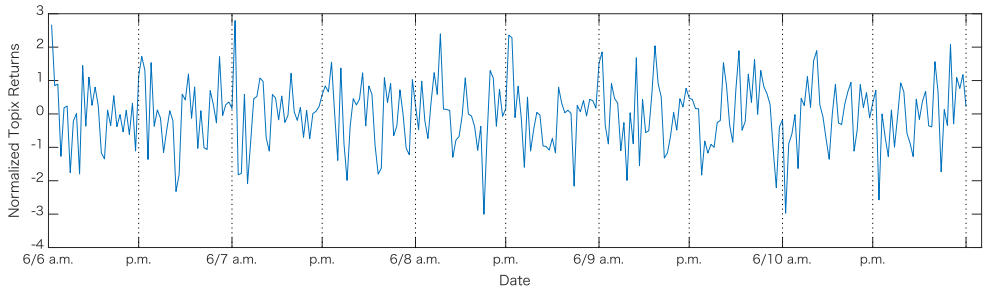


Figure 4.2: Standardized returns of the TOPIX data in the second week of June 2016

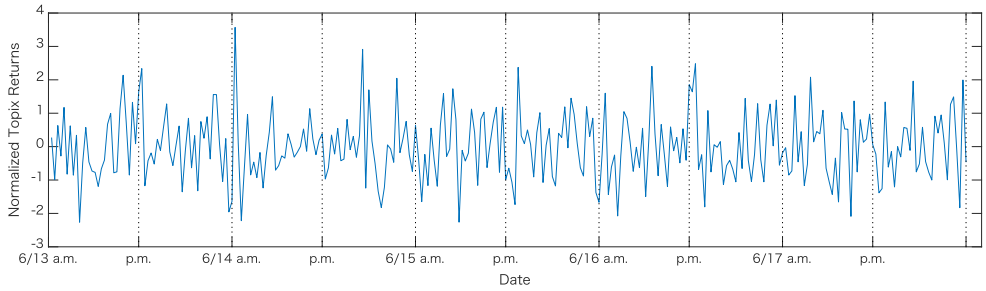


Figure 4.3: Standardized returns of the TOPIX data in the third week of June 2016

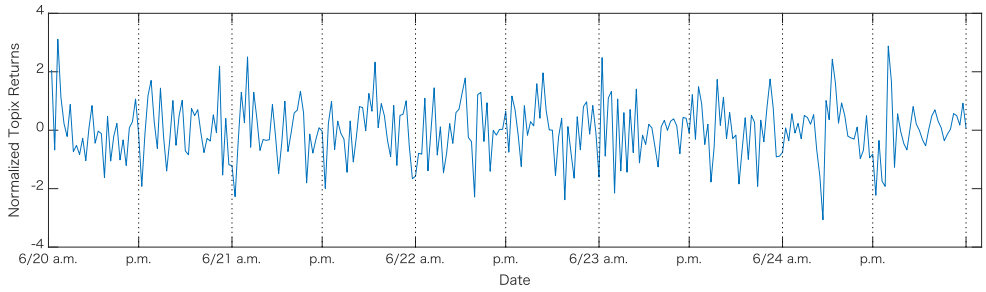


Figure 4.4: Standardized returns of the TOPIX data in the fourth week of June 2016

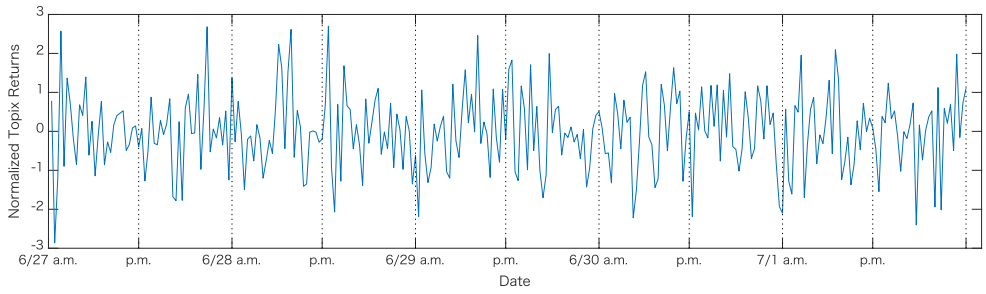


Figure 4.5: Standardized returns of the TOPIX data in the fifth week of June 2016

Table 4.2: WAIC values of TOPIX returns (week 1)

	Order 5	Order 6	Order 7	Order 8	Order 9	Order 10
SV-N	3965.1	3966.0	3962.0	3975.1	3969.6	3971.3
SV-G	3701.0	3698.5	3702.2	**3697.4**	3700.0	3701.1
SV-SG	3701.9	3701.0	3698.9	3699.0	3702.8	3700.8
SV-T	3813.2	3813.3	3813.8	3816.6	3813.9	3813.0
SV-ST	3819.3	3813.8	3816.5	3816.2	3815.7	3817.5

Note: Bold highlight means the best model according to WAIC.

Table 4.3: WAIC values of TOPIX returns (week 2)

	Order 5	Order 6	Order 7	Order 8	Order 9	Order 10
SV-N	3811.0	3813.7	3814.5	3815.7	3818.0	3819.6
SV-G	3621.0	3622.7	3621.0	3619.7	3621.2	3621.8
SV-SG	**3618.6**	3621.3	3622.0	3623.7	3623.4	3619.7
SV-T	3685.4	3684.1	3681.5	3859.0	3860.4	3684.6
SV-ST	3686.9	3684.9	3683.3	3684.1	3684.3	3686.4

Note: Bold highlight means the best model according to WAIC.

