Contributions to Large Covariance and Inverse Covariance Matrices Estimation

Xiaoning Kang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Xinwei Deng, Committee Chair

Feng Guo

Inyoung Kim

George R. Terrell

July 25, 2016

Blacksburg, Virginia

Keywords: Covariance matrix; modified Cholesky decomposition; sparse estimation.

Contributions to Large Covariance and Inverse Covariance Matrices Estimation

Xiaoning Kang

ABSTRACT

Estimation of covariance matrix and its inverse is of great importance in multivariate statistics with broad applications such as dimension reduction, portfolio optimization, linear discriminant analysis and gene expression analysis. However, accurate estimation of covariance or inverse covariance matrices is challenging due to the positive definiteness constraint and large number of parameters, especially in the high-dimensional cases. In this thesis, I develop several approaches for estimating large covariance and inverse covariance matrices with different applications.

In Chapter 2, I consider an estimation of time-varying covariance matrices in the analysis of multivariate financial data. An order-invariant Cholesky-log-GARCH model is developed for estimating the time-varying covariance matrices based on the modified Cholesky decomposition. This decomposition provides a statistically interpretable parametrization of the covariance matrix. The key idea of the proposed model is to consider an ensemble estimation of covariance matrix based on the multiple permutations of variables.

Chapter 3 investigates the sparse estimation of inverse covariance matrix for the high-dimensional data. This problem has attracted wide attention, since zero entries in the inverse covariance matrix imply the conditional independence among variables. I propose an order-invariant sparse estimator based on the modified Cholesky decomposition. The proposed estimator is obtained by assembling a set of estimates from the multiple permutations of variables. Hard thresholding is imposed on the ensemble Cholesky factor to encourage the sparsity in the estimated inverse covariance matrix. The proposed method is able to catch the correct sparse structure of the inverse covariance matrix.

Chapter 4 focuses on the sparse estimation of large covariance matrix. Traditional estimation approach is known to perform poorly in the high dimensions. I propose a positive-definite estimator for the covariance matrix using the modified Cholesky decomposition. Such a decom-

position provides a flexibility to obtain a set of covariance matrix estimates. The proposed method considers an ensemble estimator as the "center" of these available estimates with respect to Frobenius norm. The proposed estimator is not only guaranteed to be positive definite, but also able to catch the underlying sparse structure of the true matrix.

Key Words: Alternating direction method of multipliers; ensemble estimate; GARCH model; high-dimensional; hard thresholding; inverse covariance matrix; modified Cholesky decomposition; order-invariant; order of variables; sparse estimate.

Contributions to Large Covariance and Inverse Covariance Matrices Estimation

Xiaoning Kang

GENERAL AUDIENCE ABSTRACT

The covariance and inverse covariance matrices estimation is fundamental important in multivariate statistical analysis with broad applications such as imaging recognition, portfolio optimization, weather forecasting, classification and gene expression analysis. However, accurate estimation of covariance or inverse covariance matrices is a challenging problem due to the positive definiteness constraint and large number of parameters, especially in the high-dimensional cases when the number of variables is close to or larger than the sample size. In this thesis, I mainly focus on developing several approaches for estimating large covariance and inverse covariance matrices with different applications.

Among various techniques of estimating covariance and inverse covariance matrices, the modified Cholesky decomposition is a useful tool by converting the estimation of a matrix into estimating a sequence of linear regressions. Such a technique provides a re-parametrization of covariance and inverse covariance matrices from a statistical perspective with meaningful interpretation. However, this approach requires a pre-specified order among the variables, which may not be available in real applications or cannot be pre-determined in the analysis. To overcome this difficulty, I propose order-invariant estimators for the covariance and its inverse matrices under the framework of modified Cholesky decomposition. The proposed methods can be applied into applications such as the analysis of financial data and the classification of gene expression data.

In Chapter 2, I consider an accurate estimation of time-varying covariance matrices in the analysis of multivariate financial data. An order-invariant Cholesky-log-GARCH model is developed for estimating the time-varying covariance matrices. Most of the multivariate GARCH models are known to break down for large dimensions of data. To overcome the curse of dimensionality, orthogonal transformation of the vector of asset returns is a more viable method. The modified Cholesky decomposition is a popular orthogonal transformation

approach which sequentially orthogonalizes the variables (assets), and provides a statistically interpretable parametrization of the covariance matrix. The key idea of the proposed order-invariant Cholesky-log-GARCH model is to consider an ensemble estimation of covariance matrix based on the multiple permutations of variables used in the modified Cholesky decompostion. The proposed methodology not only provides accurate covariance matrix estimation, but also gives accurate prediction of covariance or volatility matrices at future time points.

Chapter 3 investigates the sparse estimation of large inverse covariance matrix for the high-dimensional data. This problem has attracted wide attention, since zero entries in the inverse covariance matrix imply the conditional independence between the corresponding variables. I propose a novel order-invariant sparse estimator based on the modified Cholesky decomposition. The proposed estimator is obtained by efficiently assembling a set of estimates obtained from the multiple permutations of variables. Hard thresholding is imposed on the ensemble Cholesky factor to encourage the sparse structure in the estimated inverse covariance matrix. The proposed method is not only able to obtain the correct sparse structure of the inverse covariance matrix, but also achieves high accuracy without the specification of variable orders used in the modified Cholesky decomposition. I applied the proposed method to the linear discriminant analysis for analyzing three classification examples.

Chapter 4 focuses on the sparse estimation of large covariance matrix. Traditional estimation approach using the sample covariance matrix is known to perform poorly in the high dimensions. I propose a positive-definite estimator for the covariance matrix based on the modified Cholesky decomposition. Such a decomposition relies on the order of variables, which provides a flexibility to obtain a set of covariance matrix estimates under multiple orders of variables. The proposed method considers an ensemble estimator as the "center" of these available estimates with respect to Frobenius norm. The proposed estimator is not only guaranteed to be positive definite, but also able to catch the underlying sparse structure of the true matrix. The merits of the proposed method are illustrated through simulation studies and one real data example.

Key Words: Alternating direction method of multipliers; ensemble estimate; GARCH model; high-dimensional; hard thresholding; inverse covariance matrix; modified Cholesky decomposition; order-invariant; order of variables; sparse estimate.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

As data-collection technology advances, there is an urgent need of developing analysis tools for high-dimensional data. The estimation of large covariance and inverse covariance matrices is of fundamental importance in multivariate analysis with various statistical applications. For example, the covariance matrix is commonly used in dimension reduction (Jolliffe, 2002), financial mathematics (Ledoit and Wolf, 2003), signal processing (Krim and Viberg, 1996), genomics (Schäfer and Strimmer, 2005), geometric functional analysis (Rudelson, 1999), multivariate volatility (Bollerslev, 1990; Engle, 2002), portfolio optimization (Deng and Tsui, 2013) and weather forecasting (Berre, 2000). The inverse covariance matrix arises in diverse applications such as graphical models (Yuan and Lin, 2007), linear discriminant analysis (McLachlan, 2004), classification of gene expression data (Dudoit, Fridlyand and Speed, 2002), fMRI analysis (Varoquaux et al., 2010), and spectroscopic imaging (Trbovic et al., 2004). In these applications, data with the number of variables close to or exceeding the sample size are commonly observed.

However, accurate estimation of covariance and inverse covariance matrices encounters great challenges due to the positive definiteness constraint and a large number of parameters in the high-dimensional cases. The number of parameters in a $p \times p$ covariance matrix increases in a quadratic order with the number of random variables p. For example of image analysis, an image of 128×128 pixels contains more than 10,000 variables in terms of intensity of pixels. The traditional methods of estimating covariance and inverse covariance matrices, such as sample covariance matrix estimation, may not work well in this situation. It is well known that the sample covariance matrix $\boldsymbol{S}$ is the maximum likelihood estimator of covariance matrix. However, such an estimate is highly unstable when the number of variables p is close to the sample size n. Even worse, the matrix $\boldsymbol{S}$ is not positive definite when $p > n$ and hence it is not appropriate to use $\boldsymbol{S}$ and $\boldsymbol{S}^{-1}$ as estimates of covariance and inverse covariance

matrices. Accordingly, alternative ways of accurate estimation are needed. In this dissertation, I focus on developing several approaches for estimating large covariance and inverse covariance matrices in the high-dimensional settings.

1.1 Covariance Matrix Estimation

The estimation of covariance matrix has been a fundamental problem in statistical inference, since the covariance matrix plays an important role in many data analysis techniques. A most natural estimate of the covariance matrix is the empirical sample covariance matrix. But it is known to behave poorly in the high-dimensional settings, where the sample eigenvalues are over-dispersed and the eigenvectors are not consistent (Johnstone, 2001). Hence, a variety of regularized approaches has been proposed in the literature to obtain accurate estimation of the covariance matrix.

Early work was developed to shrink the eigenvalues of the sample covariance matrix, including Dey and Srinivasan (1985) and Haff (1991), among many others. Ledoit and Wolf (2004) introduced a linear combination of the sample covariance with the identity matrix. Bickel and Levina (2008) considered the regularization method by thresholding the entries of the sample covariance matrix and studied its theoretical behavior when the number of variables is large. Deng and Tsui (2013) proposed a Log-ME estimation method, which estimates the covariance matrix through matrix logarithm. The matrix log-transformation provides the ability to impose a convex penalty on the transformed likelihood such that the largest and smallest eigenvalues of the covariance matrix estimate can be regularized simultaneously. Additionally, banded covariance matrix estimation (Bickel and Levina, 2008) is also developed as a special case, especially used in the autoregressive model for time series data. Furrer and Bengtsson (2007) proposed to shrink the covariance entries based on their distance from the diagonal. In most of these works, convergence rates of the estimators were well established. Some other literature of covariance matrix estimation can also be found from Perron (1992), Smith and Kohn (2002), Won et al. (2013), Cai and Yuan (2012), Fan, Liao and Mincheva

(2013), Wang and Pillai (2013), Lounici (2014) and references therein.

In addition to the aforementioned methods, the modified Cholesky decomposition (MCD) approach is developed for covariance matrix estimation by several researchers (Pourahmadi, 1999; Pourahmadi, 2000; Wu and Pourahmadi, 2003; Huang et al., 2006; Pourahmadi, Daniels and Park, 2007; Rothman et al., 2008; Chang and Tsay, 2010; Zhang and Leng, 2012). Such a technique is to reduce the challenge of estimating a covariance matrix into estimating a set of regressions. The key idea is that the covariance matrix $\boldsymbol{\Sigma}$ of a zero-mean random vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ can be diagonalised by a lower triangular matrix constructed from the regression coefficients when X_j is regressed on its predecessors $X_1, \ldots, X_{j-1}$. Precisely, for $j = 2, \ldots, p$, define

$$\begin{aligned} X_j &= \sum_{t=1}^{j-1} a_{jt} X_t + \epsilon_j \\ &= \boldsymbol{Z}_j^T \boldsymbol{a}_j + \epsilon_j, \end{aligned} \tag{1.1}$$

where $\boldsymbol{Z}_j = (X_1, \ldots, X_{j-1})^{'}$, and $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{j,j-1})^{'}$ is the corresponding vector of regression coefficients. The error term ϵ_j has population expectation $E(\epsilon_j) = 0$ and population variance $Var(\epsilon_j) = d_j^2$. Also define $d_1^2 = Var(X_1)$. Consequently, the MCD approach is written as:

$$\boldsymbol{\Sigma} = \boldsymbol{T}^{-1} \boldsymbol{D} \boldsymbol{T}'^{-1}, \tag{1.2}$$

where $\boldsymbol{T}$ is a unit lower triangular matrix with ones on its diagonal and $\boldsymbol{a}_j^{'}$ as its jth row. $\boldsymbol{D} = diag(d_1^2, \ldots, d_p^2)$ is a diagonal matrix with residual variances of these regressions on the diagonal. As a result, this approach converts the constraint entries of covariance matrix into two groups of unconstrained "regression" and "variance" parameters. It is statistically meaningful and provides a positive definite covariance matrix estimate.

However, in terms of sparse estimation of covariance matrix in high dimensions, this approach (1.2) requires an inverse of Cholesky factor matrix $\boldsymbol{T}$, and hence, it is not easy to impose

a sparse structure on the estimate of $\boldsymbol{\Sigma}$. Alternatively, Pourahmadi (1999) proposed another form of the modified Cholesky decomposition by a latent variable regression model, which enables regularization, as follows. Denote the vector of errors in (1.1) by $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$, where $\epsilon_1 = X_1$. Then one can write $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{\epsilon}$ when X_j is regressed on its previous latent variables $\epsilon_1, \ldots, \epsilon_{j-1}$, and hence $\boldsymbol{L} = (l_{jk})_{p\times p}$ is a unit lower triangular matrix constructed from the regression coefficients in the following sequential regressions:

$$X_j = \boldsymbol{l}_j^T \boldsymbol{\epsilon} = \sum_{k<j} l_{jk}\epsilon_k + \epsilon_j, \quad j = 2, \ldots, p, \tag{1.3}$$

where $\boldsymbol{l}_j = (l_{jk})$ is the jth row of $\boldsymbol{L}$. Here $l_{jj} = 1$ and $l_{jk} = 0$ for $k > j$. Consequently, the modified Cholesky decomposition provides a re-parameterization of the covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T.$$

Because this decomposition associates the covariance matrix $\boldsymbol{\Sigma}$ with a sequence of linear regressions in (1.3), regularization in sequential linear regressions shapes the covariance matrix estimate. For the sparse estimation of the covariance matrix with a banded structure, Rothman et al. (2010) proposed to band its Cholesky factor matrix and adopted a procedure similar to Gram–Schmidt process (Sajid, Ahmed and Taj, 2009) to sequentially obtain realized residuals.

1.2 Inverse Covariance Matrix Estimation

The sparse estimation of inverse covariance matrix in the high-dimensional cases is widely studied in the literature. The zero entry in the inverse covariance matrix indicates the conditional independence between the corresponding variables, given all the others. One straightforward inverse covariance estimation is to obtain an estimate of covariance matrix first, using various regularized approaches mentioned in Section 1.1, and then take its inverse as an estimate of inverse covariance matrix. However, this would definitely destroy the sparse

structure of the estimated inverse covariance matrix. Accordingly, it is desirable to obtain an inverse covariance matrix estimate directly. Suppose there are n multivariate normal observations of dimension p, with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and $\boldsymbol{S}$ be the empirical sample covariance matrix, the objective is to minimize the negative log-likelihood

$$L_n(\mathbf{\Omega}) = -\log|\mathbf{\Omega}| + \mathrm{tr}[\mathbf{\Omega}\boldsymbol{S}] \tag{1.4}$$

over definite matrix $\mathbf{\Omega}$. One commonly used technique to solve (1.4) with sparsity is Graphical Lasso (Glasso), introduced in Yuan and Lin (2007). They proposed the graphical model to simultaneously conduct inverse covariance matrix estimation and model selection by adding penalty term onto the negative log-likelihood (1.4), giving their sparse estimate as

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}} \quad -\log|\mathbf{\Omega}| + \mathrm{tr}[\mathbf{\Omega}\boldsymbol{S}] + \rho||\mathbf{\Omega}||_1,$$

where ρ is a positive tuning parameter, and $||\mathbf{\Omega}||_1$ denotes L_1 norm, the sum of the absolute values of the entries in $\mathbf{\Omega}$. Lasso type penalty is imposed on the entries of the inverse covariance matrix when minimizing the negative log-likelihood, thus encouraging some of the off-diagonal entries of the estimated inverse covariance matrix to be exact zeroes. One of the main challenges of this approach is the computational cost since estimating a positive definite matrix requires semi-definite programming, which is not very feasible in the high-dimensional cases. Glasso with fast computation has been developed by Friedman, Hastie, and Tibshirani (2008), Rocha, Zhao and Yu (2008), Rothman et al. (2008) and Yuan (2008). Yuan and Lin (2007) and Lam and Fan (2009) also developed some theoretical for the Glasso method. Glasso estimator has also been studied by Friedman, Hastie, and Tibshirani (2008), Deng and Yuan (2009), Yuan (2010), Yin and Li (2011), Liu (2013) and Danaher, Wang and Witten (2014) among others. However, Raskutti (2008) pointed out that when the number of variables is much larger than the sample size, Glasso method may not perform well.

Several other methods of inverse covariance matrix estimation have also been proposed in

the literature. Meinshausen and Bühlmann (2006) introduced a neighborhood-based approach. It estimates the inverse covariance matrix column by column using the Lasso or Dantzig selector for each variable against all of the rest variables; see Cai, Liu and Luo (2011), Cai, Liu and Zhou (2016) and Sun and Zhang (2013). In addition, Fan, Fan and Lv (2008) developed a factor model to estimate both covariance matrix and its inverse. They studied the estimation in the asymptotic framework that the dimensionality tends to infinity as the sample size increases. A multi-factor model was employed to reduce the dimensionality and to estimate the inverse covariance matrix. Drton and Perlman (2008) proposed a multiple testing procedure for simultaneously conducting hypotheses of zeroes in the inverse covariance matrix. Xue and Zou (2012) introduced a regularized rank-based estimation idea for estimating nonparametric graphical models. Sun and Zhang (2012) derived the confidence intervals for entries of the estimated inverse covariance matrix. Some recent Bayesian literature can also be found in the work of Cheng and Lenkoski (2012), Wang (2012), Bhadra and Mallick (2013), Scutari (2013) and Mohammadi and Wit (2015), among many others.

Besides the methods mentioned above, the modified Cholesky decomposition (MCD) approach provides a statistically interpretable re-parametrization of the inverse covariance matrix and is able to give a positive definite estimation. It is also applicable in the high-dimensional cases and enjoys fast computation. From (1.2), the MCD for the inverse covariance matrix $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}.$$

In terms of sparse models estimation, several methods have been proposed to regularize the Cholesky factor $\boldsymbol{T}$. A k-banded estimator of $\boldsymbol{T}$ is often obtained by regressing each variable only on its closest k predecessors; Wu and Pourahmadi (2003) proposed this estimator and suggested to select the tuning parameter k based on Akaike information criterion (AIC). Bickel and Levina (2008) demonstrated that banding the Cholesky factor $\boldsymbol{T}$ is able to produce a consistent estimator in the operator norm under weak conditions as long as $(\log\, p)/n \to 0$, and they suggested to use cross-validation for choosing k. Huang et al. (2006) proposed

to impose either an L_1 (Lasso) or an L_2 (Ridge) penalty on the entries of $\boldsymbol{T}$. The Lasso penalty creates zeroes in the Cholesky factor $\boldsymbol{T}$ in arbitrary locations, which is more flexible than banding, but the resulting estimate of the inverse covariance matrix may not have any zeroes at all. Levina, Rothman and Zhu (2008) proposed adaptive banding, using a nested lasso penalty, which allows a different k for each regression, and hence is more flexible than banding while also retaining some sparsity in the estimate of inverse covariance matrix.

Although there are various methods proposed to investigate the sparse estimation of covariance and its inverse matrices in the framework of the MCD approach, few work has contributed to solve a potential problem, the order issue. Clearly, different orders of the random variable $X_1, \ldots, X_p$ used in the MCD will lead to different estimates of regression coefficients in the model (1.1), thus resulting in different estimates of covariance and inverse covariance matrices. However, this order often cannot be determined before the analysis. Dellaportas and Pourahmadi (2012) suggested a search algorithm to choose the order according to Akaike information criterion (AIC) or Bayesian information criterion (BIC). But according to our studies, the order selected from these criteria does not necessarily lead to a better estimate of covariance matrix or its inverse. Rajaratnam and Salzman (2013) proposed the best permutation algorithm (BPA) for recovering the order of variables. However, the BPA approach only focuses on the banded and autoregressive models, thus is not suitable for the general situations. Therefore, in this dissertation, I consider the MCD approach as a launching point and develop order-invariant sparse estimators for the covariance and inverse covariance matrices without the prior knowledge of the order. The proposed method takes advantages of multiple orders over one single order of the random variables, and hard thresholding is imposed on the Cholesky factor $\boldsymbol{T}$ to encourage the sparse structures.

1.3 Multivariate GARCH Model

Time-varying covariance between the asset returns is often critical inputs for a broad spectrum of the common tasks of financial management (Engle, 2002). For example, in portfolio invest-

ment, the covariance matrix of asset returns is used to minimize the portfolio risk. Hence, many financial applications call for the estimation of large time-varying covariance matrices. One of the most popular financial tools to estimate covariance is the generalized autoregressive conditional heteroscedastic (GARCH) models. The univariate GARCH models are proposed to fit time-ordered data by Engle (1982) and Bollerslev (1986). They have been very successful for short- and middle-term volatility forecasting in financial markets. Multivariate GARCH models, originally introduced by Bollerslev, Engle and Wooldridge (1988), generalize univariate GARCH models to multivariate parameterizations; see, for example, Engle and Kroner (1995) and Engle and Mezrich (1996). However, the estimation of these models can be quite challenging, since the models have a large amount of parameters, and consequently the optimization of the likelihood becomes practicably infeasible.

Later, a variety of extensions of the GARCH models has been introduced in the finance literature. For instance, Bollerslev (1990) proposed the constant conditional correlation GARCH (CCC-GARCH) model, assuming that the time-varying covariance matrix of variables is a product of dynamic volatilities and a constant correlation matrix. The quasi maximum likelihood technique is employed to estimate the model. However, this assumption is often difficult to be satisfied in real data. Tsui and Yu (1999) found that the constant correlation can be rejected for ceratin assets. Both Bera and Kim (2002) and Tse (2000) have developed the hypothesis testings for constant correlation, the former being a bivariate test while the latter is a more general multivariate Lagrange multiplier test. Although the constant conditional correlation assumption is strong, it has been relaxed by Engle (2002) to the case where the correlation matrix is time-varying, known as dynamic conditional correlation GARCH (DCC-GARCH) model. It allows for non-constant correlations, but it is computationally expensive and has convergence difficulties in the high-dimensional cases.

Another direction of the GARCH model extensions is the principal components GARCH or orthogonal GARCH (O-GARCH) method (Alexander, 2001). It assumes that the observed data can be linearly transformed into a set of independent latent variables through an orthogonal link matrix, such that univariate GARCH models are used to model them one at a

time. Hence, the O-GARCH method provides a simple way of estimating multivariate volatilities that may be difficult to be obtained by direct GARCH estimation. However, one strong restriction imposed on this method is that it requires the matrix, which links the observed data and the independent variables, to be orthogonal. A natural generalization of O-GARCH model is to replace the orthogonal link matrix with an arbitrary invertible link matrix, giving rise to the generalized orthogonal GARCH (GO-GARCH) model (Van der Weide, 2002). The estimation of parameters in the GO-GARCH model is essentially consisted of two steps. In the first step, only part of the link matrix is identified. The rest parameters of this link matrix, together with the GARCH parameters, are then estimated in the second step by maximizing a multivariate likelihood function. However, the latter can be troublesome in the high-dimensional cases.

Recently, Pedeli, Fokianos and Pourahmadi (2015) adopted the modified Cholesky decomposition (MCD) approach (Pourahmadi, 1999, 2001) to estimate the covariance matrix in the log-GARCH model (Geweke, 1986) in the analysis of multivariate time series data. The MCD approach provides an unconstrained and statistically interpretable parametrization of a covariance matrix. However, such estimation based on the MCD approach depends on the order of the random variables, which often cannot be pre-determined. Therefore, in Chapter 2, I propose an order-invariant Cholesky-log-GARCH model for estimating the covariance matrix of multivariate time series. The proposed method provides an accurate estimate of covariance matrix, and works even better in the high-dimensional settings. Moreover, the prediction of the covariance matrix at a future time point is also developed, and the predicted covariance matrix can be used as a measure of the volatility of a portfolio.

1.4 Linear Discriminant Analysis

One popular application of using inverse covariance matrix is the discriminant analysis, which requires a reliable estimate of inverse covariance matrix in the classification rule, for example the linear discriminant analysis (LDA). Discriminant analysis required between two or more

populations commonly occurs in the daily life. In medicine, diagnosis is based on the symptoms and perhaps some clinical tests whose values vary on a continuous scale. In financial companies, a decision has to make whether or not to issue a credit card to a customer. In biomedical area, statistics methods are used for finding the relative genes that controls certain characteristic. Other applications of discrimination include social connection, public health, engineering quality control and image recognition.

One of the widely used linear discrimination is Fisher discriminant analysis (FDA), introduced by Fisher (1936). The idea is to project data from multiple dimensions onto one dimension such that the samples from different classes are separated as much as possible. Specifically, it is to search for a linear separating hyperplane $\boldsymbol{\alpha}$ that maximizes the ratio of between-class variance to the within-class variance $\frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_b\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_w\boldsymbol{\alpha}}$, thus achieving maximum discrimination, where $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_w$ are pooled between and within-class variances, respectively. Clearly, a reliable estimation of the covariance matrix plays an important role in the discriminant analysis.

Another popular classification approach is the linear discriminant analysis (LDA). Consider a classification problem with K classes. Each observation belongs to some class $k \in 1, 2, \ldots, K$. LDA assumes the data are from multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with different class means $\boldsymbol{\mu}_k$, and the same covariance matrix $\boldsymbol{\Sigma}$. Then LDA classification rule is: classify a new observation $\boldsymbol{x}$ to class k^* if $k^* = \arg\max_k \eta_k(\boldsymbol{x})$, where

$$\eta_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log\pi_k, \tag{1.5}$$

where π_k is the frequency of class k in the data set. This approach works well in the situation where the number of variables p is much smaller than the sample size n. Additionally, it is asymptotically optimal in the sense that, when p diverges to infinity at a rate slower than $\sqrt{n}$, its misclassification rate over that of the optimal rule converges to one (Shao et al. 2011). However, the classical LDA requires that the covariance matrix $\boldsymbol{\Sigma}$ in (1.5) be nonsingular. In many applications involving high-dimensional data, such as cancer classification with gene

expression, this covariance matrix can be singular. Different methods have been proposed to deal with the singularity problem, including PCA + LDA (Belhumeur, Hespanha and Kriegman 1997), regularized LDA (Guo, Hastie and Tibshirani, 2007), uncorrelated LDA (Ye, 2005) and orthogonal LDA (Ye, 2005). Howland and Park (2004) proposed generalized LDA (GLDA), which does not assume the non-singularity of $\mathbf{\Sigma}$, and thus solved the small sample size problem. It is mathematically well-founded and coincides with the classical LDA when $\mathbf{\Sigma}$ is nonsingular. Other approaches that impose regularization on the covariance structure include Guo, Hastie and Tibshirani (2007), Rothman et al. (2008), Witten and Tibshirani (2009) and Shao et al. (2011).

1.5 Outline of the Dissertation

The rest of this dissertation is organized as follows. In Chapter 2, an order-invariant Cholesky-log-GARCH model is developed to analyze the multivariate financial time series data. The proposed approach combines the log-GARCH model and the modified Cholesky decomposition method together, re-parameterizing the covariance matrix via the modified Cholesky decomposition and modeling the time-varying volatilities using log-GARCH model. It enjoys the advantage of multiple orders of random variables used in the modified Cholesky decomposition, and the resulting estimator is accurate in both estimation and prediction of the covariance matrix. In Chapter 3, I investigate an order-invariant Cholesky decomposition method for sparse estimation of inverse covariance matrix in the high-dimensional cases. This model considers ensemble estimates under multiple orders of random variables and encourages the sparsity using hard thresholding technique. The framework is also extended to the classification settings. In Chapter 4, a positive definite estimator for high-dimensional covariance matrix is developed based on the modified Cholesky decomposition. The proposed method takes advantage of multiple covariance matrix estimates obtained from different orders of variables used in the modified Cholesky decomposition, and seeks for the "center" of these estimates under Frobenius norm. The L_1 constraint is imposed on the objective func-

tion to achieve the sparsity in the estimate. An efficient algorithm is developed to solve the challenging optimization problem. Some discussions are included in Chapter 5.

Chapter 2 An Order-Invariant Cholesky-Log-GARCH Model for Multivariate Financial Time Series

2.1 Introduction

Many tasks of financial management, such as portfolio selection, option pricing and risk assessment, require modeling and prediction of a sequence of large $p \times p$ covariance (volatility) matrices $\{\mathbf{\Sigma}_t\}$ based on the (conditionally) independently $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$ – distributed data $\boldsymbol{y}_t$, $t = 1, 2, \ldots, n$. Here $\boldsymbol{y}_t$ can be viewed as the shock (innovation) at time t of the returns of p assets in a portfolio. Clearly, this is a challenging statistical problem since the number of parameters for n covariance matrices of order p is linear in n and quadratic in p. Moreover, the estimation of a single covariance matrix involves a large number of parameters and requires the positive definiteness constraint, especially in the high-dimensional cases (Deng and Tsui, 2013).

The modeling of high-dimensional covariance matrices in financial area often is to characterize the instantaneous dependence among several asset returns. A variety of extensions of the univariate generalized autoregressive conditional heteroscedastic (GARCH) models (Bollerslev, 1986) has been proposed in the finance literature, see, for example, Engle and Kroner (1995), Tse and Tsui (2002) and Ledoit, Santa-Clara and Wolf (2004). However, most of the existing methods have strong assumptions on the dynamics of the conditional correlation matrices and are computationally expensive for dimension bigger than ten or so. For instance, Bollerslev (1990) assumes constant conditional correlation for the GARCH model (CCC-GARCH), which is often not satisfied in real data. Engle (2002) assumes the dynamic conditional correlation for the GARCH model (DCC-GARCH), which is computationally expensive in high-dimensional cases, see also Tse and Tsui (2002). There are also several Bayesian approaches for GARCH models in both inference and prediction (Ardia and

Hoogerheide, 2010). For example, Jensen and Maheu (2013) proposed a dynamic component model based on time-varying Wishart distribution in multivariate GARCH models. Virbickaitė, Ausín and Galeano (2014) considered an infinite mixture of Gaussian distributions with a Dirichlet process prior to model the multivariate time series. Some other Bayesian methods can be found in Galeano and Ausín (2010), Arakelian and Dellaportas (2012), Jacquier and Polson (2012), Ausín, Galeano and Ghosh (2014), Burda (2015) and references therein. The computation of Bayesian methods for multivariate time series is also intensive in the high-dimensional cases.

Orthogonal transformation of the data vector is a viable method for overcoming the curse of dimensionality in the finance literature. An orthogonal transformation model basically assumes that a p-dimensional data vector is driven by p orthogonal latent components or variables, so that univariate GARCH models can be used to model these orthogonal components one-at-a-time or equation-by-equation. For Gaussian data, the idea of principal component analysis (PCA) of the sample covariance matrix was used by Alexander (2001) to orthogonalize the components of the vector of returns, and then univariate GARCH models were used for each principal component, giving rise to the class of so-called O-GARCH models. For non-Gaussian data, the technique of independent component analysis (ICA) was used by Van der Weide (2002) to introduce the class of generalized orthogonal GARCH (GO-GARCH) models. Combining the previous two ideas, Matteson and Tsay (2011) have introduced the dynamic orthogonal component (DOC) models, see Tsay (2014, Section 7.8) for a recent review of orthogonal transformation methods.

The modified Cholesky decomposition (MCD) approach (Pourahmadi, 1999) is another example of the orthogonal transformation technique which sequentially orthogonalizes the variables in the data vector. It also provides an unconstrained and statistically interpretable parametrization of a covariance matrix. Recently, Pedeli, Fokianos and Pourahmadi (2015) adopted the MCD idea to estimate the covariance matrix using the log-GARCH model (Geweke, 1986) for each orthogonal component. However, due to the sequential nature of the MCD, the estimation results from such a MCD-based log-GARCH model depend on the order

in the data vector, which often cannot be pre-determined before the analysis. Some early and implicit use of Cholesky-type decompositions of the covariance matrix can be found in the applications of factor models to finance, more specific examples are Vrontos, Dellaportas and Politis (2003) and Palandri (2009).

In this work, we propose an order-invariant Cholesky-log-GARCH model for estimating the covariance matrix of multivariate time series. The proposed method can provide an accurate estimate of $\mathbf{\Sigma}_t$ in high-dimensional settings. It is computationally attractive by employing penalized regression to shrink the off-diagonal elements of the Cholesky factor. Using the order-invariant MCD-based estimate of $\mathbf{\Sigma}_t$ in the log-GARCH model, the proposed method performs much better than other conventional approaches for analyzing multivariate time series data. Moreover, the prediction for the covariance matrix of the volatility at a future time point is also developed.

The rest of the chapter is organized as follows. In Section 2.2, we briefly review the MCD approach to estimate a covariance matrix. In Sections 2.3, we address the order issue of the MCD approach by considering an ensemble estimate of the covariance matrix. Combining the log-GARCH model and the ensemble estimate, we detail in Section 2.4 an order-invariant Cholesky-log-GARCH model for analyzing multivariate financial time series data. Three real examples of financial data sets are presented in Section 2.5. We conclude our work with some discussion in Section 2.6.

2.2 The Modified Cholesky Decomposition

Suppose $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is a p-dimensional vector of random variables with covariance matrix $\mathbf{\Sigma}$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent and identically distributed observations following a multivariate normal distribution with mean equal to the zero vector and covariance matrix $\mathbf{\Sigma}$, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Pourahmadi (1999) proposed a modified Cholesky decomposition technique to estimate $\mathbf{\Sigma}$, which is statistically meaningful and always provides a positive definite covariance matrix estimate. The key idea is that the covariance matrix of a zero-mean random vector

$\boldsymbol{X} = (X_1, \ldots, X_p)'$ can be diagonalised by a lower triangular matrix constructed from the regression coefficients when X_j is regressed on its predecessors $X_1, \ldots, X_{j-1}$. Specifically, for $j = 2, \ldots, p$, define

$$\begin{aligned} X_j &= \sum_{t=1}^{j-1} a_{jt} X_t + \epsilon_j \\ &= \boldsymbol{Z}_j^T \boldsymbol{a}_j + \epsilon_j, \end{aligned} \tag{2.1}$$

where $\boldsymbol{Z}_j = (X_1, \ldots, X_{j-1})'$, and $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{j,j-1})'$ is the corresponding vector of regression coefficients. The error ϵ_j has population expectation $E(\epsilon_j) = 0$ and population variance $Var(\epsilon_j) = d_j^2$. Hence, we can form a lower triangular matrix $\boldsymbol{A}$ such that

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ a_{21} & 0 & 0 & \ldots & 0 \\ a_{31} & a_{32} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{p1} & a_{p2} & \ldots & a_{p,p-1} & 0 \end{pmatrix},$$

which contains all the regression coefficients in (2.1). Also define:

$$d_j^2 = Var(\epsilon_j) = \begin{cases} Var(X_1), & j = 1, \\ Var(X_j - \boldsymbol{Z}_j^T \boldsymbol{a}_j), & j = 2, \ldots, p. \end{cases} \tag{2.2}$$

Let $\boldsymbol{I}$ be the $p \times p$ identity matrix. Denote $\boldsymbol{D} = diag(d_1^2, \ldots, d_p^2)$ a diagonal matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$. Then the MCD approach gives:

$$Var(\boldsymbol{\epsilon}) = Var(\boldsymbol{X} - \boldsymbol{A}\boldsymbol{X}) = Var[(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X}]$$
$$\boldsymbol{D} = \boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}',$$

where $\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{A}$ is a unit lower triangular matrix having ones on its diagonal. As a result, the decomposition (2.1) converts the constraint entries of $\boldsymbol{\Sigma}$ into two groups of unconstrained "regression" and "variance" parameters. Conceptually, this approach reduces the challenge of modeling a covariance matrix into modeling $(p-1)$ regression problems. A straightforward estimate $\hat{\boldsymbol{T}}$ of $\boldsymbol{T}$ can be obtained from the least squares estimates of the regression coefficients

$$\hat{\boldsymbol{a}}_j = \arg\min_{\boldsymbol{a}_j} \|\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{a}_j\|_2^2, \ j = 2, \ldots, p,$$

where $\boldsymbol{x}^{(j)}$ is the jth column of the data matrix $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, and $\mathbb{Z}^{(j)} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(j-1)})$ represents the first (j-1) columns of $\mathbb{X}$. The estimate $\hat{\boldsymbol{D}}$ of $\boldsymbol{D}$ is constructed from the corresponding residual variances according to (2.2).

