

## Weighted average ensemble for Cholesky-based covariance matrix estimation

Xiaoning Kang, Zhenguo Gao, Xi Liang & Xinwei Deng

To cite this article: Xiaoning Kang, Zhenguo Gao, Xi Liang & Xinwei Deng (09 Apr 2025): Weighted average ensemble for Cholesky-based covariance matrix estimation, Statistical Theory and Related Fields, DOI: [10.1080/24754269.2025.2484979](https://doi.org/10.1080/24754269.2025.2484979)

To link to this article: <https://doi.org/10.1080/24754269.2025.2484979>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 09 Apr 2025.



Submit your article to this journal [↗](#)



Article views: 37



View related articles [↗](#)



View Crossmark data [↗](#)



# Weighted average ensemble for Cholesky-based covariance matrix estimation

Xiaoning Kang<sup>a</sup>, Zhenguo Gao<sup>b</sup>, Xi Liang<sup>b</sup> and Xinwei Deng<sup>c</sup>

<sup>a</sup>Institute of Supply Chain Analytics, Dongbei University of Finance and Economics, Dalian, People's Republic of China; <sup>b</sup>School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, People's Republic of China;

<sup>c</sup>Department of Statistics, Virginia Tech, Blacksburg, VA, USA

## ABSTRACT

The modified Cholesky decomposition (MCD) is an efficient technique for estimating a covariance matrix. However, it is known that the MCD technique often requires a pre-specified variable ordering in the estimation procedure. In this work, we propose a weighted average ensemble covariance estimation for high-dimensional data based on the MCD technique. It can flexibly accommodate the high-dimensional case and ensure the positive definiteness property of the resultant estimate. Our key idea is to obtain different weights for different candidate estimates by minimizing an appropriate risk function with respect to the Frobenius norm. Different from the existing ensemble estimation based on the MCD, the proposed method provides a sparse weighting scheme such that one can distinguish which variable orderings employed in the MCD are useful for the ensemble matrix estimate. The asymptotically theoretical convergence rate of the proposed ensemble estimate is established under regularity conditions. The merits of the proposed method are examined by the simulation studies and a portfolio allocation example of real stock data.

## ARTICLE HISTORY

Received 27 May 2024

Revised 13 February 2025

Accepted 22 March 2025

## KEYWORDS

Ensemble estimate; modified Cholesky decomposition; portfolio strategy; variable ordering

## 1. Introduction

The estimation of covariance matrix plays an important role in the multivariate statistics with a broad spectrum of applications including dimension reduction, linear discriminant analysis, social network, remote sensing, functional magnetic resonance imaging etc. Recently, scholars have investigated the covariance estimation for a certain type of data, such as the compositional data (Cao et al., 2019), the matrix-valued data (Zhang et al., 2022) and the spatial data (Kidd & Katzfuss, 2022). However, the covariance estimation for high dimensions where the number of variables is larger than the sample size encounters challenges due to two difficulties. One is the positive definite and symmetric properties required by the covariance matrix itself. The other is a large amount of parameters involved in the model estimation in the sense that the number of parameters increases rapidly in the quadratic order of the model dimension. Therefore, the covariance matrix estimation is not an easy work, and has attracted extensive interest among many scholars.

**CONTACT** Zhenguo Gao  gaozheng@sjtu.edu.cn, gaozhenguo3@126.com  School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

To induce the sparsity for the high-dimensional covariance matrix estimate, Bickel and Levina (2008a) suggested a hard-thresholding method which directly sets the small quantities of the sample covariance matrix to be zero, and provided its theoretical results. In addition, with the assumption that variables are weakly correlated when they are far away in an ordering, Furrer and Bengtsson (2007) and Bickel and Levina (2008b) introduced the tapering and banded covariance estimates which are consistent under some conditions. Although the computational cost of such thresholding-based methods is small, their estimates cannot guarantee the property of the positive definiteness. To circumvent this problem, Bien and Tibshirani (2011) improved the covariance estimation via imposing the  $L_1$  penalty on the negative log likelihood. Furthermore, Xue et al. (2012) developed a covariance estimation with additional positive definite constraint added on their objective function. However, they are computationally intensive due to either the non-convexity of the objective function or the iterative estimation procedure. More covariance estimation studies can be found in Deng and Tsui (2013), Cai et al. (2016), C. Huang et al. (2017), Ledoit and Wolf (2020), Jenny Shi et al. (2021) and Xin and Zhao (2023), among others.

Another powerful tool for the estimation of the high-dimensional covariance matrix is the modified Cholesky decomposition (MCD) (Pourahmadi, 1999) which is efficient to deal with the two challenges mentioned above. It guarantees the positive definiteness and symmetry of the covariance estimate via the matrix decomposition, and transforms the difficult problem of matrix estimation into the easy tasks of solving linear regressions, which is thus statistically meaningful and able to accommodate large number of variables in the high-dimensional data analysis. In view of this, many papers in the literature have studied the covariance estimation via the MCD (J. Huang et al., 2006; Kang & Wang, 2021; Leng & Li, 2011; Lv et al., 2018; Pedeli et al., 2015; Wu & Pourahmadi, 2003). However, the matrix estimates by the MCD technique depend on the variable ordering when constructing the linear regressions, implying that different variable orderings will lead to different Cholesky-based estimates. Therefore, it requires a prior knowledge on the variable ordering of data in practice, e.g., the longitudinal data and the spatial data have a natural ordering. This remarkably narrows down the scope of its applications, since in most cases the real data do not have a natural ordering, or we do not have any information on the variable orderings before the analysis, such as the gene data, the stock data, the industry data and the medical data. Hence several papers have contributed to solving this ordering issue. One idea is that a certain variable ordering can be determined based on several criteria via some data driven mechanisms before applying the MCD technique. Typical criteria include the Isomap proposed by Wagaman and Levina (2009), the Bayesian Information Criterion (BIC) suggested by Dellaportas and Pourahmadi (2012), the Best Permutation Algorithm (BPA) introduced by Rajaratnam and Salzman (2013) and so forth. Although their methods extend the application scope of the MCD approach, an inappropriately identified ordering will damage the performance of the subsequent Cholesky-based estimation.

In contrast to using only one single variable ordering, H. Zheng et al. (2017) introduced another idea of employing the ensemble model via the MCD for the covariance estimation, which has been extensively studied in the later literature (Kang & Deng, 2020; Kang et al., 2025; Li et al., 2021; Liang et al., 2024). These methods have two advantages: (1) they require no prior knowledge on the variable ordering, solving the issue mentioned above; (2) the ensemble estimators combine useful information attained from different candidate models to reduce the estimation risk caused by relying on one single variable ordering which might be incorrectly selected, thus improving the estimation accuracy. The numerical studies

have empirically shown that the Cholesky-based matrix estimates by the ensemble model perform better than the estimates using one single variable ordering.

Nonetheless, such ensemble matrix estimates by the MCD technique in general have two main disadvantages. The first one is that they randomly generated a set of different variable orderings, and considered them to contribute equally in constructing the ensemble estimates. That is, they assigned the same weight to the candidate estimates that are obtained by the MCD corresponding to different variable orderings. An obvious way to improving the performance of such simply averaging estimators is to consider an ensemble model by using the optimal weights that might be different for each candidate estimate. The second disadvantage is that they took into account all the randomly generated variable orderings for the ensemble covariance estimation, without distinguishing whether the variable orderings are useful for building the covariance estimators under the framework of the MCD. To our best knowledge, these two disadvantages lie in all the existing Cholesky-based ensemble methods of matrix estimation. Especially the second one might affect the estimation accuracy if some useless variable orderings are involved into the covariance estimation. Therefore, in this paper we propose a Weighted Averaging Ensemble (WAE) for covariance estimation based on the MCD technique. We take the advantage of the MCD framework that it is able to provide multiple candidates of the covariance estimates by considering different variable orderings. The proposed method overcomes the first disadvantage by aggregating such candidates via the optimal weights, which are adaptively determined by minimizing a risk function in terms of the Frobenius norm. The proposed method also solves the second disadvantage through identifying which variable orderings are the most useful for the ensemble covariance estimates by means of imposing a penalty with respect to the weights on the objective function.

The remaining of this article is organized as follows. In Section 2, we review the MCD for the covariance estimation, then introduce the proposed WAE methods in details, and establish the theoretical properties. The simulations are conducted in Section 3, and a real case study of portfolio selection is presented in Section 4. We conclude this work with some discussions in Section 5. All the technical proofs are in the Appendix.

## 2. Methodology

### 2.1. Covariance estimation via the MCD

Without loss of generality, we assume that  $\mathbf{X} = (X_1, \dots, X_p)^\top$  is a  $p$ -dimensional random vector with mean value of  $\mathbf{0}$  and covariance matrix  $\Sigma$ . The MCD technique regresses each variable  $X_j$  on its preceding errors  $\varepsilon_1, \dots, \varepsilon_{j-1}$  for  $2 \leq j \leq p$ . Specifically, consider a series of linear regressions

$$X_j = \sum_{i=1}^{j-1} l_{ji} \varepsilon_i + \varepsilon_j, \quad j = 2, \dots, p, \quad (1)$$

where  $\mathbf{l}_j = (l_{j1}, \dots, l_{j(j-1)})^\top$  is the vector consisted of regression coefficients, and  $\varepsilon_j$  is the error term for the  $j$ th regression with its expectation of 0 and variance of  $d_j^2$ . In addition, let  $\varepsilon_1 = X_1$  and  $d_1^2 = \text{Var}(X_1)$ . Define by a diagonal matrix  $\mathbf{D} = \text{diag}(d_1^2, \dots, d_p^2)$  the variance matrix of the error vector  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)^\top$ . Let  $\mathbf{L} = (l_{ji})_{p \times p}$  be a lower triangular matrix whose diagonal elements are all ones, where  $l_{jj} = 1$  and  $l_{ji} = 0$  if  $i > j$ . As a result, the linear

regressions (1) can be written in the matrix form of  $\mathbf{X} = \mathbf{L}\boldsymbol{\varepsilon}$ . The covariance matrix can be thus written as  $\boldsymbol{\Sigma} = \text{Var}(\mathbf{X}) = \text{Var}(\mathbf{L}\boldsymbol{\varepsilon}) = \mathbf{L}\mathbf{D}\mathbf{L}^\top$ . By this means, the MCD technique reduces the difficulty of covariance estimation by decomposing it into the estimation of the Cholesky factor matrices  $\mathbf{D}$  and  $\mathbf{L}$ , which can be obtained easily by solving the  $p-1$  linear regressions in (1).

