Journal of Multivariate Analysis 212 (2026) 105553

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On convergence of regularized covariance estimator based on modified Cholesky decomposition☆

Yuli Liang [a], Deliang Dai [b,*], Shaobo Jin [c]

[a] School of Mathematics and Statistics, Guangxi Normal University, Guilin, 541004, China
[b] Department of Economics and Statistics, Linnaeus University, SE 351 95, Växjö, Sweden
[c] Department of Mathematics, Uppsala University, SE 751 06, Uppsala, Sweden

ARTICLE INFO

AMS 2020 subject classifications:
primary 62H12
62J07
secondary 62E20

Keywords:
Covariance matrix estimation
Modified Cholesky decomposition
Shrinkage
Smoothing

ABSTRACT

The regularization for covariance matrix is a widely used technique when estimating large covariance matrices. This paper examines a penalized likelihood method for constructing a statistically efficient covariance matrix estimator. Modified Cholesky decomposition (MCD) is used to parameterize the covariance matrix and the effective regularization scheme is achieved by combining both shrinkage and smoothing penalties on the Cholesky factor. The practical performance is at odds with an absence of theoretical properties of the derived estimators in the literature. In this work, we aim to fill the gap between theory and practice by establishing the convergence properties under regularity conditions. We also provide a simulation study as numerical illustrations.

1. Introduction

In recent years, there has been a significant increase in the demand for data-intensive applications. The use of high-dimensional data, where the number of variables p is comparable to the sample size n, has become commonplace in various fields such as portfolio allocation, multiple-input multiple-output (MIMO) in wireless technology, quantum physics, and analysis of gene expression data. However, the analysis of large dimensional data presents both opportunities in providing a significant amount of information and challenges in obtaining a stable and efficient estimation of the covariance matrix and its inverse, the precision matrix. It is widely recognized that the classical sample estimators do not function properly in this context, and as a result, estimators built based on the sample estimator are also not recommended. Yao and coauthors [1] provided a comprehensive discussion of these issues.

In response to the challenges posed by large dimensional data analysis, various approaches have been proposed to mitigate the bias of the estimators, including regularization. Regularization, first introduced by Hoerl and Kennard [2], is a family of methods that modifies the maximum likelihood estimator (MLE) to provide reliable answers in unstable situations [3]. Regularization achieves this by retaining the information in the data set while simplifying the regularized model through shrinking the parameters towards a known target. In recent years, the field of regularization has seen a surge in theoretical and empirical research, with numerous works examining its applications in a wide range of contexts, for example, see [4–6] and the references therein.

Regularizing the sample covariance matrix directly can provide a satisfactory estimator of the covariance matrix and its related statistics. However, regularizing the precision matrix is not easy and it is mainly because the inverse of the sample covariance matrix will become ill-conditioned when the dimension is relatively to the sample size. To address this issue, alternative methods

☆ This article is part of a Special issue entitled: 'YJMVA_Learning in Multi-Analysis' published in Journal of Multivariate Analysis.

* Corresponding author.
E-mail address: deliang.dai@lnu.se (D. Dai).

https://doi.org/10.1016/j.jmva.2025.105553
Received 31 December 2023; Received in revised form 15 May 2025; Accepted 27 November 2025

Available online 28 November 2025

have been proposed along with the development of estimating covariance matrix and precision matrix. Pourahmadi [7] developed a new parameterization of $\boldsymbol{\Sigma}$, namely, modified Cholesky decomposition (MCD). The key idea is that the covariance matrix $\boldsymbol{\Sigma} = (\phi_{ij})$ of a zero-mean random vector $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ can be decomposed by a lower triangular matrix constructed from a regression coefficients when y_t is regressed on its predecessors $y_1, \ldots, y_{t-1}$, $t \in \{2, \ldots, p\}$ [8]. More precisely, for MCD we have

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_k + \varepsilon_t, \quad \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}^\top = \mathbf{D}, \tag{1}$$

where $\mathbf{T}$ is a lower triangular matrix having ones on its diagonal and $-\phi_{tj}$ at its (t, j)th element for $j < t$, and $\mathbf{D}$ is a diagonal matrix with $\sigma_t^2 = \mathrm{var}(\varepsilon_t)$ as its diagonal entries. With the close connection between the idea of regression and the Cholesky decomposition of a covariance matrix, MCD provides unique, completely unconstrained and interpretable reparameterizations of a covariance matrix. The precision matrix $\boldsymbol{\Sigma}^{-1}$ has a similar decomposition as in (1): $\boldsymbol{\Sigma}^{-1} = \mathbf{T}^\top\mathbf{D}^{-1}\mathbf{T}$ and it means that for any estimate $(\hat{\mathbf{T}}, \hat{\mathbf{D}})$, the estimated covariance matrix $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{T}}^{-1}\hat{\mathbf{D}}(\hat{\mathbf{T}}^\top)^{-1}$ is guaranteed to be positive definite [9]. Noting, MCD relies on prior information on the variable ordering and natural ordering is typical for longitudinal data. Recently there have been some studies which do not require the pre-determined order of variables [10,11]. However, how to resolve the drawback of order dependency in the MCD approach when the natural order is not available is not the central concern of this paper.

This paper focuses on regularizing covariance matrix via MCD. Huang et al. [12] showed that shrinkage is very useful in providing a more stable estimate of a covariance matrix, especially when the dimension p is high. Bickel and Levina [13] showed that banding the Cholesky factor can produce a consistent estimator in the operator norm under weak conditions on the covariance matrix. Tai [14] proposed a penalized likelihood method for producing a statistically efficient estimator of a covariance matrix by introducing both shrinkage and smoothing on the Cholesky factor. In Tai's work, it was claimed that one drawback of the shrinkage approach in [12] is that it ignores possible continuity among neighbouring entries of the modified Cholesky factor which is a feature typical for covariance matrices of longitudinal data. The proposed regularization scheme is motivated by the fact that the Cholesky factor $\mathbf{T}$ often shows some form of continuity in longitudinal data. For example, for a first-order autoregressive covariance matrix, i.e., AR(1), there could be a considerable amount of zeros in the subdiagonals of the $\mathbf{T}$ matrix. In particular, the first subdiagonal of $\mathbf{T}$ has the same entry in all places, and the rest of the subdiagonals are either zero or contain minor values that are close to zero. For a compound symmetry (CS) covariance matrix, the elements of each row of $\mathbf{T}$ are the same, and one can expect that there are many small value elements in the last few rows of the lower triangular part of $\mathbf{T}$. Taking into account the smoothing can help provide more efficient covariance matrix estimates.

Dai et al. [15] proposed a new type of Mahalanobis distance based on Tai's regularized estimator of the precision matrix. It has been demonstrated that the approach outperforms existing methods in terms of its performance. Despite the satisfactory practical performance which has been noticed in literature [14,15], the theoretical property of this proposed estimator was not investigated. Since both smoothing and shrinkage penalties are included the asymptotically theoretical property of the derived estimators is complicated but deserves further research. Recently Dai et al. [16] developed a regularized flip-flop algorithm (RFF) for estimating large Kronecker structured covariance matrix and established theoretical results. Their work considered a specific covariance structure when data are matrix-valued, but may not for non-Kronecker structured covariance matrix. Motivated by this, in this paper we derive the consistency property of regularized covariance estimator when shrinkage and smoothing penalties are adopted. Moreover, we have implemented the corresponding algorithm of the examined method in an efficient way by using "Rcpp" [17] to integrate R and C++.

