Essays on Asset Return Modeling with Bayesian Markov Chain Monte Carlo

Sakae Oya
Graduate School of Economics, Keio University

Abstract

Since Markowitz (1952) proposed the mean-variance model for optimal portfolio selection, it has remained an important foundation in research of portfolio theory as well as in practice of portfolio management. Although numerous solutions have been proposed for various problems on the mean-variance model in decades of its history, there remains some challenging tasks; (1) $p > n$ problem in case of high-dimensional data with many assets (p: the number of assets, n: the number of data points) and (2) modeling of a skewed and fat-tailed distribution with real-world data. To deal with these challenges, we will discuss researches on new statistical modeling techniques with Bayesian Markov Chain Monte Carlo (MCMC) and their applications to real-world data in this doctoral dissertation.

The first challenge is a problem how to construct the optimal portfolio when the number of assets p exceeds the number of observations of asset returns n. In a conventional Markowitz-type portfolio, the optimal weights depend on the inverse of the covariance matrix or the precision matrix of asset returns. It is well-known that the calculation of the inverse matrix of the sample covariance matrix becomes unstable as p approaches n, and it becomes impossible when p is greater than n. Thus, to solve this $p > n$ problem, numerous dimensional compression methods have been proposed in the literature. For example, dimensional compression methods based on factor models such as Fama and French (1993) model or BARRA model have widely been used in the field of finance. In recent years, however, another type of dimension compression method that directly estimates the precision

matrix of high-dimensional asset return data with a graphical model such as graphical LASSO (e.g., Meinshausen and Bühlmann (2006), Friedman et al. (2008), Yuan and Lin (2007), Banerjee et al. (2008), Guo et al. (2011)) has been developed in the literature of machine learning. This dissertation will tackle the first challenge in Chapter 2 and 3 from the perspective of the second approach.

In Chapter 2, we focus on a serious flaw in the Bayesian graphical LASSO estimation method proposed by Wang (2012) and propose a remedy for it. In Wang (2012)'s algorithm, the precision matrix is partitioned into three parts: the i-th diagonal element ($i = 1, \ldots, p$), a $(p-1) \times 1$ vector of off-diagonal elements which is corresponding to the precision matrix's i-th column excluding the diagonal element, and the remaining $(p-1) \times (p-1)$ block of the precision matrix. Based on this partition, Wang (2012) derived the full conditional posterior distribution of each diagonal element as well as that of the corresponding off-diagonal elements and constructed a Gibbs sampling algorithm to iteratively generate the precision matrix along with the other parameters in the graphical LASSO model. Wang (2012) defined the vector of the off-diagonal elements in the i-th column of the precision matrix as β and γ as the diagonal element minus the quadratic form of β and the $(p-1) \times (p-1)$ block of the precision matrix. Note that the necessary and sufficient condition for positive definiteness of the precision matrix is that the diagonal element is larger than the corresponding quadratic form in the partition used by Wang (2012). Given β, this condition is satisfied for γ since $\gamma > 0$ always holds in Wang (2012)'s algorithm as long as we use a positive-valued distribution (e.g., gamma distribution) as the prior for γ. This is why Wang (2012) argues that the positive definiteness of the precision matrix is sufficiently guaranteed in the algorithm.

Unlike γ, however, β does not necessarily satisfy the condition for positive definiteness of the precision matrix. We demonstrate in Chapter 2 that the precision matrix is not always positive definite in the Gibbs sampler under the designs of numerical experiments adopted by Wang (2012). For some design of the precision matrix, the positive definiteness is violated for more than 20% of the iterations. The reason is simple. Wang (2012) naively supposes that the full conditional posterior distribution of β is a multivariate normal distribution, but it turns out that an additional constraint must be imposed upon the distribution of β so that the quadratic form of β must be less than the corresponding diagonal element of the precision matrix. In other words, to assure the positive definiteness of the precision matrix, the full conditional posterior distribution must be a truncated multivariate normal distribution. If we generate β from the unconstrained multivariate normal distribution, the positive definite condition can be broken at the moment β is updated.

To solve this issue, we propose a modified algorithm that samples β from the truncated multivariate normal distribution by using Bélisle et al. (1993)'s hit-and-run algorithm. To demonstrate the superiority of the proposed algorithm, we conducted the Monte Carlo experiments in which we inherited the settings from Wang (2012) and generated artificial data with $p = 30$ or $p = 100$ and six designs of the precision matrix (AR1, AR2, Block, Star, Circle, and Full) as the true structure. The results show that this modification not only stabilizes the sampling procedure but also significantly improves the performance of parameter estimation and graphical structure learning. Interestingly, the proposed algorithm also improves the performance in scenarios where Wang (2012)'s algorithm does not violate the positive definiteness.

In Chapter 3, we discuss an application of graphical models to portfolio management. The previous studies (e.g., Goto and Xu (2015), Brownlees et al. (2018), Torri et al. (2019) among others) already applied graphical models to portfolio optimization. Especially, Torri et al. (2019) examined performance of the global minimum variance portfolio constructed by graphical LASSO (glasso), Student's t-based graphical LASSO (tlasso), random matrix theory filtering (Bouchaud and Potters (2009)), Ledoit-Wolf shrinkage estimation (Ledoit and Wolf (2004)), the conventional sample covariance approach and the equal weight approach in long-term portfolio management with US stock return data. Though Torri et al. (2019)'s research was innovative, its scope of study was limited in case of $p < n$. Moreover, Torri et al. (2019) only tested non-Bayesian graphical models and Bayesian models were not included in comparison. Thus, we develop a data-driven portfolio framework based on a Bayesian graphical LASSO model proposed in Chapter 2, and try to construct the global minimum variance portfolio in case of $p > n$.

In the empirical study, we constructed the global minimum variance portfolio of 100 assets for different sample lengths with the proposed Bayesian approach, variations of non-Bayesian graphical LASSO (graphical LASSO with both diagonal and off-diagonal elements shrinkage, graphical LASSO with only off-diagonal elements shrinkage), random matrix theory filtering, Ledoit-Wolf shrinkage estimation, the conventional sample covariance approach, and the equal weight approach as a benchmark, and compared their out-of-sample performance in 10-year portfolio management from 2011 to 2020. We used monthly return data on 100 portfolios of US companies formed on size and book-to-market ratios provided by Kenneth French and test five scenarios: (p, n) = (100, 120), (100, 60), (100, 12), (100, 6) and (100, 3), which were corresponding to the sample period of 10 years, 5 years, 1 year, 2 quarters, and 1 quarter respectively. Each portfolio was rebalanced once every three months, and it was assumed that there were no short selling restrictions and no trading fees assumed for the sake of simplicity. In this experiment, we confirmed advantages of the proposed approach over the others in terms of return-risk tradeoff and portfolio composition. Both Sharpe ratios and indices of portfolio composition were relatively stable for the proposed approach while they are either unstable for non-Bayesian graphical LASSO approaches. Even in the most severe scenario where the precision matrix of 100 assets must be estimated with only 3 observations, the proposed approach was able to estimate the precision matrix and outperformed the equal weight portfolio without taking abnormal values in regard to indices of portfolio composition.

The second challenge is that the normality assumption of the traditional Markowitz's approach for asset returns is not necessarily satisfied for real-world data. Actually it is well-known that they tend to follow a fat-tailed, possibly skewed distribution as Kon (1984), Mills (1995), Markowitz and Usmen (1996), Peiró (1999) among others have pointed out. Thus, researchers have proposed numerous distributions that can express these characteristics of asset returns well. In particular, a so-called skew-t distribution is often assumed for asset return since Hansen (1994) first used it for modeling financial data. There are many types of skew-t distribution known in the literature, but arguably the most famous one is generalized hyperbolic (GH) skew-t distribution (Hansen (1994), Fernández and Steel (1998) and Aas and Haff (2006)) as a special case of the GH distribution originally proposed by Barndorff-Nielsen (1977). Especially, application of the GH distribution has been recently advanced in the field of asset price volatility models. Although the GH distribution is flexible enough to model a single asset on many occasions, it has difficulty in capturing the skewness dependency among multiple assets. Fund managers would find the skewness dependency useful in particular when the financial market crashes and almost all assets suddenly go south since such sharp price co-movement may not be captured by the second moment (i.e., correlation) only.

In Chapter 4, we examine a skew-elliptical distribution which is another type of distribution that can express these characteristics of asset return to circumvent the aforementioned shortcoming of the GH distribution. The skew-elliptical distribution was proposed by Branco and Dey (2001) as a generalization of the multivariate skew-normal distribution by Azzalini and Valle (1996) and later improved by Sahu et al. (2003). Unlike the GH distribution, it is straightforward to extend the skew-elliptical distribution to the multivariate case. The multivariate skew-normal distribution has another advantage: its Bayesian estimation can be conducted via pure Gibbs sampling. For example, Sahu et al. (2003) proposed a Gibbs sampler for a linear regression model in which the error term follows a skew-elliptical distribution without skewness dependency. Moreover, Harvey et al. (2010) improved Sahu et al. (2003)'s method, and applied it to Bayesian estimation of the multivariate skew-normal distribution as well as portfolio optimization that considers up to the third moment in the presence of skewness dependency. Harvey et al. (2010)'s research is considered to be a model that is of great interest to researchers because it is frequently cited in papers on portfolio selection with the downward risk.

In our assessment, however, the Bayesian estimation method of the multivariate skew-elliptical distribution by Harvey et al. (2010) has an identification issue about the skewness parameters due to so-called label switching. Precisely speaking, a summation in each element of matrix multiplication of the latent variable and the skewness matrix in Harvey et al. (2010)'s model is invariant in terms of permutation, the likelihood of the model takes the same value for any permutations of the columns in the skewness matrix. As a result, it is likely that the columns of the skewness matrix are randomly misaligned during the Gibbs sampler and their interpretability is lost. This problem is well-known in the field of latent factor models.

To solve the issue, we propose a modified model in which the lower-triangular constraint (e.g., Geweke and Zhou (1996), West (2003) and Lopes and West (2004)) is imposed upon the skewness matrix. Moreover, we devise an extended model with the horseshoe prior for both skewness matrix and precision matrix to further improve the estimation accuracy. In the simulation study, we compared the proposed models with the model of Harvey et al. (2010) in three structural designs of the skewness matrix; Diag, Sparse, and Dense. The results show that the proposed models with the identification constraint significantly improved the estimation accuracy of the skewness matrix.

In Chapter 5, as concluding remarks of this dissertation, we review key points of the thesis and comment on a future development.

Acknowledgments

First, I would like to sincerely thank my supervisor, Professer Teruo Nakatsuma. His enthusiastic research guidance based on his overwhelming knowledge of finance, Bayesian statistics, computation and econometrics, kind help and encouragement have been indispensable to my research activities at Graduate School of Economics, Keio University.

I am deeply grateful to my doctoral preliminary exam committee members, Professor Takahiro Hoshino and Professor Daisuke Nagakura, for their time and support. Also, I thank my colleague graduate students and members in the *Econometrics Seminar* for their discussion and providing hints of research about econometrics, finance, marketing, and Bayesian statistics.

I thank Japan Student Services Organization (JASSO) for providing me interest-free scholarship to go on to graduate school. Its financial support eased my financial strain and made it possible for me to continue research activities at the Graduate School of Economics, Keio University. I am also thankful to Keio University Doctorate Student Grant-in-Aid Program from Ushioda Memorial Fund which financially supported me in 2020 and 2021. Its support has been essential for creating a sufficient research environment.

Finally, I thank my family members for their love and enthusiastic support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This doctoral dissertation attempts to address two challenges in modeling asset return data; (1) $p > n$ problem in case of high-dimensional asset return data (p: the number of assets, n: the number of data points) and (2) modeling of a skewed and fat-tailed distribution in real-world data. To deal with these challenges, we will discuss new statistical modeling techniques with Bayesian Markov Chain Monte Carlo (MCMC) and their applications to real-world data in this doctoral dissertation.

As for the $p > n$ problem, the Bayesian graphical model may be a promising solution. However, a popular Bayesian MCMC estimation method for the graphical model by Wang (2012) has a serious problem that the positive definiteness of the precision matrix in the model is not always assured in the Gibbs sampler. This issue is not yet addressed in the literature and the proposed solution (positive-definiteness-assured Gibbs sampler) is one of the main contributions of this dissertation. In Chapter 2, we will show that the proposed algorithm can significantly improve the estimation accuracy of the precision matrix.

Furthermore, in Chapter 3, we propose a data-driven portfolio management approach based on the proposed algorithm. In long-term portfolio management experiments with real data of asset returns, we will show the superiority of the proposed approach over non-Bayesian graphical models and other dimension compression methods.

The summaries of Chapter 2 and 3 are given as follows.

- Chapter 2, "*A Positive-definiteness-assured Block Gibbs Sampler for Bayesian Graphical Models with Shrinkage Priors*"

This research focuses on the development of block Gibbs sampler for the Bayesian graphical LASSO originally proposed by Wang (2012), which has been widely applied and extended to various shrinkage priors in recent years. Our contribution is that we discover that Wang (2012)'s algorithm has a less noticeable but severe disadvantage that the positive definiteness of the precision matrix in the Gaussian graphical model is not guaranteed in each cycle of the Gibbs sampler. Specifically, if the dimension of the precision matrix exceeds the sample size, the positive definiteness of the precision matrix will be barely satisfied and the Gibbs sampler will almost surely fail. In this research, we propose modifying the original block Gibbs sampler so that the precision matrix never fails to be positive definite by sampling it exactly from the domain of the positive definiteness. As we have shown in the Monte Carlo experiments, this modification not only stabilizes the sampling procedure but also significantly improves the performance of the parameter estimation and graphical structure learning. We also apply our proposed algorithm to a graphical model of the monthly return data in which the number of stocks exceeds the sample period, demonstrating its stability and scalability. This chapter is based on Oya and Nakatsuma (2022).

- Chapter 3, "*A Bayesian Graphical Approach for Large-Scale Portfolio Management with Fewer Historical Data*"

In the field of finance, managing a large-scale portfolio with many assets is one of the most challenging tasks in the field of finance. It is partly because estimation of either covariance or precision matrix of asset returns tends to be unstable or even infeasible when the number of assets p exceeds the number of observations n. For this reason, most of the previous studies on portfolio management have focused on the case of $p < n$. To deal with the case of $p > n$, we propose to use a new Bayesian framework based on adaptive graphical LASSO for estimating the precision matrix of asset returns in a large-scale portfolio. Unlike the previous studies on graphical LASSO in the literature, our approach utilizes a Bayesian estimation method for the precision matrix proposed by Oya and Nakatsuma (2022) so that the positive definiteness of the precision matrix should be always guaranteed. As an empirical application, we construct the global minimum variance portfolio of $p = 100$ for various values of n with the proposed approach as well as the non-Bayesian graphical LASSO approach, and compare their out-of-sample performance with the equal weight portfolio as the benchmark. We also compare them with portfolios based on random matrix theory filtering and Ledoit-Wolf shrinkage estimation which were used by Torri et al. (2019). In this comparison, the proposed approach produces more stable results than the non-Bayesian approach and the other comparative approaches in terms of Sharpe ratio, portfolio composition and turnover even if n is much smaller than p. This chapter is based on Oya (2022).

Regarding the second challenge, modeling of a skewed and fat-tailed distribution, we conducted research on the parameter estimation of a multivariate skew-elliptical distribution, which can flexibly capture the skewness dependency among assets. In this research, we discovered an error in a Bayesian parameter estimation model of a previous study and constructed a revised one. In addition, we also proposed an extended model with a shrinkage prior for the key parameters and showed that the proposed models could improve the estimation accuracy of the skewness matrix and the precision matrix in a simulation study. This research is presented in Chapter 4.

- Chapter 4, "*Identification in Bayesian Estimation of the skewness Matrix in a Multivariate Skew-Elliptical Distribution*"

This research focuses on the Bayesian estimation method for a multivariate skew-elliptical distribution, one of asymmetric fat-tailed distributions. Harvey et al. (2010) extended the Bayesian estimation method by Sahu et al.

(2003) to the multivariate skew-elliptical distribution with a general skewness matrix, and applied it to Bayesian portfolio optimization with higher moments. Although their method is epochal in the sense that it can handle the skewness dependency among asset returns and incorporate higher moments into portfolio optimization, it cannot identify all elements in the skewness matrix due to label switching in the Gibbs sampler. To deal with this identification issue, we propose to modify their sampling algorithm by imposing a positive lower-triangular constraint on the skewness matrix of the multivariate skew-elliptical distribution and improved interpretability. Furthermore, we propose a Bayesian sparse estimation of the skewness matrix with the horseshoe prior to further improve the accuracy. In the simulation study, we demonstrate that the proposed method with the identification constraint can successfully estimate the true structure of the skewness dependency while the existing method suffers from the identification issue. This chapter is based on Oya and Nakatsuma (2021).

Finally, we provide summary review of the thesis and remarks on a direction of the future research in Chapter 5.

Chapter 2

A Positive-Definiteness-Assured Block Gibbs Sampler for Bayesian Graphical Models with Shrinkage Priors

2.1 Introduction

Suppose $\boldsymbol{Y}$ is an $(n \times p)$ data matrix of p variables and n observations and the t-th row vector of $\boldsymbol{Y}$, $\boldsymbol{y}_t$ $(1 \leqq t \leqq n)$, follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega} = (\omega_{ij})$, $(1 \leqq i, j \leqq p)$ is the inverse of the covariance matrix, called the precision matrix. In the multivariate normal distribution, $\omega_{ij} = 0$ implies that y_{ti} and y_{tj} are independent. Therefore, a set of nonzero off-diagonal elements in $\boldsymbol{\Omega}$ constitutes an undirected graphical structure among $(y_{t1}, \dots, y_{tp})$ that is called the Gaussian graphical model.

We may estimate $\boldsymbol{\Omega}$ by maximizing the log likelihood:

$$\ell(\boldsymbol{\Omega}) = -\frac{np}{2}\log 2\pi + \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{S}\boldsymbol{\Omega}\right), \tag{2.1}$$

where $\boldsymbol{S} = (s_{ij}) = \boldsymbol{Y}^\intercal \boldsymbol{Y}$. In practice, however, a maximum likelihood estimator (MLE) with (2.1) does not produce estimates of off-diagonal ω_{ij}'s that are exactly equal to zero. To obtain "zero estimates" of ω_{ij}'s, we may employ a LASSO-type penalized MLE:

$$\max_{\boldsymbol{\Omega} \in M^+} \frac{n}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \mathrm{tr}\,(\boldsymbol{S\Omega}) - \lambda \|\boldsymbol{\Omega}\|_1, \tag{2.2}$$

where $\|\boldsymbol{\Omega}\|_1 = \sum_{i \leq j} |\omega_{ij}|$ and M^+ are the subsets of the parameter space of $\boldsymbol{\Omega}$ in which $\boldsymbol{\Omega}$ is a positive definite precision matrix. The solution of (2.2) is called the graphical LASSO estimator, and there have been many research studies on this model in recent years, including those by Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008), and Guo et al. (2011) among others.

Note that the penalty in (2.2) is equivalent to the logarithm of

$$p(\omega_{ij}) = \begin{cases} \lambda e^{-\lambda \omega_{ii}}, & (i = j); \\ \frac{\lambda}{2} e^{-\lambda |\omega_{ij}|}, & (i \neq j). \end{cases} \tag{2.3}$$

From the viewpoint of Bayesian statistics, as in Marlin et al. (2009) and Marlin and Murphy (2009), the graphical LASSO estimator is a maximum a posteriori estimator of $\boldsymbol{\Omega}$ in which the prior distribution of each diagonal element is exponential and that of each off-diagonal element is Laplace as in (2.3). This is a natural extension of the original Bayesian LASSO by Park and Casella (2008) who have extended the LASSO regression by Tibshirani (1996) to a Bayesian counterpart.

Based on this interpretation, Wang (2012) and Khondker et al. (2013) have independently proposed Markov chain sampling algorithms to generate the precision matrix $\boldsymbol{\Omega}$ from its posterior distribution. The difference between Wang (2012) and Khondker et al. (2013) is described in Table 1. The main difference is that Wang (2012) has developed a Gibbs sampling algorithm, while Khondker et al. (2013) have devised a random walk Metropolis-Hastings algorithm. Considering an application of these algorithms to high-dimensional data, Wang (2012)'s algorithm is relatively efficient in that it does not suffer from a low acceptance rate. Moreover, Wang (2012)'s algorithm does not require parameter tuning while Khondker et al. (2013)'s algorithm requires it. Thanks to this feature, Wang (2012)'s algorithm is more scalable compared with Khondker et al. (2013)'s.

Due to these merits of Wang (2012)'s algorithm, it has become an indispensable building block for recent applied research on the Bayesian analysis of Gaussian graphical models. For example, as natural extensions of the block Gibbs sampler, Wang (2015) has extended the original algorithm to a graphical spike-and-slab model, while Li et al. (2019) have applied it to a graphical horseshoe model.

Table 2.1: Difference between the methods of the previous studies and ours

	Khondker et al. (2013)	Wang (2012)	This Paper
Algorithm	Random Walk Metropolis-Hastings	Gibbs Sampler	Gibbs Sampler
Acceptance Rate	low	**1**	**1**
Positive Definiteness of $\boldsymbol{\Omega}$	**Assured**	Insufficient	**Assured**
Parameter Tuning	Required	**Not Required**	**Not Required**
Scalability	△	○	○
Note:	The better features are boldfaced.		

Although Wang (2012)'s algorithm (block Gibbs sampler) and its variants proposed in recent years are nice and elegant, we think that an important point is overlooked in the literature. As shown in Table 2.1, these sampling algorithms cannot sufficiently assure the positive definiteness of the precision matrix $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$ generated with them is not necessarily positive definite. To explain the problem, let us briefly review the block Gibbs sampler, which we will discuss in more detail in Chapter 2.2. Wang (2012)'s block Gibbs sampler generates the i-th diagonal element ω_{ii} and the off-diagonal elements in the i-th column (or row) alternatively in the following fashion.[1]

[1] Although we have simplified the steps here for a brief overview of the algorithm, there

Block Gibbs sampler for the precision matrix

For $i = 1, \dots, p$, repeat **Step 1** to **Step 3**.

Step 1: Partition $\mathbf{\Omega}$ into the i-th diagonal element ω_{ii}, the off-diagonal elements $(\omega_{1i}, \dots, \omega_{i-1,i},\ \omega_{i+1,i}, \dots, \omega_{pi})$, and the rest.

Step 2: Generate $(\omega_{1i}, \dots, \omega_{i-1,i},\ \omega_{i+1,i}, \dots, \omega_{pi})$ from the full conditional posterior distribution.

Step 3: Generate ω_{ii} from the full conditional posterior distribution.

The violation of the positive definiteness of $\mathbf{\Omega}$ occurs because the off-diagonal elements of $\mathbf{\Omega}$ are not generated from M^+ in **Step 2**. To a varying degree, this problem occurs regardless of whether the choice of the prior distribution is LASSO (Wang [2012]), spike-and-slab prior (Wang [2015]), or horseshoe prior (Li et al. [2019]); although, a strong shrinkage prior may somehow offset the lack of positive definiteness. To demonstrate our point, herein, we run Monte Carlo experiments similar to those conducted by Wang (2012). We generate data sets with six different graph structures (AR(1), AR(2), Block, Star, Circle, and Full) and two different dimensions ($p = 30, 100$), and apply the block Gibbs sampler for the Bayesian adaptive LASSO[2] in which the shrinkage parameter λ may differ from element to element in $\mathbf{\Omega}$. The number of iterations in the block Gibbs sampler is 10,000 for each experiment. Thus, if we count every $\mathbf{\Omega}$ that is partially updated from **Step 1** to **Step 3** as distinctive, we have 300,000 ($p = 30$) or 10,000,000 ($p = 100$) replications of $\mathbf{\Omega}$ in one experiment. The results of the Monte Carlo experiments are summarized in Table 2.2. In the case of $p = 30$, violation of the positive definiteness occurs in all designs. In particular, about one quarter of the generated $\mathbf{\Omega}$'s do not satisfy the positive definiteness in the Circle design. In the case of $p = 100$, the violation of the positive definiteness is less severe for some designs, but the ratio of violation is still high (20.51%)

are other steps for sampling the shrinkage parameters. Please see Chapter 2.2 for details.

[2] We have explained the Bayesian adaptive LASSO in Chapter 2.2. In our experience, violation of the positive definiteness occurs whether it is adaptive or not.

Table 2.2: The number of violations in the positive definiteness of $\mathbf{\Omega}$

p	AR(1)	AR(2)	Block	Star	Circle	Full
30	7,644	561	12	27	77,768	14
	(2.55)	(0.19)	(0.00)	(0.01)	(25.88)	(0.00)
100	566	9	0	2,524	205,093	0
	(0.06)	(0.00)	(0.00)	(0.25)	(20.51)	(0.00)

Notes: (a) The number of generated $\mathbf{\Omega}$'s is $p \times 10,000$.
(b) The figures in parentheses are the % ratios.

in the Circle design.

To address this issue, we propose improving Wang (2012)'s block Gibbs sampler so that the generated $\mathbf{\Omega}$ will never fail to be positive definite. Although it seems too intractable to guarantee the positive definiteness of $\mathbf{\Omega}$ in each cycle of the block Gibbs sampler, the hit-and-run algorithm by Bélisle et al. (1993) is applicable to the Bayesian (adaptive) graphical LASSO in a fairly straightforward manner, and the resultant algorithm is a pure Gibbs sampler without the Metropolis-Hastings step. Therefore, our proposed algorithm enjoys the same efficiency as Wang (2012)'s but can prevent $\mathbf{\Omega}$ from violating the positive definiteness. In other words, our proposed algorithm achieves the merits of both Khondker et al. (2013) and Wang (2012) as described in Table 2.1.

The main body of this paper is organized as follows: In Chapter 2.2, we briefly review Wang (2012)'s block Gibbs sampling algorithm for the Bayesian adaptive graphical LASSO, though Wang (2012) has also derived an algorithm for the Bayesian graphical LASSO with the common shrinkage parameter. This is because the core part of the block Gibbs sampling algorithm is almost identical in both prior settings. In Chapter 2.3, we discuss why the positive definiteness of the precision matrix is violated in Wang (2012)'s algorithm and derive a modified Gibbs sampling algorithm that guarantees positive definiteness. In Chapter 2.4, we compare our proposed algorithm with Wang (2012)'s in several Monte Carlo experiments and report the results of the performance comparison. Finally, in Chapter 2.5, we state our concluding remarks.

2.2 Review of Wang (2012)'s Algorithm

In this section, we briefly review a Gibbs sampling algorithm developed by Wang (2012). Although, Wang (2012) derived it for the Bayesian graphical LASSO with the prior distribution (2.3), we consider a more general prior setting that allows λ in (2.3) to vary for each element of precision matrix $\mathbf{\Omega}$, namely

$$p(\omega_{ij}) = \begin{cases} \lambda_{ii} e^{-\lambda_{ii}\omega_{ii}}, & (i = j); \\ \frac{\lambda_{ij}}{2} e^{-\lambda_{ij}|\omega_{ij}|}, & (i \neq j), \end{cases} \tag{2.4}$$

which is called the adaptive graphical LASSO. Here, note that our expression is slightly different from Wang (2012)'s. Wang (2012) assumed that the prior distribution of each diagonal element ω_{ii} is $\frac{\lambda_{ii}}{2}\exp\left(-\frac{\lambda_{ii}}{2}\omega_{ii}\right)$ instead of $\lambda_{ii}\exp\left(-\lambda_{ii}\omega_{ii}\right)$ because Wang (2012) employed $\|\mathbf{\Omega}\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p}|\omega_{ij}|$ as the penalty, in which each off-diagonal element ω_{ij} $(i \neq j)$ appears twice. However, ours is $\|\mathbf{\Omega}\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{i}|\omega_{ij}|$, which includes the lower triangular part of $\mathbf{\Omega}$ only. Since Wang (2012) demonstrated that the Bayesian adaptive LASSO outperforms its nonadaptive counterpart in terms of parameter estimation and graphical structure learning, we will illustrate the Gibbs sampling algorithm for the adaptive LASSO in detail.