Let  $\mathbf{x}_1, \dots, \mathbf{x}_n$  be  $n$  independent observations from the random vector  $\mathbf{X}$ , and  $\mathbb{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$  be the  $n \times p$  data matrix. Define the  $j$ th column of  $\mathbb{X}$  by  $\mathbf{x}^{(j)}$ . Denote by  $\mathbf{e}^{(j)}$  the residuals of the  $j$ th regression for  $j \geq 2$ , and  $\mathbf{e}^{(1)} = \mathbf{x}^{(1)}$ . Let  $\mathbb{Z}^{(j)} = (\mathbf{e}^{(1)}, \dots, \mathbf{e}^{(j-1)})$  be the matrix containing the first  $j-1$  residuals. For the high-dimensional data, the Lasso regularization is often used to shrink the estimated regression coefficients (Chang & Tsay, 2010; J. Huang et al., 2006)

$$\hat{\mathbf{l}}_j = \arg \min_{\mathbf{l}_j} \|\mathbf{x}^{(j)} - \mathbb{Z}^{(j)}\mathbf{l}_j\|_2^2 + \lambda_j \|\mathbf{l}_j\|_1, \quad j = 2, \dots, p, \quad (2)$$

where  $\lambda_j \geq 0$  is a tuning parameter. The symbols  $\|\cdot\|_1$  and  $\|\cdot\|_2$  stand for the vector  $L_1$  and  $L_2$  norms.  $\mathbf{e}^{(j)} = \mathbf{x}^{(j)} - \mathbb{Z}^{(j)}\hat{\mathbf{l}}_j$  is used to construct the residuals for the last column of  $\mathbb{Z}^{(j+1)}$ . The estimate  $\hat{\sigma}_j^2$  is calculated as the sample variance of  $\mathbf{e}^{(j)}$ . Consequently, the covariance estimate is  $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{D}}\hat{\mathbf{L}}^\top$ , where  $\hat{\mathbf{L}}$  is constructed with  $\hat{\mathbf{l}}_j$  as its  $j$ th row, and  $\hat{\mathbf{D}} = \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2)$ .

## 2.2. The proposed estimators

Since the Cholesky factor matrix estimates  $\hat{\mathbf{L}}$  and  $\hat{\mathbf{D}}$  depend on the variable ordering  $X_1, \dots, X_p$  as one can see from Equation (1), we apply the ensemble model to solve this issue by considering multiple orderings for the MCD technique. Define a mapping  $\Pi: \{1, 2, \dots, p\} \rightarrow \{1, 2, \dots, p\}$  as  $\{1, 2, \dots, p\} \rightarrow \{\Pi(1), \dots, \Pi(p)\}$ , and let  $\mathbf{P}_\Pi$  be the corresponding permutation matrix. Let  $(\mathbf{L}_\Pi, \mathbf{D}_\Pi)$  be the matrices of  $(\mathbf{L}, \mathbf{D})$  obtained by the MCD using the variable ordering permutation  $\Pi$ . Accordingly we have a covariance estimate for  $\boldsymbol{\Sigma}$  under  $\Pi$  by transforming  $\hat{\boldsymbol{\Sigma}}_\Pi = \hat{\mathbf{L}}_\Pi \hat{\mathbf{D}}_\Pi \hat{\mathbf{L}}_\Pi^\top$  back to the original ordering in the following way

$$\hat{\boldsymbol{\Sigma}} = \mathbf{P}_\Pi \hat{\boldsymbol{\Sigma}}_\Pi \mathbf{P}_\Pi^\top = \mathbf{P}_\Pi \hat{\mathbf{L}}_\Pi \hat{\mathbf{D}}_\Pi \hat{\mathbf{L}}_\Pi^\top \mathbf{P}_\Pi^\top = (\mathbf{P}_\Pi \hat{\mathbf{L}}_\Pi \mathbf{P}_\Pi^\top) (\mathbf{P}_\Pi \hat{\mathbf{D}}_\Pi \mathbf{P}_\Pi^\top) (\mathbf{P}_\Pi \hat{\mathbf{L}}_\Pi^\top \mathbf{P}_\Pi^\top) = \hat{\mathbf{L}} \hat{\mathbf{D}} \hat{\mathbf{L}}^\top, \quad (3)$$

where  $\hat{\mathbf{L}} = \mathbf{P}_\Pi \hat{\mathbf{L}}_\Pi \mathbf{P}_\Pi^\top$  and  $\hat{\mathbf{D}} = \mathbf{P}_\Pi \hat{\mathbf{D}}_\Pi \mathbf{P}_\Pi^\top$ . Suppose we consider  $M$  different variable orderings  $\Pi_k, k = 1, \dots, M$ , and denote the corresponding estimates  $\hat{\boldsymbol{\Sigma}}, \hat{\mathbf{L}}$ , and  $\hat{\mathbf{D}}$  in (3) by  $\tilde{\boldsymbol{\Sigma}}_k, \tilde{\mathbf{L}}_k$ , and  $\tilde{\mathbf{D}}_k$ . H. Zheng et al. (2017) proposed a model averaging estimate  $\hat{\boldsymbol{\Sigma}}_{\text{zheng}} = \frac{1}{M} \sum_{k=1}^M \tilde{\boldsymbol{\Sigma}}_k$ , which has a relatively smaller variance due to the averaging and displays better performance than the Cholesky-based covariance estimate with one single variable ordering. However, since the Cholesky-based covariance estimates  $\tilde{\boldsymbol{\Sigma}}_k$ 's with different orderings may have different estimation biases, and H. Zheng et al. (2017) simply assigns equal weights to each candidate estimate  $\tilde{\boldsymbol{\Sigma}}_k$ , their method hence ignores that these estimates under different variable orderings may have different effects to the ensemble model. Therefore, we propose the following WAE estimate

$$\hat{\boldsymbol{\Sigma}}_{\text{wae}} = \sum_{k=1}^M \omega_k \tilde{\boldsymbol{\Sigma}}_k, \quad (4)$$

where  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_M)^\top$  is the vector of weights, which plays a critical role in producing a good and reliable ensemble covariance estimate. To obtain the optimal weights, we consider

to minimize the difference between the ensemble estimator and the true covariance matrix under the Frobenius norm. It is worth noting that such a strategy is also used in shrinking the estimators of covariance matrix (Ledoit & Wolf, 2003). Specifically, we employ the following quadratic loss function as the risk function to determine the optimal weights by

$$\begin{aligned} \hat{\omega} &= \arg \min_{\omega} R(\omega) = \arg \min_{\omega} E \|\omega_1 \tilde{\Sigma}_1 + \cdots + \omega_M \tilde{\Sigma}_M - \Sigma\|_F^2 \\ \text{s.t. } &\sum_{i=1}^M \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0, \end{aligned}$$

where  $E(\cdot)$  represents the expectation. Let  $\Sigma = (\sigma_{ij})_{p \times p}$  and  $\tilde{\Sigma}_k = (\tilde{\sigma}_{ij}^k)_{p \times p}$ . We decompose the risk function as

$$\begin{aligned} R(\omega) &= E \|\omega_1 \tilde{\Sigma}_1 + \cdots + \omega_M \tilde{\Sigma}_M - \Sigma\|_F^2 \\ &= \sum_{i=1}^p \sum_{j=1}^p E (\omega_1 \tilde{\sigma}_{ij}^1 + \cdots + \omega_M \tilde{\sigma}_{ij}^M - \sigma_{ij})^2 \\ &= \sum_{i=1}^p \sum_{j=1}^p \{ \text{Var}(\omega_1 \tilde{\sigma}_{ij}^1 + \cdots + \omega_M \tilde{\sigma}_{ij}^M) + [E(\omega_1 \tilde{\sigma}_{ij}^1 + \cdots + \omega_M \tilde{\sigma}_{ij}^M - \sigma_{ij})]^2 \} \\ &= \sum_{i=1}^p \sum_{j=1}^p \{ \omega_1^2 \text{Var}(\tilde{\sigma}_{ij}^1) + \cdots + \omega_M^2 \text{Var}(\tilde{\sigma}_{ij}^M) + \sum_{h \neq \ell} \omega_h \omega_\ell \text{Cov}(\tilde{\sigma}_{ij}^h, \tilde{\sigma}_{ij}^\ell) \\ &\quad + [\omega_1 E(\tilde{\sigma}_{ij}^1) + \cdots + \omega_M E(\tilde{\sigma}_{ij}^M) - \sigma_{ij}]^2 \} \\ &= \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^M \omega_k^2 [\text{Var}(\tilde{\sigma}_{ij}^k) + E^2(\tilde{\sigma}_{ij}^k)] \\ &\quad + \sum_{i=1}^p \sum_{j=1}^p \sum_{h \neq \ell} \omega_h \omega_\ell [\text{Cov}(\tilde{\sigma}_{ij}^h, \tilde{\sigma}_{ij}^\ell) + E(\tilde{\sigma}_{ij}^h)E(\tilde{\sigma}_{ij}^\ell)] \\ &\quad - \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^M 2\omega_k E(\tilde{\sigma}_{ij}^k) \sigma_{ij} + \sum_{i=1}^p \sum_{j=1}^p \sigma_{ij}^2. \end{aligned}$$

