Covariance Estimation via the Modified Cholesky Decomposition

43

Xiaoning Kang, Zhiyang Zhang, and Xinwei Deng

Contents

43.1 Introduction ... 887

43.2 Review of Modified Cholesky Decomposition ... 888
43.2.1 MCD for Precision Matrix Estimation ... 888
43.2.2 MCD for Covariance Matrix Estimation ... 889
43.2.3 MCD for Banded Matrix Estimation ... 890
43.2.4 Adaptive Banding in the MCD ... 890

43.3 Ordering Issue in the MCD ... 892

43.4 Real Applications ... 893
43.4.1 MCD for Varying Covariance Matrix Estimation in Multivariate Time Series ... 893
43.4.2 MCD for Portfolio Optimization ... 894
43.4.3 MCD for Linear Discriminant Analysis ... 895

43.5 Numerical Study ... 896

43.6 Discussion ... 897

Appendix ... 897

References ... 899

Abstract

In many engineering applications, estimation of covariance and precision matrices is of great importance, helping researchers understand the dependency and conditional dependency between variables of interest. Among various matrix estimation methods, the modified Cholesky decomposition is a commonly used technique. It has the advantage of transforming the matrix estimation task into solving a sequence of regression models. Moreover, the sparsity on the regression coefficients implies certain sparse structure on the covariance and precision matrices. In this chapter, we first overview the Cholesky-based covariance and precision matrices estimation. It is known that the Cholesky-based matrix estimation depends on a prespecified ordering of variables, which is often not available in practice. To address this issue, we then introduce several techniques to enhance the Cholesky-based estimation of covariance and precision matrices. These approaches are able to ensure the positive definiteness of the matrix estimate and applicable in general situations without specifying the ordering of variables. The advantage of Cholesky-based estimation is illustrated by numerical studies and several real-case applications.

Keywords

Banded matrix · Cholesky-based · Linear discriminant analysis · Ordering of variables · Positive definite · Portfolio optimization · Sparsity

43.1 Introduction

Estimation of covariance and precision matrices is of fundamental importance in the multivariate analysis [5]. It has received wide attentions of scholars in various engineering applications such as additive manufacturing [30, 31], biomedical engineering [17], and tissue engineering [39]. The resultant matrix estimation has also been widely used in various statistical methods and applications. For example, dimension reduction via the principal component analysis (PCA) usually relies on the estimation of covariance matrix. In the classification problem, the linear discriminant analysis (LDA) constructs the classification rule through the precision matrix. In the financial area, the portfolio optimization uses the precision matrix to minimize the portfolio risk. In signal

X. Kang
International Business College and Institute of Supply Chain Analytics, Dongbei University of Finance and Economics, Dalian, China
e-mail: kangxiaoning@dufe.edu.cn; xiaoningmike@126.com

Z. Zhang · X. Deng (✉)
Department of Statistics, Virginia Tech, Blacksburg, VA, USA
e-mail: zhiyangz@vt.edu; xdeng@vt.edu

H. Pham (ed.), *Springer Handbook of Engineering Statistics*, Springer Handbooks,
https://doi.org/10.1007/978-1-4471-7503-2_43

processing, the covariance matrix helps to distinguish between signals and noise. The covariance and precision matrices also arise in the graphical models, multivariate volatility, weather forecasting, social network, fMRI analysis, and so forth.

For the matrix estimation, a desirable property is the sparsity in the sense that some elements in an estimated matrix are zeros [13, 16, 20, 26, 37]. A sparse covariance matrix estimate is useful for the subsequent statistical analysis, such as inferring the correlation pattern among the predictor variables. A sparse precision matrix estimate often implies the conditional independence among the corresponding predictor variables. Therefore, a variety of classical approaches has been developed in literature for estimating the covariance and precision matrices with particular interest of sparse structure. Yuan and Lin [38] introduced the graphical Lasso (Glasso) model, which gives a sparse precision matrix estimate by imposing an L_1 penalty on the negative log-likelihood function. Bickel and Levina [1] proposed to threshold the small elements of the sample covariance matrix directly to zeroes with a large number of predictor variables. More work on the sparse estimation of the covariance and precision matrices can be found in [3–5, 10, 15, 21, 27, 32, 34, 36], among others.

Another important property of covariance and precision matrices is that they need to be positive definite for proper inference. The modified Cholesky decomposition (MCD) is a popular and commonly used technique for the matrix estimation, which ensures the estimated matrix to be positive definite. The MCD provides an unconstrained and statistically interpretable parameterization of a matrix by sequentially regressing the variables in a random variable vector. This method reduces the challenge of estimating a covariance or precision matrix into an easier task of solving a sequence of linear regression models. However, it is known that the Cholesky-based matrix estimation relies on the preknowledge of the ordering of variables. When the variable ordering is not naturally available, we need to consider several techniques to enhance the Cholesky-based estimation of covariance and precision matrices. Such techniques are able to ensure the positive definiteness of the matrix estimate and applicable in general situations without specifying the ordering of variables.

The remaining of this chapter is organized as follows. Sect. 43.2 provides a comprehensive overview of the MCD for the estimation of covariance and precision matrices, respectively. We then point out the variable ordering issue in the MCD and consider a couple of techniques to address the variable ordering issue in Sect. 43.3. Several real-data applications and numerical examples are presented in Sects. 43.4 and 43.5 to examine the performances of the Cholesky-based matrix estimates. We conclude this chapter with some discussion in Sect. 43.6.

43.2 Review of Modified Cholesky Decomposition

In this section, we review the modified Cholesky decomposition (MCD) in detail for estimating the covariance and precision matrices, respectively. As proposed by Pourahmadi [24], the MCD approach is statistically meaningful and guarantees the positive definiteness of a matrix estimate. The sparsity can be encouraged in the estimated matrix via the MCD technique.

43.2.1 MCD for Precision Matrix Estimation

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 $\boldsymbol{\Sigma}$. Denote by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the n independent and identically distributed observations following a multivariate distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}^{-1}$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix. The key idea of the MCD is that $\boldsymbol{\Omega}$ can be diagonalized 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

$$\begin{aligned} X_j &= \sum_{i=1}^{j-1} a_{ji} X_i + \epsilon_j \\ &= \boldsymbol{Z}_j^T \boldsymbol{a}_j + \epsilon_j, \end{aligned} \tag{43.1}$$

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 $Var(\epsilon_j) = d_j^2$. Hence, a lower triangular matrix $\boldsymbol{A}$ can be formed as

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

which contains all the regression coefficients in (43.1). Also define

$$d_j^2 = Var(\epsilon_j) = \begin{cases} Var(X_1), & j = 1, \\ Var(X_j - \boldsymbol{Z}_j^T \boldsymbol{a}_j), & j = 2, \ldots, p. \end{cases} \tag{43.2}$$

Let $\boldsymbol{D} = diag(d_1^2, \ldots, d_p^2)$ be the diagonal covariance matrix of the vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)'$. Then, move the term $\sum_{k=1}^{j-1} a_{jk} X_k$ in Eq. (43.1) to the left side and rewrite it in the following matrix form:

$$\boldsymbol{\epsilon} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X} = \boldsymbol{T}\boldsymbol{X}, \tag{43.3}$$

where $\boldsymbol{I}$ represents the $p \times p$ identity matrix and $\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{A}$ is a unit lower triangular matrix having ones on its diagonal. The matrices $\boldsymbol{T}$ and $\boldsymbol{D}$ are called the Cholesky factor matrices. By taking the variance operator on both sides of Eq. (43.3), one can easily obtain

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

and thus

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}. \tag{43.4}$$

As a result, the decomposition (43.1) 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 precision matrix into the task of solving $(p-1)$ linear regression models, which is much easier to implement.

A straightforward estimate $\hat{\boldsymbol{T}}$ of $\boldsymbol{T}$ can be obtained from the least squares estimates of the regression coefficients

$$\hat{\boldsymbol{a}}_j = \arg\min_{\boldsymbol{a}_j} \|\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{a}_j\|_2^2, \ j = 2, \ldots, p, \tag{43.5}$$

where $\boldsymbol{x}^{(j)}$ is the jth column of the data matrix $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\mathbb{Z}^{(j)} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(j-1)})$ stands for 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 (43.2). Because the optimization (43.5) uses the ordinary least squares, this approach of precision matrix estimation is only suitable in low-dimensional settings with the number of predictor variables p smaller than the sample size n. When applying the MCD to the high-dimensional situations where p is close to or even larger than n, the least squares estimation is inaccurate or not available. In such cases, a natural idea is to employ the Lasso regularization [33] to shrink the regression estimates and encourage the sparsity on the Cholesky factor matrix $\boldsymbol{T}$

$$\hat{\boldsymbol{a}}_j = \arg\min_{\boldsymbol{a}_j} \|\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{a}_j\|_2^2 + \lambda_j\|\boldsymbol{a}_j\|_1, \ j = 2, \ldots, p, \tag{43.6}$$

where $\lambda_j \geq 0$ is a tuning parameter and $\|\cdot\|_1$ stands for the vector L_1 norm. Note that the penalty in Eq. (43.6) is often suitable for data with large number of variables such as engineering data, social network data, and imaging data, since their underlying matrix is usually sparse with no specific sparse pattern. The optimization problem (43.6) can be solved by the coordinate descent algorithm [12]. The tuning parameters are determined by the cross-validation scheme for each Lasso regression. In addition, one can alternatively consider the estimation of $\boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ in the Cholesky factor matrix $\boldsymbol{T}$ under a joint fashion as

$$\hat{\boldsymbol{a}}_2, \ldots, \hat{\boldsymbol{a}}_p = \arg\min_{\boldsymbol{a}_2, \ldots, \boldsymbol{a}_p} \sum_{j=2}^{p} \|\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\boldsymbol{a}_j\|_2^2 + \lambda \sum_{j=2}^{p} \|\boldsymbol{a}_j\|_1. \tag{43.7}$$

Such a joint estimation approach is used in [6, 13, 14, 40], among others.