Table 4.4: WAIC values of TOPIX returns (week 3)

	Order 5	Order 6	Order 7	Order 8	Order 9	Order 10
SV-N	3976.9	3991.1	3996.8	3975.1	3969.6	3971.3
SV-G	**3750.2**	3752.4	3752.0	3755.2	3752.9	3754.3
SV-SG	3753.6	3753.3	3755.8	3754.1	3759.9	3752.8
SV-T	3862.0	3859.4	3861.6	3859.0	3860.4	3861.6
SV-ST	3862.1	3861.2	3861.3	3860.6	3861.5	3862.1

Note: Bold highlight means the best model according to WAIC.

Table 4.5: WAIC values of TOPIX returns (week 4)

	Order 5	Order 6	Order 7	Order 8	Order 9	Order 10
SV-N	3927.5	3924.5	3924.8	3925.8	3924.7	3927.1
SV-G	3670.3	3680.1	3679.0	**3662.8**	3675.7	3670.4
SV-SG	3663.5	3668.2	3670.3	3664.6	3672.6	3669.1
SV-T	3750.9	3751.9	3752.3	3750.8	3753.1	3753.9
SV-ST	3753.5	3754.3	3753.4	3755.3	3752.8	3751.7

Note: Bold highlight means the best model according to WAIC.

Table 4.6: WAIC values of TOPIX returns (week 5)

	Order 5	Order 6	Order 7	Order 8	Order 9	Order 10
SV-N	4114.6	4114.3	4111.3	4115.2	4113.7	4114.5
SV-G	3961.2	3959.6	3958.3	3960.0	3961.2	3960.0
SV-SG	3953.7	3955.6	3960.0	3963.3	3957.7	**3953.1**
SV-T	4033.7	4033.8	4032.7	4031.2	4032.8	4033.1
SV-ST	4032.5	4033.0	4033.4	4030.4	4032.0	4031.5

Note: Bold highlight means the best model according to WAIC.

For the selected models, we compute the posterior statistics (posterior means, standard deviations, 95% credible intervals and inefficiency factors) of the parameters and report them in Table 7–11. In these tables, the 95% credible intervals of the leverage

parameter γ and the asymmetric parameter α contain 0 for all specifications. Thus we may conclude that the error distribution of 1-minute returns of TOPIX is not asymmetric. In addition, most of the marginal posterior distribution of ϕ are concentrated near 1, even though the uniform prior is assumed for ϕ. This suggests that the latent log volatility is strongly persistent, which is consistent with findings by the previous studies on the stock markets. Regarding the tail parameter ν, its marginal posterior distribution is centered around 2–6 in most models, which indicates that the excess kurtosis of the error distribution is high.

As for the intraday seasonality, the estimates of β themselves are not of our interest. Instead we show the posterior mean and the 95% credible interval of the Bernstein polynomial $x_t'\beta$ in Figure 6. These figures show that some of the trading days exhibit the well-known U-shaped curve of intraday volatility, but others slant upward or downward. At the beginning on the day of Brexit (June 23rd), the market began with a highly volatile situation, but the volatility gradually became lower. During the afternoon session, the volatility was kept in a stable condition.

Table 4.7: Estimation results for TOPIX returns (week 1)

	γ	ϕ	τ	α	ν
SV-N (7)[a]	−0.5742[b]	0.8909	0.1903		
	[−1.4270, −0.1241][c]	[0.8214, 0.9431]	[0.1306, 0.2494]		
	3.42[d]	4.01	4.54		
SV-G (8)	−1.3565	0.9608	0.0798		2.5248
	[−3.1743, 0.8347]	[0.9320, 0.9815]	[0.0604, 0.1034]		[2.0444, 3.2647]
	4.13	3.17	4.45		3.40
SV-SG (7)	−1.4520	0.9598	0.0812	0.0014	2.5489
	[−3.2233, 0.4995]	[0.9316, 0.9802]	[0.0630, 0.1077]	[−0.0504, 0.0544]	[2.0425, 3.3462]
	3.99	3.01	4.45	1.23	3.48
SV-T (10)	−1.4209	0.9864	0.0766		4.2950
	[−4.3298, 1.5726]	[0.9745, 0.9955]	[0.0594, 0.1010]		[3.3294, 5.5097]
	3.85	2.90	4.42		3.44
SV-ST (6)	−1.6137	0.9870	0.0753	0.0006	4.2338
	[−5.0449, 1.6958]	[0.9751, 0.9957]	[0.0586, 0.0962]	[−0.0400, 0.0407]	[3.3247, 5.5254]
	3.94	2.78	4.38	1.62	3.54

Note: a: the selected Bernstein polynomial order, b: posterior mean, c: 95% credible interval and d: inefficiency factor

Table 4.8: Estimation results for TOPIX returns (week 2)

	γ	ϕ	τ	α	ν
SV-N (5)[a]	−0.2127[b]	0.7149	0.3435		
	[−0.5550, 0.0923][c]	[0.5967, 0.8106]	[0.2750, 0.4187]		
	3.04[d]	3.87	4.39		
SV-G (8)	0.0397	0.9191	0.0917		2.2475
	[−1.7042, 2.0695]	[0.8122, 0.9694]	[0.0604, 0.1442]		[2.0089, 2.7055]
	4.27	4.05	4.58		2.6736
SV-SG (5)	0.0380	0.8965	0.0991	−0.0001	2.2622
	[−1.5878, 1.9299]	[0.6761, 0.9664]	[0.0647, 0.1610]	[−0.0530, 0.0531]	[2.0102, 2.7308]
	4.24	4.46	4.61	1.24	2.85
SV-T (7)	-0.7862	0.914	0.0673		3.30
	[−4.7784, 2.8361]	[0.9825, 0.9979]	[0.0511, 0.0886]		[2.7141, 4.0299]
	3.91	2.83	4.43		3.17
SV-ST (7)	−0.6862	0.9909	0.0690	−0.0027	3.3647
	[−4.4268, 2.9433]	[0.9816, 0.9976]	[0.0535, 0.0902]	[−0.0376, 0.0323]	[2.7611, 4.0958]
	3.93	2.77	4.42	1.50	3.14