To write the risk function  $R(\omega)$  in a quadratic form of a matrix, we define a symmetric matrix  $\mathbf{A} = (a_{mn})$  and a vector  $\mathbf{b} = (b_k)$ , where

$$a_{mn} = \sum_{i=1}^p \sum_{j=1}^p \text{Cov}(\tilde{\sigma}_{ij}^m, \tilde{\sigma}_{ij}^n) + E(\tilde{\sigma}_{ij}^m)E(\tilde{\sigma}_{ij}^n) \quad \text{and} \quad b_k = -2 \sum_{i=1}^p \sum_{j=1}^p E(\tilde{\sigma}_{ij}^k) \sigma_{ij}.$$

Consequently, the risk function  $R(\omega)$  is rewritten as

$$R(\omega) = \omega^\top \mathbf{A} \omega + \mathbf{b}^\top \omega + \sum_{i=1}^p \sum_{j=1}^p \sigma_{ij}^2.$$

The optimal weight vector  $\boldsymbol{\omega}$  is estimated as the minimizer of  $R(\boldsymbol{\omega})$ , that is,

$$\widehat{\boldsymbol{\omega}} = \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} \quad \text{s.t.} \quad \sum_{i=1}^M \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0. \quad (5)$$

This optimization problem can be easily solved using the function `solve.QP(.)` in the package `quadprog` of R software.

As we mention in Section 1, not all the variable orderings would be useful in the weighted ensemble model for estimating the covariance matrix based on the MCD. Some sets of linear regressions of (1) from certain variable orderings may play no roles in the sense that using them in the MCD would increase the model complexity but not improve the estimation accuracy. It is equivalent to saying that some candidate estimates  $\widetilde{\Sigma}_k$  might contribute little or nothing to the weighted ensemble covariance estimation in Equation (4). In such case, we hence need to rule out such variable orderings that could be useless for the data analysis, and simultaneously to simplify the model complexity. It can be realized by setting the corresponding weights to zeros. To this end, we adopt the Lasso (Tibshirani, 1996) and the adaptive Lasso ideas (Zou, 2006) and suggest to impose two types of penalties on Equation (5), which yields

$$\widehat{\boldsymbol{\omega}} = \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} + \phi \sum_{i=1}^M |\omega_i| \quad \text{s.t.} \quad \sum_{i=1}^M \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0 \quad (6)$$

and

$$\widehat{\boldsymbol{\omega}}^* = \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} + \zeta \sum_{i=1}^M \theta_i |\omega_i| \quad \text{s.t.} \quad \sum_{i=1}^M \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0, \quad (7)$$

where  $\phi \geq 0$  and  $\zeta \geq 0$  are tuning parameters. A little algebra on (6) shows

$$\begin{aligned} \widehat{\boldsymbol{\omega}} &= \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} + \phi \sum_{i=1}^M \omega_i \\ &= \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} + \phi \quad \text{s.t.} \quad \sum_{i=1}^M \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0, \end{aligned}$$

which is equivalent to the optimization problem (5). It is thereby very interesting to see that the proposed WAE method with weight  $\widehat{\boldsymbol{\omega}}$  solved from (5) overcomes two disadvantages of existing methods that are mentioned in the Introduction by simultaneously completing the tasks of allocating different weights to the estimates  $\widetilde{\Sigma}_k$  and picking up the most useful variable orderings for the ensemble estimation.

For the optimization problem (7), let  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)^\top$ , and then it follows that

$$\begin{aligned} \widehat{\boldsymbol{\omega}}^* &= \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} + \zeta \sum_{i=1}^M \theta_i \omega_i \\ &= \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + \mathbf{b}^\top \boldsymbol{\omega} + \zeta \boldsymbol{\theta}^\top \boldsymbol{\omega} \\ &= \arg \min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega} + (\mathbf{b} + \zeta \boldsymbol{\theta})^\top \boldsymbol{\omega} \quad \text{s.t.} \quad \sum_{i=1}^M \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0. \end{aligned}$$

Similar to the adaptive Lasso idea, we suggest the value of  $\theta = 1/\widehat{\omega}$ , where  $\widehat{\omega}$  is the solution from the optimization problem (5). As a result, we induce a sparsity in the estimated weight vector, which automatically distinguishes and thus selects a set of the most important candidate estimates  $\widetilde{\Sigma}_k$ , or equivalently the most useful variable orderings, for the ensemble covariance estimation under the framework of MCD based on the data themselves.

The optimal value of the tuning parameter  $\zeta$  in the optimization problem (7) can be determined by minimizing the negative likelihood function of covariance matrix  $\Sigma$ . Specifically, consider a set of candidate values for  $\zeta$ , denoted as  $\mathcal{A}_\zeta = \{\zeta_1, \dots, \zeta_N\}$ . Let  $\widehat{\Sigma}_{\text{wae}^*}(\zeta_t)$  represent the proposed weighted ensemble covariance estimate computed from Equation (4) with its weights solved from the optimization (7) using tuning parameter  $\zeta_t$ ,  $t = 1, \dots, N$ . We then obtain  $Q_t = \log |\widehat{\Sigma}_{\text{wae}^*}(\zeta_t)| + \text{tr}[\widehat{\Sigma}_{\text{wae}^*}^{-1}(\zeta_t)\mathbf{S}]$ , where  $\mathbf{S}$  is the sample covariance matrix. Subsequently denote  $\tilde{t} = \arg \min_t \{Q_t, t = 1, \dots, N\}$ . Consequently, the optimal value of  $\zeta$  is  $\zeta_{\tilde{t}}$ , and the resultant covariance estimate is  $\widehat{\Sigma}_{\text{wae}^*}(\zeta_{\tilde{t}})$ . That is, the proposed covariance estimate is

$$\widehat{\Sigma}_{\text{wae}^*} = \sum_{k=1}^M \widetilde{\omega}_k^* \widetilde{\Sigma}_k,$$

where  $\widetilde{\omega}^* = (\widetilde{\omega}_1^*, \dots, \widetilde{\omega}_M^*)^\top$  is the solution of Equation (7) using tuning parameter  $\zeta_{\tilde{t}}$ .

### 2.3. Theoretical properties

This section establishes the convergence properties of the proposed WAE estimators. To facilitate the expression of the theoretical results, some assumptions are made on the true model. Let  $\Sigma_0 = \mathbf{L}_0 \mathbf{D}_0 \mathbf{L}_0^\top$  represent the underlying covariance matrix and its MCD. Similarly, let  $\Sigma_{0\Pi_k} = \mathbf{L}_{0\Pi_k} \mathbf{D}_{0\Pi_k} \mathbf{L}_{0\Pi_k}^\top$  be the MCD of the underlying covariance matrix regarding a variable order  $\Pi_k$ . Define  $\mathcal{C}_{\Pi_k} = \{(i, j) : j < i, l_{0ij}^{\Pi_k} \neq 0\}$  to be a set indexing the nonzero elements in the lower triangular part of the matrix  $\mathbf{L}_{0\Pi_k} = (l_{0ij}^{\Pi_k})$ . Then denote the maximum cardinality of  $\mathcal{C}_{\Pi_k}$  by  $s$ . Additionally, denote the singular values of matrix  $\Sigma_0$  by  $sv_1(\Sigma_0) \geq \dots \geq sv_p(\Sigma_0)$  in a decreasing order. In order to establish the theoretical property, we assume the regularity conditions as listed below.

- C1: The singular values of  $\Sigma_0$  are bounded. That is, there exist constants  $l_1$  and  $u_1$  such that  $0 < l_1 < sv_p(\Sigma_0) \leq sv_1(\Sigma_0) < u_1 < \infty$ .
- C2: The tuning parameters  $\lambda_j$ 's in (2) satisfy  $\sum_{j=1}^p \lambda_j = O(\sqrt{\log(p)/n})$ .
- C3:  $(s + p) \log(p) = o(n)$ .

The condition C1 is commonly used to guarantee the positive definiteness of  $\Sigma_0$  in the literature. The conditions C2 and C3 are used to derive the consistency properties of the proposed estimators. Let  $a_n \asymp b_n$  represent that two sequences  $a_n$  and  $b_n$  are the same order. Now we present the main results.

**Theorem 2.1:** *Suppose that the data are independently and identically distributed from  $N(\mathbf{0}, \Sigma_0)$ . Under the conditions C1–C3, we have*

$$\|\widehat{\Sigma}_{\text{wae}} - \Sigma_0\|_F \asymp \|\widehat{\Sigma}_{\text{wae}^*} - \Sigma_0\|_F = O_p\left(\sqrt{\frac{(s + p) \log(p)}{n}}\right).$$

Theorem 2.1 demonstrates the asymptotic convergence rates of the proposed WAE estimators with respect to the Frobenius norm. Such rate is parallel with that of some existing works (Lam & Fan, 2009; Rothman et al., 2008). The review paper of Cai et al. (2016) pointed out that the high-dimensional covariance estimators have the convergence rate in the form of  $O_p(\sqrt{\frac{(\kappa+p)\log(p)}{n}})$  in terms of the Frobenius norm, where  $\kappa$  is a measure of the sparsity for the underlying covariance matrix.