To derive the Gibbs sampling algorithm, Wang (2012) utilized the well-known fact that the Laplace distribution in (2.4) is expressed as a scale mixture of normal distributions with the exponential distribution:

$$\omega_{ij}|\tau_{ij} \sim \mathcal{N}(0, \tau_{ij}), \quad \tau_{ij} \sim \mathcal{E}xp\left(\frac{\lambda_{ij}^2}{2}\right). \tag{2.5}$$

By using gamma distribution $\mathcal{G}a(r, s)$ as the common prior for λ_{ij} $(1 \leqq i \leqq j \leqq p)$, we obtain the joint posterior distribution of $\boldsymbol{\omega} = \{\omega_{ij}\}_{i \leqq j}$,

$\boldsymbol{\tau} = \{\tau_{ij}\}_{i<j}$ and $\boldsymbol{\lambda} = \{\lambda_{ij}\}_{i\leqq j}$ as

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\tau}, \boldsymbol{\lambda}|\boldsymbol{Y}) \propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{S\Omega})\right] \prod_{i=1}^{p} \lambda_{ii} e^{-\lambda_{ii}\omega_{ii}} \\ \times \prod_{i<j} \frac{1}{\sqrt{2\pi\tau_{ij}}} \exp\left(-\frac{\omega_{ij}^2}{2\tau_{ij}}\right) \frac{\lambda_{ij}^2}{2} \exp\left(-\frac{\lambda_{ij}^2}{2}\tau_{ij}\right) \mathbf{1}_{M^+}(\boldsymbol{\Omega}) \\ \times \prod_{i\leqq j} \lambda_{ij}^{r-1} e^{-s\lambda_{ij}}, \end{aligned} \tag{2.6}$$

where $\mathbf{1}_{M^+}(\boldsymbol{\Omega})$ is the indicator function that will be equal to 1 if $\boldsymbol{\Omega} \in M^+$; otherwise, it is equal to 0. To construct a Gibbs sampler for the posterior distribution in (2.6), we need to derive all full conditional posterior distributions for $\boldsymbol{\omega}$, $\boldsymbol{\tau}$, and $\boldsymbol{\lambda}$.

It is straightforward to show that the full conditional posterior distribution of $1/\tau_{ij}$ $(1 \leqq i < j \leqq p)$ is the inverse Gaussian distribution:

$$\left.\frac{1}{\tau_{ij}}\right| \boldsymbol{\theta}_{-\tau_{ij}}, \boldsymbol{Y} \sim \mathcal{IG}\left(\frac{\lambda_{ij}}{|\omega_{ij}|}, \lambda_{ij}^2\right), \tag{2.7}$$

while that of λ_{ij} $(1 \leqq i \leqq j \leqq p)$ is the gamma distribution:

$$\lambda_{ij}|\boldsymbol{\theta}_{-\lambda_{ij}}, \boldsymbol{Y} \sim \mathcal{G}a\left(r+1, s+|\omega_{ij}|\right), \tag{2.8}$$

where $\boldsymbol{\theta}$ represents the vector of all parameters and latent variables in the model and expressions such as $\boldsymbol{\theta}_{-x}$ indicate that a parameter x is excluded from $\boldsymbol{\theta}$. Note that τ_{ij} is integrated out in (2.8).

To generate $\boldsymbol{\omega}$ from the full conditional posterior distribution, Wang (2012) proposed a Gibbs sampling algorithm that iteratively generates each diagonal element and the corresponding off-diagonal elements of the precision matrix $\boldsymbol{\Omega}$ from their full conditional posterior distributions, i.e., the block Gibbs sampler. The block Gibbs sampler is based on the following partition of $\boldsymbol{\Omega}$:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^{\intercal} & \omega_{22} \end{bmatrix}, \tag{2.9}$$

where $\boldsymbol{\Omega}_{11}$ is a $(p-1 \times p-1)$ matrix, $\boldsymbol{\omega}_{12}$ is a $(p-1\times 1)$ vector, and ω_{22} is a scalar. Without a loss of generality we can rearrange the rows and columns of $\boldsymbol{\Omega}$, so that the lower-right corner of $\boldsymbol{\Omega}$, ω_{22} is the diagonal element to be generated from its full conditional posterior distribution. Likewise, we can partition $\boldsymbol{S}$, $\boldsymbol{\Upsilon}$, and $\boldsymbol{\lambda}$ as

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^{\mathsf{T}} & s_{22} \end{bmatrix}, \quad \boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Upsilon}_{11} & \boldsymbol{\tau}_{12} \\ \boldsymbol{\tau}_{12}^{\mathsf{T}} & 0 \end{bmatrix}, \quad \boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_{12} \\ \lambda_{22} \end{bmatrix}, \tag{2.10}$$

where $\boldsymbol{\Upsilon}$ is a $(p \times p)$ symmetric matrix in which the off-diagonal (i, j) element is τ_{ij} and all diagonal elements are equal to zero, while λ_{22} is the element in $\boldsymbol{\lambda}$ that corresponds with the diagonal element ω_{22} in the prior distribution (2.4).

With the partition of $\boldsymbol{\Omega}$ in (2.9) and $\boldsymbol{S}$ in (2.10), we have

$$\operatorname{tr}(\boldsymbol{S\Omega}) = s_{22}\omega_{22} + 2\boldsymbol{s}_{12}^{\mathsf{T}}\boldsymbol{\omega}_{12} + \operatorname{tr}(\boldsymbol{S}_{11}\boldsymbol{\Omega}_{11}),$$

and

$$|\boldsymbol{\Omega}| = \left|\omega_{22} - \boldsymbol{\omega}_{12}^{\mathsf{T}}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}\right| |\boldsymbol{\Omega}_{11}|.$$

Then, the likelihood can be expressed as

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\Omega}) &\propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{S\Omega})\right] \\ &\propto \left|\omega_{22} - \boldsymbol{\omega}_{12}^{\mathsf{T}}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}\right|^{\frac{n}{2}} |\boldsymbol{\Omega}_{11}|^{\frac{n}{2}} \\ &\quad \times \exp\left[-\frac{1}{2}\left\{s_{22}\omega_{22} + 2\boldsymbol{s}_{12}^{\mathsf{T}}\boldsymbol{\omega}_{12} + \operatorname{tr}(\boldsymbol{S}_{11}\boldsymbol{\Omega}_{11})\right\}\right]. \end{aligned} \tag{2.11}$$

Wang (2012) reparametrized $(\omega_{22}, \boldsymbol{\omega}_{12})$ to $(\gamma, \boldsymbol{\beta})$, where

$$\gamma = \omega_{22} - \boldsymbol{\omega}_{12}^{\mathsf{T}}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}, \quad \boldsymbol{\beta} = \boldsymbol{\omega}_{\mathbf{12}}. \tag{2.12}$$

Thus, the likelihood (2.11) can be expressed as follows:

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\Omega}) &\propto \gamma^{\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{s_{22}\gamma + s_{22}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta} + 2s_{12}^{\mathsf{T}}\boldsymbol{\beta} + \operatorname{tr}(\boldsymbol{S}_{11}\boldsymbol{\Omega}_{11})\right\}\right] \\ &\propto \gamma^{\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{s_{22}\gamma + s_{22}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta} + 2s_{22}\boldsymbol{\beta}\right\}\right]. \end{aligned} \tag{2.13}$$

With the adaptive prior (2.4) and the flat prior $p(\gamma) \propto$ constant, Wang (2012) proposed using

$$\begin{aligned}
&\boldsymbol{\beta}|\boldsymbol{\theta}_{-\boldsymbol{\beta}}, \boldsymbol{Y} \sim \mathcal{N}\left(-\boldsymbol{C}\boldsymbol{s}_{12},\ \boldsymbol{C}\right), && (2.14)\\
&\boldsymbol{C} = \left\{(s_{22} + 2\lambda_{22})\boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{D}_{\boldsymbol{\tau}}^{-1}\right\}^{-1}, \quad \boldsymbol{D}_{\boldsymbol{\tau}} = \mathrm{diag}(\boldsymbol{\tau}_{12}), &&\\
&\gamma|\boldsymbol{\theta}_{-\gamma}, \boldsymbol{Y} \sim \mathcal{G}a\left(\frac{n}{2} + 1,\ \frac{s_{22}}{2} + \lambda_{22}\right) && (2.15)
\end{aligned}$$

as the full conditional posterior distribution of γ and $\boldsymbol{\beta}$.

In summary, Wang's (2012) block Gibbs sampler is given as follows:[3]

Block Gibbs sampler for all parameters

```
For $i = 1, \ldots, p$, repeat Step 1 to Step 5.
    Step 1: Rearrange $\boldsymbol{\Omega}$, $\boldsymbol{S}$, $\boldsymbol{\Upsilon}$, and $\boldsymbol{\lambda}$ so that $\omega_{ii}$ is in the place of $\omega_{22}$ in $\boldsymbol{\Omega}$
        and partition them as in (2.9) and (2.10).
    Step 2: If $i \geqq 2$, $\boldsymbol{\beta} \leftarrow \mathcal{N}(-\boldsymbol{C}\boldsymbol{s}_{12}, \boldsymbol{C})$, and set $\boldsymbol{\omega}_{12} = \boldsymbol{\beta}$.
    Step 3: $\gamma \leftarrow \mathcal{G}a\left(\frac{n}{2} + 1, \frac{s_{22}}{2} + \lambda_{22}\right)$, and set $\omega_{22} = \gamma + \boldsymbol{\omega}_{12}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}$.
    Step 4: $\lambda_{12} \leftarrow \mathcal{G}a(r + 1, s + |\boldsymbol{\omega}_{12}|)$.
    Step 5: $\upsilon \leftarrow \mathcal{IG}\left(\frac{\lambda_{12}}{|\boldsymbol{\omega}_{12}|}, \lambda_{12}^2\right)$, and set $\tau_{12} = 1/\upsilon$.
```

2.3 Proposed Algorithm

As we pointed out in the introduction, Wang (2012)'s block Gibbs sampler does not necessarily guarantee the positive definiteness of the generated $\boldsymbol{\Omega}$'s. Therefore, in this section, we propose an efficient sampling method to generate them under the positive definiteness constraint: $\boldsymbol{\Omega} \in M^{+}$.

First, let us derive the full conditional posterior distribution of γ. Here, we need to take care in choosing the prior distribution of (γ, β). Given that

[3] In Wang's (2012) study, *Step 4* and *Step 5* are calculated together outside the for loop, but since there is no essential difference, they are shown in the for loop here.

Ω from the previous iteration of the block Gibbs sampler is positive definite, the newly generated ω_{22} and $\boldsymbol{\omega}_{12}$ must satisfy

$$\omega_{22} > \boldsymbol{\omega}_{12}^{\intercal}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12} \tag{2.16}$$

to ensure that the updated $\boldsymbol{\Omega}$ is also positive definite. This condition (2.16) requires

$$\gamma = \omega_{22} - \boldsymbol{\beta}^{\intercal}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta} > 0.$$

as the prior distribution of γ. In other words, the conditional prior distribution of γ given $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}_{11}$ must be

$$p(\gamma|\boldsymbol{\beta}, \boldsymbol{\Omega}_{11}) \propto \lambda_{22} \exp\left(-\lambda_{22}\gamma\right) \mathbf{1}_{M_{\gamma}^{+}}(\gamma), \tag{2.17}$$

where $M_{\gamma}^{+} = \{\gamma : \gamma > 0\}$. Therefore, by ignoring the parts that do not depend on γ in (2.11), we obtain

$$\begin{aligned} & p(\gamma|\boldsymbol{\theta}_{-\gamma}, \boldsymbol{Y}) \\ & \quad \propto |\gamma|^{\frac{n}{2}} \exp\left(-\frac{s_{22}}{2}\gamma\right) \times \exp\left(-\lambda_{22}\gamma\right) \mathbf{1}_{M_{\gamma}^{+}}(\gamma) \\ & \quad \propto |\gamma|^{\frac{n}{2}} \exp\left[-\frac{s_{22} + 2\lambda_{22}}{2}(\gamma)\right] \mathbf{1}_{M_{\gamma}^{+}}(\gamma). \end{aligned} \tag{2.18}$$

The full conditional posterior distribution of γ in (2.18) is the gamma distribution:

$$\gamma|\boldsymbol{\theta}_{-\gamma}, \boldsymbol{Y} \sim \mathcal{G}a\left(\frac{n}{2} + 1,\ \frac{s_{22}}{2} + \lambda_{22}\right) \tag{2.19}$$

Obviously, the distribution of γ in (2.19) is equivalent to that in (2.15). Thus, (2.19) and (2.15) are basically identical to each other, and γ generated from either (2.19) or (2.15) always satisfies the positive definiteness condition (2.16) because random variables generated from the gamma distribution always have positive values.

Next, let us derive the full conditional posterior distribution of $\boldsymbol{\beta}$. For the same reason as in (2.17), the conditional prior distribution of $\boldsymbol{\beta}$ must be the following truncated multivariate normal distribution:

$$p(\boldsymbol{\beta}|\gamma, \boldsymbol{\Omega}_{11}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\intercal}\boldsymbol{D}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\beta}\right) \mathbf{1}_{M_{\beta}^{+}}(\boldsymbol{\beta}), \tag{2.20}$$

where $M_{\beta}^{+} = \{\boldsymbol{\beta} : \omega_{22} > \boldsymbol{\beta}^{\intercal}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta}\}$. As a result, the full conditional posterior distribution of $\boldsymbol{\beta}$ is also a truncated multivariate normal distribution:

$$\boldsymbol{\beta}|\boldsymbol{\theta}_{-\boldsymbol{\beta}}, \boldsymbol{Y} \sim \mathcal{N}\left(-\boldsymbol{C}\boldsymbol{s}_{12},\ \boldsymbol{C}\right)\mathbf{1}_{M_{\beta}^{+}}(\boldsymbol{\beta}). \tag{2.21}$$

However, Wang (2012) proposed using the unconstrained multivariate normal distribution (2.14), which does not impose the truncation $\mathbf{1}_{M_{\beta}^{+}}(\boldsymbol{\beta})$, to generate $\boldsymbol{\beta}$. Consequently, if we generate $\boldsymbol{\beta}$ from (2.14), there is no guarantee that the newly updated $\boldsymbol{\omega}_{12}$ will satisfy the positive definiteness condition (2.16). This is why generated $\boldsymbol{\Omega}$'s are not always positive definite, as shown in Table 2. Therefore, to ensure the positive definiteness of $\boldsymbol{\Omega}$, it is preferable to use the truncated multivariate normal distribution (2.21) in the block Gibbs sampler.

Since both the naive rejection method and Metropolis-Hastings algorithm are inefficient, even for a modest-size graphical model, we can apply the hit-and-run algorithm (Bélisle et al. (1993)) to generate $\boldsymbol{\beta}$ from the truncated multivariate normal distribution (2.19).

Hit-and-run algorithm

Step 1: Pick a point $\boldsymbol{\alpha}$ on the unit sphere randomly as $\boldsymbol{\alpha} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}$, $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

Step 2: Generate a random scalar κ from the distribution with the density

$$f(\kappa) \propto p(\boldsymbol{\beta} + \kappa\boldsymbol{\alpha})\mathbf{1}_{M_{\beta}^{+}}(\boldsymbol{\beta} + \kappa\boldsymbol{\alpha}), \tag{2.22}$$

where $p(\cdot)$ is the density of $\mathcal{N}\left(-\boldsymbol{C}\boldsymbol{s}_{12}, \boldsymbol{C}\right)$ in (2.21).

Step 3: Set $\boldsymbol{\beta} + \kappa\boldsymbol{\alpha}$ as the new $\boldsymbol{\beta}$.

It is straightforward to show that the distribution of κ in (2.22) is

$$\kappa \sim \mathcal{N}\left(\mu_{\kappa}, \sigma_{\kappa}^{2}\right)\mathbf{1}_{M_{\beta}^{+}}(\boldsymbol{\beta} + \kappa\boldsymbol{\alpha}), \tag{2.23}$$

where

$$\mu_{\kappa} = -\frac{\boldsymbol{s}_{12}^{\intercal}\boldsymbol{\alpha} + \boldsymbol{\beta}^{\intercal}\boldsymbol{C}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\intercal}\boldsymbol{C}^{-1}\boldsymbol{\alpha}}, \quad \sigma_{\kappa}^{2} = \frac{1}{\boldsymbol{\alpha}^{\intercal}\boldsymbol{C}^{-1}\boldsymbol{\alpha}}.$$

The indicator function $\mathbf{1}_{M_\beta^+}(\boldsymbol{\beta}+\kappa\boldsymbol{\alpha})$ is equal to 1 if and only if

$$(\boldsymbol{\beta}+\kappa\boldsymbol{\alpha})^\intercal\boldsymbol{\Omega}_{11}^{-1}(\boldsymbol{\beta}+\kappa\boldsymbol{\alpha})-(\gamma+\boldsymbol{\beta}^\intercal\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta})<0.$$

This means that κ must satisfy

$$\underbrace{\left(\boldsymbol{\alpha}^\intercal\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\alpha}\right)}_{a}\kappa^2+2\underbrace{\left(\boldsymbol{\beta}^\intercal\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\alpha}\right)}_{b}\kappa+\underbrace{(-\gamma)}_{c}<0.$$

Note that $a>0$, $c<0$ as long as the current $\boldsymbol{\Omega}$ is positive definite, which implies that the quadratic equation $a\kappa^2+2b\kappa+c=0$ has two distinctive real roots. Therefore, the distribution in (2.23) is the truncated univariate normal distribution on the interval:

$$R^+=\left\{\kappa:\frac{-b-\sqrt{b^2-ac}}{a}<\kappa<\frac{-b+\sqrt{b^2-ac}}{a}\right\}.$$

Thus, using the hit-and-run algorithm, sampling from the seemingly intractable distribution (2.19) is reduced to sampling from the truncated univariate normal distribution:

$$\kappa\sim\mathcal{N}\left(\mu_\kappa,\sigma_\kappa^2\right)\mathbf{1}_{R^+}(\kappa),$$

and the sampling procedure becomes much simpler.

By replacing (2.15) in **Step 2** with (2.19) and (2.14) in **Step 3** with the hit-and-run algorithm, we obtain the modified block Gibbs sampler as follows:

Modified block Gibbs sampler

```
For $i = 1, \dots, p$, repeat Step 1 to Step 5.

Step 1: Rearrange $\boldsymbol{\Omega}$, $\boldsymbol{S}$, $\boldsymbol{\Upsilon}$, and $\boldsymbol{\lambda}$ so that $\omega_{ii}$ is in the place of $\omega_{22}$ in $\boldsymbol{\Omega}$
        and partition them as in (2.9) and (2.10).

Step 2: If $i \geqq 2$,
        (a) $\boldsymbol{z} \leftarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and set $\boldsymbol{\alpha} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}$.
        (b) $\kappa \leftarrow \mathcal{N}\left(\mu_\kappa, \sigma_\kappa^2\right) \mathbf{1}_{R^+}(\kappa)$, and update the old $\boldsymbol{\beta}$ with $\boldsymbol{\beta} + \kappa\boldsymbol{\alpha}$.
            Then, set $\boldsymbol{\omega}_{12} = \beta$.

Step 3: $\gamma \leftarrow \mathcal{G}a\left(\frac{n}{2} + 1, \frac{s_{22}}{2} + \lambda_{22}\right)$ and set $\omega_{22} = \gamma + \boldsymbol{\omega}_{12}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}$.

Step 4: $\lambda_{12} \leftarrow \mathcal{G}a\left(r + 1, s + |\boldsymbol{\omega}_{12}|\right)$.

Step 5: $\upsilon \leftarrow \mathcal{IG}\left(\frac{\lambda_{12}}{|\boldsymbol{\omega}_{12}|}, \lambda_{12}^2\right)$, and set $\tau_{12} = 1/\upsilon$.
```

Since Wang (2012)'s algorithm and our proposed algorithm are Gibbs sampler, there should be no problem in calculating the posterior distribution even if the order of the steps is changed. In fact, Wang (2012) sampled *Step 4* and *Step 5* first in the code disclosed before. However, we confirmed that Wang's code disclosed before does not work if we swap *Step 2* and *Step 3*. This implies that Wang (2012)'s algorithm cannot sample from the correct posterior distribution. In contrast, we confirmed that our algorithm proposed in Chapter 2.3 works even if we exchange *Step 2* and *Step 3*.

2.4 Performance Comparison

2.4.1 Simulation Study

In this section, we report the results of the Monte Carlo experiments to compare our modified block Gibbs sampler with Wang (2012)'s original algorithm in terms of accuracy in the parameter estimation and graphical structure learning. For brevity, we shall refer to Wang (2012)'s original algorithm as the BGS (block Gibbs sampler) and our modified version as the HRS (hit-and-run sampler). Following Wang (2012), we examined the following six different specifications of the Gaussian graphical model in the Monte Carlo experiments:

(a) AR(1): $\sigma_{ij} = 0.7^{|i-j|}$.

(b) AR(2): $\omega_{ii} = 1.0$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.5$, and $\omega_{i,i-2} = \omega_{i-2,i} = 0.25$.

(c) Block: $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$ for $1 \leq i \neq j \leq p/2$, $\sigma_{ij} = 0.5$ for $p/2 + 1 \leq i \neq j \leq 10$, and $\sigma_{ij} = 0.0$ otherwise.

(d) Star: $\omega_{ii} = 1.0$, $\omega_{1,i} = \omega_{i,1} = 0.1$, and $\omega_{ij} = 0.0$ otherwise.

(e) Circle: $\omega_{ii} = 2.0$, $\omega_{i-1,i} = \omega_{i,i-1} = 1.0$, $\omega_{1p} = \omega_{p1} = 0.9$.

(f) Full: $\omega_{ii} = 2.0$, $\omega_{ij} = 1.0$ for $i \neq j$.

Here, σ_{ij} $(1 \leqq i,\ j \leqq p)$ is the (i,j) element of the covariance matrix $\boldsymbol{\Omega}^{-1}$ in the Gaussian graphical model.

The other settings for the Monte Carlo experiments also mirrored Wang (2012)'s. For each model, we generated a sample of $(p \times 1)$ random vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ independently from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1})$. We considered two cases: $(n, p) = (50, 30)$ and $(n, p) = (200, 100)$. Thus, we tried 12 $(= 6 \times 2)$ scenarios in the experiments. The hyperparameters in the prior distribution of λ_{ij} were $r = 10^{-2}$ and $s = 10^{-6}$. For both the BGS and HRS, the number of burn-in iterations were 5,000, and the Monte Carlo sample from the following 10,000 iterations was used in the Bayesian inference.[4] We repeated each simulation scenario 100 times and obtained a set of point estimates of $\boldsymbol{\Omega}$. All computations were implemented on a workstation with 64 GB RAM and a six-core 3.4 GHz Intel Xeon processor using Python 3.6.1. For the

[4] The same simulation design (specifications of $\boldsymbol{\Omega}$, combinations of (n, p), hyperparameters, burn-in iterations, and the size of the Monte Carlo sample) was used in producing the results in Table 2.2.

BGS, we rewrote Wang's disclosed MATLAB code "BayesGLassoGDP.m" into Python and used it. Although not mentioned in Wang (2012)'s study, there is a part that arbitrarily cuts a range of random number generations of λ_{12} and τ_{12} in Wang's disclosed code. We took over this adjustment in our rewritten Python code because the BGS calculation resulted in an error if we excluded the adjustment. The HRS required additional computations because it explicitly imposed the positive definite constraint $\mathbf{\Omega} \in M^+$, but we observed only a modest difference in computation time between the HRS and BGS.

To compare the HRS with the BGS in terms of accuracy in the point estimation of the precision matrix $\mathbf{\Omega}$, we computed two sample loss functions, Stein's loss and the Frobenius norm, as measurements of discrepancy between the point estimate and the true $\mathbf{\Omega}$. Table 2.3 shows the sample median loss (Stein's loss in the upper half, and the Frobenius norm in the lower half) of 100 replications in 12 scenarios for the BGS and HRS. The figures in parentheses are the standard errors. The loss was unanimously and substantially smaller in the HRS than in the BGS. This observation was valid not only for the Circle model, in which the positive definiteness of $\mathbf{\Omega}$ was most frequently violated as shown in Table 2.2, but also for the other models with different graphical structures. Interestingly, the HRS outperformed the BGS even for the Full model in which $\mathbf{\Omega}$ was not sparse and the estimation loss of the graphical LASSO was expected to be much worse. Furthermore, this tendency was unchanged in either the small ($p = 30$) or large ($p = 100$) model. All in all, the results in Table 2.3 suggest that imposing the positive definiteness constraint remarkably improved the accuracy in the point estimation of $\mathbf{\Omega}$ in the Bayesian adaptive graphical LASSO.

To assess the performance of the graphical structure learning, we checked whether the point estimate of $\mathbf{\Omega}$ could successfully restore the true structure from the simulated data. Recall that there was no connection between nodes, e.g., node i and node j ($1 \leqq i,\ j \leqq p$), if $\omega_{ij} = 0$. Like Fan et al. (2009), we used the following rule to determine whether a pair of nodes was connected

Table 2.3: Sample median loss in the point estimation of $\mathbf{\Omega}$

	AR(1)	AR(2)	Block	Star	Circle	Full
Stein's loss						
			$p = 30$			
BGS	1.78	4.28	1.36	1.52	1.73	19.15
	(0.33)	(0.43)	(0.27)	(0.24)	(0.30)	(0.82)
HRS	**0.61**	**0.76**	**0.66**	**0.86**	**0.54**	**13.71**
	(0.18)	(0.18)	(0.18)	(0.18)	(0.15)	(0.55)
			$p = 100$			
BGS	3.01	4.17	2.75	3.79	3.09	70.06
	(0.15)	(0.23)	(0.18)	(0.23)	(0.16)	(0.93)
HRS	**0.50**	**0.56**	**0.53**	**0.90**	**0.47**	**41.98**
	(0.08)	(0.07)	(0.07)	(0.10)	(0.06)	(0.66)
Frobenius norm						
			$p = 30$			
BGS	4.05	2.99	2.19	2.21	2.51	29.60
	(0.53)	(0.17)	(0.36)	(0.29)	(0.43)	(0.06)
HRS	**1.48**	**0.79**	**1.24**	**1.34**	**0.38**	**19.85**
	(0.27)	(0.14)	(0.23)	(0.20)	(0.07)	(0.53)
			$p = 100$			
BGS	4.35	2.33	2.89	3.22	2.62	99.62
	(0.26)	(0.11)	(0.12)	(0.13)	(0.19)	(0.02)
HRS	**1.27**	**0.62**	**1.04**	**1.04**	**0.24**	**47.78**
	(0.11)	(0.04)	(0.07)	(0.08)	(0.03)	(0.44)

Notes: (a) The smaller losses are boldfaced.
(b) The figures in parentheses are the standard errors.