Note: a: the selected Bernstein polynomial order, b: posterior mean, c: 95% credible interval and d: inefficiency factor

Table 4.9: Estimation results for TOPIX returns (week 3)

	γ	ϕ	τ	α	ν
SV-N (9)[a]	0.0514[b]	0.6249	0.2950		
	[−0.3345, 0.4436][c]	[0.3506, 0.8190]	[0.2042, 0.3872]		
	2.71[d]	4.39	4.54		
SV-G (5)	0.0639	0.4533	0.0919		2.2888
	[−1.8134, 1.9573]	[0.1344, 0.7393]	[0.0632, 0.1413]		[2.0155, 2.7501]
	4.20	4.33	4.57		2.65
SV-SG (10)	−0.2023	0.7992	0.0851	−0.0031	2.3419
	[−2.1985, 1.7331]	[0.2317, 0.9511]	[0.0596, 0.1250]	[−0.0546, 0.0485]	[2.0232, 2.9008]
	4.20	4.59	4.55 [	1.17	2.99
SV-T (8)	−0.2369	0.9871	0.0661		4.0539
	[−3.6943, 3.4453]	[0.9755, 0.9960]	[0.0514, 0.0837]		[3.2156, 5.1327]
	3.86	2.83	4.39		3.39
SV-ST (8)	−0.3313	0.9866	0.0670	−0.0034	4.1237
	[−4.2240, 3.6835]	[0.9730, 0.9957]	[0.0521, 0.0900]	[−0.0426, 0.0361]	[3.2114, 5.2538]
	3.98	3.07	4.42	1.59	3.44

Note: a: the selected Bernstein polynomial order, b: posterior mean, c: 95% credible interval and d: inefficiency factor

Table 4.10: Estimation results for TOPIX returns (week 4)

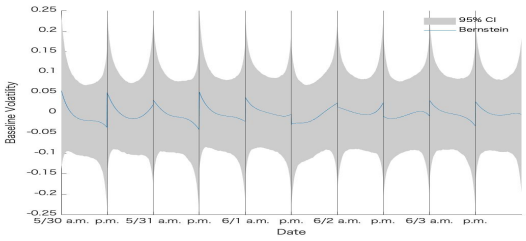
	γ	ϕ	τ	α	ν
SV-N (6)[a]	−0.6502[b]	0.9336	0.1526		
	[−1.8098, 0.2658][c]	[0.8831, 0.9704]	[0.1038, 0.2117]		
	3.53[d]	4.01	4.57		
SV-G (8)	−1.4435	0.9689	0.0853		2.9487
	[−3.6134, 0.6862]	[0.9454, 0.9859]	[0.0657, 0.1098]		[2.1879, 3.9161]
	4.11	3.10	4.43		3.63
SV-SG (5)	−1.7146	0.9694	0.0846	−0.0068	2.93
	[−4.0432,−0.4361]	[0.9458, 0.9868]	[0.0650, 0.1116]	[−0.0599, 0.0471]	[2.14, 3.94]
	4.18	3.11	4.46	1.54	3.64
SV-T (8)	−1.4043	0.9869	0.0824		4.5805
	[−4.4147, −1.6809]	[0.9747, 0.9960]	[0.0634, 0.1060]		[3.5659, 5.8472]
	3.96	2.89	4.41		3.43
SV-ST (10)	−1.6482	0.9882	0.0788	−0.0045	4.4738
	[−5.1451, 1.6390]	[0.9777, 0.9964]	[0.0623, 0.0982]	[−0.0460, 0.0362]	[3.4944, 5.7573]
	4.05	2.64	4.36	1.70	3.39

Note: a: the selected Bernstein polynomial order, b: posterior mean, c: 95% credible interval and d: inefficiency factor

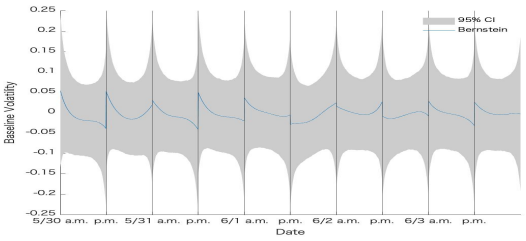
Table 4.11: Estimation results for TOPIX returns (week 5)

	γ	ϕ	τ	α	ν
SV-N (7)[a]	−0.2602[b]	0.8529	0.1428		
	[−1.5977, 0.8891][c]	[0.6617, 0.9395]	[0.0858, 0.2348]		
	3.47[d]	4.38	4.62		
SV-G (7)	−1.1175	0.8638	0.0873		3.6578
	[−3.7543, 1.4340]	[0.6609, 0.9457]	[0.0642, 0.1144]		[2.4812, 5.1414]
	4.21	4.24	4.45		3.86
SV-SG (10)	−1.0562	0.8226	0.0828	-0.0042	3.5728
	[−4.0335, 2.0096]	[0.1974, 0.9485]	[0.0587, 0.1170]	[−0.0565, 0.0490]	[2.4627, 4.8738]
	4.27	4.60	4.52	1.19	3.73
SV-T (8)	−1.2853	0.9727	0.0730		6.2435
	[−4.7632, 1.8139]	[0.9448, 0.9898]	[0.0547, 0.0998]		[4.5496, 8.6440]
	3.92	3.59	4.50		3.71
SV-ST (8)	−1.5744	0.9726	0.0735	−0.0084	6.2388
	[−5.5186, 1.9908]	[0.9459, 0.9898]	[0.0566, 0.0992]	[−0.0536, 0.0378]	[4.6016, 8.6696]
	4.04	3.56	4.46	1.84	3.74