## 2.4. Estimates of matrix $\mathbf{A}$ and vector $\mathbf{b}$

In order to solve the weights in the optimizations (5) and (7), we need to find consistent estimators for  $\text{Var}(\tilde{\sigma}_{ij}^k)$ ,  $\text{Cov}(\tilde{\sigma}_{ij}^k, \tilde{\sigma}_{ij}^\ell)$ ,  $E(\tilde{\sigma}_{ij}^k)E(\tilde{\sigma}_{ij}^\ell)$  and  $E(\tilde{\sigma}_{ij}^k)\sigma_{ij}$  in the matrix  $\mathbf{A}$  and vector  $\mathbf{b}$ . Define  $\varphi_{ij}^k = \text{AsyVar}(\sqrt{n}\tilde{\sigma}_{ij}^k)$ ,  $\rho_{ij}^{k\ell} = \text{AsyCov}(\sqrt{n}\tilde{\sigma}_{ij}^k, \sqrt{n}\tilde{\sigma}_{ij}^\ell)$ ,  $\gamma_{ij}^{k\ell} = E(\tilde{\sigma}_{ij}^k)E(\tilde{\sigma}_{ij}^\ell)$  and  $\eta_{ij}^k = E(\tilde{\sigma}_{ij}^k)\sigma_{ij}$ , where  $\text{AsyVar}$  and  $\text{AsyCov}$  represent the asymptotic variance and covariance. Let  $m_j = \sum_{i=1}^n \mathbf{x}_i^{(j)} / n$  be the mean value of the  $j$ th column of data matrix  $\mathbb{X}$ . Standard asymptotic theory provides consistent estimators for  $\varphi_{ij}^k$ ,  $\rho_{ij}^{k\ell}$ ,  $\gamma_{ij}^{k\ell}$  and  $\eta_{ij}^k$ , as shown in the following lemmas.

**Lemma 2.2:** A consistent estimator for  $\varphi_{ij}^k$  is given by

$$\hat{\varphi}_{ij}^k = \frac{1}{n} \sum_{t=1}^n [(x_{ti} - m_i)(x_{tj} - m_j) - \tilde{\sigma}_{ij}^k]^2.$$

**Proof:** Without loss of generality, we assume  $i < j$ . Because  $X_i = \sum_{k=1}^i l_{ik}\varepsilon_k$  and  $X_j = \sum_{k=1}^j l_{jk}\varepsilon_k$ , we have  $\text{Cov}(X_i, X_j) = \sum_{k=1}^i l_{ik}l_{jk}d_k^2$ . Hence  $s_{ij} = \sum_{k=1}^i \hat{l}_{ik}\hat{l}_{jk}\hat{d}_k^2 = \tilde{\sigma}_{ij}^k$  without regularization (H. Zheng et al., 2017). By the Lemma 1 in Ledoit and Wolf (2003),  $\hat{\varphi}_{ij}^k$  converges in probability to  $\varphi_{ij}^k$ . ■

**Lemma 2.3:** A consistent estimator for  $\rho_{ij}^{k\ell}$  is given by

$$\hat{\rho}_{ij}^{k\ell} = \frac{1}{n} \sum_{t=1}^n [(x_{ti} - m_i)(x_{tj} - m_j) - \tilde{\sigma}_{ij}^k][(x_{ti} - m_i)(x_{tj} - m_j) - \tilde{\sigma}_{ij}^\ell].$$

**Proof:** It is an easy extension of the result from Lemma 2.2. ■

Under conditions C1–C3 and from Theorem 2.1, estimators for  $\gamma_{ij}^{k\ell}$  and  $\eta_{ij}^k$  are  $\tilde{\sigma}_{ij}^k\tilde{\sigma}_{ij}^\ell$  and  $\tilde{\sigma}_{ij}^k s_{ij}$ , where  $s_{ij}$  is the  $(i, j)$ th element of the sample covariance matrix. Consequently, we can replace  $\varphi_{ij}^k$ ,  $\rho_{ij}^{k\ell}$ ,  $\gamma_{ij}^{k\ell}$  and  $\eta_{ij}^k$  with their corresponding estimators into the optimizations (5) and (7) to solve the weights for each  $\tilde{\Sigma}_k$ .

## 3. Simulation

In this section, the performances of the proposed WAE estimators  $\hat{\Sigma}_{\text{wae}}$  and  $\hat{\Sigma}_{\text{wae}^*}$  are evaluated by the comparison with the method developed by H. Zheng et al. (2017), denoted as

$\widehat{\Sigma}_{\text{zheng}}$ , which will demonstrate the advantages of the optimal weights solved from the optimization problems (5) and (7) over the equal weights for the average ensemble idea in the MCD. Let  $\mathbb{I}_{\{\cdot\}}$  represent the indicator function. We consider seven different setups for the true covariance matrix in the simulation studies.

Scenario-1 (Identity Structure):  $\Sigma_1 = \mathbf{I}$  is the identity matrix.

Scenario-2 (Compact Banded Structure):  $\Sigma_2 = \mathbf{B}^\top \mathbf{B}$ , where  $\mathbf{B} = (b_{st})$  with  $b_{st} = \mathbb{I}_{\{s=t\}} + 0.8\mathbb{I}_{\{s-t=1\}} + 0.6\mathbb{I}_{\{s-t=2\}}$ , is the second-order moving average structure.

Scenario-3 (Permuted Banded Structure):  $\Sigma_3$  is obtained by randomly permutating rows and corresponding columns of  $\Sigma_2$ .

Scenario-4 (Loose Banded Structure):  $\Sigma_4 = \mathbf{B}^\top \mathbf{B}$ , where  $\mathbf{B} = (b_{st})$  with  $b_{st} = \mathbb{I}_{\{s=t\}} + 0.8\mathbb{I}_{\{s-t=1\}} + 0.6\mathbb{I}_{\{2 \leq s-t \leq 5\}}$ .

Scenario-5 (Block Diagonal Structure):  $\Sigma_5 = (\sigma_{st})$ ,  $\sigma_{st} = \mathbb{I}_{\{s=t\}} + 0.5\mathbb{I}_{\{s \neq t, s \leq 20, t \leq 20\}}$ .

Scenario-6 (Dense Structure):  $\Sigma_6 = \mathbf{B}\mathbf{B}^\top$ ,  $\mathbf{B} = (b_{st})$  is a unit lower triangular matrix with  $b_{st}$  independently generated from normal distribution  $N(0, 0.2)$ .

Scenario-7 (Compound Structure):  $\Sigma_7 = (\sigma_{st})$ ,  $\sigma_{st} = \mathbb{I}_{\{s=t\}} + 0.5\mathbb{I}_{\{s \neq t\}}$ .

Scenario-1 is the identity matrix, indicating variables are independent from each other. Scenarios-2 and 4 are banded matrices, indicating each variable is only correlated with its most nearby variables. Scenario-5 has a compound structure with element values 0.5 on the upper left corner. Scenario-6 represents a randomly generated matrix. Scenario-7 indicates that each variable is correlated with others. Therefore, the true covariance matrices in this study include very sparse cases (e.g., Scenario-1, Scenario-2, Scenario-3), moderate sparse cases (e.g., Scenario-4, Scenario-5) as well as the dense cases (e.g., Scenario-6, Scenario-7). In order to measure the accuracy of covariance estimates, we consider the following three loss functions

$$F = \|\widehat{\Sigma} - \Sigma\|_F, \quad \text{MAE} = \frac{1}{p^2} \sum_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \quad \text{and} \quad L_2 = \lambda_{\max}(\widehat{\Sigma} - \Sigma),$$

where  $\widehat{\Sigma} = (\hat{\sigma}_{ij})$  stands for a covariance estimate  $\Sigma = (\sigma_{ij})$ , and  $\lambda_{\max}$  represents the maximum eigenvalue of the matrix. For each scenario, the data are independently generated from multivariate normal distribution  $N(\mathbf{0}, \Sigma)$  with the sample size  $n = 50$ , and the dimensionality  $p = 50, 100$ . The number of variable orderings  $M$  is set to be 30 as suggested by H. Zheng et al. (2017). Such 30 variable orderings are uniformly sampled, that is, they are randomly generated with equal probability from all  $p!$  possible orderings. The LASSO technique of (2) in the proposed method is implemented by the function *glmnet*( $\cdot$ ) in the R software with the value of its tuning parameter selected by the function itself. The set  $\mathcal{A}_\zeta$  for the tuning parameter  $\zeta$  in the optimization problem (7) contains an arithmetic sequence starting from 0.01 to 3 with the common difference 0.05. Table 1 shows the averaged values and corresponding standard errors of loss functions for each compared estimator over 50 replications.

