Large Covariance Matrix Estimation with Nonnegative Correlations

Yixin Yan
ShanghaiTech

Qiao Yang
ShanghaiTech

Ziping Zhao
ShanghaiTech

Abstract

Covariance matrix estimation is a fundamental problem in multivariate data analysis. In many situations, it is often observed that variables exhibit a positive linear dependency, indicating a positive linear correlation. This paper tackles the challenge of estimating covariance matrices with positive correlations in high-dimensional settings. We propose a positive definite thresholding covariance estimation problem that includes nonconvex sparsity penalties and nonnegative correlation constraints. To address this problem, we introduce a multi-stage adaptive estimation algorithm based on majorization-minimization (MM). This algorithm progressively refines the estimates by solving a weighted ℓ_1-regularized problem at each stage. Additionally, we present a comprehensive theoretical analysis that characterizes the estimation error associated with the estimates generated by the MM algorithm. The analysis reveals that the error comprises two components: the optimization error and the statistical error. The optimization error decreases to zero at a linear rate, allowing the proposed estimator to eventually reach the oracle statistical rate under mild conditions. Furthermore, we explore various extensions based on the proposed estimation technique. Our theoretical findings are supported by extensive numerical experiments conducted on both synthetic and real-world datasets.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

1 INTRODUCTION

The covariance matrix plays a crucial role in various fields of science and engineering, including dimension reduction (Bishop and Nasrabadi, 2006), variable screening (Ke et al., 2014), portfolio optimization (Markowitz and Todd, 2000; Zhao and Palomar, 2018; Zhao et al., 2019), transcription analysis (Schäfer and Strimmer, 2005), and so on. However, the covariance matrix is not directly observable. Therefore, developing efficient methods to accurately estimate this quantity is an important task in multivariate data analysis (Bai and Shi, 2011; Fan et al., 2016; Ebadi et al., 2022). One of the most commonly used estimators is the sample covariance matrix (SCM), which performs well in low-dimensional contexts where the sample size significantly exceeds the number of variables. However, SCM becomes less effective in high-dimensional settings (Zhou et al., 2011; Pourahmadi, 2013; Tong et al., 2014). To improve the accuracy of large covariance matrix estimation, structural regularization is often employed to achieve a consistent estimator (Peter J. Bickel and Elizaveta Levina, 2008; El Karoui, 2008; Lam and Fan, 2009; Rothman, 2012).

In many applications, variables frequently demonstrate positive (or nonnegative) linear dependency, indicating a positive (or nonnegative) correlation. This positive correlation establishes a distinct structure in the covariance matrix. If this structure can be effectively integrated into the estimation process, it could lead to a more accurate covariance estimator, thereby enhancing the performance of subsequent tasks. For instance, in finance, numerous empirical studies suggest that assets often exhibit similar movement patterns due to common factors such as market conditions and economic influences (Agrawal et al., 2022; Zhou et al., 2022). A better portfolio is more likely to be achieved when covariance matrix estimation takes into account these positive correlations (Agrawal et al., 2022). In psychometrics, individuals with negative self-views tend to display consistent behavior patterns, making them more predictable compared to those with positive self-views (Malle and Horowitz, 1995). Therefore, a covariance matrix that acknowledges positive

correlations can assist in the design of timely mental health interventions (Lauritzen et al., 2019). Moreover, species with similar traits often exhibit high positive correlations (Kemp and Tenenbaum, 2009), which is essential for biological reasoning. For species with similar environmental adaptations, estimating a positive covariance matrix improves classification performance (Lake and Tenenbaum, 2010).

In high-dimensional estimation, most variables are often uncorrelated or weakly correlated, implying that many entries in the true covariance matrix are zero or nearly zero. To exploit this structure, sparsity is commonly introduced as a key assumption (Bien and Tibshirani, 2011), which reduces the complexity of covariance estimation. To enable flexible sparsity selection, sparsity regularization methods have been proposed for sparse covariance matrix estimation (Xia et al., 2023; Zou and Zhao, 2024; Xia et al., 2024). In covariance estimation with a nonnegative correlation constraint, the constraint itself can enforce some zero elements (Agrawal et al., 2022). However, it cannot adjust the sparsity levels to specific values, a feature that is often sought in practice (Lam and Fan, 2009; Bien and Tibshirani, 2011). Consequently, sparsity regularization techniques have also been explored in the context of nonnegative covariance estimation (Ying et al., 2023).

In this paper, we study the problem of covariance matrix estimation with nonnegative correlations in high-dimensional contexts. The main contributions are summarized as follows:

- We consider a positive definite thresholding covariance estimation problem that features nonconvex sparsity penalties and nonnegative correlation constraints. To solve this problem, we propose an adaptive method based on majorization-minimization (MM), which consists of multiple stages. In the first stage, we solve an ℓ_1-regularized positive definite covariance estimation problem, yielding an initial estimate. In the subsequent stages, we iteratively refine this initial estimate by solving a series of adaptive ℓ_1-regularized problems. To solve the resultant regularized problems, we develop a proximal gradient descent (PGD) algorithm.

- We establish theoretical guarantees on the estimation error of the proposed method, which consists of two terms: optimization error and statistical error. The optimization error decays to zero at a linear rate, indicating that the estimate is refined iteratively in subsequent stages, while the statistical error does not decrease during iterations and takes the order of $\sqrt{\frac{s}{n}}$, i.e., the oracle rate, where n and s denote the sample size and number of nonzero elements of the underlying covariance matrix, respectively.

- Experiments on synthetic and financial time-series data demonstrate that our method outperforms state-of-the-art techniques in estimating the covariance matrix. Our method achieves a better estimation result for the covariance matrix, which further supports our theoretical analysis.

2 RELATED WORK

Sparse Covariance Matrix Estimation. Among the techniques for estimating sparse covariance matrices, thresholding is a classic and straightforward approach. This method promotes sparsity by either directly setting smaller elements in the SCM to zero (Bickel, Peter J. and Levina, Elizaveta, 2008; Cai and Liu, 2011) or by incorporating a penalty term into the least squares formulation (Fan and Li, 2001; Rothman et al., 2009; Zhang, 2010; Cai et al., 2011). While covariance estimators that utilize thresholding can effectively enhance sparsity, they do not guarantee that the covariance matrix will be positive (semi)definite with finite samples, which is an essential characteristic in many applications (Fan et al., 2016). To address this limitation, regularization techniques have been employed to ensure the positive definiteness of covariance estimators under thresholding. For instance, methods such as applying an eigenvalue constraint (Xue et al., 2012; Liu et al., 2014) or utilizing a logarithmic barrier penalty (Rothman, 2012) has been introduced to enhance a soft-thresholding estimator.