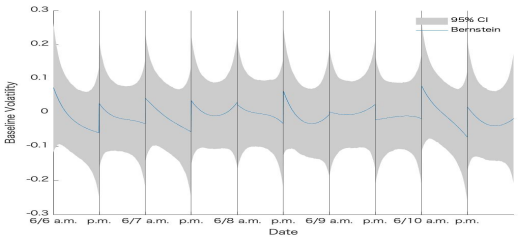
Note: a: the selected Bernstein polynomial order, b: posterior mean, c: 95% credible interval and d: inefficiency factor



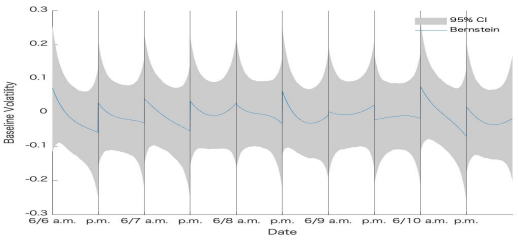
(a) SV-G of Bernstein order 8 in the first week



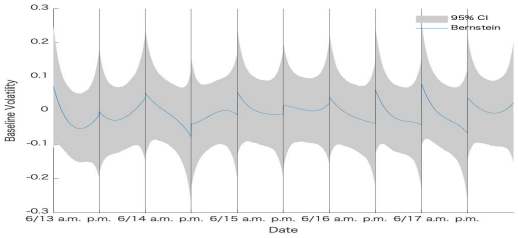
(b) SV-SG of Bernstein order 8 in the first week



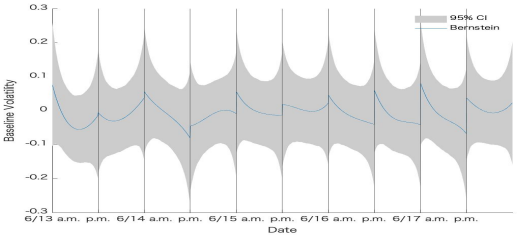
(c) SV-G of Bernstein order 5 in the second week



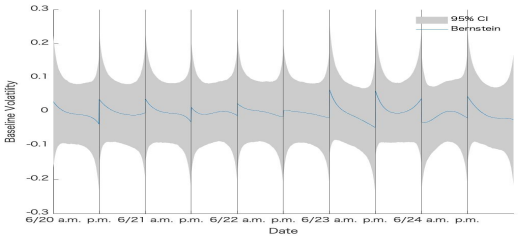
(d) SV-SG of Bernstein order 5 in the second week



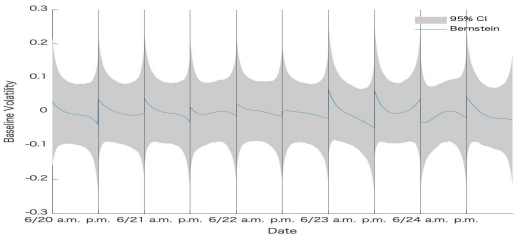
(e) SV-G of Bernstein order 5 in the third week



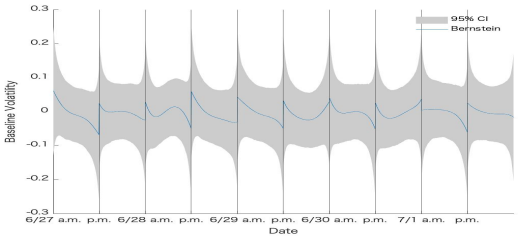
(f) SV-SG of Bernstein order 5 in the third week



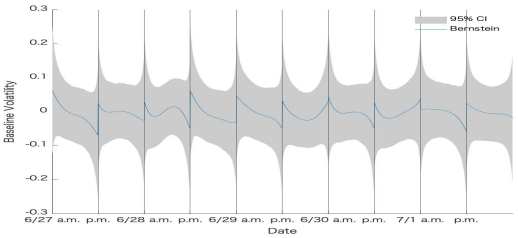
(g) SV-G of Bernstein order 8 in the fourth week



(h) SV-SG of Bernstein order 8 in the fourth week



(i) SV-G of Bernstein order 10 in the fifth week



(j) SV-SG of Bernstein order 10 in the fifth week

Figure 4.6: Intraday seasonality with Bernstein polynomial approximation

4.6 Conclusion

In this chapter, we extended the standard SV model into a more general formulation so that it could capture key characteristics of intraday high frequency stock returns such as intraday seasonality, asymmetry and excess kurtosis. Our proposed model uses a Bernstein polynomial of time stamps as the intraday seasonal component of the stock volatility and the coefficients in the Bernstein polynomial are simultaneously estimated along with the rest of the parameters in the model. To incorporate asymmetry and excess kurtosis into the standard SV model, we assume that the error distribution of stock returns in the SV model belongs to a family of generalized hyperbolic distributions. In particular, we focus on two sub-classes of this family: skew Student's t distribution and skew variance-gamma distribution. Furthermore we developed an efficient MCMC sampling algorithm for Bayesian inference on the proposed model by utilizing AWOL, ASIS and the generalized Gibbs sampler.