It can be seen that overall two proposed WAE methods are substantially superior to the estimate  $\widehat{\Sigma}_{\text{zheng}}$  in terms of  $F$  and  $L_2$  loss functions for all the considered covariance structures, and they are at least comparable with or slightly better than  $\widehat{\Sigma}_{\text{zheng}}$  with respect to the MAE loss. By comparing the results of Scenario-2 and Scenario-3, we observe that the proposed WAE methods have the similar performances when we permute the rows and

**Table 1.** Means and standard errors (in parentheses) of loss functions from each estimate for normal data.

Scenario	$(n, p) = (50, 50)$			$(n, p) = (50, 100)$		
	$\hat{\Sigma}_{\text{zheng}}$	$\hat{\Sigma}_{\text{wae}}$	$\hat{\Sigma}_{\text{wae}^*}$	$\hat{\Sigma}_{\text{zheng}}$	$\hat{\Sigma}_{\text{wae}}$	$\hat{\Sigma}_{\text{wae}^*}$
	F					
$\Sigma_1$	4.36(0.03)	3.43(0.09)	3.59(0.08)	4.88(0.02)	4.36(0.06)	4.44(0.06)
$\Sigma_2$	10.2(0.05)	7.80(0.06)	7.90(0.07)	15.3(0.05)	12.2(0.05)	12.3(0.06)
$\Sigma_3$	10.3(0.05)	7.94(0.06)	8.00(0.07)	15.3(0.05)	12.2(0.06)	12.3(0.07)
$\Sigma_4$	27.9(0.08)	16.6(0.20)	16.8(0.21)	41.4(0.08)	28.1(0.18)	28.6(0.18)
$\Sigma_5$	8.81(0.04)	6.06(0.17)	6.64(0.16)	9.27(0.03)	7.22(0.14)	7.68(0.13)
$\Sigma_6$	20.6(0.06)	16.8(0.09)	17.1(0.09)	51.5(0.08)	48.6(0.09)	49.1(0.09)
$\Sigma_7$	23.0(0.05)	15.6(0.44)	16.2(0.42)	47.9(0.05)	37.3(0.61)	37.3(0.61)
	$L_2$					
$\Sigma_1$	4.00(0.03)	2.72(0.11)	2.86(0.11)	4.15(0.02)	3.24(0.07)	3.30(0.07)
$\Sigma_2$	3.64(0.02)	3.09(0.03)	3.08(0.04)	3.84(0.02)	3.43(0.02)	3.45(0.02)
$\Sigma_3$	3.68(0.02)	3.15(0.04)	3.17(0.04)	3.84(0.02)	3.44(0.03)	3.45(0.03)
$\Sigma_4$	13.1(0.05)	9.05(0.16)	9.10(0.16)	13.8(0.03)	10.8(0.09)	10.9(0.09)
$\Sigma_5$	8.64(0.04)	5.64(0.20)	6.19(0.17)	8.90(0.03)	6.55(0.16)	6.90(0.15)
$\Sigma_6$	9.49(0.05)	7.46(0.09)	7.56(0.09)	18.0(0.07)	16.8(0.11)	16.9(0.10)
$\Sigma_7$	22.9(0.05)	15.3(0.45)	16.0(0.44)	47.8(0.05)	37.1(0.62)	37.0(0.62)
	MAE					
$\Sigma_1$	0.026(0.000)	0.024(0.001)	0.024(0.001)	0.012(0.000)	0.011(0.000)	0.011(0.000)
$\Sigma_2$	0.083(0.000)	0.082(0.001)	0.081(0.001)	0.050(0.000)	0.048(0.000)	0.048(0.000)
$\Sigma_3$	0.084(0.000)	0.083(0.001)	0.083(0.001)	0.050(0.000)	0.048(0.000)	0.048(0.000)
$\Sigma_4$	0.271(0.001)	0.203(0.002)	0.203(0.002)	0.149(0.000)	0.128(0.001)	0.128(0.001)
$\Sigma_5$	0.079(0.000)	0.056(0.002)	0.061(0.001)	0.024(0.000)	0.021(0.000)	0.022(0.000)
$\Sigma_6$	0.315(0.001)	0.266(0.001)	0.269(0.001)	0.398(0.001)	0.381(0.001)	0.385(0.001)
$\Sigma_7$	0.458(0.001)	0.299(0.009)	0.311(0.009)	0.478(0.001)	0.366(0.007)	0.364(0.007)

**Table 2.** Means and standard errors (in parentheses) for the number of non-zeros weights for the proposed estimates.

Scenario	$(n, p) = (50, 50)$		$(n, p) = (50, 100)$	
	$\hat{\Sigma}_{\text{wae}}$	$\hat{\Sigma}_{\text{wae}^*}$	$\hat{\Sigma}_{\text{wae}}$	$\hat{\Sigma}_{\text{wae}^*}$
$\Sigma_1$	7.12(0.27)	3.68(0.21)	8.46(0.33)	5.22(0.17)
$\Sigma_2$	9.02(0.30)	4.86(0.22)	9.86(0.25)	5.34(0.15)
$\Sigma_3$	8.94(0.26)	6.00(0.15)	9.76(0.28)	6.18(0.15)
$\Sigma_4$	7.64(0.27)	4.66(0.16)	8.62(0.25)	4.48(0.14)
$\Sigma_5$	5.52(0.28)	2.32(0.08)	6.32(0.26)	2.86(0.16)
$\Sigma_6$	12.1(0.27)	6.76(0.21)	10.5(0.32)	6.70(0.15)
$\Sigma_7$	5.58(0.29)	3.24(0.12)	4.62(0.21)	2.92(0.16)

columns of the banded covariance matrix randomly. Besides, by the comparison of two proposed methods, we observe that the estimate  $\hat{\Sigma}_{\text{wae}^*}$  is almost comparable with but sometimes slightly worse the estimate  $\hat{\Sigma}_{\text{wae}}$ . A possible reason can be found in Table 2, which displays the number of non-zero elements in the estimated weights  $\hat{\omega}$  and  $\hat{\omega}^*$  from the optimizations (5) and (7), respectively. We observe that the ensemble estimate  $\hat{\Sigma}_{\text{wae}^*}$  aggregates less number of candidate estimates than  $\hat{\Sigma}_{\text{wae}}$ , implying that it reduces the model complexity and thus picks up the most useful variable orderings for constructing the covariance estimate. Therefore, the estimate  $\hat{\Sigma}_{\text{wae}}$  might be a little more accurate sometimes, and the estimate  $\hat{\Sigma}_{\text{wae}^*}$  leads to a simpler model. Furthermore from Table 2 we conclude that in practice it is no need to calculate every candidate estimate  $\tilde{\Sigma}_k$  with respect to all the  $p!$  variable orderings, which is very computationally expensive, since there are only a few orderings that are the most important for constructing the ensemble matrix estimate.

Notice that the MCD technique empirically does not require the normal property of data to estimate the covariance matrix. It is thus interesting to examine the robust performance

**Table 3.** Means and standard errors (in parentheses) of loss functions from each estimate for non-normal data.

Case		$(n, p) = (50, 50)$			$(n, p) = (50, 100)$		
Case		$\widehat{\Sigma}_{\text{zheng}}$	$\widehat{\Sigma}_{\text{wae}}$	$\widehat{\Sigma}_{\text{wae}^*}$	$\widehat{\Sigma}_{\text{zheng}}$	$\widehat{\Sigma}_{\text{wae}}$	$\widehat{\Sigma}_{\text{wae}^*}$
F	(I)	16.4(1.32)	14.0(1.47)	14.4(1.53)	46.3(0.14)	37.7(2.30)	38.7(2.52)
	(II)	23.3(0.13)	17.0(0.62)	17.0(0.63)	48.0(0.07)	37.9(0.71)	37.8(0.71)
$L_2$	(I)	16.1(1.30)	12.6(1.44)	12.9(1.50)	46.0(0.14)	34.6(2.35)	35.1(2.60)
	(II)	23.1(0.06)	16.5(0.56)	16.5(0.57)	47.9(0.05)	37.7(0.71)	37.6(0.71)
MAE	(I)	0.323(0.026)	0.233(0.023)	0.238(0.024)	0.460(0.001)	0.328(0.013)	0.334(0.014)
	(II)	0.464(0.002)	0.327(0.013)	0.326(0.013)	0.479(0.001)	0.372(0.007)	0.371(0.008)

of the proposed WAE methods for the non-normal data. Now we conduct a simulation study using  $\Sigma_7$  as the true covariance matrix, and consider the following two cases of data generating processes with  $n = 50$  and  $p = 50, 100$ .

*Case (I).* Independently generate data from  $t$  distribution with degrees of freedom 4, as well as mean  $\mathbf{0}$  and  $\Sigma_7$  as its scale matrix.

*Case (II).* Independently generate data from a mixed distribution  $(1 - b)N(\mathbf{0}, \Sigma_7) + bT$ , where  $b$  is a random number following Bernoulli distribution with the probability of success equal to 0.1. Here  $T$  represents  $t$  distribution with degrees of freedom 4, as well as mean  $\mathbf{5}$  and the identity matrix as its scale matrix.

The data in Case (I) come from a heavy-tailed distribution, and Case (II) presents a set of contaminated data. Table 3 displays the averaged values and corresponding standard errors of loss functions from each competitor over 50 replications. Overall, the proposed WAE methods outperform the estimate  $\widehat{\Sigma}_{\text{zheng}}$  substantially under the considered losses, implying the advantages of the ensemble idea with the optimally selected weights for different candidate covariance estimates.

To sum up, the simulation studies demonstrate that the proposed WAE methods improve the performance of the ensemble estimation for covariance matrix under the framework of MCD technique by assigning each covariance candidate estimate with different weights that are obtained from minimizing the risk functions as in (5) and (7).

#### 4. Case study of portfolio allocation

To further explore the performances of the proposed WAE methods, we apply them to analyse a real stock market data, where the estimated covariance matrix is subsequently used for the decision on the portfolio allocation.

The common portfolio strategy (Markowitz, 1952) tries to minimize the risk of an expected return level by diversifying the investment of various assets. Let  $\mathbf{w} = (w_1, \dots, w_p)^\top$  represent the proportions of a set of assets in the portfolio, and  $\Sigma$  denote the volatility matrix of the returns in the asset pool. The minimum variance portfolio selection problem is formulated as

$$\min \mathbf{w}^\top \Sigma \mathbf{w} \quad \text{s.t.} \quad \sum_{i=1}^p w_i = 1 \quad \text{and} \quad w_i \geq 0, \quad i = 1, \dots, p. \quad (8)$$

Since the objective function in (8) involves the covariance matrix, an ill-conditioned matrix estimate may lead to an unstable solution of  $\mathbf{w}$ , and thus greatly magnifies the error of portfolio allocation. In addition, when a large amount of stocks is considered for the investment, a

**Table 4.** The performances of weekly returns of portfolio for compared methods.

	AWR	SE	AWR/SE
$\widehat{\Sigma}_{\text{zheng}}$	0.647	0.565	1.145
$\widehat{\Sigma}_{\text{wae}}$	0.764	0.604	1.265
$\widehat{\Sigma}_{\text{wae}^*}$	0.795	0.606	1.296

valid estimate of volatility matrix that accommodates for the high dimensions is needed. Furthermore, because there is no natural variable ordering among stocks, the idea of ensemble model by the MCD is suitable for this analysis. We hence apply the proposed WAE methods for the covariance estimation, which may lead to a good subsequent portfolio allocation strategy. We compare the performances of the proposed methods with  $\widehat{\Sigma}_{\text{zheng}}$  as in the simulation studies.