Table 2.4: Accuracy in graphical structure learning

	AR(1)	AR(2)	Block	Star	Circle
Specificity					
			$p = 30$		
BGS	5.97	10.39	7.09	7.17	12.79
HRS	**73.28**	**70.09**	**78.83**	**79.45**	**84.52**
			$p = 100$		
BGS	10.46	20.45	12.92	12.24	28.47
HRS	**93.35**	**92.89**	**94.23**	**95.66**	**98.44**
Sensitivity					
			$p = 30$		
BGS	100.00	99.32	100.00	**96.27**	100.00
HRS	100.00	**100.00**	100.00	91.23	100.00
			$p = 100$		
BGS	100.00	100.00	100.00	100.00	100.00
HRS	100.00	100.00	100.00	100.00	100.00
MCC					
			$p = 30$		
BGS	7.82	12.53	5.19	4.00	11.99
HRS	**46.92**	**52.94**	**35.39**	**49.19**	**61.14**
			$p = 100$		
BGS	5.89	11.20	3.89	6.43	10.86
HRS	**55.58**	**63.60**	**39.32**	**66.22**	**82.42**

Notes: (a) The better results are boldfaced.
(b) The figures are in percentages.

or not:

$$\begin{cases} |\hat{\omega}_{ij}| \geqq 10^{-3} & \text{(node } i \text{ and node } j \text{ are connected);} \\ |\hat{\omega}_{ij}| < 10^{-3} & \text{(node } i \text{ and node } j \text{ are not connected),} \end{cases} \tag{2.24}$$

where $\hat{\omega}_{ij}$ is the point estimate of ω_{ij} computed with the Monte Carlo sample of $\boldsymbol{\Omega}$ that we generated for each scenario with the HRS or BGS. Then, with the estimated graphical structures (100 in total), the accuracy in the graphical structure learning was measured with three criteria: specificity, sensitivity, and the Matthews correlation coefficient (MCC), namely

$$\begin{aligned} \text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{MCC} &= \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FN})(\text{TN}+\text{FN})}}, \end{aligned} \tag{2.25}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively, in the 100 replications.

Table 2.4 reports the calculated criteria for the 12 scenarios. As in Table 2.4, the HRS outperformed the BGS for all scenarios, except for the sensitivity of the Star model with $p = 30$, though the sensitivity of the HRS was still over 90%. Specifically, in the case of $p = 100$, the values of specificity were over 90% for the HRS, which means that most of the zero off-diagonal elements in $\boldsymbol{\Omega}$ were correctly identified. This accuracy is crucial when trying to detect the true graphical structure in practice. It seems that imposing the positive definiteness constraint also enhanced the graphical structure learning in the Bayesian adaptive graphical LASSO.

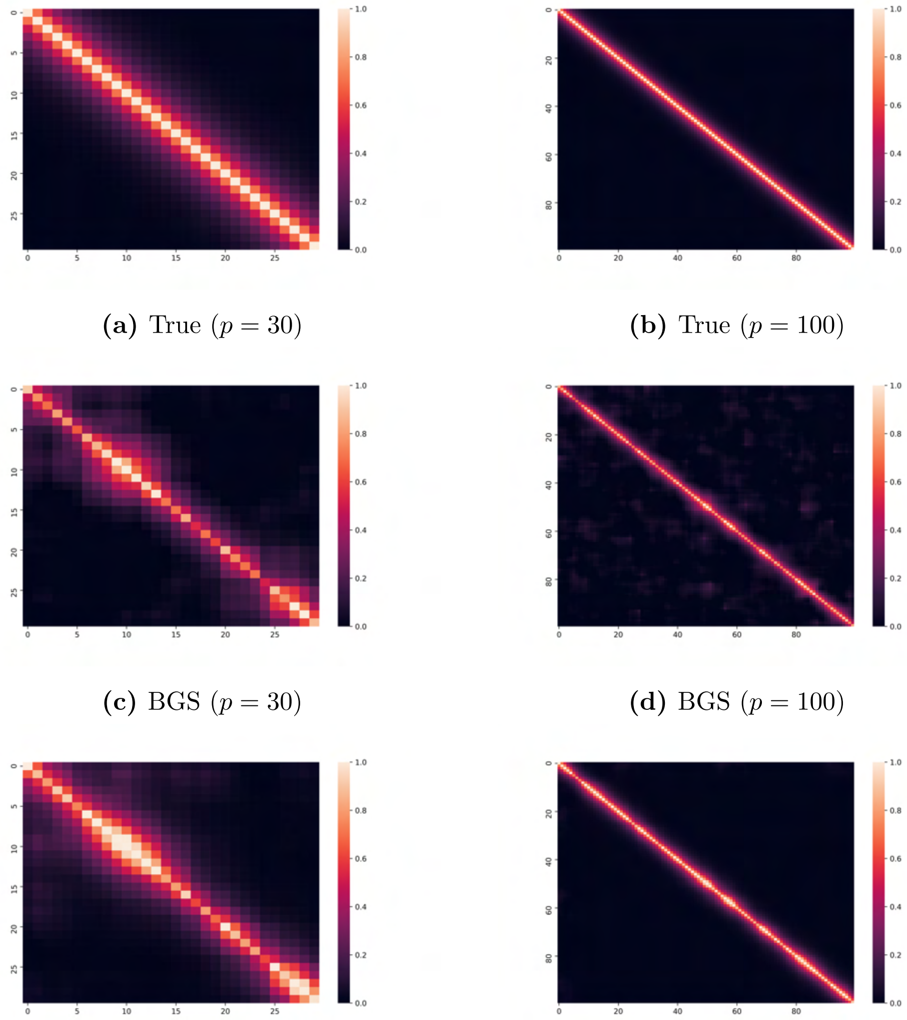
In addition, in order to make the estimation results visually easy to understand, the posterior mean Ω of each scenario in the 50th replication are shown in Figures 2.1 – 2.6. In each figure, left-half shows $p = 30$ case and right-half shows $p = 100$ case. From Figures 2.1 – 2.6, HRS adequately describes the structure of the true Ω, while BGS has a slightly broken structure. As for BGS, non-zero off-diagonal elements that are not in the true structure can be seen. Also, in the scenario of Full structure shown in Figure 2.6, the off-diagonal elements of Ω estimated by BGS shrinks to 0 while the true values are 1. This also implies that the estimation of BGS is not working well.

Finally, let us explain about a reason why the results of the BGS in Table 2.4 are far different from those in Wang (2012)'s Table 2 and BGS's performance is significantly lower than HRS. We assumed that this discrepancy was caused by the difference in the criteria for detecting connections. p882, Wang (2012) stated that "we claim $\{\omega_{ij} = 0\}$ if $\hat{\omega}_{ij} < 10^{-3}$ as Fan et al. (2009)," which means that a negative $\hat{\omega}_{ij}$, whether near or far from 0, is regarded as evidence against a connection between nodes. As a result, negative relations between nodes would be over-rejected and the estimated graphical structure would be too sparse in the sense that the precision matrix would include too many zeros in the off-diagonal elements. To confirm this conjecture, we recalculated the three criteria in (2.25) without the absolute value in (2.24) and found that the recalculated results were comparably similar to those of Wang (2012). In a disclosed Python code of this paper, the three criteria with the absolute value and without the absolute value are implemented for reference. See `https://github.com/oyakeioecon/onglasso` for more details.

2.4.2 Application to S&P500 Stock Return Data

Next, we applied the BGS and HRS to stock return data and estimated Ω. We used the standardized monthly excess return data against the S&P 500 stock index for 483 stocks continuously listed from the end of December 2013 to the end of January 2018 of 505 constituents of the S&P 500 as of February 2018 (n = 50, p = 483). The settings were the same as those for the simulation data.

Although violation of the positive definiteness after updating the off-diagonal elements reached 808,009 times (16.73%) in the BGS, it never occurred in the HRS. Figures 2.7 and 2.8 show the posterior mean of Ω by the BGS and HRS. Here, to make it easier to compare the BGS and HRS, we adjusted the scale of Ω so that the diagonal elements were one. The Ω



(a) True ($p = 30$)

(b) True ($p = 100$)

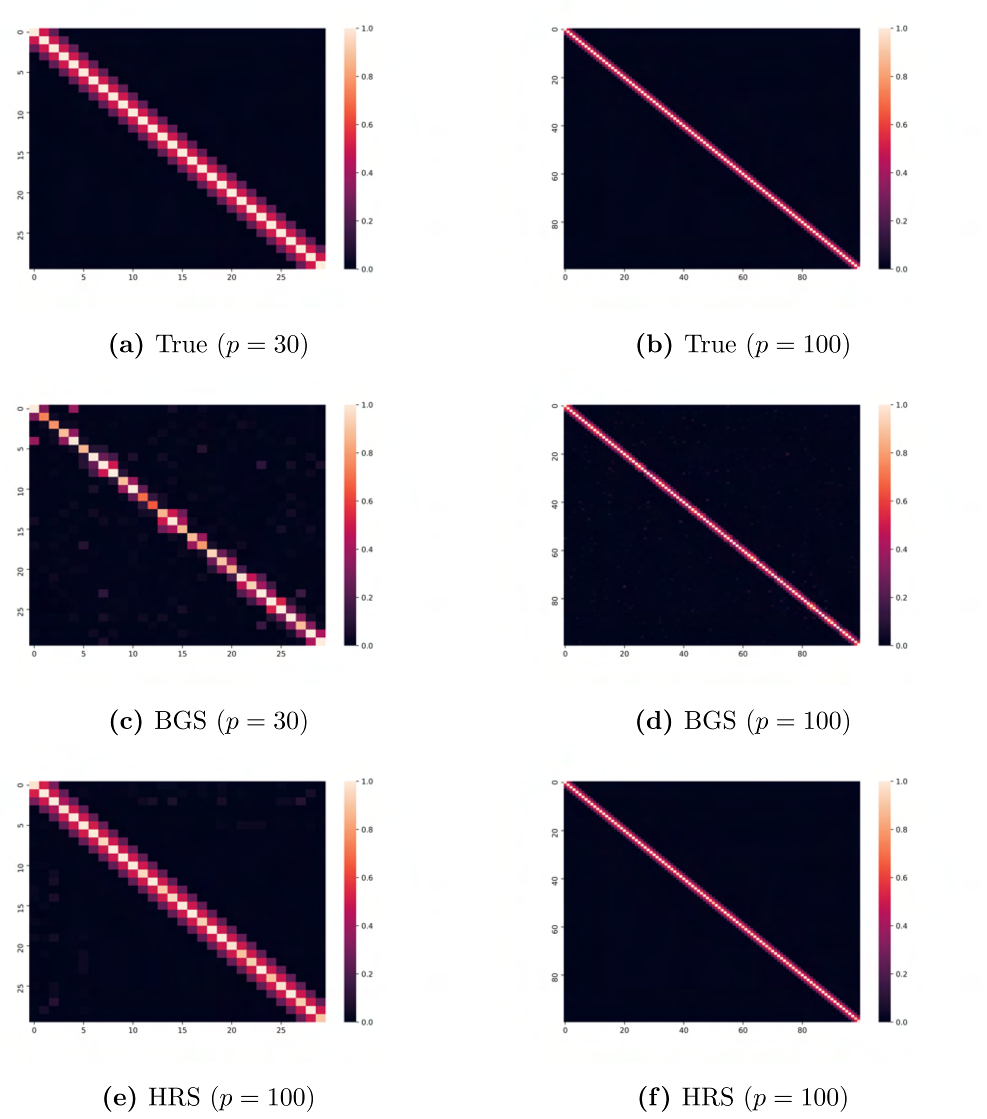
(c) BGS ($p = 30$)

(d) BGS ($p = 100$)

(e) HRS ($p = 30$)

(f) HRS ($p = 100$)

Figure 2.1: AR(1) : True Structure and estimated $\Sigma = \Omega^{-1}$



(a) True ($p = 30$)

(b) True ($p = 100$)

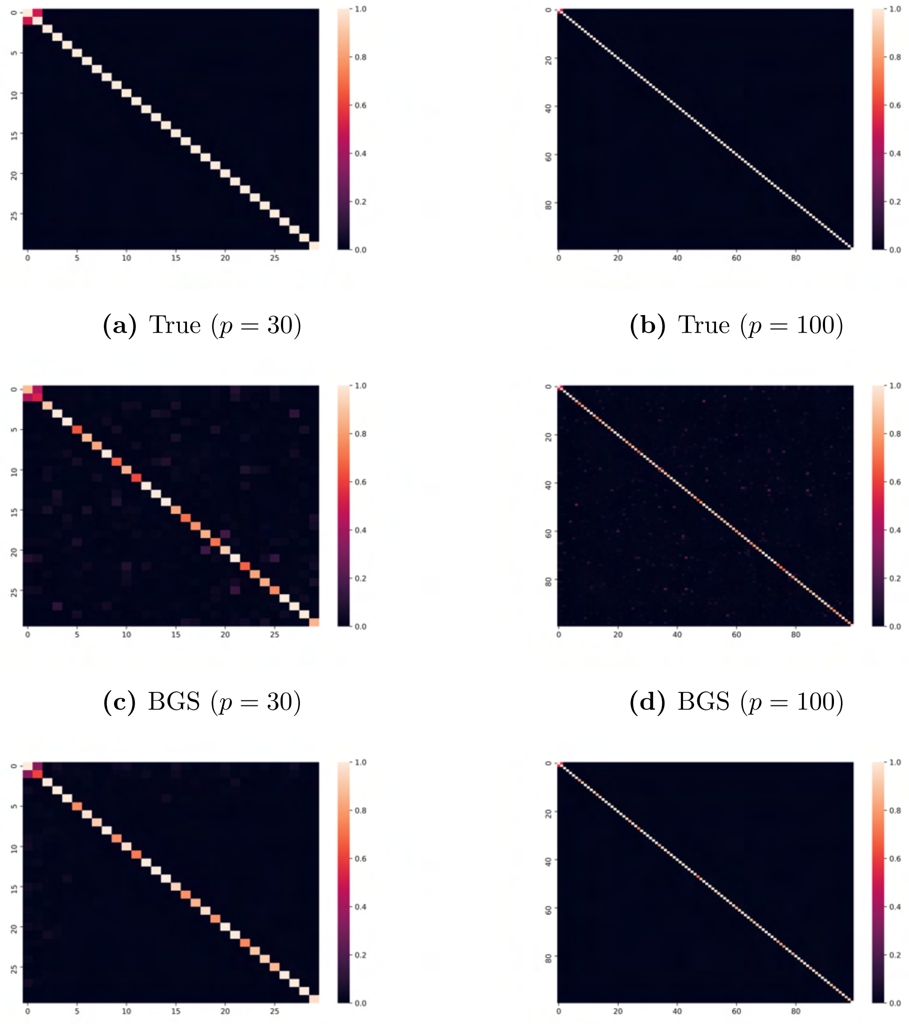
(c) BGS ($p = 30$)

(d) BGS ($p = 100$)

(e) HRS ($p = 100$)

(f) HRS ($p = 100$)

Figure 2.2: AR(2): True Structure and estimated Ω



(a) True $(p = 30)$

(b) True $(p = 100)$

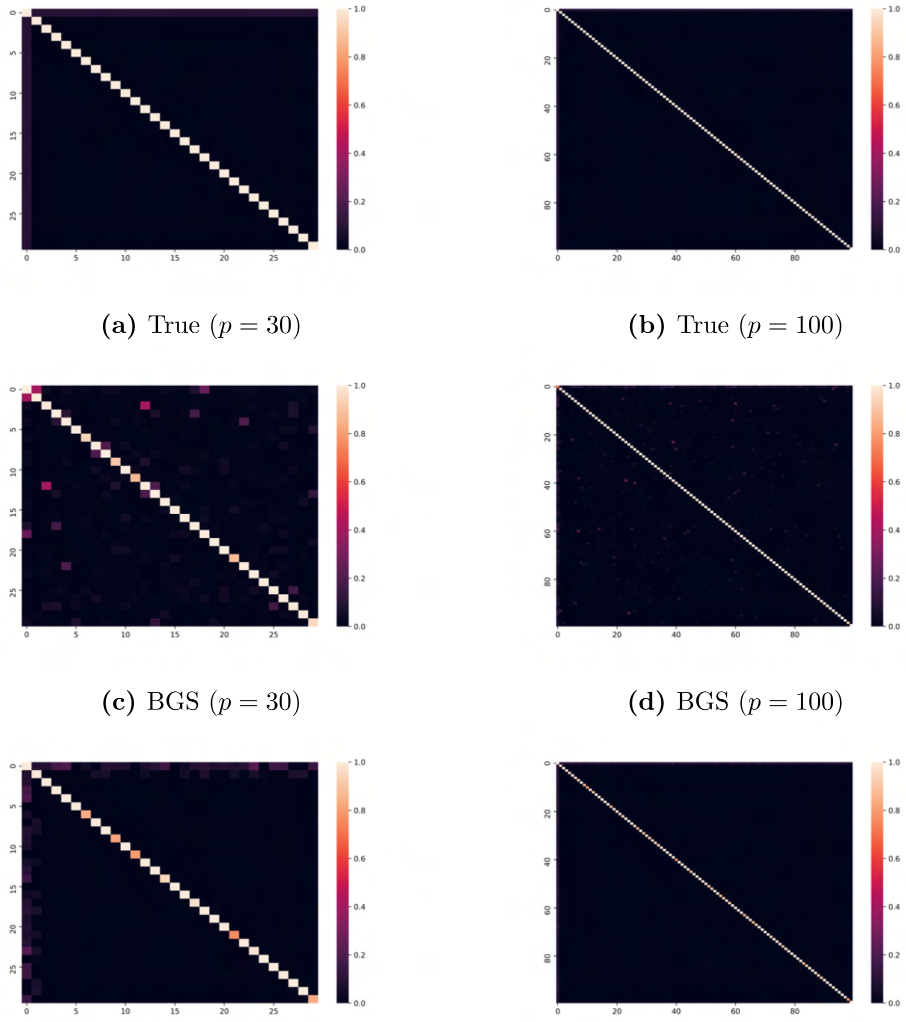
(c) BGS $(p = 30)$

(d) BGS $(p = 100)$

(e) HRS $(p = 30)$

(f) HRS $(p = 100)$

Figure 2.3: Block: True Structure and estimated $\Sigma = \Omega^{-1}$



(a) True ($p = 30$)

(b) True ($p = 100$)

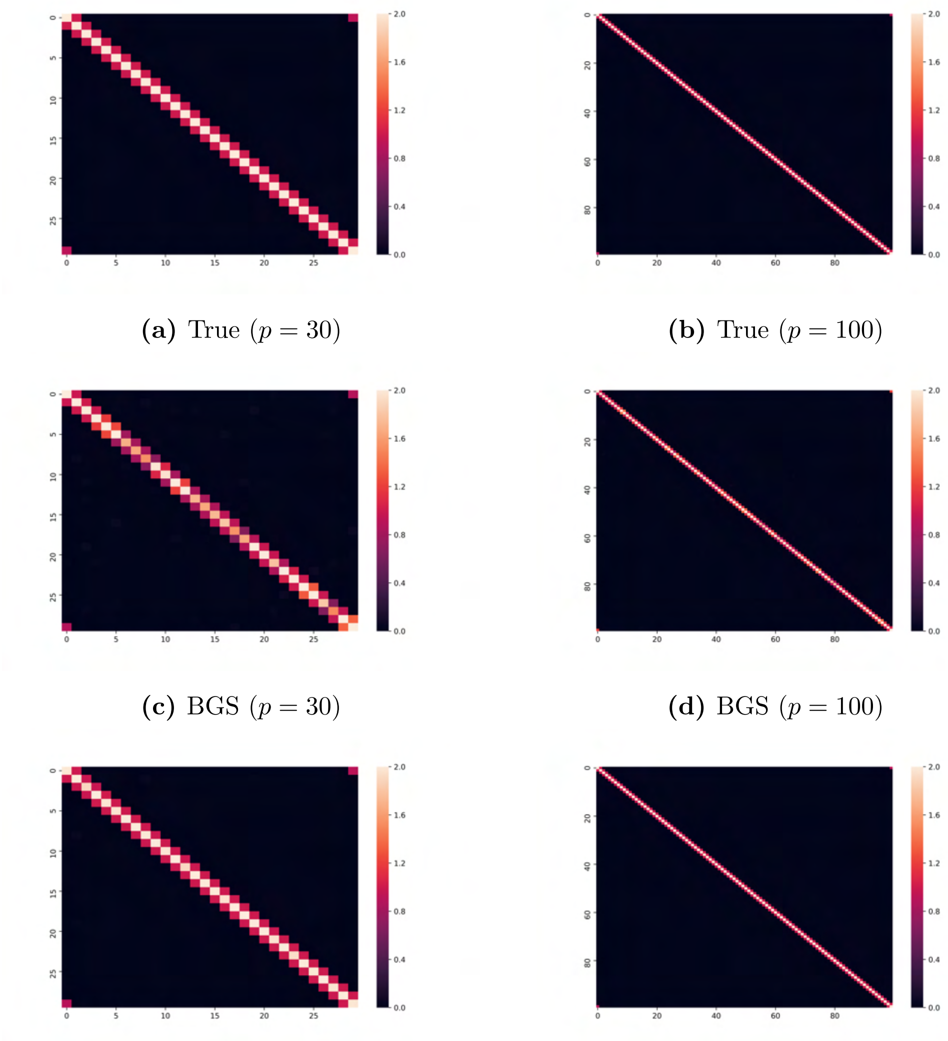
(c) BGS ($p = 30$)

(d) BGS ($p = 100$)

(e) HRS ($p = 100$)

(f) HRS ($p = 100$)

Figure 2.4: Star: True Structure and estimated Ω



(a) True ($p = 30$)

(b) True ($p = 100$)

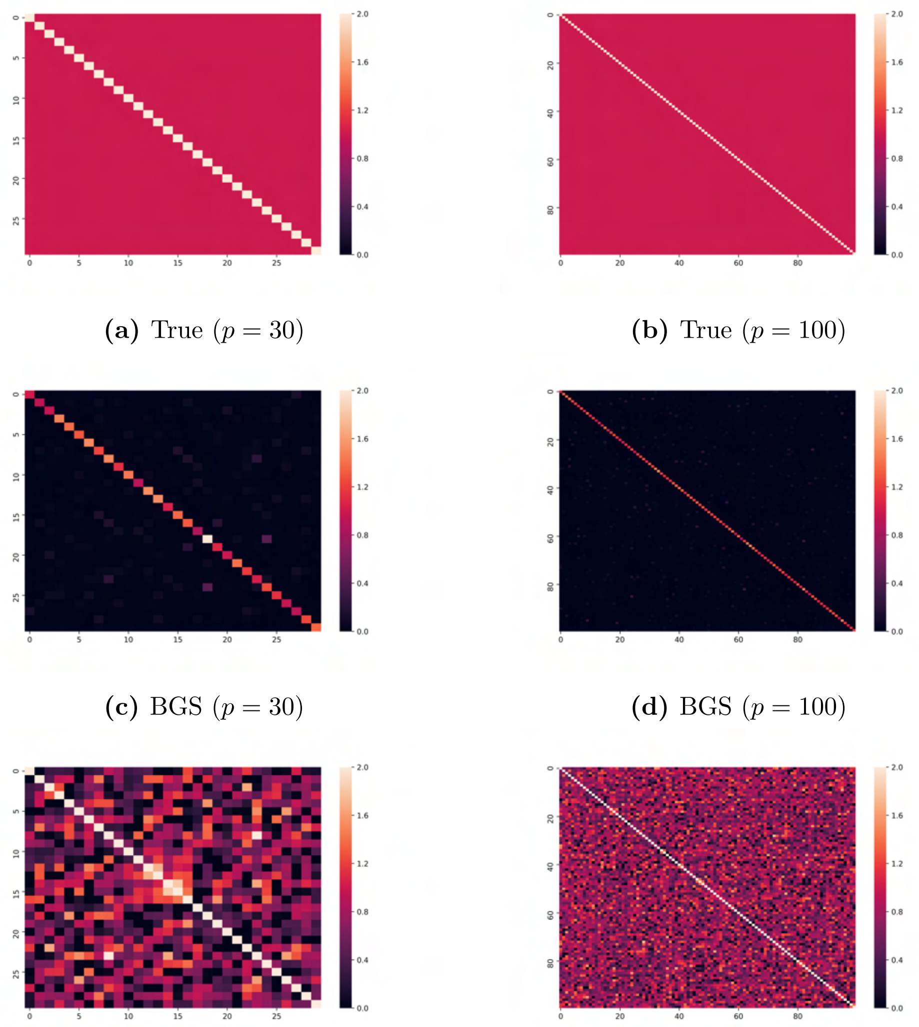
(c) BGS ($p = 30$)

(d) BGS ($p = 100$)

(e) HRS ($p = 100$)

(f) HRS ($p = 100$)

Figure 2.5: Circle: True Structure and estimated Ω



(a) True ($p = 30$)

(b) True ($p = 100$)

(c) BGS ($p = 30$)

(d) BGS ($p = 100$)

(e) HRS ($p = 100$)

(f) HRS ($p = 100$)

Figure 2.6: Full: True Structure and estimated Ω

estimated by the BGS in Figure 2.7 had many nonzero values remaining in the off-diagonal elements, while the off-diagonal elements of Ω estimated by the HRS in Figure 2.8 shrunk.

2.5 Conclusion

In Chapter 2, we proposed a modification of Wang (2012)'s block Gibbs sampling algorithm for the Bayesian graphical LASSO that we used as the primary example. Our modified algorithm guarantees the positive definiteness of the precision matrix throughout the sampling procedure by generating the off-diagonal elements of the precision matrix from a truncated multivariate normal distribution whose support is the region wherein the updated precision matrix remains positive definite. To facilitate sampling from such a complicated distribution, we proposed utilizing the hit-and-run algorithm by Bélisle et al. (1993). The derived algorithm is still a pure Gibbs sampler and maintains the efficiency and scalability of Wang (2012)'s original algorithm. In the simulation study, we showed that our modified algorithm remarkably improved the accuracy in the point estimation and graphical structure learning. We also demonstrated that our modified algorithm could estimate the precision matrix even when the dimension of the precision matrix exceeds the sample size by applying it to the monthly return data of 483 stocks over 50 months. Since the key part of the Gibbs sampling algorithm in which the precision matrix is updated is common to other graphical models with shrinkage priors, such as the spike-and-slab prior (Wang [2015]), the horseshoe prior (Li et al. [2019]), and other scale-mixture-of-normals shrinkage priors, it would be simple to incorporate our modified algorithm into the Gibbs sampling algorithm for those models.

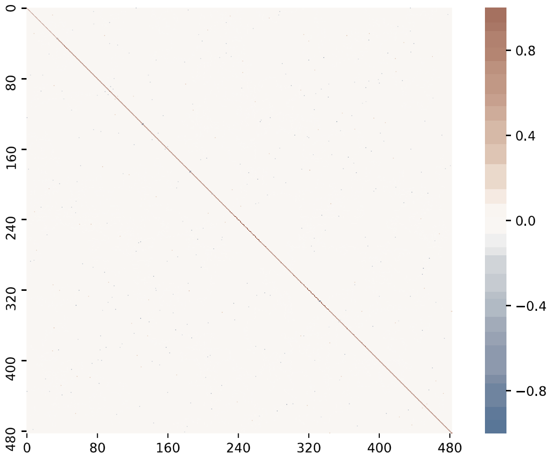


Figure 2.7: Posterior mean of Ω by the BGS

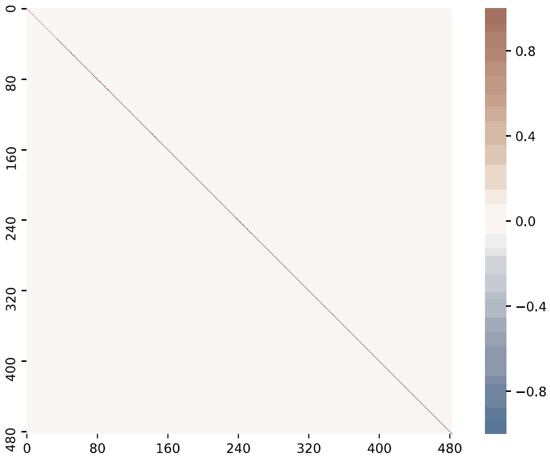


Figure 2.8: Posterior mean of Ω by the HRS

Acknowledgements

This Chapter is based on Oya and Nakatsuma (2022). The paper is first published in *Japanese Journal of Statistics and Data Science*, 2022 by Springer Nature. Reproduced with permission from "Springer Nature". The Version of Record of the paper is available online at `https://doi.org/10.1007/s42081-022-00147-1`.