As an application, we estimated the proposed SV models with 1-minute return data of TOPIX in various specifications and conducted model selection with WAIC. The model selection procedure chose the SV model with the variance-gamma-type error as the most suitable one. The estimated parameters indicated strong excess kurtosis in the error distribution of 1-minute returns, though the asymmetry was not supported since both leverage parameter γ and asymmetry parameter α were not significantly different from zero. Furthermore our proposed model successfully extracted intraday seasonal patterns in the stock volatility with Bernstein polynomial approximation, though the shape of the intraday seasonal component was not necessarily U-shaped.

Chapter 5

Summary and Future Prospects

In this doctoral dissertation, we introduced the problem of inefficient sampling in Markov chain Monte Carlo (MCMC). Although MCMC is arguably the most indispensable tool for Bayesian statisticians, effective usage of this tool relies on a proper choice of sampling scheme. Bayesian statisticians have devised numerous sampling methods that could solve the problem of slow convergence. In our own opinion, the Ancillarity-suffciency interweaving strategy (ASIS) proposed by Yu and Meng (2011) is possibly one of the most powerful of such methods. ASIS is a versatile tool that can be applied to a wide variety of data and model specifications appropriate for analyzing them. Indeed, from Chapter 2 through Chapter 4, we have shown that ASIS is capable of providing significant improvements in the efficiency of MCMC not only for simple balanced panel data, but also for more complex, high-dimensional unbalanced panel data and high-frequency time series data.

In Chapter 2, we first provided the definition and principle of ASIS. Then we developed a new ASIS scheme for the panel data regression model and compared this new scheme with the traditional non-ASIS Gibbs sampling scheme. In the empirical study with the real-world panel data used by Vella and Verbeek (1998), we showed that the ASIS scheme could improve the efficiency of MCMC sampling in terms of stability and mixing of generated random series. Without ASIS, many parameters in the panel data regression model failed to pass the convergence diagnostic test because generated random series exhibited strongly positive and persistent autocorrelation. With ASIS, on the other hand, generated random series are far more stable and their sample autocorrelation declines far

much faster. Furthermore, even though the panel data regression model includes hundreds of parameters, all of them succeeded in passing the convergence diagnostic test. In sum, the ASIS scheme archived efficient sampling for all parameters while the performance of the non-ASIS scheme was dismal.

In Chapter 3, we conducted a hierarchical Bayesian analysis of racehorse running ability and jockey skills by using the running records of all 1800-m races held by JRA from 2016 to 2018 excluding obstacle courses. Three factors were considered to affect the racehorse speed during a race, which were as follows: the individual racehorse's ability, the environment in which the race is run, and the jockey's ability to control the horse. The horserace data is higher-dimensional and more complex than in Chapter 2; 22,183 runs with 140 jockeys and 4,063 racehorses. In addition to the characteristics of the model that includes a large number of parameters as racehorse's and jockey's individual effects, the racehorse race data used in this study are unbalanced panel data in which the number of times each racehorse has run is not uniform; thus, we apply a hierarchical Bayesian analysis method following Silverman (2012). Furthermore, ASIS is applied to improve the efficiency of convergence of the random series. Moreover, WAIC is used to select combinations of explanatory variables in the regression model.

The application of ASIS leads to a lower autocorrelation of the MCMC sample, which results in a more efficient sampling. The improvement in efficiency is so great that, although the analysis of this study's model is impractical without ASIS, it is now possible to perform the analysis at a level that can be considered efficient in general.

In Chapter 4, we applied ASIS to a complex SV model using time series data; high frequency trading data of stock prices. We extended the standard SV model into a more general formulation so that it could capture key characteristics of intraday high frequency stock returns such as intraday seasonality, asymmetry and excess kurtosis. Our proposed model uses the Bernstein polynomials of time stamps as the intraday seasonal component of the stock volatility and the coefficients for the Bernstein polynomial are simultaneously estimated along with the rest of the parameters in the model. To incorporate asymmetry and excess kurtosis into the standard SV model, we assume that the error distribution of stock returns in the SV model belongs to a family of generalized hyperbolic distributions. In particular, we focus on two sub-classes of this family: skew Student's t distribution

and skew variance-gamma distribution. Furthermore, we developed an efficient MCMC sampling algorithm for Bayesian inference on the proposed model by utilizing ASIS with AWOL and the generalized Gibbs sampler.

As an application, we estimated the proposed SV models with 1-minute return data of TOPIX in various specifications and conducted model selection with WAIC. The model selection procedure chose the SV model with the variance-gamma-type error as the most suitable one. The estimated parameters indicated strong excess kurtosis in the error distribution of 1-minute returns, though the asymmetry was not supported since both leverage parameter γ and asymmetry parameter α were not significantly different from zero. Furthermore our proposed model successfully extracted intraday seasonal patterns in the stock volatility with Bernstein polynomial approximation.

The advantage of ASIS is not only its large effect on improving MCMC efficiency, but also the fact that it is a very "cost-effective" method. For example, in Chapter 2, MCMC efficiency can be improved by simply adding the process of sampling from a "centralized" version of the latent variable or missing data without changing the whole structure of the model.

The future prospects for research include: further development of ASIS to improve the efficiency of MCMC sampling; application of ASIS to other statistical models and real-world data to solve problems that could not be solved in the past; and development of new analytic tools for effective Bayesian inference based on the ideas of ASIS. Although MCMC is very powerful, it is undeniable that the degree of efficiency improvement depends on models and data we use. Thus we would like to develop ASIS that can be applicable to a variety of statistical models and real-world data. In addition, ASIS is still less known among researchers even though its practicality is outstanding. This is because there are not so many examples of its usage in applied fields. Through the future research, we aim not only to develop ASIS, but also to promote it as a widely applicable powerful tool for Bayesian inference.