After obtaining $\hat{\boldsymbol{a}}_j$, the lower triangular matrix $\hat{\boldsymbol{T}}$ is established with ones on its diagonal and $\hat{\boldsymbol{a}}_j'$ as its jth row. Meanwhile, the diagonal matrix $\hat{\boldsymbol{D}}$ has its jth diagonal element equal to $\hat{d}_j^2$, where

$$\hat{d}_j^2 = \begin{cases} \widehat{Var}(\boldsymbol{x}^{(1)}), & j = 1, \\ \widehat{Var}(\boldsymbol{x}^{(j)} - \mathbb{Z}^{(j)}\hat{\boldsymbol{a}}_j), & j = 2, \ldots, p, \end{cases}$$

where $\widehat{Var}(\cdot)$ denotes the sample variance. Consequently,

$$\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{T}}'\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{T}} \tag{43.8}$$

is a sparse estimate for the precision matrix $\boldsymbol{\Omega}$.

Remark 1 For the optimization problem in (43.6) with $\lambda_j = 0$ and $n > p$, the estimated precision matrix in (43.8) is equivalent to the inverse of the sample covariance matrix.

43.2.2 MCD for Covariance Matrix Estimation

From (43.4), it is easy to see that a Cholesky-based estimate for a covariance matrix can be expressed as

$$\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}\boldsymbol{D}\boldsymbol{T}'^{-1}. \tag{43.9}$$

The construction of the Cholesky factor matrices $(\boldsymbol{T}, \boldsymbol{D})$ through a series of linear regressions (43.1) is thoroughly discussed in Sect. 43.2.1. Such a covariance matrix estimate via the MCD can perform well if one does not require the estimated covariance matrix to be sparse. In other words, the estimate obtained based on Eq. (43.9) generally does not have sparse structure even though the Cholesky factor matrix $\boldsymbol{T}$ is sparse. This is because the matrix $\boldsymbol{T}^{-1}$ often does not inherit any sparse property from $\boldsymbol{T}$, leading to a dense estimate of covariance matrix $\boldsymbol{\Sigma}$. It is thus not convenient to impose a sparse structure on the estimate of $\boldsymbol{\Sigma}$ via Eq. (43.9).

Alternatively, one can consider a Cholesky-based latent variable regression model, which enables the regularization more easily. Write $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{\epsilon}$, implying the predictor variable X_j is regressed on its previous latent variables $\epsilon_1, \ldots, \epsilon_{j-1}$,

and hence $\boldsymbol{L} = (l_{ji})_{p\times p}$ is a unit lower triangular matrix constructed from the regression coefficients of the following sequential regressions:

$$X_j = \boldsymbol{l}_j^T \boldsymbol{\epsilon} = \sum_{i=1}^{j-1} l_{ji}\epsilon_i + \epsilon_j, \quad j = 2, \ldots, p, \tag{43.10}$$

where $\boldsymbol{l}_j = (l_{ji})$ is the jth row of $\boldsymbol{L}$. Here $l_{jj} = 1$ and $l_{ji} = 0$ for $i > j$. This decomposition is interpreted as resulting from a different 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. As a result, this form of the MCD by the latent variable regression model provides a re-parameterization of the covariance matrix

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

This decomposition connects the covariance matrix $\boldsymbol{\Sigma}$ with linear regressions (43.10), such that the Lasso penalty can be imposed on the coefficients of the linear regressions, thus conveniently encouraging the sparsity in the Cholesky factor matrix $\boldsymbol{L}$ and the estimated covariance matrix.

Denote by $\boldsymbol{x}^{(j)}$ the jth column of the data matrix $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. Let $\boldsymbol{e}^{(j)}$ represent the residuals for the linear regression when $\boldsymbol{x}^{(j)}$ is treated as the response data, $j \geq 2$, and $\boldsymbol{e}^{(1)} = \boldsymbol{x}^{(1)}$. Let $\mathbb{W}^{(j)} = (\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(j-1)})$ be the matrix containing the first $(j-1)$ residuals. Now we construct the matrix $\boldsymbol{L}$ by employing the Lasso penalty to select important predictor variables

$$\hat{\boldsymbol{l}}_j = \arg\min_{\boldsymbol{l}_j} \|\boldsymbol{x}^{(j)} - \mathbb{W}^{(j)}\boldsymbol{l}_j\|_2^2 + \lambda_j\|\boldsymbol{l}_j\|_1, \; j = 2, \ldots, p, \tag{43.11}$$

where $\lambda_j \geq 0$ is a tuning parameter and selected by cross-validation. $\boldsymbol{e}^{(j)} = \boldsymbol{x}^{(j)} - \mathbb{W}^{(j)}\hat{\boldsymbol{l}}_j$ is used to construct the residuals for the last column of $\mathbb{W}^{(j+1)}$. Then the element d_j^2 of the diagonal matrix $\boldsymbol{D} = diag(d_1^2, \ldots, d_p^2)$ is estimated as the sample variance of $\boldsymbol{e}^{(j)}$

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

Consequently,

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{D}}\hat{\boldsymbol{L}}' \tag{43.12}$$

is a sparse estimate for the covariance matrix $\boldsymbol{\Sigma}$. It is worth pointing out that one can also consider estimating $\boldsymbol{l}_2, \ldots, \boldsymbol{l}_p$ in a joint fashion similar as in (43.7). But the estimation procedure will become more complicated since it involves latent variables.

Remark 2 For the optimization problem in (43.11) with $\lambda_j = 0$ and $n > p$, the estimated covariance matrix in (43.12) is equivalent to the sample covariance matrix.

43.2.3 MCD for Banded Matrix Estimation

In some applications such as assimilation [22] and social unrest study [35], the variables are strongly correlated with the ones that are close to them in the ordering, and the variables far apart in the ordering are weakly correlated. For example, in the random variable vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$, the variable X_3 may have a strong correlation with its neighbor variables X_1, X_2, X_4, and X_5, but it could be weakly correlated with variables X_j for $j > 5$. In this situation, the covariance or precision matrices are usually assumed to have a banded structure. The kth banded structure means that the first $k < p$ sub-diagonals elements of a matrix are non-zeroes with the rest elements being zeroes. Bickel and Levina [2] proposed a banded estimate for the covariance matrix by banding the sample covariance matrix. Note that their estimate cannot guarantee the positive definiteness. The MCD technique can also be used for estimating a banded matrix by banding the Cholesky matrix factor [28]. In this section, we focus on the estimation for banded precision matrix via the MCD. The Cholesky-based estimation for the banded covariance matrix follows the similar principle by considering the banded $\boldsymbol{L}$ in (43.12).

The key idea is to band the Cholesky factor matrix $\boldsymbol{T}$. Recall that in the decomposition (43.1), the predictor variable X_j is regressed on all of its predecessors $X_1, \ldots, X_{j-1}$. To accommodate a banded estimate of precision matrix, each predictor variable X_j is regressed only on its k previous variables $X_{j-k}, \ldots, X_{j-1}$, for all $j = 2, \ldots, p$. The index $j-k$ is interpreted to mean $\max(1, j-k)$. Hence, we solve the following $(p-1)$ linear regression models instead of (43.1)

$$X_j = \sum_{i=j-k}^{j-1} a_{ji}X_i + \epsilon_j \tag{43.13}$$

to construct the Cholesky factors $(\boldsymbol{T}, \boldsymbol{D})$, As a result, the unit lower triangular matrix $\boldsymbol{T}$, with ones as its diagonal and $(0, \ldots, 0, a_{j,j-k}, \ldots, a_{j,j-1})$ as its jth row, has a banded structure on its bottom left part. Then the estimate for the precision matrix $\boldsymbol{\Omega} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}$ obtained from (43.13) has a kth banded structure. This approach is designated to estimate a banded matrix and guarantees the positive definiteness of the estimated matrix.