Chapter 3

A Bayesian Graphical Approach for Large-Scale Portfolio Management with Fewer Historical Data

3.1 Introduction

Since Markowitz (1952) proposed the mean-variance model for optimal portfolio selection, it has remained an important foundation in research of portfolio theory as well as in practice of portfolio management. For example, many of robot advisor services developed in the recent fintech boom are basically based on the mean-variance model or its variants such as the Black-Litterman approach (Black and Litterman (1991, 1992)). Although numerous solutions have been proposed for various problems in the mean-variance model over its long history, one of the most challenging tasks is how to construct the optimal portfolio when the number of assets p exceeds the number of observations of asset returns n. This issue is often referred to as the $p > n$ problem in the literature.

Let us restate this problem in the context of optimal portfolio selection.

For the sake of simplicity, we consider the variance minimization problem without setting the target level of the expected return of the portfolio, i.e.,

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^\intercal \boldsymbol{\Sigma} \boldsymbol{w} \\ \text{s.t.} \quad & \boldsymbol{\iota}^\intercal \boldsymbol{w} = 1, \end{aligned} \tag{3.1}$$

where

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \dots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \dots & \sigma_p^2 \end{bmatrix}, \ \boldsymbol{\iota} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

$\boldsymbol{w}$ is a $p \times 1$ vector of allocation weights and $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix of asset returns. $\boldsymbol{\iota}$ is a $p \times 1$ vector whose elements are all equal to one. In this setup, the solution of (3.1) is given by

$$\boldsymbol{w}^{\text{GMV}} = \frac{1}{\boldsymbol{\iota}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\iota}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\iota}, \tag{3.2}$$

where $\boldsymbol{w}^{\boldsymbol{GMV}}$ is called the global minimum variance portfolio.

Note that $\boldsymbol{w}^{\boldsymbol{GMV}}$ depends on $\boldsymbol{\Sigma}^{-1}$, the inverse matrix of the covariance matrix $\boldsymbol{\Sigma}$ or the precision matrix. For later use, we define $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Since the precision matrix $\boldsymbol{\Omega}$ is unknown in practice, we need to estimate it with asset return data before we apply the formula in (3.2). The simplest estimation method for $\boldsymbol{\Omega}$ is to replace each element of $\boldsymbol{\Sigma}$ with its sample analog, which is called the sample covariance matrix, and then compute the inverse of the sample covariance matrix to obtain an estimate of $\boldsymbol{\Omega}$. In principle, however, we cannot compute the inverse of the sample covariance matrix when the number of assets p is larger than the number of asset return observations n. Moreover, even if p is smaller than n, it is well known that computation of the inverse of the sample covariance matrix tends to be unstable as p approaches n. Therefore, in order to replace $\boldsymbol{\Sigma}$ in (3.2) with the sample covariance matrix, we need to prepare a sufficiently larger number of asset return observations than the number of assets in the portfolio.

This requirement would be a great hindrance in practice. For example, when we want to manage 500 assets in a fund, we need more than two years' worth of data in case of daily returns and more than 42 years' worth of data in case of monthly returns. The latter is rather impractical, of course.

Let us point out another situation related to the $p > n$ problem. Suppose a large-scale IPO is made due to a privatization of a state-owned enterprise or a large-cap firm merges with another one and relists itself. For those stocks with large market-cap, available historical data after IPO or merger is too few. Thus it is difficult to evaluate the risk of a portfolio which includes those stocks because of the $p > n$ problem.

Besides the problem of data availability, using long historical asset return data may cause another problem. Suppose we observe large market turmoil in the past few days. Although it is preferable to incorporate this recent shock into the estimation of the precision matrix, influence of older observations well before the shock is still so dominating in long historical data that we could underestimate the risk of the new shock and it might result in under-performance. To avoid this issue, we may cut out most of the older observations. But then again we are faced with the $p > n$ problem due to a shorten sample period.

In the literature, many researchers have tackled the $p > n$ problem in different contexts and proposed many possible solutions. Among them, one of the most widely explored methods is dimension compression. In particular, factor models are quite popular as tools for dimension compression in the field of finance. A typical factor model assumes that the variation in asset returns is mostly explained by a set of common factors and the residuals are independent of each other as well as the common factors. In this way, the covariance matrix of asset returns $\mathbf{\Sigma}$ is decomposed into two matrices: the covariance matrix induced by the common factors and the diagonal covariance matrix of the residuals. It is well known that $\mathbf{\Sigma}$ in a factor model will be non-singular as long as the common factors are linearly independent. In a typical factor model, the number of the common factors, say k, is much smaller than the number of observations n. For most cases, this property itself is sufficient to guarantee the linear independence among the common factors and the existence of the precision matrix $\mathbf{\Omega}$. Furthermore, since $k < p$, the dimension of the parameter space is largely reduced. So it is regarded as a type of dimension compression method. There are many variants of factor models including dynamic ones, but arguably the most famous one is the three-factor model proposed by Fama and French (1993).

As the factor model gains popularity in both academia and business, the number of potential candidates for common factors has been exploding. According to Harvey et al. (2016), it reaches 316 and counting. Thus it is necessary to select appropriate factors among a huge set of candidates in practice. Here again, we can utilize dimension compression for this purpose. In the context of linear regression with many explanatory variables, the penalized regression method is widely used for simultaneously selecting appropriate variables and estimating the corresponding coefficients as a convenient way to apply dimension compression. It includes least absolute shrinkage and selection operator or LASSO (Tibshirani (1996)), elastic net (Zou and Hastie (2005)), adaptive LASSO (Zou (2006)), Bayesian LASSO (Park and Casella (2008)), horseshoe prior (Carvalho et al. (2009, 2010a)), Bayesian adaptive LASSO (Alhamzawi et al. (2012), Leng et al. (2014)), generalized double Pareto (Armagan et al. (2013)) among others. In relation to the $p > n$ problem, Guhaniyogi and Dunson (2015) proposed a new approach called "Bayesian compressed regression" which randomly compresses a scaled predictor vector prior to analysis. Guhaniyogi and Dunson (2015) tested a case of $p = 25,000$ and $n = 110$ in simulation, though it has not yet been applied to asset management as far as we know. For the application of BCR to finance field, Koop et al. (2019) extended BCR to a multivariate VAR model and applied it to macroeconomic variable prediction, and Luo and Chen (2020) applied BCR to a realized volatility model.

Another popular approach[1] is to use a Gaussian graphical model[2] for di-

[1] In addition to these two major approaches we mention here, alternative approaches such as Ledoit and Wolf (2004), Laloux et al. (1999) among others are known in the literature.

[2] Although we separately explain the factor model and the Gaussian graphical model

rect estimation of the precision matrix $\mathbf{\Omega}$. The Gaussian graphical model take advantage of the fact that all elements in $\mathbf{\Omega}$ can be treated as unknown parameters in the likelihood function if asset returns are supposed to jointly follow a p-dimensional multivariate normal distribution with the zero mean vector and the covariance matrix $\mathbf{\Sigma}$. The term “graphical” comes from a property of the multivariate normal distribution that any pair of normal random variables are independent if and only if the corresponding off-diagonal element in $\mathbf{\Omega}$ is zero. Therefore $\mathbf{\Omega}$ gives a network of dependence among the random variables where non-zero off-diagonal elements are regarded as links connecting random variables (nodes in the context of graphical modeling) and any zero off-diagonal element indicates no link between two nodes.

Since the number of elements in $\mathbf{\Omega}$ is $p(p+1)/2$, the number of parameters to be estimated will be considerably high for a large-scale graphical model. Therefore it is important to force weak and unessential links to be zero so that the estimated structure of network would become more sparse and interpretable. To penalize inclusion of such redundant links in the model, aforementioned LASSO proposed by Tibshirani (1996) has been applied to the Gaussian graphical model by Meinshausen and Bühlmann (2006), Friedman et al. (2008), Yuan and Lin (2007), Banerjee et al. (2008), Guo et al. (2011) among others. This type of LASSO is called graphical LASSO or glasso. Alternatively, Finegold and Drton (2011) proposed tlasso which used the multivariate Student’s t distribution in place of the multivariate normal distribution in the likelihood function.

In this chapter, to deal with the $p > n$ problem in large-scale portfolio management, we pursue the second approach and propose to utilize graphical LASSO to cull unnecessary dependence among assets in $\mathbf{\Omega}$ so that the global minimum variance portfolio (3.2) should be stabilized even in case of $p > n$. The Gaussian graphical model was already applied to portfolio optimization by Goto and Xu (2015), Brownlees et al. (2018), Torri et al. (2019) among

in this introduction, distinction between them is rather arbitrary. In practice, we can combine both approaches together as we do in Chapter 3.3.

others. Especially, Torri et al. (2019) examined performance of the global minimum variance portfolio (3.2) constructed by both glasso and tlasso in long-term asset management[3], though Torri et al. (2019) limited the scope of their study in case of $p < n$. Instead, we try to construct (3.2) in case of $p > n$ and push the envelope of graphical LASSO. As far as we know, ours is the first attempt to estimate $\boldsymbol{\Omega}$ in case of $p > n$ and use it in performance comparison of long-term asset management.

For this purpose, we develop a data-driven portfolio framework based on a Bayesian version of graphical LASSO. From the Bayesian perspective, graphical LASSO is regarded as a maximum a posteriori (MAP) estimator with the Laplace prior for each element in $\boldsymbol{\Omega}$. Therefore it is natural for Bayesian statisticians to extend graphical LASSO by introducing a hierarchical structure among priors such as adaptive graphical LASSO (Wang (2012)) and covariance LASSO (Khondker et al. (2013)), or replacing the Laplace prior with alternative shrinkage priors such as the spike-and-slab prior (Wang (2015)) and the horseshoe prior (Li et al. (2019)). These aforementioned previous studies on Bayesian graphical LASSO except for Khondker et al. (2013) relies on the block Gibbs sampler proposed by Wang (2012) which is used to generate a pseudo-random sample of $\boldsymbol{\Omega}$ for Monte Carlo integration. Wang (2012)'s algorithm is a pure-and-simple Gibbs sampler and easy to implement, but Oya and Nakatsuma (2022) pointed out that it could not exactly guarantee the positive-definiteness of generated $\boldsymbol{\Omega}$. Instead they developed a positive-definiteness-assured block Gibbs sampler for Bayesian adaptive graphical LASSO. We apply their algorithm to estimate $\boldsymbol{\Omega}$ and construct the optimal portfolio.

As Torri et al. (2019) mentioned, the previous studies (e.g., DeMiguel et al. (2009) and Fan et al. (2012)) suggested that the expected return of any asset could not be reliably estimated. Thus we will focus on the global

[3] Torri et al. (2019) also constructed portfolios based on random matrix theory filtering (Bouchaud and Potters (2009)) and Ledoit-Wolf shrinkage estimation (Ledoit and Wolf (2004)) as comparative approaches. We also test these approaches in Chapter 3.3.

minimum variance portfolio (3.2) which does not require any estimate of the expected return. In performance comparison, we construct the global minimum variance portfolio of 100 assets for different sample lengths with our new approach or commonly-used non-Bayesian graphical LASSO, and compare their out-of-sample performance in long-term asset management.

The main body of this chapter is organized as follows. In Chapter 3.2, we briefly review the basic idea of graphical LASSO as well as its Bayesian interpretation, and explain the Markov chain sampling algorithm for Bayesian adaptive graphical LASSO by Oya and Nakatsuma (2022). In Chapter 3.3, we report the results of experiments on long-term portfolio management with asset return data of 100 assets. Lastly, we state our concluding remarks in Chapter 3.4.

3.2 Bayesian Adaptive Graphical LASSO

In this section, we introduce the basic framework of graphical LASSO and outline the Bayesian graphical LASSO approach based on Oya and Nakatsuma (2022) which we employ for portfolio management with many assets.

Suppose $\boldsymbol{Y}$ is a $n \times p$ matrix of asset return data with p assets and n observations and each row vector of $\boldsymbol{Y}$ follows the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1})$ where $\boldsymbol{\Omega} = (\omega_{ij})$, $(1 \leqq i, j \leqq p)$ is the precision matrix. Then graphical LASSO is formulated as the following penalized maximum likelihood estimation:

$$\max_{\boldsymbol{\Omega} \in M^{+}} \frac{n}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \mathrm{tr}\,(\boldsymbol{S}\boldsymbol{\Omega}) - \lambda \|\boldsymbol{\Omega}\|_1, \tag{3.3}$$

where $\|\boldsymbol{\Omega}\|_1 = \sum_{i \leqq j} |\omega_{ij}|$[4] and M^{+} are subsets of the parameter space of $\boldsymbol{\Omega}$ in which $\boldsymbol{\Omega}$ is positive definite. $\boldsymbol{S}$ is defined as $\boldsymbol{S} = \boldsymbol{Y}^{\intercal}\boldsymbol{Y}$ and called the scatter

[4] We introduce a general version of non-Bayesian graphical LASSO which shrinks all the elements of $\boldsymbol{\Omega}$ here, but there is another type of non-Bayesian graphical LASSO which shrinks only the off-diagonal elements with $\|\boldsymbol{\Omega}\|_1 = \sum_{i<j} |\omega_{ij}|$. Because the Bayesian graphical LASSO does not shrink the diagonal elements, we also report results for the latter type of non-Bayes graphical LASSO in Chapter 3.3 for comparison.

matrix. The first two term of the objective function in (3.3) is corresponding to the log likelihood function of the multivariate normal distribution. The last term in (3.3) is the penalty for complexity of the graph structure and λ, which is called the shrinkage parameter, dictates the magnitude of penalty for adding an extra link to the graph structure.

As Park and Casella (2008) pointed out, the penalty in (3.3) is equivalent to the log density of a Laplace distribution:

$$p(\omega_{ij}) = \frac{\lambda}{2} e^{-\lambda|\omega_{ij}|}, \quad (1 \leqq i \leqq j \leqq p). \tag{3.4}$$

Note that $\lambda/2$ can be ignored because it does not affect the solution of (3.3). With this interpretation, the graphical LASSO (3.3) is regarded as a maximum a posteriori (MAP) estimator. Pushing further to this line of thinking, we can conduct a fully Bayesian analysis of the Gaussian graphical model.

In a general Bayesian framework, we first assume the probability distribution of the data generating process and set up the likelihood function. In graphical LASSO (3.3), the data generating process is the multivariate normal distribution and the likelihood function is

$$p(\boldsymbol{Y}|\boldsymbol{\omega}) \propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{S\Omega})\right] \mathbf{1}_{M^+}(\boldsymbol{\Omega}), \tag{3.5}$$

where $\boldsymbol{\omega} = \{\omega_{ij}\}_{1\leqq i\leqq j\leqq p}$ is a $p(p+1)/2 \times 1$ vector of elements in the upper or lower triangular part of $\boldsymbol{\Omega}$ and $\mathbf{1}_{M^+}(\boldsymbol{\Omega})$ is an indicator function to check whether $\boldsymbol{\Omega}$ is positive definite or not.

Next, we set up the probability distribution of unknown parameters in the likelihood function, which is called the prior distribution or the prior to be short, to express non-data information about the parameters. Although we may use the original Laplace prior for our Bayesian analysis, we instead propose to use the following prior for each element in $\boldsymbol{\Omega}$:

$$p(\omega_{ij}|\lambda_{ij}) = \begin{cases} \lambda_{ii} e^{-\lambda_{ii}\omega_{ii}}, & (i = j); \\ \frac{\lambda_{ij}}{2} e^{-\lambda_{ij}|\omega_{ij}|}, & (i \neq j), \end{cases} \tag{3.6}$$

which means that the prior of each diagonal element is exponential while that of each off-diagonal element is Laplace. Note that we allow the shrinkage parameter λ_{ij} to differ from element to element. Unlike the original graphical LASSO (3.3) where the Laplace prior is also assumed for the diagonal elements in $\boldsymbol{\Omega}$, the exponential prior is assumed in (3.6). The exponential prior is not a shrinkage prior, but this will not cause any problems because, in principle, we dot no have to force the diagonal elements in $\boldsymbol{\Omega}$ to be zero because they represent links to nodes themselves and must be non-zero. Therefore, in order to achieve sparsity of $\boldsymbol{\Omega}$, we only need to apply a shrinkage prior to the off-diagonal elements[5].

Furthermore, we treat each shrinkage parameter λ_{ij} as an unknown parameter and assume the common prior for all λ_{ij}'s as

$$p(\lambda_{ij}) = \frac{s^r}{\Gamma(r)} \lambda_{ij}^{r-1} e^{-s\lambda_{ij}}, \tag{3.7}$$

which is a gamma distribution. This type of “prior of priors” is called the hierarchical prior.

Lastly, we derive the posterior distribution, which incorporate both data-related information in the likelihood function and non-data information in the prior, with Bayes' theorem:

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\lambda}|\boldsymbol{R}) &\propto p(\boldsymbol{R}|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}), \\ p(\boldsymbol{\omega}|\boldsymbol{\lambda}) &= \prod_{1 \leqq i \leqq j \leqq p} p(\omega_{ij}|\lambda_{ij}), \\ p(\boldsymbol{\lambda}) &= \prod_{1 \leqq i \leqq j \leqq p} p(\lambda_{ij}), \end{aligned} \tag{3.8}$$

[5] Alternative shrinkage priors have been developed for Bayesian graphical LASSO in the literature. The spike-and-slab prior (Wang (2015)) is a widely applied shrinkage prior in particular for variable selection in a regression model. Basically, it is a mixture of two distributions; one is normal with large variance and another is the Dirac delta at zero. When the latter is realized in the mixture of distributions, the corresponding link will be excluded from the graph structure. Another popular choice is the horseshoe prior (Li et al. (2019)) which assumes that each off-diagonal element in $\boldsymbol{\Omega}$ follows a half-Cauchy distribution.

where $\boldsymbol{\lambda} = \{\lambda_{ij}\}_{1\leqq i\leqq j\leqq p}$ is a $p(p+1)/2\times 1$ vector of the shrinkage parameters. If all λ_{ij}'s in (3.6) take a common value λ (or we may use two different values; one for the diagonal elements and another for the off-diagonal elements), the posterior distribution (3.8) is reduced to a simpler model called Bayesian graphical LASSO. In case they can take different values for any elements in $\boldsymbol{\Omega}$ as we assume in (3.6), it is called Bayesian adaptive graphical LASSO. By adding an additional layer of uncertainty, the adaptive version of graphical LASSO can flexibly adjust the shape of the posterior distribution and may hopefully capture the reality in the financial market better.

Unfortunately, we cannot analytically evaluate the posterior distribution (3.8), which means that we need to use a numerical approximation method to obtain an estimate of $\boldsymbol{\Omega}$. In this paper, we employ the Markov chain sampling algorithm by Oya and Nakatsuma (2022) for generating a pseudo-random sample of $\boldsymbol{\Omega}$, $\{\boldsymbol{\Omega}^{(t)}\}_{t=1}^{T}$, along with other parameters from the posterior distribution (3.8), and use the generated sample in Monte Carlo integration to approximate the posterior mean of $\boldsymbol{\Omega}$ as its point estimate, i.e.,

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Omega}^{(t)}. \tag{3.9}$$

In the rest of this section, we briefly describe the Markov chain sampling algorithm by Oya and Nakatsuma (2022). In general, the term "Markov chain sampling" refers to a generic random number generating method which utilizes the convergence of a Markov chain to its invariant distribution. The basic principle of the Markov chain sampling is rather simple. If a Markov chain is convergent to the invariant distribution, any sequence of pseudo-random numbers drawn from such a Markov chain will eventually converge to the invariant distribution. Furthermore, under some mild conditions, the law of large numbers is applicable to the drawn sequence even though it is not an independent process. Therefore, if we can construct a Markov chain whose invariant distribution is the posterior distribution such as (3.8), we will obtain a pseudo-random sample of the parameters by drawing them repeatedly from the Markov chain until the sequence will be stabilized.

One of the popular algorithms for Markov chain sampling is the Gibbs sampler. Suppose it is possible to draw each parameter in the posterior distribution from its conditional distribution given the rest of the parameters, which is called the full conditional posterior distribution in the literature. Then the Gibbs sampler is defined as an iterative algorithm which repeatedly draws each parameter from its full conditional posterior distribution and replaces the previous value of the parameter with the new one before the next parameter will be drawn from its full conditional posterior distribution. In this way, the new values of the parameters are obtained at the end of each cycle of the Gibbs sampler. By construction, any sequence of pseudo-random numbers generated with the Gibbs sampler is a Markov chain whose invariant distribution is the posterior distribution, and this Markov chain will convergent to the posterior distribution in most applications including the Gaussian graphical model we study here.

To derive the Gibbs sampling algorithm for the posterior distribution (3.8), we make use of the fact that the Laplace distribution in (3.6) is expressed as a scale mixture of normal distributions with the exponential distribution:

$$\begin{aligned} p(\omega_{ij}|\tau_{ij}) &= \frac{1}{\sqrt{2\pi\tau_{ij}}}\exp\left(-\frac{\omega_{ij}^2}{2\tau_{ij}}\right), \\ p(\tau_{ij}) &= \frac{\lambda_{ij}^2}{2}\exp\left(-\frac{\lambda_{ij}^2}{2}\tau_{ij}\right). \end{aligned} \tag{3.10}$$

Define $\boldsymbol{\tau} = \{\tau_{ij}\}_{1\leqq i<j\leqq p}$. Then the posterior distribution (3.8) is rewritten as the joint distribution of $\boldsymbol{\omega}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$:

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\tau}, \boldsymbol{\lambda}|\boldsymbol{Y}) \propto\ & |\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{S\Omega})\right]\mathbf{1}_{M^+}(\boldsymbol{\Omega}) \times \prod_{i=1}^{p}\lambda_{ii}e^{-\lambda_{ii}\omega_{ii}} \\ & \times \prod_{1\leqq i<j\leqq p}\frac{1}{\sqrt{2\pi\tau_{ij}}}\exp\left(-\frac{\omega_{ij}^2}{2\tau_{ij}}\right)\frac{\lambda_{ij}^2}{2}\exp\left(-\frac{\lambda_{ij}^2}{2}\tau_{ij}\right) \\ & \times \prod_{1\leqq i\leqq j\leqq p}\lambda_{ij}^{r-1}e^{-s\lambda_{ij}}. \end{aligned} \tag{3.11}$$

To derive the full conditional posterior distribution of $\boldsymbol{\omega}$, we consider the following partition of the precision matrix $\boldsymbol{\Omega}$:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^\intercal & \omega_{22} \end{bmatrix}, \tag{3.12}$$

where $\boldsymbol{\Omega}_{11}$ is a $(p-1)\times(p-1)$ matrix, $\boldsymbol{\omega}_{12}$ is a $(p-1)\times 1$ vector, and ω_{22} is a scalar. Without a loss of generality we can rearrange rows and columns of $\boldsymbol{\Omega}$ so that the lower-right corner of $\boldsymbol{\Omega}$, ω_{22}, is the diagonal element to be generated from its full conditional posterior distribution. Likewise, we can partition $\boldsymbol{S}$, $\boldsymbol{\Upsilon}$, and $\boldsymbol{\lambda}$ as

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^\intercal & s_{22} \end{bmatrix}, \quad \boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Upsilon}_{11} & \boldsymbol{\tau}_{12} \\ \boldsymbol{\tau}_{12}^\intercal & 0 \end{bmatrix}, \quad \boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_{12} \\ \lambda_{22} \end{bmatrix}, \tag{3.13}$$

where $\boldsymbol{\Upsilon}$ is a $p \times p$ symmetric matrix in which the off-diagonal (i,j) element is τ_{ij} and all diagonal elements are equal to zero, while λ_{22} is the element in $\boldsymbol{\lambda}$ that corresponds with the diagonal element ω_{22} in the prior distribution (3.6). According to Oya and Nakatsuma (2022), the full conditional posterior distribution of ω_{22} is derived as the shifted gamma distribution:

$$\omega_{22} = \gamma + \boldsymbol{\omega}_{12}^\intercal \boldsymbol{\Omega}_{11} \boldsymbol{\omega}_{12}, \quad \gamma \sim \text{Gamma}\left(\frac{n}{2}+1,\ \frac{s_{22}}{2}+\lambda_{22}\right), \tag{3.14}$$

while that of $\boldsymbol{\omega}_{\mathbf{12}}$ is obtained as the truncated multivariate normal distribution:

$$\boldsymbol{\omega}_{12} \sim \text{Normal}\left(-\boldsymbol{C}\boldsymbol{s}_{12},\ \boldsymbol{C}\right) \mathbf{1}_{M_\omega^+}(\boldsymbol{\omega}_{12}), \tag{3.15}$$

where

$$\boldsymbol{C} = \left\{(s_{22}+2\lambda_{22})\boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{D}_{\boldsymbol{\tau}}^{-1}\right\}^{-1}, \quad \boldsymbol{D}_{\boldsymbol{\tau}} = \text{diag}(\boldsymbol{\tau}_{12}),$$

and the indicator function $\mathbf{1}_{M_\omega^+}(\boldsymbol{\omega}_{12})$ implies that the domain of the distribution is truncated within

$$M_\omega^+ = \{\boldsymbol{\omega}_{12} :\ \omega_{22} > \boldsymbol{\omega}_{12}^\intercal \boldsymbol{\Omega}_{11} \boldsymbol{\omega}_{12}\}. \tag{3.16}$$

The constraint (3.16) imposed on $\boldsymbol{\omega}_{12}$ is the key to assure the positive definiteness of $\boldsymbol{\Omega}$. Oya and Nakatsuma (2022) suggested using the Hit-and-Run algorithm to draw $\boldsymbol{\omega}_{12}$ from the truncated multivariate normal distribution

(3.15). See Oya and Nakatsuma (2022) for more details on the derivation of their algorithm. It turns out that we can easily construct a Gibbs sampler for the posterior distribution in the form of (3.11).

Moreover, it is straightforward to show that the full conditional posterior distribution of λ_{ij} $(1 \leqq i \leqq j \leqq p)$ is the gamma distribution:

$$\lambda_{ij} \sim \text{Gamma}\left(r+1, s+|\omega_{ij}|\right), \tag{3.17}$$

while that of $1/\tau_{ij}$ $(1 \leqq i < j \leqq p)$ is the inverse Gaussian distribution:

$$\frac{1}{\tau_{ij}} \sim \text{Inverse Gaussian}\left(\frac{\lambda_{ij}}{|\omega_{ij}|}, \lambda_{ij}^2\right). \tag{3.18}$$

The outline of the Gibbs sampler is summarized as follows.

Gibbs sampler for Bayesian adaptive graphical LASSO

For $i = 1, \ldots, p$, repeat *Step 1* to *Step 5*.

Step 1: Rearrange $\boldsymbol{\Omega}$, $\boldsymbol{S}$, $\boldsymbol{\Upsilon}$, and $\boldsymbol{\lambda}$ so that ω_{ii} is in the place of ω_{22} in $\boldsymbol{\Omega}$ and partition them as in (3.12) and (3.13).

Step 2: If $i \geqq 2$, $\boldsymbol{\omega}_{12} \leftarrow \text{Normal}\left(-\boldsymbol{C}\boldsymbol{s}_{12}, \boldsymbol{C}\right) \mathbf{1}_{M_\omega^+}(\boldsymbol{\omega}_{12})$.

Step 3: $\gamma \leftarrow \text{Gamma}\left(\frac{n}{2}+1, \frac{s_{22}}{2}+\lambda_{22}\right)$ and set $\omega_{22} = \gamma + \boldsymbol{\omega}_{12}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\omega}_{12}$.