Sparse Precision Matrix Estimation. In addition to directly pursuing sparsity in the covariance matrix, another common approach involves imposing structural assumptions on the inverse covariance matrix (Dempster, 1972; Yuan, 2010; Cai et al., 2016), which is referred to as the precision matrix. For example, the graphical lasso (Friedman et al., 2008; Mazumder and Hastie, 2012), which is based on the assumption of a sparse precision matrix, allows for the estimation of a covariance matrix in high-dimensional settings. Under the assumption of Gaussian maximum likelihood (ML) estimation, the problem is also known as the Gaussian graphical model (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Banerjee et al., 2008). Unlike methods that estimate sparse covariance matrices by leveraging the sparse correlation properties among variables, sparse precision matrix estimation relies on the assumption that the partial correlations among variables are sparse. However, this approach does not necessarily yield a sparse covariance matrix.

Precision Matrix Estimation with Nonpositive

Partial Correlations. An intriguing property related to positive correlation is multivariate total positivity of order 2 (MTP_2) (Fortuin et al., 1971; Karlin and Rinott, 1980; Colangelo et al., 2005). A Gaussian random variable is MTP_2 if its precision matrix is a symmetric M-matrix, meaning all off-diagonal elements are nonpositive (S. Karlin and Y. Rinott, 1983), which corresponds to nonnegative partial correlations. Notably, this structure directly implies that the covariance matrix has nonnegative off-diagonal elements, ensuring nonnegative correlations. This property has driven extensive research on (sparse) precision matrix estimation under the MTP_2 constraint (Lauritzen et al., 2019; Soloff et al., 2020; Agrawal et al., 2022; Ying et al., 2023).

A precision matrix satisfying the MTP_2 constraint is a sufficient but not necessary condition for ensuring nonnegative correlations, except in the bivariate case. As a result, imposing the MTP_2 constraint in precision matrix estimation may overly restrict the feasible domain when the primary objective is to obtain a covariance matrix with nonnegative correlations. This distinction can be illustrated through the following example. Consider a nonnegative covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}.$$

When we compute its inverse, we obtain:

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 0.625 & 0.125 & -0.250 \\ 0.125 & 0.625 & -0.250 \\ -0.250 & -0.250 & 0.500 \end{bmatrix}.$$

In this case, not all partial correlations are nonpositive, highlighting the limitations of the MTP_2 constraint in dimensions larger than two.

Covariance Matrix Estimation with Nonnegative Correlations. In (Zhou et al., 2022; Fatima et al., 2022), the authors proposed a method for estimating a covariance matrix with nonnegative correlations using Gaussian ML estimation. In (Zhou et al., 2022), a low-rank structure was further applied to the covariance matrix to improve the estimation performance. However, the ML-based formulation is generally not well-suited for high-dimensional estimation scenarios, where the sample covariance matrix is singular. This singularity causes the objective function to be unbounded from below. Additionally, the Gaussian ML estimation problem is nonconvex, which poses significant challenges in analyzing the properties of its solutions.

3 NOTATION

Lower-case and upper-case letters represent scalars. Boldface lower-case and upper-case letters denote vectors and matrices, respectively. X_{ij} refers to the (i,j)-th element of the matrix $\boldsymbol{X}$. $\mathbb{R}^n$ denotes the set of $n\times 1$ vectors, and $\mathbb{R}^{m\times n}$ denotes the set of $m \times n$ matrices. $\mathbf{0}$ stands for the all-zero vector or matrix. $\mathbf{1}$ stands for the all-one vector. $\boldsymbol{I}$ stands for the identity matrix.

$\boldsymbol{X} \succeq \mathbf{0}$ denotes that $\boldsymbol{X}$ is symmetric positive semidefinite, while $\boldsymbol{X} \succ \mathbf{0}$ denotes that $\boldsymbol{X}$ is symmetric positive definite. $\boldsymbol{X} \geq \mathbf{0}$ represents each element in $\boldsymbol{X}$ is nonnegative. $\boldsymbol{X}^\top$, $\boldsymbol{X}^{-1}$, and $\det(\boldsymbol{X})$ denote the transpose, inverse, and determinant of $\boldsymbol{X}$, respectively. $\lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X})$ represent the maximum and minimum eigenvalues of $\boldsymbol{X}$, respectively. $\|\boldsymbol{X}\|_{\text{F}}$ and $\|\boldsymbol{X}\|_2$ denote the Frobenius and spectral norms, respectively. $\|\boldsymbol{X}\|_{\max}$ and $\|\boldsymbol{X}\|_{\min}$ are used to denote the max-absolute-value and minimum-absolute-value norms, respectively. The Frobenius inner product is defined as $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \sum_{i,j} X_{ij}Y_{ij}$.

For an index set $\mathcal{S}$, we use $\text{card}(\mathcal{S})$ to represent its cardinality, $\mathcal{S}^c$ to denote its complement, and $\boldsymbol{X}_{\mathcal{S}}$ to denote the matrix whose (i,j)-th element is equal to X_{ij} if $(i,j) \in \mathcal{S}$, and zero, otherwise.

f' represents the derivative of a univariate function f. ∇f denotes the gradient of a multivariate function f. Given functions $f(x)$ and $g(x)$, we use $f(x) \gtrsim g(x)$ if $f(x) \geq cg(x)$, $f(x) \lesssim g(x)$ if $f(x) \leq cg(x)$, and $f(x) \asymp g(x)$ if $cg(x) \leq f(x) \leq Cg(x)$ for some positive constants c and C. $O_p(\cdot)$ is used to denote being bounded in probability.

4 PROBLEM FORMULATION

Consider a zero-mean random vector $\boldsymbol{x} \in \mathbb{R}^d$ following a Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$. Given n independent and identically distributed observed data instances $\{\boldsymbol{x}_i\}_{i=1}^n$, the SCM estimator is computed as $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$. Our goal is to estimate $\boldsymbol{\Sigma}^\star$ which has nonnegative correlations. We consider the following estimation problem:

$$\begin{aligned} \min_{\boldsymbol{\Sigma}} \quad & \frac{1}{2}\|\boldsymbol{\Sigma} - \boldsymbol{S}\|_{\text{F}}^2 - \tau \log\det(\boldsymbol{\Sigma}) + \sum_{i\neq j} p_\lambda(\Sigma_{ij}) \\ \text{s.t.} \quad & \boldsymbol{\Sigma} \succeq \mathbf{0}, \quad \boldsymbol{\Sigma} \geq \mathbf{0}, \end{aligned} \tag{1}$$

The objective in (1) consists of three terms: the first term is a data fidelity term; the second one is a log-barrier regularizer with $\tau \geq 0$ to ensure the positive definiteness of the estimates, i.e., $\boldsymbol{\Sigma} \succ \mathbf{0}$; and the third one is a sparsity-inducing regularizer, where p_λ with $\lambda \geq 0$ is a nonconvex function satisfying the following

assumptions. The SCM in (1) serves as a pilot estimator, which can be substituted with other alternatives depending on the specific context (Avella-Medina et al., 2018).