Throughout the paper, we use the following notations. For a matrix $\mathbf{A}$, $\mathbf{A}^\top$ denotes the transpose, $\mathrm{tr}(\mathbf{A})$ is the trace, $\mathbf{A}^{-1}$ is the inverse matrix, $\det(\mathbf{A})$ is the determinant, $\mathrm{diag}(\mathbf{A})$ is the diagonalizing operator and $\mathrm{vec}(\mathbf{A})$ is the vectorizing operator of $\mathbf{A}$, respectively. Moreover, $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^\top\mathbf{A})$, operators $O(\bullet)$ and is the infinitely large quantity, and $O_p(\bullet)$ denote that relationships hold with probability tending to 1.

The rest of the paper is organized as follows. A brief introduction to the MCD method and the corresponding penalized likelihood functions with shrinkage and smoothing penalties as well as the choice of tuning parameters are introduced in Section 2. In Section 3, we present the main results and the consistency properties of the derived estimators. The simulation studies are reported in Section 4. The paper is concluded with some reflections and future studies in Section 5.

2. Regularized estimation of the covariance matrix

In this section, we introduce the designs of the shrinkage and smoothing penalties on the Cholesky factor $\mathbf{T}$. Eq. (1) shows that $\hat{y}_t$, the linear least squares predictor of y_t, can be estimated based on its predecessors $y_{t-1}, \ldots, y_1$ with prediction error $\varepsilon_t = y_t - \hat{y}_t$ and variance $\sigma_t^2 = \mathrm{var}\left(\varepsilon_t\right)$. The statistical advantage of modelling covariance matrix with MCD is first and foremost, it transforms inverse covariance matrix estimation into a regression problem, and hence regularization approaches for regression can be applied [18]. Like the mean, the precision matrix can be estimated by using the coefficients of the regression from the linear regression scheme. Moreover, the regression formulation in (1) shows the possibility of performing regularized estimation that combines both shrinkage and smoothing to matrix $\mathbf{T}$.

Let λ_{sk} and λ_{sm} be the penalty parameters that correspond to shrinkage and smoothing penalties respectively. The shrinkage penalty on the likelihood function of the component of $\boldsymbol{\Sigma}$ is given as follows:

$$\lambda_{sk} \sum_{t=2}^{p} \sum_{j=1}^{t-1} |\phi_{tj}|, \tag{2}$$

which is also known as L_1 penalty. The L_1 penalty shrinks the estimates of ϕ_{tj}'s toward zero and helps to produce more stable estimates. When $\lambda_{sk} = 0$, the minimization of the penalized log-likelihood degenerates back to the classical MLE. Without loss of generality, we assume $\mathbf{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n) \sim N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n)$, i.e., matrix $\mathbf{Y}$ is normally distributed and consists of n independent random vectors with length p and the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{y}_i, i \in \{1, \ldots, n\}$. Huang et al. [12] considered a statistically efficient estimator based on the penalized normal likelihood with an L_1 penalty:

$$-2\ell(\boldsymbol{\Sigma}, \mathbf{Y}) + \lambda_{sk} \sum_{t=2}^{p} \sum_{j=1}^{t-1} |\phi_{tj}|, \tag{3}$$

where $-2\ell(\boldsymbol{\Sigma}, \mathbf{Y}) = \sum_{t=1}^{p} \left(n \log \sigma_t^2 + \sum_{i=1}^{n} \epsilon_{it}^2 / \sigma_t^2 \right)$, $\varepsilon_{i1} = y_{i1}$ and $\varepsilon_{it} = y_{it} - \sum_{j=1}^{t-1} y_{ij}\phi_{tj}$, $t \in \{2, \ldots, p\}$.

Smoothing is a method that links one observation to its previous observation in a simple functional form. The advantages of using smoothing are mainly twofold. Firstly, smoothing is simple to implement. It adds a minor computational burden to the estimating procedure, especially when the dataset has a relatively high amount of observations and variables. The results only depend on a limited amount of previous observations which reduces the complication of estimation. Secondly, the smoothing penalty is a data-driven technique that can distinguish meaningful signals from noises by adjusting the tuning parameter. The smoothing penalty is given as follows,

$$\lambda_{sm} \sum_{t=2}^{p-2} \sum_{j=1}^{t-1} (\Delta_{diag}^2 \phi_{t+2,t+2-j})^2, \tag{4}$$

where $\Delta_{diag}^2 \phi_{t+2,t+2-j} = (\phi_{t+2,t+2-j} - \phi_{t+1,t+1-j}) - (\phi_{t+1,t+1-j} - \phi_{t,t-j})$ is the second-order difference in the lower triangle matrix $\mathbf{T}$. Here we implement the differencing along the sub-diagonal of the Cholesky factor and use smoothing spline as the smoothing penalty for the likelihood function. The smoothing function in (4) is a bidirectional operation which can be applied in the diagonal direction both forward and backwards. Combining both shrinkage and smoothing in (2) and (4), the penalized log-likelihood function is defined as follows,

$$-2\ell(\boldsymbol{\Sigma}, \mathbf{Y}) + \lambda_{sk} \sum_{t=2}^{p} \sum_{j=1}^{t-1} |\phi_{tj}| + \lambda_{sm} \sum_{t=2}^{p-2} \sum_{j=1}^{t-1} (\Delta_{diag}^2 \phi_{t+2,t+2-j})^2, \tag{5}$$

where $-2\ell(\boldsymbol{\Sigma}, \mathbf{Y})$ is presented in (3). The estimated $\mathbf{T}$ and $\mathbf{D}$ that minimize the penalized log-likelihood can be obtained numerically, i.e. the minimization of (5) with regards to estimators $\hat{\sigma}_t^2$ and $\hat{\boldsymbol{\phi}}_{tj}$ are derived from the pseudo-algorithm of the MCD in Appendix A.

Note that there are two tuning parameters λ_{sk} and λ_{sm} in the objective function (5). To the choice of the tuning parameters, like any other penalty-based method, cross-validation and information criteria such as BIC (Bayesian information criterion) [19] are often suggested. In this work we employ five-fold cross-validation (CV) [20]. The procedure starts by randomly splitting the dataset into 5 parts of the same size. Each time we use 4 parts as training data to estimate the model parameters and the resting part as testing data for validation. This process will be repeated 5 times in order to find out the optimal values of λ_{sk} and λ_{sm}. Let $\hat{\boldsymbol{\Sigma}}_{-k}$ and $\hat{\boldsymbol{\Sigma}}_{-k}^{-1}$ be the estimators of the covariance matrix and its inverse which are based on the training data in the kth replication, $1 \le k \le 5$. The optimal values of tuning parameters are chosen by

$$(\hat{\lambda}_{sk}, \hat{\lambda}_{sm}) = \underset{\lambda_{sk}, \lambda_{sm}}{\arg\min} \frac{1}{5} \sum_{k=1}^{5} \left(n_k \log |\hat{\boldsymbol{\Sigma}}_{-k}| + \sum_{i \in testing} \boldsymbol{y}_i{}^{\top} \hat{\boldsymbol{\Sigma}}_{-k}^{-1} \boldsymbol{y}_i \right),$$

where n_k is the sample size of the testing data. The first term in the summation penalizes conditional covariance estimates (through the log-determinant), while the second term measures the quadratic form of the testing observations under the fitted model, reflecting the goodness-of-fit [12]. The minimization is done through the grid-searching for each combination of the values of λ_{sk} and λ_{sm}.

3. Theoretical results

We show the main results of the consistency property of the estimator in this section. We make the following assumptions about the true model.

A1: The true Cholesky factor $\mathbf{T}_0$ is a lower triangular matrix and $\mathbf{D}_0$ is a diagonal matrix such that the covariance matrix $\boldsymbol{\Sigma}_0 = \mathbf{T}_0^{-1}\mathbf{D}_0(\mathbf{T}_0^{\top})^{-1}$.