However, the least squares estimate $\hat{\boldsymbol{a}}_j$ of $\boldsymbol{a}_j$ is not available when the dimension p is larger than the sample size n in the high-dimensional settings. Furthermore, the MCD approach needs a pre-specified order of $X_1, \ldots, X_p$ when constructing the matrices $\boldsymbol{T}$ and $\boldsymbol{D}$. A discussion of the order issue for the MCD-based approach is thus important. Wagaman and Levina (2009) proposed an Isomap method to discover the order of variables for the estimation of banded covariance matrix. Rajaratnam and Salzman (2013) introduced a so-called best permutation algorithm to recover the natural order of variables in autoregressive models for banded covariance matrix estimation, by minimizing the sum of the diagonals of $\boldsymbol{D}$ in the MCD approach. Dellaportas and Pourahmadi (2012) suggested a search algorithm to choose the order of variables for MCD based on Akaike information criterion (AIC) or Bayesian information criterion (BIC). However, the order chosen by such a search algorithm may not give a

better estimate of the covariance matrix. In the next section, we propose an ensemble estimate of the covariance matrix based on MCD, which can lead to an accurate and order-invariant estimate of the covariance matrix.

2.3 Estimation of a Single Covariance Matrix $\boldsymbol{\Sigma}$

2.3.1 Ensemble Estimate of $\boldsymbol{\Sigma}$

Note that the MCD-based covariance matrix estimation for $\boldsymbol{\Sigma}$ depends on the order of $X_1, \ldots, X_p$. Chang and Tsay (2010) pointing out that the MCD approach is not order invariant, investigated the sensitivity of MCD to order by randomly permuting the variables before estimation. To address this order issue, we take advantage of permutation to gain the flexibility such that we can ensemble the multiple estimates under different permutations of order.

Define a permutation mapping $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$, which gives

$$(\pi(1), \pi(2), \ldots, \pi(p)) \,. \tag{2.3}$$

Define the corresponding permutation matrix $\boldsymbol{P}_\pi$ of which the entries in the jth column are all 0 except taking 1 at position $\pi(j)$. Therefore, the transformed data matrix is

$$\mathbb{X}_\pi = \mathbb{X}\boldsymbol{P}_\pi = (\boldsymbol{x}_\pi^{(1)}, \ldots, \boldsymbol{x}_\pi^{(p)}),$$

where $\boldsymbol{x}_\pi^{(j)}$ is the jth column of $\mathbb{X}_\pi$, $j = 1, 2, \ldots, p$. Now we apply the Lasso technique (Tibshirani, 1996) for the situation where p is close to n or even larger than n. The idea of Lasso-type estimator for the Cholesky factor has been used by Huang et al. (2006), Rothaman et al. (2010) and Chang and Tsay (2010). Under a given permutation π, let

$$\hat{\boldsymbol{a}}_{\pi(j)} = \arg\min_{\boldsymbol{a}_{\pi(j)}} \|\boldsymbol{x}_\pi^{(\pi(j))} - \mathbb{Z}_\pi^{(\pi(j))}\boldsymbol{a}_{\pi(j)}\|_2^2 + \lambda_{\pi(j)}\|\boldsymbol{a}_{\pi(j)}\|_1, \text{ for } \pi(j) \neq 1, \tag{2.4}$$

and

$$\hat{d}^2_{\pi(j)} = \begin{cases} \widehat{Var}(\boldsymbol{x}^{(1)}_{\pi}), & \pi(j) = 1, \\ \widehat{Var}(\boldsymbol{x}^{(\pi(j))}_{\pi} - \mathbb{Z}^{(\pi(j))}_{\pi}\hat{\boldsymbol{a}}_{\pi(j)}), & \text{otherwise}, \end{cases}$$

where $\mathbb{Z}^{(j)}_{\pi}$ represents the first (j-1) columns of $\mathbb{X}_{\pi}$, $\lambda \geq 0$ is a tuning parameter, and $\|\cdot\|_1$ stands for the vector L_1 norm. Here $\widehat{Var}(\cdot)$ denotes the sample variance. Then we can obtain the lower triangular matrix $\hat{\boldsymbol{T}}_{\pi}$ with ones on its diagonal and $\hat{\boldsymbol{a}}'_{\pi(j)}$ as its $\pi(j)$th row. Meanwhile, the diagonal matrix $\hat{\boldsymbol{D}}_{\pi}$ has its $\pi(j)$th diagonal element equal to $\hat{d}^2_{\pi(j)}$. Correspondingly, $\hat{\boldsymbol{\Sigma}}_{\pi} = \hat{\boldsymbol{T}}_{\pi}^{-1}\hat{\boldsymbol{D}}_{\pi}\hat{\boldsymbol{T}}'^{-1}_{\pi}$ will be a covariance matrix estimate under π. Transforming back to the original order, we can estimate $\boldsymbol{\Sigma}$ as

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= \boldsymbol{P}_{\pi}\hat{\boldsymbol{\Sigma}}_{\pi}\boldsymbol{P}'_{\pi} \\ &= \boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}_{\pi}^{-1}\hat{\boldsymbol{D}}_{\pi}\hat{\boldsymbol{T}}'^{-1}_{\pi}\boldsymbol{P}'_{\pi} \\ &= (\boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}_{\pi}^{-1}\boldsymbol{P}'_{\pi})(\boldsymbol{P}_{\pi}\hat{\boldsymbol{D}}_{\pi}\boldsymbol{P}'_{\pi})(\boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}'^{-1}_{\pi}\boldsymbol{P}'_{\pi}) \\ &\triangleq \hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{D}}\hat{\boldsymbol{T}}'^{-1}. \end{aligned} \tag{2.5}$$

Note that $\hat{\boldsymbol{T}} = \boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}_{\pi}\boldsymbol{P}'_{\pi}$ may no longer be a lower triangular matrix. Suppose we generate M permutation mappings π_k, $k = 1, \ldots, M$. Accordingly, we obtain the corresponding estimates $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{T}}$, and $\hat{\boldsymbol{D}}$ in (2.5), denoted as $\hat{\boldsymbol{\Sigma}}_k$, $\hat{\boldsymbol{T}}_k$, and $\hat{\boldsymbol{D}}_k$ for the permutation π_k.

If using multiple covariance matrix estimates $\hat{\boldsymbol{\Sigma}}_k$, $k = 1, \ldots, M$, a naive ensemble estimation of $\boldsymbol{\Sigma}$ can be $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k$. However, such an estimate may not be very accurate since the estimation error of $\hat{\boldsymbol{\Sigma}}_k$ is aggregated by the estimation error of $\hat{\boldsymbol{T}}_k$ and $\hat{\boldsymbol{D}}_k$. Alternatively, we propose the ensemble estimate as

$$\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{T}}^{-1}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{T}}'^{-1} \text{ with } \tilde{\boldsymbol{T}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{T}}_k, \ \tilde{\boldsymbol{D}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{D}}_k. \tag{2.6}$$

Compared to the naive ensemble method, the estimate in (2.6) can achieve better accuracy.

It can reduce the variability in the estimates $\tilde{\boldsymbol{T}}$ and $\tilde{\boldsymbol{D}}$ more directly, hence leading to a small variability in $\tilde{\boldsymbol{\Sigma}}$. Note that to implement this method, one needs to choose the number of permutations M. It is known that the number of all possible permutations increases rapidly as the number of variables p increases. To choose an appropriate number of permutations M for efficient computation, we have tried $M = 30, 50, 80, 100$ and 150 as the number of randomly selected permutations from all the possible permutations. The performance results are quite comparable when M is larger than 50. Hence, in this chapter we choose a sample of $M = 100$ permutations for the proposed order-invariant Cholesky-log-GARCH model.

2.3.2 Statistical Interpretation of Cholesky Factor

In this section, we investigate the statistical meanings of $\boldsymbol{T}_k$ and $\boldsymbol{D}_k$ for a given permutation π_k. First, Proposition 1 shows that $\boldsymbol{T}_k$ and $\boldsymbol{D}_k$ are unique under π_k, which guarantees that $\tilde{\boldsymbol{T}}$, $\tilde{\boldsymbol{D}}$ and the proposed ensemble estimate $\tilde{\boldsymbol{\Sigma}}$ in the method (2.6) are identified.

Proposition 1. *In* (2.5)*, there is a unique* $\boldsymbol{P}_\pi$ *such that* $\boldsymbol{P}_\pi \hat{\boldsymbol{D}}_\pi \boldsymbol{P}'_\pi = \hat{\boldsymbol{D}}$.

Proof. Suppose if there exists another matrix $\boldsymbol{Q}$ such that $\boldsymbol{Q}\hat{\boldsymbol{D}}_\pi\boldsymbol{Q}' = \hat{\boldsymbol{D}}$, we are going to show $\boldsymbol{Q} = \boldsymbol{P}_\pi$. First, $\boldsymbol{P}_\pi \hat{\boldsymbol{D}}_\pi \boldsymbol{P}'_\pi = \hat{\boldsymbol{D}}$ together with $\boldsymbol{Q}\hat{\boldsymbol{D}}_\pi\boldsymbol{Q}' = \hat{\boldsymbol{D}}$ leads to

$$\begin{aligned} \boldsymbol{P}_\pi \hat{\boldsymbol{D}}_\pi \boldsymbol{P}'_\pi &= \boldsymbol{Q}\hat{\boldsymbol{D}}_\pi\boldsymbol{Q}' \\ \hat{\boldsymbol{D}}_\pi &= \boldsymbol{P}'_\pi\boldsymbol{Q}\hat{\boldsymbol{D}}_\pi\boldsymbol{Q}'\boldsymbol{P}_\pi \triangleq \boldsymbol{B}\hat{\boldsymbol{D}}_\pi\boldsymbol{B}', \end{aligned}$$

where $\boldsymbol{B} = (b_{ij})_{p\times p} = \boldsymbol{P}'_\pi\boldsymbol{Q}$. Second, since $\hat{\boldsymbol{D}}_\pi = diag(\hat{d}^2_{\pi(1)}, \ldots, \hat{d}^2_{\pi(p)})$ is a diagonal matrix, so $\hat{d}^2_{\pi(k)} = \sum_{j=1}^p b^2_{kj}\hat{d}^2_{\pi(j)}$, for $k = 1, \ldots, p$. Solving these p equations yields $b_{ii} = \pm 1$ for $i = 1, \ldots, p$ and $b_{kj} = 0$ if $k \neq j$. Therefore, $\boldsymbol{B} = \boldsymbol{P}'_\pi\boldsymbol{Q} = \pm\boldsymbol{I}$, where the notation $\pm\boldsymbol{I}$ represents the identity matrix with each diagonal element taking values 1 or -1. So we have $\boldsymbol{Q} = \pm\boldsymbol{I}\boldsymbol{P}_\pi$. In the sense of reordering the columns of $\hat{\boldsymbol{D}}_\pi$ back to $\hat{\boldsymbol{D}}$, $\boldsymbol{Q} = \boldsymbol{P}_\pi$. □

Second, $\boldsymbol{T}_k = \boldsymbol{P}_{\pi_k}\boldsymbol{T}_{\pi_k}\boldsymbol{P}'_{\pi_k}$ may not be a lower triangular matrix anymore. But its pattern can be determined in the following. Check any two numbers of π_k in sequence, denoted by

$[i, j]$. If the former number i is larger than the latter number j, then $\boldsymbol{T}_k$ is obtained by switching the elements of $\boldsymbol{T}_{\pi_k}$ in the positions $[i, j]$ and $[j, i]$. For example, in the specific order $\pi_* : \{1, 2, 3, 4\} \to \{3, 1, 4, 2\}$, $\boldsymbol{T}_*$ is obtained by switching the elements of $\boldsymbol{T}_{\pi_*}$ in the positions [3, 1] and [1, 3], [3, 2] and [2, 3], as well as [2, 4] and [4, 2]. Precisely, let

$$\boldsymbol{T}_{\pi_*} = \begin{pmatrix} 1 & & & \\ t_{(21)} & 1 & & \\ t_{(31)} & t_{(32)} & 1 & \\ t_{(41)} & t_{(42)} & t_{(43)} & 1 \end{pmatrix},$$

where $t_{(ij)}, i = 2, 3, 4,\ j = 1, \ldots, i-1$, represents the coefficients associated with the jth variable when the ith variable in π_* is regressed on its previous variables. As a result,

$$\boldsymbol{T}_* = \begin{pmatrix} 1 & & t_{(31)} & \\ t_{(21)} & 1 & t_{(32)} & t_{(42)} \\ & & 1 & \\ t_{(41)} & & t_{(43)} & 1 \end{pmatrix}.$$

Based on this pattern of $\boldsymbol{T}_*$, the elements of the diagonal matrix $\boldsymbol{D}_* = diag(d^2_{(1)}, d^2_{(2)}, d^2_{(3)}, d^2_{(4)})$ can be interpreted as: $d^2_{(1)}$ is the residual variance when the first variable is regressed on the third variable in π_*; $d^2_{(2)}$ is the residual variance when the second variable is regressed on the first, third and fourth variables in π_*; $d^2_{(3)}$ is the variance of the third variable in π_*; $d^2_{(4)}$ is the residual variance when the fourth variable is regressed on the first and third variables in π_*.

2.4 The Order-Invariant Cholesky-Log-GARCH Models

2.4.1 Cholesky-Log-GARCH Model Estimation

In modern financial management, many tasks can be reduced to the estimation of a sequence of large $p \times p$ covariance matrices $\{\mathbf{\Sigma}_t\}$ based on the (conditionally) independently $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$ – distributed data $\boldsymbol{y}_t$, $t = 1, 2, \ldots, n$, where $\boldsymbol{y}_t$ is the shock (innovation) at time t of a multivariate time series of returns of p assets in a portfolio.

For a univariate time series $\{\epsilon_t\}$ where $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$, the (asymmetric) log-GARCH model (u, v) (Geweke, 1986) is defined as

$$\begin{aligned} \epsilon_t &= \sigma_t \eta_t \\ \log\, \sigma_t^2 &= \beta_0 + \sum_{i=1}^{v} (\alpha_{i+} 1_{\{\epsilon_{t-i}>0\}} + \alpha_{i-} 1_{\{\epsilon_{t-i}<0\}}) \log\, \epsilon_{t-i}^2 + \sum_{k=1}^{u} \beta_k \log\, \sigma_{t-k}^2, \end{aligned}$$

where $1_{\{\cdot\}}$ represents the indicator function, and η_t is a sequence of independent and identically distributed variables with $E\eta_0 = 0$ and $E\eta_0^2 = 1$. The usual symmetric log-GARCH model corresponds to the case $\alpha_{i+} = \alpha_{i-}$ for $i = 1, \ldots, v$.

By using the modified Cholesky decomposition in Section 2.2, we can obtain an orthogonal transformation of $\boldsymbol{y}_t$ such that $\boldsymbol{T}_t \boldsymbol{y}_t \equiv \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}_t)$ where $\boldsymbol{D}_t = \mathrm{diag}(d_{1;t}^2, \ldots, d_{p;t}^2)$. Therefore, for each $d_{j;t}^2$, $j = 1, 2, \ldots, p$, we consider to model $\log\, d_{j;t}^2$ by using log-GARCH(u,v) model defined recursively in time by

$$\log\, d_{j;t}^2 = \beta_0^{(j)} + \sum_{i=1}^{v} (\alpha_{i+}^{(j)} 1_{\{\epsilon_{j;t-i}>0\}} + \alpha_{i-}^{(j)} 1_{\{\epsilon_{j;t-i}<0\}}) \log\, \epsilon_{j;t-i}^2 + \sum_{k=1}^{u} \beta_k^{(j)} \log\, d_{j;t-k}^2. \tag{2.7}$$

The log-GARCH model (2.7) allows for asymmetric effects between positive and negative latent factors for variance estimate. Moreover, the model also incorporates information from the past reflecting the time-varying nature of the financial data.

To estimate the parameters in the log-GARCH model (2.7), we employ a quasi-maximum likelihood approach similar to that in Francq and Zakoïan (2010). It requires the initial

values $\tilde{d}^2_{j;t}$ of $d^2_{j;t}$ and $\tilde{\boldsymbol{\epsilon}}^{(j)}$ of $\boldsymbol{\epsilon}^{(j)} = (\epsilon_{j;1}, \ldots, \epsilon_{j;n})'$. We obtain $\tilde{d}^2_{j;t}$ as in Pedeli, Fokianos and Pourahmadi (2015) based on the moving block approach (Lopes, McCullogh and Tsay, 2012). More precisely, at time t, a moving block is constructed with m observations that are centered at t. At both left and right end of the data range, the block size m is truncated when it exceeds the observed time window. Then $\tilde{d}^2_{j;t}$ is the residual variance when the jth variable is regressed on all the other regressors using observations $\boldsymbol{y}_{\langle t-\frac{m-1}{2}\rangle}, \ldots, \boldsymbol{y}_{\langle t+\frac{m-1}{2}\rangle}$, where $\langle z\rangle = 1$ if $z \leq 1$, $\langle z\rangle = n$ if $z \geq n$, and otherwise equals the largest integer not greater than z. More precisely, with the data matrix $\mathbb{Y}_t = (\boldsymbol{y}_{\langle t-\frac{m-1}{2}\rangle}, \ldots, \boldsymbol{y}_{\langle t+\frac{m-1}{2}\rangle})'$, define its jth column to be $\boldsymbol{y}^{(j)}_t$. Let $\mathbb{Y}^{(-j)}_t$ be $\mathbb{Y}_t$ without the column $\boldsymbol{y}^{(j)}_t$, then for each j we have

$$\tilde{d}^2_{j;t} = \widehat{Var}(\boldsymbol{y}^{(j)}_t - \mathbb{Y}^{(-j)}_t \hat{\boldsymbol{b}}^{(j)}_t), \tag{2.8}$$

where $\hat{\boldsymbol{b}}^{(j)}_t = \underset{\boldsymbol{b}^{(j)}_t}{\arg\min} \|\boldsymbol{y}^{(j)}_t - \mathbb{Y}^{(-j)}_t \boldsymbol{b}^{(j)}_t\|^2_2$. In addition, the initial value $\tilde{\boldsymbol{\epsilon}}^{(j)}$ in (2.7) is the residual from (2.1) of the MCD approach using all the observations. Precisely, with the data matrix $\mathbb{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$, define its jth column to be $\boldsymbol{y}^{(j)}$, and $\mathbb{W}^{(j)} = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(j-1)})$ represents the first (j-1) columns of $\mathbb{Y}$. We have

$$\tilde{\boldsymbol{\epsilon}}^{(j)} = \begin{cases} \boldsymbol{y}^{(1)}, & j = 1, \\ \boldsymbol{y}^{(j)} - \mathbb{W}^{(j)} \hat{\boldsymbol{c}}^{(j)}, & j = 2, \ldots, p, \end{cases} \tag{2.9}$$

where $\hat{\boldsymbol{c}}^{(j)} = \underset{\boldsymbol{c}^{(j)}}{\arg\min} \|\boldsymbol{y}^{(j)} - \mathbb{W}^{(j)} \boldsymbol{c}^{(j)}\|^2_2$. With $\tilde{d}^2_{j;t}$ and $\tilde{\boldsymbol{\epsilon}}^{(j)}$ available, the parameter vector $\boldsymbol{\phi}^{(j)} = (\beta^{(j)}_0, \beta^{(j)}_1, \ldots, \beta^{(j)}_u, \alpha^{(j)}_{1+}, \ldots, \alpha^{(j)}_{v+}, \alpha^{(j)}_{1-}, \ldots, \alpha^{(j)}_{v-})'$ can be estimated by fitting the log-GARCH model in (2.7).

From the fitted log-GARCH model, we are able to obtain the estimate $\hat{d}^2_{j;t}$ of $d^2_{j;t}$. The sample variance of the first $l = 5$ values of $\tilde{\boldsymbol{\epsilon}}^{(j)}$ is used as $\hat{d}^2_{j;1}$ (Francq, Wintenberger and Zakoian, 2013). The estimate $\hat{d}^2_{j;t}$, $t = 2, 3, \ldots, n$, can be obtained using the fitted log-GARCH model recursively. If a permutation mapping π is under consideration, j represents

the jth variable of the sequence $(\pi(1), \pi(2), \ldots, \pi(p))$. In addition, to reduce large number of parameters in $\boldsymbol{T}_t$, we consider a time-invariant lower triangular matrix $\boldsymbol{T}_t = \boldsymbol{T}$ (Pedeli, Fokianos and Pourahmadi, 2015). For the proposed order-invariant Cholesky-log-GARCH model, the $\tilde{\boldsymbol{T}}$ is thus constructed as in (2.6), while the diagonal of $\boldsymbol{D}_t$ is estimated by the log-GARCH model (2.7), and consequently we construct $\tilde{\boldsymbol{\Sigma}}_t = \tilde{\boldsymbol{T}}^{-1}\tilde{\boldsymbol{D}}_t\tilde{\boldsymbol{T}'}^{-1}$. An algorithm for estimating $\boldsymbol{\Sigma}_t$ for multivariate time series data is summarized as follows:

Algorithm 1. (*Estimation*)

Step 1: *Input centered time series data* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$.

Step 2: *Generate* M *permutation mappings* π_k *as in* (2.3), $k = 1, 2, \ldots, M$.

Step 3: *For each permutation* π_k, *construct* $\hat{\boldsymbol{T}}_{\pi_k}$ *from the estimates of regression coefficients in* (2.4) *using* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. *At each time* t, *the diagonal of* $\hat{\boldsymbol{D}}_{k;t}$ *is obtained from the log-GARCH model* (2.7) *using* $\tilde{d}^2_{j;t}$ *in* (2.8) *and* $\tilde{\epsilon}_{j;t}$ *in* (2.9).

Step 4: *Transform* $\hat{\boldsymbol{T}}_{\pi_k}$ *back to the original order:* $\hat{\boldsymbol{T}}_k = \boldsymbol{P}_{\pi_k}\hat{\boldsymbol{T}}_{\pi_k}\boldsymbol{P}'_{\pi_k}$.

Step 5: $\tilde{\boldsymbol{T}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{T}}_k$, $\tilde{\boldsymbol{D}}_t = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{D}}_{k;t}$ *as in* (2.6).

Step 6: *At each time* t, $\tilde{\boldsymbol{\Sigma}}_t = \tilde{\boldsymbol{T}}^{-1}\tilde{\boldsymbol{D}}_t\tilde{\boldsymbol{T}'}^{-1}$.

One point we would like to emphasize in Algorithm 1 is that before applying the log-GARCH model in Step 3, the initial values $\tilde{\epsilon}^2_{j;t}$ and $\tilde{d}^2_{j;t}$ need to be arranged to the original order based on j such that the order of variables in $\hat{\boldsymbol{D}}_{k;t}$ is the same as that in $\hat{\boldsymbol{T}}_k$ computed in Step 4.

2.4.2 Volatility Prediction

Prediction of the volatility given the past information is of the central importance in the financial markets. In this section we develop an approach to predict the volatility (covariance) matrices using the Cholesky-log-GARCH models.

With n observations $\boldsymbol{y}_t$, $t = 1, 2, \ldots, n$, the goal is to predict the covariance matrix at time $t = n + h$. To start, all the observations are used to estimate the parameter vector $\boldsymbol{\phi}^{(j)}$ in (2.7), and the fitted model can be used to predict $d^2_{j;n+1}$. Then the h-step ahead prediction is

easily implemented by recursively generating $\hat{\epsilon}_{j;t} = \hat{d}_{j;t}\eta_t$ and calculating $\hat{d}^2_{j;t}$ with h-1 times in t to obtain $\hat{d}^2_{j;n+h}$ to form the diagonal of $\hat{\boldsymbol{D}}_{k;n+h}$. Incorporating this process into the framework of Algorithm 1 leads to the following h-step ahead prediction for time series data by the proposed order-invariant Cholesky-log-GARCH model:

Algorithm 2. (*Prediction*)

Step 1: *For each* π_k, *use* $\hat{\boldsymbol{\phi}}^{(j)}$ *to predict* $\hat{d}^2_{j;n+1}$ *by the log-GARCH model* (2.7).

Step 2: *Generate* $\hat{\epsilon}_{j;n+1} = \hat{d}_{j;n+1}\eta_{n+1}$, *where* $\eta_{n+1} \sim \mathcal{N}(0,1)$, $j = 1, 2, \ldots, p$.

Step 3: *Predict* $\hat{d}^2_{j;n+2}$ *using* $\hat{\boldsymbol{\phi}}^{(j)}$, $\hat{d}^2_{j;n+1}$ *and* $\hat{\epsilon}_{j;n+1}$ *in the log-GARCH model* (2.7).

Step 4: *Recursively repeat the mechanism of Step 2 - 3 in time until* $\hat{d}^2_{j;n+h}$ *is obtained.*

Step 5: *Construct* $\hat{\boldsymbol{D}}_{k;n+h}$ *with* $\hat{d}^2_{j;n+h}$ *as its diagonal elements and transform* $\hat{\boldsymbol{T}}_{\pi_k}$ *back to the original order:* $\hat{\boldsymbol{T}}_k = \boldsymbol{P}_{\pi_k}\hat{\boldsymbol{T}}_{\pi_k}\boldsymbol{P}'_{\pi_k}$.

Step 6: $\tilde{\boldsymbol{T}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{T}}_k$, $\tilde{\boldsymbol{D}}_{n+h} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{D}}_{k;n+h}$ *as in* (2.6).

Step 7: $\tilde{\boldsymbol{\Sigma}}_{n+h} = \tilde{\boldsymbol{T}}^{-1}\tilde{\boldsymbol{D}}_{n+h}\tilde{\boldsymbol{T}'}^{-1}$.

Algorithm 2 not only provides a way to predict the covariance matrix at time $t = n+h$, but also enables us to compute the covariance matrices at times $t = n+1, n+2, \ldots, n+h-1$, by using $\hat{\boldsymbol{D}}_{k;n+1}, \hat{\boldsymbol{D}}_{k;n+2}, \ldots, \hat{\boldsymbol{D}}_{k;n+h-1}$ instead of $\hat{\boldsymbol{D}}_{k;n+h}$ in Step 5.

2.4.3 The Competing Models

This section reviews three classes of multivariate GARCH models where their empirical performances are compared later using three data sets. These methods have the common feature of converting the problem of estimating a covariance matrix into two separate subproblems: modeling the Cholesky factor and the diagonal components or modeling variances and the correlations.

The first class is composed of the proposed method and the hyperspherical specification approach (Pedeli, Fokianos and Pourahmadi, 2015). The proposed method consists of two versions, denoted by M1 and M2, which represent the methods in (2.6) with the lower triangular $\boldsymbol{T}_k$ estimated using the Lasso and least squares, respectively. This provides a contrast between

the Lasso and least squares estimation of $\hat{\boldsymbol{T}}_k$. The hyperspherical specification approach for log-GARCH model, denoted as HS, relies on the standard Cholesky factor of the correlation matrix by the hyperspherical parameterization (Rebonato and Jäckel, 2000). Specifically, decompose $\boldsymbol{\Sigma} = \boldsymbol{HRH}$, where $\boldsymbol{H} = diag(\sigma_1, \ldots, \sigma_p)$ is the diagonal matrix of standard deviations of variable vector $\boldsymbol{X}$, and $\boldsymbol{R}$ is the correlation matrix of $\boldsymbol{X}$. Then consider the standard Cholesky decomposition $\boldsymbol{R} = \boldsymbol{BB}'$, where $\boldsymbol{B} = (b_{ij})_{p\times p}$ is a lower triangular matrix. GARCH models are used to estimate the diagonal elements of matrix $\boldsymbol{H}$. The entries of $\boldsymbol{B}$ are parameterized by the hyperspherical parameterization as $b_{11} = 1, b_{i1} = \cos(\theta_{i1}), i = 2, \ldots, p$ and

$$b_{ij} = \begin{cases} \cos(\theta_{ij})\Pi_{k=1}^{j-1}\sin(\theta_{ik}),\ j = 2, \ldots, i-1; i = 3, \ldots, p; \\ \\ \Pi_{k=1}^{j-1}\sin(\theta_{ik}),\ j = i; i = 2, \ldots, p, \end{cases}$$

where θ_{ij} are unconstrained parameters in the range of $[0, \pi)$ (Rapisarda, Brigo and Mercurio, 2007).

The next class of benchmark models which would compare the role of the variable orders, includes the Cholesky-log-GARCH model using the original order, the Cholesky-log-GARCH model with the order selected by BIC (Dellaportas and Pourahmadi, 2012; Pedeli, Fokianos and Pourahmadi, 2015) and the Cholesky-log-GARCH model with the order selected by the best permutation algorithm (Rajaratnam and Salzman, 2013). We denote these three models by ORIG, BIC and BPA respectively. The Cholesky factor matrix $\boldsymbol{T}$ is modeled using Lasso estimates. The BIC method determines the order of variables in the MCD in a forward selection fashion. That is, in each step, it selects a new variable having the smallest value of BIC when regressing this variable on the rest candidate variables. For example, let $\mathcal{C} = \{X_{i_1}, \ldots, X_{i_k}\}$ be the candidate set of variables and there are $p - k$ variables already being chosen in an order. By regressing each X_j, $j = i_1, \ldots, i_k$ onto the rest variables in $\mathcal{C}$, we can assign the variable corresponding to the minimum BIC value among the k regressions to the kth position of the order. In addition, The BPA selects the order of variables such that $||\boldsymbol{D}||_F^2$

is minimized, where $||\cdot||_F$ denotes the Frobenius norm, and $\boldsymbol{D}$ is the diagonal matrix in the MCD approach.

The last class consists of two well-known models in the finance literature. They are the constant conditional correlation (CCC) and the dynamic conditional correlation (DCC) models of order (1,1), denoted by CCC and DCC, respectively. The CCC model assumes the conditional correlation matrix is time-invariant, while the DCC model imposes a simple dynamic structure on the conditional correlation matrices. For estimation of the latter two models, we use the *eccc.estimation(·)* and *dcc.estimation(·)* functions in R (R version 3.0.3).

2.5 Application

In this section, three financial time series data sets are used to illustrate and evaluate the performance of the proposed order-invariant Cholesky-log-GARCH model. The first data set is the monthly stock returns of 12 U.S. bluechips with $n = 251$ observations. The second data set is composed of $n = 436$ weekly returns of 97 stocks in the Standard and Poor's 100 index (S&P100). The third data set considers a higher dimension with $p = 200$, which is an extension of the second data set with additional 103 stocks selected from the Standard and Poor's 500 index (S&P500). Denote by $\hat{\boldsymbol{\Sigma}}_t = (\hat{\omega}_{ij;t})_{p\times p}$ the estimate for the covariance matrix $\boldsymbol{\Sigma}_t = (\omega_{ij;t})_{p\times p}$, $t = 1, \ldots, n$. To measure the accuracy of the covariance matrix estimate $\hat{\boldsymbol{\Sigma}}_t$, we consider the entropy loss Δ_{1t}, the Kullback-Leibler loss Δ_{2t}, two quadratic loss functions Δ_{3t} and Δ_{4t} (up to some scale) as follows,

$$
\begin{aligned}
\Delta_{1t} &= \frac{1}{p^2}\,[\mathrm{tr}[\boldsymbol{\Sigma}_t^{-1}\hat{\boldsymbol{\Sigma}}_t] - \log|\boldsymbol{\Sigma}_t^{-1}\hat{\boldsymbol{\Sigma}}_t| - p],\\
\Delta_{2t} &= \frac{1}{p^2}\,[\mathrm{tr}[\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\Sigma}_t] - \log|\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\Sigma}_t| - p],\\
\Delta_{3t} &= \frac{1}{p^2}\,[\mathrm{tr}(\boldsymbol{\Sigma}_t^{-1}\hat{\boldsymbol{\Sigma}}_t - \boldsymbol{I})]^2,\\
\Delta_{4t} &= \frac{1}{p^2}\,[\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\Sigma}_t - \boldsymbol{I})]^2.
\end{aligned}
$$

We also use the mean absolute error and mean squared error given by

$$\text{MAE}_t = \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}|\hat{\omega}_{ij;t} - \omega_{ij;t}| \quad \text{and} \quad \text{MSE}_t = \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\omega}_{ij;t} - \omega_{ij;t})^2.$$

For each loss function above, we report the averages of the performance measures over time t, i.e., $\text{MAE} = \sum_{t=1}^{n}\text{MAE}_t/n$, $\text{MSE} = \sum_{t=1}^{n}\text{MSE}_t/n$ and $\Delta_i = \sum_{t=1}^{n}\Delta_{it}/n$, $i = 1, 2, 3, 4$.

Since the true covariance matrix is unavailable, we employ the moving block technique to get a reliable proxy (Lopes, McCullogh and Tsay, 2012). That is, a sample covariance matrix is calculated within each moving block as a benchmark to measure the accuracy of a covariance matrix estimate. In practice, the block size m is selected from a pre-specified set $\{m_1, \ldots, m_B\}$ such that it can stabilize at least two loss measures in $\hat{\Delta}_1, \ldots, \hat{\Delta}_4$. More specifically, the average loss functions $\hat{\Delta}_i = \sum_{t=1}^{n}\hat{\Delta}_{it}/n$, $i = 1, 2, 3, 4$ are calculated for each $m_j, j = 1, \ldots, B$. The optimal m_k is chosen to satisfy that the relative difference $|\hat{\Delta}_i^{(m_k)} - \hat{\Delta}_i^{(m_{k-1})}|/\hat{\Delta}_i^{(m_{k-1})}$ does not change significantly for at least two loss functions. Using this procedure, $m = 130$ is selected for the first data set and $m = 300$ for the both second and third data sets.

2.5.1 Monthly Stock Returns of 12 U.S. Bluechips

This data set with $n = 251$ returns and $p = 12$ stocks was monthly stock prices from January 1990 to December 2010. The data are multiplied by 10 for the practical purposes (Pedeli, Fokianos and Pourahmadi, 2015). Here, the log-GARCH (1, 1) model with $u = 1$ and $v = 1$ in (2.7) is used. The appropriateness of using log-GARCH (1,1) model for estimating $\boldsymbol{D}_t$ is examined via Figure 2.1 by plotting $\log \tilde{d}_{j;t}^2$ against its lag-1 values $\log \tilde{d}_{j;t-1}^2$, $j = 1, 2, \ldots, 12$. As each plot shows a roughly linear relationship between $\log \tilde{d}_{j;t}^2$ and $\log \tilde{d}_{j;t-1}^2$, the log-GARCH (1,1) model appears to be a reasonable choice.