Assumption 1. *The function p_λ, defined on the domain $[0,+\infty)$, satisfies the following conditions:*

1. *$p_\lambda(t)$ is monotonically nondecreasing and smooth on $[0,+\infty)$ with $p_\lambda(0)=0$, and is differentiable almost everywhere on $[0,+\infty)$.*
2. *$p'_\lambda(t)$ is monotonically nonincreasing on $[0,+\infty)$ with $p'_\lambda(0)=\lambda$.*
3. *There exists a constant α such that $p'_\lambda(\alpha\lambda) \geq \frac{\sqrt{2}}{2}\lambda$, and another constant $\gamma > 0$ such that $p'_\lambda(t)=0$ for any $t \geq \gamma\lambda$.*

Functions satisfying the above assumptions are commonly referred to as folded concave penalty functions. Common examples include the smooth clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010).

5 OPTIMIZATION ALGORITHM

In this section, we propose an algorithm based on the generic majorization-minimization (MM) framework (Hunter and Lange, 2004; Sun et al., 2016; Fan et al., 2018) to address the estimation problem (1). The algorithm resolves the original nonconvex estimation problem by a series of convex subproblems, whose objective serves as an upper bound for the original objective.

5.1 A Majorization-Minimization Algorithm

Based on the MM procedure, we can derive a weighted ℓ_1-norm surrogate for the sparsity inducing regularizer at each iteration, leading to a convex subproblem. By repeating this process, multiple stages are established, where each one refines the estimate from previous stage. Specifically, at the k-th stage, we obtain the estimate $\boldsymbol{\Sigma}^k$ by solving the following convex optimization problem:[1]

$$\begin{aligned} \min_{\boldsymbol{\Sigma}} \quad & \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_{\mathrm{F}}^2 - \tau \log\det(\boldsymbol{\Sigma}) + \sum_{i\neq j} \Lambda_{ij}^k \Sigma_{ij} \\ \text{s.t.} \quad & \boldsymbol{\Sigma} \succeq \mathbf{0}, \quad \boldsymbol{\Sigma} \geq \mathbf{0}, \end{aligned} \tag{2}$$

where, for $i \neq j$, $\Lambda_{ij}^k = p'_\lambda\left(\Sigma_{ij}^{k-1}\right)$ with $\boldsymbol{\Sigma}^{k-1}$ denoting the solution from the $(k-1)$-th stage. It is easy to see that whenever the estimate, $\boldsymbol{\Sigma}^{k-1}$, exhibits a larger value for its (i,j)-th element, according to Assumption 1, a smaller Λ_{ij}^k should be allocated in the next stage. This multi-stage convex relaxation algorithm is summarized in Algorithm 1. We choose $\boldsymbol{\Sigma}^0 = \boldsymbol{I}$, which implies $\Lambda_{ij}^1 = p'_\lambda\left(\Sigma_{ij}^0\right) = \lambda$. In this case, this optimization problem becomes an ℓ_1-regularized positive definite nonnegative covariance matrix estimation problem.

[1] Since Σ_{ij} is nonnegative, the absolute value sign is reduced.

Algorithm 1 MM-based multistage algorithm for (1)

Input: $\boldsymbol{S}$, τ, λ;
Initialize: $\boldsymbol{\Sigma}^0 = \boldsymbol{I}$;
for $k = 1, 2, \ldots, K$ **do**
  $\Lambda_{ij}^k = p'_\lambda\left(\Sigma_{ij}^{k-1}\right)$;
  obtain $\boldsymbol{\Sigma}^k$ by solving (2);
  $k = k+1$;
end
Output: $\boldsymbol{\Sigma}^K$.

5.2 Proximal Gradient Descent

The problem (2) within the multistage algorithm is convex, which can be solved by many methods. In this paper, we adopt the PGD algorithm (Rolfs et al., 2012) with a backtracking line search procedure to solve it.

We first rewrite the objective in problem (2) into the following compact form:

$$f_k(\boldsymbol{\Sigma}) = \frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{S}+\boldsymbol{\Lambda}^k\right\|_{\mathrm{F}}^2 - \tau\log\det(\boldsymbol{\Sigma}),$$

where we define $\Lambda_{ii}^k = 0$ in $\boldsymbol{\Lambda}^k$. Denote the t-th iterate of the PGD algorithm in solving the k-th stage problem as $\boldsymbol{\Sigma}_t^k$. In the $(t+1)$-th iteration, an isotropic quadratic approximation of $f_k(\boldsymbol{\Sigma})$ at $\boldsymbol{\Sigma}_t^k$ is derived as:

$$\begin{aligned} f_{k,t}(\boldsymbol{\Sigma}) &= f_k\left(\boldsymbol{\Sigma}_t^k\right) + \left\langle \nabla f_k\left(\boldsymbol{\Sigma}_t^k\right), \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t^k\right\rangle \\ &\quad + \frac{\phi_t}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t^k\right\|_{\mathrm{F}}^2 \\ &= \frac{\phi_t}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t^k+\phi_t^{-1}\nabla f_k\left(\boldsymbol{\Sigma}_t^k\right)\right\|_{\mathrm{F}}^2 + \text{const.}, \end{aligned}$$

where $\phi_t > 0$. We update $\boldsymbol{\Sigma}_{t+1}^k$ by solving the following problem:

$$\begin{aligned} \min_{\boldsymbol{\Sigma}} \quad & \frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t^k+\phi_t^{-1}\nabla f_k\left(\boldsymbol{\Sigma}_t^k\right)\right\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \boldsymbol{\Sigma} \succ \mathbf{0}, \quad \boldsymbol{\Sigma} \geq \mathbf{0}. \end{aligned} \tag{3}$$

To guarantee the positive definiteness of $\boldsymbol{\Sigma}_{t+1}^k$, we solve (3) based on a backtracking line search procedure on ϕ_t, i.e., finding a ϕ_t such that $f_{k,t}\left(\boldsymbol{\Sigma}_{t+1}^k\right) \geq f_k\left(\boldsymbol{\Sigma}_{t+1}^k\right)$. The obtained ϕ_t ensures $\boldsymbol{\Sigma}_{t+1}^k$ is positive definite. Then, the update of $\boldsymbol{\Sigma}_{t+1}^k$ is expressed as:

$$\boldsymbol{\Sigma}_{t+1}^k = \max\left(\boldsymbol{\Sigma}_t^k - \phi_t^{-1}\nabla f_k\left(\boldsymbol{\Sigma}_t^k\right), \mathbf{0}\right), \tag{4}$$

Algorithm 2 Proximal gradient descent for (2)