Appendix A

Derivations of Conditional Posterior Distributions

From the Bayesian perspective, there is no essential difference between the following panel data regression models:

- the regression model of balanced panel data with the one-way individual effect in Chapter 2;
- the regression model of unbalanced panel data with the two-way individual effect in Chapter 3,

because the two models share almost identical model structures, variables and parameters (the definitions of vectors and matrices in the formulae are a little different between them). Hence we will derive the conditional posterior distributions for the latter since the former is nested in the latter.

Conditional posterior distribution of $\boldsymbol{\delta}$

The conditional posterior distribution of $\boldsymbol{\delta}$ is

$$p(\boldsymbol{\delta}|\mathcal{D}, \boldsymbol{\theta}_{-\boldsymbol{\delta}}) \propto \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\delta})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\delta}) + (\boldsymbol{\delta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}-\boldsymbol{\mu})\right\}\right], \qquad \text{(A.1)}$$

with likelihood (3.4) and prior distribution (3.6). Applying the square completion formula, we obtain

$$\begin{aligned}
&\sigma_\epsilon^{-2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\delta})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\delta})+(\boldsymbol{\delta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}-\boldsymbol{\mu}) \\
&=\boldsymbol{\delta}'\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\delta}-2\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)'\boldsymbol{\delta}+\text{Constant} \\
&=\left(\boldsymbol{\delta}-\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right) \\
&\quad\times\left(\boldsymbol{\delta}-\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)+\text{Constant}.
\end{aligned}$$

where "Constant" means a term that does not depend on $\boldsymbol{\delta}$. The conditional posterior distribution of (A.1) is

$$\begin{aligned}
&p(\boldsymbol{\delta}|\mathcal{D},\boldsymbol{\theta}_{-\boldsymbol{\delta}}) \\
&\propto\exp\Bigg[-\frac{1}{2}\Big(\boldsymbol{\delta}-\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\Big)\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right) \\
&\qquad\qquad\times\Big(\boldsymbol{\delta}-\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\sigma_\epsilon^{-2}\boldsymbol{Z}'\boldsymbol{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\Big)\Bigg]. \qquad \text{(A.2)}
\end{aligned}$$

This is equivalent to the probability density function of the distribution in (3.11).

Conditional posterior distribution of μ_α

The conditional posterior distribution of μ_α is

$$\begin{aligned}
p(\mu_\alpha|\mathcal{D},\boldsymbol{\theta}_{-\mu_\alpha}) &\propto\exp\left[-\frac{\sum_{i=1}^N(\alpha_i-\mu_\alpha)^2}{2\sigma_\alpha^2}-\frac{(\mu_\alpha-\varphi_\alpha)^2}{2\tau_\alpha^2}\right] \\
&\propto\exp\left[-\frac{1}{2}\left\{\left(\sigma_\alpha^{-2}N+\tau_\alpha^{-2}\right)\mu_\alpha^2-2\left(\sigma_\alpha^{-2}\sum_{i=1}^N\alpha_i+\tau_\alpha^{-2}\varphi_\alpha\right)\mu_\alpha\right\}\right] \\
&\propto\exp\left[-\frac{1}{2}\left(\sigma_\alpha^{-2}N+\tau_\alpha^{-2}\right)\left(\mu_\alpha-\frac{\sigma_\alpha^{-2}\sum_{i=1}^N\alpha_i+\tau_\alpha^{-2}\varphi_\alpha}{\sigma_\alpha^{-2}N+\tau_\alpha^{-2}}\right)^2\right], \qquad \text{(A.3)}
\end{aligned}$$

from (3.8) and (3.9). This is equivalent to the probability density function of the distribution in (3.12).

Conditional posterior distribution of μ_γ

The conditional posterior distribution of μ_γ is

$$\begin{aligned} p(\mu_\gamma|\mathcal{D},\boldsymbol{\theta}_{-\mu_\gamma}) &\propto \exp\left[-\frac{\sum_{i=1}^{M}(\gamma_i-\mu_\gamma)^2}{2\sigma_\gamma^2}-\frac{(\mu_\gamma-\varphi_\gamma)^2}{2\tau_\gamma^2}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\left(\sigma_\gamma^{-2}M+\tau_\gamma^{-2}\right)\mu_\gamma^2-2\left(\sigma_\gamma^{-2}\sum_{i=1}^{M}\gamma_i+\tau_\gamma^{-2}\varphi_\gamma\right)\mu_\gamma\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\sigma_\gamma^{-2}M+\tau_\gamma^{-2}\right)\left(\mu_\gamma-\frac{\sigma_\gamma^{-2}\sum_{i=1}^{M}\gamma_i+\tau_\gamma^{-2}\varphi_\gamma}{\sigma_\gamma^{-2}M+\tau_\gamma^{-2}}\right)^2\right], \end{aligned} \tag{A.4}$$

from (3.8) and (3.9). This is equivalent to the probability density function of the distribution in (3.13).