Table 2.1 reports the average loss measures of the estimates $\hat{\boldsymbol{\Sigma}}_t$ over time t and their standard errors (in parenthesis) for eight methods in Section 2.4.3. From the results, it is seen that our proposed methods M1 and M2 considerably outperform other approaches, and is comparable to the BPA. The proposed methods are better than the BPA in terms of Δ_2, Δ_4

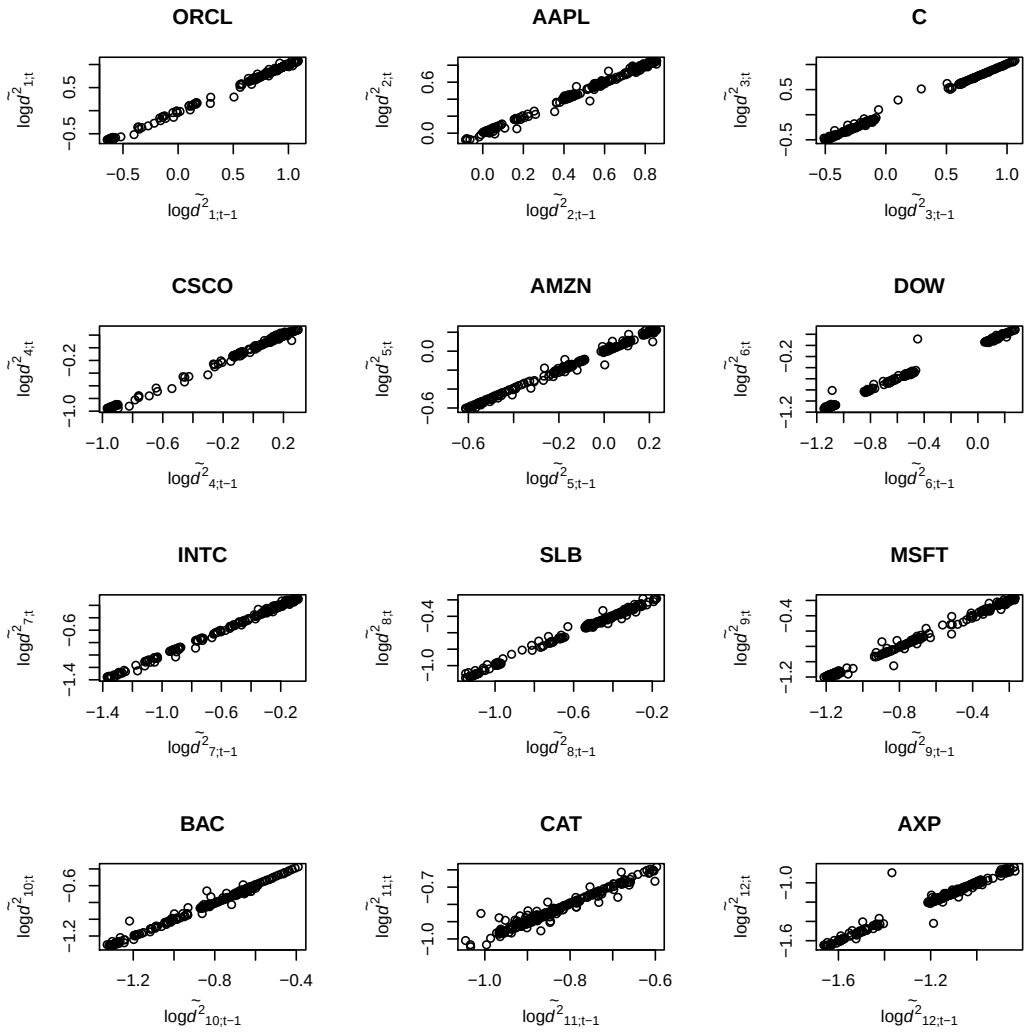

Figure 2.1: Scatter plots between log $\tilde{d}^2_{j;t}$ and log $\tilde{d}^2_{j;t-1}$, $j = 1, 2, \ldots, 12$, for the monthly returns of 12 U.S. bluechips.

and comparable under Δ_1, MAE and MSE. Note that the MCD approach with the original order (i.e., ORIG) is comparable with that based on the BIC order selection (i.e., BIC). It implies that the BIC-based order selection of the MCD approach for the log-GARCH model may not be helpful for improving the estimation accuracy. In addition, the HS, the CCC and DCC models perform better than ORIG and BIC under all loss measures except Δ_3 and MSE. By addressing the order issue of the MCD-based approach, our proposed order-invariant Cholesky-log-GARCH models perform much better than the HS, the CCC and DCC models. It is worth pointing out that the performances of the M1 and M2 are very comparable in this

example. One explanation is that the number of variable $p = 12$ is relatively small for $n = 251$ observations so that the Lasso technique may not be able to show its full advantage.

Table 2.1: The averages and standard errors (in parenthesis) of six loss measures of the multivariate series of monthly returns of 12 U.S. bluechips.

	Δ_1	Δ_2	Δ_3	Δ_4	MAE	MSE
ORIG	8.413 (0.112)	9.607 (0.221)	0.596 (0.027)	0.843 (0.064)	0.320 (0.002)	0.186 (0.003)
BIC	8.850 (0.150)	8.909 (0.186)	0.653 (0.037)	0.820 (0.059)	0.321 (0.004)	0.192 (0.004)
BPA	2.250 (0.056)	2.218 (0.088)	0.124 (0.009)	0.111 (0.011)	0.170 (0.003)	0.059 (0.002)
HS	4.072 (0.148)	3.026 (0.109)	0.544 (0.050)	0.156 (0.015)	0.236 (0.005)	0.120 (0.006)
CCC	4.023 (0.376)	4.165 (0.232)	0.705 (0.212)	0.508 (0.052)	0.220 (0.008)	0.226 (0.052)
DCC	4.770 (0.455)	5.183 (0.330)	0.954 (0.268)	0.754 (0.089)	0.239 (0.011)	0.355 (0.098)
M1	2.354 (0.084)	1.407 (0.033)	0.258 (0.017)	0.037 (0.002)	0.176 (0.003)	0.055 (0.002)
M2	2.379 (0.085)	1.404 (0.033)	0.263 (0.018)	0.038 (0.002)	0.193 (0.004)	0.065 (0.003)

Next, we employ the log-GARCH (1,1) model to predict future covariance matrices. We consider one-step, two-step and five-step ahead predictions based on Algorithm 2. Specifically, we use the first k observations to estimate the model and then predict the covariance matrices at times $t = k + h$, $h = 1, 2, 5$. By varying $k = 200, 201, \ldots, 250$, Table 2.2 shows the results of one-step, two-step and five-step ahead predictions in terms of average loss measures for different methods. We do not include the ORIG here since its performance is very similar to the BIC. Note that the HS can only predict one-step ahead because its prediction at a future time point t^* requires the observations by time t^*-1. From the results in Table 2.2, it is clear that the proposed methods generally outperform the BIC, HS, CCC and DCC models and are comparable with the BPA for the one-step and two-step ahead predictions in most of the performance measures. However, in the case of five-step ahead prediction, the BPA seems to perform slightly better than the proposed methods. Also in terms of Δ_3 and MSE, the proposed methods do not appear to be as good as the BIC for the five-step ahead prediction. One possible explanation is that the data in 2009 could behave quite differently due to the financial crisis. Consequently, the proposed methods could be inferior to some

extent when conducting prediction at those time points, especially for the five-step ahead prediction. Extension of the proposed methods for robustness will be discussed in Section 2.6.

Table 2.2: The averages and standard errors (in parenthesis) of six loss measures for predictions of the multivariate series of monthly returns of 12 U.S. bluechips.

		Δ_1	Δ_2	Δ_3	Δ_4	MAE	MSE
One	BIC	0.062 (0.002)	0.102 (0.007)	0.589 (0.081)	2.248 (0.328)	0.398 (0.006)	0.277 (0.011)
	HS	1.192 (1.139)	0.060 (0.010)	1.246 (0.219)	0.871 (0.203)	0.322 (0.015)	0.255 (0.027)
	BPA	0.045 (0.001)	0.045 (0.002)	0.422 (0.043)	0.271 (0.033)	0.257 (0.003)	0.129 (0.007)
	CCC	0.073 (0.013)	0.076 (0.008)	3.101 (1.148)	1.846 (0.376)	0.384 (0.025)	0.740 (0.191)
	DCC	0.068 (0.009)	0.084 (0.011)	2.376 (0.751)	2.683 (0.619)	0.390 (0.029)	0.809 (0.237)
	M1	0.035 (0.002)	0.039 (0.004)	0.541 (0.090)	0.360 (0.061)	0.274 (0.011)	0.178 (0.018)
	M2	0.036 (0.002)	0.038 (0.003)	0.564 (0.092)	0.327 (0.056)	0.285 (0.014)	0.190 (0.023)
Two	BIC	0.065 (0.002)	0.102 (0.007)	0.618 (0.066)	2.213 (0.321)	0.401 (0.006)	0.278 (0.010)
	BPA	0.046 (0.002)	0.044 (0.002)	0.457 (0.062)	0.285 (0.044)	0.254 (0.004)	0.132 (0.009)
	CCC	0.066 (0.011)	0.072 (0.007)	2.398 (0.918)	1.559 (0.293)	0.373 (0.025)	0.694 (0.216)
	DCC	0.086 (0.015)	0.080 (0.010)	4.276 (1.591)	2.239 (0.491)	0.448 (0.048)	1.626 (0.591)
	M1	0.039 (0.003)	0.042 (0.004)	0.727 (0.137)	0.402 (0.069)	0.308 (0.018)	0.267 (0.046)
	M2	0.040 (0.002)	0.040 (0.003)	0.739 (0.134)	0.337 (0.056)	0.322 (0.021)	0.290 (0.054)
Five	BIC	0.070 (0.002)	0.108 (0.007)	0.703 (0.086)	2.396 (0.360)	0.399 (0.005)	0.282 (0.009)
	BPA	0.046 (0.002)	0.047 (0.003)	0.424 (0.047)	0.354 (0.069)	0.251 (0.005)	0.127 (0.008)
	CCC	0.075 (0.013)	0.077 (0.008)	3.067 (1.099)	1.715 (0.392)	0.390 (0.030)	0.962 (0.314)
	DCC	0.110 (0.018)	0.086 (0.011)	6.143 (1.707)	2.436 (0.646)	0.518 (0.057)	2.649 (0.788)
	M1	0.075 (0.015)	0.049 (0.004)	3.701 (1.597)	0.455 (0.072)	0.443 (0.055)	0.986 (0.322)
	M2	0.080 (0.018)	0.044 (0.003)	4.505 (2.069)	0.335 (0.050)	0.471 (0.064)	1.106 (0.371)

2.5.2 Weekly Stock Returns of 97 Stocks in the S&P100

The second data set comprises of $n = 436$ observations and $p = 97$ stocks in the S&P100 weekly recorded from August 23, 2004 to December 12, 2012. The data are multiplied by 100 for the practical purpose. The CCC and DCC models are not included for comparison due to their convergence issues of the R functions *eccc.estiomation(·)* and *dcc.estimation(·)*, caused

by the large dimensionality of p. Here we also employ the log-GARCH (1, 1) model for the estimate of $\boldsymbol{D}_t$. To justify the properness of employing log-GARCH (1,1) model, Figure 2.2 reports the scatter plots between $\log \tilde{d}^2_{j;t}$ and $\log \tilde{d}^2_{j;t-1}$ for nine randomly selected stocks. The rest of the stocks have similar linear patterns and hence their plots are omitted.

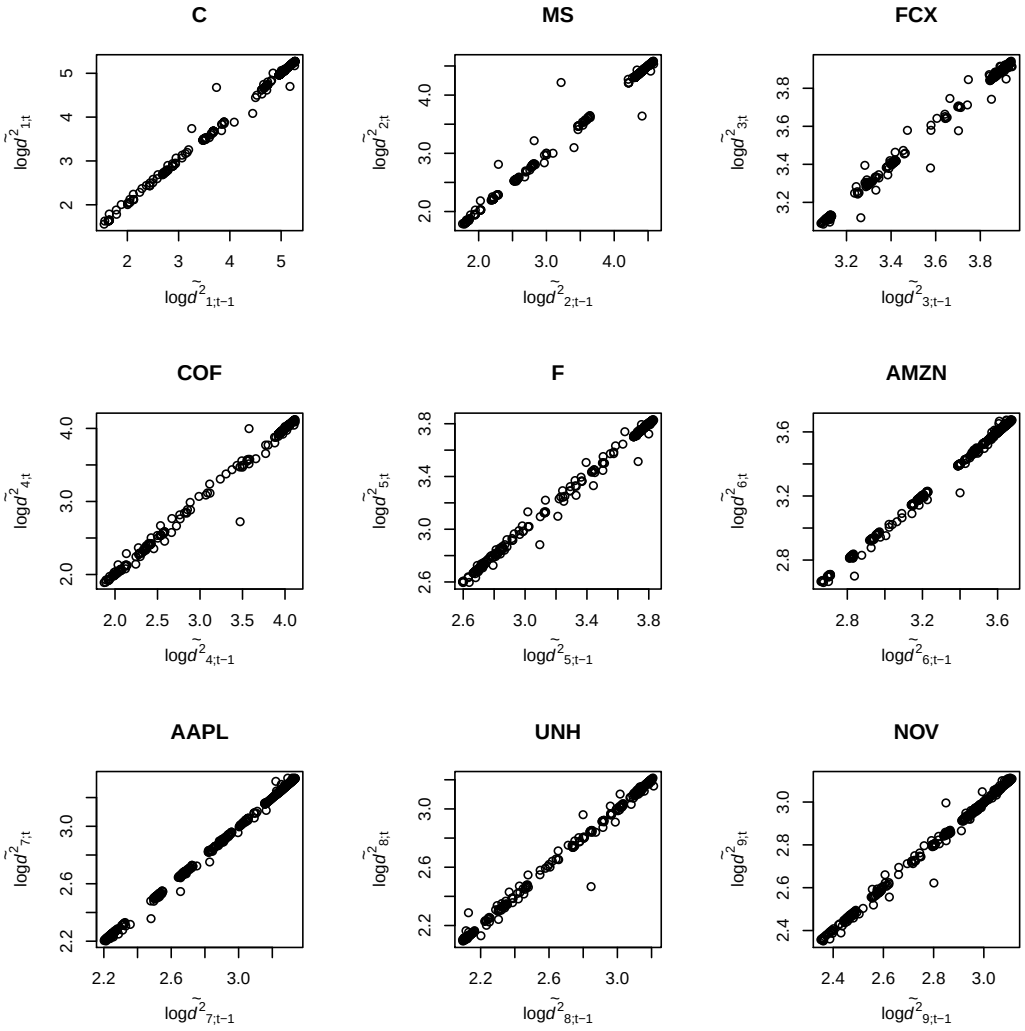

Figure 2.2: Scatter plots between $\log \tilde{d}^2_{j;t}$ and $\log \tilde{d}^2_{j;t-1}$, $j = 1, \ldots, 9$, for the weekly returns of 9 randomly selected stocks in the S&P100.

Table 2.3 reports the results for the loss measures Δ_1, Δ_2, Δ_3, Δ_4, MAE and MSE. It is clear that the proposed order-invariant Cholesky-log-GARCH models give more accurate estimation than other approaches. It also shows that the ORIG has very similar performance as the BIC. Although the HS outperforms the ORIG and BIC, its performance is not as good

Table 2.3: The averages and standard errors (in parenthesis) of six loss measures of the weekly returns of 97 stocks.

	Δ_1	Δ_2	Δ_3	Δ_4	MAE	MSE
ORIG	0.0350 (0.0006)	0.0330 (0.0004)	13.91 (0.583)	10.49 (0.292)	6.380 (0.089)	90.61 (2.167)
BIC	0.0311 (0.0008)	0.0348 (0.0005)	12.55 (0.974)	12.12 (0.337)	9.004 (0.486)	364.0 (87.92)
BPA	0.0077 (0.0003)	0.0046 (0.0001)	1.316 (0.088)	0.154 (0.010)	3.877 (0.096)	36.12 (1.594)
HS	0.0078 (0.0003)	0.0100 (0.0004)	1.458 (0.146)	2.403 (0.213)	4.688 (0.169)	72.33 (6.354)
M1	0.0083 (0.0003)	0.0035 (0.0001)	1.842 (0.114)	0.045 (0.002)	2.963 (0.081)	28.69 (1.933)
M2	0.0087 (0.0005)	0.0023 (0.0001)	2.821 (0.241)	0.032 (0.002)	14.99 (0.241)	364.1 (15.88)

as M1 and M2. The BPA gains advantage over other methods in terms of Δ_3. Additionally, M1 using the Lasso technique appears to enjoy the best performance among all methods. In particular, it significantly dominates all other approaches in terms of MAE and MSE. Note that these two criteria focus on the element-wise errors of $\hat{\mathbf{\Sigma}}$ and the Lasso can play an effective role in shrinking each element in the estimated covariance matrix.

Based on the log-GARCH (1,1) model, we also evaluate the prediction of one-, two-, and five-step ahead for different methods in comparison. Since the performance of the ORIG is very similar to that of BIC, we only include the results of the BIC in the table. With the estimated model from the first k observations, we make prediction of the covariance matrices at time $t = k + h$, where $h = 1, 2, 5$. By varying $k = 350, 351, \ldots, 435$, Table 2.4 summarizes the results from different methods by averaging the loss measures. The results further show that M1 works substantially better than other methods, especially in terms of MAE and MSE. The BPA performs the best under Δ_1 and Δ_3.

Next, we investigate how the size of observations n affects the performance of the proposed approaches. We use the S&P100 data set with the number of variables being fixed at $p = 20$. Specifically, we randomly choose 20 stocks and set $n = 50, 75, 100, 150, 200$ in this study. Here we confine our attention to the MCD-based methods, including the BIC, BPA, M1 and M2. Figure 2.3 plots the loss measures against the number of observations n given fixed $p = 20$. Clearly, the proposed methods have a great advantage when the size of observations n is small.

Table 2.4: The averages and standard errors (in parenthesis) of six loss measures for predictions of the weekly returns of 97 stocks.

		Δ_1	Δ_2	Δ_3	Δ_4	MAE	MSE
One	BIC	0.0415 (0.0010)	0.0314 (0.0013)	20.18 (1.014)	9.194 (0.954)	6.212 (0.266)	83.52 (7.579)
	BPA	0.0111 (0.0005)	0.0065 (0.0001)	1.974 (0.197)	0.337 (0.042)	3.682 (0.264)	38.87 (4.936)
	HS	0.0130 (0.0005)	0.0074 (0.0003)	2.543 (0.180)	0.564 (0.110)	3.146 (0.245)	31.25 (4.874)
	M1	0.0152 (0.0006)	0.0047 (0.0001)	4.581 (0.337)	0.033 (0.003)	2.592 (0.199)	20.56 (2.903)
	M2	0.0182 (0.0010)	0.0042 (0.0001)	6.635 (0.532)	0.043 (0.003)	13.95 (0.328)	281.4 (11.33)
Two	BIC	0.0404 (0.0011)	0.0315 (0.0013)	19.35 (1.043)	9.700 (0.965)	6.220 (0.275)	83.14 (7.822)
	BPA	0.0117 (0.0005)	0.0065 (0.0001)	2.248 (0.220)	0.300 (0.040)	3.555 (0.261)	36.67 (4.785)
	M1	0.0164 (0.0007)	0.0048 (0.0001)	5.273 (0.381)	0.031 (0.002)	2.657 (0.188)	20.68 (2.739)
	M2	0.0194 (0.0010)	0.0043 (0.0001)	7.479 (0.592)	0.044 (0.002)	14.58 (0.345)	310.9 (12.70)
Five	BIC	0.0438 (0.0012)	0.0328 (0.0015)	22.60 (1.186)	10.30 (1.218)	6.337 (0.277)	83.86 (7.873)
	BPA	0.0156 (0.0007)	0.0068 (0.0001)	4.059 (0.372)	0.203 (0.029)	3.114 (0.238)	39.78 (4.102)
	M1	0.0237 (0.0010)	0.0053 (0.0001)	10.19 (0.667)	0.059 (0.004)	3.652 (0.113)	38.13 (2.071)
	M2	0.0267 (0.0013)	0.0049 (0.0001)	13.04 (0.959)	0.082 (0.004)	18.37 (0.435)	536.5 (23.11)

The BPA performs comparable with the proposed methods M1 and M2. However, the BIC performs poorly when the number of observation is as small as 50. As the size n increases, the gap between the proposed methods and BIC becomes smaller, but the proposed methods still perform better than BIC.

In addition, we also study the impact of the number of variables p on the performance of the MCD-based methods. Here we fix the data with the first $n = 150$ observations. We randomly selected $p = 20, 40, 60, 70$ stocks among 97 stocks, respectively. Figure 2.4 displays the loss measures against the number of variables p. The results from Figure 2.4 further confirm that our proposed methods perform substantially better than the BIC in terms of all measures of accuracy. The BPA is comparable with the proposed methods. As the number of variables p increases, all the loss measures increase, as expected, the curve for the BIC rises rapidly as p becomes large. In contrast, our proposed methods do not change much in terms of Δ_2 and Δ_4. Moreover, for M1 using the Lasso technique appears to have its advantage in

Table 2.5: The averages and standard errors (in parenthesis) of six loss measures of the weekly returns of 200 stocks.

	Δ_1	Δ_2	Δ_3	Δ_4	MAE	MSE
ORIG	4.8791 (0.3986)	0.5657 (0.0108)	976.30 (79.73)	112.65 (2.174)	41.969 (0.567)	85.606 (1.936)
BIC	0.7780 (0.0531)	0.4720 (0.0108)	156.05 (10.63)	93.946 (2.169)	52.067 (8.873)	91.231 (14.44)
BPA	0.2942 (0.0274)	0.0073 (0.0002)	1671.8 (216.9)	1.6353 (0.199)	5.3089 (0.129)	131.96 (43.74)
HS	0.3457 (0.0472)	0.0096 (0.0003)	69.562 (9.462)	1.5940 (0.074)	6.0723 (0.243)	20.892 (0.543)
M1	0.3119 (0.0279)	0.0061 (0.0001)	63.235 (5.592)	0.4445 (0.026)	5.0369 (0.124)	20.167 (1.072)
M2	0.5890 (0.0591)	0.0033 (0.0001)	118.65 (11.84)	0.2132 (0.004)	28.335 (0.280)	37.537 (0.413)

MAE and MSE for larger values of p.

2.5.3 Weekly Stock Returns of 200 Stocks from the S&P500

To evaluate the performance of the proposed methods in a much higher dimensional situation, the third data set combines the second data set with additional 103 stocks chosen from the S&P500, weekly recorded from August 23, 2004 to December 12, 2012. It has $n = 436$ observations and $p = 200$ variables. The data are multiplied by 100 for the practical purpose. The log-GARCH (1, 1) model is used in the estimation.

Table 2.5 presents six loss measures Δ_1, Δ_2, Δ_3, Δ_4, MAE and MSE obtained for each method in comparison. The conclusions are very similar to that in Section 2.5.2. The proposed M1 appears to be the best among all approaches. It outperforms the other four methods regarding Δ_3, MAE and MSE. In terms of Δ_2 and Δ_4, the performance of the proposed M2 dominates the performances of all other approaches and the M1 is the second best. In addition, when the number of variables p is large, the performance of the BPA is not very promising. In contrast, the proposed order-invariant Cholesky-log-GARCH model work consistently well in the high-dimensional settings.

As the computational efficiency is of great importance in analyzing multivariate time series, we also evaluate the computational time of different methods in comparison. Specifically, we compare the computational time for the BIC, BPA, HS, M1 and M2 approaches, which are

implemented in R program (R version 3.0.3). The implementation are carried out on an Intel Core 2.50 GHz processor. Figure 2.5 shows the computational time of the five methods using $p = 12, 30, 50, 97, 150, 200$ stocks, where the stocks are randomly selected from the aforementioned 200 stocks with weekly returns from August 23, 2004 to December 12, 2012.

It can be seen that the computational time of the proposed methods (M1 and M2) appears to be linear with respect to the number of variables p. While the computational time of the BIC, BPA and HS heavily depends on the values of p. For the BIC, BPA and HS methods, they are fast to compute in the low-dimensional cases. But their computational times dramatically rise as p increases, especially when p is larger than 100. For example, the computational time for the BIC is roughly $O(p^2)$ when $p < 100$, but $O(p^3)$ when $p > 100$. Also, the computational time of the HS method increases exponentially with respect to p, especially in the high dimensions. A possible explanation is that the HS method encounters convergence issue due to the high dimensionality. The computational time of the BPA is almost linear in low dimensions, but increases rapidly when $p > 150$. Finally, it is worth pointing out that the computational time of the proposed methods is closely associated with the choice of M, the number of permutations used in the MCD. Choosing a small value of M can largely reduce the computational time of the proposed methods. Actually, based on our studies, the estimation accuracy of the proposed methods with $M = 50$ does not decrease much from that with $M = 100$.

2.6 Discussion

In this chapter, we introduced an order-invariant Cholesky-log-GARCH model to analyze the multivariate financial time series data. The proposed estimator is properly assembled from a set of multiple estimates of $\boldsymbol{T}$ and $\boldsymbol{D}$ under different orders used in MCD. The proposed method not only provides accurate covariance matrix estimation, but also gives accurate prediction of the covariance or volatility matrices at future time points. The analysis of three real data examples of growing dimensions shows the superior performance of our proposed

order-invariant Cholesky-log-GARCH model in terms of both estimation and prediction. We verified the appropriateness of log-GARCH (1, 1) in the proposed method using certain lag-scatter plots of innovation variances. How to properly choose the values of u and v in a log-GARCH (u, v) model in our context can be an interesting topic for the future research.

Finally, we would like to remark that, for the multivariate time series in financial applications, the data may include abnormal observations, especially due to the financial crisis. In this application, the robustness of the proposed method is an important research topic. We alternatively considered the ensemble estimate by the element-wise median of $\hat{\boldsymbol{T}}_k$ and $\hat{\boldsymbol{D}}_k$ instead of taking average. This idea was applied to the 12 U.S. bluechips data set in Section 2.5.1. It appears that such an extension leads to a robust covariance matrix estimator and performs very well in terms of Δ_3 and MSE. We will further investigate the robustness of our proposed method in the future research.

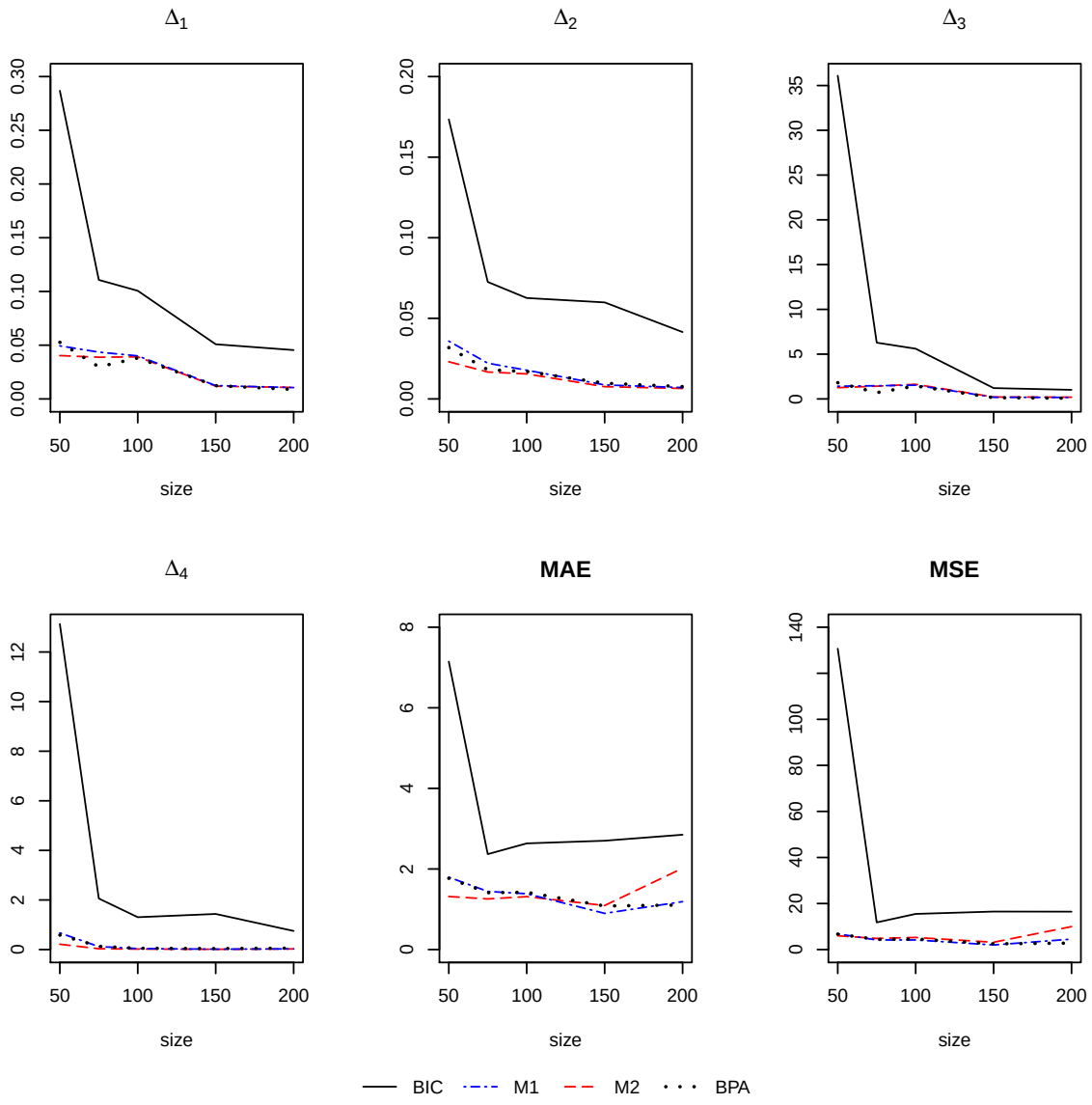

Figure 2.3: Plots of six loss measures with respect to the the size of observations n for $p = 20$ randomly selected stocks in the S&P100.

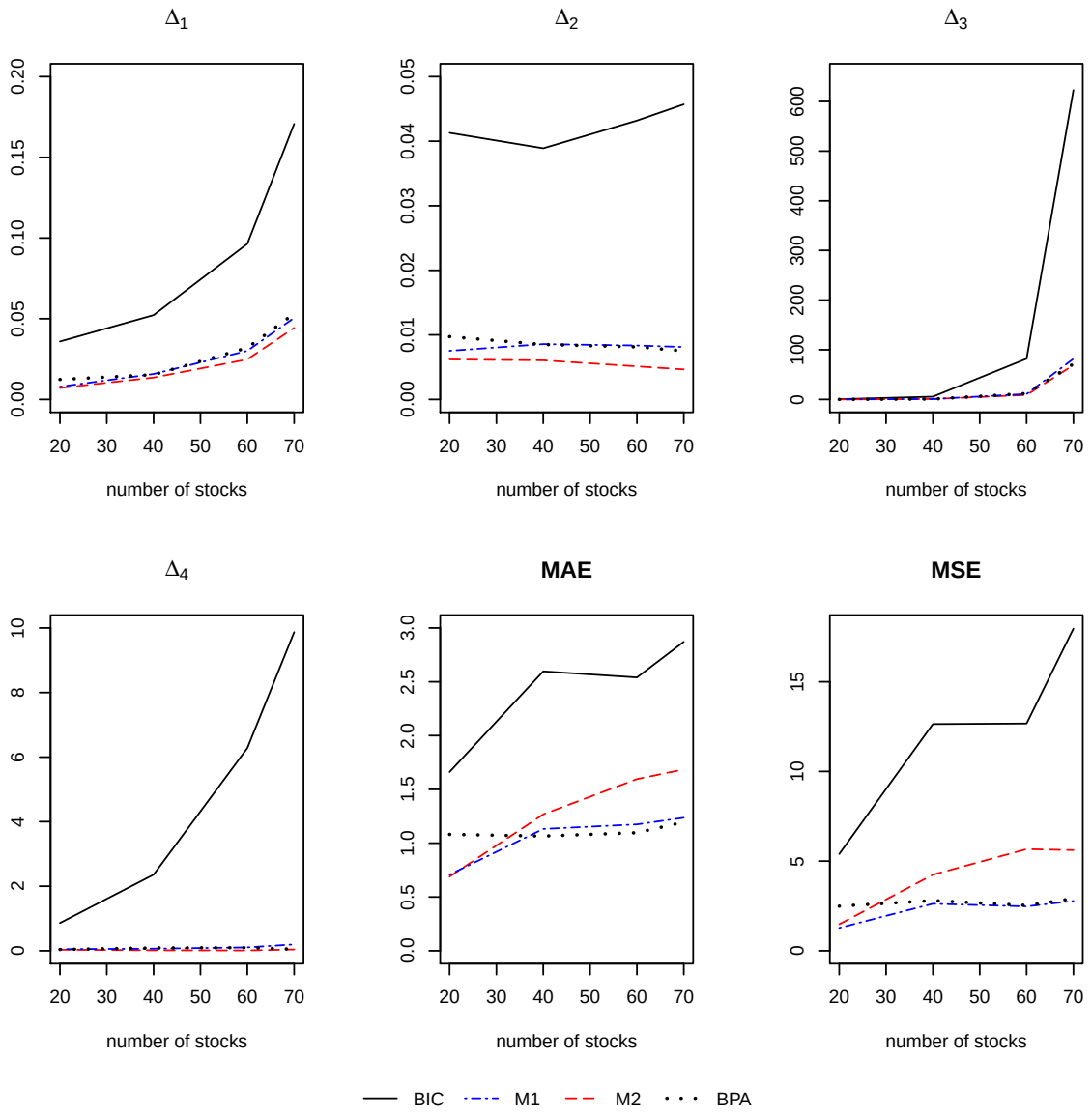

Figure 2.4: Plots of six loss measures with respect to the dimensionality of variables p for the first $n = 150$ observations in the S&P100.

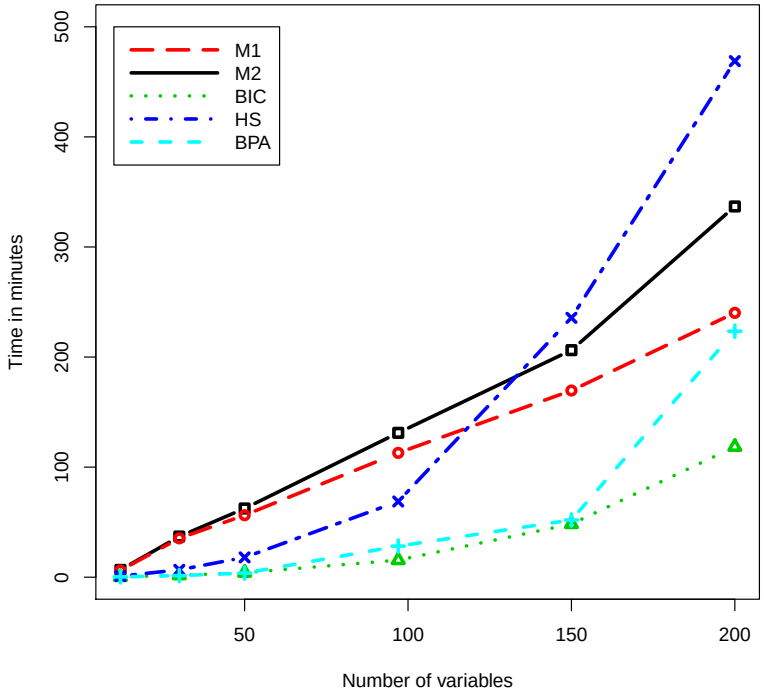

Figure 2.5: The computational time of the methods using $p = 12, 30, 50, 97, 150, 200$ stocks.