Input: $\boldsymbol{\Sigma}^{k-1}, \boldsymbol{\Lambda}^{k-1}$;
Initialize: $\boldsymbol{\Sigma}_0^k = \boldsymbol{\Sigma}^{k-1}$, $\phi_0 > 0$, $\zeta > 1$, $t = 0$;
repeat
  repeat
    $\boldsymbol{\Sigma}_{t+1}^k = \max\left(\boldsymbol{\Sigma}_t^k - \phi_t^{-1}\nabla f_k\left(\boldsymbol{\Sigma}_t^k\right), \mathbf{0}\right)$;
    if $f_{k,t}\left(\boldsymbol{\Sigma}_{t+1}^k\right) < f_k\left(\boldsymbol{\Sigma}_{t+1}^k\right)$ or $\boldsymbol{\Sigma}_{t+1}^k \nsucc \mathbf{0}$ **then**
      $\phi_t = \zeta\phi_t$;
    end
  until $f_{k,t}\left(\boldsymbol{\Sigma}_{t+1}^k\right) \geq f_k\left(\boldsymbol{\Sigma}_{t+1}^k\right)$ and $\boldsymbol{\Sigma}_{t+1}^k \succ \mathbf{0}$;
  $\phi_{t+1} = \max\left(\phi_0, \zeta^{-1}\phi_t\right)$;
  $t = t + 1$;
until some convergence criterion is met;
Output: $\boldsymbol{\Sigma}^k = \boldsymbol{\Sigma}_t^k$.

where the max operator $\max(\cdot,\cdot)$ is applied elementwise. The overall PGD algorithm is summarized in Algorithm 2. At the k-th stage, we set $\boldsymbol{\Sigma}_0^k = \boldsymbol{\Sigma}^{k-1}$, which acts as warm start for the inner problem (2) to accelerate the convergence. The minimizer for problem (2) is guaranteed by the globally strong convexity of the problem (2). The theoretical convergence property of Algorithm 2 is described in the following theorem.

Theorem 1. *The sequence $\left\{\boldsymbol{\Sigma}_t^k\right\}_{t\geq 0}$ established by Algorithm 2 converges to the optimal solution of problem (2) for $k = 1, \ldots, K$.*

6 ESTIMATION ERROR ANALYSIS

In this section, we first introduce some necessary assumptions for theoretical analysis. Then, we provide the statistical properties of the proposed estimator.

6.1 Technical Assumptions

Let $\boldsymbol{\Sigma}^\star$ denote the true covariance matrix. We define the support set of $\boldsymbol{\Sigma}^\star$ as $\mathcal{S} = \left\{(i,j) \mid \Sigma_{ij}^\star \neq 0, i \neq j\right\}$, and use s to represent its cardinality, i.e., $s = \operatorname{card}(\mathcal{S})$. Moreover, we denote the set of diagonal elements in $\boldsymbol{\Sigma}^\star$ as $\mathcal{I} = \{(i,i) \mid i \in \{1,2,\ldots,d\}\}$. In the following, we impose a mild assumption on the true covariance matrix $\boldsymbol{\Sigma}^\star$.

Assumption 2. *The true covariance matrix $\boldsymbol{\Sigma}^\star$ satisfies*

$$\min_{(i,j)\in\mathcal{S}} \Sigma_{ij}^\star \geq (\alpha + \gamma)\lambda \gtrsim \lambda,$$

where α and γ are constants defined in Assumption 1.

Assumption 2, commonly referred to as minimum signal strength assumption, has been widely applied in the analysis of nonconvex penalized problems (Sun et al., 2018; Wei and Zhao, 2023; Ying et al., 2023). It is rather mild, as the regularization parameter λ in our statistical analysis takes the order of $\sqrt{\frac{\log d}{n}}$, which could be very small when the sample size n increases.

6.2 Theoretical Results

We now present the main theorems related to the statistical convergence rate of our proposed estimator, which demonstrates the estimation error of solution path $\left\{\boldsymbol{\Sigma}^k\right\}_{k\geq 1}$ generated from Algorithm 1.

Theorem 2. *Under Assumptions 1 and 2, taking the regularization parameter $\lambda \asymp \sqrt{\frac{\log d}{n}}$, then the $\boldsymbol{\Sigma}^k$ generated by Algorithm 1 satisfies the following property:*

$$\left\|\boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq \underbrace{\delta^{k-1}\left\|\boldsymbol{\Sigma}^1 - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}}_{\textit{optimization error}} + \underbrace{\frac{1}{1-\delta}\left(\left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}} + \tau\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}}\right)}_{\textit{statistical error}},$$

for $k = 1, \ldots, K$ with high probability, where $\delta \in (0,1)$ is the contraction parameter.

Theorem 2 establishes that the estimation error between the estimated covariance matrix $\boldsymbol{\Sigma}^k$ and the true covariance matrix $\boldsymbol{\Sigma}^\star$ can be upper bounded by two terms: optimization error and statistical error. Now, we provide the explicit statistical rate of convergence under the sub-Gaussian case.

Corollary 3. *Let $\boldsymbol{x}$ be a zero-mean sub-Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}^\star$. Under the same conditions in Theorem 2, if $\lambda \asymp \sqrt{\frac{\log d}{n}}$, $\tau \lesssim \sqrt{\frac{s}{n}}\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}}^{-1}$, then the solution $\boldsymbol{\Sigma}^1$ satisfies*

$$\left\|\boldsymbol{\Sigma}^1 - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{s\log d}{n}},$$

with high probability.

Corollary 3 is a direct consequence of Theorem 2 for $k = 1$. It is known that $\boldsymbol{\Sigma}^1$ is the optimal solution of the first subproblem of (1). Due to the contraction parameter δ induced by the MM-based convex relaxation algorithm, to achieve the oracle rate, we should make the number of stages K large enough. Then the following corollary can be provided.

Corollary 4. *Let $\boldsymbol{x}$ be a zero-mean sub-Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}^\star$. Under the same conditions in Theorem 2, if $\lambda \asymp \sqrt{\frac{\log d}{n}}$, $\tau \lesssim \sqrt{\frac{s}{n}}\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}}^{-1}$, and $K \gtrsim \log(\lambda\sqrt{n}) \gtrsim \log\log d$, then the solution $\boldsymbol{\Sigma}^K$ satisfies*

$$\left\|\boldsymbol{\Sigma}^K - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} = O_p\left(\sqrt{\frac{s}{n}}\right).$$

Corollary 4 implies that under weak assumptions, we just need to solve no more than approximately $\log\log d$ convex optimization problems in (2) to achieve the statistical rate of $O_p\left(\sqrt{\frac{s}{n}}\right)$, which is the oracle rate[2].

7 EXTENSIONS

7.1 Correlation Matrix Estimation

Compared to the covariance matrix, the correlation matrix is scale-invariant because its diagonal elements are fixed to be one. Hence, directly estimating the correlation matrix is more efficient for capturing linear correlations (Rothman et al., 2008; Lam and Fan, 2009; Cui et al., 2016; Liu et al., 2014). In this section, we show that the covariance estimation method discussed in previous sections can be extended to estimate correlation matrices with nonnegative correlations. We define the true correlation matrix as $\boldsymbol{\Gamma}^\star = \mathrm{Diag}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\boldsymbol{\Sigma}^\star\mathrm{Diag}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}$. Given the SCM $\boldsymbol{S}$, the sample correlation matrix is computed as $\boldsymbol{R} = \mathrm{Diag}\left(\boldsymbol{S}\right)^{-1/2}\boldsymbol{S}\mathrm{Diag}\left(\boldsymbol{S}\right)^{-1/2}$. Then, we consider the following correlation matrix estimation problem:

$$\begin{aligned}\min_{\boldsymbol{\Gamma}} \quad & \frac{1}{2}\left\|\boldsymbol{\Gamma}-\boldsymbol{R}\right\|_{\mathrm{F}}^2 - \tau\log\det\left(\boldsymbol{\Gamma}\right) + \sum_{i\neq j} p_\lambda\left(\Gamma_{ij}\right) \\ \text{s.t.} \quad & \boldsymbol{\Gamma}\succeq\mathbf{0}, \quad \boldsymbol{\Gamma}\geq\mathbf{0}, \quad \mathrm{Diag}(\boldsymbol{\Gamma})=\boldsymbol{I},\end{aligned} \tag{5}$$

where the constraint $\mathrm{Diag}(\boldsymbol{\Gamma})=\boldsymbol{I}$ is imposed to guarantee the result $\boldsymbol{\Gamma}$ to be a correlation matrix.