A2: Let s be the number of nonzero elements in the lower left part of $\mathbf{T}_0$ and $(s+p)\log(p)/n = o(1)$ as $n \to \infty$.

A3: There exists a constant $h > 1$ such that the eigenvalues of the true covariance matrix are bounded, i.e., $1/h < \psi_{min}(\boldsymbol{\Sigma}_0) \le \psi_{max}(\boldsymbol{\Sigma}_0) < h$, where $\psi_{max}(\boldsymbol{\Sigma}_0)$ and $\psi_{min}(\boldsymbol{\Sigma}_0)$ are the largest and smallest eigenvalues of $\boldsymbol{\Sigma}_0$, respectively.

The theoretical consistency results for the penalized modified Cholesky decomposition rely on three key assumptions regarding the model structure, sparsity, and regularity of the underlying covariance matrix. Assumption A1 states that the true covariance matrix $\boldsymbol{\Sigma}_0$ admits a modified Cholesky decomposition (MCD). This structural assumption guarantees the unique and stable decomposition $\boldsymbol{\Sigma}_0 = \mathbf{T}_0^{-1}\mathbf{D}_0(\mathbf{T}_0^{\top})^{-1}$, which facilitates separate estimation of the conditional dependence structure through $\mathbf{T}_0$ and the marginal variances through $\mathbf{D}_0$. Assumption A2 imposes both a sparsity condition and a dimensionality constraint. The assumption requires that the effective model complexity, quantified by $(s+p)\log(p)/n$, tends to a constant as the sample size n increases. This condition allows both the number of variables p and the sparsity s to grow with n at a proper speed. Assumption A3 specifies

eigenvalue conditions for the true covariance matrix $\mathbf{\Sigma}_0$, that all eigenvalues remain bounded away from both zero and infinity. Bounded eigenvalues imply that the variances and correlations among the variables are of reasonable scale, which is essential for the theoretical guarantees of penalized likelihood methods, and controlling the estimation error in both $\mathbf{\Sigma}_0$ and its inverse $\mathbf{\Sigma}_0^{-1}$.

Now we restate the following theorem, which was given as Theorem 9 in [21]. It presents the rates of convergence of the estimators $\hat{\mathbf{T}}_{sk}$, $\hat{\mathbf{D}}_{sk}$ and $\hat{\mathbf{\Sigma}}_{sk}$ under Frobenius norm.

Remark 1. Under assumptions A1–A3, if $\lambda_{sk} = O_p(\log(p)/n)$, there exists a local minimizer $\hat{\mathbf{T}}_{sk}$ and $\hat{\mathbf{D}}_{sk}$ in (3) such that

$$\|\hat{\mathbf{T}}_{sk}-\mathbf{T}_0\|_F=O_p(\sqrt{s\log(p)/n}),\quad \|\hat{\mathbf{D}}_{sk}-\mathbf{D}_0\|_F=O_p(\sqrt{p\log(p)/n}),\quad \|\hat{\mathbf{\Sigma}}_{sk}-\mathbf{\Sigma}_0\|_F=O_p(\sqrt{(s+p)\log(p)/n}).$$

Remark 1 provides the convergence rates for the estimators $\hat{\mathbf{T}}_{sk}$ and $\hat{\mathbf{D}}_{sk}$ under regularization. Under assumptions A1–A3 and with the tuning parameter λ_{sk} chosen of order $O_p(\log(p)/n)$, there exists a local minimizer of the penalized objective function such that the estimated triangular matrix $\hat{\mathbf{T}}_{sk}$ converges to the true matrix $\mathbf{T}_0$ at the rate $O_p(\sqrt{s\log(p)/n})$, and the diagonal matrix estimate $\hat{\mathbf{D}}_{sk}$ converges to $\mathbf{D}_0$ at the rate $O_p(\sqrt{p\log(p)/n})$ in Frobenius norm. Consequently, the overall covariance matrix estimate $\hat{\mathbf{\Sigma}}_{sk}$ converges to the true covariance $\mathbf{\Sigma}_0$ at the combined rate $O_p(\sqrt{(s+p)\log(p)/n})$. These rates reflect the interplay between sparsity (s), dimensionality (p), and sample size (n), demonstrating that consistent estimation is possible while sparsity and regularization are properly controlled.

Remark 2. Let $\mathbf{\Sigma}_0 = \mathbf{T}_0^{-1}\mathbf{D}_0(\mathbf{T}_0^{\top})^{-1}$ be the MCD of the true covariance matrix. If the eigenvalues of $\mathbf{\Sigma}_0$ are bounded, so that there exist constants θ_1 and θ_2 such that $0 < \theta_1 < \psi_{min}(\mathbf{\Sigma}_0) \le \psi_{max}(\mathbf{\Sigma}_0) < \theta_2 < \infty$, then there exist constants h_1 and h_2 such that $h_1 < \psi_{min}(\mathbf{T}_0) \le \psi_{max}(\mathbf{T}_0) < h_2$ and $h_1 < \psi_{min}(\mathbf{D}_0) \le \psi_{max}(\mathbf{D}_0) < h_2$.

Remark 2 was given in [22]. It shows that the eigenvalues of the corresponding matrices $\mathbf{T}_0$ and $\mathbf{D}_0$ are all bounded once the eigenvalues of the covariance matrix $\mathbf{\Sigma}$ are bounded. The result establishes the boundedness of eigenvalues for the underlying Cholesky components of the covariance matrix. Specifically, when the true covariance matrix $\mathbf{\Sigma}_0$ has eigenvalues bounded between positive constants θ_1 and θ_2, then the corresponding Cholesky factors $\mathbf{T}_0$ and $\mathbf{D}_0$ similarly exhibit eigenvalues bounded in constants h_1 and h_2. This result, derived from [22], guarantees numerical stability and ensures the modified Cholesky decomposition remains well-conditioned.

Based on the results above, we show the consistency of the derived estimators in the following theorem.

Theorem 1. *Assume $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^{\top} \sim N(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$, under the conditions A1–A3, if $\lambda_{sk} = O_p(\log(p)/n)$ and $\lambda_{sm} = O_p(\log(p)/n)$, there exists a local minimizer $\hat{\mathbf{T}}_{skm}$ and $\hat{\mathbf{D}}_{skm}$ of the penalized likelihood function in (5) such that*

$$\|\hat{\mathbf{T}}_{skm} - \mathbf{T}_0\|_F = O_p(\sqrt{s\log(p)/n}), \qquad \|\hat{\mathbf{D}}_{skm} - \mathbf{D}_0\|_F = O_p(\sqrt{p\log(p)/n}).$$

Proof. The details of the proof are deferred to Appendix B. □

Theorem 1 establishes the convergence properties of the Cholesky factor estimates obtained from the joint penalized likelihood framework defined in (5). Under assumptions A1–A3, and provided that the tuning parameters satisfy $\lambda_{sk} = O_p(\log(p)/n)$ and $\lambda_{sm} = O_p(\log(p)/n)$, the theorem guarantees the existence of a local minimizer $(\hat{\mathbf{T}}_{skm}, \hat{\mathbf{D}}_{skm})$ achieving the estimation error in Frobenius norm converges at rate $O_p(\sqrt{s\log(p)/n})$ for the lower triangular matrix $\mathbf{T}_0$, and at rate $O_p(\sqrt{p\log(p)/n})$ for the diagonal matrix $\mathbf{D}_0$. These rates reflect how sparsity s and the dimension p jointly influence estimation accuracy. The result shows that consistent estimation of both Cholesky factors is feasible as long as the sparsity level is controlled and the regularization parameters are properly scaled with $\log(p)/n$.