Step 4: $\lambda_{ij} \leftarrow \text{Gamma}\left(r+1, s+|\boldsymbol{\omega}_{ij}|\right)$ for $j = i, \ldots, p$.

Step 5: $\upsilon_{ij} \leftarrow$ Inverse Gaussian $\left(\frac{\lambda_{ij}}{|\boldsymbol{\omega}_{ij}|}, \lambda_{ij}^2\right)$ and set $\tau_{ij} = 1/\upsilon_{ij}$ for $j = i+1, \ldots, p$.

From now on we refer to Oya and Nakatsuma (2022)'s positive-definiteness-assured Bayesian adaptive graphical LASSO approach as Bada-PD.

3.3 Performance Comparison in Long-term Portfolio Management

We compare Bada-PD with non-Bayesian graphical LASSO (glasso) in terms of long-run portfolio management with the dataset of portfolios provided by Kenneth French[6]. Following Torri et al. (2019), we choose monthly return data on 100 portfolios of US companies formed on size and book-to-market ratio. According to the description given in Kenneth French's website, these portfolios are the intersections of 10 portfolios formed on size (market equity, ME) and 10 portfolios formed on the ratio of book equity to market equity (BE/ME). Although Torri et al. (2019) used the original portfolio return data, we use the OLS residuals in the Fama-French three-factor model of these portfolio returns[7]. This is because the precision matrix of the original portfolio returns is not sparse, possibly due to the existence of common factors. The data used for estimating the Fama-French three-factor model are also retrieved from Kenneth French's website.

In our empirical study, we test five scenarios: (p, n) = (100, 120), (100, 60), (100, 12), (100, 6) and (100, 3), which are corresponding to the sample period of 10 years, 5 years, 1 year, 2 quarters, and 1 quarter respectively. The description of these scenarios is summarized in Table 3.1. The p/n ratio of the scenarios ranges from $5/6 = 0.833\ldots$ to $100/3 = 33.333\ldots$ and Case (b) – (e) are corresponding to the $p > n$ problem. In particular, Case (e) is an extreme scenario in which we construct a portfolio of 100 assets with only 3 observations.

The backtesting for performance comparison is conducted as follows. We form the global minimum variance portfolio (3.2) with an estimate of $\boldsymbol{\Omega}$. For glasso, we estimate $\boldsymbol{\Omega}$ by using the functionality of `GraphicalLassoCV` in a Python package `sklearn.covariance`. For Bada-PD, we estimate $\boldsymbol{\Omega}$ with its posterior mean[8] in (3.8). To obtain a stable sequence of the Markov chain,

[6] The dataset is available at Kenneth French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

[7] We first estimate the three-factor model for the whole sample period ($n = 120$) and use the residuals in (a) – (e) to remove the influence of the common factors.

[8] When optimizing a portfolio in a Bayesian approach, we generally use the predictive distribution of asset returns which is derived with the posterior distribution of the unknown parameters. As far as we apply the mean-variance model to portfolio selection, however, we do not have to evaluate the predictive distribution explicitly. As an example relevant to our study, let us consider the case of the multivariate normal distribution Normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In

Table 3.1: Descriptive Statistics of the Dataset

	p	n	p / n	Time Period	Data Frequency
Case : $p < n$					
(a)	100	120 (10Y)	0.833	01/2001 - 09/2020	monthly
Case : $p > n$					
(b)	100	60 (5Y)	1.667	01/2006 - 09/2020	monthly
(c)	100	12 (1Y)	8.333	01/2010 - 09/2020	monthly
(d)	100	6 (2Q)	16.667	07/2010 - 09/2020	monthly
(e)	100	3 (1Q)	33.333	10/2010 - 09/2020	monthly

we first iterate the Gibbs sampler 15,000 times for Case (a) and 5,000 times for Case (b) – (e) as burn-in and store pseudo-random numbers drawn in the next 20,000 iterations[9]. In addition to glasso and Bada-PD, we form the equal weight portfolio (EW) as the benchmark. We also tested the non-Bayesian graphical LASSO with $\lambda_{ii} = 0$ graphical LASSO[10] (glasso -

this case, the covariance matrix of the predictive distribution is expressed as the sum of the posterior covariance matrix of the mean vector $\boldsymbol{\mu}$ and the posterior mean of the covariance matrix $\boldsymbol{\Sigma}$. If we assume $\boldsymbol{\mu} = \mathbf{0}$, the covariance matrix of the predictive distribution is identical to the posterior mean of $\boldsymbol{\Sigma}$. Thus evaluating the posterior mean of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$ will suffice to construct the global minimum variance portfolio (3.2).

[9] We also conducted an additional performance test of the proposed approach with simulated data in a case where the p/n ratio is much higher than the experiment on long-term portfolio management. In this simulation study, it is also shown how we examined the convergence of the proposed model. See Appendix for more details. Since it is impossible to know a true structure of a precision matrix here unlike the simulation study in Appendix, we alternatively confirmed the convergence of the top 10 eigenvalues in Chapter 3.3. Based on results of the convergence diagnostic, it is figured that we need to increase the number of burn-in to 15,000 times for Case (a) although 5,000 times burn-in is sufficient for Case (b) – (e).

[10] We used `QuicGraphLassoCV` in a Python package `skggm` based on Hsieh et al. (2014). This package is available at `https://github.com/skggm/skggm`. This algorithm is based on the second order method. Thus, strictly speaking, it is different from the algorithms of the first order method models such as glasso and Rothman et al. (2008)'s SPICE with

$\lambda_{ii} = 0$) as another type of non-Bayesian graphical LASSO algorithm that does not impose penalties on diagonal elements as mentioned in Chapter 3.2. Random matrix theory filtering[11] (RMT, Bouchaud and Potters (2009)) and Ledoit–Wolf shrinkage estimation[12] (LW, Ledoit and Wolf (2004)) are also compared as in Torri et al. (2019). Moreover, we try to estimate the covariance matrix with its sample analog for Case (a) in which the sample covariance matrix is non-singular since $p < n$. The out-of-sample period is from January 2011 to December 2020 for all cases. We examine out-sample-performance of each portfolio strategy by using a rolling window approach by rebalancing the portfolios once every three months. For the sake of simplicity, we ignore transaction fees and selling restrictions. All computations are implemented with Python codes on a desktop PC with 128GB RAM and 3.8GHz i7-10700K Intel processor.

First, we examine 10-year performance of each portfolio strategy in terms of risk-return tradeoff. Table 3.2 shows three performance measures of portfolios: mean return, standard deviation and Sharpe ratio. As for the performance comparison among EW, RMT, LW, glasso - $\lambda_{ii} = 0$, glasso and Bada-PD, we focus on the Sharpe ratio in the fourth column of Table 3.2. EW achieves the lowest Sharpe ratio in all cases because of its large standard deviation in Case (a) – (d). In Case (a) where p is less than n, the Sharpe ratio is the highest for the global minimum variance portfolio with the sample covariance matrix, though it cannot be applicable to Case (b) – (e) because the sample covariance matrix is singular in those cases.

The graphical LASSO approaches (Bada-PD, glasso and glasso - $\lambda_{ii} = 0$) consistently out-perform EW in Case (a) – (d) and their Sharpe ratios even exceed 1. However, the standard deviation increases sharply for both in Case (e) where the p/n ratio reaches 33 and, as a result, the Sharpe ratio declines

$\lambda_{ii} = 0$. Since SPICE has problems related to computational load and complexity as Duchi et al. (2008) pointed out and implementation of Hsieh et al. (2014) is relatively easy, we use this algorithm as an example of non-Bayesian graphical LASSO with $\lambda_{ii} = 0$.

[11] We used a Python package `pyRMT` (`https://github.com/GGiecold/pyRMT`).

[12] We used `Ledoit-Wolf` in a Python package `sklearn.covariance`.

below 1. As for non-Bayesian models, glasso significantly dropped to 0.235, underperforming EW. Glasso - $\lambda_{ii} = 0$ fails to estimate the precision matrix. Bada-PD, on the other hand, succeeds in estimating the precision matrix even in Case (e) and still out-performs EW in terms of the Sharpe ratio. LW achieves constantly high performance in Case (a) – (d). Even in Case (e), LW's performance is comparable to EW although it does not reach Bada-PD. RMT out-performs the graphical LASSO approaches in Case (a) and achieves approximately the similar performance as glasso - $\lambda_{ii} = 0$ in Case (b) – (d). In Case (e), RMT slightly underperforms EW.

Next, we compare the portfolio strategies in terms of portfolio composition such as shorting, diversification and turnover. Following Torri et al. (2019), we calculate the summary statistics for portfolio composition and report them in Table 3.3. "Gross exp." in the second column of Table 3.3 is the gross exposure which is defined as the sum of the absolute values of the weights $\sum_i |w_i|$. "Short exp." in the third column means the short exposure, i.e., the total amount of the short position. "Max short" in the fourth column indicates the maximum negative exposure of individual assets. HDI in the fifth column is the modified Herfindahl diversification index corrected to account for short portfolio:

$$\mathrm{HDI} = \sum_{i=1}^{p} w_i^{*2}, \quad w_i^* = \frac{w_i}{\sum_{i=1}^{p} |w_i|}.$$

This index measures a level of diversification of the portfolio. "Active pos." in the sixth column means the percentage of active position of the portfolio. Since we do not impose any restrictions on the weights in this study, all values are equal to 100%. "Turnover" in the seventh column indicates the turnover ratio.

First, let us examine the risk exposures of each portfolio strategy. For the global minimum variance portfolio with the sample covariance matrix, the gross exposure is over 23 times higher than the initial endowment and the short exposure is almost 11, in spite of the fact that the p/n ratio is less than 1 in Case (a). Thus we may conclude that using a plain sample estimate

Table 3.2: Out-of-sample Performance of the Portfolios

Portfolio	Mean Return	Standard Deviation	Sharpe Ratio
EW	0.148	0.254	0.499
$p < n$			
(a) p = 100, n = 120			
sample covariance	0.299	0.119	2.332
glasso	0.152	0.103	1.265
glasso - $\lambda_{ii} = 0$	0.187	0.150	1.108
Bada-PD	0.160	0.107	1.295
RMT	0.171	0.114	1.311
LW	0.218	0.088	2.225
$p > n$			
(b) p = 100, n = 60			
glasso	0.160	0.104	1.327
glasso - $\lambda_{ii} = 0$	0.181	0.142	1.124
Bada-PD	0.168	0.109	1.343
RMT	0.149	0.109	1.171
LW	0.185	0.086	1.890
(c) p = 100, n = 12			
glasso	0.157	0.105	1.287
glasso - $\lambda_{ii} = 0$	0.186	0.149	1.103
Bada-PD	0.172	0.099	1.521
RMT	0.142	0.106	1.138
LW	0.211	0.123	1.548
(d) p = 100, n = 6			
glasso	0.188	0.129	1.294
glasso - $\lambda_{ii} = 0$	0.187	0.138	1.197
Bada-PD	0.185	0.111	1.481
RMT	0.166	0.125	1.156
LW	0.188	0.130	1.283
(e) p = 100, n = 3			
glasso	0.086	0.274	0.235
glasso - $\lambda_{ii} = 0$	NA	NA	NA
Bada-PD	0.134	0.184	0.611
RMT	0.125	0.220	0.469
LW	0.119	0.198	0.492

Note: NA means we cannot estimate by the model because of ill condition.

Table 3.3: Statistics of Portfolio Composition[1]

Portfolio	Gross exp. [2]	Short exp. [3]	Max short [4]	Active pos.	HDI[5]	Turnover
EW	1.000	0.000	0.000	100%	0.010	0.000
$p < n$						
(a) p = 100, n = 120						
sample covariance	23.195	11.098	-1.626	100%	0.017	11.727
glasso	5.976	2.488	-0.220	100%	0.020	1.144
glasso - $\lambda_{ii} = 0$	5.670	2.335	-0.380	100%	0.019	1.399
Bada-PD	6.014	2.507	-0.454	100%	0.021	1.792
RMT	8.290	3.645	-0.315	100%	0.019	1.914
LW	9.450	4.225	-0.471	100%	0.017	2.676
$p > n$						
(b) p = 100, n = 60						
glasso	5.553	2.276	-0.216	100%	0.021	1.149
glasso - $\lambda_{ii} = 0$	5.465	2.233	-0.355	100%	0.019	1.354
Bada-PD	5.795	2.397	-0.747	100%	0.022	1.850
RMT	7.497	3.249	-0.397	100%	0.020	1.895
LW	6.626	2.813	-0.375	100%	0.018	2.001
(c) p = 100, n = 12						
glasso	5.302	2.151	-0.627	100%	0.025	2.172
glasso - $\lambda_{ii} = 0$	5.251	2.125	-0.350	100%	0.019	1.226
Bada-PD	5.266	2.133	-0.590	100%	0.024	2.309
RMT	6.854	2.927	-0.666	100%	0.024	2.716
LW	5.078	2.039	-0.305	100%	0.018	1.996
(d) p = 100, n = 6						
glasso	5.251	2.126	-0.580	100%	0.030	3.925
glasso - $\lambda_{ii} = 0$	4.965	1.982	-0.367	100%	0.019	1.464
Bada-PD	4.829	1.914	-0.858	100%	0.027	3.202
RMT	6.676	2.838	-0.769	100%	0.029	4.227
LW	4.684	1.842	-0.324	100%	0.019	2.508
(e) p = 100, n = 3						
glasso	4.503	1.751	-4.136	100%	0.139	7.266
glasso - $\lambda_{ii} = 0$	NA[6]	NA	NA	NA	NA	NA
Bada-PD	3.592	1.296	-0.559	100%	0.033	4.429
RMT	4.530	1.765	-0.856	100%	0.063	6.511
LW	3.581	1.291	-0.343	100%	0.022	3.700

[1] These statistics are averaged across all rebalancing periods.

[2] Gross exp. is $\sum_{i=1}^{p} |w_i|$.

[3] Short exp. is the total amount of the short position.

[4] Max short is the maximum negative exposure of each asset.

[5] HDI = $\sum_{i=1}^{p} w_i^{*2}$ where $w_i^* = w_i / (\sum_{i=1}^{p} |w_i|)$.

[6] NA means that we fail to estimate the precision matrix because of the ill condition.

of the covariance matrix has a tendency to take extreme short positions. For the graphical LASSO approaches, on the other hand, the gross exposure is considerably lower than that of the sample covariance case and gradually decreases as the p/n ratio increases. As for RMT and LW, LW takes relatively larger gross exposure than the graphical LASSO approaches in Case (a) – (b), however, the gross exposure decreases as the p/n ratio increases and becomes the same level as them. For RMT, gross exposure decreases as the p/n ratio increases, but is relatively higher than the other approaches except sample covariance. Note that EW takes only long position for all asset by definition and the gross exposure is always equal to 1.

Next, let us check the degree of portfolio diversification. As for HDI, we do not observe any noticeable differences between the graphical LASSO approaches except for Case (e) in which the HDI of glasso is 4 times higher than that of Bada-PD and glasso - $\lambda_{ii} = 0$ cannot estimate the precision matrix. In Case (e), the p/n ratio is beyond 33 and *Max short* of glasso is much more extreme than the other cases. We speculate that this is because the estimation procedure for the precision matrix may become unstable if the p/n ratio is too high. On the other hand, Bada-PD seems to construct portfolios with stable HDI even in Case (e). Glasso - $\lambda_{ii} = 0$ stably takes a low HDI in Case (a) – (d), but fails to obtain the estimate of the precision matrix in Case (e). LW remarkably takes the most stable HDI, whereas RMT takes a large HDI in Case (e).

Finally, let us look into the turnover ratio. Obviously, the turnover ratio of EW is equal to zero and portfolios with the sample covariance matrix have the highest turnover ratio. For both Bada-PD and glasso, the turnover ratio tends to increase as the p/n ratio gets higher, though the rate of increment is less severe for Bada-PD than glasso. Glasso - $\lambda_{ii} = 0$ stably takes the low turnover ratio, but fails to estimate the precision matrix in Case (e). LW takes a low turnover ratio in Case (c) – (e), although takes lager turnover ratio in Case (a) – (b) than the other models except sample covariance. RMT tends to take a large turnover as the p / n ratio increases, showing no significant differences compared to other models. Unlike Bada-PD and LW, the turnover rises a little steeply in Case (e) like glasso.

3.4 Conclusion

Limited availability of historical data on asset returns has been a hindrance to asset management because the sample covariance matrix of asset returns is singular when the number of asset p exceeds the number of observations n. In this paper, we explored a possible solution to this so-called $p > n$ problem in large-scale portfolio management. To solve this problem, we proposed a new data-driven portfolio framework based on Bayesian adaptive graphical LASSO with the Markov chain sampling algorithm proposed by Oya and Nakatsuma (2022). The proposed approach can directly estimate the precision matrix of asset returns even in case of $p < n$ as we demonstrated in Chapter 3.3. We tested out-of-sample performance of the proposed approach in long-term portfolio management by using monthly return data of 100 portfolios available at Kenneth French's website in various scenarios. We also compare them with portfolios based on another type of non-Bayesian graphical LASSO which shrinks only the off-diagonal elements, random matrix theory filtering and Ledoit-Wolf shrinkage estimation.

In this experiment, we confirmed advantages of the proposed approach over the conventional sample covariance approach, non-Bayesian graphical LASSO and other approaches in comparison in terms of return-risk tradeoff and portfolio composition. Both Sharpe ratios and indices of portfolio composition were relatively stable for the proposed approach while they were either unstable for non-Bayesian graphical LASSO. Even in the most severe scenario where the precision matrix of 100 assets must be estimated with only 3 observations, the proposed approach was able to estimate the precision matrix and outperformed the equal weight portfolio without taking abnormal values in regard to indices of portfolio composition.

3.5 Appendix

In Chapter 3.3, we tested the proposed approach in case of $p = 100$ due to the data constraint. In addition, we test whether the proposed model works properly and converges in a more high-dimensional and more severe p/n environment using simulated data. As the true design of the precision matrix $\mathbf{\Omega}$, we assume a AR(4) model:

- AR(4): $\omega_{ii} = 1.0$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.8$, $\omega_{i,i-2} = \omega_{i-2,i} = 0.6$, $\omega_{i,i-3} = \omega_{i-3,i} = 0.4$, and $\omega_{i,i-4} = \omega_{i-4,i} = 0.2$.

Then, artificial data are generated from a multivariate normal distribution with zero mean and the precision matrix of the AR(4) model. We also assume $(p, n) = (300, 3)$ where p/n is 100. Based on the artificial data, we calculate the posterior statistics of $\mathbf{\Omega}$ with Bada-PD and see if the true structure of $\mathbf{\Omega}$ can be estimated well. We used the Monte Carlo sample from 10,000 iterations after 5,000 burn-in iterations in the Bayesian inference because the proposed model with the artificial data converges faster than that with the real data in Chapter 3.3. Settings for hyperpriors are the same as in Chapter 3.3.

To assess estimation accuracy, we consider performance of the graphical structure learning as Fan et al. (2009) and Oya and Nakatsuma (2022). In the context of graphical modeling, non-zero elements of $\mathbf{\Omega}$ are regarded as links which connect the corresponding nodes. We use the following criteria to determine whether the nodes of $\mathbf{\Omega}$ are connected or not:

$$\begin{cases} |\hat{\omega}_{ij}| \geqq 10^{-3} & \text{(node } i \text{ and node } j \text{ are connected);} \\ |\hat{\omega}_{ij}| < 10^{-3} & \text{(node } i \text{ and node } j \text{ are not connected),} \end{cases} \tag{3.19}$$

where $\hat{\omega}_{ij}$ is the point estimate of ω_{ij} computed with the Monte Carlo sample of $\mathbf{\Omega}$ with Bada-PD. With the estimated graphical structure, we measure the accuracy in the graphical structure learning with three indexes: specificity, sensitivity, and the Matthews correlation coefficient (MCC), namely

$$\begin{aligned} \text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{MCC} &= \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FN})(\text{TN}+\text{FN})}}, \end{aligned} \tag{3.20}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. Table 3.4 shows the graph structure learning performance of $\mathbf{\Omega}$ estimated with Bada-PD. All of 2,680 non-zero elements of the true $\mathbf{\Omega}$ were correctly identified, and only 4 out of 87,320 zero elements were misidentified. Also, it can be seen from Figures 3.1 and 3.2 that HRS is able to properly estimate the structure of true Ω.

Table 3.4: Accuracy in graphical structure learning

$(p,n)=(300,3)$	TN	FP	FN	TP	Specificity	Sensitivity	MCC
AR(4)	87,316	4	0	2,680	99.9954	100.0000	99.9232

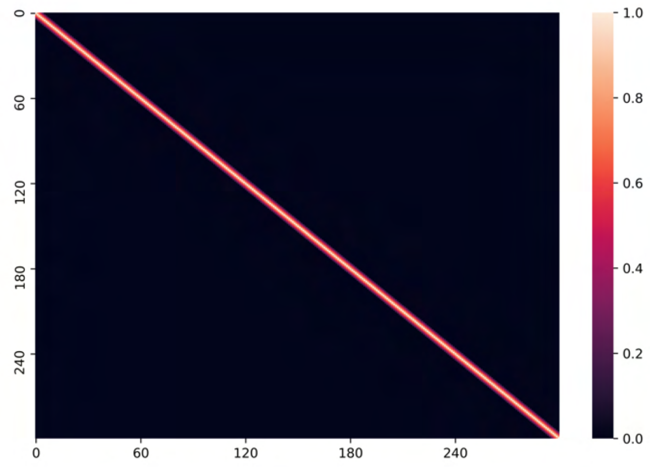


Figure 3.1: True Ω

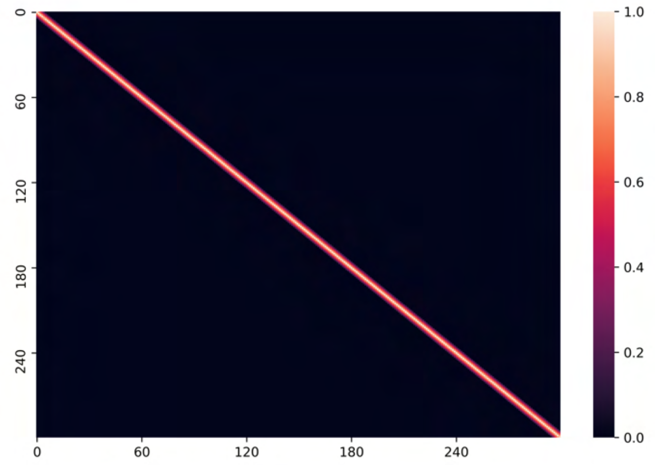


Figure 3.2: Posterior Mean of Ω by Bada-PD

Finally, we examine whether the generated sample of $\boldsymbol{\Omega}$ has converged properly. $\boldsymbol{\Omega}$ has 90,000 elements. Even considering the symmetry of $\boldsymbol{\Omega}$, we have to confirm the convergence of $p(p+1)/2 = 45,150$ elements. Although convergence tests for MCMC simulation with a high-dimensional distribution have been proposed (e.g., VanDerwerken and Schmidlerb (2017)) in recent years, there is no definitive method yet to be devised. Therefore, instead of the elements of $\boldsymbol{\Omega}$, we check the convergence of the eigenvalues of $\boldsymbol{\Omega}$. We first apply principal component analysis[13] for the true $\boldsymbol{\Omega}$ and calculate the contribution rate of each component. Then, we count the number of components until the cumulative contribution exceeds 80%. For the AR(4) model, the number of such components is 42. We compute the eigenvalues corresponding the top 42 components for simulated $\boldsymbol{\Omega}$ in each iteration of MCMC and check the convergence of these eigenvalues with the convergence diagnostic proposed by Gelman and Rubin (1992). Following Gelman and Rubin (1992), we conclude that the convergence is achieved if the convergence

[13] We used `PCA` in a Python package `sklearn.decomposition`.

diagnostic is between 0.99 and 1.01. Table 3.5 shows the posterior statistics and Gelman-Rubin diagnostic for the top 42 eigenvalues. The results clearly indicate that the MCMC samples of these eigenvalues are converged. Figures 3.3 – 3.6 show the posterior densities and traceplots for the top 42 eigenvalues. We can also confirm the convergence of the top 42 eigenvalues from them.

Acknowledgements

This Chapter is based on Oya (2022). The paper is first published *Asia-Pacific Financial Markets*, 2022, by Springer Nature. Reproduced with permission from "Springer Nature". The Version of Record of the paper is available online at https://doi.org/10.1007/s10690-022-09358-8.

Table 3.5: Posterior Statistics and Convergence Diagnostics of top 42 eigenvalues

	Mean	Std	95% HPDI-lower [1]	95% HPDI-upper	Gelman-Rubin [2]
1st	0.0858	0.0018	0.0843	0.0884	1.0002
2nd	0.0846	0.0004	0.0840	0.0854	0.9999
3rd	0.0838	0.0002	0.0835	0.0842	1.0000
4th	0.0835	0.0001	0.0833	0.0837	1.0005
5th	0.0832	0.0001	0.0830	0.0834	1.0011
6th	0.0828	0.0001	0.0825	0.0831	1.0000
7th	0.0823	0.0002	0.0820	0.0826	0.9999
8th	0.0817	0.0002	0.0815	0.0821	0.9999
9th	0.0810	0.0002	0.0807	0.0814	0.9999
10th	0.0803	0.0002	0.0800	0.0806	1.0000
11th	0.0794	0.0002	0.0791	0.0798	0.9999
12th	0.0786	0.0002	0.0784	0.0790	1.0000
13th	0.0778	0.0002	0.0775	0.0782	0.9999
14th	0.0769	0.0002	0.0766	0.0774	0.9999
15th	0.0759	0.0002	0.0756	0.0763	0.9999
16th	0.0747	0.0002	0.0744	0.0750	1.0000
17th	0.0735	0.0002	0.0732	0.0738	0.9999
18th	0.0726	0.0002	0.0722	0.0730	0.9999
19th	0.0713	0.0002	0.0709	0.0717	1.0000
20th	0.0699	0.0002	0.0695	0.0703	0.9999
21st	0.0686	0.0002	0.0682	0.0690	0.9999
22nd	0.0674	0.0002	0.0671	0.0678	0.9999
23rd	0.0659	0.0002	0.0656	0.0664	1.0000
24th	0.0646	0.0002	0.0642	0.0650	0.9999
25th	0.0631	0.0002	0.0627	0.0635	1.0001
26th	0.0616	0.0002	0.0613	0.0620	1.0002
27th	0.0602	0.0002	0.0599	0.0606	0.9999
28th	0.0587	0.0002	0.0583	0.0592	1.0000
29th	0.0571	0.0002	0.0568	0.0575	0.9999
30th	0.0556	0.0002	0.0553	0.0559	0.9999
31st	0.0541	0.0002	0.0537	0.0545	1.0001
32nd	0.0527	0.0002	0.0523	0.0531	0.9999
33rd	0.0510	0.0002	0.0506	0.0514	1.0001
34th	0.0495	0.0002	0.0491	0.0499	1.0000
35th	0.0478	0.0002	0.0475	0.0482	0.9999
36th	0.0463	0.0002	0.0460	0.0466	0.9999
37th	0.0448	0.0002	0.0444	0.0452	1.0000
38th	0.0431	0.0002	0.0428	0.0435	1.0000
39th	0.0415	0.0002	0.0412	0.0418	0.9999
40th	0.0401	0.0002	0.0398	0.0404	0.9999
41st	0.0385	0.0002	0.0382	0.0388	0.9999
42nd	0.0370	0.0002	0.0367	0.0374	0.9999

[1] HPDI indicates the highest posterior density interval.

[2] The number of chains to calculate the Gelman-Rubin statistic in Table 3.5 is 2. To investigate the effect of the number of the chains on the Gelman-Rubin statistic, we changed it from 2 to 10 and confirmed that there is no problem in all cases.