The MM-based multistage algorithm in Algorithm 1 is applicable to problem (5), which generates a sequence $\left\{\boldsymbol{\Gamma}^k\right\}_{k\geq 1}$. We start the algorithm with $\boldsymbol{\Gamma}^0=\boldsymbol{I}$, leading to $p'_\lambda\left(\Gamma^0_{ij}\right)=\lambda$. For each $\boldsymbol{\Gamma}^k$, a subproblem of (5) is solved by the PGD algorithm, which generates $\left\{\boldsymbol{\Gamma}^k_t\right\}_{t\geq 0}$. In the $(t+1)$-th iteration of PGD, the update of $\boldsymbol{\Gamma}^k_{t+1}$ is given by

$$\begin{aligned}&\left[\boldsymbol{\Gamma}^k_{t+1}\right]_{ij} \\ &= \begin{cases}\max\left(\left[\boldsymbol{\Gamma}^k_t\right]_{ij} - \phi_t^{-1}\left[\nabla f_k\left(\boldsymbol{\Gamma}^k_t\right)\right]_{ij}, 0\right) & i\neq j \\ 1 & i=j\end{cases}.\end{aligned}$$

7.2 Covariance Matrix Estimation for Heavy-Tailed Distributions

The covariance estimator we have discussed so far relies on the assumption of Gaussian or sub-Gaussian distributions, which is typical in many cases. However, certain types of real-world data are believed to follow heavy-tailed distributions (Catoni, 2012; Sun et al., 2015; Wei and Minsker, 2017). For instance, in finance, the returns of financial assets often exhibit heavy-tailed distributions, indicating that the probability of extreme returns is significantly higher than what a normal distribution would predict (Cont, 2001). Similarly, in genetic analysis, gene expression levels exhibit a heavy-tailed distribution, with a small subset of genes displaying significantly higher expression levels than the majority, leading to extreme values (Liu et al., 2003). Consequently, developing robust covariance estimation procedures that mitigate sensitivity to these distributional characteristics is crucial.

In this section, we address the problem of covariance estimation for heavy-tailed distributions. We assume $\boldsymbol{x}$ follows a heavy-tailed distribution with an unknown mean $\boldsymbol{\mu}^\star$ and covariance matrix $\boldsymbol{\Sigma}^\star$. Specifically, we represent $\boldsymbol{x}$ as

$$\boldsymbol{x}=\boldsymbol{\mu}^\star+\boldsymbol{e},$$

where $\boldsymbol{e}$ is a heavy-detailed error term. Given n observations $\{\boldsymbol{x}_i\}_{i=1}^n$, we construct a new sample using pairwise differences: $\{\boldsymbol{x}_i-\boldsymbol{x}_j\}_{1\leq i<j\leq n} = \{\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\ldots,\tilde{\boldsymbol{x}}_{\tilde{n}}\}$, where $\tilde{n}=n\left(n-1\right)/2$. Each $\tilde{\boldsymbol{x}}_i$ follows a heavy-detailed distribution with mean zero and covariance $2\boldsymbol{\Sigma}^\star$. Using this transformed sample, we define the following loss function (Ke et al., 2019):

$$L_\beta\left(\boldsymbol{\Sigma}\right)=\frac{1}{2\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{k,l=1}^{d} h_\beta\left(\Sigma_{kl}-\frac{1}{2}\tilde{x}_{ik}\tilde{x}_{il}\right),$$

where h_β is the Huber loss function (Huber and Ronchetti, 2011), defined as

$$h_\beta\left(x\right)=\begin{cases}\frac{x^2}{2}, & |x|\leq\beta, \\ \beta\left|x\right|-\frac{\beta^2}{2}, & |x|>\beta,\end{cases}$$

for some nonnegative parameter β.

The covariance matrix is estimated based on the following problem:

$$\begin{aligned}\min_{\boldsymbol{\Sigma}} \quad & L_\beta\left(\boldsymbol{\Sigma}\right) - \tau\log\det\left(\boldsymbol{\Sigma}\right) + \sum_{i\neq j} p_\lambda\left(\Sigma_{ij}\right) \\ \text{s.t.} \quad & \boldsymbol{\Sigma}\succeq\mathbf{0}, \quad \boldsymbol{\Sigma}\geq\mathbf{0}.\end{aligned} \tag{6}$$

The algorithm we proposed in Section 5 is also applicable to this problem.

7.3 Covariance Matrix Estimation with General Correlation Prior

While the focus of this work has been on estimating covariance matrices under nonnegative correlation constraints, the proposed estimation framework can be extended to accommodate more general prior information on correlation signs. In many applications, certain variables are expected to exhibit either positive

[2]The oracle rate refers to the statistical convergence rate of the estimator, which knows the true support set $\mathcal{S}\cup\mathcal{I}$ in advance.

or negative correlations due to domain-specific knowledge. For example, in genetics, gene expressions may be governed by regulatory mechanisms that impose structured sign constraints on the correlations (Jezequel et al., 2013; Anastasiadi et al., 2018).

To incorporate such general correlation sign priors, one can modify the constraint set in problem (3) to explicitly enforce both positive and negative correlation structures. Specifically, for a given sign prior matrix $\boldsymbol{M} \in \{-1, 0, 1\}^{d\times d}$, where $M_{ii} = 1$, $i = 1, \ldots, d$, and for $i \neq j$, $M_{ij} = 1$ indicates a positive correlation, $M_{ij} = -1$ enforces a negative correlation, and $M_{ij} = 0$ imposes no sign restriction, the covariance estimation problem can be formulated as:

$$\begin{aligned} \min_{\boldsymbol{\Sigma}} \quad & \frac{1}{2}\|\boldsymbol{\Sigma} - \boldsymbol{S}\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma} + \sum_{i\neq j} p_\lambda(\Sigma_{ij}) \\ \text{s.t.} \quad & \boldsymbol{\Sigma} \succeq \mathbf{0}, \quad \boldsymbol{M} \odot \boldsymbol{\Sigma} \geq \mathbf{0}, \end{aligned} \tag{7}$$

where $\odot$ denotes the Hadamard product. The above formulation naturally generalizes our previous approach, allowing for a broader class of structured constraints beyond purely nonnegative correlations. The MM-based multistage estimation algorithm can be directly extended to solve this problem by incorporating the sign priors into the PGD steps.