The convergence rates of the matrices $\hat{\mathbf{T}}_{skm}$ and $\hat{\mathbf{D}}_{skm}$ are presented separately in Theorem 1. In what follow-up, we extend these results to derive the convergence rates for the covariance matrix and the corresponding precision matrix. Specifically, we provide the following convergence rates for these matrices.

Theorem 2. *Following the same assumptions as in Theorem 1, there exists a local minimizer $\hat{\mathbf{\Sigma}}_{skm}$ and $\hat{\mathbf{\Sigma}}_{skm}^{-1}$ of the penalized likelihood function in (5) such that*

$$\|\hat{\mathbf{\Sigma}}_{skm} - \mathbf{\Sigma}_0\|_F = O_p(\sqrt{(s+p)\log(p)/n}), \qquad \|\hat{\mathbf{\Sigma}}_{skm}^{-1} - \mathbf{\Sigma}_0^{-1}\|_F = O_p(\sqrt{(s+p)\log(p)/n}).$$

Proof. The details of the proof are deferred to Appendix B. □

Building on these results, Theorem 2 establishes convergence guarantees for both the full covariance matrix $\mathbf{\Sigma}_0$ and its inverse (the precision matrix). Specifically, the estimated matrices $\hat{\mathbf{\Sigma}}_{skm}$ and $\hat{\mathbf{\Sigma}}_{skm}^{-1}$ achieve an $O_p(\sqrt{(s+p)\log(p)/n})$ convergence rate in the Frobenius norm. The dependence on both s and p reflects the contributions of estimating the structured sparsity in $\mathbf{T}_0$ as well as the scale in $\mathbf{D}_0$. Note the convergence rates are the same for the covariance matrix and its inverse. It implies that the consistency properties are proved for shrinkage and smoothing on both the covariance matrix and the precision matrix. Moreover, the convergence rates are only related to the values of λ_{sk} and λ_{sm}. More details can be found in the proof that is given in Appendix B.

Table 1
Risk functions based on the entropy (EL) and quadratic loss (QL) for the estimated AR(1) covariance matrix $\rho \in \{0.3, 0.5, 0.7\}$ combining shrinkage and smoothing penalties. The results are the averages and standard errors (in parenthesis) of the loss functions over 1000 simulation runs.

ρ=0.3	QL	EL	ρ=0.5	QL	EL	ρ=0.7	QL	EL
5	1.48 (1.99)	1.32 (3.07)	5	1.94 (2.56)	1.64 (3.61)	5	6.86 (8.68)	9.49 (34.46)
10	4.44 (3.41)	3.35 (5.65)	10	4.41 (3.40)	3.29 (5.57)	10	4.36 (3.39)	3.15 (5.44)
20	3.26 (1.71)	1.96 (1.73)	20	3.44 (1.92)	1.97 (1.91)	20	4.24 (2.70)	2.18 (2.55)
60	3.05 (1.00)	1.89 (1.10)	60	3.05 (1.00)	1.88 (1.10)	60	3.06 (1.00)	1.89 (1.10)
80	3.86 (1.03)	2.13 (0.99)	80	2.37 (0.69)	2.50 (0.80)	80	4.49 (1.41)	2.11 (1.32)
100	3.56 (0.89)	2.12 (1.10)	100	3.57 (0.88)	2.13 (1.10)	100	3.71 (1.00)	2.15 (1.08)

Table 2
Risk functions based on the entropy (EL) and quadratic loss (QL) for the estimated AR(1) covariance matrix $\rho \in \{0.3, 0.5, 0.7\}$ with only shrinkage penalty. The results are the averages and standard errors (in parenthesis) of the loss functions over 1000 simulation runs.

ρ=0.3	QL	EL	ρ=0.5	QL	EL	ρ=0.7	QL	EL
5	7.08 (9.94)	9.08 (36.76)	5	7.29 (11.05)	9.86 (41.13)	5	7.68 (12.94)	11.20 (48.84)
10	4.50 (3.64)	3.40 (6.77)	10	4.67 (3.97)	3.36 (6.58)	10	5.01 (4.67)	3.33 (6.58)
20	3.64 (1.86)	2.06 (2.01)	20	3.82 (2.09)	2.06 (2.15)	20	4.24 (2.71)	2.18 (2.55)
60	3.06 (1.05)	1.99 (1.31)	60	3.25 (1.16)	2.00 (1.41)	60	4.02 (1.52)	2.17 (1.57)
80	3.92 (1.05)	2.11 (0.99)	80	4.12 (1.16)	2.12 (1.10)	80	4.51 (1.45)	2.19 (1.41)
100	3.64 (0.90)	2.08 (0.93)	100	3.85 (1.00)	2.09 (1.02)	100	4.28 (1.22)	1.91 (1.12)

4. Numerical illustrations

In this section, we compare the performance of the MCD estimations with simulations. The dataset are generated from multivariate normal distribution with $\boldsymbol{y}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, \ldots, n$. Here the covariance matrix $\boldsymbol{\Sigma}$ (it is also the correlation matrix in this case) has following two types of structure:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix},$$

i.e., $\boldsymbol{\Sigma}$ that follows an auto-regressive (AR) structure or a compound symmetric (CS) structure with different values $\rho \in \{0.3, 0.5, 0.7\}$. We estimate the covariance matrix under penalty of either shrinkage or both shrinkage and smoothing. The ℓ_1 norms are employed here for the regularization. Since the difference between ℓ_1 and ℓ_2 norms is negligible, we skip the results for the ℓ_2 norms and present only the ℓ_1 in the paper. The dimensions for $p \in \{5, 10, 20, 60, 80, 100\}$ while n is fixed to 10 since we are more interested in the performance of the estimator when the ratio p/n changes below and above 1 rather than the values of the dimensions. The ratio p/n is considered in different scenarios such that $p/n < 1$, $p/n > 1$ and $p/n = 1$.

The entropy loss (EL) and quadratic loss (QL) functions are used here to evaluate the estimator. The loss functions are defined as follows:

$$f_E\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = p^{-1}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right) - \ln\left[\det\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right)\right] - p\right), \qquad f_Q\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = p^{-1}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\right)\right)^2. \tag{6}$$

The results are reported by their averages and corresponding standard errors (in parenthesis) of the loss functions over 1000 repetitions. The candidate values of the tuning parameters $\lambda_{sk}, \lambda_{sm}$ are chosen from the interval of $(0.1, 10^2)$ with an increment of 1 each time. The results are given in Tables 1–4.

Notably, the combined use of shrinkage and smoothing penalties generally outperforms shrinkage alone across most scenarios. This improvement is particularly evident when $n > p$ under both quadratic and entropy loss functions in the AR cases. Furthermore, the proposed method appears robust, performing well in most of the settings. It should be noted that as the value of ρ increases, the estimator's performance deteriorates for both AR and CS structures. One potential explanation is that a stronger correlation among the variables diminish the effectiveness of shrinkage and smoothing. A stronger penalty may be required to reduce the bias of the estimator.

The most challenging scenario occurs when $n = p$. In these cases, the loss functions yield relatively higher values, particularly for larger correlation levels ρ under both quadratic and entropy loss criteria. However, the bias decreases as p diverges from n, meaning that the performance improves substantially as p either decreases or increases.

The standard errors are stable around 1 for the AR(1) structure while it varies between 2 and 3 for the CS structure. The CS structure leads to higher variability in standard errors compared to the AR(1) structure, reflecting the inherent estimation challenges posed by its uniform correlation pattern.