43.2.4 Adaptive Banding in the MCD

Note that the kth banded matrix estimation in Sect. 43.2.3 considers the banding width k to be the same for each vari-

able. In this section, we discuss a more flexible case where the band for each variable to be different. That is, we allow $k = k_j$ in the jth linear regression of (43.13), where the jth variable is regressed on its k_j closest predecessors with k_j depending on j. We call such a procedure as adaptive banding, which is useful when each variable may depend on an unknown number of its predecessors. It preserves sparsity in the resulting estimate of precision matrix and produces a better estimate by being able to adapt to the data. Next we introduce two types of adaptive banding. They use two different techniques to obtain an adaptively banded estimate for the Cholesky factor $\boldsymbol{T}$.

Adaptive Banding: Levina and Zhu [19] applied a nested Lasso penalty imposed on the negative normal log-likelihood function to produce matrix $\boldsymbol{T}$ with adaptive banding k_j. Specially, the negative log-likelihood of the data, up to a constant, is

$$\begin{aligned}\ell(\boldsymbol{\Sigma}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &= n \log |\boldsymbol{\Sigma}| + \sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i \\ &= n \log |\boldsymbol{D}| + \sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{T}' \boldsymbol{D}^{-1} \boldsymbol{T} \boldsymbol{x}_i \\ &= \sum_{j=1}^{p} \ell_j(d_j, \boldsymbol{a}_j, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n),\end{aligned}$$

where $\boldsymbol{a}_j$ stands for the vector of regression coefficients for the jth regression in the MCD and

$$\ell_j(d_j, \boldsymbol{a}_j, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = n \log d_j^2 + \sum_{i=1}^{n} \frac{1}{d_j^2} \left(x_{ij} - \sum_{l=1}^{j-1} a_{jl} x_{il} \right)^2.$$

Minimizing the negative log-likelihood $\ell(\boldsymbol{\Sigma}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is equivalent to minimizing each $\ell_j(d_j, \boldsymbol{a}_j, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Then one can consider to minimize

$$\ell_j(\boldsymbol{\Sigma}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) + J(\boldsymbol{a}_j), \tag{43.14}$$

where $J(\boldsymbol{a}_j)$ is the nested Lasso penalty as

$$J(\boldsymbol{a}_j) = \lambda \left(|a_{j,j-1}| + \frac{|a_{j,j-2}|}{|a_{j,j-1}|} + \frac{|a_{j,j-3}|}{|a_{j,j-2}|} + \cdots + \frac{|a_{j,1}|}{|a_{j,2}|} \right),$$

where $\lambda \geq 0$ is a tuning parameter, and we define $0/0 = 0$. The effect of variable selection by the nested Lasso penalty is that if the lth variable is not included in the jth regression ($a_{jl} = 0$), then all the subsequent variables ($l - 1$ through 1) are also excluded. Hence, the jth linear regression only has $k_j \leq j - 1$ closest predecessors, with values of k_j varying for each regression.

Since the expression in (43.14) is highly nonconvex and nonlinear, an iterative algorithm is developed to minimize it for constructing the Cholesky factor matrices $\boldsymbol{T}$ and $\boldsymbol{D}$. Let $\hat{\boldsymbol{a}}_j^{(m)}$ and $\hat{d}_j^{(m)}$ represent the estimates of $\boldsymbol{a}_j$ and d_j at the mth iteration. Then we repeat the following steps 1 and 2 until convergence.

Step 1: Given $\hat{\boldsymbol{a}}_j^{(m)}$, solve for $\hat{d}_j^{(m)}$

$$(\hat{d}_j^{(m)})^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \sum_{l=1}^{j-1} \hat{a}_{jl}^{(m)} x_{il} \right)^2.$$

Step 2: Given $\hat{\boldsymbol{a}}_j^{(m)}$ and $\hat{d}_j^{(m)}$, solve for $\hat{\boldsymbol{a}}_j^{(m+1)}$. Here we use the local quadratic approximation [9, 13]

$$\left| a_{jl}^{(m+1)} \right| \approx \frac{\left(a_{jl}^{(m+1)} \right)^2}{2 \left| a_{jl}^{(m)} \right|} + \frac{\left| a_{jl}^{(m)} \right|}{2}.$$

Then the estimate of $\boldsymbol{a}_j$ at iteration $m + 1$ is

$$\begin{aligned}\hat{\boldsymbol{a}}_j^{(m+1)} = \arg\min_{\boldsymbol{a}_j} & \frac{1}{(\hat{d}_j^{(m)})^2} \sum_{i=1}^{n} \left(x_{ij} - \sum_{l=1}^{j-1} a_{jl} x_{il} \right)^2 \\ & + \frac{\lambda}{2} \left(\frac{a_{j,j-1}^2}{|\hat{a}_{j,j-1}^{(m)}|} + \sum_{l=1}^{j-2} \frac{a_{jl}^2}{|\hat{a}_{jl}^{(m)}| \cdot |\hat{a}_{j,l+1}^{(m)}|} \right).\end{aligned}$$

This is a quadratic optimization problem, which can be solved in closed form. Note that the algorithm needs an initial value for $\boldsymbol{a}_j$. One could use the least squares estimates when $p < n$. If $p > n$, we initialize with $\hat{a}_{jl}^{(0)} = \hat{a}_{jl}^{\star}$, which are found by regressing X_j on X_l alone, for $l = 1, \ldots, j - 1$.

Forward Adaptive Banding: Instead of imposing a nested Lasso penalty on the likelihood function, Leng and Li [18] proposed a forward adaptive banding approach to determine k_j for each linear regression by minimizing their modified Bayesian information criterion (BIC). Operationally, for the jth variable, we fit $j - 1$ nested linear models by regressing X_j on $X_{j-1}, \ldots, X_{j-k_j}$ for $k_j = 1, \ldots, j - 1$. Then the optimal k_j is chosen to minimize

$$\begin{aligned}\text{BIC} &= n \log |\hat{\boldsymbol{\Sigma}}| + \sum_{i=1}^{n} \boldsymbol{x}_i' \hat{\boldsymbol{T}}' \hat{\boldsymbol{D}}^{-1} \hat{\boldsymbol{T}} \boldsymbol{x}_i + C_n \log(n) \sum_{j=1}^{p} k_j \\ &= \sum_{j=1}^{p} \left\{ n \log \hat{d}_j^2 + \sum_{i=1}^{n} \frac{1}{\hat{d}_j^2} \left(x_{ij} - \sum_{l=1}^{k_j} a_{j,j-l} x_{i,j-l} \right)^2 + C_n \log(n) k_j \right\},\end{aligned} \tag{43.15}$$

for all $k_j \leq \min\{n/(\log n)^2, j-1\}(j=2,\ldots,p)$ with some diverging C_n. The expression in (43.15) implies that the banding width k_j can be chosen separately for each j. The penalty coefficient $C_n \log n$ is set to different values to accommodate the diverging number of predictor variables p. Leng and Li [18] proved that this modified BIC is model selection consistent under some regular conditions. The advantage of this approach is that we determine the value of k_j by only fitting a sequence of linear models. Therefore, it can be easily and efficiently implemented.

43.3 Ordering Issue in the MCD

Although the MCD approach has been widely investigated for the matrix estimation, only a few work contributed to solve a potential problem, the ordering issue. From the decompositions (43.1) and (43.10), it is clear to see that different orderings of the predictor variables $X_1,\ldots,X_p$ used in the MCD would lead to different linear regressions. If a regularization is adopted to shrink the coefficients of the linear regressions, such as the Lasso penalty in (43.6), the Cholesky factor matrices estimates would be different under different variable orderings, thus leading to the different estimates for the covariance and precision matrices. As a statistical decision problem, ordering variables is quite challenging. In practice, the MCD method can be suitably used to estimate a matrix without this ordering issue when the predictor variables have a natural ordering among themselves, as in time series, longitudinal data, or spatial data. However, there are a large number of applications where such a natural ordering is not available or the variable ordering cannot be determined before the analysis, for example, gene expressions, financial, and economic data. In these cases, one may need to first determine a proper ordering among the variables before employing the MCD method. Next, we introduce three commonly used techniques to tackle this issue. In this part, we focus on the ordering issue in the MCD for the estimation of precision matrix. Ordering issue for Cholesky-based estimation of covariance matrix can be addressed in a similar fashion.

BIC: A search algorithm for ordering the predictor variables based on the BIC criterion can be easily implemented. In each step, a new variable is selected into the ordering with the smallest value of BIC when regressing it on the rest of the candidate variables. Specifically, suppose we want to construct an ordering for p variables $X_1,\ldots,X_p$. In the first step, each variable X_j, $j=1,\ldots,p$ is regressed on the rest of variables, producing p values of BIC from p linear regressions. Then the first variable determined into the ordering is the response variable of the linear regression that gives the minimum value of BIC. This selected variable is denoted by X_{i_p}, and it is assigned to the pth position of the ordering. All the variables excluding the selected variable X_{i_p} consist of the candidate set $\mathcal{C}$. In the second step, from the set $\mathcal{C}$, each variable is regressed on the rest of variables, producing $p-1$ values of BIC from $p-1$ linear regressions. Then the second variable, denoted by $X_{i_{p-1}}$, determined into the ordering is the response variable of the linear regression that gives the minimum value of BIC. The variable $X_{i_{p-1}}$ is assigned to the $(p-1)$th position of the ordering. Then all the variables excluding X_{i_p} and $X_{i_{p-1}}$ compose of the candidate set for the next round.