8 NUMERICAL EXPERIMENTS

In this section, we evaluate our proposed estimator against existing methods in high-dimensional settings using both synthetic and real financial data. The following benchmarks are used for comparison.

- SCM: the sample covariance matrix estimator;
- PDTE_L1: the positive definite thresholding estimator with the ℓ_1 penalty (Rothman, 2012);
- PDTE_FC: the positive definite thresholding estimator with a folded concave penalty (Wei and Zhao, 2023);
- MLE_L1: the ML estimator with the ℓ_1-norm penalty (Bien and Tibshirani, 2011);
- MLE_FC: the ML estimator with a folded concave penalty (Phan et al., 2017);
- MLE_MTP2: the ML estimator with MTP_2 constraint (Slawski and Hein, 2015);
- MLE_MTP2_L1: the ML estimator with MTP_2 constraint and the ℓ_1 penalty (Cai et al., 2024);
- MLE_MTP2_FC: the ML estimator with MTP_2 constraint and a folded concave penalty (Ying et al., 2023);
- MLE_NN: the ML estimator with the nonnegative correlation constraint (Zhou et al., 2022);
- PDTE_NN (prop.): the proposed positive definite thresholding estimator with the nonnegative correlation constraint;
- PDTE_NN_FC (prop.): the proposed positive definite thresholding estimator with the nonnegative correlation constraint and a folded concave penalty;

In terms of the folded concave penalty in the above methods, we choose the MCP function[3] and set the parameter $\xi = 2.7$ in all experiments. In the simulation, the SCM $\boldsymbol{S}$ is singular. Since the existence of solutions for MLE_L1, MLE_FC, and MLE_NN relies $\boldsymbol{S}$ to be positive definite we replace $\boldsymbol{S}$ in these methods by $\boldsymbol{S} + \epsilon\boldsymbol{I}$, where ϵ is a positive tuning parameter. We choose the parameters τ, λ, and ϵ by five-fold cross-validation. All estimation methods are implemented in Python and executed on a system with an Intel i7 2.90 GHz CPU.

8.1 Synthetic Data

We generate synthetic samples of size $n = 50$ based on Gaussian distributions with dimensions $d = \{100, 200\}$, which has zero mean and covariance matrices $\boldsymbol{\Gamma}^\star$ with nonnegative correlations. We consider the following three structures for $\boldsymbol{\Gamma}^\star$, which are commonly used in the literature (Rothman et al., 2009; Xue et al., 2012; Cui et al., 2016):

- Banded structure:

$$\Gamma^\star_{ij} = \begin{cases} 1 - \frac{|i-j|}{10} & |i-j| \leq 10, \\ 0 & \text{otherwise.} \end{cases}$$

- Block structure: The indices $1, 2, \ldots, d$ are partitioned into 10 equal-sized ordered groups with

$$\Gamma^\star_{ij} = \begin{cases} 1 & i = j, \\ 0.6 & i \text{ and } j\,(i \neq j) \text{ are in the same group,} \\ 0 & \text{otherwise.} \end{cases}$$

- Toeplitz structure: $\Gamma^\star_{ij} = 0.75^{|i-j|}$.

Among the three structures, the first two covariance structures are sparse, while the third one is approximately sparse. The effectiveness of the estimators is evaluated by the estimation error between the estimated and true covariance matrices, using both the Frobenius norm and the spectral norm. Additionally, the performance of variable selection is evaluated using the false positive rate (FPR) and the true positive

[3] The MCP function p_λ is defined as: $p_\lambda(x) = \lambda \int_0^{|x|} \max\left(1 - \frac{u}{\lambda\xi}, 0\right) du$ with $\xi \geq 1$.

Table 1: Quantitative comparison among different methods for the banded matrix structure.

	SCM	PDTE_L1	PDTE_FC	MLE_L1	MLE_FC	MLE_MTP2	MLE_MTP2_L1	MLE_MTP2_FC	MLE_NN	PDTE_NN (prop.)	PDTE_NN_FC (prop.)
						$d = 100, n = 50$					
$\|\cdot\|_F$	14.3198 (0.1167)	9.4292 (0.1050)	9.0993 (0.0991)	9.7818 (0.1112)	9.3832 (0.1319)	26.9979 (0.0019)	24.4005 (0.0021)	23.4577 (0.0016)	9.8749 (0.1023)	9.9850 (0.1153)	**8.5741 (0.0936)**
$\|\cdot\|_2$	7.7214 (0.1485)	4.4837 (0.0882)	4.3388 (0.0910)	4.9231 (0.0869)	4.3042 (0.1458)	23.3511 (0.0011)	20.4007 (0.0018)	19.1509 (0.0021)	5.5013 (0.1231)	5.7964 (0.1229)	**4.1201 (0.0968)**
FPR	NA	0.1596 (0.0049)	0.0592 (0.0038)	0.2374 (0.0022)	0.0601 (0.0431)	0.9999 (0.0001)	0.9999 (0.0001)	0.9999 (0.0001)	0.4132 (0.0103)	0.5162 (0.0051)	**0.0279 (0.0068)**
TPR	NA	0.8693 (0.0403)	0.8905 (0.0037)	0.8297 (0.0123)	0.8797 (0.0364)	**0.9999 (0.0001)**	**0.9999 (0.0001)**	**0.9999 (0.0001)**	0.9698 (0.0035)	0.9714 (0.0016)	0.9060 (0.0032)
Time	–	6.21	19.88	12.63	51.26	37.13	4.39	17.92	81.2356	**0.98**	1.02
						$d = 200, n = 50$					
$\|\cdot\|_F$	28.0354 (0.1408)	16.4932 (0.1525)	14.2286 (0.1081)	14.5659 (0.1809)	14.2989 (0.1043)	41.1309 (0.0013)	30.5522 (0.0025)	29.6237 (0.0019)	19.8766 (0.2163)	21.1721 (0.1765)	**13.6935 (0.1151)**
$\|\cdot\|_2$	12.1778 (0.1606)	5.5429 (0.1880)	5.2532 (0.0994)	5.7399 (0.0684)	5.3537 (0.0871)	36.2046 (0.0021)	23.5039 (0.0039)	22.0879 (0.0022)	6.9981 (0.1568)	9.0911 (0.2175)	**5.2259 (0.1233)**
FPR	NA	0.1706 (0.0026)	0.0210 (0.0038)	0.1334 (0.0082)	0.0240 (0.0213)	0.9999 (0.0001)	0.9999 (0.0001)	0.9999 (0.0001)	0.4283 (0.0031)	0.5350 (0.0024)	**0.0147 (0.0019)**
TPR	NA	0.8274 (0.0021)	0.8669 (0.0030)	0.8031 (0.0125)	0.8049 (0.0256)	**0.9999 (0.0001)**	**0.9999 (0.0001)**	**0.9999 (0.0001)**	0.9536 (0.0005)	0.9632 (0.0009)	0.8829 (0.0022)
Time	–	16.26	29.06	43.46	143.65	485.71	28.80	131.74	689.53	**2.53**	5.21

rate (TPR), which are defined as follows:

$$\text{FPR} = \frac{\#\left\{(i,j) : \Gamma_{ij} \neq 0 \,\&\, \Gamma^\star_{ij} = 0\right\}}{\#\left\{(i,j) : \Gamma^\star_{ij} = 0\right\}},$$
$$\text{TPR} = \frac{\#\left\{(i,j) : \Gamma_{ij} > 0 \,\&\, \Gamma^\star_{ij} > 0\right\}}{\#\left\{(i,j) : \Gamma^\star_{ij} > 0\right\}}.$$

All the reported results are averaged over 100 Monte Carlo simulations.