Table 3
Risk functions based on the entropy (EL) and quadratic loss (QL) for the estimated CS covariance matrix $\rho \in \{0.3, 0.5, 0.7\}$ combining shrinkage and smoothing penalties. The results are the averages and standard errors (in parenthesis) of the loss functions over 1000 simulation runs.

ρ=0.3	QL	EL	ρ=0.5	QL	EL	ρ=0.7	QL	EL
5	7.19 (11.28)	10.29 (44.54)	5	7.48 (13.02)	11.55 (51.78)	5	7.92 (14.93)	12.85 (59.09)
10	3.91 (3.14)	2.40 (3.06)	10	4.18 (3.98)	2.41 (3.85)	10	4.61 (4.97)	2.42 (4.80)
20	3.41 (2.29)	1.95 (2.22)	20	3.70 (3.18)	1.95 (3.00)	20	4.17 (4.16)	1.95 (3.89)
60	3.48 (2.04)	2.16 (2.24)	60	3.80 (3.06)	2.16 (3.26)	60	4.30 (4.15)	2.15 (4.37)
80	4.06 (2.16)	2.11 (2.00)	80	4.39 (3.35)	2.11 (3.10)	80	4.89 (4.60)	2.10 (4.25)
100	4.12 (2.13)	2.07 (1.99)	100	4.45 (3.35)	2.07 (3.12)	100	4.92 (4.61)	2.07 (4.28)

Table 4
Risk functions based on the entropy (EL) and quadratic loss (QL) for the estimated CS covariance matrix $\rho \in \{0.3, 0.5, 0.7\}$ with only shrinkage penalty. The results are the averages and standard errors (in parenthesis) of the loss functions over 1000 simulation runs.

ρ=0.3	QL	EL	ρ=0.5	QL	EL	ρ=0.7	QL	EL
5	7.38 (9.92)	8.86 (32.10)	5	7.65 (11.32)	9.64 (36.74)	5	8.07 (12.95)	10.47 (41.52)
10	4.58 (4.04)	3.42 (6.97)	10	4.82 (4.79)	3.32 (6.53)	10	5.23 (5.77)	3.24 (6.74)
20	3.78 (2.49)	2.05 (2.45)	20	4.08 (3.43)	2.05 (3.22)	20	4.54 (4.49)	2.04 (4.12)
60	3.27 (1.91)	2.00 (2.11)	60	3.59 (2.89)	1.99 (3.10)	60	4.09 (3.93)	1.98 (4.17)
80	4.14 (2.18)	2.11 (2.04)	80	4.46 (3.39)	2.11 (3.17)	80	4.96 (4.65)	2.12 (4.35)
100	4.69 (4.41)	2.08 (4.32)	100	4.19 (3.20)	2.08 (3.14)	100	4.95 (4.63)	2.09 (4.34)

5. Conclusion and discussion

This study advances the theoretical understanding of MCD-based regularized covariance estimation by proving the consistency of estimators derived from combined shrinkage and smoothing penalties. Our results show that the convergence rates depend explicitly on the sparsity of the true covariance matrix, which is analogous to results for Kronecker structured covariance [16], where the rates of convergence also depend on how sparse the true covariance components are.

There are some directions for future research. First, MCD requires an ordering of the variables, hence the data without natural ordering cannot be applied. A potential way to address this issue is to consider permutation invariant covariance matrix [23]. Second, under a multilevel multivariate model, for example a so-called doubly multivariate model [24] as a special case, the covariance matrices will exhibit certain block structures. In this situation a block-type MCD deserves to be investigated.

CRediT authorship contribution statement

Yuli Liang: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Software. **Deliang Dai:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Software. **Shaobo Jin:** Conceptualization, Investigation, Validation.

Acknowledgments

The authors are thankful to the Editor, an Associate Editor and the anonymous reviewer for their insightful comments which has resulted in significant improvement of this manuscript. Yuli Liang and Deliang Dai thank Linnaeus University, Sweden for giving the opportunity and support for this paper. Liang's work is also supported by Center for Applied Mathematics of Guangxi (Guangxi Normal University), and Key Laboratory of Interdisciplinary Research for Data Science, Education Department of Guangxi Zhuang Autonomous Region.

Appendix A. Regularized algorithm combining shrinkage and smoothing

Denote $\mathbf{E}_t$ be the $(p-2-t)\times(p-t)$ matrix representations of the difference operator Δ^2_{diag},

$$\begin{pmatrix} 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}$$

such that $\sum_{k=1}^{t-1}(\Delta^2_{diag}\phi_{t+2,t+2-k})^2 = \boldsymbol{\phi}_t^\top \mathbf{E}_t^\top \mathbf{E}_t \boldsymbol{\phi}_t$. We summarized the following algorithm for the regularized estimates based on the MCD.

Algorithm 1 Shrinkage and smoothing sub-diagonals of Cholesky factor [14].

Initialize $\left\{\sigma_t^{2(1)}, t \in \{1,\dots,p\}\right\}$ and $\left\{\boldsymbol{\phi}_t^{(1)}, t \in \{1,\dots,p-1\}\right\}$
if $n > p$ **then**
  use MLE as initial values of $\sigma_t^{2(1)}, \boldsymbol{\phi}_t^{(1)}$
else if $n \le p$ **then**
  use L_2 penalized likelihood estimate given in [14] as initial values of $\sigma_t^{2(1)}, \boldsymbol{\phi}_t^{(1)}$
end if
Set $tol = 10^{-5}$

$$\mathbf{H}_t = \operatorname{diag}\left(\frac{1}{\gamma_{t+1}^2}\sum_{i=1}^{n} y_{i1}^2, \frac{1}{\gamma_{t+2}^2}\sum_{i=1}^{n} y_{i2}^2, \dots, \frac{1}{\gamma_n^2}\sum_{i=1}^{n} y_{i,p-t}^2\right)_{(p-t)\times(p-t)},$$

$$\mathbf{g}_t^{\top} = \frac{1}{\sigma_{t+1}^2}\sum_{i=1}^{n} y_{i1}\left(y_{i,t+1} - \sum_{j=2}^{t} y_{ij}\phi_{t+1,j}\right), \dots, \frac{1}{\sigma_n^2}\sum_{i=1}^{n} y_{i,n-t}\left(y_{in} - \sum_{\substack{1\le j\le t-1\\ j\ne p-t}} y_{ij}\phi_{nj}\right).$$

for $i \in \{1,\dots,10^4\}$ **do**
  (1) $\boldsymbol{\phi}_t^{(i+1)} = \left(\mathbf{H}_t^{(i)} + \lambda_{sh}\mathbf{I}_t + \lambda_{sm}\mathbf{E}_t^{\top}\mathbf{E}_t\right)^{-1}\mathbf{g}_t^{(i)}$;
  (2) $\sigma_t^{2(i+1)} = n^{-1}\sum_{i=1}^{n}\left(y_{it} - \sum_{j=1}^{t-1} y_{ij}\phi_{tj}^{(i+1)}\right)^2$.
  if $\min(|\sigma_t^{2(i+1)} - \sigma_t^{2(i)}|, |\boldsymbol{\phi}_t^{(i+1)} - \boldsymbol{\phi}_t^{(i)}|) \ge tol$ **then**
    Repeat (1) and (2)
  else
    $\boldsymbol{\phi}_t^{(i+1)} = \boldsymbol{\phi}_t^{(i)}$
    $\sigma_t^{2(i+1)} = \sigma_t^{2(i)}$
  end if
end for
  return $\boldsymbol{\phi}_t^{(i+1)}$ and $\sigma_t^{2(i+1)}$