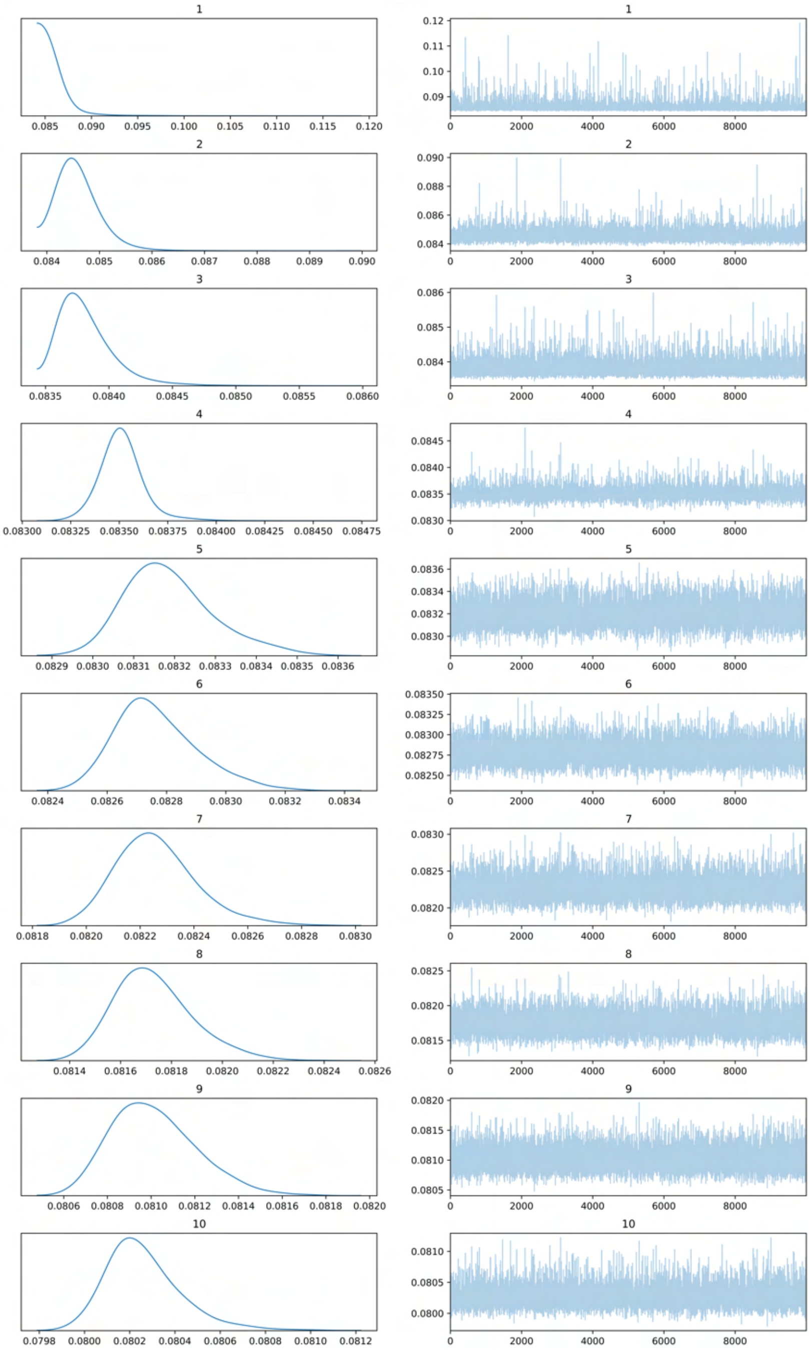


Figure 3.3: The posterior density and traceplot for top 1 to 10 eigenvalues

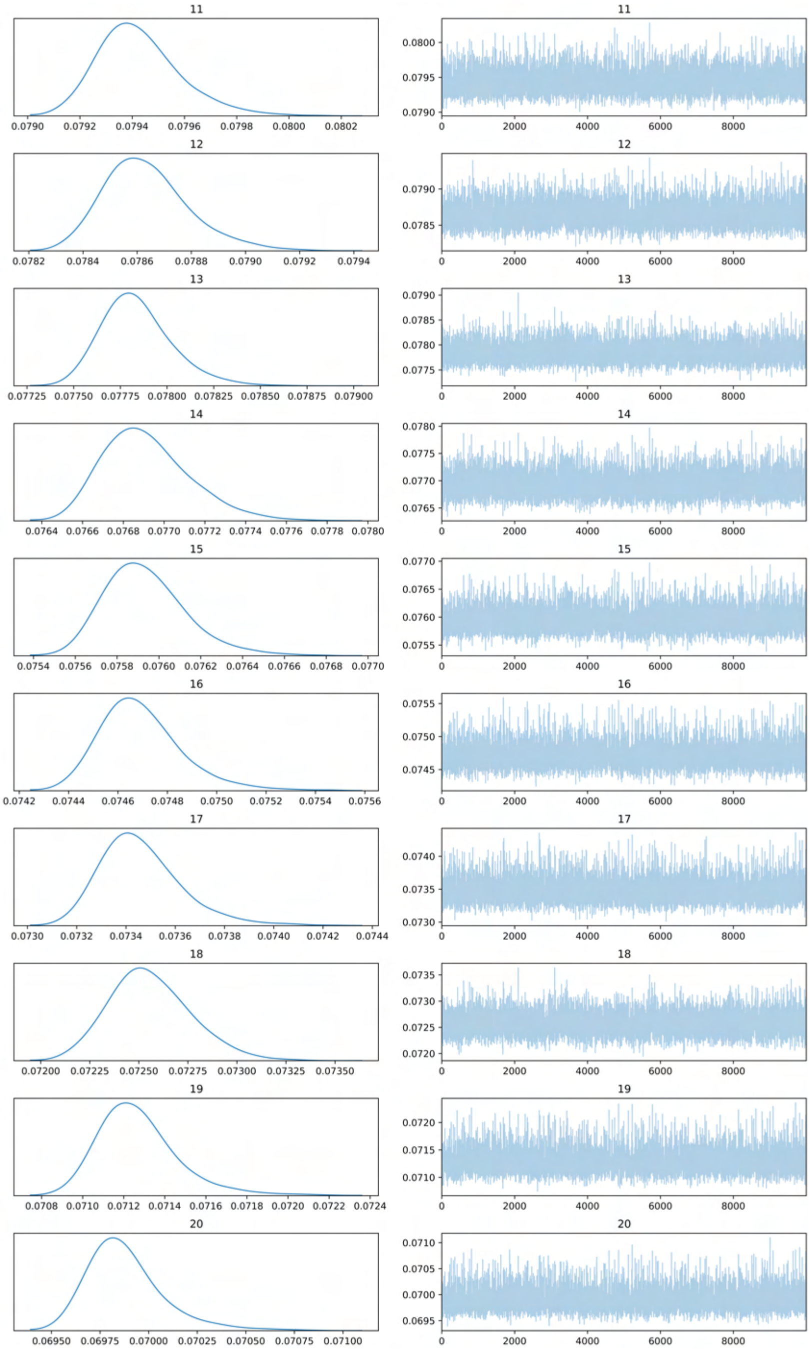


Figure 3.4: The posterior density and traceplot for top 11 to 20 eigenvalues

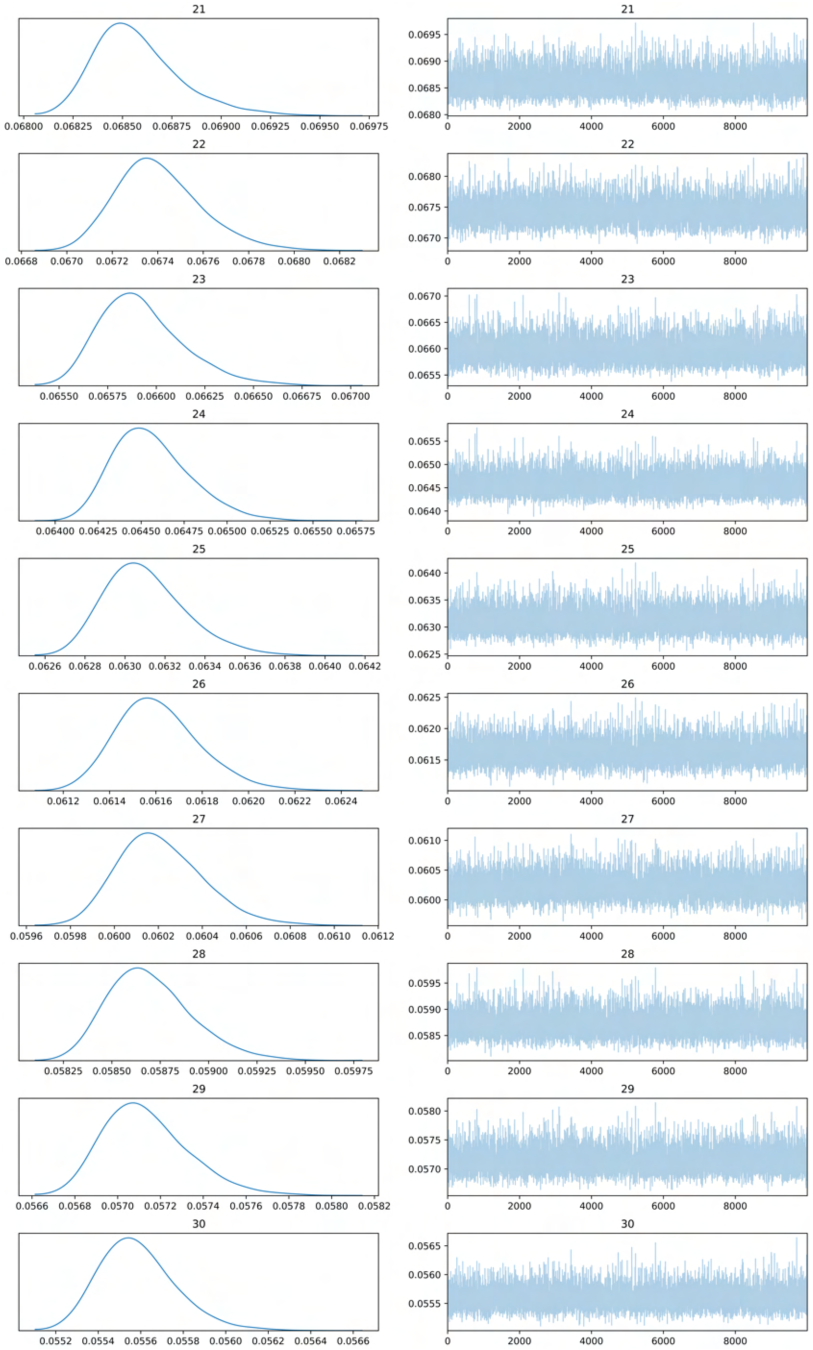


Figure 3.5: The posterior density and traceplot for top 21 to 30 eigenvalues

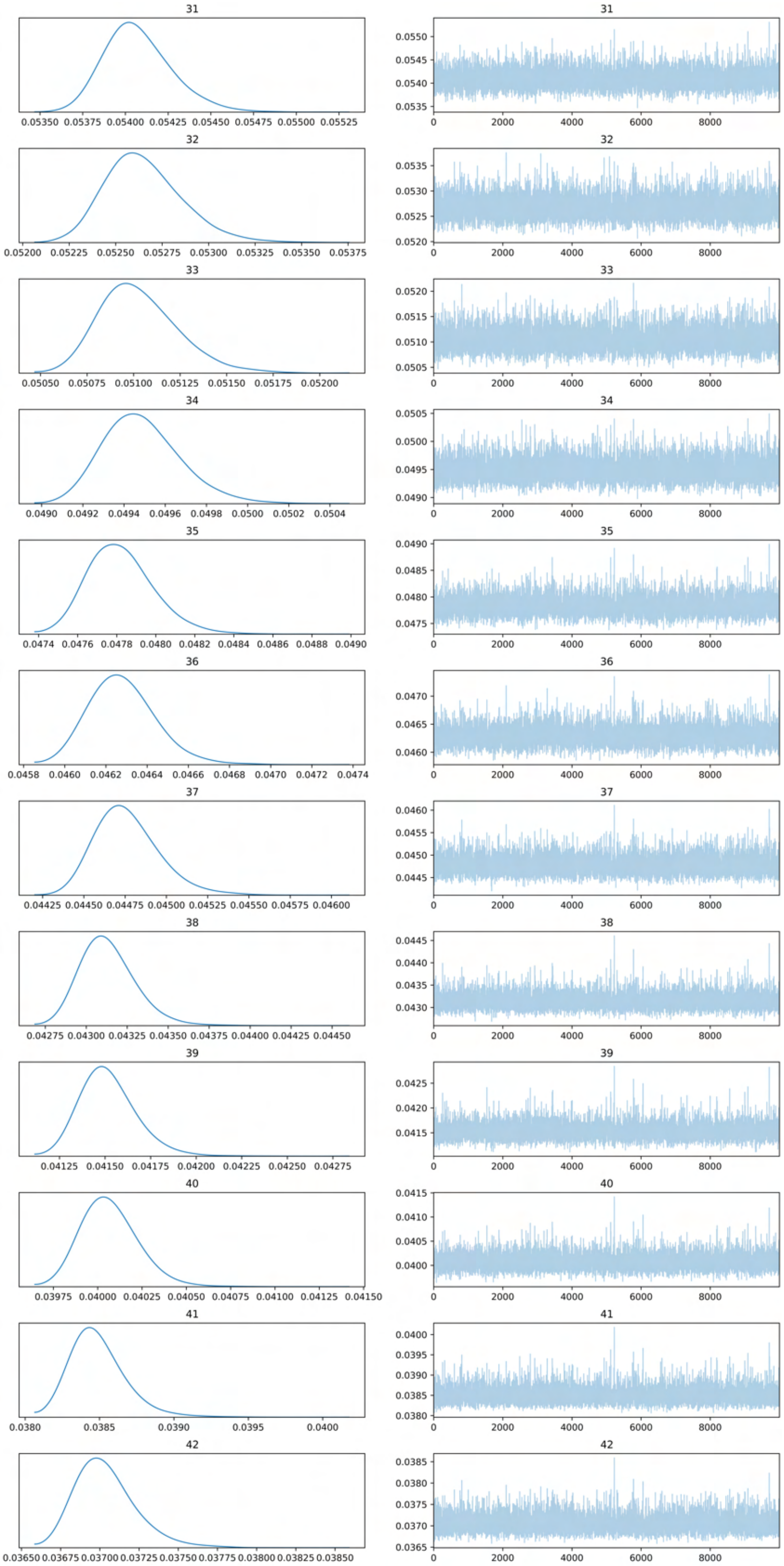


Figure 3.6: The posterior density and traceplot for top 31 to 42 eigenvalues

Chapter 4

Identification in Bayesian Estimation of the Skewness Matrix in a Multivariate Skew-Elliptical Distribution

4.1 Introduction

The mean-variance approach proposed by Markowitz (1952) still plays the central role in portfolio management even today. One of the key assumptions of this approach is that asset returns jointly follow a multivariate normal distribution, though it is well-known that they tend to follow a fat-tailed, possibly skewed distribution as Kon (1984), Mills (1995), Markowitz and Usmen (1996), Peiró (1999) among others have pointed out. Therefore, researches have proposed numerous distributions that can express these characteristics of asset returns well. In particular, a so-called skew-t distribution is often assumed for asset returns since Hansen (1994) first used it for modeling financial data. There are various types of skew-t distribution known in the literature, but arguably the most famous one is based on the generalized hyperbolic (GH) distribution.

The GH distribution, which was originally introduced by Barndorff-Nielsen (1977), can flexibly describe many distributions including the normal distribution, hyperbolic distribution, normal inverse Gaussian (NIG) distribution, Student's t distribution, and skew-t distribution. The skew-t distribution as a special case of the GH distribution is called the GH skew-t distribution. Hansen (1994), Fernández and Steel (1998) and Aas and Haff (2006) assumed the GH skew-t distribution for asset returns. Especially, application of the GH distribution has been recently advanced in the field of asset price volatility models. For example, Nakajima and Omori (2012) assumed the GH skew-t distribution for the error distribution of the stochastic volatility (SV) model and proposed a Bayesian Markov chain Monte Carlo (MCMC) method while Nakajima (2017) constructed a sparse estimation method for the skewness parameter of the GH skew-t distribution in the SV model and demonstrated that it could improve prediction accuracy.

Although the GH distribution is flexible enough to model a single asset on many occasions, it has difficulty in capturing the skewness dependency among multiple assets. Fund managers would find the skewness dependency useful in particular when the financial market crashes and almost all assets suddenly go south since such sharp price co-movement may not be captured by the second moment (i.e., correlation) only.

To circumvent this shortcoming of the GH distribution, we propose to use the skew-elliptical distribution, which was proposed by Branco and Dey (2001) as a generalization of the multivariate skew-normal distribution by Azzalini and Valle (1996) and later improved by Sahu et al. (2003)[1]. The skew-elliptical distribution includes the normal distribution, Student's t dis-

[1] Although we take up the skew-elliptical distribution based on Sahu et al. (2003) with application to portfolio management in mind, there are alternative skew-elliptical-type distributions known in the literature. Research on skew-elliptical-type distributions to financial data is very active (e.g., Barbi and Romagnoli (2018), Carmichael and Coën (2013), Alodat and Al-Rawwash (2014)). Adcock and Azzalini (2020) reviews the recent development in this field and explains relationship among various types of skew-elliptical distribution in detail.

tribution and their skewed counterparts: skew-normal and skew-t distribution. Unlike the GH distribution, it is straightforward to extend the skew-elliptical distribution to the multivariate case. The multivariate skew-normal distribution has another advantage: Its Bayesian estimation can be conducted via pure Gibbs sampling. For example, Sahu et al. (2003) proposed a Gibbs sampler for a linear regression model in which the error term follows a skew-elliptical distribution without skewness dependency. Moreover, Harvey et al. (2010) improved Sahu et al. (2003)'s method, and applied it to Bayesian estimation of the multivariate skew-normal distribution as well as portfolio optimization that considers up to the third moment in the presence of skewness dependency.

In our assessment, however, the Bayesian estimation method of the multivariate skew-elliptical distribution by Harvey et al. (2010) has an identification issue about the skewness parameters due to so-called label switching. To elaborate on our point, let us look into the definition of a multivariate skew-elliptical distribution. For simplicity, we only consider the multivariate skew-normal distribution[2]. Suppose a $p \times 1$ random vector Y_t $(t = 1, \dots, n)$ of asset returns follows a multivariate skew-normal distribution such that

$$\begin{aligned} Y_t &= \mu + \Delta Z_t + \epsilon_t, \\ Z_t &\sim \mathcal{N}^+(0, I_p), \quad \epsilon_t \sim \mathcal{N}(0, \Omega^{-1}), \\ Z_t &\perp \epsilon_t, \end{aligned} \tag{4.1}$$

where each element in Z_t is supposed to independently follow a positive half normal distribution with the scale parameter equal to 1. Δ and Ω are the skewness matrix[3] and the precision matrix[4] respectively. Harvey et al. (2010)

[2] In essence, any skew-elliptical distributions have the same identification issue. So we start with the skew-normal distribution as a representative example. It is straightforward to extend our argument to any skew-elliptical distributions including the skew-t distribution which we will deal with in Appendix.

[3] Here we call Δ the skewness matrix though it does not match the skewness of the distribution in this model. For more information, see Section 2.2 and Appendix A of Harvey et al. (2010).

[4] Harvey et al. (2010) used the covariance matrix in their specification of the skew-

did not impose any restriction and assumed Δ is full matrix :

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \cdots & \delta_{1p} \\ \delta_{21} & \delta_{22} & \delta_{23} & \cdots & \delta_{2p} \\ \delta_{31} & \delta_{32} & \delta_{33} & \cdots & \delta_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ \delta_{p1} & \delta_{p2} & \delta_{p3} & \cdots & \delta_{pp} \end{bmatrix}.$$

By defining

$$Y = \begin{bmatrix} Y_1^\intercal \\ \vdots \\ Y_n^\intercal \end{bmatrix}, \quad \tilde{Y} = \begin{bmatrix} \tilde{Y}_1^\intercal \\ \vdots \\ \tilde{Y}_n^\intercal \end{bmatrix} = \begin{bmatrix} (Y_1 - \mu)^\intercal \\ \vdots \\ (Y_n - \mu)^\intercal \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1^\intercal \\ \vdots \\ Z_n^\intercal \end{bmatrix}, \quad E = \begin{bmatrix} \epsilon_1^\intercal \\ \vdots \\ \epsilon_n^\intercal \end{bmatrix},$$

(4.1) can be rewritten as

$$\tilde{Y} = Z\Delta^\intercal + E \tag{4.2}$$

Note that $Z\Delta^\intercal$ in (4.2) is

$$Z\Delta^\intercal = \begin{bmatrix} Z_{11} & Z_{12} & Z_{13} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & Z_{23} & \cdots & Z_{2p} \\ Z_{31} & Z_{32} & Z_{33} & \cdots & Z_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ Z_{n1} & Z_{n2} & Z_{n3} & \cdots & Z_{np} \end{bmatrix} \begin{bmatrix} \delta_{11} & \delta_{21} & \delta_{31} & \cdots & \delta_{p1} \\ \delta_{12} & \delta_{22} & \delta_{32} & \cdots & \delta_{p2} \\ \delta_{13} & \delta_{23} & \delta_{33} & \cdots & \delta_{p3} \\ \vdots & \vdots & \vdots & & \vdots \\ \delta_{1p} & \delta_{2p} & \delta_{3p} & \cdots & \delta_{pp} \end{bmatrix} \tag{4.3}$$

$$= \begin{bmatrix} Z_{11}\delta_{11} + Z_{12}\delta_{12} + \cdots + Z_{1p}\delta_{1p} & \cdots & Z_{11}\delta_{p1} + Z_{12}\delta_{p2} + \cdots + Z_{1p}\delta_{pp} \\ Z_{21}\delta_{11} + Z_{22}\delta_{12} + \cdots + Z_{2p}\delta_{1p} & \cdots & Z_{21}\delta_{p1} + Z_{22}\delta_{p2} + \cdots + Z_{2p}\delta_{pp} \\ \vdots & & \vdots \\ Z_{n1}\delta_{11} + Z_{n2}\delta_{12} + \cdots + Z_{np}\delta_{1p} & \cdots & Z_{n1}\delta_{p1} + Z_{n2}\delta_{p2} + \cdots + Z_{np}\delta_{pp} \end{bmatrix}.$$

Since the summation in each element of (4.3) is invariant in terms of permutation, the likelihood of Δ in the model (4.2) takes the same value

normal distribution and assumed the inverse-Wishart prior for it, which is equivalent to assuming the Wishart prior for the precision matrix in our specification. We use the precision matrix because we later examine the extended model that incorporates sparsity into the graphical structure among asset returns.

for any permutations of the columns in Δ. As a result, it is likely that the columns of Δ are randomly misaligned during the Gibbs sampler and their interpretability is lost. This problem is well-known in the field of latent factor models, which have a structure similar to the model (4.2).

As far as we know, no research[5] has examined the identification issue of Harvey et al. (2010)'s model due to the label switching problem yet. Therefore we aim to construct a modified model in which the identification issue of Δ is resolved and the interpretability is assured. Moreover, we also propose an extended model assuming a shrinkage prior to further to improve the estimation accuracy.

This chapter is organized as follows. In Chapter 4.2, we briefly review the estimation method by Harvey et al. (2010) and propose the modified method that solves the identification issue. Then we extend our proposed method by applying the shrinkage prior to the co-skewness. In Chapter 4.3, we perform simulation studies in multiple settings of the structure of Δ and verify whether proposed methods can properly estimate the true structure. The conclusion is given in Chapter 4.4.

4.2 Proposed Method

First we review the Bayesian MCMC method proposed by Harvey et al. (2010). Based on (4.1) and (4.2), two equivalent expressions of the joint

[5] Panagiotelis and Smith (2010) pointed out that, in the model by Sahu et al. (2003) or Azzalini and Capitanio (2003), it becomes difficult to identify the parameter when the skewness parameter approaches to 0, and proposed the improved model with sparsity. The identification issue we point out in this paper still occurs regardless of the magnitude of the skewness parameter when the co-skewness is taken into consideration as in Harvey et al. (2010). Note that this is a separate issue from Panagiotelis and Smith (2010). In this paper as well, we will study an extended model with sparsity of co-skewness in Chapter 4.2.

conditional density of Y given Z is obtained as:

$$p(Y|\mu, \Delta, \Omega, Z) \propto |\Omega|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{n}(Y_t - \mu - \Delta Z_t)^{\intercal}\Omega(Y_t - \mu - \Delta Z_t)\right] \tag{4.4}$$

$$\propto |\Omega|^{\frac{n}{2}} \exp\left[-\frac{1}{2}tr\left\{\Omega(\tilde{Y} - Z\Delta^{\intercal})^{\intercal}(\tilde{Y} - Z\Delta^{\intercal})\right\}\right]. \tag{4.5}$$

Harvey et al. (2010) assumed the following normal-Wishart prior for μ, δ and Ω[6]:

$$\mu \sim \mathcal{N}(b_\mu, A_\mu^{-1}), \quad \Delta \sim \mathcal{N}(b_\Delta, A_\Delta^{-1}), \quad \Omega \sim \mathcal{W}(S_\Omega^{-1}, \nu_\Omega). \tag{4.6}$$

We refer to the skew elliptical distribution with the normal- Wishart prior (4.6) as Full-NOWI. With Bayes' theorem, the posterior distribution of (μ, Δ, Ω) is obtained as

$$p(\mu, \Delta, \Omega|Y) \propto \int_0^\infty \cdots \int_0^\infty p(Y|\mu, \Delta, \Omega, Z)p(Z_1)dZ_1 \cdots p(Z_n)dZ_n p(\mu)p(\Delta)p(\Omega). \tag{4.7}$$

Since the multiple integral in (4.7) is intractable, we employ Monte Carlo integration to compute the summary statistics of parameters in the posterior distribution (4.7). For this purpose, we apply a Markov chain sampling method to draw the latent variables $(Z_1, \ldots, Z_n)$ along with the parameters (μ, Δ, Ω) from the posterior distribution (4.7).