In Table **??**, we present the estimation results for the banded covariance matrix structure, with standard errors provided in parentheses. Our proposed estimators show the best performance in both cases. In Figure 1, we illustrate the estimation performance using correlation graphs. Each node represents a variable, with blue lines indicating a positive association between variables and red lines indicating a negative relationship. The intensity of the color reflects the strength of the association, with darker colors signifying stronger connections. Compared with other types of estimators, the proposed PDTE_NN_FC estimator achieves better estimation results.

8.2 Real Data

We further conduct experiments on financial time-series data. A commonly used approach to assess the quality of the estimated covariance matrix is by evaluating the risk of portfolios constructed from it (Markowitz, 1952). A covariance estimate is considered of higher quality if it results in lower portfolio return volatility (i.e., standard deviation). Following (Xue et al., 2012), we focus on the global minimum variance portfolio (GMVP) under the no-short sales constraint, formulated as:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$$
$$\text{s.t.} \quad \boldsymbol{w}^\top \mathbf{1} = 1, \quad \boldsymbol{w} \geq \mathbf{0}.$$

This optimization problem can be efficiently solved using CVX (Grant and Boyd, 2014).

We collect historical monthly stock prices for the S&P 100 Index components over a 240-month period (December 2002 to December 2022). After removing missing data, we obtain monthly returns for 78 companies ($d = 78$). To ensure the robustness of our results, we construct 100 datasets by selecting a random starting date, each containing the monthly returns of 78 stocks over 120 consecutive trading months. We then conduct experiments on each dataset using a rolling window scheme, where 60 months are used for training and one month for testing. The performance is evaluated by comparing the monthly volatility of portfolio returns.

Figure 2 presents the boxplot of the monthly volatility obtained using the GMVP across different estimators, along with the uniform portfolio ($\boldsymbol{w} = \mathbf{1}/d$), which serves as a heuristic baseline. From the figure, we observe that the uniform portfolio exhibits the highest monthly volatility, making it the least effective strategy. Additionally, the SCM, which lacks regularization, significantly underperforms compared to regularized methods. Among the regularized estimators, the proposed PDTE_NN_FC achieves lower volatility, demonstrating a notable advantage over the alternatives.

9 CONCLUSION

This paper has addressed the problem of covariance matrix estimation with nonnegative correlations in high-dimensional settings. We have proposed a positive definite thresholding method incorporating nonconvex sparsity penalties and nonnegative correlation constraints. To solve this problem, we have developed a multistage adaptive estimation algorithm based on the majorization-minimization principle. Theoretical

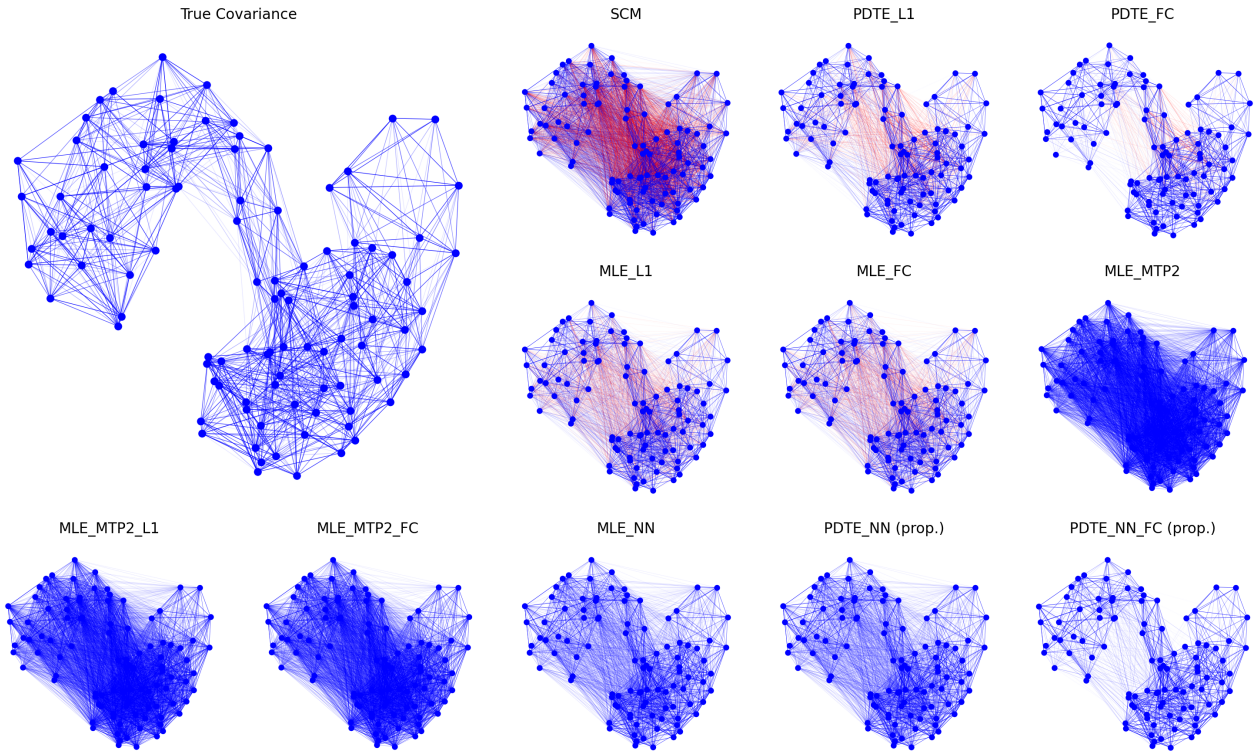

Figure 1: Correlation graph for the banded matrix structure with $d = 100$ and $n = 50$.