Appendix B. Proofs of Theorems 1 and 2

Proof of Theorem 1. We prove the theorem using the idea found in [11] which was enlightened by the proof in [22]. For $m = 1$, let us reconsider the objective function in (5),

$$Q(\mathbf{T},\mathbf{D}) = \log|\mathbf{D}| + \operatorname{tr}(\mathbf{T}^{\top}\mathbf{D}^{-1}\mathbf{T}\mathbf{S}) + \lambda_{sk}\sum_{t=2}^{p}\sum_{j=1}^{t-1}|\phi_{tj}| + \lambda_{sm}\sum_{t=2}^{p-2}\sum_{j=1}^{t-1}(\Delta_{diag}^2\phi_{t+2}, \phi_{t+2-j})^2.$$

Consider the difference $G(\Delta_{\mathbf{T}}, \Delta_{\mathbf{D}}) = Q(\mathbf{T}_0 + \Delta_{\mathbf{T}}, \mathbf{D}_0 + \Delta_{\mathbf{D}}) - Q(\mathbf{T}_0, \mathbf{D}_0)$, where $\mathbf{T}_0$ and $\mathbf{D}_0$ are the true matrices of $\mathbf{T}$ and $\mathbf{D}$, we define the sets

$$\mathcal{A} = \left\{\Delta_{\mathbf{T}} : \|\Delta_{\mathbf{T}}\|_F^2 \le M^2 s\log(p)/n\right\}, \quad \mathcal{B} = \left\{\Delta_{\mathbf{D}} : \|\Delta_{\mathbf{D}}\|_F^2 \le N^2 p\log(p)/n\right\},$$

where M and N are constants. We would like to prove that for each $\Delta_{\mathbf{T}} \in \partial\mathcal{A}$ and $\Delta_{\mathbf{D}} \in \partial\mathcal{B}$,

$$\Pr\left(G\left(\Delta_{\mathbf{T}}, \Delta_{\mathbf{D}}\right) > 0\right) \to 1,$$

as $n \to \infty$ for sufficiently large M and N, where $\partial\mathcal{A}$ and $\partial\mathcal{B}$ are the boundaries of $\mathcal{A}$ and $\mathcal{B}$, respectively. This implies that the minimal point of $G(\Delta_{\mathbf{T}}, \Delta_{\mathbf{D}})$ is achieved when $\Delta_{\mathbf{T}} \in \mathcal{A}$ and $\Delta_{\mathbf{D}} \in \mathcal{B}$, that is

$$\|\Delta_{\mathbf{T}}\|_F = O_p(\sqrt{s\log(p)/n}), \quad \|\Delta_{\mathbf{D}}\|_F = O_p(\sqrt{p\log(p)/n}),$$

for sufficiently large n. To simplify the notation, write $\mathbf{T} = \mathbf{T}_0 + \Delta_{\mathbf{T}}$ and $\mathbf{D} = \mathbf{D}_0 + \Delta_{\mathbf{D}}$. First the difference $G(\Delta_{\mathbf{T}}, \Delta_{\mathbf{D}})$ can be written as

$$\begin{aligned}
G(\Delta_{\mathbf{T}}, \Delta_{\mathbf{D}}) =& Q(\mathbf{T},\mathbf{D}) - Q(\mathbf{T}_0,\mathbf{D}_0)\\
=& \log|\mathbf{D}| - \log|\mathbf{D}_0| + \operatorname{tr}\left(\mathbf{T}^{\top}\mathbf{D}^{-1}\mathbf{T}\mathbf{S}\right) - \operatorname{tr}\left(\mathbf{T}_0^{\top}\mathbf{D}_0^{-1}\mathbf{T}_0\mathbf{S}\right) + \lambda_{sk}\sum_{t=2}^{p}\sum_{j=1}^{t-1}\left(\left|\phi_{tj}\right| - \left|\phi_{0tj}\right|\right)\\
&+ \lambda_{sm}\sum_{t=2}^{p-2}\sum_{j=1}^{t-1}\left[(\Delta_{diag}^2\phi_{t+2}, \phi_{t+2-j})^2 - (\Delta_{diag}^2\phi_{0(t+2)}, \phi_{0(t+2-j)})^2\right]\\
=& \log|\mathbf{D}| - \log\left|\mathbf{D}_0\right| + \operatorname{tr}\left[\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\right)\mathbf{D}_0\right] - \operatorname{tr}\left[\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\right)\mathbf{D}_0\right]\\
&+ \operatorname{tr}\left(\mathbf{T}^{\top}\mathbf{D}^{-1}\mathbf{T}\mathbf{S}\right) - \operatorname{tr}\left(\mathbf{T}_0\mathbf{D}_0^{-1}\mathbf{T}_0\mathbf{S}\right) + \lambda_{sk}\sum_{t=2}^{p}\sum_{j=1}^{t-1}(|\phi_{tj}| - |\phi_{0tj}|)
\end{aligned}$$

$$
\begin{aligned}
&+ \lambda_{sm} \sum_{t=2}^{p-2} \sum_{j=1}^{t-1} \left[(\Delta_{diag}^2 \phi_{t+2}, \phi_{t+2-j})^2 - (\Delta_{diag}^2 \phi_{0(t+2)}, \phi_{0(t+2-j)})^2 \right] \\
&= \log |\mathbf{D}| - \log |\mathbf{D}_0| + \operatorname{tr} \left[\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1} \right) \mathbf{D}_0 \right] - \operatorname{tr} \left[\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1} \right) \mathbf{D}_0 \right] \\
&+ \operatorname{tr} \left(\mathbf{T}^\top \mathbf{D}^{-1} \mathbf{T} \mathbf{S} \right) - \operatorname{tr} \left(\mathbf{T}_0 \mathbf{D}_0^{-1} \mathbf{T}_0 \mathbf{S} \right) + \lambda_{sk} \sum_{t,j \in Z^c} |\phi_{tj}| + \lambda_{sk} \sum_{t,j \in Z} (|\phi_{tj}| - |\phi_{0tj}|) \\
&+ \lambda_{sm} \sum_{t,j} (\Delta_{diag}^2 \phi_{t+2}, \phi_{t+2-j})^2 = M_1 + M_2 + M_3 + M_4 + M_5,
\end{aligned}
$$

where $Z = \{(t,j) : t < j, \phi_{0tj} \neq 0\}$, ϕ_{0tj} represents the (t,j)-th element of the matrix $\boldsymbol{T}_0$, and

$$
\begin{aligned}
&M_1 = \log |\mathbf{D}| - \log |\mathbf{D}_0| + \operatorname{tr} \left[\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1} \right) \mathbf{D}_0 \right], \\
&M_2 = \operatorname{tr} \left(\mathbf{T}^\top \mathbf{D}^{-1} \mathbf{T} \mathbf{S} \right) - \operatorname{tr} \left(\mathbf{T}_0 \mathbf{D}_0^{-1} \mathbf{T}_0 \mathbf{S} \right) - \operatorname{tr} \left[\left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1} \right) \mathbf{D}_0 \right], \\
&M_3 = \lambda_{sk} \sum_{t,j \in Z^c} |\phi_{tj}|, \quad M_4 = \lambda_{sk} \sum_{t,j \in Z} (|\phi_{tj}| - |\phi_{0tj}|), \\
&M_5 = \lambda_{sm} \sum_{t,j} \left[(\Delta_{diag}^2 \phi_{t+2}, \phi_{t+2-j})^2 - (\Delta_{diag}^2 \phi_{0(t+2)}, \phi_{0(t+2-j)})^2 \right].
\end{aligned}
$$