To sum up this procedure, let $\mathcal{C}=\{X_{i_1},\ldots,X_{i_k}\}$ be the candidate set of variables, and there are $p-k$ variables already chosen into the ordering. By regressing each $X_j, j=i_1,\ldots,i_k$ on the rest of the variables in $\mathcal{C}$, the variable corresponding to the minimum BIC value among the k regressions is assigned to the kth position of the ordering. Then the ordering created by this procedure is used for the MCD.

BPA: A popular method to recover the variable ordering for the autoregressive model is the best permutation algorithm (BPA) developed by Rajaratnam and Salzman [25]. It is formulated as a well-defined optimization problem where the optimal ordering is determined as the one minimizing the sum of squared diagonal entries of the Cholesky factor matrix $\boldsymbol{D}$ in the MCD. For the convenience of presentation, define a permutation mapping $\pi:\{1,\ldots,p\}\to\{1,\ldots,p\}$ by

$$(\pi(1),\pi(2),\ldots,\pi(p))\,. \tag{43.16}$$

Denote the corresponding permutation matrix by $\boldsymbol{P}_\pi$ of which the entries in the jth column are all 0 except taking 1 at position $\pi(j)$. Let S_p be the symmetric group of all permutations of the integers $1,\ldots,p$. For a given $\pi\in S_p$, let $\boldsymbol{\Omega}_\pi$ be the precision matrix corresponding to the variable ordering π and

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

be its MCD. Since the conditional variances can be used as natural measures to quantify the extent to which the variability of a random variable is explained by the ones that precede it. Hence, the conditional variances give a sense of closeness between variables. From this viewpoint, the BPA is to find an ordering π^* in S_p to minimize $||\boldsymbol{D}_{\pi^*}||_F^2$, where $||\cdot||_F$ represents the Frobenius norm. Then $\hat{\boldsymbol{\Omega}} = \boldsymbol{P}_{\pi^*}\hat{\boldsymbol{\Omega}}_{\pi^*}\boldsymbol{P}'_{\pi^*} = \boldsymbol{P}_{\pi^*}\hat{\boldsymbol{T}}'_{\pi^*}\hat{\boldsymbol{D}}_{\pi^*}^{-1}\hat{\boldsymbol{T}}_{\pi^*}\boldsymbol{P}'_{\pi^*}$ is an estimate of precision matrix based on BPA. Rajaratnam and Salzman [25] showed the consistency of this approach in recovering the natural order of variables in underlying autoregressive models.

Ordering-averaged method (OAM): To address the problem that the ordering of variables is often not available in practice, we consider a Cholesky-based model averaging idea [40] by averaging a representative set of individual matrix estimates obtained from random permutations of the variable orderings. This method does not require any prior knowledge

of the orderings of variables; hence, it is suitable for the case where there is no natural ordering among the variables, or such an ordering cannot be easily determined. Use the definition of permutation mapping π in (43.16), a precision matrix estimate under π is $\boldsymbol{\Omega}_\pi = \boldsymbol{T}_\pi' \boldsymbol{D}_\pi^{-1} \boldsymbol{T}_\pi$. Transforming back to the original ordering, we can estimate $\boldsymbol{\Omega}$ as

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{P}_\pi \hat{\boldsymbol{\Omega}}_\pi \boldsymbol{P}_\pi' = \boldsymbol{P}_\pi \boldsymbol{T}_\pi' \boldsymbol{D}_\pi^{-1} \boldsymbol{T}_\pi \boldsymbol{P}_\pi'. \tag{43.17}$$

By incorporating several permutations of πs, one can obtain a pool of precision matrix estimates. Taking the average of these estimates leads to an ordering-averaged estimation. In practice, a modest number of permutations are sufficient to serve this purpose. Therefore, we randomly generate multiple permutations π_k, $k = 1, \ldots, M$ and obtain the corresponding estimates $\hat{\boldsymbol{\Omega}}$ in (43.17), denoted by $\hat{\boldsymbol{\Omega}}_k$ for the permutation π_k. The model averaging estimate of $\boldsymbol{\Omega}$ is

$$\hat{\boldsymbol{\Omega}}_{OAM} = \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Omega}}_k. \tag{43.18}$$

Similarly, it is very convenient to apply the OAM for the estimation of the covariance matrix. Combining Eq. (43.9) and the averaging idea of (43.18), we have the model averaging estimate of $\boldsymbol{\Sigma}$ as

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{OAM} &= \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\Sigma}}_k = \frac{1}{M} \sum_{k=1}^{M} \boldsymbol{P}_k \hat{\boldsymbol{\Sigma}}_k \boldsymbol{P}_k' \\ &= \frac{1}{M} \sum_{k=1}^{M} \boldsymbol{P}_k \hat{\boldsymbol{T}}_k^{-1} \hat{\boldsymbol{D}}_k (\hat{\boldsymbol{T}}_k')^{-1} \boldsymbol{P}_k', \end{aligned}$$

where $\hat{\boldsymbol{T}}_k$, $\hat{\boldsymbol{D}}_k$, and $\hat{\boldsymbol{\Sigma}}_k$ represent the estimates of $\boldsymbol{T}$, $\boldsymbol{D}$, and $\boldsymbol{\Sigma}$ under the permutation π_k. According to the finite population sampling survey theory [7], the selection of permutations π_k is not essential when we use a reasonable size M. Although choosing a larger M would further reduce the variability of the OAM estimate, Zheng et al. [40] showed that a modest number $M = 30$ is seen to lead to stable results.

43.4 Real Applications

The covariance and precision matrices have been widely used in various areas such as portfolio selection, risk assessment, principle component analysis, social network, graphical models, classification, and so forth. In this section, real-data examples are used to illustrate the application of the MCD approach for the estimation of covariance and precision matrices.

43.4.1 MCD for Varying Covariance Matrix Estimation in Multivariate Time Series

In the financial management with multivariate time series, a major task is to estimate the time-varying covariance matrices $\{\boldsymbol{\Sigma}_t\}$ based on the (conditionally) independently distributed data $\boldsymbol{x}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_t)$, $t = 1, 2, \ldots, n$. The data $\boldsymbol{x}_t$ can be viewed as the returns of p assets in a portfolio at time t.

By the decomposition (43.9), the estimation of time-varying covariance matrix is given by

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

where the Cholesky factor matrix $\boldsymbol{T} = \boldsymbol{T}_t$ for all $t = 1, 2, \ldots, n$ is assumed to be time-invariant to reduce a large number of parameters. For each element of the diagonal matrix $\boldsymbol{D}_t = \mathrm{diag}(d_{1;t}^2, \ldots, d_{p;t}^2)$, the $\log d_{j;t}^2, j = 1, 2, \ldots, p$, is modeled using the log-GARCH(u, v) defined recursively in time as

$$\begin{aligned} \log d_{j;t}^2 = \beta_0^{(j)} &+ \sum_{i=1}^{v} \left(\alpha_{i+}^{(j)} 1_{\{\epsilon_{j;t-i} > 0\}} + \alpha_{i-}^{(j)} 1_{\{\epsilon_{j;t-i} < 0\}} \right) \log \epsilon_{j;t-i}^2 \\ &+ \sum_{k=1}^{u} \beta_k^{(j)} \log d_{j;t-k}^2. \end{aligned} \tag{43.19}$$

where $1_{\{\cdot\}}$ is the indicator function and $\beta_0^{(j)}$, $\beta_k^{(j)}$, $\alpha_{i+}^{(j)}$, $\alpha_{i-}^{(j)}$ are corresponding coefficients. The quasi-maximum likelihood approach in Francq and Zakoïan [11] is used to fit the model (43.19). Therefore, we combine the MCD method and the log-GARCH model to analyze a stock return data from the Standard and Poor's 100 index (S&P100). The data set comprises of $n = 436$ returns and $p = 97$ stocks weekly recorded from August 23, 2004, to December 12, 2012. For simplicity, the log-GARCH (1, 1) model is used to estimate $\boldsymbol{D}_t$.

To measure the accuracy of the covariance matrix estimates, we consider the following loss functions: the entropy loss Δ_{1t}, the Kullback-Leibler loss Δ_{2t}, and the quadratic loss functions Δ_{3t} (up to some scale) defined as

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

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

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

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

where $\hat{\boldsymbol{\Sigma}}_t = (\hat{\omega}_{ij;t})_{p\times p}$ stands for the estimate of the covariance matrix $\boldsymbol{\Sigma}_t = (\omega_{ij;t})_{p\times p}$, $t = 1, \ldots, n$. For each loss measure, we report their averages over the time t as $\text{MAE} = \sum_{t=1}^{n}\text{MAE}_t/n$, $\text{MSE} = \sum_{t=1}^{n}\text{MSE}_t/n$, and $\Delta_i = \sum_{t=1}^{n}\Delta_{it}/n$, $i = 1, 2, 3$.