Conditional posterior distributions of σ_α^2 and ξ_α

The conditional posterior distribution of σ_α^2 is

$$\begin{aligned} &p(\sigma_\alpha^2|\mathcal{D},\boldsymbol{\theta}_{-\sigma_\alpha^2},\xi_\alpha) \\ &\propto (\sigma_\alpha^2)^{-\frac{N}{2}}\exp\left[-\frac{\sum_{i=1}^{N}(\alpha_i-\mu_\alpha)^2}{2\sigma_\alpha^2}\right]\times(\sigma_\alpha^2)^{-\left(\frac{1}{2}+1\right)}\exp\left(-\frac{1}{\xi_\alpha\sigma_\alpha^2}\right) \\ &\propto (\sigma_\alpha^2)^{-\left(\frac{N+1}{2}+1\right)}\exp\left[-\frac{\frac{1}{2}\sum_{i=1}^{N}(\alpha_i-\mu_\alpha)^2+\xi_\alpha^{-1}}{\sigma_\alpha^2}\right], \end{aligned} \tag{A.5}$$

given the latent variable ξ_α from equations (3.8) and (3.17). This is equivalent to the probability density function of the distribution of σ_α^2 in equation (3.14). Whereas, the conditional posterior distribution of ξ_α is obtained as

$$\begin{aligned} p(\xi_\alpha|\sigma_\alpha^2) &\propto \xi_\alpha^{-\frac{1}{2}}(\sigma_\alpha^2)^{-\left(\frac{1}{2}+1\right)}\exp\left(-\frac{1}{\xi_\alpha\sigma_\alpha^2}\right)\times\xi_\alpha^{-\left(\frac{1}{2}+1\right)}\exp\left(-\frac{1}{s_\alpha^2\xi_\alpha}\right) \\ &\propto \xi_\alpha^{-(1+1)}\exp\left(-\frac{\sigma_\alpha^{-2}+s_\alpha^{-2}}{\xi_\alpha}\right). \end{aligned} \tag{A.6}$$

This is also equivalent to the probability density function of the distribution of ξ_α^2 in equation (3.14).

Conditional posterior distributions of σ_γ^2 and ξ_γ

The conditional posterior distribution of σ_γ^2 is

$$\begin{aligned} & p(\sigma_\gamma^2|\mathcal{D}, \boldsymbol{\theta}_{-\sigma_\gamma^2}, \xi_\gamma) \\ & \propto (\sigma_\gamma^2)^{-\frac{M}{2}} \exp\left[-\frac{\sum_{i=1}^{M}(\gamma_i - \mu_\gamma)^2}{2\sigma_\gamma^2}\right] \times (\sigma_\gamma^2)^{-\left(\frac{1}{2}+1\right)} \exp\left(-\frac{1}{\xi_\gamma \sigma_\gamma^2}\right) \\ & \propto (\sigma_\gamma^2)^{-\left(\frac{M+1}{2}+1\right)} \exp\left[-\frac{\frac{1}{2}\sum_{i=1}^{M}(\gamma_i - \mu_\gamma)^2 + \xi_\gamma^{-1}}{\sigma_\gamma^2}\right], \end{aligned} \tag{A.7}$$

given the latent variable ξ_γ from (3.8) and (3.17). This is equivalent to the probability density function of the distribution of σ_γ^2 in (3.15). Whereas, the conditional posterior distribution of ξ_α in (3.15) is derived by replacing σ_α^2 and s_α^2 by σ_γ^2 and s_γ^2 in (A.6), respectively.

Conditional posterior distribution of σ_ϵ^2 and ξ_ϵ

The conditional posterior distribution of σ_ϵ^2 is

$$\begin{aligned} & p(\sigma_\epsilon^2|\mathcal{D}, \boldsymbol{\theta}_{-\sigma_\epsilon^2}) \\ & \propto (\sigma_\epsilon^2)^{-\frac{T}{2}} \exp\left(-\frac{\sum_{i=}^{N}\sum_{t=1}^{T_i} e_{it}^2}{2\sigma_\epsilon^2}\right) \times (\sigma_\epsilon^2)^{-\left(\frac{1}{2}+1\right)} \exp\left(-\frac{1}{\xi_\epsilon \sigma_\epsilon^2}\right) \\ & \propto (\sigma_\epsilon^2)^{-\left(\frac{T+1}{2}+1\right)} \exp\left(-\frac{\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T_i} e_{it}^2 + \xi_\epsilon^{-1}}{\sigma_\epsilon^2}\right), \end{aligned} \tag{A.8}$$

from (3.5) and (3.17). This is equivalent to the probability density function of the distribution in (3.16). Whereas, the conditional posterior distribution of ξ_ϵ in (3.16) is derived by replacing σ_α^2 and s_α^2 by σ_ϵ^2 and s_ϵ^2 in (A.6) respectively.

Bibliography

[1] C. A. Abanto-Valle, V. H. Lachos and D. K. Dey, Bayesian estimation of a skew-Student-t stochastic volatility model, *Methodology and Computing in Applied Probability* **17** (2015), 721–738.

[2] T. G. Andersen and T. Bollerslev, Intraday periodicity and volatility persistence in financial markets, *Journal of Empirical Finance* **17** (1997), 115–158.

[3] O. Barndorff-Nielsen, Exponentially decreasing distributions for the logarithm of particle size, *Proceedings of the Royal Society A. Mathematical, Physical and Engineering Sciences* **353** (1997), 401–419.

[4] W. Benter, Computer Based Horse Race Handicapping and Wagering Systems: A Report, *Efficiency of Racetrack Betting Markets* (2008), 183–198.

[5] J. C. C. Chan and I. Jeliazkov, Efficient simulation and integrated likelihood estimation in state space models, *Journal Abbreviation* **1** (2009), 101–120.

[6] K. Chan, K. C. Chan and G. A. Karolyi, Intraday volatility in the stock index and stock index futures markets, *The Review of Financial Studies* **4** (1001), 657–684.

[7] S. Chib, Markov chain Monte Carlo methods: computation and inference. In *Handbook of Econometrics*; Heckman, J. J., Leamer, E. E., Eds.; Publishing House: North Holland, Amsterdam, (2001), 3569–3649.

[8] W-C. Chung, C-Y. Chang and C-C. Ko, A SVM-Based committee machine for prediction of Hong Kong horse racing, *2017 10th International Conference on Ubi-media Computing and Workshops* **1** (2017), 147–150.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39(1)** (1977), 1–38.