The data are consisted of the stock returns collected weekly from years 2020 to 2021 from the CSI300 Index, which contain  $p = 100$  randomly selected stocks with  $n = 104$  observations (52 weeks for each year). The first  $j$  observations are used to construct the covariance estimate  $\widehat{\Sigma}_j$ ,  $j = 52, \dots, 103$ , for each method. Then the estimated portfolio weight  $\widehat{\mathbf{w}}_{j+1}$  is the solution of the optimization problem (8) by replacing  $\Sigma$  with its estimate  $\widehat{\Sigma}_j$ . We measure the performance of the estimated portfolio by the realized averaged weekly return (AWR) for the year of 2021, which is defined as

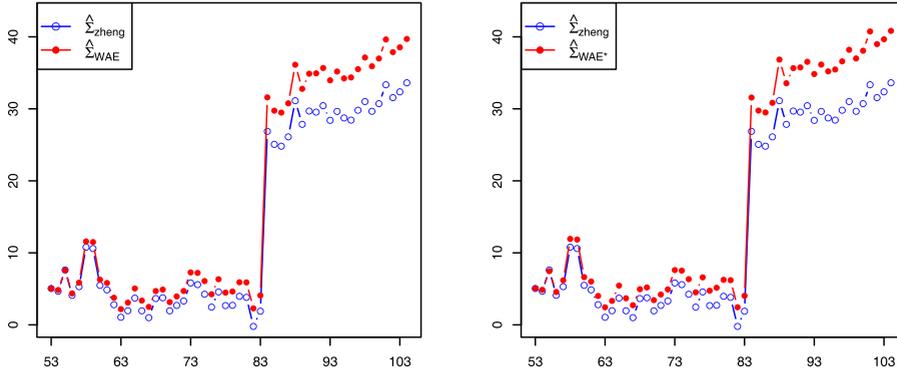
$$\text{AWR} = \frac{1}{52} \sum_{j=52}^{103} \widehat{\mathbf{w}}_{j+1}^{\top} \mathbf{x}_{j+1},$$

as well as their standard errors (SE) and the information ratio AWR/SE. Although a larger AWR is the main pursue for the investors since it is the profit, a higher information ratio is also desired as it reflects both profit and risk.

Table 4 presents the compared portfolio performances obtained by the proposed WAE methods and  $\widehat{\Sigma}_{\text{zheng}}$  for year 2021. We observe that although the proposed models produce a slightly larger SE, they have higher AWR and AWR/SD values compared with the method of  $\widehat{\Sigma}_{\text{zheng}}$ , which indicates that the proposed methods gain more realized returns. Furthermore, Figure 1 are the plots of the cumulative weekly returns for different compared methods. Because the most parts of lines produced by the estimates  $\widehat{\Sigma}_{\text{wae}^*}$  and  $\widehat{\Sigma}_{\text{wae}}$  overlap each other as they behave similarly in this study, we plot their cumulative weekly returns separately in two plots in order to see them clearly in comparison with the estimate  $\widehat{\Sigma}_{\text{zheng}}$ . From the plots it is observed that the cumulative weekly returns of the proposed WAE methods are consistently higher than that of the estimate  $\widehat{\Sigma}_{\text{zheng}}$ . Especially as the investment time increases, the advantages of the proposed methods over the estimate  $\widehat{\Sigma}_{\text{zheng}}$  are much more evidenced. Therefore the proposed WAE methods are able to provide a valid strategy for the portfolio allocation for this set of stock data, verifying the merits of using different weights in the ensemble model of MCD for this study.

## 5. Discussion

In this work, we propose a weighted average ensemble covariance matrix estimation via the MCD technique. The weights are adaptively determined by minimizing a risk function in terms of the Frobenius norm. The risk function can be transformed into a matrix quadratic



**Figure 1.** The cumulative weekly returns from compared methods for portfolio study.

form with respect to the weights such that it can be solved easily by the off-the-shelf softwares. At the same time, the proposed methods are able to automatically distinguish which variable orderings utilized in the MCD are useful, and thus preclude from the ensemble model the candidate covariance estimates corresponding to the useless variable orderings to reduce the model complexity. The asymptotic convergence rates of the proposed methods are established, and the numerical studies demonstrate the merits of the proposed methods that use the optimal weights under the MCD framework.

There are several research directions along this line in the future work. Firstly, we would like to point out that although a simulation is performed to examine the robustness of the proposed methods, they are not designed for the non-normal data. One may employ some robust techniques such as Huber method instead of LASSO in Equation (2) to decrease the impact of the outliers or heavy-tailed data on the linear models, hence leading to a robust covariance estimate (Wang et al., 2023). Secondly, we might improve the proposed WAE by constructing an appropriate objective function regarding the Cholesky factor matrices  $\mathbf{L}$  and  $\mathbf{D}$  rather than  $\Sigma$  for the weights optimization. Such weights are used to compute the weighted averaging values on  $\tilde{\mathbf{L}}_k$  and  $\tilde{\mathbf{D}}_k$ , i.e.,  $\hat{\mathbf{L}} = \sum_{k=1}^M \omega_k \tilde{\mathbf{L}}_k$  and  $\hat{\mathbf{D}} = \sum_{k=1}^M \omega_k \tilde{\mathbf{D}}_k$ . Then a covariance estimate is  $\hat{\Sigma} = \hat{\mathbf{L}}\hat{\mathbf{D}}\hat{\mathbf{L}}^\top$  which would have a smaller variance than  $\hat{\Sigma}_{\text{wae}}$ , hence leading to a more accurate estimate (Kang & Deng, 2020). However, solving such objective functions in terms of  $\mathbf{L}$  and  $\mathbf{D}$  is more complicated. Another idea to determine a covariance matrix estimate from matrices  $\tilde{\Sigma}_k$  is based on the concepts of Bures-Wasserstein distance and the Fréchet mean, which are used to measure the central tendency of a set of positive semi-definite matrices (J. Zheng et al., 2023). However, whether such technique is suitable for the Cholesky-based covariance estimation still needs further investigation.

## Acknowledgments

The authors thank the editor, associate editors and referees for their constructive comments.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Kang's work was supported by National Natural Science Foundation of China [grant numbers 72232001 and 72371059]. Gao's research was supported by National Natural Science Foundation of China [grant number 12001365].

## Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

- Bickel, P. J., & Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6), 2577–2604.
- Bickel, P. J., & Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1), 199–227. <https://doi.org/10.1214/009053607000000758>
- Bien, J., & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4), 807–820. <https://doi.org/10.1093/biomet/asr054>
- Cai, T. T., Ren, Z., & Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1), 1–59.
- Cao, Y., Lin, W., & Li, H. (2019). Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association*, 114(526), 759–772. <https://doi.org/10.1080/01621459.2018.1442340>
- Chang, C., & 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. <https://doi.org/10.1016/j.jspi.2010.04.048>
- Dellaportas, P., & Pourahmadi, M. (2012). Cholesky-GARCH models with applications to finance. *Statistics and Computing*, 22(4), 849–855. <https://doi.org/10.1007/s11222-011-9251-2>
- Deng, X., & Tsui, K. W. (2013). Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics*, 22(2), 494–512. <https://doi.org/10.1080/10618600.2012.715556>
- Furrer, R., & Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2), 227–255. <https://doi.org/10.1016/j.jmva.2006.08.003>
- Huang, C., Farewell, D., & Pan, J. (2017). A calibration method for non-positive definite covariance matrix in multivariate data analysis. *Journal of Multivariate Analysis*, 157, 45–52. <https://doi.org/10.1016/j.jmva.2017.03.001>
- Huang, J., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1), 85–98. <https://doi.org/10.1093/biomet/93.1.85>
- Jenny Shi, W., Hannig, J., Lai, R. C., & Lee, T. C. (2021). Covariance estimation via fiducial inference. *Statistical Theory and Related Fields*, 5(4), 316–331. <https://doi.org/10.1080/24754269.2021.1877950>
- Kang, X., & Deng, X. (2020). An improved modified cholesky decomposition approach for precision matrix estimation. *Journal of Statistical Computation and Simulation*, 90(3), 443–464. <https://doi.org/10.1080/00949655.2019.1687701>
- Kang, X., & Deng, X. (2021). On variable ordination of Cholesky-based estimation for a sparse covariance matrix. *Canadian Journal of Statistics*, 49(2), 283–310. <https://doi.org/10.1002/cjs.v49.2>
- Kang, X., Lian, J., & Deng, X. (2025). On block Cholesky decomposition for sparse inverse covariance estimation. *Statistica Sinica*, 35(3), 1–22.
- Kang, X., & Wang, M. (2021). Ensemble sparse estimation of covariance structure for exploring genetic disease data. *Computational Statistics & Data Analysis*, 159, 107220. <https://doi.org/10.1016/j.csda.2021.107220>