Since the true realized covariance matrix $\boldsymbol{\Sigma}_t$ is unknown, we employ a moving blocks approach to get a reliable proxy [23]. Table 43.1 reports the averaged loss measures over time t and their standard errors in parenthesis for methods including ORIG, BIC, BPA, DCC, and OAM. ORIG represents the Cholesky-based method for the estimation of covariance matrix based on the original ordering of variables. Similarly, BIC and BPA are the Cholesky-based methods for the estimation of covariance matrix based on the BIC and BPA to determine the ordering of variables, respectively. OAM stands for the Cholesky-based estimate of covariance matrix using the ordering-averaged model. DCC represents for the dynamic conditional correlation GARCH model, which is a popular tool to fit time series data in finance. It imposes a dynamic structure on the conditional correlation matrices.

It is clear to see from Table 43.1 that the OAM estimate based on the MCD gives the best performance regarding all the loss functions. The possible reason is that the stocks have no natural ordering among themselves. For example, it is not reasonable to order the stocks of Apple Inc., Dow Chemical Co., Microsoft Corp, and Bank of America Corp. Hence, the ordering-averaged model shows a relatively accurate estimation. The second best is the BPA, which is slightly inferior to the OAM. The DCC model does not provide accurate estimates for the time-varying covariance matrices compared with the OAM and BPA for this data set. The ORIG and BIC methods produce large losses, especially in terms of the quadratic loss Δ_3 and MSE. Overall, the MCD approach is suitable for the time-varying covariance matrix estimation for this time-series data.

43.4.2 MCD for Portfolio Optimization

Now we consider a portfolio optimization process that determines the portfolio allocation of multiple assets to minimize the portfolio variance. The risk of a portfolio $\boldsymbol{w} = (w_1, \ldots, w_p)$ is measured by the variance $\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ of its return, where $w_i \geq 0$ and $\sum_{i=1}^{p} w_i = 1$. The estimated minimum variance portfolio optimization problem is formulated as

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\hat{\boldsymbol{\Sigma}}\boldsymbol{w} \qquad (43.20)$$
$$\text{s.t.} \sum_{i=1}^{p} w_i = 1,$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of the true covariance matrix $\boldsymbol{\Sigma}$ of asset returns.

We illustrate the performances of each method, including ORIG, BIC, BPA, OAM, and DCC, applied into the portfolio optimization problem using the same data set of 97 stock returns from S&P100 as in Sect. 43.4.1. The first t observations are used to estimate the Cholesky factor matrices $(\boldsymbol{T}, \boldsymbol{D}_t)$ and then predict the covariance matrices $\hat{\boldsymbol{\Sigma}}_{t+1} = \hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{D}}_{t+1}\hat{\boldsymbol{T}}^{'-1}$ at time $t+1$, $t = 350, 351, \ldots, 435$. Note that the value of $d^2_{j;t+1}$, which is the diagonal element of $\boldsymbol{D}_{t+1}$, can be estimated from Eq. (43.19) by generating $\hat{\epsilon}_{j;t} = \hat{d}_{j;t}\eta_t$, where η_t is a random variable with mean 0. For this application, we choose $\eta_t \sim t_{df=5}$ distribution. The estimated portfolio $\hat{\boldsymbol{w}}_{t+1}$ is the solution of (43.20) by replacing $\hat{\boldsymbol{\Sigma}}$ with $\hat{\boldsymbol{\Sigma}}_{t+1}$. In practice, the researchers care not only the portfolio risk in (43.20) but also the reward and the information ratio (reward to risk). Hence, the performance measures of interest are the average annual realized return

$$\text{AVG} = \frac{1}{86}\sum_{t=350}^{435} 52 * \hat{\boldsymbol{w}}'_{t+1}\boldsymbol{x}_{t+1},$$

their standard deviation (SD), and the information ratio AVG/SD. Notice that the optimization (43.20) is designed to minimize the portfolio variance rather than to maximize the expected return or the information ratio. Therefore, any portfolio should be primarily evaluated by how successfully it achieves the minimum SD. A large realized return and a high value of information ratio are naturally also desirable

Table 43.1 The averages and standard errors (in parenthesis) of loss measures for the weekly returns of 97 stocks

	Δ_1	Δ_2	Δ_3	MAE	MSE
ORIG	3.397 (0.054)	3.196 (0.036)	1017 (28.35)	6.380 (0.089)	90.61 (2.167)
BIC	3.017 (0.077)	3.376 (0.049)	1176 (32.69)	9.004 (0.486)	364.0 (87.92)
BPA	0.747 (0.029)	0.446 (0.010)	14.94 (0.970)	3.877 (0.096)	36.12 (1.594)
DCC	0.902 (0.049)	0.825 (0.029)	142.3 (11.35)	5.338 (0.223)	172.8 (27.32)
OAM	0.681 (0.024)	0.335 (0.005)	6.183 (0.344)	3.687 (0.102)	35.27 (1.897)

Table 43.2 The comparison of portfolio performances for the weekly returns of 97 stocks

	ORIG	BIC	BPA	DCC	OAM
AVG	8.078	10.93	10.70	8.738	9.840
SD	14.59	17.09	8.007	8.800	7.230
AVG/SD	0.554	0.640	1.336	0.993	1.361

Table 43.3 Misclassification rates (in percentage) under 50 times randomly splitting for hand movement data

Method	BIC	BPA	Glasso	GLDA	DLDA	OAM
Misclassification	40.1	39.0	39.1	51.0	46.3	38.9
SE	0.5	0.5	0.6	0.5	0.6	0.5

but should be considered of secondary importance from the point of view of evaluating the quality of a covariance matrix estimate.

Table 43.2 summarizes the portfolio performances for each method in terms of AVG, SD, and the information ratio. The OAM estimate provides the smallest SD and the largest information ratio, followed by the BPA method. Although the BIC produces the highest value of the realized return, it has the worst portfolio risk among all the approaches, hence resulting in a low information ratio. The ORIG gives inferior performance with relatively small AVG and large SD. The DCC model does not perform well as the OAM and BPA, but it is better than the BIC and ORIG for this set of data.

43.4.3 MCD for Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a commonly used classification method in statistics and machine learning to construct a decision boundary via a linear combination of predictor variables that separates two or more classes of objects. For a multiple-class discriminant problem, each observation $\boldsymbol{x}$ belongs to some class $k \in 1, 2, \ldots, K$. Let $Y \in \{1, 2, \ldots, K\}$ represent K classes. Under the assumption that the conditional density function $f(\boldsymbol{x}|Y = k)$ follows a normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, the LDA classification rule is

$$\eta_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k, \qquad (43.21)$$

where π_k is the prior probability for class k. The observation $\boldsymbol{x}$ is assigned to the class $k^\star$ if $k^* = \arg\max_k \eta_k(\boldsymbol{x})$.

There are several unknown parameters in the classification rule (43.21), the population mean $\boldsymbol{\mu}_k$, precision matrix $\boldsymbol{\Sigma}^{-1}$, and prior probability π_k. We estimate them from the training data set. Let C_k be the set composed of the training observations belonging to the class k. Denote by $\hat{\boldsymbol{\mu}}_k$ the $p \times 1$ vector of the sample mean for the training data in class k. Let $\hat{\boldsymbol{\Sigma}}_{LDA} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i\in C_k}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'$ be the estimated within-class covariance matrix based on the training data. Then the estimated LDA classification rule is

$$\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}_{LDA}^{-1}\hat{\boldsymbol{\mu}}_k - \frac{1}{2}\hat{\boldsymbol{\mu}}_k'\hat{\boldsymbol{\Sigma}}_{LDA}^{-1}\hat{\boldsymbol{\mu}}_k + \log\hat{\pi}_k,$$

where $\hat{\pi}_k$ is the frequency of class k in the training data set. It is clear that the estimation accuracy of the covariance or precision matrix will have a profound effect on the accuracy of the classification accuracy of the LDA methods. See the derivation in the appendix for the details. For the high-dimensional data, the $\hat{\boldsymbol{\Sigma}}_{LDA}$ is often unstable or singular. Hence some other estimates of $\boldsymbol{\Sigma}^{-1}$ are used such as the generalized inverse of the within-class covariance matrix (GLDA) or Glasso estimate [38]. In this section, we consider $\boldsymbol{\Sigma}^{-1}$ estimated by some Cholesky-based precision matrix estimation methods such as BIC, OAM, and BPA and examine their performances under the LDA framework. We also present the classification results obtained by estimating $\boldsymbol{\Sigma}^{-1}$ from Glasso, GLDA, and DLDA methods. The DLDA assumes the off-diagonal elements of the precision matrix to be zeroes.

We apply each method into the hand movement data set [29], which contains 15 classes with 24 observations in each class. Every class refers 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 predictor variables. The whole data set is randomly split into a training set with 160 observations and testing set with the rest 200 observations. The 160 observations in the training set are used to estimate the population means $\boldsymbol{\mu}_k$, the population covariance matrix $\boldsymbol{\Sigma}$, and the prior probability π_k in the classification rule (43.21). Then the testing data are used to compute the misclassification rate for each method. The above procedure is repeated 50 times, and Table 43.3 reports the averages and standard errors (SE) of the misclassification rate in percentage for each method.