Chapter 3 An Improved Modified Cholesky Decomposition Method for Inverse Covariance Matrix Estimation

3.1 Introduction

As data-collection technique advances, high-dimensional data are commonly observed over a wide range of fields such as gene expression, image recognition, social connection and climate forecasting. Hence, there is an urgent need to develop statistical analysis tools for high-dimensional data. The estimation of large sparse inverse covariance matrix is of fundamental importance in the multivariate analysis and various statistical applications. For example, in the classification problem, linear discriminant analysis (LDA) needs the inverse covariance matrix to compute the classification rule. In financial applications, portfolio optimization often utilizes the inverse covariance matrix for minimizing the portfolio risk. A sparse estimate of inverse covariance matrix not only provides a parsimonious model structure, but also gives meaningful interpretation since zeroes in the inverse covariance matrix indicate the conditional independence among the variables.

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is a p-dimensional vector of random variables with an unknown covariance matrix $\boldsymbol{\Sigma}$. Without loss of generality, we assume that the expectation of $\boldsymbol{X}$ is zero. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be the n independently and identically distributed observations following a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with mean equal to the zero vector and covariance matrix $\boldsymbol{\Sigma}$. In the high-dimensional cases, often the number of variables p is close to or larger than the sample size n. The goal of this work is to estimate the inverse covariance matrix $\boldsymbol{\Omega} = (\omega_{ij})_{p\times p} = \boldsymbol{\Sigma}^{-1}$. Particular interest is to identify zero entries of ω_{ij}, since $\omega_{ij} = 0$ implies the conditional independence between X_i and X_j given all the other random variables. Although one can estimate the covariance matrix and then obtain its inverse, the inverse is often computationally intensive in the high-dimensional cases. Moreover, the inverse of a

sparse covariance matrix often would not result in sparse structure for the inverse covariance matrix. Therefore, it is desirable to obtain a sparse inverse covariance matrix estimate directly.

The estimation of sparse inverse covariance matrix has attracted great attention from different researchers. Yuan and Lin (2007) proposed a Graphical Lasso (Glasso) method, which gives a sparse and shrinkage estimator of $\mathbf{\Omega}$ by penalizing the negative log-likelihood as

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}} \quad -\log|\mathbf{\Omega}| + \mathrm{tr}[\mathbf{\Omega}\boldsymbol{S}] + \rho||\mathbf{\Omega}||_1,$$

where $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$ is the sample covariance matrix, $\rho \geq 0$ is a tuning parameter, and $||\mathbf{\Omega}||_1$ denotes L_1 norm. As the L_1 penalty is imposed on the off-diagonal entries of the inverse covariance matrix when minimizing the negative log-likelihood, it can encourage some of the off-diagonal entries of the estimated $\mathbf{\Omega}$ to be exact zeroes. Different variation of the Glasso formulation has also been later studied by Friedman, Hastie, and Tibshirani (2008), Rocha, Zhao and Yu (2008), Rothman et al. (2008), Yuan (2008), Deng and Yuan (2009), and Yuan (2010). Some theoretical properties of Glasso method are developed by Yuan and Lin (2007), Raskutti et al. (2008), Rothman et al. (2008) and Lam and Fan (2009). In particular, the results from Raskutti et al. (2008) and Rothman et al. (2008) suggest that, although better than the sample covariance matrix, the Glasso estimate may not perform well when p is larger than the sample size n.

Several other methods of sparse inverse covariance matrix estimation have also been proposed in the literature. Pourahmadi (1999, 2001) developed the modified Cholesky decomposition (MCD) approach to estimate an inverse matrix. This method reduces the challenge of modeling an inverse covariance matrix into modeling a sequence of regression problems, and provides an unconstrained and statistically interpretable parametrization of an inverse covariance matrix. In addition, a neighborhood-based approach was introduced by Meinshausen and Bühlmann (2006). It first estimates each column of the inverse covariance matrix by the scaled Lasso or Dantzig selector, and then adjusts the matrix estimator to be symmetric. Based on the same framework, Sun and Zhang (2013) proved that the scaled Lasso method

guarantees the fastest rate of convergence under mild conditions. Fan, Fan and Lv (2008) developed a factor model to estimate both covariance matrix and its inverse. They also studied the estimation in the asymptotic framework that both the dimension p and the sample size n go to infinity. Xue and Zou (2012) introduced a rank-based estimation for estimating high-dimensional nonparametric graphical models under a strong sparsity assumption that the true inverse covariance matrix has only a few nonzero entries. There are also a few work focusing on the inference for the inverse covariance matrix estimation. Drton and Perlman (2008) proposed a new method for model selection in Gaussian graphical models based on simultaneous hypotheses testings of the conditional independence between variables. Sun and Zhang (2012) derived a residual-based estimator to construct confidence intervals for entries of the estimated inverse covariance matrix. Some recent Bayesian literature can also be found in the work of Cheng and Lenkoski (2012), Wang (2012), Bhadra and Mallick (2013), Scutari (2013) and Mohammadi and Wit (2015), among many others.

In this paper, we propose a sparse inverse covariance matrix estimate based on the MCD approach for the high-dimensional data. Although the MCD approach is statistically meaningful and applicable in the high dimensions, the resultant estimate often depends on the order of the random variables $X_1, \ldots, X_p$. In many applications, the variable order is often not available or cannot be pre-determined before the analysis. To overcome this difficulty, the proposed sparse estimate considers an ensemble estimation for the inverse covariance matrix under multiple permutations of the variable orders. Specifically, we take average on the multiple estimates of the Cholesky factors, and consequently construct the final estimate of the inverse covariance matrix. Such an estimator has small variability and is order-invariant because of the ensemble effort. Since the ensemble estimate of the Cholesky factor matrix may not have sparse structure, we adopt the hard thresholding technique on the ensemble Cholesky factor matrix to obtain the sparsity, thus leading to the sparse structure in the estimated inverse covariance matrix.

The rest of the paper is organized as follows. In Section 3.2, we briefly review the MCD approach to estimate the inverse covariance matrix. In Section 3.3, we address the order issue

of the MCD approach and propose an ensemble sparse estimate of $\mathbf{\Omega}$. Simulation studies are reported in Section 3.4 and illustrative examples of real data are presented in Section 3.5. We conclude our work with some discussion in Section 3.6.

3.2 Modified Cholesky Decomposition of Ω

The key idea of the modified Cholesky decomposition approach is that the inverse covariance matrix $\mathbf{\Omega}$ can be decomposed using a unique lower triangular matrix $\boldsymbol{T}$ and a unique diagonal matrix $\boldsymbol{D}$ with positive diagonal entries such that (Pourahmadi, 1999)

$$\mathbf{\Omega} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}.$$

The entries of $\boldsymbol{T}$ and the diagonal of $\boldsymbol{D}$ are unconstrained and interpretable as regression coefficients and corresponding variances when one variable X_j is regressed on its predecessors $X_1, \ldots, X_{j-1}$. Clearly, here an order for variables $X_1, \ldots, X_p$ is pre-specified. Specifically, consider $X_1 = \epsilon_1$, and for $j = 2, \ldots, p$, define

$$\begin{aligned} X_j &= \sum_{t=1}^{j-1} a_{jt} X_t + \epsilon_j \\ &= \boldsymbol{Z}_j^T \boldsymbol{a}_j + \epsilon_j, \end{aligned} \tag{3.1}$$

where $\boldsymbol{Z}_j = (X_1, \ldots, X_{j-1})'$, and $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{j,j-1})'$ is the corresponding vector of regression coefficients. The error ϵ_j is assumed to be independent with zero mean and variance d_j^2. Denote $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$ and $\boldsymbol{D} = Cov(\boldsymbol{\epsilon}) = diag(d_1^2, \ldots, d_p^2)$. Then the p regression models in (3.1) can be expressed in the matrix form $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\epsilon}$, where $\boldsymbol{A}$ is a lower triangular matrix with a_{jt} in the $(j,\ t)$th position, and 0 as its diagonal entries. Thus one can easily write $\boldsymbol{T}\boldsymbol{X} = \boldsymbol{\epsilon}$ with $\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{A}$ to derive the expression of $\mathbf{\Omega} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}$. The MCD approach therefore reduces the challenge of modeling a covariance matrix into the task of modeling $(p-1)$ regression problems.

Note that in the MCD approach, it requires the regression of one variable on its predecessors. It means that the order of $X_1, \ldots, X_p$ needs to be pre-determined for the estimation of $\boldsymbol{T}$ and $\boldsymbol{D}$ matrices. Obviously, different orders of variables would lead to different estimates of $\boldsymbol{T}$ and $\boldsymbol{D}$, and consequently different estimates of $\boldsymbol{\Omega}$. For example, to see this clearly, we generate 20 observations from a 4-dimensional normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega}$ is a sparse matrix with 1 as its diagonal and $\boldsymbol{\Omega}_{13} = \boldsymbol{\Omega}_{31} = 0.5$. We consider two different variable orders $\pi_1 = (1, 2, 3, 4)$ and $\pi_2 = (1, 4, 3, 2)$, and obtain the corresponding estimates $\hat{\boldsymbol{\Omega}}_1$ and $\hat{\boldsymbol{\Omega}}_2$ based on the MCD (3.1) as follows, with the regression coefficients $\boldsymbol{a}_j$ estimated according to (3.3)

$$\hat{\boldsymbol{\Omega}}_1 = \begin{pmatrix} 1.80 & -0.13 & 0.75 & 0.06 \\ -0.13 & 1.94 & 0.24 & 0.07 \\ 0.75 & 0.24 & 0.83 & 0.08 \\ 0.06 & 0.07 & 0.08 & 1.41 \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Omega}}_2 = \begin{pmatrix} 0.85 & 0.22 & 0.64 & 0.08 \\ 0.22 & 1.82 & -0.11 & 0.08 \\ 0.64 & -0.11 & 1.64 & 0.05 \\ 0.08 & 0.08 & 0.05 & 1.41 \end{pmatrix}.$$

Clearly, the estimates $\hat{\boldsymbol{\Omega}}_1$ and $\hat{\boldsymbol{\Omega}}_2$ are different. Hence, it is important to address the issue of the variable order for the MCD-based approach. Wagaman and Levina (2009) proposed an Isomap method to find the order of variables based on their correlations prior to applying banding techniques. Rajaratnam and Salzman (2013) introduced a so-called "best permutation algorithm" to recover the natural order of variables in autoregressive models for banded covariance matrix estimation, by minimizing the sum of the diagonals of $\boldsymbol{D}$ in the MCD approach. Dellaportas and Pourahmadi (2012) suggested a search algorithm to choose the order based on Akaike information criterion (AIC) or Bayesian information criterion (BIC). However, a natural variable order of $\boldsymbol{X}$ may not be available in practice, such as in the gene expression data or stock data. Moreover, the order chosen using the aforementioned criteria may not necessarily give an accurate estimate of the inverse covariance matrix. In the next section, we propose an ensemble estimate of the inverse covariance matrix based on MCD,

which can lead to a sparse and order-invariant estimate of the inverse covariance matrix.

3.3 Proposed Sparse Estimate of $\boldsymbol{\Omega}$

Note that MCD-based inverse covariance matrix estimation for $\boldsymbol{\Omega}$ depends on the order of $X_1, \ldots, X_p$. Chang and Tsay (2010) pointing out that the MCD approach is not order invariant, investigated the sensitivity of MCD to order by randomly permuting the variables before estimation. To address this order issue and obtain an accurate estimate $\hat{\boldsymbol{\Omega}} = (\hat{\omega}_{ij})_{p\times p}$, we take advantage of permutations to gain the flexibility such that we can ensemble the multiple estimates under different orders.

Define a permutation mapping $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$, which gives

$$(\pi(1), \pi(2), \ldots, \pi(p)) \,. \tag{3.2}$$

Define the corresponding permutation matrix $\boldsymbol{P}_\pi$ of which the entries in the jth column are all 0 except taking 1 at position $\pi(j)$. Therefore, the transformed data matrix is

$$\mathbb{X}_\pi = \mathbb{X}\boldsymbol{P}_\pi = (\boldsymbol{x}_\pi^{(1)}, \ldots, \boldsymbol{x}_\pi^{(p)}),$$

where $\boldsymbol{x}_\pi^{(j)}$ is the jth column of $\mathbb{X}_\pi$, $j = 1, 2, \ldots, p$. The Lasso technique (Tibshirani, 1996) is employed for the shrinkage purpose and for the situation where p is close to n or even larger than n. The idea of Lasso-type estimator for the Cholesky factor has been used by Huang et al. (2006), Rothaman et al. (2010) and Chang and Tsay (2010). Under a given permutation π, obtain

$$\hat{\boldsymbol{a}}_{\pi(j)} = \arg\min_{\boldsymbol{a}_{\pi(j)}} \|\boldsymbol{x}_\pi^{(\pi(j))} - \mathbb{Z}_\pi^{(\pi(j))}\boldsymbol{a}_{\pi(j)}\|_2^2 + \lambda_{\pi(j)}\|\boldsymbol{a}_{\pi(j)}\|_1, \text{ for } \pi(j) \neq 1, \tag{3.3}$$

and

$$\hat{d}^2_{\pi(j)} = \begin{cases} \widehat{Var}(\boldsymbol{x}^{(1)}_{\pi}), & \pi(j) = 1, \\ \widehat{Var}(\boldsymbol{x}^{(\pi(j))}_{\pi} - \mathbb{Z}^{(\pi(j))}_{\pi}\hat{\boldsymbol{a}}_{\pi(j)}), & \text{otherwise}, \end{cases} \tag{3.4}$$

where $\mathbb{Z}^{(j)}_{\pi}$ represents the first (j-1) columns of $\mathbb{X}_{\pi}$, and $\lambda \geq 0$ is a tuning parameter. Here $\widehat{Var}(\cdot)$ denotes the sample variance. Then we can model the lower triangular matrix $\hat{\boldsymbol{T}}_{\pi}$ with ones on its diagonal and $\hat{\boldsymbol{a}}'_{\pi(j)}$ as its $\pi(j)$th row. Meanwhile, the diagonal matrix $\hat{\boldsymbol{D}}_{\pi}$ has its $\pi(j)$th diagonal element equal to $\hat{d}^2_{\pi(j)}$. Correspondingly, $\hat{\boldsymbol{\Omega}}_{\pi} = \hat{\boldsymbol{T}}'_{\pi}\hat{\boldsymbol{D}}^{-1}_{\pi}\hat{\boldsymbol{T}}_{\pi}$ will be a sparse inverse covariance matrix estimate under π. Transforming back to the original order, we can estimate $\boldsymbol{\Omega}$ as

$$\begin{aligned} \hat{\boldsymbol{\Omega}} &= \boldsymbol{P}_{\pi}\hat{\boldsymbol{\Omega}}_{\pi}\boldsymbol{P}'_{\pi} \\ &= \boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}'_{\pi}\hat{\boldsymbol{D}}^{-1}_{\pi}\hat{\boldsymbol{T}}_{\pi}\boldsymbol{P}'_{\pi} \\ &= (\boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}'_{\pi}\boldsymbol{P}'_{\pi})(\boldsymbol{P}_{\pi}\hat{\boldsymbol{D}}^{-1}_{\pi}\boldsymbol{P}'_{\pi})(\boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}_{\pi}\boldsymbol{P}'_{\pi}) \\ &\triangleq \hat{\boldsymbol{T}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{T}}. \end{aligned} \tag{3.5}$$

Chapter 2 discussed the statistical interpretation of $\boldsymbol{T}_{\pi}$ and $\boldsymbol{D}_{\pi}$, and proved that there is a unique $\boldsymbol{P}_{\pi}$ such that $\boldsymbol{P}_{\pi}\hat{\boldsymbol{D}}^{-1}_{\pi}\boldsymbol{P}'_{\pi} = \hat{\boldsymbol{D}}^{-1}$, which guarantees $\hat{\boldsymbol{T}}$ and $\hat{\boldsymbol{D}}$ are unique under a specified order π. In addition, note that $\hat{\boldsymbol{T}} = \boldsymbol{P}_{\pi}\hat{\boldsymbol{T}}_{\pi}\boldsymbol{P}'_{\pi}$ may no longer be a lower triangular matrix, but it still contains the sparse structure. Suppose we generate M permutation mappings π_k, $k = 1, \ldots, M$. Accordingly, we obtain the corresponding estimates $\hat{\boldsymbol{\Omega}}$, $\hat{\boldsymbol{T}}$, and $\hat{\boldsymbol{D}}$ in (3.5), denoted as $\hat{\boldsymbol{\Omega}}_k$, $\hat{\boldsymbol{T}}_k$, and $\hat{\boldsymbol{D}}_k$ for the permutation π_k.

Based on the multiple estimates $\hat{\boldsymbol{T}}_k$'s and $\hat{\boldsymbol{D}}_k$'s, we consider the ensemble estimate of $\boldsymbol{\Omega}$ as follows

$$\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{T}}'\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{T}} \;\text{ with }\; \tilde{\boldsymbol{T}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{T}}_k,\; \tilde{\boldsymbol{D}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{D}}_k. \tag{3.6}$$

The estimate in (3.6) is able to achieve good estimation accuracy since it reduces the variability in the estimates of $\tilde{\boldsymbol{T}}$ and $\tilde{\boldsymbol{D}}$. It is worth pointing out that we do not consider the averaged estimate $\bar{\boldsymbol{\Omega}}$ based on the ensemble of $\hat{\boldsymbol{\Omega}}_k$

$$\bar{\boldsymbol{\Omega}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Omega}}_k. \tag{3.7}$$

The reason is that the estimation error of $\hat{\boldsymbol{\Omega}}_k$ is already aggregated by the estimation error of $\hat{\boldsymbol{T}}_k$ and $\hat{\boldsymbol{D}}_k$. As shown in the simulations in Section 3.4, the estimate $\bar{\boldsymbol{\Omega}}$ does not give good performance on the estimation.

Although the method (3.6) is able to produce an accurate estimate $\tilde{\boldsymbol{\Omega}}$ with small variability, it fails to capture any sparse structure of the true inverse covariance matrix, especially in the high-dimensional settings, since $\tilde{\boldsymbol{T}}$ in (3.6) does not contain the sparsity. To illustrate this point, we generate 50 observations from normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega}$ is a 15×15 banded structure with main diagonal 1, the first sub-diagonal 0.5 and the second sub-diagonal 0.3. The first three panels of Figure 3.1 display the heat maps for the true inverse covariance matrix $\boldsymbol{\Omega}$, the estimates $\tilde{\boldsymbol{\Omega}}$ in (3.6) and $\bar{\boldsymbol{\Omega}}$ in (3.7). Clearly, there are many non-zeroes in the off-diagonal positions of estimates $\tilde{\boldsymbol{\Omega}}$ and $\bar{\boldsymbol{\Omega}}$.

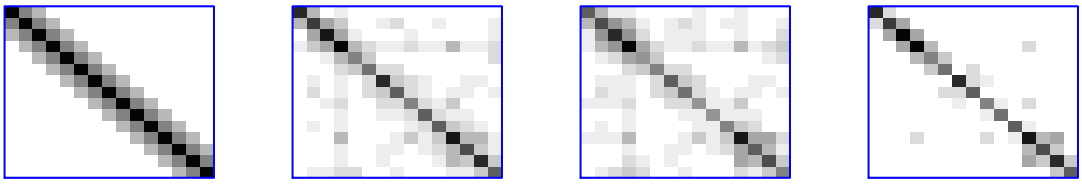

Figure 3.1: Heat maps for the true inverse covariance matrix $\boldsymbol{\Omega}$, the estimates $\tilde{\boldsymbol{\Omega}}$, $\bar{\boldsymbol{\Omega}}$ and the proposed estimate $\tilde{\boldsymbol{\Omega}}_{\delta_{opt}}$. Darker colour indicates higher density; lighter colour indicates lower density.

Therefore, to encourage the sparse structure in the estimate of $\boldsymbol{T}$, we impose a hard thresholding on each entry of the ensemble estimate $\tilde{\boldsymbol{T}}$ in (3.6). The hard thresholding will result in the sparsity of $\tilde{\boldsymbol{T}}$, hence leading to a sparse estimate of $\boldsymbol{\Omega}$ with only a little cost of

losing accuracy to some acceptable extent. The resultant estimate of $\mathbf{\Omega}$ not only enjoys the sparse structure in high dimensions, but also requires no information of the order of variables in the MCD before the analysis.

The hard thresholding procedure is described as follows. let $\tilde{\boldsymbol{T}} = (\tilde{t}_{ij})_{p\times p}$ be the ensemble estimate obtained from method (3.6) and a hard thresholding is denoted by δ. Then $\tilde{\boldsymbol{T}}_\delta = (\tilde{t}_{ij}^{(\delta)})_{p\times p}$ is defined as

$$\tilde{t}_{ij}^{(\delta)} = \begin{cases} \tilde{t}_{ij}, \text{ if } |\tilde{t}_{ij}| > \delta, \\ \\ 0, \text{ if } |\tilde{t}_{ij}| \leq \delta, \end{cases} \tag{3.8}$$

then the sparse estimate $\tilde{\mathbf{\Omega}}_\delta = \tilde{\boldsymbol{T}}_\delta' \tilde{\boldsymbol{D}}^{-1} \tilde{\boldsymbol{T}}_\delta$. So a large hard thresholding will definitely improve the performance of capturing sparse structure, but reduce the accuracy. Now a natural question arises as how to decide an appropriate hard thresholding, and we suggest to use BIC. That is, for a given hard thresholding δ_l, $l = 1, \ldots, H$, the corresponding BIC(δ_l) (Yuan 2007) is computed by

$$\text{BIC}(\delta_l) = -\log|\tilde{\mathbf{\Omega}}_{\delta_l}| + \text{tr}[\tilde{\mathbf{\Omega}}_{\delta_l}\boldsymbol{S}] + \frac{\log n}{n}\sum_{i\leq j}\tilde{e}_{ij}(l), \tag{3.9}$$

where $\tilde{\mathbf{\Omega}}_{\delta_l} = (\tilde{\omega}_{ij}^{(\delta_l)})_{p\times p} = \tilde{\boldsymbol{T}}_{\delta_l}' \tilde{\boldsymbol{D}}^{-1} \tilde{\boldsymbol{T}}_{\delta_l}$ using the hard thresholding δ_l. $\tilde{e}_{ij}(l) = 0$ if $\tilde{\omega}_{ij}^{(\delta_l)} = 0$, and $\tilde{e}_{ij}(l) = 1$ otherwise. The optimal hard thresholding δ_{opt} is chosen as that produces the minimum BIC, and our proposed order-invariant sparse inverse covariance estimate is

$$\tilde{\mathbf{\Omega}}_{\delta_{opt}} = \tilde{\boldsymbol{T}}_{\delta_{opt}}' \tilde{\boldsymbol{D}}^{-1} \tilde{\boldsymbol{T}}_{\delta_{opt}}. \tag{3.10}$$

Clearly, the method (3.6) can be viewed as a special case of the proposed estimate with hard thresholding $\delta = 0$. The fourth panel in Figure 3.1 shows the heat map of the proposed estimate $\tilde{\mathbf{\Omega}}_{\delta_{opt}}$ in (3.10). It has much less off-diagonal non-zeroes compared with $\tilde{\mathbf{\Omega}}$ and $\bar{\mathbf{\Omega}}$.

The algorithm of proposed order-invariant sparse estimate for inverse covariance matrix $\mathbf{\Omega}$

based on the MCD is summarized as follows:

Algorithm 3.

Step 1*: Input centered data.*

Step 2*: Generate M permutation mappings π_k as in* (3.2)*, $k = 1, 2, \ldots, M$.*

Step 3*: Under each permutation mapping π_k, construct $\hat{\boldsymbol{T}}_{\pi_k}$ from the estimates of regression coefficients in* (3.3)*. Obtain $\hat{\boldsymbol{D}}_{\pi_k}$ from the corresponding residual variances in* (3.4)*.*

Step 4*: Transform to the original order: $\hat{\boldsymbol{T}}_k = \boldsymbol{P}_{\pi_k} \hat{\boldsymbol{T}}_{\pi_k} \boldsymbol{P}'_{\pi_k}$ and $\hat{\boldsymbol{D}}_k = \boldsymbol{P}_{\pi_k} \hat{\boldsymbol{D}}_{\pi_k} \boldsymbol{P}'_{\pi_k}$.*

Step 5*: $\tilde{\boldsymbol{T}} = \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{T}}_k$, $\tilde{\boldsymbol{D}} = \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{D}}_k$ as in* (3.6)*.*

Step 6*: Obtain $\tilde{\boldsymbol{T}}_{\delta_{opt}}$ from* (3.8) *by applying δ_{opt} to $\tilde{\boldsymbol{T}}$, where δ_{opt} is selected by* (3.9)*.*

Step 7*: $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{T}}'_{\delta_{opt}} \tilde{\boldsymbol{D}}^{-1} \tilde{\boldsymbol{T}}_{\delta_{opt}}$ as in* (3.10)*.*

As seen in Algorithm 3, the proposed method attempts to balance between the accuracy and sparsity of the estimate for $\boldsymbol{\Omega}$. Meanwhile, we would like to point out that Algorithm 3 is also very flexible with respect to the objective in practice. If there is no clear evidence to find sparse estimate, one can set the hard thresholding $\delta = 0$ for the estimation of $\boldsymbol{\Omega}$. As shown in Section 3.4, such an estimator has good performance in certain setting of covariance structure.

Note that the proposed method needs to choose the number of permutations M. It is known that the number of all possible permutations increases rapidly as the number of variables p increases. To choose an appropriate number of permutations M for efficient computation, we have tried $M = 10, 30, 50, 80, 100, 120$ and 150 as the number of randomly selected permutations from all the possible permutations. The performance results are quite comparable when M is larger than 30. Hence, in this paper we choose $M = 100$ for the proposed order-invariant MCD method.

3.4 Simulation

In this section, we conduct a simulation study to evaluate the performance of the proposed method in comparison with several existing approaches. Two versions of the proposed method

are considered, denoted by M1 and M2. The proposed method M1 represents the estimate in (3.6) with hard thresholding $\delta = 0$. The proposed method M2 stands for the estimate in (3.10) with hard thresholding chosen by the BIC criterion as in (3.9). Among the comparison methods, the first one is the inverse sample covariance matrix, denoted as $\boldsymbol{S}^{-1}$. It serves as a benchmark estimate of $\boldsymbol{\Omega}$. The second one is the MCD method for estimating $\boldsymbol{\Omega}$ with the order chosen by BIC criterion (Dellaportas and Pourahmadi, 2012), denoted as BIC. The key idea of such an approach is to determine the order of variables in the MCD in a forward selection fashion. That is, in each step, it selects a new variable having the smallest value of BIC when regressing this variable on the variables in the candidate set. For example, let $\mathcal{C} = \{X_{i_1}, \ldots, X_{i_k}\}$ be the candidate set of variables and there are $p-k$ variables already being chosen in an order. By regressing each X_j, $j = i_1, \ldots, i_k$ onto the rest variables in $\mathcal{C}$, we can assign the variable having the order k if it gives the minimum BIC value in the k regressions. The third approach is the Best Permutation Algorithm (Rajaratnam and Salzman, 2013), denoted by BPA. It selects the order of variables such that $||\boldsymbol{D}||_F^2$ is minimized, where $||\cdot||_F$ denotes the Frobenius norm, and $\boldsymbol{D}$ is the diagonal matrix in the MCD approach. The fourth method is an naive ensemble estimate $\bar{\boldsymbol{\Omega}}$ in (3.7), denoted by AVE. The last method for comparison is the Graphical Lasso (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman, Hastie and Tibshirani, 2008), denoted as Glasso.

Denote by $\hat{\boldsymbol{\Omega}} = (\hat{\omega}_{ij})_{p\times p}$ an estimate for the covariance matrix $\boldsymbol{\Omega} = (\omega_{ij})_{p\times p}$. To measure the accuracy of an inverse covariance matrix estimate, we consider the Kullback-Leibler loss Δ_1, the entropy loss Δ_2 and the quadratic loss Δ_3 (up to some scale) as follows,

$$\begin{aligned}
\Delta_1 &= \frac{1}{p}\,(\mathrm{tr}[\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}] - \log|\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}| - p),\\
\Delta_2 &= \frac{1}{p}\,(\mathrm{tr}[\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}] - \log|\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}| - p),\\
\Delta_3 &= \frac{1}{p}\,[\mathrm{tr}(\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}} - \boldsymbol{I})]^2.
\end{aligned}$$

We also use the mean absolute error and mean squared error given by

$$\text{MAE} = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}|\hat{\omega}_{ij} - \omega_{ij}| \quad \text{and} \quad \text{MSE} = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\omega}_{ij} - \omega_{ij})^2.$$

In addition, to gauge the performance of the estimates in capturing the sparse structure, the false selection loss (FLS) are used, which is the summation of false positive (FP) and false negative (FN). We say a FP occurs if a nonzero element in the true matrix is incorrectly estimated as a zero. Similarly, a FN occurs if a zero element in the true matrix is incorrectly identified as a nonzero. The FSL is computed in percentage as (FP + FN) / p^2. For each loss function above, we report the averages of the performance measures over 50 simulations.

We consider the following six inverse covariance matrix structures.

Model 1. $\mathbf{\Omega}_1 = \text{MA}(0.5, 0.3)$. The main diagonal elements are 1 with first sub-diagonal elements 0.5 and second sub-diagonal elements 0.3.

Model 2. $\mathbf{\Omega}_2$ is generated by randomly permuting rows and corresponding columns of $\mathbf{\Omega}_1$.

Model 3. $\mathbf{\Omega}_3 = \begin{pmatrix} \text{CS}(0.5) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix}$, where CS(0.5) represents a 10×10 compound structure matrix with diagonal elements 1 and others 0.5. $\mathbf{0}$ indicates a matrix with all elements 0.

Model 4. $\mathbf{\Omega}_4 = \text{AR}(0.5)$. The conditional covariance between any two random variables X_i and X_j is fixed to be $0.5^{|i-j|}$, $1 \le i, j \le p$.

Model 5. $\mathbf{\Omega}_5^{-1} = \text{diag}(p, p-1, p-2, \ldots, 1)$.

Model 6. $\mathbf{\Omega}_6 = \boldsymbol{B}'\boldsymbol{H}^{-1}\boldsymbol{B}$, where $\boldsymbol{H} = 0.01 \times \boldsymbol{I}$, and $\boldsymbol{B} = (-\phi_{t,s})$ with $\phi_{t,t} = 1$, $\phi_{t+1,t} = 0.8$, and $\phi_{t,s} = 0$ otherwise.

Model 1 is a sparse banded structure. *Model* 2 permutates the rows and corresponding columns of *Model* 1 randomly. *Model* 3 is a block compound structure on the upper left corner. *Model* 4 is a autoregressive structure that has homogeneous variances and correlations declining with distance. This model is more dense than the other models. The structures of *Model* 5 and *Model* 6 are also used in Huang et al. (2006). For each model, we generate

normally distributed data with three settings of sample sizes and variable sizes: (1) $n = 50, p = 30$; (2) $n = 50, p = 50$ and (3) $n = 50, p = 100$. Table 3.1 to Table 3.3 report the loss measures of the estimates averaged over 50 simulations and their corresponding standard errors (in parenthesis) for different approaches. For each model, the lowest averages regarding each measure are shown in bold.

Table 3.1 reports the averages and corresponding standard errors (in parenthesis) of different loss measures obtained from each method when $p = 30$. From the results it can be seen that, by addressing the order issue, the proposed methods M1 and M2 considerably outperform other approaches with respect to all the loss measures. Overall, the M1 performs the best under Δ_1, Δ_3 and MSE criteria, followed by M2. The M2 produces the minimum MAE in all the six models. It also significantly dominates all the other approaches in terms of FSL except *Model* 3, where the M2 is the second best and inferior to the Glasso. Nevertheless, the M2 substantially outperforms the Glasso in *Model* 3 regarding all the other loss measures. Additionally, although the AVE gives the best performance for the loss function Δ_2, the M1 is much comparable. Particularly, from the perspective of models, the M2 generally gives the superior performance to the other methods in the sparse *Model* 5, and also shows advantage in *Model* 6. Moreover, from the perspective of variation, the proposed methods M1 and M2 result in a much smaller variability of the estimates for all the models in terms of Δ_1, Δ_3 and MSE. The AVE has comparable standard errors regarding Δ_2, and the Glasso gives the smallest standard errors under MAE.

Compared with the proposed methods, the MCD approach based on the BIC order selection (i.e., BIC) does not do as well as M1 and M2, which implies that using a single variable order in the MCD approach is not helpful to improve the estimation accuracy, while the multiple orders would lead to a more accurate estimate. Also, the inferior performance of the AVE to the proposed methods implies that the way of assembling the available estimates obtained from multiple orders is important, i.e., the method (3.6) with the ensemble estimates $\tilde{\boldsymbol{T}}$ and $\tilde{\boldsymbol{D}}$ performs better than the method (3.7) of the ensemble estimate $\bar{\boldsymbol{\Omega}}$.

Table 3.2 and Table 3.3 present the comparison results regarding the loss measures Δ_1,

Table 3.1: The averages and standard errors (in parenthesis) of estimates for p = 30.