Based on the proof of Theorem 3.1 in [22], it has been showed that $M_1 \geq \|\Delta_{\boldsymbol{D}}\|_F^2/(8h^4)$ and $M_2 = M_2^{(1)} + M_2^{(2)} + M_2^{(3)}$, where

$$
\begin{aligned}
&M_2^{(1)} = \operatorname{tr} \left(\mathbf{D}^{-1} - \mathbf{D}_0^{-1} \right) \left[\mathbf{T} \left(\mathbf{S} - \boldsymbol{\Sigma}_0 \right) \mathbf{T}^\top \right], \quad M_2^{(2)} = \operatorname{tr} \left[\mathbf{D}_0^{-1} \left(\mathbf{T} \left(\mathbf{S} - \boldsymbol{\Sigma}_0 \right) \mathbf{T}^\top - \mathbf{T}_0 \left(\mathbf{S} - \boldsymbol{\Sigma}_0 \right) \mathbf{T}_0^\top \right) \right], \\
&M_2^{(3)} = \operatorname{tr} \left[\mathbf{D}^{-1} \left(\mathbf{T} \boldsymbol{\Sigma}_0 \mathbf{T}^\top - \mathbf{T}_0 \boldsymbol{\Sigma}_0 \mathbf{T}_0^\top \right) \right].
\end{aligned}
$$

It is also showed that, there exists $V_1 > 0$ and $V_2 > 0$ such that $|M_2^{(1)}| \leq V_1 \sqrt{p \log(p)/n} \|\Delta_{\mathbf{D}}\|_F$ and

$$
M_2^{(3)} - |M_2^{(2)}| > \frac{1}{2h^4} \|\Delta_{\mathbf{T}}\|_F^2 - V_2 \sqrt{\log(p)/n} \sum_{(t,j) \in Z^c} |\phi_{tj}| - V_2 \sqrt{s \log(p)/n} \|\Delta_{\mathbf{T}}\|_F.
$$

For the shrinkage term, we have

$$
|M_4| = |\lambda_{sk} \sum_{t,j \in Z} (|\phi_{tj}| - |\phi_{0tj}|)| \leq \lambda_{sk} \sum_{t,j \in Z} ||\phi_{tj}| - |\phi_{0tj}|| \leq \lambda_{sk} \sum_{t,j \in Z} |\phi_{tj} - \phi_{0tj}| \leq \lambda_{sk} \sqrt{s} \|\Delta_{\mathbf{T}}\|_F.
$$

For the smoothing penalty term, set $A_{tj} = \phi_{t+2,t+2-j} - \phi_{t+1,t+1-j}$, $B_{tj} = \phi_{t+1,t+1-j} - \phi_{t,t-j}$, $A_{tj,0} = \phi_{t+2,t+2-j,0} - \phi_{t+1,t+1-j,0}$ and $B_{tj,0} = \phi_{t+1,t+1-j,0} - \phi_{t,t-j,0}$, and they have the same sign, i.e., all are positive or negative,

$$
\begin{aligned}
|M_5| &= \lambda_{sm} \sum_{t,j} | \left(A_{tj} - B_{tj} \right)^2 - \left(A_{tj,0} - B_{tj,0} \right)^2 | \\
&= \lambda_{sm} \sum_{t,j} | \left(A_{tj} - B_{tj} - A_{tj,0} + B_{tj,0} \right) \left(A_{tj} - B_{tj} + A_{tj,0} - B_{tj,0} \right) | \\
&\leq \lambda_{sm} \sqrt{\sum_{t,j} |A_{tj} - B_{tj} - A_{tj,0} + B_{tj,0}|^2} \sqrt{\sum_{t,j} |A_{tj} - B_{tj} + A_{tj,0} - B_{tj,0}|^2} \\
&\leq \frac{\lambda_{sm}}{2} \sum_{t,j} \left[|A_{tj} - A_{tj,0} - B_{tj} + B_{tj,0}|^2 + |A_{tj} - A_{tj,0} - B_{tj} + B_{tj,0} + 2 \left(A_{tj,0} - B_{tj,0} \right)|^2 \right] \\
&\leq \frac{\lambda_{sm}}{2} \sum_{t,j} \left[|A_{tj} - A_{tj,0} - B_{tj} + B_{tj,0}|^2 + 2|A_{tj} - A_{tj,0} - B_{tj} + B_{tj,0}|^2 + 8|A_{tj,0} - B_{tj,0}|^2 \right] \\
&= \frac{\lambda_{sm}}{2} \sum_{t,j} \left[3|A_{tj} - A_{tj,0} - \left(B_{tj} - B_{tj,0} \right)|^2 + 8|A_{tj,0} - B_{tj,0}|^2 \right].
\end{aligned} \tag{7}
$$

If $A_{tj} - A_{tj,0}$ and $B_{tj} - B_{tj,0}$ have the same sign, i.e., both are positive or negative,

$$
(7) \leq \frac{\lambda_{sm}}{2} \sum_{t,j} \left[3|A_{tj} - A_{tj,0}|^2 + 6|A_{tj} - A_{tj,0}||B_{tj} - B_{tj,0}| + 3|B_{tj} - B_{tj,0}|^2 + 8|A_{tj,0} - B_{tj,0}|^2 \right].
$$

If $A_{tj} - A_{tj,0}$ and $B_{tj} - B_{tj,0}$ have different sign, we get

$$
(7) \leq \frac{\lambda_{sm}}{2} \sum_{t,j} \left[3|A_{tj} - A_{tj,0}|^2 + 3|B_{tj} - B_{tj,0}|^2 + 8|A_{tj,0} - B_{tj,0}|^2 \right].
$$

Note,

$$
\begin{aligned}
\sum_{t,j} |A_{tj} - A_{tj,0}|^2 \leq & \sum_{t,j} |\phi_{t+2,t+2-j} - \phi_{t+2,t+2-j,0}|^2 + \sum_{t,j} |\phi_{t+1,t+1-j} - \phi_{t+1,t+1-j,0}|^2 \\
&+ 2 \sum_{t,j} |\phi_{t+2,t+2-j} - \phi_{t+2,t+2-j,0}||\phi_{t+1,t+1-j} - \phi_{t+1,t+1-j,0}| \\
\leq & 2\|\Delta_{\mathbf{T}}\|_F^2 + 2 \sum_{t,j} |\phi_{t+2,t+2-j} - \phi_{t+2,t+2-j,0}||\phi_{t+1,t+1-j} - \phi_{t+1,t+1-j,0}|
\end{aligned}
$$

$$\leq 2\|\Delta_{\mathbf{T}}\|_F^2 + 2\sqrt{\sum_{t,j}|\phi_{t+2,t+2-j} - \phi_{t+2,t+2-j,0}|^2}\sqrt{\sum_{t,j}|\phi_{t+1,t+1-j} - \phi_{t+1,t+1-j,0}|^2}$$
$$\leq 4\|\Delta_{\mathbf{T}}\|_F^2.$$

Likewise, $\sum_{t,j}|B_{tj} - B_{tj,0}|^2 \leq 4\|\Delta_{\mathbf{T}}\|_F^2$. By Cauchy–Schwarz inequality,

$$\sum_{t,j}|A_{tj} - A_{tj,0}||B_{tj} - B_{tj,0}| \leq \sqrt{\sum_{t,j}|A_{tj} - A_{tj,0}|^2}\sqrt{\sum_{t,j}|B_{tj} - B_{tj,0}|^2} \leq 4\|\Delta_{\mathbf{T}}\|_F^2.$$

Then, we further obtain $|M_5| \leq \lambda_{sm}\left[12\|\Delta_T\|_F^2 + 4\sum_{t,j}|A_{tj,0} - B_{tj,0}|^2\right]$.