- Kidd, B., & Katzfuss, M. (2022). Bayesian nonstationary and nonparametric covariance estimation for large spatial data (with discussion). *Bayesian Analysis*, 17(1), 291–351. <https://doi.org/10.1214/21-BA1273>
- Lam, C., & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B), 4254–4278. <https://doi.org/10.1214/09-AOS720>
- Ledoit, O., & 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. [https://doi.org/10.1016/S0927-5398\(03\)00007-0](https://doi.org/10.1016/S0927-5398(03)00007-0)
- Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5), 3043–3065. <https://doi.org/10.1214/19-AOS1921>
- Leng, C., & Li, B. (2011). Forward adaptive banding for estimating large covariance matrices. *Biometrika*, 98(4), 821–830. <https://doi.org/10.1093/biomet/asr045>
- Li, C., Yang, M., Wang, M., Kang, H., & Kang, X. (2021). A Cholesky-based sparse covariance estimation with an application to genes data. *Journal of Biopharmaceutical Statistics*, 31(5), 603–616. <https://doi.org/10.1080/10543406.2021.1931270>
- Liang, W., Zhang, Y., Wang, J., Wu, Y., & Ma, X. (2024). A new approach for ultrahigh dimensional precision matrix estimation. *Journal of Statistical Planning and Inference*, 232, 106164. <https://doi.org/10.1016/j.jspi.2024.106164>
- Lv, J., Guo, C., & Wu, J. (2018). Smoothed empirical likelihood inference via the modified Cholesky decomposition for quantile varying coefficient models with longitudinal data. TEST 1–34.
- Markowitz, H. (1952). Portfolio selection\*. *The Journal of Finance*, 7(1), 77–91.
- Pedeli, X., Fokianos, K., & Pourahmadi, M. (2015). Two Cholesky-log-GARCH models for multivariate volatilities. *Statistical Modeling*, 15(3), 233–255. <https://doi.org/10.1177/1471082X14551246>
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3), 677–690. <https://doi.org/10.1093/biomet/86.3.677>
- Rajaratnam, B., & Salzman, J. (2013). Best permutation analysis. *Journal of Multivariate Analysis*, 121, 193–223. <https://doi.org/10.1016/j.jmva.2013.03.001>
- Rothman, A. J., Bickel, P. J., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2, 494–515. <https://doi.org/10.1214/08-EJS176>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Wagaman, A., & Levina, E. (2009). Discovering sparse covariance structures with the isomap. *Journal of Computational and Graphical Statistics*, 18(3), 551–572. <https://doi.org/10.1198/jcgs.2009.08021>
- Wang, S., Xie, C., & Kang, X. (2023). A novel robust estimation for high-dimensional precision matrices. *Statistics in Medicine*, 42(5), 656–675. <https://doi.org/10.1002/sim.v42.5>
- Wu, W., & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4), 831–844. <https://doi.org/10.1093/biomet/90.4.831>
- Xin, H., & Zhao, S. D. (2023). A compound decision approach to covariance matrix estimation. *Biometrics*, 79(2), 1201–1212. <https://doi.org/10.1111/biom.13686>
- Xue, L., Ma, S., & Zou, H. (2012). Positive-definite  $L_1$ -penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500), 1480–1491. <https://doi.org/10.1080/01621459.2012.725386>
- Zhang, Y., Shen, W., & Kong, D. (2022). Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, 117, 1–12. <https://doi.org/10.1080/01621459.2021.1906684>
- Zheng, J., Huang, H., Yi, Y., Li, Y., & Lin, S. C. (2023). Barycenter estimation of positive semi-definite matrices with Bures-Wasserstein distance. arXiv:2302.14618.
- Zheng, H., Tsui, K. W., Kang, X., & Deng, X. (2017). Cholesky-based model averaging for covariance matrix estimation. *Statistical Theory and Related Fields*, 1(1), 48–58. <https://doi.org/10.1080/24754269.2017.1336831>
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429. <https://doi.org/10.1198/016214506000000735>

## Appendix. Proof of Theorem 2.1

In this section, we provide the technical proofs of the main theoretical results in the paper. To prove Theorem 2.1, we need to present three lemmas.

**Lemma A.1:** *Assume a positive definite matrix  $\Sigma$  has its modified Cholesky decomposition  $\Sigma = \mathbf{L}\mathbf{D}\mathbf{L}^\top$ . Under condition C1 (that is, the singular values of  $\Sigma$  are bounded), there exist constants  $l_2$  and  $u_2$  such that*

$$\begin{aligned} l_2 &< sv_p(\mathbf{L}) \leq sv_1(\mathbf{L}) < u_2, \\ l_2 &< sv_p(\mathbf{D}) \leq sv_1(\mathbf{D}) < u_2. \end{aligned}$$

The proof of Lemma A.1 is similar to that of Lemma 3 in Kang et al. (2025), which is omitted here.

**Lemma A.2:** *Suppose that the data are independently and identically distributed from  $N(\mathbf{0}, \Sigma_0)$ . Under the conditions C1 and C2, the estimates of Cholesky factor matrices  $\widehat{\mathbf{L}}_\Pi$  and  $\widehat{\mathbf{D}}_\Pi$  corresponding to a permutation  $\Pi$  have the following consistent properties*

$$\begin{aligned} \|\widehat{\mathbf{L}}_\Pi - \mathbf{L}_{0\Pi}\|_F &= O_p(\sqrt{s \log(p)/n}), \\ \|\widehat{\mathbf{D}}_\Pi - \mathbf{D}_{0\Pi}\|_F &= O_p(\sqrt{p \log(p)/n}). \end{aligned}$$

**Proof:** To simplify the notations, we prove it without the symbol  $\Pi$  since the conclusions hold for any variable orderings. From Equation (1), it is easy to see  $\boldsymbol{\epsilon} = \mathbf{L}^{-1}\mathbf{X} \sim N(\mathbf{0}, \mathbf{D})$ . Hence, the penalized negative log likelihood, up to some constant, is

$$\Phi(\mathbf{D}, \mathbf{L}) = (\log |\mathbf{D}| + \text{tr} [(\mathbf{L}^\top)^{-1} \mathbf{D}^{-1} \mathbf{L}^{-1} \mathbf{S}]) + \sum_{j=1}^p \lambda_j \sum_{k < j} |l_{jk}|.$$

Define  $G(\Delta_L, \Delta_D) = \Phi(\mathbf{D}_0 + \Delta_D, \mathbf{L}_0 + \Delta_L) - \Phi(\mathbf{D}_0, \mathbf{L}_0)$ . Let  $\mathcal{B}_1 = \{\Delta_L : \|\Delta_L\|_F^2 \leq \kappa_1 s \log(p)/n\}$  and  $\mathcal{B}_2 = \{\Delta_D : \|\Delta_D\|_F^2 \leq \kappa_2 p \log(p)/n\}$ , where  $\kappa_1 \geq 0$  and  $\kappa_2 \geq 0$  are constants. Denote  $\partial\mathcal{B}_1$  and  $\partial\mathcal{B}_2$  as the boundaries of  $\mathcal{B}_1$  and  $\mathcal{B}_2$ . We will show that for each  $\Delta_L \in \partial\mathcal{B}_1$  and  $\Delta_D \in \partial\mathcal{B}_2$ , probability  $\Pr(G(\Delta_L, \Delta_D) > 0) \rightarrow 1$  as  $n \rightarrow \infty$  for sufficiently large  $\kappa_1$  and  $\kappa_2$ . Additionally, since  $G(\Delta_L, \Delta_D) = 0$  when  $\Delta_L = 0$  and  $\Delta_D = 0$ ,  $G(\Delta_L, \Delta_D)$  achieves its minimum value when  $\Delta_L \in \mathcal{B}_1$  and  $\Delta_D \in \mathcal{B}_2$ . That is  $\|\Delta_L\|_F^2 = O_p(s \log(p)/n)$  and  $\|\Delta_D\|_F^2 = O_p(p \log(p)/n)$ .

Assume  $\|\Delta_L\|_F^2 = \kappa_1 s \log(p)/n$  and  $\|\Delta_D\|_F^2 = \kappa_2 p \log(p)/n$ . According to Lemma A.1 and condition C1, there exists a constant  $\delta$  satisfying  $0 < 1/\delta < sv_p(\mathbf{L}_0) \leq sv_1(\mathbf{L}_0) < \delta < \infty$  and  $0 < 1/\delta < sv_p(\mathbf{D}_0) \leq sv_1(\mathbf{D}_0) < \delta < \infty$ . Let  $\mathbf{D} = \mathbf{D}_0 + \Delta_D$  and  $\mathbf{L} = \mathbf{L}_0 + \Delta_L$ , and then we decompose  $G(\Delta_L, \Delta_D)$  into five parts and then bound them separately.

$$\begin{aligned} G(\Delta_L, \Delta_D) &= \Phi(\mathbf{D}, \mathbf{L}) - \Phi(\mathbf{D}_0, \mathbf{L}_0) \\ &= \log |\mathbf{D}| - \log |\mathbf{D}_0| + \text{tr} \left( (\mathbf{L}^\top)^{-1} \mathbf{D}^{-1} \mathbf{L}^{-1} \mathbf{S} \right) - \text{tr} \left( (\mathbf{L}_0^\top)^{-1} \mathbf{D}_0^{-1} \mathbf{L}_0^{-1} \mathbf{S} \right) \\ &\quad + \sum_{j=1}^p \lambda_j \sum_{k < j} |l_{jk}| - \sum_{j=1}^p \lambda_j \sum_{k < j} |l_{0jk}| \\ &= \log |\mathbf{D}| - \log |\mathbf{D}_0| + \text{tr} [(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}) \mathbf{D}_0] - \text{tr} [(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}) \mathbf{D}_0] \\ &\quad + \text{tr} \left[ (\mathbf{L}^\top)^{-1} \mathbf{D}^{-1} \mathbf{L}^{-1} \mathbf{S} \right] - \text{tr} \left[ (\mathbf{L}_0^\top)^{-1} \mathbf{D}_0^{-1} \mathbf{L}_0^{-1} \mathbf{S} \right] \\ &\quad + \sum_{j=1}^p \lambda_j \sum_{k < j} |l_{jk}| - \sum_{j=1}^p \lambda_j \sum_{k < j} |l_{0jk}| \\ &= \log |\mathbf{D}| - \log |\mathbf{D}_0| + \text{tr} [(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}) \mathbf{D}_0] + \text{tr} \left[ (\mathbf{L}^\top)^{-1} \mathbf{D}^{-1} \mathbf{L}^{-1} \mathbf{S} \right] \end{aligned}$$