We see that the Cholesky-based estimates of BIC, OAM, BPA, and Glasso estimate perform comparably with respect to the misclassification rate. The DLDA produces an inferior performance result, possibly due to the reason that the underlying precision matrix of the 90 predictor variables might not be a diagonal structure. In other words, some variables are not conditional independent with each other. The GLDA does not perform well for this data set. It only has a half chance of assigning a new object to the correct class. Note that the misclassification rate of each method is relatively large, since the data contain 15 classes, which makes the discriminant analysis more difficult.

43.5 Numerical Study

In this section, we provide several numerical examples to further illustrate the Cholesky-based estimation of covariance and precision matrices. Here we consider the following two covariance and two precision matrix structures.

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

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

Model 3. $\mathbf{\Omega}_2$ is a diagonal matrix with its diagonal elements the inverse of vector $(p, p-1, p-2, \ldots, 1)'$.

Model 4. $\mathbf{\Omega}_1 = \text{AR}(0.5)$. 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 1 is a banded sparse matrix, while **Model** 2 is an unstructured sparse matrix. **Model** 3 is a diagonal matrix. **Model** 4 is an autoregressive structure that has homogeneous variances and correlations declining with distance. This model is more dense than the other models. For each model, data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are generated independently from the normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ with $n = 50$ and $p \in \{30, 50\}$. We use the same loss functions as in the application section to evaluate the accuracy of the estimates. Besides, to examine the performances of the estimates in catching the sparse structure, the false selection loss (FSL) is used, which is the summation of false positive (FP) and false negative (FN). The FSL is computed in percentage as $(\text{FP} + \text{FN}) / p^2$. Tables 43.4 and 43.5 report the averages of loss measures and their corresponding standard errors (in parenthesis), respectively, for each method based on 50 replications. The dashed lines in the tables represent the corresponding values not available due to matrix singularity.

In Table 43.4, we compare the performances of the sample covariance matrix $\boldsymbol{S}$, BIC, BPA, OAM, and RLZ. The RLZ represents the covariance estimator proposed by Rothman et al. [28] (see details in Sect. 43.2.3), which is designated for the banded matrix estimation. From the table, it is seen that the ordering-averaged method OAM gives better results than the sample covariance matrix, BIC and BPA with respect to Δ_1, Δ_2, Δ_3, MAE, and MSE, since it takes advantage of multiple variable orderings over one single ordering. However, the OAM does not capture the sparse structure, which is destroyed by the average operation in (43.18). The BIC and BPA can have some sparsity regarding FSL due to the Lasso regularization when constructing the Cholesky factors. But their FSL are worse than that of RLZ, which directly forces most of elements in the Cholesky factors to be zeroes. In addition, it is not surprising to observe that the RLZ performs better for the covariance matrix $\mathbf{\Sigma}_1$ than $\mathbf{\Sigma}_2$, since $\mathbf{\Sigma}_2$ is no longer a banded matrix after random permutations of rows and columns.

Table 43.5 presents the results obtained by the BIC, BPA, OAM, and Glasso methods. The ordering-averaged method OAM performs generally well but produces no sparsity for the matrix. In contrast, the Glasso estimate is good at catching the sparse structure as seen from FSL. The Glasso method [38] imposes an L_1 type penalty on the negative likelihood function of $\mathbf{\Omega}$, hence encouraging the sparsity in the esti-

Table 43.4 The averages and standard errors of estimates for covariance matrix $\mathbf{\Sigma}$

			Δ_1	Δ_2	Δ_3	MAE	MSE	FSL(%)
$\mathbf{\Sigma}_1$	$p = 30$	$\boldsymbol{S}$	12.3 (0.12)	38.7 (1.11)	89.9 (4.54)	3.51 (0.04)	19.9 (0.46)	84.0 (0.02)
		BIC	7.11 (0.12)	11.0 (0.53)	8.11 (0.84)	1.74 (0.02)	10.6 (0.19)	52.0 (0.80)
		BPA	5.94 (0.10)	7.97 (0.28)	4.46 (0.35)	1.53 (0.02)	8.78 (0.19)	45.9 (0.91)
		OAM	4.60 (0.07)	5.24 (0.11)	1.58 (0.11)	1.56 (0.02)	7.88 (0.17)	83.6 (0.04)
		RLZ	9.34 (0.09)	6.58 (0.06)	0.35 (0.03)	1.04 (0.01)	8.51 (0.08)	6.22 (0.01)
	$p = 50$	$\boldsymbol{S}$	–	–	–	5.77 (0.04)	53.0 (0.70)	90.2 (0.01)
		BIC	15.7 (0.30)	52.8 (8.35)	147 (49.0)	1.96 (0.02)	20.3 (0.24)	42.6 (0.70)
		BPA	13.0 (0.26)	20.3 (1.36)	16.2 (2.62)	1.82 (0.03)	18.4 (0.33)	41.1 (0.93)
		OAM	9.31 (0.10)	10.0 (0.16)	2.88 (0.19)	1.79 (0.02)	15.6 (0.25)	89.2 (0.06)
		RLZ	16.2 (0.10)	11.4 (0.08)	0.65 (0.07)	1.06 (0.01)	14.3 (0.13)	3.84 (0.01)
$\mathbf{\Sigma}_2$	$p = 30$	$\boldsymbol{S}$	12.4 (0.12)	38.6 (1.11)	88.2 (4.56)	3.53 (0.03)	19.8 (0.37)	84.0 (0.01)
		BIC	7.31 (0.14)	11.3 (0.51)	8.49 (0.84)	1.75 (0.02)	10.3 (0.17)	52.3 (0.66)
		BPA	5.81 (0.12)	7.58 (0.31)	4.06 (0.43)	1.49 (0.02)	8.40 (0.18)	46.5 (1.02)
		OAM	4.61 (0.08)	5.14 (0.11)	1.43 (0.12)	1.54 (0.01)	7.50 (0.16)	83.5 (0.05)
		RLZ	17.0 (0.17)	10.2 (0.05)	0.28 (0.04)	1.68 (0.01)	18.9 (0.07)	15.6 (0.01)
	$p = 50$	$\boldsymbol{S}$	–	–	–	5.73 (0.04)	52.7 (0.74)	90.2 (0.01)
		BIC	16.0 (0.32)	50.1 (7.06)	121 (33.6)	1.97 (0.02)	20.4 (0.24)	43.2 (0.75)
		BPA	12.8 (0.23)	20.3 (1.51)	16.6 (3.15)	1.83 (0.03)	18.5 (0.29)	40.9 (0.96)
		OAM	9.39 (0.09)	10.3 (0.15)	3.08 (0.19)	1.81 (0.01)	15.9 (0.23)	89.2 (0.05)
		RLZ	31.6 (0.25)	19.8 (0.08)	0.75 (0.09)	1.90 (0.01)	36.5 (0.08)	11.4 (0.01)

Table 43.5 The averages and standard errors of estimates for precision matrix $\boldsymbol{\Omega}$

			Δ_1	Δ_2	Δ_3	MAE	MSE	FSL(%)
$\boldsymbol{\Omega}_1$	$p = 30$	BIC	2.79 (0.26)	1.84 (0.09)	1.34 (0.23)	0.10 (0.01)	0.45 (0.09)	27.4 (2.09)
		BPA	2.45 (0.13)	1.72 (0.06)	0.99 (0.12)	0.13 (0.01)	0.42 (0.05)	21.7 (1.69)
		OAM	1.99 (0.14)	1.40 (0.06)	1.19 (0.16)	0.10 (0.01)	0.27 (0.04)	79.8 (0.84)
		Glasso	2.86 (0.06)	5.48 (0.16)	1.65 (0.08)	0.07 (0.01)	0.80 (0.01)	8.24 (0.34)
	$p = 50$	BIC	14.0 (3.78)	3.97 (0.22)	22.6 (12.8)	0.26 (0.09)	4.64 (2.57)	31.0 (2.41)
		BPA	5.32 (0.47)	3.22 (0.15)	2.62 (0.39)	0.13 (0.02)	0.79 (0.34)	19.3 (1.21)
		OAM	4.36 (0.28)	2.73 (0.09)	3.11 (0.26)	0.10 (0.01)	0.27 (0.03)	73.2 (1.28)
		Glasso	5.45 (0.07)	11.9 (0.23)	2.74 (0.12)	0.05 (0.01)	1.00 (0.01)	7.97 (0.34)
$\boldsymbol{\Omega}_2$	$p = 30$	BIC	9.03 (0.43)	6.05 (0.11)	5.92 (0.61)	3.34 (0.20)	42.3 (5.52)	45.7 (0.31)
		BPA	6.29 (0.19)	5.20 (0.09)	2.84 (0.23)	2.29 (0.06)	17.5 (0.87)	44.8 (0.48)
		OAM	5.51 (0.17)	3.98 (0.06)	3.68 (0.27)	2.15 (0.05)	12.9 (0.63)	46.7 (0.04)
		Glasso	6.08 (0.10)	14.0 (0.37)	1.71 (0.07)	2.28 (0.01)	24.0 (0.24)	45.2 (0.17)
	$p = 50$	BIC	32.6 (5.21)	12.8 (0.26)	61.8 (30.7)	8.09 (1.93)	221 (29.9)	55.3 (0.59)
		BPA	14.2 (0.87)	10.8 (0.16)	8.54 (1.46)	3.21 (0.39)	81.1 (42.4)	44.0 (0.55)
		OAM	14.1 (0.59)	8.63 (0.12)	13.0 (1.06)	3.47 (0.21)	49.7 (7.84)	65.4 (0.03)
		Glasso	11.7 (0.12)	29.0 (0.47)	3.52 (0.09)	2.40 (0.01)	44.1 (0.25)	30.6 (0.07)