Combining all the terms above together, for any ϵ, with a probability greater than $1 - 2\epsilon$, we have

$$\begin{aligned}
&|G(\Delta_T, \Delta_D)| \geq M_1 - |M_2^{(1)}| + M_2^{(3)} - |M_2^{(2)}| + M_3 - |M_4| - |M_5| \\
\geq &\frac{|\Delta_D|_F^2}{8h^4} - V_1\sqrt{p\log(p)/n}\,|\Delta_D|_F + \frac{|\Delta_T|_F^2}{2h^4} - V_2\sqrt{\log(p)/n}\sum_{(t,j)\in Z^c}|\phi_{tj}| \\
&- V_2\sqrt{s\log(p)/n}\|\Delta_{\mathbf{T}}\|_F + \lambda_{sk}\sum_{t,j\in Z^c}|\phi_{tj}| - \lambda_{sk}\sqrt{s}\,\|\Delta_{\mathbf{T}}\|_F - \lambda_{sm}\left(24\|\Delta_{\mathbf{T}}\|_F^2 + 4\sum_{t,j}|A_{tj,0} - B_{tj,0}|^2\right) \\
= &\frac{N^2}{8h^4}p\log(p)/n - V_1 N p\log(p)/n + \frac{M^2}{2h^4}s\log(p)/n - V_2\sqrt{\log(p)/n}\sum_{(t,j)\in Z^c}|\phi_{tj}| \\
&- V_2 M s\log(p)/n + \lambda_{sk}\sum_{t,j\in Z^c}|\phi_{tj}| - \lambda_{sk}Ms\sqrt{\log(p)/n} - \lambda_{sm}\left(24M^2 s\log(p)/n + 4\sum_{t,j}|A_{tj,0} - B_{tj,0}|^2\right) \\
= &Np\log(p)/n\left(\frac{N}{8h^4} - V_1 - 4\tau s/(Np)\right) + \sum_{t,j\in Z^c}|\phi_{tj}|\left(\lambda_{sk} - V_2\sqrt{\log(p)/n}\right) \\
&+ s\log(p)/n\left[\left(\frac{1}{2h^4} - 24\lambda_{sm}\right)M^2 - \left(V_2 + \frac{\lambda_{sk}}{\sqrt{\log(p)/n}}\right)M\right],
\end{aligned}$$

where $\tau = s\iota\sum_{t,j}|A_{tj,0} - B_{tj,0}|^2$ and ι is the boundary due to $\lambda_{sm} = O_p(\log(p)/n)$. Here V_1 and V_2 are only related to n and ϵ. Note that, for a sufficiently large n, $(2h^4)^{-1} > 24\lambda_{sm}$ if $\lambda_{sm} = o(1)$. Hence, we can choose $N > 8h^4V_1$ for sufficient large p, $\lambda_{sk} = K_{sk}\sqrt{\log(p)/n}$ with $K_{sk} > V_2$, $\lambda_{sm} = o(1)$ and $\tau = O_p(1)$. By choosing M such that $M > \left(V_2 + K_{sk}\right)/\left(\frac{1}{2h^4} - 24\lambda_{sm}\right)$, we get $G(\Delta_T, \Delta_D) > 0$, which completes the proof of Theorem 1.

Proof of Theorem 2. Since the eigenvalues of Σ_0 are bounded, it implies that $|\mathbf{T}_0| = O(1)$ and $|\mathbf{D}_0| = O(1)$. Moreover, given $\|\mathbf{T} - \mathbf{T}_0\|_F^2 = O_p(s\log(p)/n)$ and $\|\mathbf{D} - \mathbf{D}_0\|_F^2 = O_p(p\log(p)/n)$, we have

$$\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_0^{-1}\|_F^2 = O_p\left(\|\Delta_{\mathbf{T}}\|_F^2\right) + O_p\left(\|\Delta_{\mathbf{D}}\|_F^2\right) = O_p((s+p)\log(p)/n).$$

Consequently, we have

$$\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_0^{-1}\|_F = O_p(\sqrt{(s+p)\log(p)/n}).$$

The same argument also applies to the covariance matrices $\mathbf{\Sigma}$. Thus, the following also holds,

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F = O_p(\sqrt{(s+p)\log(p)/n}),$$

which establishes Theorem 2.

References

[1] J. Yao, S. Zheng, Z. Bai, Large Sample Covariance Matrices and High-Dimensional Data Analysis, Cambridge University Press, Cambridge, 2015.
[2] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[3] P.J. Bickel, B. Li, Regularization in statistics, Test 15 (2006) 271–344.
[4] R.F. Engle, O. Ledoit, M. Wolf, Large dynamic covariance matrices, J. Bus. Econom. Statist. 37 (2) (2019) 363–375.
[5] O. Ledoit, M. Wolf, Analytical nonlinear shrinkage of large-dimensional covariance matrices, Ann. Statist. 48 (5) (2020) 3043–3065.
[6] O. Bodnar, T. Bodnar, N. Parolya, Recent advances in shrinkage-based high-dimensional inference, J. Multivariate Anal. 188 (2022) 104826.
[7] M. Pourahmadi, Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation, Biometrika 86 (1999) 677–690.
[8] W.B. Wu, M. Pourahmadi, Nonparametric estimation of large covariance matrices of longitudinal data, Biometrika 90 (4) (2003) 831–844.
[9] M. Pourahmadi, High-Dimensional Covariance Estimation: with High-Dimensional Data, John Wiley & Sons, Hoboken, New Jersey, 2013.
[10] X. Kang, X. Deng, An improved modified Cholesky decomposition approach for precision matrix estimation, J. Stat. Comput. Simul. 90 (2020) 443–464.
[11] X. Kang, X. Deng, On variable ordination of Cholesky-based estimation for a sparse covariance matrix, Canad. J. Statist. 49 (2) (2021) 283–310.
[12] J.Z. Huang, N. Liu, M. Pourahmadi, L. Liu, Covariance matrix selection and estimation via penalised normal likelihood, Biometrika 93 (2006) 85–98.
[13] P.J. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. (2008) 2577–2604.
[14] W. Tai, Regularized Estimation of Covariance Matrices for Longitudinal Data through Smoothing and Shrinkage (Ph.D. thesis), Columbia University, 2009.
[15] D. Dai, J. Pan, Y. Liang, Regularized estimation of the Mahalanobis distance based on modified Cholesky decomposition, Commun. Stat.: Case Stud. Data Anal. Appl. 8 (4) (2022) 559–573.

[16] D. Dai, C. Hao, S. Jin, Y. Liang, Regularized estimation of kronecker structured covariance matrix using modified cholesky decomposition, J. Stat. Comput. Simul. 95 (5) (2025) 905–930.
[17] D. Eddelbuettel, R. François, J. Allaire, K. Ushey, Q. Kou, N. Russel, J. Chambers, D. Bates, Rcpp: Seamless R and C++ integration, J. Stat. Softw. 40 (2011) 1–18.
[18] A.J. Rothman, E. Levina, J. Zhu, A new approach to Cholesky-based covariance regularization in high dimensions, Biometrika 97 (3) (2010) 539–550.
[19] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (1) (2007) 19–35.
[20] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79.
[21] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Ann. Stat. 37 (6B) (2009).
[22] X. Jiang, Joint Estimation of Covariance Matrix via Cholesky Decomposition (Ph.D. thesis), Department of Statistics and Applied Probability, National University of Singapore, 2012.
[23] S. Andersson, Invariant normal models, Ann. Statist. 3 (1) (1975) 132–154.
[24] K. Filipiak, M. John, D. Klein, Testing independence under a block compound symmetry covariance structure, Statist. Papers 64 (2) (2023) 677–704.