		Δ_1	Δ_2	Δ_3	MAE	MSE	FSL (%)
	$\boldsymbol{S}^{-1}$	1.321 (0.041)	0.416 (0.005)	93.442 (5.290)	17.530 (0.628)	19.927 (1.590)	83.973 (0.012)
	M1	**0.177 (0.004)**	0.168 (0.003)	**0.898 (0.099)**	1.667 (0.015)	**0.340 (0.005)**	83.293 (0.066)
	M2	0.278 (0.006)	0.246 (0.004)	2.932 (0.223)	**1.397 (0.013)**	0.460 (0.008)	**7.640 (0.189)**
Model 1	BIC	0.390 (0.016)	0.244 (0.005)	9.411 (0.705)	3.692 (0.293)	1.784 (0.324)	71.489 (0.814)
	BPA	0.274 (0.016)	0.198 (0.004)	5.201 (0.808)	2.347 (0.255)	0.880 (0.350)	54.151 (1.203)
	AVE	0.256 (0.009)	**0.158 (0.003)**	6.483 (0.475)	2.289 (0.103)	0.527 (0.053)	83.764 (0.031)
	Glasso	0.323 (0.008)	0.845 (0.035)	2.031 (0.132)	2.086 (0.011)	0.948 (0.017)	12.862 (0.563)
	$\boldsymbol{S}^{-1}$	1.263 (0.033)	0.412 (0.004)	86.233 (4.158)	16.914 (0.550)	18.685 (1.371)	83.987 (0.008)
	M1	**0.175 (0.003)**	0.167 (0.002)	**0.859 (0.069)**	1.653 (0.010)	**0.335 (0.004)**	83.347 (0.044)
	M2	0.283 (0.005)	0.250 (0.004)	3.004 (0.171)	**1.391 (0.013)**	0.463 (0.008)	**7.502 (0.175)**
Model 2	BIC	0.371 (0.011)	0.239 (0.003)	8.443 (0.539)	3.539 (0.248)	1.654 (0.301)	71.320 (0.888)
	BPA	0.271 (0.007)	0.201 (0.004)	4.723 (0.272)	2.319 (0.097)	0.645 (0.056)	55.391 (1.225)
	AVE	0.248 (0.005)	**0.157 (0.002)**	6.072 (0.300)	2.235 (0.068)	0.490 (0.030)	83.804 (0.031)
	Glasso	0.329 (0.006)	0.862 (0.026)	2.039 (0.098)	2.097 (0.007)	0.967 (0.011)	12.071 (0.247)
	$\boldsymbol{S}^{-1}$	1.289 (0.034)	0.416 (0.004)	89.310 (4.118)	17.212 (0.645)	19.257 (1.754)	86.658 (0.006)
	M1	0.086 (0.002)	0.161 (0.004)	**0.572 (0.056)**	2.024 (0.015)	0.794 (0.009)	84.920 (0.125)
	M2	**0.081 (0.001)**	0.215 (0.003)	0.604 (0.056)	**1.766 (0.007)**	0.848 (0.005)	10.209 (0.070)
Model 3	BIC	0.230 (0.019)	0.156 (0.005)	3.939 (0.653)	4.326 (0.533)	3.099 (0.933)	56.898 (2.788)
	BPA	0.144 (0.006)	0.235 (0.005)	1.388 (0.147)	2.492 (0.076)	1.074 (0.035)	35.267 (1.878)
	AVE	0.111 (0.005)	**0.116 (0.005)**	1.738 (0.164)	2.323 (0.097)	**0.752 (0.054)**	86.013 (0.072)
	Glasso	0.099 (0.002)	0.331 (0.005)	1.904 (0.076)	1.869 (0.004)	0.901 (0.003)	**9.667 (0.109)**
	$\boldsymbol{S}^{-1}$	1.283 (0.035)	0.417 (0.004)	88.84 (4.403)	16.913 (0.571)	18.372 (1.514)	46.676 (0.006)
	M1	**0.136 (0.002)**	0.141 (0.002)	**0.665 (0.063)**	1.886 (0.011)	**0.380 (0.004)**	46.529 (0.068)
	M2	0.171 (0.003)	0.189 (0.004)	1.253 (0.094)	**1.764 (0.011)**	0.478 (0.007)	**44.382 (0.219)**
Model 4	BIC	0.301 (0.014)	0.202 (0.004)	5.925 (0.609)	3.341 (0.197)	1.409 (0.184)	45.649 (0.309)
	BPA	0.210 (0.006)	0.174 (0.003)	2.838 (0.235)	2.294 (0.056)	0.582 (0.029)	44.818 (0.479)
	AVE	0.184 (0.006)	**0.133 (0.002)**	3.683 (0.267)	2.149 (0.048)	0.429 (0.021)	46.671 (0.044)
	Glasso	0.203 (0.003)	0.467 (0.012)	1.707 (0.072)	2.279 (0.007)	0.801 (0.008)	45.173 (0.168)
	$\boldsymbol{S}^{-1}$	1.313 (0.036)	0.419 (0.004)	92.325 (4.667)	1.644 (0.048)	0.388 (0.028)	96.502 (0.029)
	M1	0.047 (0.002)	0.038 (0.001)	0.546 (0.066)	0.070 (0.003)	0.007 (0.001)	70.636 (0.755)
	M2	**0.033 (0.001)**	**0.027 (0.001)**	**0.481 (0.057)**	**0.436 (0.072)**	**0.006 (0.001)**	**3.920 (0.651)**
Model 5	BIC	0.093 (0.009)	0.061 (0.003)	1.337 (0.227)	0.097 (0.010)	0.015 (0.003)	27.422 (2.092)
	BPA	0.082 (0.004)	0.057 (0.002)	0.992 (0.116)	0.129 (0.006)	0.014 (0.002)	21.680 (1.690)
	AVE	0.066 (0.005)	0.047 (0.002)	1.186 (0.164)	0.096 (0.005)	0.009 (0.001)	79.827 (0.838)
	Glasso	0.095 (0.002)	0.183 (0.005)	1.645 (0.079)	0.070 (0.000)	0.027 (0.000)	8.240 (0.344)
	Sample	1.276 (0.031)	0.413 (0.004)	87.703 (3.726)	27.212 (0.827)	47.563 (2.938)	90.218 (0.004)
	M1	**0.129 (0.002)**	0.157 (0.003)	**0.357 (0.048)**	1.793 (0.012)	0.619 (0.011)	89.578 (0.046)
	M2	0.230 (0.004)	0.188 (0.003)	0.404 (0.037)	**1.366 (0.014)**	**0.579 (0.011)**	**10.071 (0.294)**
Model 6	BIC	0.261 (0.015)	0.163 (0.004)	5.583 (0.611)	3.572 (0.360)	2.798 (0.674)	58.284 (1.745)
	BPA	0.171 (0.009)	0.110 (0.004)	2.663 (0.264)	2.167 (0.171)	0.974 (0.162)	42.089 (1.527)
	AVE	0.162 (0.006)	**0.101 (0.002)**	3.744 (0.284)	2.203 (0.086)	0.698 (0.054)	89.920 (0.038)
	Glasso	0.138 (0.002)	0.189 (0.005)	0.837 (0.087)	1.738 (0.016)	0.708 (0.025)	29.289 (0.555)

Table 3.2: The averages and standard errors (in parenthesis) of estimates for p = 50.

		Δ_1	Δ_2	Δ_3	MAE	MSE	FLS (%)
Model 1	M1	**0.218 (0.003)**	**0.198 (0.002)**	**2.780 (0.148)**	1.894 (0.010)	**0.392 (0.004)**	89.043 (0.064)
	M2	0.316 (0.004)	0.274 (0.004)	6.779 (0.259)	**1.483 (0.010)**	0.509 (0.006)	**5.115 (0.097)**
	BIC	0.986 (0.101)	0.332 (0.006)	100.669 (22.202)	11.774 (1.775)	35.285 (13.292)	76.197 (0.827)
	BPA	0.383 (0.012)	0.249 (0.004)	15.434 (0.987)	3.172 (0.169)	1.155 (0.190)	53.430 (0.970)
	AVE	0.436 (0.017)	0.209 (0.003)	28.162 (1.925)	4.268 (0.272)	1.556 (0.246)	90.024 (0.024)
	Glasso	0.363 (0.003)	1.023 (0.016)	4.090 (0.126)	2.170 (0.004)	1.037 (0.005)	7.232 (0.055)
Model 2	M1	**0.217 (0.002)**	**0.202 (0.002)**	**2.502 (0.135)**	1.882 (0.009)	**0.404 (0.004)**	88.944 (0.052)
	M2	0.321 (0.003)	0.285 (0.003)	6.527 (0.261)	**1.516 (0.009)**	0.529 (0.005)	**5.160 (0.087)**
	BIC	0.814 (0.112)	0.321 (0.005)	83.611 (33.313)	9.104 (2.227)	35.013 (25.822)	74.122 (0.943)
	BPA	0.352 (0.012)	0.247 (0.004)	12.129 (0.903)	2.850 (0.173)	0.949 (0.165)	49.699 (1.079)
	AVE	0.407 (0.018)	0.203 (0.003)	24.151 (1.766)	4.203 (0.379)	1.685 (0.400)	89.984 (0.029)
	Glasso	0.360 (0.002)	0.998 (0.011)	3.896 (0.093)	2.164 (0.003)	1.031 (0.004)	7.181 (0.060)
Model 3	M1	0.075 (0.001)	0.134 (0.002)	1.146 (0.079)	1.538 (0.013)	0.563 (0.005)	88.010 (0.299)
	M2	**0.065 (0.001)**	0.142 (0.002)	**1.078 (0.075)**	**1.164 (0.005)**	**0.551 (0.004)**	**3.888 (0.032)**
	BIC	0.709 (0.284)	0.160 (0.005)	240.564 (204.822)	12.454 (5.710)	143.518 (122.060)	45.938 (2.656)
	BPA	0.141 (0.006)	0.174 (0.003)	2.889 (0.253)	2.049 (0.063)	0.856 (0.031)	24.037 (1.070)
	AVE	0.174 (0.024)	**0.133 (0.003)**	7.471 (1.964)	3.515 (0.563)	2.121 (0.808)	93.526 (0.102)
	Glasso	0.083 (0.001)	0.230 (0.003)	3.490 (0.090)	1.240 (0.003)	0.577 (0.002)	3.904 (0.053)
Model 4	M1	**0.161 (0.002)**	**0.170 (0.002)**	**1.698 (0.120)**	2.143 (0.009)	**0.448 (0.004)**	64.654 (0.075)
	M2	0.193 (0.002)	0.216 (0.003)	2.745 (0.157)	**1.888 (0.008)**	0.530 (0.005)	**29.626 (0.079)**
	BIC	0.651 (0.104)	0.256 (0.005)	61.784 (30.722)	8.095 (1.927)	24.901 (18.629)	55.331 (0.592)
	BPA	0.285 (0.017)	0.216 (0.003)	8.544 (1.462)	3.206 (0.387)	1.621 (0.849)	43.960 (0.553)
	AVE	0.283 (0.012)	0.173 (0.002)	13.04 (1.058)	3.468 (0.212)	0.994 (0.157)	65.379 (0.033)
	Glasso	0.233 (0.002)	0.581 (0.010)	3.520 (0.093)	2.400 (0.004)	0.882 (0.005)	30.557 (0.070)
Model 5	M1	0.047 (0.002)	0.038 (0.001)	1.035 (0.077)	0.049 (0.001)	**0.003 (0.000)**	46.182 (0.746)
	M2	**0.034 (0.001)**	**0.027 (0.001)**	**0.922 (0.071)**	**0.021 (0.001)**	**0.003 (0.000)**	**0.224 (0.038)**
	BIC	0.280 (0.076)	0.079 (0.004)	22.586 (12.778)	0.258 (0.094)	0.093 (0.051)	31.037 (2.411)
	BPA	0.107 (0.009)	0.064 (0.003)	2.622 (0.389)	0.126 (0.016)	0.016 (0.007)	19.256 (1.214)
	AVE	0.087 (0.006)	0.055 (0.002)	3.110 (0.255)	0.096 (0.008)	0.006 (0.001)	73.165 (1.283)
	Glasso	0.109 (0.001)	0.239 (0.005)	2.736 (0.119)	0.053 (0.000)	0.020 (0.000)	7.973 (0.338)
Model 6	M1	**0.148 (0.002)**	0.163 (0.002)	**0.253 (0.041)**	1.882 (0.008)	0.626 (0.007)	92.797 (0.044)
	M2	0.267 (0.004)	0.196 (0.002)	0.691 (0.081)	**1.374 (0.009)**	**0.587 (0.007)**	**6.059 (0.152)**
	BIC	0.552 (0.163)	0.202 (0.004)	93.435 (75.770)	9.558 (4.628)	208.371 (203.098)	57.403 (1.252)
	BPA	0.223 (0.008)	0.131 (0.003)	6.909 (0.478)	2.561 (0.116)	1.069 (0.084)	39.218 (1.120)
	AVE	0.243 (0.012)	**0.123 (0.002)**	12.312 (0.978)	3.611 (0.373)	1.964 (0.563)	93.768 (0.029)
	Glasso	0.196 (0.003)	0.306 (0.009)	2.832 (0.181)	2.085 (0.012)	1.094 (0.025)	21.902 (0.360)

Table 3.3: The averages and standard errors (in parenthesis) of estimates for p = 100.

		Δ_1	Δ_2	Δ_3	MAE	MSE	FLS (%)
Model 1	M1	**0.275 (0.002)**	**0.248 (0.002)**	**9.296 (0.354)**	2.180 (0.009)	**0.484 (0.003)**	92.118 (0.077)
	M2	0.360 (0.003)	0.319 (0.002)	17.52 (0.474)	**1.628 (0.005)**	0.588 (0.003)	**2.991 (0.034)**
	BIC	5.549 (0.618)	0.473 (0.007)	5463.851 (1019.529)	67.248 (7.801)	1610.774 (308.144)	73.237 (0.668)
	BPA	1.043 (0.200)	0.343 (0.004)	349.150 (162.716)	10.927 (2.649)	100.857 (57.985)	47.570 (0.994)
	AVE	1.587 (0.070)	0.372 (0.006)	499.315 (37.701)	15.332 (0.699)	18.726 (2.345)	94.940 (0.009)
	Glasso	0.392 (0.002)	1.171 (0.012)	9.690 (0.205)	2.222 (0.002)	1.086 (0.003)	3.750 (0.016)
Model 2	M1	**0.271 (0.002)**	**0.250 (0.002)**	**8.470 (0.273)**	2.180 (0.009)	**0.487 (0.003)**	92.128 (0.075)
	M2	0.354 (0.002)	0.320 (0.002)	16.103 (0.372)	**1.630 (0.005)**	0.588 (0.004)	**2.998 (0.029)**
	BIC	5.238 (0.610)	0.482 (0.007)	5015.974 (1062.565)	58.675 (6.936)	1226.585 (248.565)	74.513 (0.642)
	BPA	0.795 (0.107)	0.340 (0.005)	153.276 (51.777)	7.778 (1.569)	31.015 (16.681)	46.890 (1.020)
	AVE	1.668 (0.073)	0.374 (0.007)	536.125 (39.379)	16.182 (0.702)	22.304 (2.673)	94.924 (0.010)
	Glasso	0.389 (0.002)	1.157 (0.012)	9.672 (0.200)	2.219 (0.002)	1.083 (0.003)	3.732 (0.016)
Model 3	M1	0.071 (0.001)	0.093 (0.001)	3.194 (0.130)	1.274 (0.012)	0.3778 (0.004)	85.552 (0.443)
	M2	**0.058 (0.001)**	**0.089 (0.001)**	**2.938 (0.122)**	**0.739 (0.005)**	**0.344 (0.003)**	**1.162 (0.019)**
	BIC	1.559 (0.414)	0.162 (0.004)	1161.856 (643.166)	24.432 (7.157)	710.784 (395.258)	40.689 (1.361)
	BPA	0.266 (0.027)	0.150 (0.003)	21.652 (4.107)	3.545 (0.429)	4.181 (1.567)	22.407 (1.081)
	AVE	0.976 (0.087)	0.201 (0.006)	223.731 (33.502)	16.309 (1.326)	35.368 (6.228)	97.729 (0.036)
	Glasso	0.078 (0.001)	0.166 (0.002)	8.309 (0.151)	0.781 (0.002)	0.352 (0.002)	1.212 (0.021)
Model 4	M1	**0.195 (0.002)**	**0.205 (0.001)**	**5.732 (0.270)**	2.415 (0.009)	**0.517 (0.003)**	77.789 (0.155)
	M2	0.224 (0.002)	0.247 (0.002)	8.074 (0.330)	**2.006 (0.005)**	0.584 (0.003)	**16.170 (0.027)**
	BIC	2.949 (0.465)	0.352 (0.007)	2143.118 (608.875)	37.093 (6.194)	772.251 (225.959)	57.449 (0.865)
	BPA	0.449 (0.029)	0.276 (0.004)	38.051 (5.745)	4.646 (0.380)	3.652 (1.510)	38.676 (0.590)
	AVE	1.105 (0.058)	0.301 (0.005)	261.304 (23.585)	12.519 (0.617)	12.556 (1.557)	81.760 (0.013)
	Glasso	0.259 (0.002)	0.690 (0.007)	8.239 (0.148)	2.489 (0.003)	0.943 (0.003)	16.559 (0.018)
Model 5	M1	0.054 (0.002)	0.041 (0.001)	2.952 (0.183)	0.032 (0.001)	**0.002 (0.000)**	22.498 (0.454)
	M2	**0.040 (0.001)**	**0.031 (0.001)**	**2.655 (0.163)**	**0.014 (0.001)**	**0.002 (0.000)**	**1.216 (0.481)**
	BIC	2.065 (0.567)	0.121 (0.006)	2099.179 (963.162)	0.983 (0.298)	2.124 (1.075)	31.034 (1.683)
	BPA	0.230 (0.020)	0.088 (0.003)	16.510 (2.034)	0.147 (0.013)	0.011 (0.002)	18.975 (0.780)
	AVE	0.779 (0.077)	0.147 (0.006)	159.517 (28.408)	0.504 (0.049)	0.182 (0.051)	82.719 (1.102)
	Glasso	0.119 (0.001)	0.274 (0.002)	2.798 (0.116)	0.037 (0.000)	0.012 (0.000)	9.786 (0.172)
Model 6	M1	**0.177 (0.002)**	0.171 (0.001)	**0.128 (0.020)**	1.982 (0.007)	0.629 (0.006)	93.839 (0.073)
	M2	0.301 (0.003)	0.203 (0.002)	2.879 (0.199)	**1.386 (0.007)**	**0.593 (0.006)**	**3.242 (0.062)**
	BIC	0.847 (0.068)	0.261 (0.005)	144.316 (26.191)	10.432 (1.044)	31.858 (6.171)	52.967 (0.897)
	BPA	0.397 (0.035)	0.162 (0.004)	40.191 (7.931)	4.760 (0.619)	6.980 (2.771)	34.222 (0.806)
	AVE	0.378 (0.011)	**0.161 (0.003)**	52.509 (2.354)	4.454 (0.138)	1.921 (0.129)	96.417 (0.025)
	Glasso	0.313 (0.004)	0.617 (0.012)	12.965 (0.368)	2.475 (0.007)	1.743 (0.019)	12.325 (0.172)

Δ_2, Δ_3, MAE, MSE and FSL for $p = 50$ and $p = 100$, respectively. The inverse sample covariance $\boldsymbol{S}^{-1}$ is excluded since it is singular and hence not a legitimate estimator for $\boldsymbol{\Omega}$. Tables show the similar conclusions as $p = 30$. The proposed methods generally give superior performances to the other approaches. As the number of variables p increases, the proposed methods work even more promising as expected. For example, the M2 performs better and better for *Model* 3 as the number of variables p increases, since this model is becoming more and more sparse. Compared with AVE, the proposed methods result in much smaller losses and standard errors in terms of Δ_2 when $p = 100$. In addition, the M1 performs the best in the dense *Model* 4 in all the settings of p, since the M1 is able to give an accurate estimate when the true model is not sparse.

Moreover, to investigate the impact of the choice of the number of orders M on the performance of the proposed methods, we compute six loss measures for different values of $M = 10, 30, 50, 80, 100, 120$ and 150. Figure 3.2 and 3.3 display the corresponding results obtained from the proposed methods M1 and M2 for *Model* 2. The solid line, dashed line and dotted line represent three situations where $p = 30, 50$ and 100, respectively. Overall, it is clear to see that almost all of the lines are very stable over different values of M except that they are decreasing in the range of $M = (10, 30)$. This indicates that the performance of the proposed methods are quite comparable when M is larger than 30.

In a brief summary, the numerical results show that the proposed methods give a superior performance over other conventional approaches. The M2 provides accurate estimate of $\boldsymbol{\Omega}$ and catches the underlying sparse structure of the inverse covariance matrix. While for the method M1, we can see that it also gives reasonable estimation accuracy, especially when the true $\boldsymbol{\Omega}$ is not sparse. The simulation study also suggests that a proper choice of the value for M should be larger than 30.

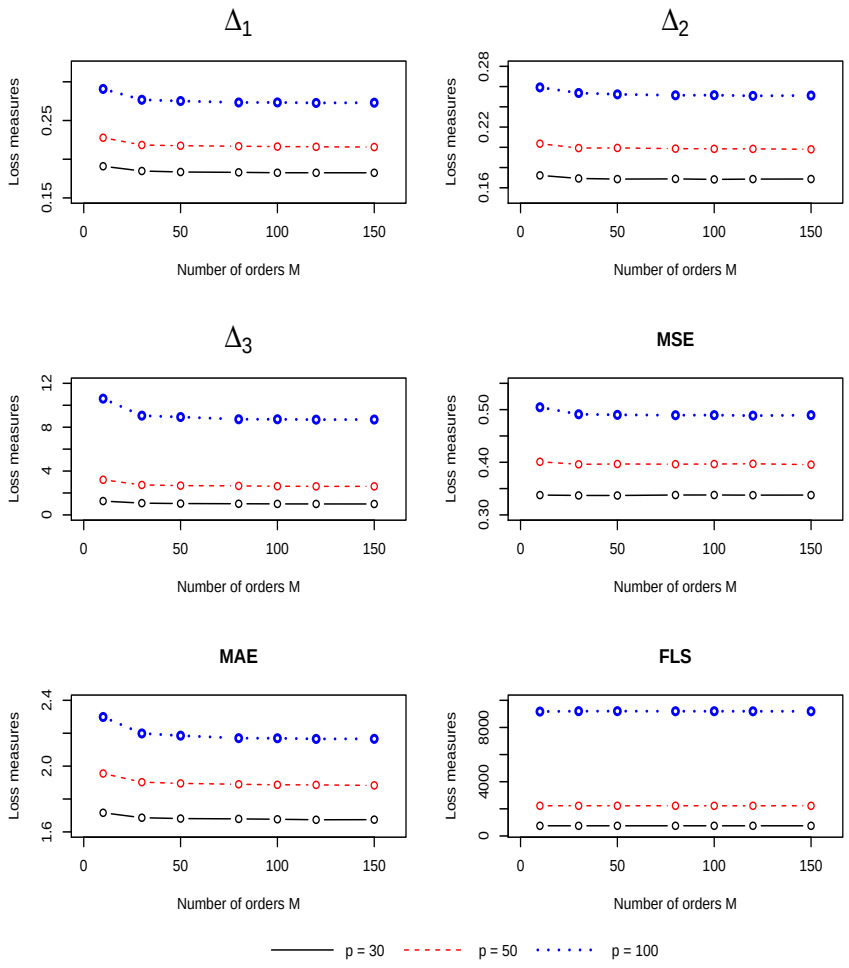

Figure 3.2: Plot of six loss measures of the proposed M1 against the number of orders M for $Model$ 2

3.5 Application

In this section, we apply the proposed method of estimating $\mathbf{\Omega}$ for the linear discriminant analysis (LDA). To overcome the drawback of the classic LDA in high-dimensional data, we consider a new classification rule by using the proposed sparse inverse covariance estimate. A sonar data set, a gene expression data set and hand movement data are used to evaluate the performance of the proposed classification rule.

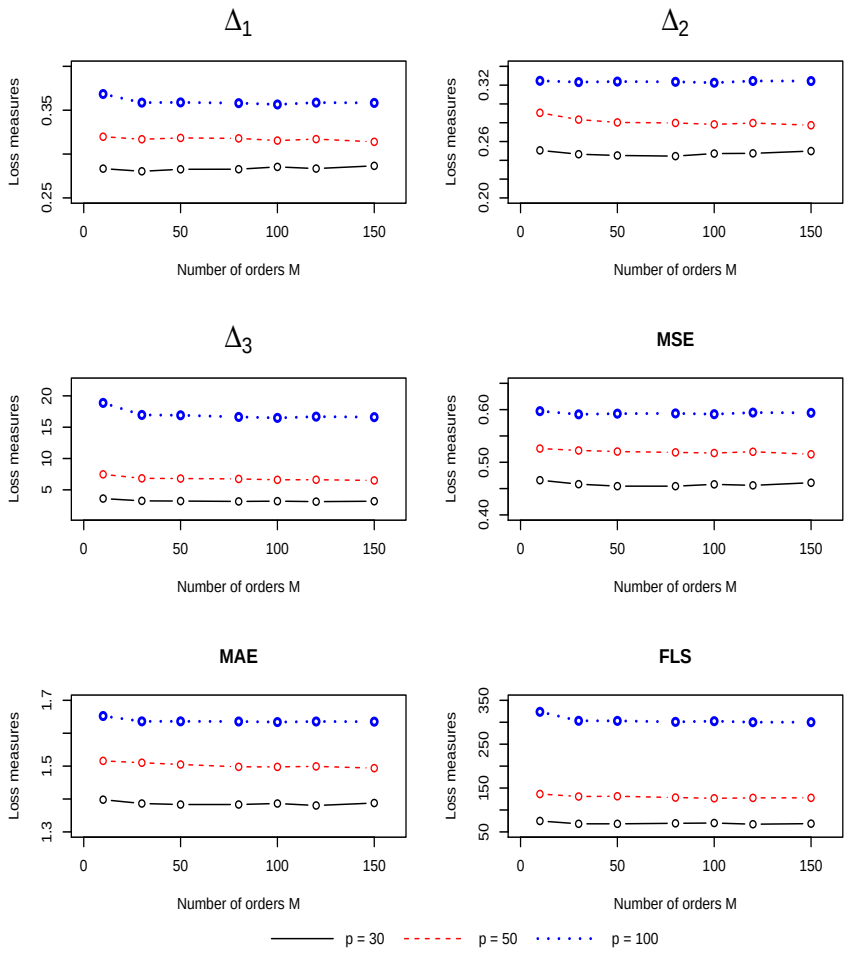

Figure 3.3: Plot of six loss measures of the proposed M2 against the number of orders M for *Model* 2

3.5.1 LDA via the Proposed Estimate of Ω

In the classification problem, LDA is one commonly used technique. Consider a classification problem with K classes. Each observation belongs to some class $k \in 1, 2, \ldots, K$. Denote by C_k the class of training set observation $\boldsymbol{x}_i$. Let $\hat{\boldsymbol{\mu}}_k$ be the $p \times 1$ vector of the sample mean of the training data in class k, and $\hat{\boldsymbol{\Sigma}}_{LDA} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'$ be the estimated within-class covariance matrix based on the training data. Then LDA classification rule is:

classify a test observation $\boldsymbol{x}$ to class k^* if $k^* = \arg\max_k \eta_k(\boldsymbol{x})$, where

$$\eta_k(\boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{LDA}^{-1}\hat{\boldsymbol{\mu}}_k - \frac{1}{2}\hat{\boldsymbol{\mu}}_k'\hat{\boldsymbol{\Sigma}}_{LDA}^{-1}\hat{\boldsymbol{\mu}}_k + \log\pi_k \tag{3.11}$$

and π_k is the frequency of class k in the training data set. This method works well if the training sample size n is larger than the number of random variables p. However, when p is close to n, Bickel and Levina (2004) showed that LDA is asymptotically as bad as random guessing. Even worse, when $n < p$, the within-class covariance matrix $\hat{\boldsymbol{\Sigma}}_{LDA}$ is singular and the classical LDA breaks down. There are different approaches developed to address these problems, such as Friedman (1989), Howland and Park (2004), Guo, Hastie and Tibshirani (2007), Fan and Fan (2008), and Shao et al. (2011).

To overcome the singular issue, we suggest a classification rule using the proposed sparse estimate instead of $\hat{\boldsymbol{\Sigma}}_{LDA}^{-1}$ in (3.11). An accurate estimation of inverse within-class covariance matrix is expected to lead to accurate classification performance. Moreover, since the proposed ensemble estimate depends on different permutations, it enables us to evaluate the uncertainty for the misclassification rate. Specifically, We first randomly select M permutations from all the possible permutations and obtain the misclassification rate e_1. Then randomly select M permutations again from all the possible permutations and obtain the misclassification rate e_2. Repeat the above procedure until e_m is obtained, such that the confidence interval can be constructed based on $e_1, e_2, \ldots, e_m$. We set $m = 100$ throughout this paper.

In the following subsections, three real classification data sets are used to evaluate the performance of the proposed estimate, obtained respectively from M1 and M2, in comparison with other approaches, including BIC, BPA, AVE and Glasso. Apart from these, the generalized LDA (Howland and Park, 2004), C5.0 (Quinlan, 1993) and diagonal linear discriminant analysis (Dudoit, Fridlyand and Speed, 2002) are also considered, denoted by GLDA, C5 and DLDA. The GLDA replaces $\boldsymbol{\Sigma}_{LDA}^{-1}$ in (3.11) with the generalized inverse covariance matrix. C5 builds decision trees from a set of training data, using the concept of entropy. On each iteration of the algorithm, it iterates through every unused variable and calculates the entropy.

It then selects the variable which has the smallest entropy value. The DLDA is a modification to LDA, where the off-diagonal elements of the pooled sample covariance matrix are set to be zeroes.

3.5.2 Sonar Data

This classification data set was obtained by bouncing sonar signals off metal cylinders and roughly cylindrical rocks from a variety of different angles. Data are available online at http://sci2s.ugr.es/keel/dataset.php?cod=85. It contains two classes with 208 samples and 60 variables, 97 of which are rocks, while the other 121 samples are metals. Each number represents the energy within a particular frequency band, integrated over a certain period of time. We randomly split the samples into two groups: training set of 80 samples and testing set of 128 samples. Table 3.4 shows the misclassification rate of different methods and the corresponding 95% confidence interval if available. Overall, the M1 performs the best with the smallest misclassification rate, as well as the lowest and narrowest confidence interval. In addition, the BIC has a larger misclassification rate than AVE, and gives a wider confidence interval, which confirms that using multiple orders is more efficient than one single order in the MCD approach.

Table 3.4: Misclassification rate (in percentage) comparison for proposed methods with other approaches from Sonar data.

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	17.2	25	25	20.3	26.6	25.8	32.8	26.6	31.3
95% C.I.	(16.4, 18.8)	(23.8, 27.7)	(17.6, 27.3)	(18.8, 21.9)	–	–	–	–	–

Furthermore, we randomly partition 80 observations of the samples as a new training data set and the remaining 128 observations as a new testing data set. Then the misclassification rate is calculated based on these new training and testing data sets under each method. Table 3.5 displays the averages and corresponding standard errors (SE) of the misclassification rate obtained from each approach for this procedure over 50 times. The M1 still works the best,

Table 3.5: Misclassification rate (in percentage) comparison for proposed methods with other approaches under the randomly splitting training and testing data from Sonar data

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	25.3	31.8	28.1	26.5	27.7	29.4	34.5	30.8	31.3
SD	0.5	0.8	0.6	0.5	0.5	0.7	0.8	0.7	0.7

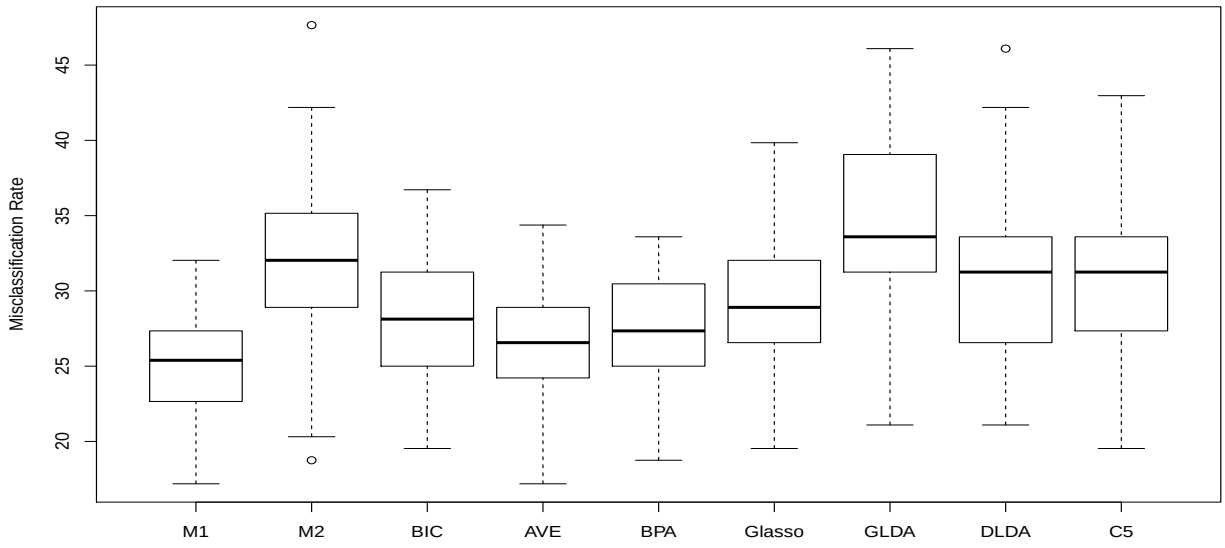

Figure 3.4: Boxplot of misclassification rate (in percentage) comparison for proposed methods with other approaches under the randomly splitting training and testing data from Sonar data

followed by the AVE, which confirms that the way of manipulating the available estimates under multiple orders is important. Compared with the BIC, the AVE has a smaller variability. This is to be expected since the AVE takes advantage of multiple orders over only one single order. Additionally, the M2 and DLDA do not appear to be as good as the M1 and AVE, possibly because the true underlying inverse covariance matrix in this example is not sparse. The Glasso is comparable with BPA, C5 and DLDA. The GLDA does not provide accurate classification. The same results of boxplot are displayed in Figure 3.4.

3.5.3 Lymphoma Data

The second classification data set include two classes. It contains expression values for 2647 genetic probes and 77 samples, 58 of which are obtained from patients suffering from diffuse large B-cell lymphoma, while the remaining 19 samples are derived from follicular lymphoma type. Data are available online at http://ico2s.org/datasets/microarray.html. We randomly split the samples into two groups: training set of 35 samples and testing set of 42 samples. Then the variable screening procedure is performed through two sample t-test. Specifically, for each variable, t-test is conducted against the two classes of the training data such that variables with large values of test statistics are ranked as significant variables, and the top 50 significant variables are selected for data classification. The results of misclassification rate and corresponding confidence interval are summarized in Table 3.6. Overall, the proposed methods are better than other approaches. The M2 performs the best with the minimum misclassification rate and lowest confidence interval. Additionally, the M1 and M2 provide much smaller confidence intervals than the BIC and AVE. Moreover, the AVE gives superior performance to the BIC in term of smaller misclassification rate and narrower confidence interval as expected. The BIC and GLDA do not give the accurate classification.

Table 3.6: Misclassification rate (in percentage) of the proposed methods compared with other approaches for Lymphoma data.