$$\begin{aligned}
 & -\operatorname{tr}\left[(\mathbf{L}^\top)^{-1}\mathbf{D}_0^{-1}\mathbf{L}^{-1}\mathbf{s}\right] + \operatorname{tr}\left[(\mathbf{L}^\top)^{-1}\mathbf{D}_0^{-1}\mathbf{L}^{-1}\mathbf{s}\right] \\
 & + \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{jk}| - \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{0jk}| \\
 = & \log|\mathbf{D}| - \log|\mathbf{D}_0| + \operatorname{tr}\left[(\mathbf{D}^{-1} - \mathbf{D}_0^{-1})\mathbf{D}_0\right] \\
 & + \operatorname{tr}\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\right)\left[\mathbf{L}^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}^\top)^{-1}\right] \\
 & + \operatorname{tr}\left[\mathbf{D}_0^{-1}(\mathbf{L}^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}^\top)^{-1} - \mathbf{L}_0^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}_0^\top)^{-1})\right] \\
 & + \operatorname{tr}\left[\mathbf{D}_0^{-1}(\mathbf{L}^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}^\top)^{-1} - \mathbf{L}_0^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}_0^\top)^{-1})\right] \\
 & + \operatorname{tr}\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\right)\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}^\top)^{-1} - \mathbf{D}_0\right) \\
 & + \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{jk}| - \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{0jk}| \\
 = & \log|\mathbf{D}| - \log|\mathbf{D}_0| + \operatorname{tr}\left[(\mathbf{D}^{-1} - \mathbf{D}_0^{-1})\mathbf{D}_0\right] \\
 & + \operatorname{tr}\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\right)\left[\mathbf{L}^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}_0^\top)^{-1}\right] \\
 & + \operatorname{tr}\left[\mathbf{D}_0^{-1}(\mathbf{L}^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}^\top)^{-1} - \mathbf{L}_0^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}_0^\top)^{-1})\right] \\
 & + \operatorname{tr}\left[\mathbf{D}^{-1}(\mathbf{L}^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}^\top)^{-1} - \mathbf{L}_0^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}_0^\top)^{-1})\right] \\
 & + \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{jk}| - \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{0jk}| \\
 \triangleq & R_1 + R_2 + R_3 + R_4 + R_5,
 \end{aligned}$$

where

$$\begin{aligned}
 R_1 &= \log|\mathbf{D}| - \log|\mathbf{D}_0| + \operatorname{tr}\left[(\mathbf{D}^{-1} - \mathbf{D}_0^{-1})\mathbf{D}_0\right], \\
 R_2 &= \operatorname{tr}\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\right)\left[\mathbf{L}^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}^\top)^{-1}\right], \\
 R_3 &= \operatorname{tr}\left[\mathbf{D}_0^{-1}(\mathbf{L}^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}^\top)^{-1} - \mathbf{L}_0^{-1}(\mathbf{s} - \boldsymbol{\Sigma}_0)(\mathbf{L}_0^\top)^{-1})\right], \\
 R_4 &= \operatorname{tr}\left[\mathbf{D}^{-1}(\mathbf{L}^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}^\top)^{-1} - \mathbf{L}_0^{-1}\boldsymbol{\Sigma}_0(\mathbf{L}_0^\top)^{-1})\right], \\
 R_5 &= \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{jk}| - \sum_{j=1}^p \lambda_j \sum_{k<j} |l_{0jk}|.
 \end{aligned}$$

From the proof of Lemma 3 in Kang and Deng (2021), we can derive

$$(1) R_1 \geq 1/8\delta^4 \|\Delta_D\|_F^2;$$

(2) for any  $\epsilon > 0$ , there exist constants  $V_1 > 0$  and  $V_2 > 0$  such that  $|R_2| \leq V_1 \sqrt{p \log(p)/n} \|\Delta_D\|_F$ ; and  $R_4 - |R_3| > 1/2\delta^4 \|\Delta_L\|_F^2 - V_2 \sqrt{\log(p)/n} \sum_{(j,k) \in W^c} |l_{jk}| - V_2 \sqrt{s \log(p)/n} \|\Delta_L\|_F$ , where the set  $W = \{(j, k) : k < j, l_{0jk} \neq 0\}$ .

Next, the term  $R_5$  is further decomposed as

$$R_5 = \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W^c} |l_{jk}| + \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W} (|l_{jk}| - |l_{0jk}|) = R_5^{(1)} + R_5^{(2)},$$

where  $R_5^{(1)} = \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W^c} |l_{jk}|$ , and

$$\left| R_5^{(2)} \right| = \left| \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W} (|l_{jk}| - |l_{0jk}|) \right| \leq \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W} |l_{jk} - l_{0jk}| \leq \sum_{j=1}^p \lambda_j \sqrt{s} \|\Delta_L\|_F.$$

Combining all the bounded terms  $R_1$  to  $R_5$  together, with probability greater than  $1 - 2\epsilon$ , we have

$$\begin{aligned} & |G(\Delta_L, \Delta_D)| \\ & \geq R_1 - |R_2| + R_4 - |R_3| + R_5^{(1)} - |R_5^{(2)}| \\ & \geq \frac{1}{8\delta^4} \|\Delta_D\|_F^2 - V_1 \sqrt{p \log(p)/n} \|\Delta_D\|_F + \frac{1}{2\delta^4} \|\Delta_L\|_F^2 - V_2 \sqrt{\log(p)/n} \sum_{(j,k) \in W^c} |l_{jk}| \\ & \quad - V_2 \sqrt{s \log(p)/n} \|\Delta_L\|_F + \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W^c} |l_{jk}| - \sum_{j=1}^p \lambda_j \sqrt{s} \|\Delta_L\|_F \\ & = \frac{\kappa_2 p \log(p)}{8\delta^4 n} - \frac{V_1 \sqrt{\kappa_2 p \log(p)}}{n} + \frac{\kappa_1 s \log(p)}{2\delta^4 n} - V_2 \sqrt{\log(p)/n} \sum_{(j,k) \in W^c} |l_{jk}| \\ & \quad - \frac{V_2 \sqrt{\kappa_1 s \log(p)}}{n} + \sum_{j=1}^p \lambda_j \sum_{(j,k) \in W^c} |l_{jk}| - s \sqrt{\kappa_1 \log(p)/n} \sum_{j=1}^p \lambda_j \\ & = \frac{\sqrt{\kappa_2 p \log(p)}}{n} \left( \frac{\sqrt{\kappa_2}}{8\delta^4} - V_1 \right) + \frac{\sqrt{\kappa_1 s p \log(p)}}{n} \left( \frac{\sqrt{\kappa_1}}{2\delta^4} - \frac{\sum_{j=1}^p \lambda_j}{\sqrt{\log(p)/n}} - V_2 \right) \\ & \quad + \sum_{(j,k) \in W^c} |l_{jk}| \left( \sum_{j=1}^p \lambda_j - V_2 \sqrt{\log(p)/n} \right). \end{aligned}$$

Assume  $\sum_{j=1}^p \lambda_j = V_3 (\sqrt{\log(p)/n})$  where  $V_3 > V_2$ , and choose  $\kappa_1 > 4\delta^8 (K + V_2)^2$  as well as  $\kappa_2 > 64\delta^8 V_1^2$ . Then  $G(\Delta_L, \Delta_D) > 0$ , which completes the proof.  $\blacksquare$

**Lemma A.3:** Suppose that the data are independently and identically distributed from  $N(\mathbf{0}, \Sigma_0)$ . Under the conditions C1 and C2, we have

$$\|\tilde{\Sigma}_k - \Sigma_0\|_F = O_p \left( \sqrt{\frac{(s+p) \log(p)}{n}} \right), \quad k = 1, \dots, M.$$

**Proof:** From the proof of Theorem 2 in Kang and Deng (2021), it follows that

$$\|\tilde{\Sigma}_k - \Sigma_0\|_F^2 = O_p \left( \|\tilde{L}_k - L_0\|_F^2 \right) + O_p \left( \|\tilde{D}_k - D_0\|_F^2 \right) = O_p \left( \frac{(s+p) \log(p)}{n} \right),$$

where the second equation results from Lemma A.2.  $\blacksquare$

Lemma A.3 establishes the convergence rate for each candidate estimate  $\tilde{\Sigma}_k$  for any variable orderings, based on which we prove the main results of Theorem 2.1.

**Proof of Theorem 2.1:** By Equation (4), we obtain

$$\|\hat{\Sigma}_{\text{wae}} - \Sigma_0\|_F = \left\| \sum_{k=1}^M \omega_k \tilde{\Sigma}_k - \Sigma_0 \right\|_F = \left\| \sum_{k=1}^M \omega_k (\tilde{\Sigma}_k - \Sigma_0) \right\|_F.$$

Based on the property of Frobenius norm, together with Lemma A.3, we have

$$\|\widehat{\Sigma}_{\text{wae}} - \Sigma_0\|_F \leq \sum_{k=1}^M \|\omega_k(\widetilde{\Sigma}_k - \Sigma_0)\|_F = \sum_{k=1}^M \omega_k \|\widetilde{\Sigma}_k - \Sigma_0\|_F = O_p\left(\sqrt{\frac{(s+p)\log(p)}{n}}\right).$$

The proofs for the convergence rate of the estimator  $\widehat{\Sigma}_{\text{wae}^*}$  follow the same principles. ■