The full conditional posterior distribution of μ, Δ, Ω, and Z_t are derived

[6] While Harvey et al. (2010) sampled μ and Δ together by jointly assuming the multivariate normal prior for them, we describe μ and Δ separately because we will later extend our proposed method to the model with a shrinkage prior for Δ.

as follows.

$$\mu|\cdot \sim \mathcal{N}\left(\hat{A}_\mu^{-1}\hat{b}_\mu, \hat{A}_\mu^{-1}\right), \quad \hat{A}_\mu = A_\mu + n\Omega, \quad \hat{b}_\mu = A_\mu b_\mu + \Omega(Y - Z\Delta^\intercal)^\intercal \iota, \tag{4.8}$$

$$\Delta|\cdot \sim \mathcal{N}(\hat{A}_\Delta^{-1}\hat{b}_\Delta, \hat{A}_\Delta^{-1}), \quad \hat{A}_\Delta = A_\Delta + Z^\intercal\tilde{\Omega}Z, \quad \hat{b}_\Delta = A_\Delta b_\Delta + Z^\intercal\tilde{\Omega}y, \tag{4.9}$$

$$\Omega|\cdot \sim \mathcal{W}\left(\hat{S}^{-1}, \hat{\nu}\right), \quad \hat{\nu} = \nu_\Omega + n, \quad \hat{S} = S_\Omega + S, \quad S = (\tilde{Y} - Z\Delta^\intercal)^\intercal(\tilde{Y} - Z\Delta^\intercal), \tag{4.10}$$

$$Z_t|\cdot \sim \mathcal{N}^+\left(\hat{A}_z^{-1}\hat{b}_z, \hat{A}_z^{-1}\right), \quad \hat{A}_z = I_p + \Delta^\intercal\Omega\Delta, \quad \hat{b}_z = \Delta^\intercal\Omega(Y_t - \mu), \tag{4.11}$$

where $Z^\intercal\tilde{\Omega}Z = \sum_{t=1}^n Z_t^\intercal \Omega Z_t$ and $Z^\intercal\tilde{\Omega}y = \sum_{t=1}^n Z_t^\intercal \Omega \tilde{Y}_t$. Since it is difficult to jointly draw Z_t from (4.11), the element-wise Gibbs sampler can be applied to (4.11). Without loss of generality, we partition Z_t, $\mu_z = \hat{A}_z^{-1}\hat{b}_z$ and $\hat{A}_z$ as

$$Z_t = \begin{bmatrix} z_{1t} \\ Z_{2t} \end{bmatrix}, \quad \mu_z = \begin{bmatrix} \mu_{z1} \\ \mu_{z2} \end{bmatrix}, \quad \hat{A}_z = \begin{bmatrix} a_{11} & a_{21}^\intercal \\ a_{21} & A_{22} \end{bmatrix},$$

where z_{1t}, μ_{z1} and a_{11} are scalars, Z_{2t}, μ_{z2} and a_{21} are $(p-1) \times 1$ vectors, and A_{22} is an $(p-1) \times (p-1)$ matrix. Then the full conditional posterior distribution of z_{1t} is

$$z_{1t}|\cdot \sim \mathcal{N}^+\left(\mu_{z1} - \frac{1}{a_{11}} a_{21}^\intercal (Z_{2t} - \mu_{z2}), \frac{1}{a_{11}}\right). \tag{4.12}$$

The full conditional posterior distribution of the second to the last element of Z_t can be derived in the same manner as (4.12). Then we can construct the element-wise Gibbs sampler for Z_t by drawing each element of Z_t sequentially from its full conditional posterior distribution.

Since columns in Δ are not identified without imposing any constraints as we confirmed in the introduction, we use a positive lower-triangular constraint (PLT, Geweke and Zhou (1996), West (2003) and Lopes and West (2004))[7] on Δ which is often used in econometric field. Assume upper-

[7] Although, Frühwirth-Schnatter and Lopes (2018) recently proposed a generalized lower triangular condition that generalizes the positive lower-triangular (GLT) condition, but in the case of the multivariate skew-elliptical distribution, the GLT condition matches the PLT condition since Δ is square matrix. Therefore, we used the PLT condition in this research.

triangular above the main diagonal of Δ equals to zero as:

$$\Delta = \begin{bmatrix} \delta_{11} & & & & \\ \delta_{21} & \delta_{22} & & & \\ \delta_{31} & \delta_{32} & \delta_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \delta_{p1} & \delta_{p2} & \delta_{p3} & \cdots & \delta_{p} \end{bmatrix}. \tag{4.13}$$

By defining

$$\Delta Z_t = W_t \delta, \quad W_t = \underbrace{\begin{bmatrix} z_{1t} & & & & & & \\ & z_{1t} & z_{2t} & & & & \\ & & & \ddots & & & \\ & & & & z_{1t} & \cdots & z_{pt} \end{bmatrix}}_{p\times\frac{p(p+1)}{2}}, \quad \delta_t = \underbrace{\begin{bmatrix} \delta_{11} \\ \delta_{21} \\ \delta_{22} \\ \vdots \\ \delta_{p1} \\ \vdots \\ \delta_{pp} \end{bmatrix}}_{\frac{p(p+1)}{2}\times 1},$$

we can rewrite (4.2) as

$$y = W\delta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tilde{\Omega}^{-1}), \tag{4.14}$$

where

$$y = \underbrace{vec(\tilde{Y}^{\mathsf{T}})}_{pn\times 1}, \quad W = \underbrace{\begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix}, \quad \epsilon = \underbrace{vec(E^{\mathsf{T}})}_{pn\times\frac{p(p+1)}{2}},}_{pn\times 1} \quad \tilde{\Omega} = \underbrace{I_n \otimes \Omega}_{pn\times pn}.$$

Using W and δ, (4.6) can be rewritten as:

$$\mu \sim \mathcal{N}(b_\mu, A_\mu^{-1}), \quad \delta \sim \mathcal{N}(b_\delta, A_\delta^{-1}), \quad \Omega \sim \mathcal{W}(S_\Omega^{-1}, \nu_\Omega). \tag{4.15}$$

We refer to the multivariate skew-elliptical distribution with the lower-triangle constraint (4.13) and the normal-Wishart prior (4.15) as LT-NOWI.

With (4.14) and (4.15), the full conditional posterior distribution of δ is derived as:

$$\delta|\cdot \sim \mathcal{N}(\hat{A}_\delta^{-1}\hat{b}_\delta, \hat{A}_\delta^{-1}), \quad \hat{A}_\delta = A_\delta + W^\intercal\tilde{\Omega}W, \quad \hat{b}_\delta = A_\delta b_\delta + W^\intercal\tilde{\Omega}y, \tag{4.16}$$

where $W^\intercal\tilde{\Omega}W = \sum_{t=1}^n W_t^\intercal \Omega W_t$ and $W^\intercal\tilde{\Omega}y = \sum_{t=1}^n W_t^\intercal \Omega \tilde{Y}_t$. The posterior distribution of Ω, μ, Z are the same as in (4.10), (4.8) and (4.11).

It is known that, when the normal-Wishart prior is used, the posterior distribution may not have a sharp peak around zero even if the true value is exactly equal to zero. In order to make the posterior distribution shrink toward zero and improve the estimation accuracy, we propose an extended model with a shrinkage prior for Δ and Ω.

To non-zero elements in Δ, we apply the horseshoe prior (Carvalho et al. (2010b)):

$$\delta_j \sim \mathcal{N}(0, \lambda_j^2\tau^2), \quad \lambda_j \sim C^+(0,1), \quad \tau \sim C^+(0,1), \quad \left(j = 1, \dots, \frac{p(p+1)}{2}\right), \tag{4.17}$$

where δ_j is the j-th element in δ and $C^+(\cdot)$ stands for the half Cauchy distribution. Note that the half Cauchy distribution in (4.17) is expressed as a mixture of inverse gamma distributions as in Makalic and Schmidt (2016):

$$\lambda_j^2|\nu_j \sim IG\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \quad \tau^2|\xi \sim IG\left(\frac{1}{2}, \frac{1}{\xi}\right), \quad \nu_j, \xi \sim IG\left(\frac{1}{2}, 1\right). \tag{4.18}$$

For computation, we first randomly generate ν_j, ξ from the inverse Gaussian distribution. Then, we also randomly generate λ_j^2, τ^2 and set them as initial values. The derivation of the full conditional posterior distribution of δ is straightforward. Given $\lambda_1, \dots, \lambda_p, \tau$, the prior distribution of δ is

$$\delta|\lambda_1, \dots, \lambda_{p^2}, \tau \sim \mathcal{N}\left(0, \tau^2 \mathrm{diag}\left(\lambda_1^2, \dots, \lambda_{p^2}^2\right)\right).$$

Thus, the full conditional posterior distribution of δ is identical to (4.16) except

$$A_\delta = \frac{1}{\tau^2}\mathrm{diag}\left(\frac{1}{\lambda_1^2}, \dots, \frac{1}{\lambda_{p^2}^2}\right), \quad b_\delta = 0.$$

The full conditional posterior distributions of λ_j^2 and τ^2 in (4.17) are

$$\lambda_j^2|\cdot \sim IG\left(1, \frac{1}{\nu_j} + \frac{\delta_j^2}{2\tau^2}\right), \quad (j = 1, \dots, p^2), \tag{4.19}$$

$$\tau^2|\cdot \sim IG\left(\frac{p^2+1}{2}, \frac{1}{\xi} + \frac{1}{2}\sum_{j=1}^{p^2} \frac{\delta_j^2}{\lambda_j^2}\right), \tag{4.20}$$

while those of the auxiliary variables are

$$\nu_j|\cdot \sim IG\left(1, 1 + \frac{1}{\lambda_j^2}\right), \quad (j = 1, \dots, p^2), \tag{4.21}$$

$$\xi|\cdot \sim IG\left(1, 1 + \frac{1}{\tau^2}\right). \tag{4.22}$$

We also apply the graphical horseshoe prior to the off-diagonal elements in Ω as in Li et al. (2019). Although it is tempting to use a horseshoe prior such as

$$\omega_{ij} \sim \mathcal{N}(0, \rho_{ij}^2\psi^2), \quad (1 \leqq i < j \leqq p), \tag{4.23}$$

$$\rho_{ij}^2|\upsilon_{ij} \sim IG\left(\frac{1}{2}, \frac{1}{\upsilon_{ij}}\right), \quad \psi^2|\zeta \sim IG\left(\frac{1}{2}, \frac{1}{\zeta}\right), \quad \upsilon_{ij}, \zeta \sim IG\left(\frac{1}{2}, 1\right), \tag{4.24}$$

where ω_{ij} is the (i, j) element in Ω, (4.23) is not appropriate for our purpose because the support of $(\omega_{ij})_{i<j}$ in (4.23) includes points where Ω is not positive definite. Thus we need to put the positive definiteness constraint upon (4.23). In this paper, we refer to the multivariate skew-elliptical distribution with the lower-triangle constraint (4.13), the horseshoe prior for the skewness matrix Δ (4.17) and the positive-definiteness-assured graphical horseshoe prior for the precision matrix Ω (4.23)–(4.24) as LT-HSGHS.

To assure the positive definiteness of Ω in the course of sampling, we apply a block Gibbs sampler by Oya and Nakatsuma (2022). To illustrate the block Gibbs sampler, we introduce the following partition of Ω and S:

$$\Omega = \begin{bmatrix} \Omega_{11} & \omega_{12} \\ \omega_{12}^\intercal & \omega_{22} \end{bmatrix}, \quad S = \begin{bmatrix} S_{11} & s_{12} \\ s_{12}^\intercal & s_{22} \end{bmatrix}, \tag{4.25}$$

where ω_{22} and s_{22} are scalars, ω_{12} and s_{12} are $(p-1) \times 1$ vectors, and Ω_{11} and S_{11} are $(p-1) \times (p-1)$ matrices. In each step of the block Gibbs sampler, we draw a diagonal element ω_{22} and off-diagonal elements ω_{12} from their full conditional posterior distributions. Without loss of generality, rows and columns of Ω can be rearranged so that the lower-right corner of Ω, ω_{22}, should be the diagonal element to be drawn from its full conditional posterior distribution. By using Ω and S in (4.25), we have

$$\mathrm{tr}\left(\Omega S\right) = s_{22}\omega_{22} + 2s_{12}^{\intercal}\omega_{12} + \mathrm{tr}\left(\Omega_{11}S_{11}\right),$$

and

$$|\Omega| = \left|\omega_{22} - \omega_{12}^{\intercal}\Omega_{11}^{-1}\omega_{12}\right| \left|\Omega_{11}\right|.$$

Then (4.5) is rewritten as

$$\begin{aligned} p(Y|\mu,\delta,\Omega,Z) &\propto |\Omega|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\Omega S)\right] \\ &\propto \left|\omega_{22} - \omega_{12}^{\intercal}\Omega_{11}^{-1}\omega_{12}\right|^{\frac{n}{2}} \left|\Omega_{11}\right|^{\frac{n}{2}} \\ &\quad \times \exp\left[-\frac{1}{2}\left\{s_{11}\omega_{22} + 2s_{12}^{\intercal}\omega_{12} + \mathrm{tr}\left(\Omega_{11}S_{11}\right)\right\}\right]. \end{aligned} \tag{4.26}$$

Furthermore, following Wang (2012), we reparameterize $(\omega_{22}, \omega_{12})$ to (η, ω_{12}) where

$$\eta = \omega_{22} - \omega_{12}^{\intercal}\Omega_{11}^{-1}\omega_{12}.$$

Finally we have

$$p(Y|\mu,\delta,\Omega,Z) \propto \eta^{\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{s_{22}\eta + s_{22}\omega_{12}^{\intercal}\Omega_{11}^{-1}\omega_{12} + 2s_{12}^{\intercal}\omega_{12}\right\}\right], \tag{4.27}$$

where we ignore the parts that do not depend on η nor ω_{12}.

We need to be careful in choosing the prior distribution of (η, ω_{12}). Given that Ω from the previous iteration of the block Gibbs sampler is positive definite, newly generated ω_{22} and ω_{12} must satisfy

$$\omega_{22} > \omega_{12}^{\intercal}\Omega_{11}^{-1}\omega_{12}, \tag{4.28}$$

to ensure that the updated Ω is also positive definite. This condition (4.28) requires

$$\eta = \omega_{22} - \omega_{12}^{\intercal}\Omega_{11}^{-1}\omega_{12} > 0.$$

Hence, we can use a gamma distribution:

$$\eta \sim Ga(a_\eta, b_\eta), \tag{4.29}$$

as the prior distribution of η. Moreover, we suppose the prior distribution of off-diagonal elements ω_{12} is a truncated multivariate normal distribution:

$$p(\omega_{12}|\omega_{22}, \Omega_{11}) \propto \exp\left(-\frac{1}{2}\omega_{12}^\intercal A_\omega \omega_{12}\right) \mathbf{1}_{M^+}(\omega_{12}), \tag{4.30}$$

where

$$A_\omega = \frac{1}{\psi^2}\mathrm{diag}\left(\frac{1}{\rho_{12}^2}, \dots, \frac{1}{\rho_{1p}^2}\right), \quad M^+ = \{\omega_{12} : \omega_{22} < \omega_{12}^\intercal \Omega_{11}^{-1}\omega_{12}\},$$

in order to assure that the condition (4.28) holds in the course of sampling. Applying Bayes' theorem to (4.29) and (4.27), we have

$$\eta|\cdot \sim Ga\left(a_\eta + \frac{n}{2}, b_\eta + \frac{s_{22}}{2}\right). \tag{4.31}$$

With (4.30), (4.24) and (4.27), the full conditional posterior distribution of ω_{12} is derived as

$$\omega_{12}|\cdot \sim \mathcal{N}\left(-\hat{A}_\omega^{-1}s_{12},\ \hat{A}_\omega^{-1}\right) \mathbf{1}_{M^+}(\omega_{12}), \quad \hat{A}_\omega = A_\omega + s_{22}\Omega_{11}^{-1}. \tag{4.32}$$

In order to draw ω_{12} from (4.32), we apply the Hit-and-Run algorithm (Bélisle et al. (1993)) as in Oya and Nakatsuma (2022).

Step 1: Pick a point α on the unit sphere randomly as $\alpha = \frac{u}{\|u\|}$, $u \sim \mathcal{N}(0, I)$.

Step 2: Draw a random scalar κ from $\mathcal{N}\left(\mu_\kappa, \sigma_\kappa^2\right)\mathbf{1}_{R^+}(\kappa)$ where

$$\begin{aligned}
\mu_\kappa &= -\frac{s_{12}^\intercal\alpha + \omega_{12}^\intercal\hat{A}_\omega\alpha}{\alpha^\intercal\hat{A}_\omega\alpha}, \quad \sigma_\kappa^2 = \frac{1}{\alpha^\intercal\hat{A}_\omega\alpha}, \\
R^+ &= \left\{\kappa : \frac{-b_\kappa - \sqrt{b_\kappa^2 - a_\kappa c_\kappa}}{a_\kappa} < \kappa < \frac{-b_\kappa + \sqrt{b_\kappa^2 - a_\kappa c_\kappa}}{a_\kappa}\right\}, \\
a_\kappa &= \alpha^\intercal\Omega_{11}^{-1}\alpha, \quad b_\kappa = \omega_{12}^\intercal\Omega_{11}^{-1}\alpha, \quad c_\kappa = \omega_{12}^\intercal\Omega_{11}^{-1}\omega_{12} - \omega_{22}.
\end{aligned}$$

Step 3: Update the old ω_{12} with $\omega_{12} + \kappa\alpha$.

Finally it is straightforward to derive the full conditional posterior distributions of hyper-parameters and auxiliary variables:

$$\rho_{ij}^2|\cdot \sim IG\left(1, \frac{1}{\upsilon_{ij}} + \frac{\omega_{ij}^2}{2\psi^2}\right), \quad (1 \leqq i < j \leqq p), \tag{4.33}$$

$$\psi^2|\cdot \sim IG\left(\frac{p(p-1)}{4} + \frac{1}{2}, \ \frac{1}{\zeta} + \frac{1}{2}\sum_{j=2}^{p}\sum_{i=1}^{j-1}\frac{\omega_{ij}^2}{\rho_{ij}^2}\right), \tag{4.34}$$

$$\upsilon_{ij}|\cdot \sim IG\left(1, 1 + \frac{1}{\rho_{ij}^2}\right), \quad (1 \leqq i < j \leqq p), \tag{4.35}$$

$$\zeta|\cdot \sim IG\left(1, 1 + \frac{1}{\psi^2}\right). \tag{4.36}$$

Although we will examine the skew-normal distribution in the next section in order to simply compare the Harvey et al. (2010)'s sampling method in terms of identification of Δ, we can easily extend the multivariate skew-normal model (4.1) to the multivariate skew-t distribution. See Appendix.

4.3 Performance Comparisons with Simulation

In this section, we report results of Monte Carlo experiments to compare three models (Full-NOWI, LT-NOWI and LT-HSGHS), which are summarized in Table 4.1, in terms of accuracy in the parameter estimation.

Table 4.1: Overview of comparative models

	Constraint for Δ	Prior for Δ	Prior for Ω
Full-NOWI	Nothing	Normal	Wishart
LT-NOWI	Positive Lower-Triangular	Normal	Wishart
LT-HSGHS	Positive Lower-Triangular	Horseshoe	Graphical Horseshoe

We assume the following three designs of Δ in this simulation:

1. Δ-Diag: $\Delta_{ii} = 2.0 \quad (i = 2p-1)$, $\Delta_{ii} = -2.0 \quad (i = 2p)$, otherwise 0.0.

2. Δ-Sparse: $\Delta_{ii} = 2.0 \quad (i = 2p - 1)$, $\Delta_{ii} = -2.0 \quad (i = 2p)$, $\Delta_{i,i-1} = -1.0$, otherwise 0.0.

3. Δ-Dense: $\Delta_{ii} = 2.0 \quad (i = 2p{-}1)$, $\Delta_{ii} = -2.0 \quad (i = 2p)$, $\Delta_{i,i-1} = -1.0$, otherwise the elements in lower-triangular equals to 1.0.

Since our main purpose is to compare estimation of Δ, we set a simple assumption for the other parameter; Ω is the identity matrix, μ is fixed to zero. We generate artificial data of $n = 1,500$ and $p = 15$ from the multivariate skew-elliptical distribution with each specification and evaluate the posterior statistics of each parameter via MCMC. The hyper-parameters in the prior distributions are set up as follows.

Full-NOWI $b_\mu = 0$, $A_\mu = 0.01I$, $b_\Delta = 0$, $A_\Delta = 0.01I$, $S_\Omega = pI$, $\nu_\Omega = p$ in (4.6).

LT-NOWI $b_\mu = 0$, $A_\mu = 0.01I$, $b_\delta = 0$, $A_\delta = 0.01I$, $S_\Omega = pI$, $\nu_\Omega = p$ in (4.15).

LT-HSGHS $b_\mu = 0$, $A_\mu = 0.01I$, $b_\delta = 0$, $A_\delta = 0.01I$ in (4.15); $a_\eta = 1.0$ and $b_\eta = 0.0$ in (4.29); $A_\omega = 0.01I$ in (4.30)

In all cases, the number of burn-in iterations were 50,000, and the Monte Carlo sample from the following 100,000 iterations was used in the Bayesian inference. Also, we repeated simulations 30 times for each setup and obtained a set of point estimates of Δ and Ω. All computations are implemented with Python 3.7.0 on a desktop PC with 128GB RAM, 8GB GPU and eight-core 3.8GHz i7-10700K Intel processor.

To compare the three models in terms of accuracy in the point estimation of Δ and Ω, we computed the Frobenius norm, as measurement of discrepancy between the point estimate and the true structure. Tables 4.2 and 4.3 show the sample median loss with 30 replications for three models. The figures in parentheses are the standard errors. The smaller the value of the Frobenious norm, the closer the estimated structure is to the true one. In addition, in order to make the estimation results visually easy to understand, the posterior averages of Δ and Ω of each model in the 30th replication are shown in Figures 4.1 – 4.6.

First, regarding Δ, the Frobenious norm of the proposed models (LT-NOWI, LT-HSGHS) have decreased to 1/8 or less of the Full-NOWI model for all designs and the estimation accuracy has remarkably improved in Table 4.2. This is because the columns of Δ is not identified at all in Full-NOWI. On the other hand, this identification issue is resolved in LT-NOWI and LT-HSGHS and the structure of Δ can be estimated well with the proposed method as shown in Figures 4.1, 4.3 and 4.5. Furthermore, for the Δ-Diag case and the Δ-Sparse case, Table 4.2 reports that the Frobenious norm of LT-HSGHS is less than half the value of LT-NOWI. This is because that the horseshoe prior in LT-HSGHS contributes to the estimation performance by shrinking non-essential elements to zero, while a large amount of non-zero entries still remain in Δ for LT-NOWI as shown in Figures 4.1 and 4.3. However, the difference in the Frobenious norm between LT-NOWI and LT-HSGHS becomes smaller in the Δ-Dense design because the sparse assumption of Δ is not satisfied in this case.

Next, let take a look at results on Ω. Note that the true structure of Ω is the identity matrix. We examine how the estimation accuracy of Ω changes across the structural designs of Δ. For all designs of Δ, the estimation accuracy is significantly improved in LT-HSGHS, the value of Frobenious norm is 1/3 or less in Table 4.3 compared with Full-NOWI and LT-NOWI. In fact, there are a lot of non-zero entries in the off-diagonal elements in Full-NOWI and LT-NOWI in Figures 4.2, 4.4 and 4.6. On the other hand, the posterior mean of Ω in LT-HSGHS becomes the diagonal matrix thanks to the shrinkage effect. Also, comparing LT-NOWI with Full-NOWI, the Frobenious norm is slightly smaller in LT-NOWI for all Δ designs. These findings suggest that the posterior distribution of Ω is affected by the estimation of Δ as shown in (4.10).

Table 4.2: Sample median loss in the point estimation of $\boldsymbol{\Delta}$

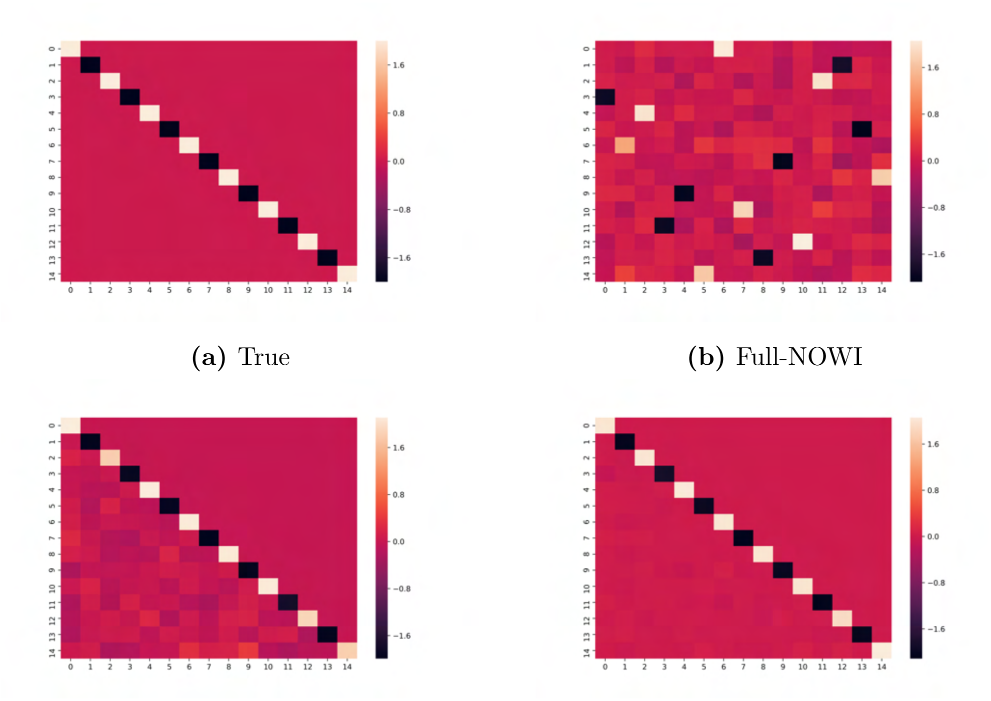
	Δ-Diag	Δ-Sparse	Δ-Dense
Frobenius norm			
Full-NOWI	10.587	11.588	12.620
	(0.534)	(0.743)	(0.723)
LT-NOWI	1.344	1.362	1.420
	(0.083)	(0.115)	(0.108)
LT-HSGHS	**0.380**	**0.617**	**1.214**
	(0.070)	(0.075)	(0.138)

Notes: (a) The smaller losses are boldfaced.
(b) The figures in parentheses are the standard errors.

Table 4.3: Sample median loss in the point estimation of $\boldsymbol{\Omega}$

	Δ-Diag	Δ-Sparse	Δ-Dense
Frobenius norm			
Full-NOWI	2.550	2.447	2.352
	(0.139)	(0.138)	(0.136)
LT-NOWI	2.255	2.238	2.210
	(0.179)	(0.147)	(0.155)
LT-HSGHS	**0.393**	**0.479**	**0.724**
	(0.101)	(0.160)	(0.288)

Notes: (a) The smaller losses are boldfaced.
(b) The figures in parentheses are the standard errors.