[10] D. Edelman, Adapting support vector machine methods for horserace odds prediction, *Annals of Operations Research* **151** (2007), 325–336.

[11] M. Fičura and J. Witzany, Estimating Stochastic Volatility and Jumps Using High-Frequency Data and Bayesian Methods, SSRN. Available online: https://ssrn.com/abstract=2551807 (accessed on 8th December 2020).

[12] M. Fičura and J. Witzany, Bayesian estimation of stochastic-volatility jump diffusion models on intraday price returns, *Applications of Mathematics & Statistics in Economics 2015*, Available online: http://amse2015.cz/doc/Ficura_witzany.pdf (accessed on 8th December 2020).

[13] A. E. Gelfand and A. F. M. Smith, Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association* **85** (1990), 398–409.

[14] A. Gelman, Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), *Bayesian Analysis* **1(3)** (2006), 515–534.

[15] A. Gelman, J. Hwang, and A. Vehtari, Understanding predictive information criteria for Bayesian models, *Statistics and Computing* **24(6)** (2014), 997–1016.

[16] S. Geman and D. Geman, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (1984), 721–741.

[17] M. Gramm and R. Marksteiner, The Effect of Age on Thoroughbred Racing Performance, *Journal of Equine Science* **21(4)** (2010), 73–78.

[18] J. Geweke, Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *Bayesian Statistics* **4** (1992), 169–194.

[19] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications , *Biometrika* **57(1)** (1970), 97–109.

[20] G. Kastner and S. Frühwirth-Schnatter, Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models, *Computational Statistics & Data Analysis* **76** (2014), 408–423.

[21] S. Lessmann, M-C. Sung and J. E. V. Johnson, Adapting least-square support vector regression models to forecast the outcome of horseraces, *The Journal of Prediction Markets* **1(3)** (2007), 43–59.

[22] S. Lessmann, M-C. Sung and J. E. V. Johnson, Identifying winners of competitive events: A SVM-based classification model for horserace prediction, *European Journal of Operational Research* **196(2)** (2009), 569–577.

[23] S. Lessmann, M-C. Sung and J. E. V. Johnson, Alternative methods of predicting competitive events: An application in horserace betting markets, *International Journal of Forecasting* **26(3)** (2010), 518–536.

[24] J. S. Liu and C. Sabatti, Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation, *Biometrika* **87(2)** (2000), 353–369.

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

[26] W. J. McCausland, S. Miller and D. Pelletier, Simulation smoothing for state-space models: A computational efficiency analysis, *Computational Statistics & Data Analysis* **55(1)** (2011), 199–212.

[27] N. Metropolis et al., Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics* **21** (1953), 1087–1092.

[28] J. Nakajima, Bayesian analysis of generalized autoregressive conditional heteroskedasticity and stochastic volatility: Modeling leverage, jumps and heavy-tails for financial time series. *Japanese Economic Review* **63(1)** (2012), 81–103.

[29] J. Nakajima and Y. Omori, Leverage, heavy-tails and correlated jumps in stochastic volatility models, *Computational Statistics & Data Analysis* **53(6)** (2009), 2335–2353.

[30] J. Nakajima and Y. Omori, Stochastic volatility model with leverage and asymmetrically heavy-tailed error using GH skew Student's t-distribution, *Computational Statistics & Data Analysis* **56** (2012), 3690–3704.

[31] M. Nakakita and T. Nakatsuma, Bayesian Analysis of Intraday Stochastic Volatility Models of High-Frequency Stock Returns with Skew Heavy-Tailed Errors, *Journal of Risk and Financial Management* **14(4)** (2021), 145.

[32] Y. Omori, S. Chib, N. Shephard and J. Nakajima, Stochastic volatility with leverage: Fast and efficient likelihood inference, *Journal of Econometrics* **140(2)** (2007), 425–449.

[33] Y. Omori and T. Watanabe, Block sampler and posterior mode estimation for asymmetric stochastic volatility models, *Computational Statistics & Data Analysis* **52(6)** (2008), 2892–2910.

[34] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, (2004).

[35] H. Rue, Fast sampling of Gaussian Markov random fields, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63(2)** (2001), 325–338.

[36] N. Silverman, A hierarchical Bayesian analysis of horse racing, *The Journal of Prediction Markets* **6(3)** (2012), 1–13.

[37] N. Silverman and M. A. Suchard, Predicting horse race winners through regularized conditional logistic regression with frailty, *The Journal of Prediction Markets* **7(1)** (2013), 43–53.

[38] J. R. Stroud and M. S. Johannes, Bayesian modeling and forecasting of 24-hour high-frequency volatility, *Journal of the American Statistical Association* **109** (2014), 1368–1384.

[39] M. Sung and J. E. V. Johnson, Comparing the effectiveness of one- and two-step conditional logit models for predicting outcomes in a speculative market, *The Journal of Prediction Markets* **1(1)** (2007), 43–59.

[40] M. A. Tanner and W. H. Wong, The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association* **82** (1987), 528–540.

[41] G. Tsiotas, On generalised asymmetric stochastic volatility models, *Computational Statistics & Data Analysis* **56(1)** (2012), 151–172.

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

[43] F. Vella and M. Verbeek, Whose Wages do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men, *Journal of Applied Econometrics* **13(2)** (1998), 163–183.

[44] S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11** (2010), 3571–3594.

[45] T. Watanabe, On sampling the degree-of-freedom of Student's-t disturbances. *Statistics & Probability Letters* **52** (2001), 177–181.

[46] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th ed., Cengage Learning, (2019).

[47] Y. Yu and X-L. Meng, To center or not to center: that is not the question – An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency, *Journal of Computational and Graphical Statistics* **20** (2011), 531–570.