mated precision matrix. Because the underlying precision matrix $\boldsymbol{\Omega}_2$ is denser than $\boldsymbol{\Omega}_1$, the Glasso method appears to perform better with respect to FSL for $\boldsymbol{\Omega}_1$. Additionally, the Glasso also performs well regarding the loss measures Δ_3 and MAE. We also note that the BPA gives a superior performance than BIC estimate. This superiority is more evident for the precision matrix $\boldsymbol{\Omega}_2$ compared with $\boldsymbol{\Omega}_1$, since $\boldsymbol{\Omega}_2$ is an autoregressive model.

43.6 Discussion

This chapter reviews the modified Cholesky decomposition (MCD) method for the estimation of covariance and precision matrices. It is seen that the MCD method has the flexibility of handling matrix estimation by transforming the problem into a sequence of regression-based problems. The sparsity can be easily imposed on the Cholesky factor matrices via the linear regressions and hence encouraging the sparse structure in the matrix estimate. The Cholesky-based methods guarantee the positive definiteness of the estimated matrix. Note that such Cholesky-based methods often require the knowledge on the ordering of variables when estimating a sequence of regressions. To address this issue, we thoroughly discuss the ordering issue in the MCD approach and examine several solutions to this ordering issue from the literature.

It is worth pointing out that the use of Cholesky-based approach is not restricted for estimating covariance and precision matrices under the conventional setting. The idea of MCD can also be used in constructing the covariance function in spatial analysis such as Kriging and Gaussian process [8], especially when involving the qualitative input variables. Because of the Cholesky decomposition on the matrix, the induced covariance matrix would have attractive property in the Gaussian process modeling. Another direction of using the MCD is the multi-response regression, where the multivariate responses have certain dependency structures. Under this situation, the MCD method of estimating the covariance or precision matrix for the multivariate responses will be coupled with the estimation of regression coefficients.

Acknowledgments The authors would like to thank the editor and reviewers for the constructive and insightful comments, which have significantly enhanced the quality of this article.

Appendix

Proof of Remark 1 and 2 Since the conclusions of Remarks 1 and 2 are much similar, we only provide the proof of Remark 2 here. Assume that there are n independent and identically distributed observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, which are centered. Let $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ be the sample covariance matrix and assume that $\boldsymbol{S}$ is non-singular since $n > p$. We denote $\hat{\boldsymbol{\Sigma}}_0$ as the estimated covariance matrix from (43.12) with tuning parameters equal to zeroes in (43.11). Then Remark 2 states that $\hat{\boldsymbol{\Sigma}}_0 = \boldsymbol{S}$ in spite of any permutation of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Below is the proof.

Based on the sequential regression of (43.11), it is known that the first step is $X_1 = \epsilon_1$. It means that

$$e_{i1} = x_{i1},\ 1 \le i \le n, \text{ and } \hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} e_{i1}^2$$

Then the second step is to consider $X_2 = l_{21}\epsilon_1 + \epsilon_2$, which provides

$$\hat{l}_{21} = \frac{\sum_{i=1}^{n} x_{i2}e_{i1}}{\sum_{i=1}^{n} e_{i1}^2},\ e_{i2} = x_{i2} - \hat{l}_{21}e_{i1},\ 1 \le i \le n$$

$$\hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n} e_{i2}^2,\ \sum_{i=1}^{n} e_{i2}e_{i1} = 0$$

In general, the jth step is to consider the regression problem as

$$X_j = \sum_{k<j} l_{jk}\epsilon_k + \epsilon_j$$

and we can obtain

$$\hat{l}_{j1} = \frac{\sum_{i=1}^{n} x_{ij}e_{i1}}{\sum_{i=1}^{n} e_{i1}^2}, \ldots, \hat{l}_{jk} = \frac{\sum_{i=1}^{n} x_{ij}e_{ik}}{\sum_{i=1}^{n} e_{ik}^2}, \ldots, \hat{l}_{j,j-1} = \frac{\sum_{i=1}^{n} x_{ij}e_{i,j-1}}{\sum_{i=1}^{n} e_{i,j-1}^2}$$

$$e_{ij} = x_{ij} - \sum_{k<j} \hat{l}_{jk}e_{ik},\ 1 \le i \le n$$

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} e_{ij}^2,\ \sum_{i=1}^{n} e_{ij}e_{i1} = 0,\ \ldots,\ \sum_{i=1}^{n} e_{ij}e_{i,j-1} = 0$$

Therefore, we can express the (s, t) entry of the covariance matrix estimate using the regression coefficients as

$$(\hat{\boldsymbol{\Sigma}})_{st} = (\hat{\boldsymbol{L}}\hat{\boldsymbol{D}}\hat{\boldsymbol{L}}^T)_{st} = \sum_{u=1}^{\min(s,t)} \hat{l}_{su}\hat{l}_{tu}\hat{\sigma}_u^2 \quad (\hat{l}_{uu} = 1).$$

Note that

$$x_{is} = \sum_{u=1}^{s} \hat{l}_{su}e_{iu} \quad (\hat{l}_{uu} = 1), \quad 1 \le i \le n,$$

$$x_{it} = \sum_{v=1}^{t} \hat{l}_{tv}e_{iv} \quad (\hat{l}_{vv} = 1), \quad 1 \le i \le n,$$

and the (s, t) entry of the sample covariance matrix is

$$\begin{aligned}(\boldsymbol{S})_{st} &= \frac{1}{n}\sum_{i=1}^{n} x_{is}x_{it} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{u=1}^{s}\hat{l}_{su}e_{iu}\right)\left(\sum_{v=1}^{t}\hat{l}_{tv}e_{iv}\right)\\ &= \frac{1}{n}\sum_{u=1}^{s}\sum_{v=1}^{t}\hat{l}_{su}\hat{l}_{tv}\left(\sum_{i=1}^{n} e_{iu}e_{iv}\right)\\ &= \sum_{u=1}^{\min(s,t)} \hat{l}_{su}\hat{l}_{tu}\hat{\sigma}_u^2 \quad (\hat{l}_{uu} = 1).\end{aligned}$$

The last equality holds because of

$$\sum_{i=1}^{n} e_{iu}e_{iv} = \begin{cases} n\sigma_u^2 & u = v;\\ 0 & u \ne v.\end{cases}$$

Thus, we can establish the result

$$\boldsymbol{S} = \hat{\boldsymbol{L}}\ \mathrm{diag}\left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2\right)\ \hat{\boldsymbol{L}}^T.$$

□

Conditional Misclassification Error of LDA

Without loss of generality, we consider a two-class classification problem here. Suppose the binary classifier function for LDA is $g(\boldsymbol{x}) = \log[P(Y = 1|\boldsymbol{X} = \boldsymbol{x})/P(Y = 2|\boldsymbol{X} = \boldsymbol{x})]$. Then

$$\begin{aligned}g(\boldsymbol{x}) \triangleq \boldsymbol{a}^T\boldsymbol{x} - b &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\\ &\quad - \left[\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \log\frac{\pi_1}{\pi_2}\right],\end{aligned}$$

where π_1 and π_2 are the prior probabilities for class 1 and 2, respectively, i.e., $\pi_1 = P(Y = 1)$ and $\pi_2 = P(Y = 2)$. For a new observation $\boldsymbol{x}$, we predict its class $Y = 1$ if $g(\boldsymbol{x}) > 0$, and $Y = 2$ otherwise. Then the conditional misclassification error is

$$\begin{aligned}&P(g(\boldsymbol{x}) = 1|Y = 2)P(Y = 2) + P(g(\boldsymbol{x}) = 2|Y = 1)P(Y = 1)\\ &\quad = P(a^T\boldsymbol{x} - b > 0|Y = 2)\pi_2 + P(a^T\boldsymbol{x} - b \le 0|Y = 1)\pi_1.\end{aligned}$$