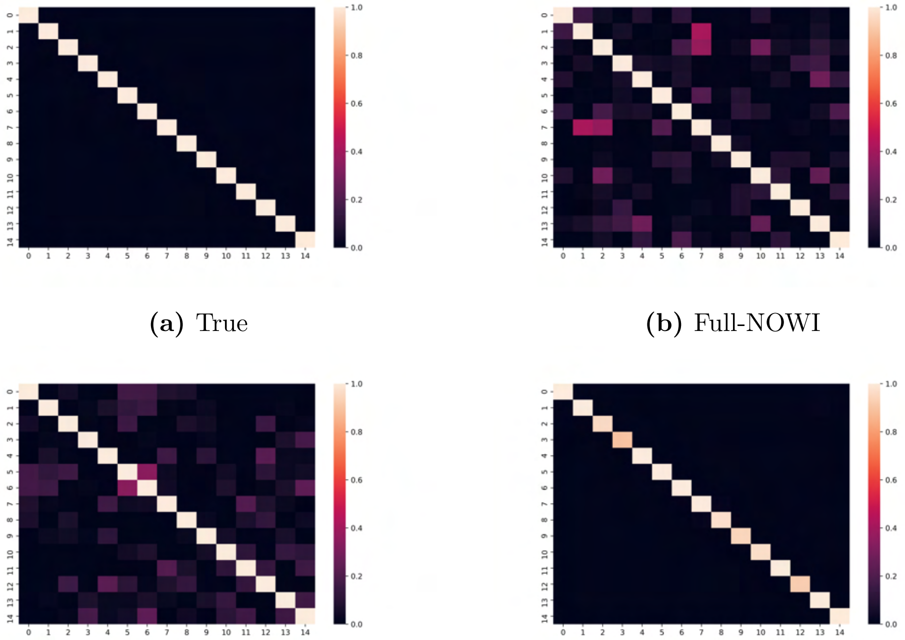
(a) True

(b) Full-NOWI

(c) LT-NOWI

(d) LT-HSGHS

Figure 4.1: Δ-Diag: True Structure of Δ and estimated Δ



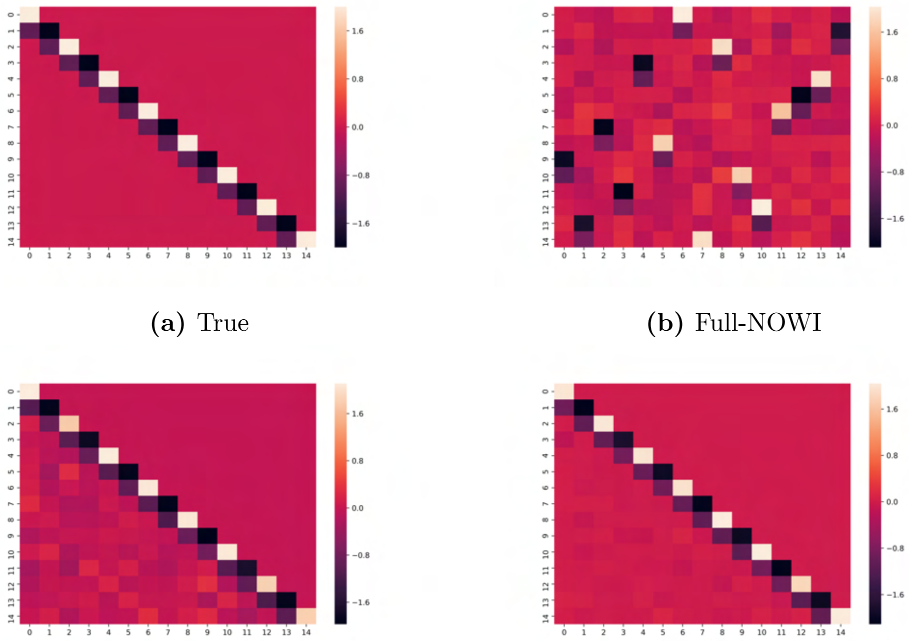
(a) True

(b) Full-NOWI

(c) LT-NOWI

(d) LT-HSGHS

Figure 4.2: Δ-Diag: True Structure of Ω and estimated Ω



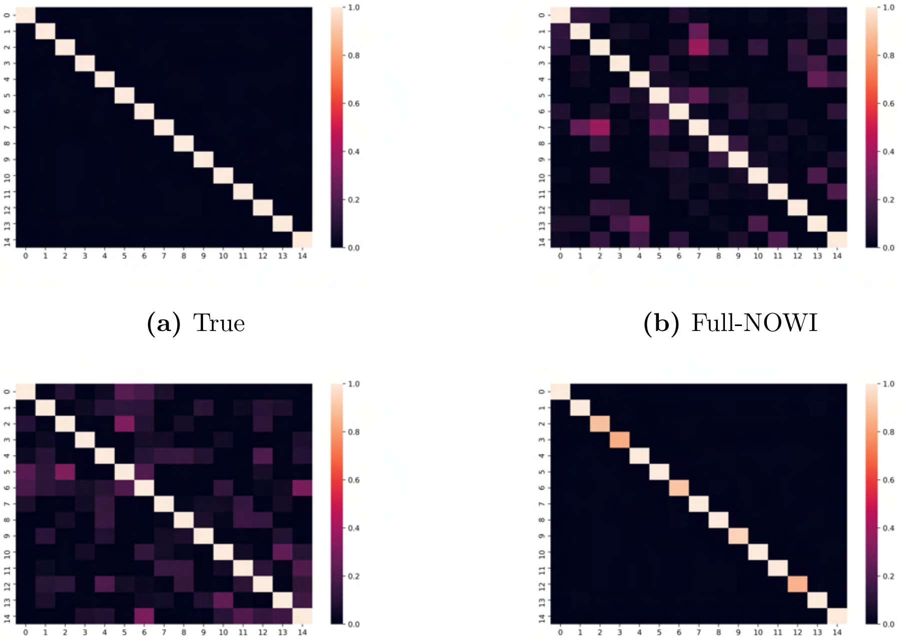
(a) True

(b) Full-NOWI

(c) LT-NOWI

(d) LT-HSGHS

Figure 4.3: Δ-Sparse: True Structure of Δ and estimated Δ



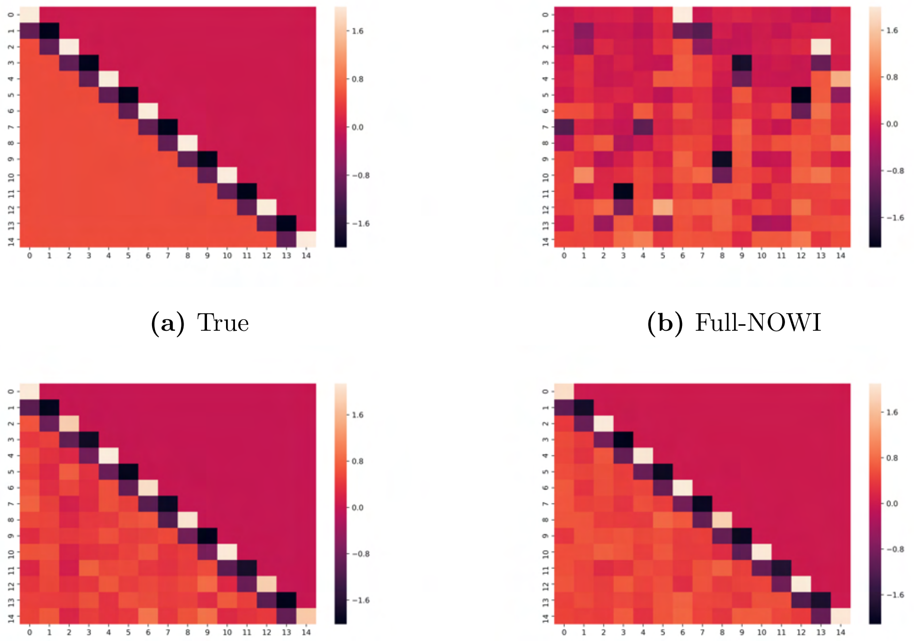
(a) True

(b) Full-NOWI

(c) LT-NOWI

(d) LT-HSGHS

Figure 4.4: Δ-Sparse: True Structure of Ω and estimated Ω



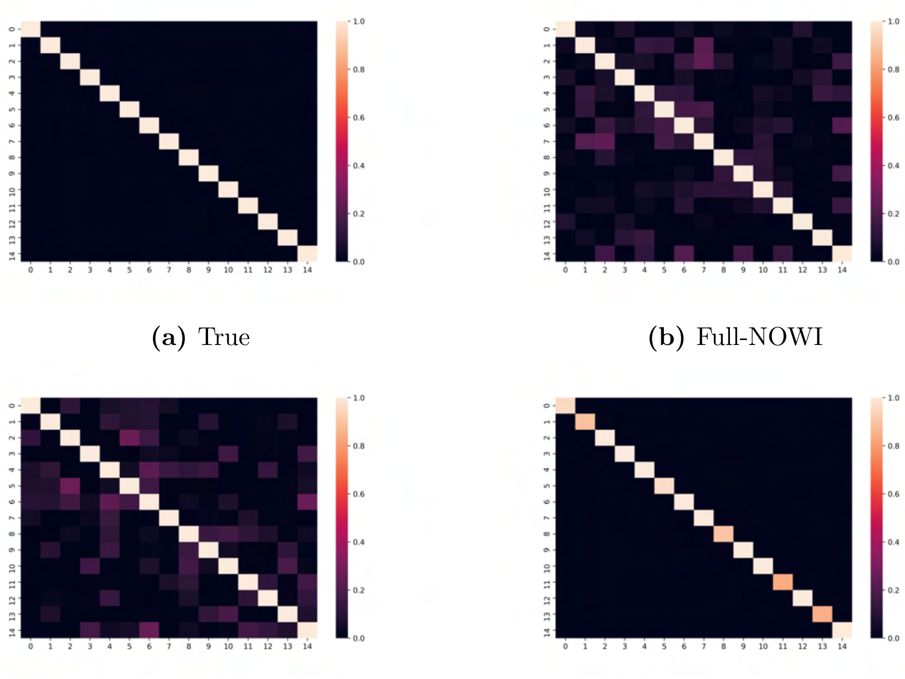
(a) True

(b) Full-NOWI

(c) LT-NOWI

(d) LT-HSGHS

Figure 4.5: Δ-Dense: True Structure of Δ and estimated Δ



(a) True

(b) Full-NOWI

(c) LT-NOWI

(d) LT-HSGHS

Figure 4.6: Δ-Dense: True Structure of Ω and estimated Ω

4.4 Conclusion

In this paper, we have raised a possible identification issue on the skewness matrix of the skew-elliptical distribution in the Bayesian MCMC method proposed by Harvey et al. (2010) due to label switching. To avoid this issue, we proposed a modified model in which the lower-triangular constraint was imposed upon the skewness matrix. Moreover, we devised an extended model with the horseshoe prior for both skewness matrix and precision matrix to further improve the estimation accuracy.

In the simulation study, we compared the proposed models with the model of Harvey et al. (2010) in the three structural designs of the skewness matrix and found that the proposed models with the identification constraint significantly improved the estimation accuracy of the skewness matrix.

4.5 Appendix: Extension to multivariate skew-t distribution

As we mentioned before, Harvey et al. (2010) developed the Gibbs sampling algorithm for the multivariate skew-normal distribution (4.1), but it is straightforward to extend it to the multivariate skew-t distribution as Sahu et al. (2003) showed. Since it is expressed as a scale mixture of multivariate skew-normal distributions, skew-t distributed Y_t is expressed as

$$\begin{aligned}
&Y_t = \mu + \Delta Z_t + \epsilon_t, \\
&Z_t \sim \mathcal{N}^+\left(0, \frac{1}{\gamma_t} I_p\right), \quad \epsilon_t \sim \mathcal{N}\left(0, (\gamma_t \Omega)^{-1}\right), \quad \gamma_t \sim Ga\left(\frac{\varphi}{2}, \frac{\varphi}{2}\right), \qquad (4.37) \\
&Z_t \perp \epsilon_t \perp \gamma_t,
\end{aligned}$$

Given γ_t, the sampling algorithms for δ, Ω, μ and Z_t in (4.37) are almost identical to the multivariate skew-normal case except that

δ: redefine $\hat{A}_\delta$ and $\hat{b}_\delta$ in (4.16) as

$$\hat{A}_\delta = A_\delta + \sum_{t=1}^{n} \gamma_t W_t^\intercal \Omega W_t, \quad \hat{b}_\delta = A_\delta b_\delta + \sum_{t=1}^{n} \gamma_t W_t^\intercal \Omega \tilde{Y}_t.$$

Ω**:** redefine S in (4.10) as

$$S = \sum_{t=1}^{n} \gamma_t (Y_t - \mu - \Delta Z_t)(Y_t - \mu - \Delta Z_t)^{\intercal}.$$

μ**:** redefine $\hat{A}_\mu$ and $\hat{b}_\mu$ in (4.8) as

$$\hat{A}_\mu = A_\mu + \sum_{t=1}^{n} \gamma_t \Omega, \quad \hat{b}_\mu = A_\mu b_\mu + \sum_{t=1}^{n} \gamma_t \Omega (Y_t - \Delta Z_t).$$

Z_t**:** redefine $\hat{A}_z$ and $\hat{b}_z$ in (4.11) as

$$\hat{A}_z = \gamma_t \left(I_p + \Delta^{\intercal} \Omega \Delta\right), \quad \hat{b}_z = \gamma_t \Delta^{\intercal} \Omega (Y_t - \mu).$$

Finally, with the prior $\varphi \sim Ga(a_\varphi, b_\varphi)$, the full conditional posterior distribution of φ is derived as

$$\begin{aligned} p(\varphi|\cdot) &\propto \prod_{t=1}^{n} \frac{\left(\frac{\varphi}{2}\right)^{\frac{\varphi}{2}}}{\Gamma\left(\frac{\varphi}{2}\right)} \gamma_t^{\frac{\varphi}{2}-1} \exp\left(-\frac{\varphi\gamma_t}{2}\right) \times \varphi^{a_\varphi - 1} \exp(-b_\varphi \varphi) \\ &\propto \frac{\left(\frac{\varphi}{2}\right)^{\frac{\varphi n}{2}}}{\Gamma\left(\frac{\varphi}{2}\right)^n} \left(\prod_{t=1}^{n} \gamma_t\right)^{\frac{\varphi}{2}-1} \exp\left(-\frac{\varphi}{2}\sum_{t=1}^{n}\gamma_t\right) \times \varphi^{a_\varphi - 1} \exp(-b_\varphi \varphi) \\ &\propto \exp\left[\left(\frac{\varphi n}{2} + a_\varphi - 1\right)\log\varphi - n \log\Gamma\left(\frac{\varphi}{2}\right) - \hat{b}_\varphi \varphi\right], \qquad (4.38) \end{aligned}$$

where

$$\hat{b}_\varphi = b_\varphi + \frac{\log 2}{2} n + \frac{1}{2}\sum_{t=1}^{n} \left(\gamma_t - \log \gamma_t\right).$$

Following Watanabe (2001), we may apply a Metropolis-Hastings algorithm to draw φ from (4.38). For this purpose, we consider the second-order Taylor approximation of

$$f(\varphi) = \left(\frac{\varphi n}{2} + a_\varphi - 1\right)\log\varphi - n \log\Gamma\left(\frac{\varphi}{2}\right) - \hat{b}_\varphi \varphi,$$

within the exponential function of (4.38), that is,

$$f(\varphi) \approx f(\varphi^*) + \nabla f(\varphi^*)(\varphi - \varphi^*) + \frac{1}{2}\nabla^2 f(\varphi^*)(\varphi - \varphi^*),$$

where

$$\begin{aligned}\nabla f(\varphi) &= \frac{n}{2}\log\varphi + \frac{n}{2} + \frac{a_\varphi - 1}{\varphi} - \frac{n}{2}\nabla\log\Gamma\left(\frac{\varphi}{2}\right) - \hat{b}_\varphi,\\ \nabla^2 f(\varphi) &= \frac{n}{2}\left(\frac{1}{\varphi} - \frac{1}{2}\nabla^2\log\Gamma\left(\frac{\varphi}{2}\right)\right) - \frac{a_\varphi - 1}{\varphi^2}.\end{aligned}$$

Note that f is globally concave and has a unique mode. If we take the mode of f as φ^*, we have $\nabla f(\varphi^*) = 0$. Thus the pdf of the full conditional posterior distribution (4.38) is approximated as

$$p(\varphi|\cdot) \approx \mathcal{K}\exp\left[\frac{1}{2}\nabla^2 f(\varphi^*)(\varphi - \varphi^*)\right],$$

Therefore we can use

$$\varphi \sim \mathcal{N}^+\left(\varphi^*,\ \left\{-\nabla^2 f(\varphi^*)\right\}^{-1}\right)$$

as the proposal distribution of φ in the Metropolis-Hastings algorithm.

Chapter 5

Concluding Remarks

In this doctoral dissertation, we tackled two challenges in modeling asset returns; (1) $p > n$ problem in case of high-dimensional data with many assets and (2) modeling of a skewed and fat-tailed distribution with real-world data. Chapter 2 and 3 examined a solution for the former with Bayesian graphical models while Chapter 4 focused on the latter.

In Chapter 2, we pointed out a serious issue on the Gibbs sampling algorithm for the Bayesian graphical LASSO model proposed by Wang (2012). As demonstrated with the simulation study in Chapter 2, Wang (2012)'s algorithm cannot guarantee the positive definiteness of the precision matrix in the graphical model in each cycle of the Gibbs sampling procedure because the positive definite condition of the precision matrix is satisfied only for the diagonal elements, but not always for the off-diagonal elements. This is due to the fact that Wang (2012)'s algorithm generates the off-diagonal elements from the incorrect full conditional posterior distribution, i.e., the unconstrained multivariate normal distribution.

To solve this issue, we proposed a modification of the original algorithm so that it should guarantee the positive definiteness of the precision matrix. The modified algorithm generates the off-diagonal elements from a truncated multivariate normal distribution whose support is the region wherein the updated precision matrix remains positive definite by utilizing Bélisle et al.

(1993)'s hit-and-run algorithm. It was shown that the proposed algorithm dramatically improved accuracy in point estimation and graphical structure learning in the simulation study for all 12 scenarios.

In Chapter 3, we developed a data-driven portfolio framework based on Bayesian graphical LASSO proposed in Chapter 2. We conducted 10-year out-of-sample portfolio management experiments from 2011 to 2020 with monthly return data of US 100 portfolios provided by Kenneth French. We tested five scenarios: (p, n) = (100, 120), (100, 60), (100, 12), (100, 6) and (100, 3) by changing the estimation period to analyze portfolio performance in different p/n ratios. In the experiment, we compared the global minimum variance portfolios based on the proposed Bayesian LASSO, two types of non-Bayesian graphical LASSO (graphical LASSO with both diagonal and off-diagonal elements shrinkage), portfolios based on other types of dimension compression method such as random matrix theory filtering (Bouchaud and Potters (2009)) and Ledoit-Wolf shrinkage estimation (Ledoit and Wolf (2004)), the traditional sample covariance approach, and the equal weight approach as a benchmark. As far as We know, this is the first attempt to apply Bayesian graphical LASSO in case of $p > n$ in the literature of portfolio selection, although there have been researches (e.g., Torri et al. (2019)) that applied non-Bayesian graphical models to portfolio selection in case of $p < n$.

In terms of return-risk tradeoff and portfolio composition, the results showed the advantages of the proposed Bayesian approach over the others. Both Sharpe ratios and indices of portfolio composition were relatively stable for the proposed approach while they were either unstable for non-Bayesian graphical LASSO. It is also considered as an important contribution that the proposed approach is a simple method that combines a factor model, which is already widely used in the field of finance, with the Bayesian graphical model, and it is applicable without changing the current business process of investors significantly.

In Chapter 4, we discussed parameter estimation of the multivariate skew-elliptical distribution which could describe the characteristics of asset returns such as skewness and fat-tail. The multivariate skew-elliptical distribution can represent the skewness dependency among assets unlike the GH distribution, which is important for application to portfolio management. We introduced Harvey et al. (2010)'s Bayesian estimation method as a pioneer research in this field, but raised a problem that Harvey et al. (2010)'s method cannot identify all elements of the skewness matrix because the likelihood of the Harvey et al. (2010)'s model takes the same value for any permutations of the columns in the skewness matrix. Once the columns of the skewness matrix are randomly misaligned in the Gibbs sampler, their interpretability is lost.

To solve the issue, we proposed a modified model with the lower-triangular constraint (Geweke and Zhou (1996), West (2003) and Lopes and West (2004)) on the skewness matrix. In addition, we extended the model with the horseshoe prior for the skewness matrix and the precision matrix. In the simulation study, we compared the proposed models with the model of Harvey et al. (2010) in three structural designs of the skewness matrix; Diag, Sparse, and Dense. It was shown that the proposed models with the identification constraint could successfully estimate the true structure of the skewness dependency in all designs while the Harvey et al. (2010)'s model suffered from the identification issue. Moreover, the extended model with the horse prior achieved further improvement in the estimation accuracy for both Diag and Sparse.

To conclude this dissertation, we point out remaining challenges in large-scale portfolio management for future research. The new algorithm for Bayesian graphical model and the portfolio approach based on it proposed in Chapters 2 and 3 enables us to stably estimate a large-scale precision matrix with high estimation accuracy. In our opinion, however, there are two remaining challenges.

The first one is a heavy computation load of the the proposed model, though non-Bayesian graphical models may also suffer from the same problem in case of adaptive LASSO. When estimation of the proposed model is implemented in a programming language with slow matrix computation like Python, a practical upper limit of the dimension of the precision matrix is about 1,100 even if we use a high-performance desktop PC with 128GB RAM, 8GB GPU and eight-core 3.8GHz i7-10700K Intel processor. It may be improved to some extent by using a fast matrix computing language such as Julia or MATLAB, but scalability needs to be significantly improved, given the fact that the number of stocks traded in the world is more than 40,000.

The second one is that graphical models theoretically assume that the precision matrix to be estimated is a sparse matrix, but the precision matrix of raw asset return data is largely influenced by common factors and, as a result, it is not sparse. Therefore, it is necessary to combine it with a factor model in advance to remove the influence of common factors before estimating the precision matrix. Although the proposed model itself can work even in case of $p > n$, it is still difficult to make an accurate estimation when a general factor model is combined with the preprocessing[1]. As a solution to this issue, we may apply Bayesian compressed regression (BCR) proposed by Guhaniyogi and Dunson (2015), which was introduced as one of the related studies in Chapter 3.

As for the proposed Bayesian estimation of the multivariate skew-elliptical distribution in Chapter 4, its stable computation can be performed only in case of $p << n$. This limitation may hinder us from applying it to large-scale portfolio management, though this issue arises not only for the skew-elliptical distribution but also for other distributions in the literature of skew distribution. Its heavy computation load for sampling latent variables would be an obstacle to its wider acceptance among practitioners. These problems may be improved by imposing another kind of identification constraint proposed in recent years on the skewness matrix. Moreover, since the likelihood of

[1] In the experiment in Chapter 3, we first estimated the three-factor model for the whole sample period and used the residuals in each case, but such operation is difficult in practice.

the skew-elliptical distribution is similar to that of a latent variable regression model, it has a possibility that we can reduce the computational load by applying the Bayesian compressed regression by Guhaniyogi and Dunson (2015).

Bibliography

K. Aas and I. H. Haff. The generalized hyperbolic skew student's t-distribution. *Journal of Financial Econometrics*, 4(2):275–309, 2006.

C. Adcock and A. Azzalini. A selective overview of skew-elliptical and related distributions and of their applications. *Symmetry*, 12(1), 2020.

R. Alhamzawi, K. Yu, and D. F. Benoit. Bayesian adaptive lasso quantile regression. *Statistical Modelling*, 12(3):279–297, 2012.

M. T. Alodat and M. Y. Al-Rawwash. The extended skew gaussian process for regression. *METRON*, 72(3):317–330, 2014.

A. Armagan, D. B. Dunson, and J. Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013.

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 367–389, 2003.

A. Azzalini and D. Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

M. Barbi and S. Romagnoli. Skewness, basis risk, and optimal futures demand. *International Review of Economics & Finance*, 58:14–29, 2018.

O. E. Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. A. Mathmatical and Physical Sciences*, 353:401–419, 1977.

C. J. P. Bélisle, H. E. Romeijn, and R. L. Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathmatics of Operations Research*, 18(2):255–266, 1993.

F. Black and R. Litterman. Asset allocation: Combining investor views with market equilibrium. *Journal of Fixed Income*, 1:7–18, 1991.

F. Black and R. Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

J.P. Bouchaud and M. Potters. Financial applications of random matrix theory: a short review. *arXiv:0910.1205*, 2009.

M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

C. Brownlees, E. Nualart, and Y. Sun. Realized networks. *Journal of Applied Econometrics*, 33(7):986–1006, 2018.

B. Carmichael and A. Coën. Asset pricing with skewed-normal return. *Finance Research Letters*, 10(2):50–57, 2013.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, 5:73–80, 2009.

C. M. Carvalho, N. G. Polson, and J. G. Scott. *The Horseshoe Estimator for Sparse Signals. Biometrika*, 97:465–480, 2010a.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010b.

V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. In *In Proceedings of the twenty-fourth conference on uncertainty in artificial intelligence*, pages 153–160. AUAI Press, 2008.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *Annals of Applied Statistics*, 3(2):521–541, 2009.

J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498): 592–606, 2012.

C. Fernández and M. F. J. Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Annals of Applied Statistics*, 5(2A):1057–1080, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 5(1):432–441, 2008.

S. Frühwirth-Schnatter and H. F. Lopes. Sparse bayesian factor analysis when the number of factors is unknown. *arXiv: 1804.04231*, 2018.

A Gelman and D Rubin. Inference from iterative simulation using multiple sequences" (with discussion). *Statistical Science*, 7(4):457–511, 1992.

J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9(2):557–587, 1996.

S. Goto and Y. Xu. Improving mean variance optimization through sparse hedging restrictions. *Journal of Financial and Quantative Analysis*, 50(6): 1415–1441, 2015.

R. Guhaniyogi and D. B. Dunson. Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. *Joint estimation of multiple graphical models. Biometrika*, 98(1):1–15, 2011.

B. E. Hansen. Autoregressive conditional density estimation. *International Economic Review*, 35(3):705–730, 1994.

C. R. Harvey, J. C. Liechty, M. W. Liechty, and P. Müller. Portfolio selection with higher moments. *Quantitative Finance*, 10(5):469–485, 2010.

C. R. Harvey, Y. Liu, and H. Zhu. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

C. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15:2911–2947, 2014.

Z. Khondker, H. Zhu, W. Lin, and J. Ibrahim. The bayesian covariance lasso. *Statistics and Its Inference*, 6(2):243–259, 2013.

S. J. Kon. Models of stock returns—a comparison. *The Journal of Finance*, 39(1):147–165, 1984.

G. Koop, D. Korobilis, and D. Pettenuzzo. Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210:135–154, 2019.

L. Laloux, P. Cizeau, JP. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1469, 1999.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

C. Leng, MN. Tran, and D. Nott. Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66:221–244, 2014.

Y. Li, B. A. Craig, and A. Bhadra. The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757, 2019.

H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67, 2004.

J. Luo and L. Chen. Realized volatility forecast with the bayesian random compressed multivariate har model. *International Journal of Forecasting*, 36(3):781–799, 2020.

E. Makalic and D. F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2016.

H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

H. Markowitz and N. Usmen. The likelihood of various stock market return distributions, part 2: Empirical results. *Journal of Risk and Uncertainty*, 13(3):221–247, 1996.

B. M. Marlin and K. P. Murphy. Sparse gaussian graphical models with unknown block structure. In *Proceedings of the 26th Annual International Conference on Machine Learning*, number ICML ’09, pages 705–712, New York, NY, USA, 2009. Association for Computing Machinery.

B. M. Marlin, M. Schmidt, and K. P. Murphy. Group sparse priors for covariance estimation. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, number UAI ’09, pages 383–392, Arlington, Virginia, USA, 2009. AUAI Press.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

T.C. Mills. Modelling skewness and kurtosis in the london stock exchange ft-se index return distributions. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(3):323–332, 1995.

J. Nakajima. Bayesian analysis of multivariate stochastic volatility with skew return distribution. *Econometric Reviews*, 36(5):546–562, 2017.

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(11):3690–3704, 2012.

S. Oya. A bayesian graphical approach for large-scale portfolio management with fewer historical data. *Asia-Pacific Financial Markets*, 2022.

S. Oya and T. Nakatsuma. Identification in bayesian estimation of the skewness matrix in a multivariate skew-elliptical distribution. *arXiv:2108.04019*, 2021.

S. Oya and T. Nakatsuma. A positive-definiteness-assured block gibbs sampler for bayesian graphical models with shrinkage priors. *Japanese Journal of Statistics and Data Science*, 2022.

A. Panagiotelis and M. Smith. Bayesian skew selection for multivariate models. *Computational Statistics & Data Analysis*, 54(7):1824–1839, 2010.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

A. Peiró. Skewness in financial returns. *Journal of Banking & Finance*, 23 (6):847–862, 1999.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

S. K. Sahu, D. K. Dey, and M.D. Branco. A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150, 2003.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

G. Torri, R. Giacometti, and S. Paterlini. Sparse precision matrices for minimum variance portfolios. *Computational Management Science*, 16: 375–400, 2019.

D. VanDerwerken and S. C. Schmidlerb. Monitoring joint convergence of mcmc samplers. *Journal of computational and graphical statistics*, 26(3): 558–568, 2017.

H. Wang. Bayesian graphical lasso methods and efficient posterior computation. *Bayesian Analysis*, 7:867–886, 2012.

H. Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

T. Watanabe. On sampling the degree-of-freedom of student's-t disturbances. *Statistics & Probability Letters*, 52(2):177–181, 2001.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 7:733–742, 2003.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.