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	16.7	14.3	38.1	26.2	21.4	19.0	30.1	16.7	21.4
95% C.I.	(14.3, 16.7)	(11.9, 16.7)	(15.5, 41.7)	(20.2, 35.7)	–	–	–	–	–

Furthermore, we randomly partition 35 observations of the samples as a new training data set and the remaining 42 observations as a new testing data set. Table 3.7 shows the averages and corresponding standard errors (SE) of the misclassification rate by repeating the above procedure over 50 times based on the top 50 significant gene expressions. It is clear that the M1, M2 and DLDA are the best, followed by C5, Glasso and AVE, which further confirms that an efficient way of organizing the available estimates will lead to a small misclassification

Table 3.7: Misclassification rate (in percentage) comparison for proposed methods with other approaches under the randomly splitting training and testing data from Lymphoma data

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	14.0	14.1	30.9	24.3	27.2	17.2	28.0	14.8	19.7
SD	0.7	0.7	1.4	1.1	1.4	0.9	1.1	0.8	1.1

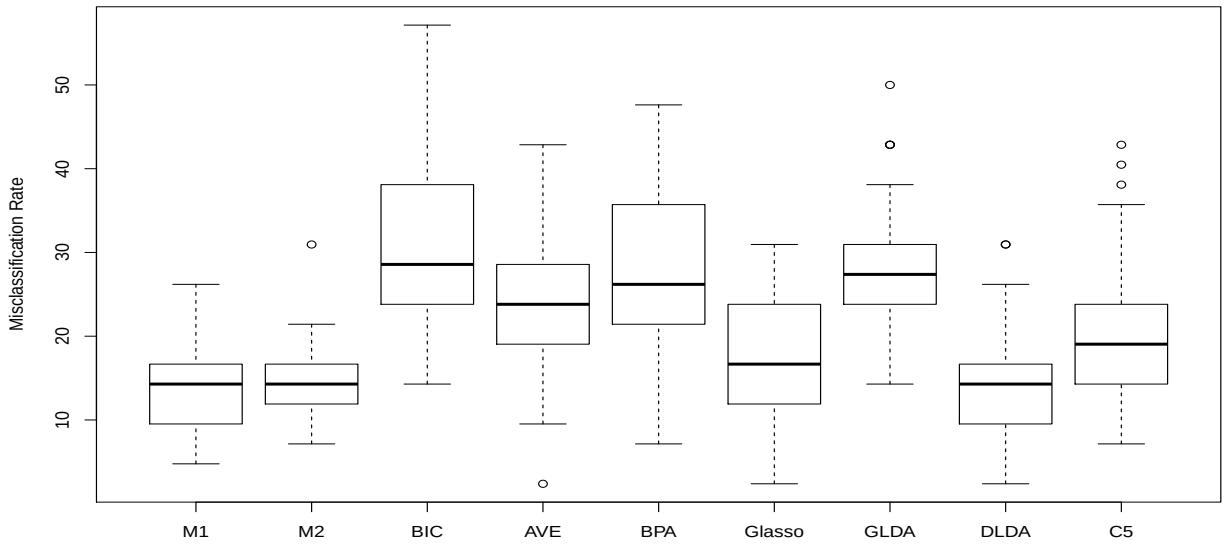

Figure 3.5: Boxplot of misclassification rate (in percentage) comparison for proposed methods with other approaches under the randomly splitting training and testing data from Lymphoma data

rate. In this example, both the M2 and DLDA perform quite well due to the conditional independence existing in large numbers among the gene expressions. M2 appears to be better because of its smaller misclassification rate and standard error. In addition, the AVE performs better than BIC due to the superiority of the multiple orders over one order. The BIC, BPA and GLDA are not as good as other approaches. Figure 3.5 presents the boxplot of the same results.

Although DLDA shows comparable performance as M2 in the lymphoma example, we will demonstrate its drawbacks in the following. The reason DLDA works well, we believe, is that

the contribution of every single variable of top 50 is relatively much more significant than the contribution of their interactions. As a result, the underlying inverse covariance matrix might be a diagonal matrix. To confirm our opinion, we assess the performance of each method using lymphoma data set based on a new group of 50 variables, which are randomly selected from all the 2647 gene expressions. Obviously, a single variable from these randomly selected 50 variables would not play a role as great as the top 50 significant variables. Therefore, their interactions are supposed to make some contribution, hence resulting in a sparse but not a diagonal inverse covariance matrix. In practice, we use the same partitioned training and testing data sets used for Figure 3.5, and randomly select 50 gene expressions for each training set. Table 3.8 reports the averages and standard errors (SE) of the misclassification rate of each method from the above procedure. It is clear that the M2 performs much better than DLDA. The M1 still gives a good performance due to its accurate estiamte. The corresponding visualization is presented in Figure 3.6.

Table 3.8: Misclassification rate (in percentage) comparison for proposed methods with other approaches under randomly selected 50 gene expressions from Lymphoma data

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	13.1	13.8	28.1	20.6	26.1	20.0	24.5	21.2	25.0
SD	0.7	0.8	1.4	1.2	1.8	1.2	1.3	1.0	1.2

3.5.4 Hand Movement Data

To evaluate the performance of the proposed methods in multiple classification problems, the third data set contains 15 classes of 24 observations each with each class referring to a hand movement type. The hand movement is represented as a two dimensional curve performed by the hand in a period of time, where each curve is mapped in a representation with 90 variables. The data are available online at https://archive.ics.uci.edu/ml/datasets/Libra+Movement. The data set is randomly split into the training set of 160 observations and testing set of 200 observations. Table 3.9 reports the misclassification rate and corresponding confidence

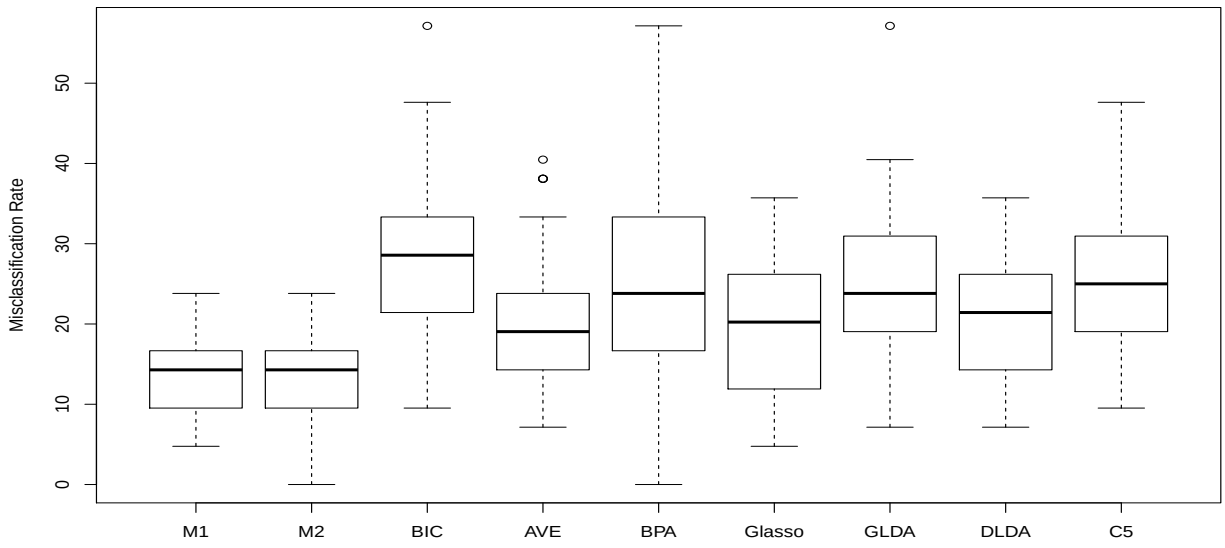

Figure 3.6: Boxplot of misclassification rate (in percentage) comparison for proposed methods with other approaches under randomly selected 50 gene expressions from Lymphoma data

interval if available for each approach. The proposed M1 dominates all the other methods attributed to the accurate inverse covariance matrix estimate. Although BIC produces the same misclassification rate as AVE, it has a wider confidence interval. M2, especially DLDA, performs not well possibly due to the non-sparse structure of the underlying inverse covariance matrix. BPA and Glasso are comparable with BIC and AVE. GLDA does not provide a accurate classification.

Table 3.9: Misclassification rate (in percentage) of the proposed methods compared with other approaches for Hand Movement data.

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	27.5	37.5	37.0	37.0	38.5	36.5	48.5	42.0	42.5
95% C.I.	(26.5, 28.5)	(34.5, 40.5)	(35.5, 39.0)	(37.0, 37.5)	–	–	–	–	–

Moreover, we randomly partition 160 observations as a new training data set and the remaining 200 observations as a new testing data set. Table 3.10 presents the averages and standard errors (SE) of the misclassification rate by repeating the above procedure over 50

Table 3.10: Misclassification rate (in percentage) comparison for proposed methods with other approaches under the randomly splitting training and testing data from Hand Movement data.

Method	M1	M2	BIC	AVE	BPA	Glasso	GLDA	DLDA	C5
Error	33.0	43.2	40.1	38.9	39.0	39.1	51.0	46.3	46.2
SD	0.5	0.6	0.5	0.5	0.5	0.6	0.5	0.6	0.5

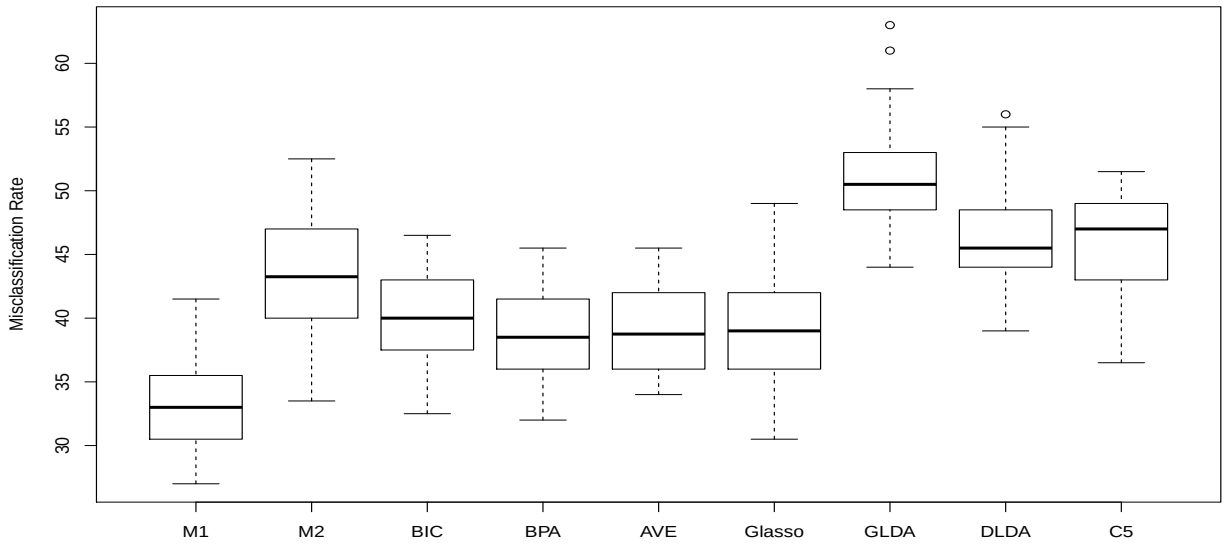

Figure 3.7: Boxplot of misclassification rate (in percentage) comparison for proposed methods with other approaches under the randomly splitting training and testing data from Hand Movement data.

times. Figure 3.7 displays the corresponding boxplot. The proposed M1 outperforms the other approaches. BIC, BPA, AVE and Glasso are comparable with each other. The performances of M2, C5 and DLDA are not as well as others. GLDA gives the highest misclassification rate. This 15 classes data example demonstrates that the proposed method works consistently well in the multiple classification settings.

3.6 Discussion

In this paper, we have introduced an order-invariant ensemble method for the high-dimensional sparse inverse covariance matrix estimation. Based on the modified Cholesky decomposition of an inverse covariance matrix, the proposed estimator is properly assembled from a set of multiple estimates of $\boldsymbol{T}$ and $\boldsymbol{D}$ under different orders of random variables. Hard thresholding technique is applied to the ensemble estimate of Cholesky factor matrix $\boldsymbol{T}$ to encourage the sparse structure. The resulting estimator does not require the prior knowledge of the order of variables, and reasonably works well in high-dimensional cases. The simulation studies show the superior performance of our proposed method in terms of loss measures and ability of capturing sparsity. The advantage of considering multiple orders over one single order is illustrated by comparison of the proposed method with the BIC approach. We have also extended this framework to classifications. The performance of the resulting method is demonstrated through the analysis of three classification data examples.

Finally, we would like to remark that sometimes real data may include abnormal observations. Hence, robustness is a very important property we need to consider when proposing an estimator. Compared to the method (3.6), alternatively, we consider the ensemble estimate by the element-wise median of $\hat{\boldsymbol{T}}$ and $\hat{\boldsymbol{D}}$ instead of taking average

$$\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{T}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{T}} \quad \text{with} \quad \hat{\boldsymbol{T}} = \text{med}(\hat{\boldsymbol{T}}_k),\ \hat{\boldsymbol{D}} = \text{med}(\hat{\boldsymbol{D}}_k).$$

This estimator is supposed to be more robust than the proposed method. It is also able to provide an order-invariant sparse estimate for the inverse covariance matrix, and is applicable in high dimensions. We will further investigate the robustness of this estimator in the future work.

Chapter 4 Positive Definite Sparse Ensemble Estimation of High-dimensional Covariance Matrix

4.1 Introduction

As the data collection advances, high-dimensional data are widely occurred in many scientific areas, such as bioinformatics, imaging recognition, weather forecasting, and financial service. Estimation of large covariance matrix or its inverse from the high-dimensional data is thus an important and challenge problem in the multivariate data analysis. For example, dimension reduction using principal component analysis usually relies on accurate estimation of covariance matrix. Under the context of graphical models, the estimation of covariance matrix or its inverse is often used to infer the network structure of the graph. However, conventional estimation of covariance matrix is known to perform poorly due to the high dimension problems when the number of variables is close to or larger than the sample size (Johnstone, 2001). To overcome this curse of dimensionality, a variety of methods proposed in literature assumes some pattern of sparsity for the covariance and its inverse matrices.

In this work, our focus is on the estimation of sparse covariance matrix for high-dimensional data. Early work on covariance matrix estimation includes shrinking eigenvalues of the sample covariance matrix (Dey and Srinivasan, 1985; Haff, 1991), a linear combination of the sample covariance and a proper diagonal matrix (Ledoit and Wolf, 2004), improving the estimation based on matrix condition number (Aubry et al., 2012; Won et al., 2013), and regularizing the eigenvectors of the matrix logarithm of the covariance matrix (Deng and Tsui, 2013). However, the above mentioned methods do not explore the sparse structure of the covariance matrix. A sparse covariance matrix estimate can be useful for subsequent inference, such as inferring the correlation pattern among the variables. Bickel and Levina (2008) proposed to threshold the small entries of the sample covariance matrix to zeroes and studied its theoretical

behavior when the number of variables is large. Rothman, Levina, and Zhu (2009) considered to threshold the sample covariance matrix with more general thresholding functions. Cai and Yuan (2012) proposed a covariance matrix estimation through block thresholding. Their estimator is constructed by dividing the sample covariance matrix into blocks and then simultaneously estimating the entries in a block by thresholding. However, the threshold-based estimator is not guaranteed to be positive definite. To make the estimate being sparse and positive-definite, Bien and Tibshirani (2011) considered a penalized likelihood method with a Lasso penalty (Tibshirani, 1996) on the entries of the covariance matrix. Their idea is similar to the graphical lasso for inverse covariance matrix estimation in the literature (Yuan and Lin, 2007; Friedman, Hastie, and Tibshirani, 2008; Rocha, Zhao, and Yu, 2008; Rothman et al., 2008; Yuan, 2008; Deng and Yuan, 2009; and Yuan, 2010), but the computation is much more complicated due to the non-convexity of the objective function. Xue, Ma and Zou (2012) developed a sparse covariance matrix estimator for high-dimensional data based on a convex objective function with positive definite constraint and L_1 penalty. They also derived a fast algorithm to solve the constraint optimization problem.

Another direction of sparse covariance matrix estimation is to take advantage of matrix decomposition. One popular and effective decomposition is the modified Cholesky decomposition (MCD) (Pourahmadi, 1999; Wu and Pourahmadi, 2003; Pourahmadi, Daniels and Park, 2007; Rothman, Levina and Zhu, 2009; Dellaportas and Pourahmadi, 2012; Xue, Ma and Zou, 2012; Rajaratnam and Salzman, 2013). It assumes that the variables have a natural order, and variables that far apart from each other are weekly correlated. By imposing the sparse or banded structure on the Cholesky factor, it can result in certain sparse structure on the estimated covariance matrix. Unlike banding the covariance matrix itself (Bickel and Levina, 2004; Cai, Zhang and Zhou, 2010), it is guaranteed to be positive definite. Wu and Pourahmadi (2003) proposed a k-banded estimator of Cholesky factor, which can be obtained by regressing each variable only on its closest k predecessors. Bickel and Levina (2008) showed that banding the Cholesky factor produces a consistent estimator in the operator norm under weak conditions on the covariance matrix. One can also consider to impose an L_1 (Lasso)

penalty on the entries of the Cholesky factor for estimating the sparse covariance matrix (Huang et al., 2006). However, the MCD-based approach for estimating covariance matrix depends on the order of variables $X_1, \ldots, X_p$. Such an assumption on the variable order may not hold in practice. Actually, the variable order is often not available or cannot be pre-determined before the analysis in many applications, i.e., the gene expression data and stock marketing data.

In this paper, we adopt the MCD approach for estimating the large covariance matrix. Different from other MCD-based approach, the proposed estimate does not depend on the variable order, while maintains the positive-definite and sparse properties. Specifically, using the permutation idea, we first obtain a number of estimates of covariance matrix calculated from different orders of variables used in the MCD approach. With these available estimates, the proposed ensemble estimator is obtained as the "center" of them under the Frobenius norm through a L_1 penalized objective function. The L_1 regularization is imposed to achieve the sparsity of the estimate. Such an estimator takes advantages of multiple orders of variables due to the ensemble effort. An efficient algorithm with guaranteed convergence is also developed to make the computation attractive for obtaining the estimator. The proposed method performs much better than other existing approaches in terms of both accurate estimation and sparsity capturing.

The remainder of this work is organized as follows. Section 4.2 briefly reviews the MCD approach to estimate the covariance matrix. Section 4.3 introduces the proposed method by addressing the order issue. An efficient algorithm is also developed to solve the objective function. In Section 4.4, the convergence property is presented. The simulation study and one real data example are reported in Section 4.5 and 4.6, respectively. We conclude the paper in Section 4.7.

4.2 The Modified Cholesky Decomposition

Without loss of generality, suppose that $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is a p-dimensional random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent and identically distributed observations following $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Pourahmadi (1999) proposed the modified Cholesky decomposition (MCD) for the estimation of a covariance matrix, which is statistically meaningful and grantees the positive definiteness of the estimate. This decomposition arises from regressing each variable X_j on its predecessors $X_1, \ldots, X_{j-1}$ for $2 \leq j \leq p$. Specifically, consider to fit a series of regressions

$$X_j = \sum_{k=1}^{j-1} (-t_{jk}) X_k + \epsilon_j = \hat{X}_j + \epsilon_j,$$

where ϵ_j is the error term for the jth regression with $E\epsilon_j = 0$ and $Var(\epsilon_j) = d_j^2$. Let $\epsilon_1 = X_1$ and $\boldsymbol{D} = diag(d_1^2, \ldots, d_p^2)$ be the diagonal covariance matrix of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$. Construct the unit lower triangular matrix $\boldsymbol{T} = (t_{jk})_{p \times p}$ with ones on its diagonal and regression coefficients $(t_{j1}, \ldots, t_{j,j-1})'$ as its jth row. Then one can have

$$\boldsymbol{D} = Var(\boldsymbol{\epsilon}) = Var(\boldsymbol{X} - \hat{\boldsymbol{X}}) = Var(\boldsymbol{T}\boldsymbol{X}) = \boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}',$$

and thus

$$\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}\boldsymbol{D}\boldsymbol{T}'^{-1}. \tag{4.1}$$

The MCD approach reduces the challenge of modeling a covariance matrix into the task of modeling $(p-1)$ regression problems, and is applicable in high dimensions. However, directly imposing the sparse structure on Cholesky factor matrix $\boldsymbol{T}$ in (4.1) does not imply the sparse pattern of covariance matrix $\boldsymbol{\Sigma}$ since it requires an inverse of $\boldsymbol{T}$. Thus the formulation (4.1) is not convenient to impose a sparse structure on the estimation of $\boldsymbol{\Sigma}$. Alternatively, one can consider a latent variable regression model based on the MCD. Writing $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{\epsilon}$ would lead

to

$$Var(\boldsymbol{X}) = Var(\boldsymbol{L}\boldsymbol{\epsilon})$$
$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}'. \tag{4.2}$$

This decomposition can be interpreted as resulting from a new sequence of regressions, where each variable X_j is regressed on all the previous latent variable $\epsilon_1, \ldots, \epsilon_{j-1}$ rather than themselves. It gives a sequence of regressions

$$X_j = \boldsymbol{l}_j^T \boldsymbol{\epsilon} = \sum_{k<j} l_{jk}\epsilon_k + \epsilon_j, \quad j = 2, \ldots, p, \tag{4.3}$$

where $\boldsymbol{l}_j = (l_{jk})$ is the jth row of $\boldsymbol{L}$. Here $l_{jj} = 1$ and $l_{jk} = 0$ for $k > j$.

With the data matrix $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, define its jth column to be $\boldsymbol{x}^{(j)}$. Let $\boldsymbol{e}^{(j)}$ denote the residuals of $\boldsymbol{x}^{(j)}$, $j \geq 2$, and $\boldsymbol{e}^{(1)} = \boldsymbol{x}^{(1)}$. Let $\mathbb{Z}^{(j)} = (\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(j-1)})$ be the matrix containing the first $(j-1)$ residuals. Now we consider a sparse covariance matrix estimate $\hat{\boldsymbol{\Sigma}}$ by encouraging the sparsity on $\hat{\boldsymbol{L}}$. One approach to achieving such sparsity is to use the Lasso for the regression (Tibshirani, 1996)

$$\hat{\boldsymbol{l}}_j = \arg\min_{\boldsymbol{l}_j} \|\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{l}_j\|_2^2 + \eta_j\|\boldsymbol{l}_j\|_1, \ j = 2, \ldots, p, \tag{4.4}$$

where $\eta_j \geq 0$ is a tuning parameter and selected by cross validation. $\|\cdot\|_1$ stands for the vector L_1 norm. $\boldsymbol{e}^{(j)} = \boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{l}_j$ is used to construct the residuals for the last column of $\mathbb{Z}^{(j+1)}$. Then d_j^2 is estimated as the sample variance of $\boldsymbol{e}^{(j)}$

$$\hat{d}_j^2 = \widehat{Var}(\hat{\boldsymbol{e}}^{(j)}) = \widehat{Var}(\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\hat{\boldsymbol{l}}_j) \tag{4.5}$$

when constructing matrix $\hat{\boldsymbol{D}} = diag(\hat{d}_1^2, \ldots, \hat{d}_p^2)$. Hence, $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{D}}\hat{\boldsymbol{L}}'$ will be a sparse covariance matrix estimate.

4.3 The Proposed Method

Clearly, the estimate $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{D}}\hat{\boldsymbol{L}}'$ depends on the order of random variables $X_1, \ldots, X_p$. It means that different orders would lead to different sparse estimates of $\boldsymbol{\Sigma}$. To address this order-dependent issue, we consider an ensemble estimation of $\boldsymbol{\Sigma}$ by using the idea of permutation. Specifically, we generate M different permutations of $\{1, \ldots, p\}$ as the orders of the random variables, denoted by π_k's, $k = 1, \ldots, M$. Let $\boldsymbol{P}_{\pi_k}$ be the corresponding permutation matrix. Under a variable order π_k, the estimate $\hat{\boldsymbol{\Sigma}}_{\pi_k} = \hat{\boldsymbol{L}}_{\pi_k}\hat{\boldsymbol{D}}_{\pi_k}\hat{\boldsymbol{L}}_{\pi_k}^{'}$, where $\hat{\boldsymbol{L}}_{\pi_k}$ and $\hat{\boldsymbol{D}}_{\pi_k}$ are calculated based on (4.4) and (4.5). Then transforming back to the original order, we have $\hat{\boldsymbol{\Sigma}}_k = \boldsymbol{P}_{\pi_k}\hat{\boldsymbol{\Sigma}}_{\pi_k}\boldsymbol{P}'_{\pi_k}$. To obtain an ensemble estimator for $\boldsymbol{\Sigma}$, a naive solution would be $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k$. However, such an estimate may not be sparse since the sparse structure in $\hat{\boldsymbol{\Sigma}}_k$ is destroyed by the average.

In order to simultaneously achieve the positive definiteness and sparsity for the estimator, we propose to consider the estimate

$$\hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Sigma}\succeq\nu\boldsymbol{I}} \frac{1}{2M}\sum_{k=1}^{M}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_k\|_F^2 + \lambda|\boldsymbol{\Sigma}|_1, \tag{4.6}$$

where $\|\cdot\|_F$ stands for the Frobenius norm, $\lambda \geq 0$ is a tuning parameter, and $|\cdot|_1$ is L_1 norm for all the off-diagonal elements. Here ν is some positive arbitrarily small number. The constraint $\boldsymbol{\Sigma} \succeq \nu\boldsymbol{I}$ is to guarantee the positive-definiteness of the estimate. The penalty term is to encourage the sparse pattern in $\hat{\boldsymbol{\Sigma}}$. It is worth pointing out that, if $\lambda = 0$ in (4.6), the solution of $\hat{\boldsymbol{\Sigma}}$ would be $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k$; if without the constraint $\boldsymbol{\Sigma} \succeq \nu\boldsymbol{I}$ in (4.6), the solution of $\hat{\boldsymbol{\Sigma}}$ would be the soft-threshold estimate of $\bar{\boldsymbol{\Sigma}}$. Therefore, the proposed estimate is to pursue the "center" of the multiple estimates $\hat{\boldsymbol{\Sigma}}_k$, while maintain the properties of being positive-definite and sparse.

To efficiently solve the optimization in (4.6), we employ the alternating direction method of multipliers (ADMM). The ADMM technique has been widely used in solving the convex optimization under the content of L_1 penalized covariance matrix estimation (Xue, Ma and

Zou, 2012). ADMM does not require the differentiability assumption of the objective function and it is easy to implement. Specifically, let us first introduce a new variable $\mathbf{\Phi}$ and an equality constraint as follows

$$(\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Phi}}) = \arg\min_{\mathbf{\Sigma}, \mathbf{\Phi}} \{\frac{1}{2M} \sum_{k=1}^{M} \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_k\|_F^2 + \lambda |\mathbf{\Sigma}|_1 : \mathbf{\Sigma} = \mathbf{\Phi}, \mathbf{\Phi} \succeq \nu \boldsymbol{I}\}. \tag{4.7}$$

Note that the solution of (4.7) gives solution to (4.6). Then minimize its augmented Lagrangian function for some given penalty parameter τ

$$L(\mathbf{\Sigma}, \mathbf{\Phi}; \mathbf{\Lambda}) = \frac{1}{2M} \sum_{k=1}^{M} \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_k\|_F^2 + \lambda |\mathbf{\Sigma}|_1 - \langle \mathbf{\Lambda}, \mathbf{\Phi} - \mathbf{\Sigma} \rangle + \frac{1}{2\tau} \|\mathbf{\Phi} - \mathbf{\Sigma}\|_F^2, \tag{4.8}$$

where $\mathbf{\Lambda}$ is the Lagrangian multiplier and $\langle \cdot, \cdot \rangle$ stands for the inner product. ADMM iteratively solves the following steps sequentially for $i = 0, 1, 2, \ldots$ till convergence

$$\mathbf{\Phi} \text{ step} : \mathbf{\Phi}^{i+1} = \arg\min_{\mathbf{\Phi} \succeq \nu \boldsymbol{I}} L(\mathbf{\Sigma}^i, \mathbf{\Phi}; \mathbf{\Lambda}^i) \tag{4.9}$$

$$\mathbf{\Sigma} \text{ step} : \mathbf{\Sigma}^{i+1} = \arg\min_{\mathbf{\Sigma}} L(\mathbf{\Sigma}, \mathbf{\Phi}^{i+1}; \mathbf{\Lambda}^i) \tag{4.10}$$

$$\mathbf{\Lambda} \text{ step} : \mathbf{\Lambda}^{i+1} = \mathbf{\Lambda}^i - \frac{1}{\tau}(\mathbf{\Phi}^{i+1} - \mathbf{\Sigma}^{i+1}). \tag{4.11}$$

Assume the eigenvalue decomposition of a matrix $\boldsymbol{Z}$ is $\sum_{i=1}^{p} \lambda_i \boldsymbol{\xi}' \boldsymbol{\xi}$, and define $(\boldsymbol{Z})_+ = \sum_{i=1}^{p} \max(\lambda_i, \nu) \boldsymbol{\xi}_i' \boldsymbol{\xi}_i$. Then we develop the closed form for (4.9) as

$$\frac{\partial L(\mathbf{\Sigma}^i, \mathbf{\Phi}; \mathbf{\Lambda}^i)}{\partial \mathbf{\Phi}} = -\mathbf{\Lambda}^i + \frac{1}{\tau}(\mathbf{\Phi} - \mathbf{\Sigma}^i) \triangleq 0$$

$$\mathbf{\Phi} = \mathbf{\Sigma}^i + \tau \mathbf{\Lambda}^i$$

$$\mathbf{\Phi}^{i+1} = (\mathbf{\Sigma}^i + \tau \mathbf{\Lambda}^i)_+.$$

Next, define an element-wise soft threshold for each entry z_{ij} in matrix $\boldsymbol{Z}$ as $\boldsymbol{s}(\boldsymbol{Z}, \delta) = \{\boldsymbol{s}(z_{ij}, \delta)\}_{1 \le i,j \le p}$ with

$$\boldsymbol{s}(z_{ij}, \delta) = \text{sign}(z_{ij})\max(|z_{ij}| - \delta, 0) I_{\{i \neq j\}} + z_{ij} I_{\{i=j\}}.$$

Then the solution of (4.10) is derived as

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}^{i+1}; \boldsymbol{\Lambda}^i)}{\partial \boldsymbol{\Sigma}} &= \frac{1}{M}\sum_{k=1}^{M}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_k) + \boldsymbol{\Lambda}^i + \frac{1}{\tau}(\boldsymbol{\Sigma} - \boldsymbol{\Phi}^{i+1}) + \lambda \text{sign}^*(\boldsymbol{\Sigma}) \triangleq 0 \\
(\tau + 1)\boldsymbol{\Sigma} &= \tau(\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i) + \boldsymbol{\Phi}^{i+1} - \lambda\tau \text{sign}^*(\boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma}^{i+1} &= \{\boldsymbol{s}(\tau(\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i) + \boldsymbol{\Phi}^{i+1}, \lambda\tau)\}/(\tau + 1),
\end{aligned}$$

where $\text{sign}^*(\boldsymbol{\Sigma})$ means $\text{sign}(\boldsymbol{\Sigma})$ with the diagonal elements replaced by $\mathbf{0}$ vector. Algorithm 4 summarizes the developed procedure for solving (4.6) by ADMM.

Algorithm 4.

Step 1*: Input initial values* $\boldsymbol{\Sigma}^0$, $\boldsymbol{\Lambda}^0$ *and* τ.

Step 2*:* $\boldsymbol{\Phi}^{i+1} = (\boldsymbol{\Sigma}^i + \tau\boldsymbol{\Lambda}^i)_+$.

Step 3*:* $\boldsymbol{\Sigma}^{i+1} = \{\boldsymbol{s}(\tau(\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^i) + \boldsymbol{\Phi}^{i+1}, \lambda\tau)\}/(\tau + 1)$.

Step 4*:* $\boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\tau}(\boldsymbol{\Phi}^{i+1} - \boldsymbol{\Sigma}^{i+1})$.

Step 5*: Repeat Step 2 - 4 till convergence.*

This algorithm converges fast and produces the optimal solution of $\arg\min L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}; \boldsymbol{\Lambda})$ in (4.8). In practice, the initial estimate $\boldsymbol{\Sigma}^0$ is set to be the naive estimate $\bar{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k$. The $\boldsymbol{\Lambda}^0$ is set to be zero matrix, and $\tau = 2$ and $\nu = 10^{-4}$ (Xue, Ma and Zou, 2012). The optimal value of tuning parameter λ in (4.8) is chosen based on BIC (Yuan and Lin, 2007)

$$\text{BIC}(\lambda) = -\log|\hat{\boldsymbol{\Sigma}}_\lambda^{-1}| + \text{tr}[\hat{\boldsymbol{\Sigma}}_\lambda^{-1}\boldsymbol{S}] + \frac{\log n}{n}\sum_{i \le j}\hat{e}_{ij}(\lambda),$$

where $\boldsymbol{S}$ is the sample covariance matrix, $\hat{\boldsymbol{\Sigma}}_\lambda = (\hat{\sigma}_{ij}^{(\lambda)})_{p \times p}$ indicates the estimate of $\boldsymbol{\Sigma}$ obtained

by applying our algorithm with tuning parameter λ. $\hat{e}_{ij}(\lambda) = 0$ if $\hat{\sigma}_{ij}^{(\lambda)} = 0$, and $\hat{e}_{ij}(\lambda) = 1$ otherwise.

Note that the implementation of the proposed method also requires the choice of M, the number of permutation orders. Obviously, the number of all possible permeation orders is $p!$, which increases rapidly as the number of variables p increases. To get an appropriate value for M for efficient computation, we have tried $M = 10, 30, 50, 100$ and 150 as the potential number of random orders. The performances were quite comparable for M larger than 30. Hence, we choose $M = 100$ as the number of permutation orders for the proposed method in this paper.

4.4 Convergence Property

In this section, we can prove that the sequence $(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i)$ generated by Algorithm 4 from any starting point converges to an optimal minimizer $(\hat{\boldsymbol{\Sigma}}^+, \hat{\boldsymbol{\Phi}}^+, \hat{\boldsymbol{\Lambda}}^+)$ of (4.8), where $\hat{\boldsymbol{\Lambda}}^+$ is the optimal dual variable. To facilitate the proof, we first introduce some notation. Define a $2p$ by $2p$ matrix $\boldsymbol{J}$ as

$$\boldsymbol{J} = \begin{pmatrix} \tau \boldsymbol{I}_{p\times p} & 0 \\ 0 & \tau^{-1}\boldsymbol{I}_{p\times p} \end{pmatrix}.$$

Let the notation $\|\cdot\|_J^2$ be $\|\boldsymbol{U}\|_J^2 = \langle \boldsymbol{U}, \boldsymbol{J}\boldsymbol{U}\rangle$ and the inner product associated with $\boldsymbol{J}$ be $\langle \boldsymbol{U}, \boldsymbol{V}\rangle_J = \langle \boldsymbol{U}, \boldsymbol{J}\boldsymbol{V}\rangle$. Now we present the following lemma and theory. The proof is given in the Appendix.

Lemma 1. *Assume that* $(\hat{\boldsymbol{\Sigma}}^+, \hat{\boldsymbol{\Phi}}^+)$ *is an optimal solution of* (4.7) *and* $\hat{\boldsymbol{\Lambda}}^+$ *is the corresponding optimal dual variable with the equality constraint* $\boldsymbol{\Sigma} = \boldsymbol{\Phi}$*, then the sequence* $(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i)$ *generated by Algorithm 4 satisfies*

$$\|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2 - \|\boldsymbol{W}^+ - \boldsymbol{W}^{i+1}\|_J^2 \geq \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2,$$

where $\boldsymbol{W}^+ = (\hat{\boldsymbol{\Lambda}}^+, \hat{\boldsymbol{\Sigma}}^+)'$ *and* $\boldsymbol{W}^i = (\boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i)'$.

Theorem 1. *Suppose* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are* n *independent and identically distributed observations from* $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. *Then the sequence* $(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i)$ *generated by Algorithm 4 from any starting point converges to an optimal minimizer of the optimization function* (4.8).

Theorem 1 demonstrates the convergence of the proposed method. It automatically indicates that the sequence $\boldsymbol{\Sigma}^i, i = 1, 2, \ldots$, produced by Algorithm 4 converges to an optimal solution of the objective (4.6).

4.5 Simulation Study

In this section, we conduct a comprehensive simulation study to evaluate the proposed method. Suppose that data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are generated independently from the normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Here we consider the following four covariance matrix structures.