Since $\boldsymbol{x}|Y = 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, and $\boldsymbol{x}|Y = 2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, obviously $\boldsymbol{a}^T\boldsymbol{x}|Y = 1 \sim N(\boldsymbol{a}^T\boldsymbol{\mu}_1, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})$, and $\boldsymbol{a}^T\boldsymbol{x}|Y = 2 \sim N(\boldsymbol{a}^T\boldsymbol{\mu}_2, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})$. Therefore,

$$P(\boldsymbol{a}^T\boldsymbol{x} - b > 0|Y = 2) = \Phi\left(\frac{\boldsymbol{a}^T\boldsymbol{\mu}_2 - b}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}}}\right),$$

$$P(a^T\boldsymbol{x} - b \le 0|Y = 1) = \Phi\left(-\frac{\boldsymbol{a}^T\boldsymbol{\mu}_1 - b}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. As a result, the conditional misclassification error is

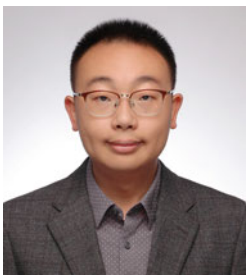
$$\pi_2\Phi\left(\frac{\boldsymbol{a}^T\boldsymbol{\mu}_2 - b}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}}}\right) + \pi_1\Phi\left(-\frac{\boldsymbol{a}^T\boldsymbol{\mu}_1 - b}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}}}\right).$$

Assume $\pi_1 = \pi_2 = 1/2$. Then with the estimates of $\boldsymbol{a}$ and b through $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}$, the conditional misclassification error $\gamma(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)$ is

$$\begin{aligned}&\gamma(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)\\ &\quad = \frac{1}{2}\Phi\left(\frac{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}{\sqrt{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}}\right)\\ &\quad\quad + \frac{1}{2}\Phi\left(-\frac{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}{\sqrt{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}}\right).\end{aligned}$$

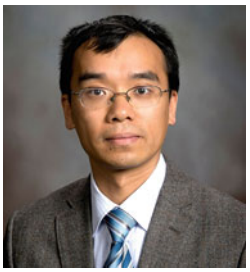
References

1. Bickel, P.J., Levina, E.: Covariance regularization by thresholding. Ann. Stat. **36**(6), 2577–2604 (2008a)
2. Bickel, P.J., Levina, E.: Regularized estimation of large covariance matrices. Ann. Stat. **36**(1), 199–227 (2008b)
3. Bien, J., Tibshirani, R.J.: Sparse estimation of a covariance matrix. Biometrika **98**(4), 807–820 (2011)
4. Cai, T.T., Yuan, M.: Adaptive covariance matrix estimation through block thresholding. Ann. Stat. **40**(40), 2014–2042 (2012)
5. Cai, T.T., Ren, Z., Zhou, H.H.: Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. Electronic Journal of Statistics **10**(1), 1–59 (2016)
6. Chang, C., Tsay, R.S.: Estimation of covariance matrix via the sparse Cholesky factor with lasso. J. Stat. Plann. Inference **140**(12), 3858–3873 (2010)
7. Cochran, W.G.: Sampling Techniques. Wiley, New York (1977)
8. Deng, X., Lin, C.D., Liu, K.-W., Rowe, R.K.: Additive Gaussian process for computer models with qualitative and quantitative factors. Technometrics **59**(3), 283–292 (2017)
9. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96** 1348–1360 (2001)
10. Fan, J., Liao, Y., Liu, H.: An overview of the estimation of large covariance and precision matrices. Econ. J. **19**(1), 1–32 (2016)
11. Francq, C., Zakoïan, J. M.: Estimating multivariate volatility models equation by equation. J. R. Stat. Soc. Ser. B (Stat Methodol.) **78**(3), 613–635 (2016)
12. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)
13. Huang, J.Z., Liu, N., Pourahmadi, M., Liu, L.: Covariance matrix selection and estimation via penalised normal likelihood. Biometrika **93**, 85–98 (2006)
14. Kang, X., Xie, C., Wang, M.: A Cholesky-based estimation for large-dimensional covariance matrices. J. Appl. Stat. **47**, 1017–1030 (2020)
15. Kang, X., Deng, X., Tsui, K., Pourahmad,i M.: On variable ordination of modified Cholesky decomposition for estimating time-varying covariance matrices. Int. Stat. Rev., **88**(3), 616–641 (2020)
16. Lam, C., Fan, J.: Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Stat. **37**, 4254–4278 (2009)
17. Lan, Q., Sun, H., Robertson, J., Deng, X., Jin, R.: Non-invasive assessment of liver 1uality in transplantation based on thermal imaging analysis. Comput. Methods Prog. Biomed. **164**, 31–47 (2018)
18. Leng, C., Li, B.: Forward adaptive banding for estimating large covariance matrices. Biometrika **98**(4), 821–830 (2011)
19. Levina, E., Zhu, R.J.: Sparse estimation of large covariance matrices via a nested lasso penalty. Ann. Appl. Stat. **2**(1), 245–263 (2008)
20. Liu, H., Wang, L., Zhao, T.: Sparse covariance matrix estimation with eigenvalue constraints. J. Comput. Graph. Stat. **23**(2), 439–459 (2014)
21. Mohammadi, A., Wit, E.C.: Bayesian structure learning in sparse Gaussian graphical models. Bayesian Anal. **10**(1), 109–138 (2015)
22. Nino-Ruiz, E.D., Sandu, A., Deng, X.: An ensemble Kalman filter implementation based on modified cholesky decomposition for inverse covariance matrix estimation. SIAM J. Sci. Comput. **40**(2), A867–CA886 (2018)
23. Pedeli, X., Fokianos, K., Pourahmadi, M.: Two Cholesky-log-GARCH models for multivariate volatilities. Stat. Model. **15**, 233–255 (2015)
24. Pourahmadi, M.: Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika **86**, 677–690 (1999)
25. Rajaratnam, B., Salzman, J.: Best permutation analysis. J. Multivar. Anal. **121**, 193–223 (2013)
26. Rigollet, P., Tsybakov, A.: Estimation of covariance matrices under sparsity constraints. Probl. Inf. Transm. **51**(4), 32–46 (2012)
27. Rothman, A.J., Levina, E., Zhu, J.: Generalized thresholding of large covariance matrices. J. Am. Stat. Assoc. **104**(485), 177–186 (2009)
28. Rothman, A.J., Levina, E., Zhu, J.: A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika **97**(3), 539–550 (2010)
29. Sapsanis, C., Georgoulas, G., Tzes, A., Lymberopoulos, D.: Improving EMG based classification of basic hand movements using EMD. In: Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5754–5757 (2013)
30. Sun, H., Huang, S., Jin, R.: Functional graphical models for manufacturing process modeling. IEEE Trans. Autom. Sci. Eng. **14**(4), 1612–1621 (2017)
31. Sun, H., Rao, P.K., Kong, Z., Deng, X., Jin, R.: Functional quantitative and qualitative models for quality modeling in a fused deposition modeling process. IEEE Trans. Autom. Sci. Eng. **15**(1), 393–403 (2018)
32. Tan, L.S., Nott, D.J.: Gaussian variational approximation with sparse precision matrices. Stat. Comput. **28**(2), 259–275 (2018)
33. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Stat Methodol.) **58**, 267–288 (1996)
34. Wagaman, A., Levina, E.: Discovering sparse covariance structures with the Isomap. J. Comput. Graph. Stat. **18**(3), 551–572 (2009)
35. Wu, H., Deng, X., Ramakrishnan, N.: Sparse estimation of multivariate poisson log-normal model and inverse covariance for counting data. Stat. Anal. Data Min. **11**, 66–77 (2018)
36. Xue, L., Ma, S., Zou, H.: Positive-definite L_1-penalized estimation of large covariance matrices. J. Am. Stat. Assoc. **107**(500), 1480–1491 (2012)
37. Yu, P.L.H., Wang, X., Zhu, Y.: High dimensional covariance matrix estimation by penalizing the matrix-logarithm transformed likelihood. Comput. Stat. Data Anal. **114**, 12–25 (2017)
38. Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94**, 19–35 (2007)
39. Zeng, L., Deng, X., Yang, J.: A constrained Gaussian process approach to modeling tissue-engineered scaffold degradation. IISE Trans. **50**(5), 431–447 (2018)
40. Zheng, H., Tsui, K.-W., Kang, X., Deng, X.: Cholesky-based model averaging for covariance matrix estimation. Stat. Theory Relat. Fields **1**(1), 48–58 (2017)



Xiaoning Kang received his Ph.D. degree in statistics from Virginia Tech. He is now an associate professor at International Business College and Institute of Supply Chain Analytics in Dongbei University of Finance and Economics, China. His research interests include high-dimensional matrix estimation, mixed responses data analysis, Bayesian hierarchical modeling, and discriminant analysis.

Zhiyang Zhang is an instructor in the in the Department of Statistics at Virginia Tech. She received her master's degree in statistics and Ph.D. degree in chemistry from Virginia Tech. Her research interests focus on engineering statistics, data mining, and experimental design for webpage optimization.



Xinwei Deng is an associate professor in the Department of Statistics at Virginia Tech. He received his bachelor's degree in mathematics from Nanjing University and Ph.D. degree in industrial engineering from Georgia Tech. His research interests focus on statistical modeling and data analysis, including high-dimensional classification, graphical model estimation, and the interface between experimental design and machine learning. He is an elected member of ISI and a member of INFORMS and ASA.