Model 1. $\boldsymbol{\Sigma}_1 = \mathrm{MA}(0.5, 0.3)$. The diagonal elements are 1 with the first sub-diagonal elements 0.5 and the seconde sub-diagonal elements 0.3.

Model 2. $\boldsymbol{\Sigma}_2 = \mathrm{AR}(0.5)$. The conditional covariance between any two random variables X_i and X_j is fixed to be $0.5^{|i-j|}$, $1 \leq i, j \leq p$.

Model 3. $\boldsymbol{\Sigma}_3$ is generated by randomly permuting rows and corresponding columns of $\boldsymbol{\Sigma}_1$.

Model 4. $\boldsymbol{\Sigma}_4$ is generated by randomly permuting rows and corresponding columns of $\boldsymbol{\Sigma}_2$.

Note that **Models** 1-2 consider the banded or nearly-banded structure for the covariance matrix. While the sparse pattern of the covariance matrix in **Models** 3-4 is not structured due to the random permutation. For each case, we generate the data set with three settings of sample sizes and variable sizes: (1) $n = 50, p = 30$; (2) $n = 50, p = 50$ and (3) $n = 50, p = 100$.

The performance of the proposed estimator is examined in comparison with several other approaches, which are divided into three classes. The first class is the sample covariance matrix $\boldsymbol{S}$ that serves as the benchmark. The second class is composed of three methods that deal with the variable order used in the MCD, including the MCD-based method with BIC order selection (BIC) (Dellaportas and Pourahmadi, 2012), the best permutation algorithm

(BPA) (Rajaratnam and Salzman, 2013) and the proposed method (Proposed). The BIC method determines the order of variables in the MCD (4.3) in a forward selection fashion. That is, in each step, it selects a new variable having the smallest value of BIC when regressing it on its previous residuals. For example, suppose that $\mathcal{C} = \{X_{i_1}, \ldots, X_{i_k}\}$ is the candidate set of variables and there are $(p-k)$ variables already chosen and ordered. By regressing each X_j, $j = i_1, \ldots, i_k$ on the residuals $[\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{p-k}]$, we can assign the variable corresponding to the minimum BIC value among the k regressions to the $(p-k+1)$th position of the order. The BPA selects the order of variables used in the MCD (4.2) such that $||\boldsymbol{D}||_F^2$ is minimized.

The third class of competing methods consists of four approaches that estimate the sparse covariance matrix directly without considering the variable order, including Bien and Tibshirani's estimate (BT) (Bien and Tibshirani, 2011), Bickel and Levina's estimate (BL) (Bickel and Levina, 2008), Xue, Ma and Zou's estimate (XMZ) (Xue, Ma and Zou, 2012) and Rothman, Levina and Zhu's estimate (RLZ) (Rothman, Levina and Zhu, 2010). The BT estimate minimizes the negative log-likelihood function with L_1 penalty on the entries of the covariance matrix, that is,

$$\hat{\boldsymbol{\Sigma}}_{BT} = \arg\min_{\boldsymbol{\Sigma} \succ 0} \Big\{ -\log|\boldsymbol{\Sigma}^{-1}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}S) + \eta|\boldsymbol{\Sigma}|_1 \Big\},$$

where $\eta \geq 0$ is the tuning parameter. The optimization is solved by using the majorization-minimization algorithm (Lange, Hunter and Yang, 2000). The BL estimate is obtained by imposing hard thresholding (Bickel and Levina, 2008) on the entries of the sample covariance matrix, so the resultant estimate may not be positive definite. The XMZ estimate is obtained from encouraging the sparsity on the sample covariance matrix $\boldsymbol{S}$ as well as maintaining the property of positive definiteness. The resultant estimate is

$$\hat{\boldsymbol{\Sigma}}_{XMZ} = \arg\min_{\boldsymbol{\Sigma} \succeq \nu \boldsymbol{I}} \frac{1}{2}\|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F^2 + \eta|\boldsymbol{\Sigma}|_1.$$

The RLZ estimate is to introduce sparsity in the Cholesky factor matrix $\boldsymbol{L}$ by estimating

the first k sub-diagonals of $\boldsymbol{L}$ and set the rest to zeroes. It means that each variable is only regressed on the k previous residuals in (4.3). The tuning parameter k can be chosen by AIC or cross validation.

To evaluate the performance of covariance matrix estimate $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})_{p\times p}$ obtained from each approach, we consider the entropy loss (EN), quadratic loss (QL), L_1 norm and mean absolute error (MAE), defined as follows:

$$\begin{aligned}
\text{EN} &= \text{tr}[\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}] - \log|\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}| - p, \\
\text{QL} &= \frac{1}{p}\ \text{tr}[\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma} - \boldsymbol{I}]^2, \\
L_1 \text{ norm} &= \max_j \sum_i |\hat{\sigma}_{ij} - \sigma_{ij}|, \\
\text{MAE} &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}|\hat{\sigma}_{ij} - \sigma_{ij}|.
\end{aligned}$$

Here we have not included the Kullback-Leibler loss (Kullback and Leibler, 1951), since it is more suitable in measuring inverse covariance matrix estimates rather than covariance matrix estimates (Levina et al., 2008). In addition, to gauge the performance of capturing sparse structure, we consider the false selection loss (FSL), which is the summation of false positive (FP) and false negative (FN). Here we say a FP occurs if a nonzero element in the true matrix is incorrectly estimated as a zero. Similarly, a FN occurs if a zero element in the true matrix is incorrectly identified as a nonzero. The FSL is computed in percentage as (FP + FN) / p^2. For each loss function above, Table 4.1 to Table 4.4 report the averages of the performance measures and their corresponding standard errors in the parentheses over 100 replicates. For each model, the two methods with lowest averages regarding each measure are shown in bold. Dashed lines in the tables represent the corresponding values not available due to matrix singularity.

For a short summary of Tables 1-4, the numerical results show that the proposed method generally provides a superior performance over other approaches. It is able to accurately catch the underlying sparse structure of the covariance matrix. When the underlying covari-

ance matrix is banded or tapered, the proposed method is comparable to the RLZ method, and performs better than other approaches. Note that the RLZ method targets on the banded covariance matrix. When the underlying structure of covariance matrix is more general without any specification, the proposed method performs evidently well. As in the high-dimensional cases, the advantage of the proposed method is even evident.

Specifically, Table 4.1 and 4.2 summarize the comparison results for **Models** 1 and 2, respectively. From the perspective of competing methods, the sample covariance matrix $\boldsymbol{S}$, serving as a benchmark approach, does not give the sparse structure and performs poorly under all the loss measures. The BIC and BPA in the second class of approaches provide sparse covariance matrix estimates compared with $\boldsymbol{S}$, but their false selection loss are considerably larger than the proposed method. Moreover, the proposed method greatly outperforms the BIC and BPA regarding QL, L_1 and MAE for all settings of $p = 30, 50$ and 100. Although the proposed method is comparable to the BIC and BPA methods under EN criterion when $p = 30$, it performs slightly better when $p = 50$ and much better in the case of $p = 100$.

For the BT in the third class of approaches, the proposed method significantly outperforms the BT approach in capturing the sparse structure for the cases of $p = 50$ and $p = 100$. Furthermore, the proposed method gives superior performance to the BT with respect to all the other loss criteria, especially QL measure. In comparison with the BL method, the proposed method performs similarly to the BL approach. It is known that the BL method is asymptotically optimal for sparse covariance matrix (Bickel and Levina, 2008). However, the estimated covariance matrix obtained from the BL method does not guarantee to be positive definite, which would result in matrix singularity in computing EN and QL losses. Compared with XMZ approach, the proposed method is superior or comparable with respect to all the loss measures except QL criterion in the settings of $p = 30$ and 50. For the high dimension case when $p = 100$, the proposed method performs much better than the XMZ approach. Finally, it is seen that the performance of the proposed method and the RLZ approach is comparable under **Models** 1-2 regarding these loss criteria. This is not surprising, since the covariance matrices of **Models** 1-2 are banded and tapered respectively, and the RLZ

Table 4.1: The averages and standard errors (in parenthesis) of estimates for **Model** 1.

		EN	QL	L_1	MAE	FSL (%)
$p = 30$	$\boldsymbol{S}$	12.42 (0.08)	88.11 (2.83)	5.21 (0.07)	3.51 (0.03)	83.96 (0.01)
	BIC	7.22 (0.08)	8.59 (0.59)	2.96 (0.04)	1.76 (0.02)	52.32 (0.55)
	BPA	**6.02 (0.09)**	4.38 (0.23)	2.74 (0.05)	1.55 (0.02)	46.53 (0.68)
	BT	7.78 (0.03)	19.44 (0.29)	2.14 (0.01)	1.85 (0.00)	6.92 (0.10)
	BL	-	-	2.33 (0.05)	**1.17 (0.02)**	6.81 (0.14)
	XMZ	10.61 (0.17)	**0.13 (0.02)**	**1.88 (0.01)**	1.25 (0.01)	7.13 (0.18)
	RLZ	9.36 (0.07)	**0.35 (0.03)**	**1.55 (0.01)**	**1.04 (0.01)**	**6.22 (0.00)**
	Proposed	**7.10 (0.11)**	0.38 (0.03)	1.92 (0.02)	1.22 (0.01)	**6.75 (0.15)**
$p = 50$	$\boldsymbol{S}$	-	-	8.26 (0.07)	5.75 (0.03)	90.19 (0.01)
	BIC	15.77 (0.21)	125.73 (28.51)	3.85 (0.06)	1.98 (0.01)	43.30 (0.47)
	BPA	**12.98 (0.16)**	15.67 (1.46)	3.62 (0.08)	1.84 (0.02)	41.49 (0.63)
	BT	15.07 (0.28)	106.19 (8.57)	2.42 (0.02)	1.93 (0.02)	10.68 (0.46)
	BL	-	-	2.41 (0.05)	**1.27 (0.01)**	4.80 (0.06)
	XMZ	20.45 (0.29)	**0.14 (0.02)**	**1.98 (0.01)**	1.36 (0.01)	5.13 (0.07)
	RLZ	16.20 (0.07)	**0.62 (0.05)**	**1.63 (0.01)**	**1.06 (0.00)**	**3.84 (0.00)**
	Proposed	**13.68 (0.10)**	1.10 (0.06)	2.02 (0.01)	1.35 (0.01)	**4.36 (0.06)**
$p = 100$	$\boldsymbol{S}$	-	-	16.15 (0.12)	11.43 (0.03)	95.01 (0.00)
	BIC	42.56 (0.45)	177503 (149636)	5.26 (0.07)	2.24 (0.01)	32.75 (0.36)
	BPA	35.68 (0.38)	3223.74 (1749.81)	5.29 (0.11)	2.20 (0.02)	33.60 (0.38)
	BT	**28.78 (0.33)**	97.08 (3.21)	2.40 (0.03)	1.87 (0.02)	4.08 (0.33)
	BL	-	-	2.67 (0.06)	**1.42 (0.01)**	2.83 (0.02)
	XMZ	364.15 (0.18)	Inf (Inf)	16.14 (0.12)	11.42 (0.03)	94.96 (0.01)
	RLZ	33.40 (0.12)	**1.28 (0.06)**	**1.69 (0.01)**	**1.08 (0.00)**	**1.96 (0.00)**
	Proposed	**31.28 (0.14)**	**4.04 (0.12)**	**2.12 (0.01)**	1.49 (0.00)	**2.38 (0.02)**

Table 4.2: The averages and standard errors (in parenthesis) of estimates for **Model** 2.

		EN	QL	L_1	MAE	FSL (%)
$p = 30$	$\boldsymbol{S}$	12.58 (0.08)	92.44 (2.69)	5.10 (0.07)	3.49 (0.03)	46.66 (0.01)
	BIC	5.35 (0.09)	5.55 (0.40)	2.94 (0.04)	1.91 (0.01)	43.47 (0.30)
	BPA	4.60 (0.07)	3.64 (0.22)	2.82 (0.04)	1.78 (0.01)	42.53 (0.31)
	BT	5.40 (0.03)	19.36 (0.32)	2.52 (0.01)	2.19 (0.00)	43.82 (0.08)
	BL	-	-	2.61 (0.04)	**1.57 (0.01)**	43.70 (0.20)
	XMZ	5.82 (0.11)	**0.12 (0.01)**	**2.27 (0.01)**	1.64 (0.01)	**42.24 (0.20)**
	RLZ	**3.16 (0.03)**	**0.43 (0.04)**	**1.89 (0.01)**	**1.34 (0.01)**	43.56 (0.00)
	Proposed	**4.09 (0.06)**	0.48 (0.04)	2.30 (0.01)	1.66 (0.01)	**41.82 (0.16)**
$p = 50$	$\boldsymbol{S}$	-	-	8.54 (0.08)	5.84 (0.03)	65.59 (0.01)
	BIC	11.08 (0.20)	42.50 (7.50)	3.93 (0.08)	2.18 (0.01)	42.28 (0.24)
	BPA	9.17 (0.15)	11.93 (0.86)	3.70 (0.06)	2.08 (0.02)	41.73 (0.28)
	BT	10.07 (0.18)	52.46 (3.38)	2.69 (0.02)	2.26 (0.01)	30.34 (0.13)
	BL	-	-	2.91 (0.05)	**1.68 (0.01)**	29.45 (0.07)
	XMZ	11.26 (0.20)	**0.12 (0.01)**	**2.36 (0.01)**	1.76 (0.01)	28.85 (0.07)
	RLZ	**5.48 (0.04)**	**0.64 (0.04)**	**1.96 (0.01)**	**1.38 (0.00)**	**28.48 (0.00)**
	Proposed	**7.22 (0.06)**	0.98 (0.05)	2.40 (0.01)	1.77 (0.01)	**28.60 (0.07)**
$p = 100$	$\boldsymbol{S}$	-	-	16.04 (0.12)	11.43 (0.04)	81.86 (0.00)
	BIC	29.70 (0.55)	17694 (11352)	5.32 (0.08)	2.47 (0.01)	33.77 (0.25)
	BPA	23.65 (0.36)	854.23 (232.83)	5.16 (0.11)	2.44 (0.02)	34.41 (0.31)
	BT	21.09 (0.36)	95.23 (3.68)	2.80 (0.03)	2.24 (0.02)	17.01 (0.23)
	BL	-	-	3.04 (0.05)	**1.82 (0.01)**	16.07 (0.02)
	XMZ	369.27 (0.19)	Inf (Inf)	16.03 (0.12)	11.42 (0.04)	81.83 (0.00)
	RLZ	**11.03 (0.07)**	**1.33 (0.06)**	**2.05 (0.02)**	**1.41 (0.00)**	**15.12 (0.00)**
	Proposed	**16.29 (0.09)**	**3.46 (0.12)**	**2.49 (0.01)**	1.90 (0.00)	**15.63 (0.02)**

Table 4.3: The averages and standard errors (in parenthesis) of estimates for **Model** 3.

		EN	QL	L_1	MAE	FSL (%)
$p = 30$	$\boldsymbol{S}$	12.46 (0.09)	91.29 (3.34)	5.08 (0.06)	3.48 (0.02)	83.96 (0.01)
	BIC	7.28 (0.10)	9.37 (0.70)	2.93 (0.04)	1.76 (0.01)	52.71 (0.51)
	BPA	**5.87 (0.09)**	4.54 (0.38)	2.60 (0.05)	1.52 (0.02)	46.59 (0.78)
	BT	7.77 (0.04)	19.38 (0.34)	2.14 (0.01)	1.85 (0.00)	6.94 (0.10)
	BL	-	-	2.24 (0.05)	**1.14 (0.01)**	**6.62 (0.15)**
	XMZ	10.53 (0.14)	**0.16 (0.02)**	**1.87 (0.01)**	1.23 (0.01)	7.00 (0.15)
	RLZ	16.75 (0.11)	**0.37 (0.03)**	2.28 (0.02)	1.68 (0.00)	15.55 (0.00)
	Proposed	**6.96 (0.10)**	0.38 (0.03)	**1.89 (0.02)**	**1.21 (0.01)**	**6.84 (0.13)**
$p = 50$	$\boldsymbol{S}$	-	-	8.28 (0.08)	5.75 (0.03)	90.19 (0.01)
	BIC	16.00 (0.21)	113.06 (22.87)	3.85 (0.08)	1.98 (0.02)	43.07 (0.50)
	BPA	**12.92 (0.16)**	15.30 (1.73)	3.55 (0.08)	1.83 (0.02)	40.90 (0.65)
	BT	15.58 (0.26)	114.94 (9.63)	2.45 (0.02)	1.96 (0.01)	11.21 (0.47)
	BL	-	-	2.46 (0.05)	**1.27 (0.01)**	**4.79 (0.06)**
	XMZ	20.40 (0.32)	**0.17 (0.02)**	**1.97 (0.01)**	**1.36 (0.01)**	5.18 (0.07)
	RLZ	31.67 (0.17)	**0.69 (0.05)**	2.43 (0.01)	1.90 (0.00)	11.36 (0.00)
	Proposed	**13.74 (0.11)**	1.13 (0.07)	**2.00 (0.01)**	1.36 (0.01)	**4.39 (0.07)**
$p = 100$	$\boldsymbol{S}$	-	-	16.18 (0.10)	11.43 (0.03)	95.01 (0.00)
	BIC	42.88 (0.52)	29998 (13821)	5.36 (0.10)	2.26 (0.01)	32.65 (0.37)
	BPA	35.37 (0.43)	2794.90 (1780.94)	5.12 (0.12)	2.19 (0.02)	33.30 (0.41)
	BT	**28.78 (0.37)**	99.03 (3.40)	**2.36 (0.02)**	1.89 (0.02)	3.91 (0.30)
	BL	-	-	2.58 (0.05)	**1.41 (0.01)**	**2.83 (0.02)**
	XMZ	364.17 (0.16)	Inf (Inf)	16.17 (0.10)	11.42 (0.03)	94.96 (0.00)
	RLZ	64.68 (0.22)	**1.28 (0.06)**	2.50 (0.01)	1.92 (0.00)	5.72 (0.00)
	Proposed	**31.33 (0.15)**	**3.81 (0.11)**	**2.10 (0.01)**	**1.49 (0.00)**	**2.39 (0.02)**

Table 4.4: The averages and standard errors (in parenthesis) of estimates for **Model** 4.

		EN	QL	L_1	MAE	FSL (%)
$p = 30$	$\boldsymbol{S}$	12.48 (0.08)	89.99 (2.78)	5.16 (0.07)	3.49 (0.03)	46.66 (0.01)
	BIC	5.22 (0.08)	5.30 (0.39)	2.89 (0.04)	1.90 (0.01)	43.03 (0.33)
	BPA	**4.48 (0.08)**	3.55 (0.23)	2.79 (0.04)	1.76 (0.01)	42.66 (0.35)
	BT	5.36 (0.04)	19.32 (0.35)	2.50 (0.01)	2.19 (0.00)	43.79 (0.08)
	BL	-	-	2.69 (0.06)	**1.58 (0.02)**	43.86 (0.20)
	XMZ	5.86 (0.09)	**0.10 (0.01)**	**2.24 (0.01)**	**1.64 (0.01)**	**42.31 (0.19)**
	RLZ	11.57 (0.10)	**0.46 (0.04)**	2.67 (0.02)	2.15 (0.00)	50.67 (0.00)
	Proposed	**4.06 (0.05)**	0.48 (0.04)	**2.27 (0.01)**	1.66 (0.01)	**42.06 (0.16)**
$p = 50$	$\boldsymbol{S}$	-	-	8.24 (0.08)	5.74 (0.03)	65.58 (0.01)
	BIC	11.16 (0.21)	352.49 (280.45)	3.95 (0.08)	2.17 (0.01)	42.10 (0.24)
	BPA	**9.22 (0.15)**	21.30 (8.30)	3.72 (0.08)	2.07 (0.01)	41.28 (0.30)
	BT	10.41 (0.21)	60.15 (5.09)	2.71 (0.01)	2.27 (0.01)	30.31 (0.14)
	BL	-	-	2.79 (0.06)	**1.68 (0.01)**	29.50 (0.07)
	XMZ	10.96 (0.19)	**0.10 (0.01)**	**2.37 (0.01)**	**1.75 (0.01)**	**28.78 (0.08)**
	RLZ	19.21 (0.11)	**0.72 (0.04)**	2.77 (0.01)	2.24 (0.00)	33.60 (0.00)
	Proposed	**7.29 (0.06)**	1.13 (0.06)	**2.41 (0.01)**	1.78 (0.01)	**28.56 (0.07)**
$p = 100$	$\boldsymbol{S}$	-	-	16.10 (0.13)	11.44 (0.04)	81.85 (0.00)
	BIC	30.15 (0.51)	21511.89 (6496.05)	5.39 (0.10)	2.46 (0.01)	33.63 (0.25)
	BPA	23.53 (0.35)	7354.89 (5272.40)	5.39 (0.13)	2.43 (0.01)	33.92 (0.29)
	BT	**20.62 (0.33)**	90.39 (3.13)	**2.73 (0.02)**	2.24 (0.02)	16.43 (0.17)
	BL	-	-	3.03 (0.06)	**1.82 (0.01)**	**16.08 (0.02)**
	XMZ	369.42 (0.22)	Inf (Inf)	16.09 (0.13)	11.43 (0.04)	81.82 (0.01)
	RLZ	40.19 (0.21)	**1.30 (0.08)**	2.89 (0.01)	2.31 (0.00)	18.44 (0.00)
	Proposed	**16.39 (0.08)**	**3.41 (0.13)**	**2.49 (0.01)**	**1.90 (0.00)**	**15.64 (0.02)**

approach is designated to estimate such covariance matrix structures.

Table 4.3 and 4.4 present the comparison results for **Models** 3 and 4. Different from **Models** 1-2, the covariance matrices under **Models** 3-4 are unstructured. This implies that the RLZ approach does not have the advantage. Hence, it is clearly seen that the proposed method performs much better than the RLZ approach, especially at capturing the sparse structure and with respect to EN loss. Generally, the proposed method provides superior performance to other approaches, with similar comparison results as described under **Models** 1-2.

4.6 Application

In this section, a real prostate cancer data set (Glaab et al., 2012) is used to evaluate the performance of the proposed method by comparison with other approaches used in Section 4.5. It contains two classes with 50 normal samples and 52 prostate cancer samples, and 2135 gene expression values recorded for each sample. Data are available online at http://ico2s.org/datasets/microarray.html. Since it includes a large number of variables, the variable screening procedure is performed through two sample t-test. Specifically, for each variable, t-test is conducted against the two classes of the prostate cancer data such that the variables corresponding to large values of test statistics are ranked as significant variables. Then the top 50 significant variables as group 1 and the top 50 nonsignificant variables as group 2 are selected for data analysis. As mentioned in Xue, Ma and Zou (2012), there would be some correlations within each group of variables, but no correlations between group 1 variables and group 2 variables. Data are centered within each class and then used for the analysis. In the section, to make each variable at the same scale, we focus on the correlation matrix rather than the covariance matrix.

Figure 4.1 shows the heatmaps of the absolute values of the estimated correlation matrices obtained from each method. It is clear to see that, overall, the estimated sparse pattern of the proposed method, BT, and XMZ well matches the expected sparse pattern. They shrink

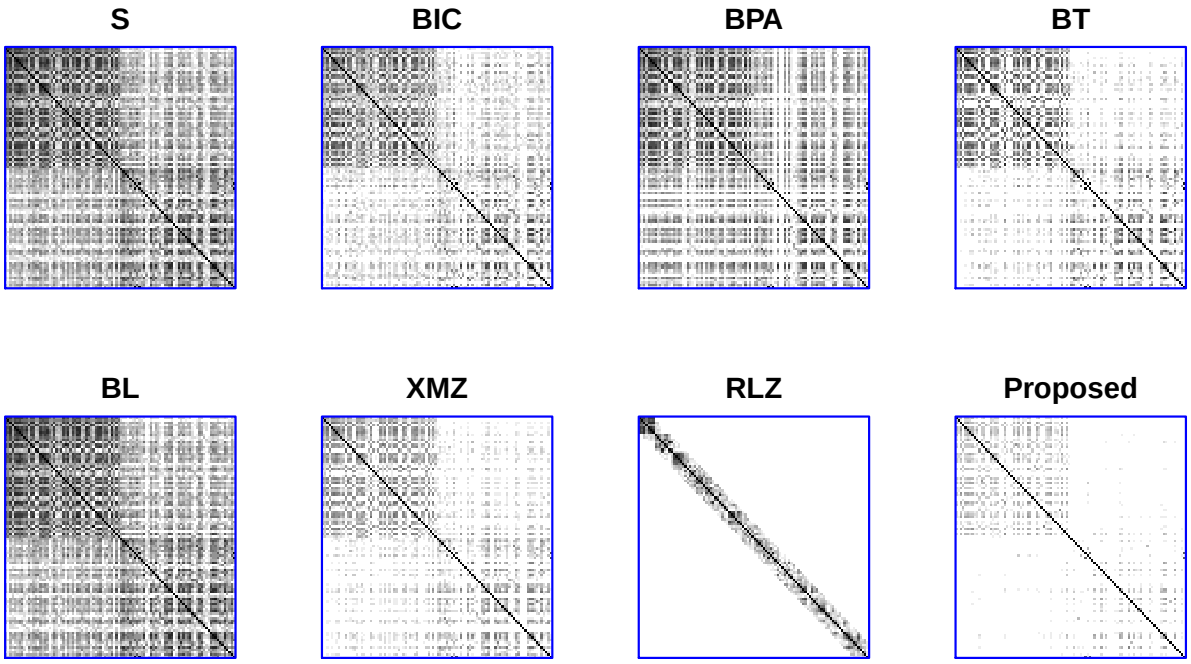

Figure 4.1: Heatmaps of the absolute values of the correlation matrices obtained from the proposed method and other approaches for prostate cancer data. Darker colour indicates higher density; lighter colour indicates lower density.

the non-zero correlations whereas the others (except RLZ) do not. The proposed method and XMZ perform the best in capturing the sparse structure with two diagonal blocks, followed by BT producing a little denser structure. All the rest approaches either result in a much sparser matrix as diagonal matrix (i.e. RLZ) or fail to catch the sparsity pattern (i.e. $\boldsymbol{S}$, BIC, BPA and BL).

In addition, Figure 4.2 displays the eigenvalues in the decreasing order for each estimate. We see that the sample covariance matrix has the most spread out eigenvalues, followed by BL and BPA. The eigenvalues from the proposed method and RLZ have the least spread. The largest and smallest eigenvalues obtained from each approach are summarized in Table 4.5. BT estimator yields a negative definite matrix with negative eigenvalues, while the other estimators guarantee the positive definiteness. Although RLZ performs slightly better than the proposed method in terms of eigenvalues spread, it produces an incorrect sparsity pattern

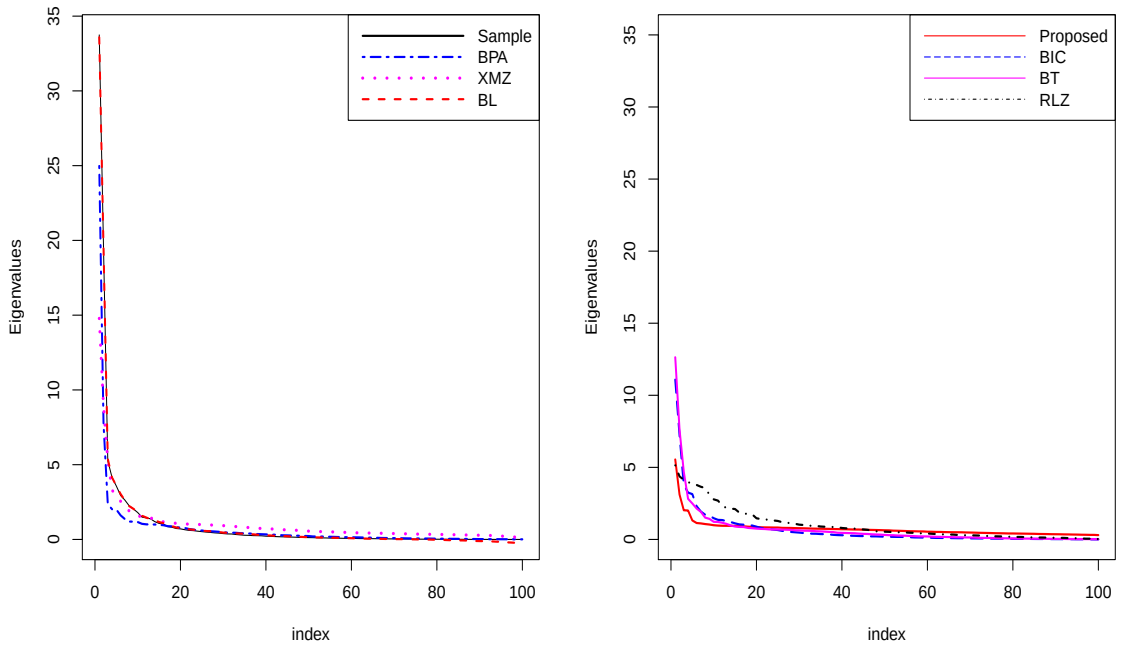

Figure 4.2: Scree plot for correlation matrices obtained from the proposed method and other approaches for prostate cancer data.

shown in Figure 4.1.

4.7 Discussion

In this paper, we have introduced a positive definite ensemble estimate of covariance matrix for the high-dimensional data. The proposed method takes advantages of the multiple estimates obtained from the modified Cholesky decomposition (MCD) approach under different variable orders. The positive definite constraint and L_1 penalty are added to the objective function to guarantee the positive definiteness and encourage the sparse structure of the estimated covariance matrix. An efficient algorithm is developed to solve the constraint optimization problem. The proposed estimator does not require the prior knowledge of the variable order used in the MCD, and performs well in the high-dimensional cases. An interesting question is that whether the proposed estimator for the covariance matrix is consistent. The solution for the optimization problem (4.6) without the positive definite constraint $\boldsymbol{\Sigma} \succeq \nu \boldsymbol{I}$ would be the

Table 4.5: The largest and smallest eigenvalues of correlation matrices obtained from each approach for prostate data.

Method	S	BIC	BPA	BT	BL	XMZ	RLZ	Proposed
Largest	33.7	11.1	25.0	12.6	33.6	14.8	5.1	5.5
Smallest	1.2e-05	3.3e-03	6.8e-03	1.6e-02	-2.9e-01	4.9e-04	2.9e-02	3.0e-01

soft-threshold estimate of $\bar{\mathbf{\Sigma}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k$. On one hand, it is known that $\bar{\mathbf{\Sigma}}$ is consistent. On the other hand, the soft-threshold operator maintains the consistency property under some mild conditions. Hence, it is not hard to prove the consistency property of the proposed estimator. A strict proof is left for the future work.

4.8 Appendix

Proof of Lemma 1

Since $(\mathbf{\Sigma}^+, \mathbf{\Phi}^+, \mathbf{\Lambda}^+)$ is the optimal minimizer of (4.8), based on the KKT conditions we have

$$(-\hat{\mathbf{\Sigma}}^+ + \frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k - \hat{\mathbf{\Lambda}}^+)_{jl} \in \lambda\partial|\hat{\mathbf{\Sigma}}^+_{jl}|,\ j = 1,\ldots,p, l = 1,\ldots,p, \text{and } j \neq l \tag{4.12}$$

$$(-\hat{\mathbf{\Sigma}}^+ + \frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k)_{jj} + \hat{\mathbf{\Lambda}}^+_{jj} = 0,\ j = 1,\ldots,p \tag{4.13}$$

$$\hat{\mathbf{\Phi}}^+ = \hat{\mathbf{\Sigma}}^+ \tag{4.14}$$

$$\hat{\mathbf{\Phi}}^+ \succeq \nu\boldsymbol{I}, \tag{4.15}$$

and

$$\langle \hat{\mathbf{\Lambda}}^{+}, \mathbf{\Phi} - \hat{\mathbf{\Phi}}^{+} \rangle \leq 0, \ \forall \mathbf{\Phi} \succeq \nu \boldsymbol{I}. \tag{4.16}$$

(4.12) and (4.13) result from the stationarity. (4.14) and (4.15) are due to the primal feasibility. By the optimality conditions of the problem (4.9) with respect to $\mathbf{\Phi}$, we obtain

$$\langle \mathbf{\Lambda}^{i} - \frac{1}{\tau}(\mathbf{\Phi}^{i+1} - \mathbf{\Sigma}^{i}), \mathbf{\Phi} - \mathbf{\Phi}^{i+1} \rangle \leq 0, \ \forall \mathbf{\Phi} \succeq \nu \boldsymbol{I}.$$

This, together with $\mathbf{\Lambda}$ step (4.11), yields

$$\langle \mathbf{\Lambda}^{i+1} - \frac{1}{\tau}(\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^{i}), \mathbf{\Phi} - \mathbf{\Phi}^{i+1} \rangle \leq 0, \ \forall \mathbf{\Phi} \succeq \nu \boldsymbol{I}. \tag{4.17}$$

Now setting $\mathbf{\Phi} = \mathbf{\Phi}^{i+1}$ in (4.16) and $\mathbf{\Phi} = \hat{\mathbf{\Phi}}^{+}$ in (4.17) respectively leads to

$$\langle \hat{\mathbf{\Lambda}}^{+}, \mathbf{\Phi}^{i+1} - \hat{\mathbf{\Phi}}^{+} \rangle \leq 0, \tag{4.18}$$

and

$$\langle \mathbf{\Lambda}^{i+1} - \frac{1}{\tau}(\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^{i}), \hat{\mathbf{\Phi}}^{+} - \mathbf{\Phi}^{i+1} \rangle \leq 0. \tag{4.19}$$

Summing (4.18) and (4.19) gives

$$\langle (\mathbf{\Lambda}^{i+1} - \hat{\mathbf{\Lambda}}^{+}) - \frac{1}{\tau}(\mathbf{\Sigma}^{i+1} - \mathbf{\Sigma}^{i}), \mathbf{\Phi}^{i+1} - \hat{\mathbf{\Phi}}^{+} \rangle \geq 0. \tag{4.20}$$

On the other hand, by the optimality conditions of the problem (4.10) with respect to $\mathbf{\Sigma}$, we have

$$0 \in [\frac{1}{M}\sum_{k=1}^{M}(\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}_k) + \mathbf{\Lambda}^{i} + \frac{1}{\tau}(\mathbf{\Sigma}^{i+1} - \mathbf{\Phi}^{i+1})]_{jl} + \lambda \partial |\mathbf{\Sigma}_{jl}^{i+1}|, \ j = 1, \ldots, p, l = 1, \ldots, p, \text{and } j \neq l, \tag{4.21}$$

and

$$[\frac{1}{M}\sum_{k=1}^{M}(\mathbf{\Sigma}^{i+1}-\hat{\mathbf{\Sigma}}_k)+\mathbf{\Lambda}^i+\frac{1}{\tau}(\mathbf{\Sigma}^{i+1}-\mathbf{\Phi}^{i+1})]_{jj}=0,\ j=1,\ldots,p, \tag{4.22}$$

Plugging $\mathbf{\Lambda}$ step (4.11) into (4.21) and (4.22) respectively results in

$$(-\mathbf{\Sigma}^{i+1}+\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k-\mathbf{\Lambda}^{i+1})_{jl}\in\lambda\partial|\mathbf{\Sigma}_{jl}^{i+1}|,\ j=1,\ldots,p,l=1,\ldots,p,\text{and } j\neq l, \tag{4.23}$$

and

$$(\mathbf{\Sigma}^{i+1}-\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k)_{jj}+\mathbf{\Lambda}_{jj}^{i+1}=0,\ j=1,\ldots,p. \tag{4.24}$$

Since $\partial|\cdot|$ is monotonically non-decreasing, (4.12) and (4.23) yield for $j\neq l$

$$(-\mathbf{\Sigma}^{i+1}+\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k-\mathbf{\Lambda}^{i+1})_{jl}\begin{cases}\geq(-\hat{\mathbf{\Sigma}}^{+}+\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k-\hat{\mathbf{\Lambda}}^{+})_{jl},\ \text{if}\ \ \mathbf{\Sigma}_{jl}^{i+1}\geq\hat{\mathbf{\Sigma}}_{jl}^{+}\\ \leq(-\hat{\mathbf{\Sigma}}^{+}+\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Sigma}}_k-\hat{\mathbf{\Lambda}}^{+})_{jl},\ \text{if}\ \ \mathbf{\Sigma}_{jl}^{i+1}<\hat{\mathbf{\Sigma}}_{jl}^{+}\end{cases},$$

that is,

$$(\hat{\mathbf{\Sigma}}^{+}-\mathbf{\Sigma}^{i+1}+\hat{\mathbf{\Lambda}}^{+}-\mathbf{\Lambda}^{i+1})_{jl}\begin{cases}\geq 0,\ \text{if}\ \ \mathbf{\Sigma}_{jl}^{i+1}\geq\hat{\mathbf{\Sigma}}_{jl}^{+}\\ \leq 0,\ \text{if}\ \ \mathbf{\Sigma}_{jl}^{i+1}<\hat{\mathbf{\Sigma}}_{jl}^{+}\end{cases}.$$

As a result, we obtain

$$(\mathbf{\Sigma}^{i+1}-\hat{\mathbf{\Sigma}}^{+})_{jl}(\hat{\mathbf{\Sigma}}^{+}-\mathbf{\Sigma}^{i+1}+\hat{\mathbf{\Lambda}}^{+}-\mathbf{\Lambda}^{i+1})_{jl}\geq 0,\ j=1,\ldots,p,l=1,\ldots,p,\text{and } j\neq l. \tag{4.25}$$

In addition, subtracting (4.24) from (4.13) implies

$$(\hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^{i+1} + \hat{\mathbf{\Lambda}}^{+} - \mathbf{\Lambda}^{i+1})_{jj} = 0,\ j = 1, \ldots, p. \tag{4.26}$$

Then combining (4.25) and (4.26) leads to

$$\langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{+}, \hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^{i+1} + \hat{\mathbf{\Lambda}}^{+} - \mathbf{\Lambda}^{i+1} \rangle \geq 0. \tag{4.27}$$

By summing (4.20) and (4.27), we have

$$\langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{+}, \hat{\mathbf{\Lambda}}^{+} - \mathbf{\Lambda}^{i+1} \rangle + \langle \mathbf{\Lambda}^{i+1} - \hat{\mathbf{\Lambda}}^{+}, \mathbf{\Phi}^{i+1} - \hat{\mathbf{\Phi}}^{+} \rangle - \frac{1}{\tau} \langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{i}, \mathbf{\Phi}^{i+1} - \hat{\mathbf{\Phi}}^{+} \rangle \geq \|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{+}\|_F^2.$$

This, together with (4.14) and $\mathbf{\Phi}^{i+1} = \tau(\mathbf{\Lambda}^{i} - \mathbf{\Lambda}^{i+1}) + \mathbf{\Sigma}^{i+1}$ from $\mathbf{\Lambda}$ step (4.11), gives

$$\begin{aligned} \tau \langle \mathbf{\Lambda}^{i+1} - \hat{\mathbf{\Lambda}}^{+}, \mathbf{\Lambda}^{i} - \mathbf{\Lambda}^{i+1} \rangle + \frac{1}{\tau} \langle \mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{+}, \mathbf{\Sigma}^{i} - \mathbf{\Sigma}^{i+1} \rangle \\ \geq \|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{+}\|_F^2 - \langle \mathbf{\Lambda}^{i} - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^{i} - \mathbf{\Sigma}^{i+1} \rangle \end{aligned} \tag{4.28}$$

By $\hat{\mathbf{\Phi}}^{+} - \mathbf{\Phi}^{i+1} = (\hat{\mathbf{\Phi}}^{+} - \mathbf{\Phi}^{i}) + (\mathbf{\Phi}^{i} - \mathbf{\Phi}^{i+1})$ and $\hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^{i+1} = (\hat{\mathbf{\Sigma}}^{+} - \mathbf{\Sigma}^{i}) + (\mathbf{\Sigma}^{i} - \mathbf{\Sigma}^{i+1})$, (4.28) is reduced to

$$\begin{aligned} \tau \langle \mathbf{\Lambda}^{i} - \hat{\mathbf{\Lambda}}^{+}, \mathbf{\Lambda}^{i} - \mathbf{\Lambda}^{i+1} \rangle + \frac{1}{\tau} \langle \mathbf{\Sigma}^{i} - \hat{\mathbf{\Sigma}}^{+}, \mathbf{\Sigma}^{i} - \mathbf{\Sigma}^{i+1} \rangle \geq \tau \|\mathbf{\Lambda}^{i} - \mathbf{\Lambda}^{i+1}\|_F^2 \\ + \frac{1}{\tau} \|\mathbf{\Sigma}^{i} - \mathbf{\Sigma}^{i+1}\|_F^2 + \|\mathbf{\Sigma}^{i+1} - \hat{\mathbf{\Sigma}}^{+}\|_F^2 - \langle \mathbf{\Lambda}^{i} - \mathbf{\Lambda}^{i+1}, \mathbf{\Sigma}^{i} - \mathbf{\Sigma}^{i+1} \rangle \end{aligned} \tag{4.29}$$

Using the notations $\boldsymbol{W}^+$ and $\boldsymbol{W}^i$, the left hand side of (4.29) becomes

$$\begin{aligned}
&\langle(\boldsymbol{\Lambda}^i - \hat{\boldsymbol{\Lambda}}^+, \boldsymbol{\Sigma}^i - \hat{\boldsymbol{\Sigma}}^+)', [\tau(\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}), \frac{1}{\tau}(\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1})]'\rangle \\
&= \langle(\boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i)' - (\hat{\boldsymbol{\Lambda}}^+, \hat{\boldsymbol{\Sigma}}^+)', \boldsymbol{J}[(\boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i)' - (\hat{\boldsymbol{\Lambda}}^{i+1}, \hat{\boldsymbol{\Sigma}}^{i+1})']\rangle \\
&= \langle \boldsymbol{W}^i - \boldsymbol{W}^+, \boldsymbol{J}(\boldsymbol{W}^i - \boldsymbol{W}^{i+1})\rangle \\
&= \langle \boldsymbol{W}^i - \boldsymbol{W}^+, \boldsymbol{W}^i - \boldsymbol{W}^{i+1}\rangle_J.
\end{aligned}$$

The first two terms on the right side of (4.29) becomes

$$\begin{aligned}
\tau\|\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}\|_F^2 + \frac{1}{\tau}\|\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\|_F^2 &= \tau\langle\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}\rangle + \frac{1}{\tau}\langle\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle \\
&= \langle(\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1})', [\tau(\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}), \frac{1}{\tau}(\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1})]'\rangle \\
&= \langle(\boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i)' - (\boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i+1})', \boldsymbol{J}[(\boldsymbol{\Lambda}^i, \boldsymbol{\Sigma}^i)' - (\boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^{i+1})']\rangle \\
&= \langle \boldsymbol{W}^i - \boldsymbol{W}^{i+1}, \boldsymbol{J}(\boldsymbol{W}^i - \boldsymbol{W}^{i+1})\rangle \\
&= \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2.
\end{aligned}$$

As a result, (4.29) can be rewritten as

$$\langle \boldsymbol{W}^i - \boldsymbol{W}^+, \boldsymbol{W}^i - \boldsymbol{W}^{i+1}\rangle_J \geq \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 + \|\boldsymbol{\Sigma}^{i+1} - \hat{\boldsymbol{\Sigma}}^+\|_F^2 - \langle\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1}, \boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1}\rangle. \tag{4.30}$$

Note a fact that

$$\|\boldsymbol{W}^+ - \boldsymbol{W}^{i+1}\|_J^2 = \|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2 - 2\langle\boldsymbol{W}^+ - \boldsymbol{W}^i, \boldsymbol{W}^{i+1} - \boldsymbol{W}^i\rangle_J + \|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2.$$

Therefore,

$$\begin{aligned}
&\|\boldsymbol{W}^{+}-\boldsymbol{W}^{i}\|_{J}^{2}-\|\boldsymbol{W}^{+}-\boldsymbol{W}^{i+1}\|_{J}^{2}\\
&=2\langle\boldsymbol{W}^{+}-\boldsymbol{W}^{i},\boldsymbol{W}^{i+1}-\boldsymbol{W}^{i}\rangle_{J}-\|\boldsymbol{W}^{i}-\boldsymbol{W}^{i+1}\|_{J}^{2}\\
&\geq 2\|\boldsymbol{W}^{i}-\boldsymbol{W}^{i+1}\|_{J}^{2}+2\|\boldsymbol{\Sigma}^{i+1}-\hat{\boldsymbol{\Sigma}}^{+}\|_{F}^{2}-2\langle\boldsymbol{\Lambda}^{i}-\boldsymbol{\Lambda}^{i+1},\boldsymbol{\Sigma}^{i}-\boldsymbol{\Sigma}^{i+1}\rangle-\|\boldsymbol{W}^{i}-\boldsymbol{W}^{i+1}\|_{J}^{2}\\
&=\|\boldsymbol{W}^{i}-\boldsymbol{W}^{i+1}\|_{J}^{2}+2\|\boldsymbol{\Sigma}^{i+1}-\hat{\boldsymbol{\Sigma}}^{+}\|_{F}^{2}+2\langle\boldsymbol{\Lambda}^{i+1}-\boldsymbol{\Lambda}^{i},\boldsymbol{\Sigma}^{i}-\boldsymbol{\Sigma}^{i+1}\rangle. \qquad (4.31)
\end{aligned}$$

Hence, next we only need to show $\langle\boldsymbol{\Lambda}^{i+1}-\boldsymbol{\Lambda}^{i},\boldsymbol{\Sigma}^{i}-\boldsymbol{\Sigma}^{i+1}\rangle\geq 0$. Now replacing i instead of $i+1$ in (4.23) and (4.24) yields

$$(-\boldsymbol{\Sigma}^{i}+\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_{k}-\boldsymbol{\Lambda}^{i})_{jl}\in\lambda\partial|\boldsymbol{\Sigma}_{jl}^{i}|,\ j=1,\ldots,p,l=1,\ldots,p,\text{and } j\neq l, \qquad (4.32)$$

and

$$(\boldsymbol{\Sigma}^{i}-\frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_{k})_{jj}+\boldsymbol{\Lambda}_{jj}^{i}=0,\ j=1,\ldots,p. \qquad (4.33)$$

So (4.23), (4.24), (4.32) and (4.33), together with the monotonically non-decreasing property of $\partial|\cdot|$, imply

$$\langle\boldsymbol{\Sigma}^{i}-\boldsymbol{\Sigma}^{i+1},\boldsymbol{\Lambda}^{i+1}-\boldsymbol{\Lambda}^{i}+\boldsymbol{\Sigma}^{i+1}-\boldsymbol{\Sigma}^{i}\rangle\geq 0. \qquad (4.34)$$

After a simple algebra of (4.34), we have

$$\langle\boldsymbol{\Sigma}^{i}-\boldsymbol{\Sigma}^{i+1},\boldsymbol{\Lambda}^{i+1}-\boldsymbol{\Lambda}^{i}\rangle\geq\|\boldsymbol{\Sigma}^{i+1}-\boldsymbol{\Sigma}^{i}\|_{F}^{2}\geq 0.$$

Hence the last two terms on the right hand side of (4.31) are both non-negative, which proves Lemma 1.

Proof of Theorem 1

According to Lemma 1, we have

(a) $\|\boldsymbol{W}^i - \boldsymbol{W}^{i+1}\|_J^2 \to 0$, as $i \to +\infty$;

(b) $\|\boldsymbol{W}^+ - \boldsymbol{W}^i\|_J^2$ is non-increasing and thus bounded.

The result (a) indicates that $\boldsymbol{\Sigma}^i - \boldsymbol{\Sigma}^{i+1} \to 0$ and $\boldsymbol{\Lambda}^i - \boldsymbol{\Lambda}^{i+1} \to 0$. Based on (4.11) it is easy to see that $\boldsymbol{\Phi}^i - \boldsymbol{\Sigma}^i \to 0$. On the other hand, (b) indicates that $\boldsymbol{W}^i$ lies in a compact region. Accordingly, there exists a subsequence $\boldsymbol{W}^{i_j}$ of $\boldsymbol{W}^i$ such that $\boldsymbol{W}^{i_j} \to \boldsymbol{W}^* = (\boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}^*)$. In addition, we also have $\boldsymbol{\Phi}^{i_j} \to \boldsymbol{\Phi}^* \triangleq \boldsymbol{\Sigma}^*$. Therefore, $\lim_{i\to\infty}(\boldsymbol{\Sigma}^i, \boldsymbol{\Phi}^i, \boldsymbol{\Lambda}^i) = (\boldsymbol{\Sigma}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Lambda}^*)$.

Next we show that $(\boldsymbol{\Sigma}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Lambda}^*)$ is an optimal solution of (4.6). By letting $i \to +\infty$ in (4.23), (4.24) and (4.17), we have

$$(-\boldsymbol{\Sigma}^* + \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Lambda}^*)_{jl} \in \lambda\partial|\boldsymbol{\Sigma}^*_{jl}|,\ j = 1, \ldots, p, l = 1, \ldots, p, \text{and } j \neq l, \tag{4.35}$$

$$(\boldsymbol{\Sigma}^* - \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\Sigma}}_k)_{jj} + \boldsymbol{\Lambda}^*_{jj} = 0,\ j = 1, \ldots, p, \tag{4.36}$$

and

$$\langle \boldsymbol{\Lambda}^*, \boldsymbol{\Phi} - \boldsymbol{\Phi}^* \rangle \leq 0,\ \forall \boldsymbol{\Phi} \succeq \nu \boldsymbol{I}. \tag{4.37}$$

(4.35), (4.36) and (4.37), together with $\boldsymbol{\Phi}^* = \boldsymbol{\Sigma}^*$, imply that $(\boldsymbol{\Sigma}^*, \boldsymbol{\Phi}^*, \boldsymbol{\Lambda}^*)$ is an optimal solution of $\arg\min L(\boldsymbol{\Sigma}, \boldsymbol{\Phi}; \boldsymbol{\Lambda})$ in (4.8). Hence, we prove that the sequence produced by Algorithm 4 from any starting point converges to an optimal minimizer of (4.8).

Chapter 5 General Conclusion and Future Work

5.1 Conclusion

In this thesis, I develop an order-invariant Cholesky decomposition model for the covariance and inverse covariance matrices estimation. The proposed estimators are obtained by efficiently assembling a set of multiple estimates of Cholesky factor $\boldsymbol{T}$ and $\boldsymbol{D}$ under different orders of variables, thus solving the order issue in the modified Cholesky decomposition approach.

In Chapter 2, combining the modified Cholesky decomposition approach with GARCH model, the proposed order-invariant methodology is able to provide both accurate estimation and prediction of the covariance matrix in the analysis of multivariate financial time series. Based on the modified Cholesky decomposition of a covariance matrix, one can obtain an ensemble estimate of covariance matrix by taking advantage of multiple estimates of Cholesky factor matrix under different orders of random variables. Log-GARCH model is employed to model the variances, since the variances are often time dependent in the time series data. Hence, the proposed model can be applied to a wide range of time-related data such as medical record, therapy trials, geographical information systems and climate study. Three applications of stock data sets with growing number of variables up to 200 illustrate the merits of the proposed order-invariant Cholesky-log-GARCH model, and show that it also works well in the high-dimensional settings.

In Chapter 3, I propose an order-invariant sparse estimator for an inverse covariance matrix in high-dimensional cases. The approach takes the modified Cholesky decomposition of an inverse covariance matrix as a launching point, and reasonably assembles a set of estimates of Cholesky factor matrix $\boldsymbol{T}$ and $\boldsymbol{D}$ obtained from different orders of variables. I adopt the hard thresholding technique to encourage the sparse structures in the resulting estimator. The proposed methodology can also be extended to the classification problems, such as applications of biomedical imaging, signal processing and medical diagnosis. The performance of proposed

model is examined by the simulation studies and analysis of three real data sets. Results of the simulation studies and applications demonstrate that the proposed method generally outperforms other conventional approaches.

In Chapter 4, I investigate the sparse estimation of high-dimensional covariance matrix. The proposed approach takes advantage of a number of covariance matrix estimates $\hat{\boldsymbol{\Sigma}}_k$, $k = 1, 2, \ldots, M$, obtained by the modified Cholesky decomposition using different orders of variables, and considers the estimator as the "center" of $\hat{\boldsymbol{\Sigma}}_k$. In order to simultaneously achieve the positive definiteness and efficiently exploit the sparsity in the estimated covariance matrix, the positive definite constraint $\boldsymbol{\Sigma} \succeq \nu \boldsymbol{I}$ and L_1 regularization are added to the objective function. An efficient algorithm is derived to solve this optimization problem and its convergence properties are established. The finite-sample performance of the proposed method is demonstrated by both simulations and a real data example.

5.2 Future Work

The order-invariant modified Cholesky decomposition model for estimating the covariance matrix and its inverse has been demonstrated through the thesis. However, in some applications, the data may be invariably heavy-tailed and include unexpected outliers. Hence the robustness of the modified Cholesky decomposition method is an important concern. This is can be addressed by considering the ensemble estimator using element-wise median of $\hat{\boldsymbol{T}}_k$ and $\hat{\boldsymbol{D}}_k$ instead of taking their averages. I have applied this idea to the 12 U.S. bluechips data set, resulting in a good performance regarding Δ_3 and MSE losses. Further investigation are still needed.

Additionally, when I obtain the multiple estimates of factor matrices $\boldsymbol{T}$ and $\boldsymbol{D}$ under each order of variable in the MCD, M permutations are randomly generated and then used. However, in practice, the variables may have relations among themselves, i.e. causal relationship, which means that some orders of variables are meaningful and reflect such relations, while others not. This is can be clearly seen by comparing the ORIG with BIC and BPA methods

in Chapter 2. Hence, ruling out the meaningless orders and only using the meaningful orders of variables would definitely improve the performance of the proposed method. How to implement this idea in the real data needs further study.

Moreover, when I estimate the covariance matrices for the multivariate financial time series using the order-invariant Cholesky-log-GARCH model in Chapter 2, the appropriateness of using log-GARCH (u, v) with $u = 1$ and $v = 1$ is verified by the lag-scatter plots of innovation variances. Hence, it will be interesting to investigate how to develop an automatic scheme to properly choose the values of u and v in a log-GARCH (u, v) model.

Finally, in the prediction Algorithm 2 of Chapter 2, the prediction results are not unique, as they are influenced by the randomness generated by $\eta_{n+1} \sim \mathcal{N}(0, 1)$. Here other distributions may also be used, such as t-distribution. Therefore, how to choose a proper distribution, or how to develop a framework to make prediction performance of the proposed method more stable will be an open topic for the future research.

References

Alexander, C. (2001). Orthogonal GARCH. *Mastering Risk*, **2**, 21–38.

Arakelian, V., and Dellaportas, P. (2012). Contagion Determination via Copula and Volatility Threshold Models. *Quantitative Finance*, **12(2)**, 295–310.

Ardia, D., and Hoogerheide, L. F. (2010). Efficient Bayesian Estimation and Combination of GARCH-Type Models. *Rethinking Risk Measurement and Reporting: Examples and Applications from Finance*, **2**, 1–22.

Aubry, A., De Maio, A., Pallotta, L., and Farina, A. (2012). Maximum Likelihood Estimation of a Structured Covariance Matrix with a Condition Number Constraint. *Signal Processing, IEEE Transactions on*, **60(6)**, 3004–3021.

Ausín, M. C., Galeano, P., and Ghosh, P. (2014). A Semiparametric Bayesian Approach to the Analysis of Financial Time Series with Applications to Value at Risk Estimation. *European Journal of Operational Research*, **232(2)**, 350–358.

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19(7)**, 711–720.

Bera, A. K., and Kim, S. (2002). Testing Constancy of Correlation and Other Specifications of the BGARCH Model with an Application to International Equity Returns. *Journal of Empirical Finance*, **9(2)**, 171–195.

Berre, L. (2000). Estimation of Synoptic and Mesoscale Forecast Error Covariances in a Limited-area Model. *Monthly Weather Review*, **128(3)**, 644–667.

Bhadra, A., and Mallick, B. K. (2013). Joint High-dimensional Bayesian Variable and Covariance Selection with an Application to eQTL Analysis. *Biometrics*, **69(2)**, 447–457.

Bickel, P. J., and Levina, E. (2004). Some Theory of Fishers Linear Discriminant Function, Naive Bayes, and Some Alternatives when There Are Many More Variables than Observations. *Bernoulli*, **10**, 989–1010.

Bickel, P. J., and Levina, E. (2008). Covariance Regularization by Thresholding. *The Annals of Statistics*, 2577–2604.

Bickel, P. J., and Levina, E. (2008). Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics*, 199–227.

Bien, J., and Tibshirani, R. J. (2011). Sparse Estimation of a Covariance Matrix. *Biometrika*, **98(4)**, 807–820.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.

Bollerslev, T., Engle, R., and Wooldridge, J. (1988). A Capital Asset Pricing Model with Time-Varying Covariances. *Journal of Political Economy*, **96**, 116–131.

Bollerslev, T. (1990). Modeling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Approach. *Review of Economics and Statistics*, **72**, 498–505.

Burda, M. (2015). Constrained Hamiltonian Monte Carlo in BEKK GARCH with Targeting. *Journal of Time Series Econometrics*, **7(1)**, 95–113.

Cai, T. T., and Yuan, M. (2012). Adaptive Covariance Matrix Estimation through Block Thresholding. *The Annals of Statistics*, **40(4)**, 2014–2042.

Cai, T., Liu, W., and Luo, X. (2011). A Constrained ℓ-1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, **106(494)**, 594–607.

Cai, T. T., Liu, W., and Zhou, H. H. (2016). Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation. *The Annals of Statistics*, **44(2)**, 455–488.

Cai, T. T., Zhang, C. H., and Zhou, H. H. (2010). Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics*, **38(4)**, 2118–2144.

Chang, C., and Tsay, R. S. (2010). Estimation of Covariance Matrix via the Sparse Cholesky Factor with Lasso. *Journal of Statistical Planning and Inference*, **140(12)**, 3858–3873.

Cheng, Y., and Lenkoski, A. (2012). Hierarchical Gaussian Graphical Models: Beyond Reversible Jump. *Electronic Journal of Statistics*, **6**, 2309–2331.

Danaher, P., Wang, P., and Witten, D. M. (2014). The Joint Graphical Lasso for Inverse Covariance Estimation across Multiple Classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76(2)**, 373–397.

Dellaportas, P., and Pourahmadi M. (2012). Cholesky-GARCH Models with Applications to Finance. *Statistics and Computing*, **22**, 849–855.

Deng, X., and Tsui, K. W. (2013). Penalized Covariance Matrix Estimation Using a Matrix-Logarithm Transformation. *Journal of Computational and Graphical Statistics*, **22**, 494–512.

Deng, X., and Yuan, M. (2009). Large Gaussian Covariance Matrix Estimation with Markov Structure, *Journal of Computational and Graphical Statistics*, **18(3)**, 640–657.

Dey, D. K., and Srinivasan, C. (1985). Estimation of a Covariance Matrix under Steins Loss. *The Annals of Statistics*, **13**, 1581–1591.

Drton, M., and Perlman, M. D. (2008). A SINful Approach to Gaussian Graphical Model Selection. *Journal of Statistical Planning and Inference*, **138(4)**, 1179–1200.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, **97(457)**, 77–87.

Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, **50(4)**, 987–1007.

Engle, R., and Kroner, K.F. (1995). Multivariate Simultaneous Generalized ARCH. *Econometric Theory*, **11**, 122–150.

Engle, R., and Mezrich, J. (1996). GARCH for Groups. *Risk*, **9(8)**, 36–40.

Engle, R. (2002). Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. *Journal of Business and Economic Statistics*, **20**, 339–350.

Fan, J., and Fan, Y. (2008). High-Dimensional Classification Using Features Annealed Independence Rules. *The Annals of Statistics*, **36**, 2605–2637.

Fan, J., Fan, Y., and Lv, J. (2008). High Dimensional Covariance Matrix Estimation Using a Factor Model. *Journal of Econometrics*, **147**, 186–197.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75(4)**, 603–680.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7(2)**, 179–188.

Francq, C., and Zakoïan J.-M, (2010). *GARCH Models*. John Wiley & Sons, Chichester, UK.

Francq, C., Wintenberger, O., and Zakoian, J. (2013). GARCH Models without Positivity Constraints: Exponential or Log GARCH? *Journal of Econometrics*, **177**, 34–46.

Friedman, J. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84**, 165–175.

Friedman, J., Hastie, T., and Tibshirani, T. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**, 432–441.

Furrer, R., and Bengtsson, T. (2007). Estimation of High-dimensional Prior and Posterior Covariance Matrices in Kalman Filter Variants. *Journal of Multivariate Analysis*, **98(2)**, 227–255.

Geweke, J. (1986). Modeling the Persistence of Conditional Variances: A Comment. *Econometric Review*, **5**, 57–61.

Glaab, E., Bacardit, J., Garibaldi, J. M., and Krasnogor, N. (2012). Using Rule-based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data. *PloS one*, **7(7)**, e39932.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized Linear Discriminant Analysis and Its Application in Microarrays. *Biostatistics*, **8**, 86–100.

Haff, L. R. (1991). The Variational Form of Certain Bayes Estimators. *The Annals of Statistics*, **19**, 1163–1190.

Howland, P., and Park, H. (2004). Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 995–1006.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance Matrix Selection and Estimation via Penalised Normal Likelihood. *Biometrika*, **93**, 85–98.

Jacquier, E., and Polson, N. G. (2012). Asset Allocation in Finance: A Bayesian Perspective. *Bayesian Theory and Applications*, **25**, 501–516.

Jensen, M. J., and Maheu, J. M. (2013). Bayesian Semiparametric Multivariate GARCH Modeling. *Journal of Econometrics*, **176(1)**, 3–17.

Johnstone, I. M. (2001). On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics*, **29(2)**, 295–327.

Jolliffe, I. (2002). *Principal Component Analysis*. John Wiley & Sons, Chichester, UK.

Krim, H., and Viberg, M. (1996). Two Decades of Array Signal Processing Research: the Parametric Approach. *IEEE Signal Processing Magazine*, **13(4)**, 67–94.

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *The annals of mathematical statistics*, **22(1)**, 79–86.

Lam, C., and Fan, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation. *The Annals of Statistics*, **37**, 4254–4278.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization Transfer Using Surrogate Objective Functions. *Journal of computational and graphical statistics*, **9(1)**, 1–20.

Ledoit, O., Santa-Clara, P., and Wolf, M. (2004). Flexible Multivariate GARCH Modeling with an Application to International Stock Markets. *Review of Economics and Statistics*, **85**, 735–747.

Ledoit, O., and Wolf, M. (2003). Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection. *Journal of Empirical Finance*, **10(5)**, 603–621.

Ledoit, O., and Wolf, M. (2004). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Levina, E., Rothman, A., and Zhu, J. (2008). Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty. *The Annals of Applied Statistics*, **2(1)**, 245–263.

Liu, W. (2013). Gaussian Graphical Model Estimation with False Discovery Rate Control. *The Annals of Statistics*, **41(6)**, 2948–2978.

Lopes, H., McCullogh, R., and Tsay R. (2012). Cholesky Stochastic Volatility Models for High-Dimensional Time Series. *Technical Report*.

Lounici, K. (2014). High-dimensional Covariance Matrix Estimation with Missing Observations. *Bernoulli*, **20(3)**, 1029–1058.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York.

Meinshausen, N., and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, **34**, 1436–1462.

Mohammadi, A., and Wit, E. C. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, **10(1)**, 109–138.

Pedeli, X., Fokianos, K., and Pourahmadi, M. (2015). Two Cholesky-Log-GARCH Models for Multivariate Volatilities. *Statistical Modelling*, **15(3)**, 233–255.

Perron, F. (1992). Minimax Estimators of a Covariance Matrix. *Journal of Multivariate Analysis*, **43**, 16–28.

Pourahmadi, M. (1999). Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation. *Biometrika*, **86**, 677–690.

Pourahmadi, M. (2000). Maximum Likelihood Estimation of Generalised Linear Models for Multivariate Normal Covariance Matrix. *Biometrika*, **87(2)**, 425–435.

Pourahmadi, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. Wiley, New York.

Pourahmadi, M., Daniels, M. J., and Park, T. (2007). Simultaneous Modelling of the Cholesky Decomposition of Several Covariance Matrices. *Journal of Multivariate Analysis*, **98(3)**, 568–587.

Quinlan, R. C. (1993). *4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc. San Francisco, USA.

Rapisarda, F., Brigo, D., and Mercurio, F. (2007). Parameterizing Correlations: a Geometric Interpretation. *IMA Journal of Management Mathematics*, **18(1)**, 55–73.

Rajaratnam, B., and Salzman, J. (2013). Best Permutation Analysis. *Journal of Multivariate Analysis*, **121**, 193–223.

Raskutti, G., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2008). Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1 Regularized MLE. *In Advances in Neural Information Processing Systems*, 1329–1336.

Rebonato, R., and Jäckel, P. (2000). The most General Methodology for Creating a Valid Correlation Matrix for Risk Management and Option Pricing Purposes. *Journal of Risk*, **2**, 17–27.

Rocha, G. V., Zhao, P., and Yu, B. (2008). A Path Following Algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation. *Technical Report*.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008). Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*, **2**, 494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized Thresholding of Large Covariance Matrices. *Journal of the American Statistical Association*, **104(485)**, 177–186.

Rothman, A. J., Levina, E., and Zhu, J. (2010). A New Approach to Cholesky-based Covariance Regularization in High Dimensions. *Biometrika*, **97(3)**, 539–550.

Rudelson, M. (1999). Random Vectors in the Isotropic Position. *Journal of Functional Analysis*, **164(1)**, 60–72.

Sajid, I., Ahmed, M. M., and Taj, I. (2009). Time Efficient Face Recognition Using Stable Gram-Schmidt Orthonormalization. *International Journal of Signal and Image Processing and Pattern Recognition*, **1(2)**. 35–48.

Schäfer, J., and Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, **4(1)**, 1175–1189.

Scutari, M. (2013). On the Prior and Posterior Distributions Used in Graphical Modelling. *Bayesian Analysis*, **8(3)**, 505–532.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse Linear Discriminant Analysis by Thresholding for High Dimensional Data. *The Annals of Statistics*, **39(2)**, 1241–1265.

Smith, M., and Kohn, R. (2002). Parsimonious Covariance Matrix Estimation for Longitudinal Data. *Journal of the American Statistical Association*, **97(460)**, 1141–1153.

Sun, T., and Zhang, C. H. (2012). Comment: Minimax Estimation of Large Covariance Matrices under ℓ-1 Norm. *Statistical Sinica*, **22**, 1354–1358.

Sun, T., and Zhang, C. H. (2013). Sparse Matrix Inversion with Scaled Lasso. *The Journal of Machine Learning Research*, **14(1)**, 3385–3418.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.

Trbovic, N., Smirnov, S., Zhang, F., and Brschweiler, R. (2004). Covariance NMR Spectroscopy by Singular Value Decomposition. *Journal of Magnetic Resonance*, **171(2)**, 277–283.

Tsay, R. S. (2014). *Multivariate Time Series Analysis: With R and Financial Applications.* John Wiley & Sons, Chichester, UK.

Tse, Y. K. (2000). A Test for Constant Correlations in a Multivariate GARCH Model. *Journal of Econometric*, **98**, 107–127.

Tse, Y. K., and Tsui, A.K. (2002). A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-Varying Correlations. *Journal of Business and Economic Statistics*, **20**, 351–362.

Tsui, A., and Yu, Q. (1999). Constant Conditional Correlation in a Bivariate GARCH Model: Evidence from the Stock Market in China. *Mathematics and Computers in Simulation*, **48**, 503–509.

Van der Weide, R. (2002). Go-Garch: A Multivariate Generalized Orthogonal GARCH Model. *Journal of Applied Econometrics*, **58**, 549–564.

Varoquaux, G., Gramfort, A., Poline, J. B., and Thirion, B. (2010). Brain Covariance Selection: Better Individual Functional Connectivity Models Using Population Prior. *In Advances in Neural Information Processing Systems*, 2334–2342.

Virbickaitė, A., Ausín, M. C., and Galeano, P. (2014). A Bayesian Non-Parametric Approach to Asymmetric Dynamic Conditional Correlation Model with Application to Portfolio Selection. *Computational Statistics and Data Analysis, In Press.*

Vrontos, I.D., Dellaportas, P., and Politis, D.N. (2003). A Full-Factor Multivariate GARCH Model. *Econmetrics Journal*, **6**, 312–334.

Wagaman, A. S., and Levina, E. (2009). Discovering Sparse Covariance Structures with the Isomap. *Journal of Computational and Graphical Statistics*, **18(3)**, 551–572.

Wang, H. (2012). Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis*, **7(4)**, 867–886.

Wang, H., and Pillai, N. S. (2013). On a Class of Shrinkage Priors for Covariance Matrix Estimation. *Journal of Computational and Graphical Statistics*, **22(3)**, 689–707.

Witten, D. M., and Tibshirani, R. (2009). Covariance-regularized Regression and Classification for High Dimensional Problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71(3)**, 615–636.

Won, J. H., Lim, J., Kim, S. J., and Rajaratnam, B. (2013). Condition - number - regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75(3)**, 427–450.

Wu, W. B., and Pourahmadi, M. (2003). Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data. *Biometrika*, **90(4)**, 831–844.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite L_1-penalized Estimation of Large Covariance Matrices. *Journal of the American Statistical Association*, **107(500)**, 1480–1491.

Xue, L., and Zou, H. (2012). Regularized Rank-based Estimation of High-dimensional Nonparanormal Graphical Models. *The Annals of Statistics*, **40(5)**, 2541–2571.

Ye, J. (2005). Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems. *Journal of Machine Learning Research*, 483–502.

Yin, J., and Li, H. (2011). A Sparse Conditional Gaussian Graphical Model for Analysis of Genetical Genomics Data. *The Annals of Applied Statistics*, **5(4)**, 2630–2650.

Yuan, M., and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, **94**, 19–35.

Yuan, M. (2008). Efficient Computation of the ℓ_1 Regularized Solution Path in Gaussian Graphical Models. *Journal of Computational and Graphical Statistics*, **17**, 809–826.

Yuan, M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, **11**, 2261–2286.

Zhang, W., and Leng, C. (2012). A Moving Average Cholesky Factor Model in Covariance Modelling for Longitudinal Data. *Biometrika*, **99(1)**, 141–150.