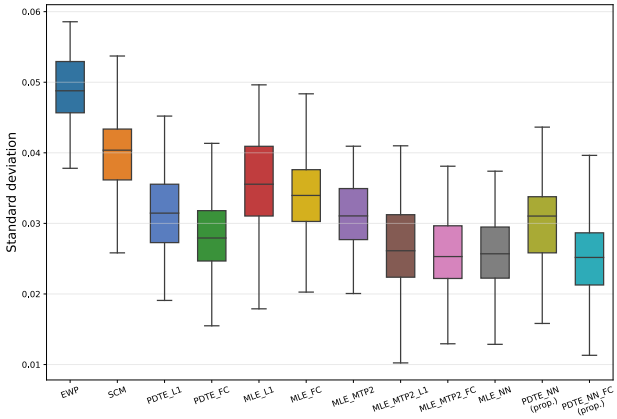

Figure 2: Boxplot of the monthly volatility of GMVP obtained using different estimators.

analysis has shown that the estimation error consists of an optimization error and a statistical error, with the former diminishing at a linear rate, allowing the estimator to achieve the oracle statistical rate under mild conditions. Numerical experiments on synthetic and real-world datasets have validated the effectiveness of our approach, demonstrating its advantages over existing methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under the Young Scientists Fund (Grant Nos. 62001295 and 72103137).

References

Raj Agrawal, Uma Roy, and Caroline Uhler. Covariance matrix estimation under total positivity for portfolio selection. *Journal of Financial Econometrics*, 20(2):367–389, 2022.

Dafni Anastasiadi, Anna Esteve-Codina, and Francesc Piferrer. Consistent inverse correlation between dna methylation of the first intron and gene expression across tissues and species. *Epigenetics & chromatin*, 11:1–17, 2018.

Marco Avella-Medina, Heather S Battey, Jianqing Fan, and Quefeng Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018.

Jushan Bai and Shuzhong Shi. Estimating high dimensional covariance matrices and its applications. *Annals of Economics and finance*, 12(2):199–215, 2011.

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

Bickel, Peter J. and Levina, Elizaveta. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604, 2008.

Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4): 807–820, 2011.

Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

Jian-Feng Cai, José Vinícius de Miranda Cardoso, Daniel Palomar, and Jiaxi Ying. Fast projected newton-like method for precision matrix estimation under total positivity. *Advances in Neural Information Processing Systems*, 36, 2024.

T. Tony Cai, Weidong Liu, and Harrison H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455 – 488, 2016.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Antonio Colangelo, Marco Scarsini, and Moshe Shaked. Some notions of multivariate positive dependence. *Insurance: Mathematics and Economics*, 37(1):13–26, 2005.

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

Ying Cui, Chenlei Leng, and Defeng Sun. Sparse estimation of high-dimensional correlation matrices. *Computational Statistics & Data Analysis*, 93: 390–403, 2016.

Arthur P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

Mohsen Ebadi, Shojaeddin Chenouri, Dennis KJ Lin, and Stefan H. Steiner. Statistical monitoring of the covariance matrix in multivariate processes: A literature review. *Journal of Quality Technology*, 54(3): 269–289, 2022.

Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. 2008.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of Statistics*, 46(2):814, 2018.

Ghania Fatima, Prabhu Babu, and Petre Stoica. Covariance matrix estimation under positivity constraints with application to portfolio selection. *IEEE Signal Processing Letters*, 29:2487–2491, 2022.

Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22:89–103, 1971.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

David R. Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1): 30–37, 2004.

Pascal Jezequel, Jean-Sebastien Frenel, Loïc Campion, Catherine Guerin-Charbonnel, Wilfried Gouraud, Gabriel Ricolleau, and Mario Campone. bc-genexminer 3.0: new mining module computes breast cancer gene expression correlation analyses. *Database*, 2013:bas060, 2013.

Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

Tracy Ke, Jiashun Jin, and Jianqing Fan. Covariance assisted screening and estimation. *Annals of Statistics*, 42(6):2202, 2014.

Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

Charles Kemp and Joshua B Tenenbaum. Structured statistical models of inductive reasoning. *Psychological Review*, 116(1):20, 2009.

Brenden Lake and Joshua Tenenbaum. Discovering structure by learning sparse graphs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254, 2009.

Steffen Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. *The Annals of Statistics*, 47(4):pp. 1835–1863, 2019.

Han Liu, Lie Wang, and Tuo Zhao. Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, 23(2): 439–459, 2014.

Li Liu, Douglas M Hawkins, Sujoy Ghosh, and S Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172, 2003.

Bertram F Malle and Leonard M Horowitz. The puzzle of negative self-views: An exploration using the schema concept. *Journal of Personality and Social Psychology*, 68(3):470, 1995.

Harry Markowitz. Modern portfolio theory. *Journal of Finance*, 7(11):77–91, 1952.

Harry Markowitz and Peter Todd. *Mean-variance analysis in portfolio choice and capital markets*, volume 66. John Wiley & Sons, 2000.

Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of Statistics*, 6:2125, 2012.

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

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. 2008.

Duy Nhat Phan, Hoai An Le Thi, and Tao Pham Dinh. Sparse covariance matrix estimation by DCA-based algorithms. *Neural computation*, 29(11): 3040–3077, 2017.

Mohsen Pourahmadi. *High-dimensional Covariance Estimation: With High-Dimensional Data*, volume 882. John Wiley & Sons, 2013.

Benjamin Rolfs, Bala Rajaratnam, Dominique Guillot, Ian Wong, and Arian Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. *Advances in Neural Information Processing Systems*, 25, 2012.

Adam J. Rothman. Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740, 2012.

Adam J. Rothman, Elizaveta Levina, and ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494 – 515, 2008.

Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

S. Karlin and Y. Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, 52:419–438, 1983.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

Martin Slawski and Matthias Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015.

Jake A. Soloff, Adityanand Guntuboyina, and Michael I. Jordan. Covariance estimation with nonnegative partial correlations. *arXiv preprint arXiv:2007.15252*, 2020.

Qiang Sun, Kean Ming Tan, Han Liu, and Tong Zhang. Graphical nonconvex optimization via an adaptive convex relaxation. In *International Conference on Machine Learning*, pages 4810–4817. PMLR, 2018.

Ying Sun, Prabhu Babu, and Daniel P. Palomar. Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions. *IEEE Transactions on Signal Processing*, 63(12):3096–3109, 2015.

Ying Sun, Prabhu Babu, and Daniel P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3): 794–816, 2016.

Tiejun Tong, Cheng Wang, and Yuedong Wang. Estimation of variances and covariances for high-dimensional data: a selective review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(4): 255–264, 2014.

Quan Wei and Ziping Zhao. Large covariance matrix estimation with oracle statistical rate via majorization-minimization. *IEEE Transactions on Signal Processing*, 71:3328–3342, 2023.

Xiaohan Wei and Stanislav Minsker. Estimation of the covariance structure of heavy-tailed distributions. *Advances in neural information processing systems*, 30, 2017.

Wenfu Xia, Ziping Zhao, and Ying Sun. C-ISTA: Iterative shrinkage-thresholding algorithm for sparse covariance matrix estimation. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pages 215–219. IEEE, 2023.

Wenfu Xia, Ziping Zhao, and Ying Sun. Distributed sparse covariance matrix estimation. In *2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE, 2024.

Lingzhou Xue, Shiqian Ma, and Hui Zou. Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491, 2012.

Jiaxi Ying, José Vinícius De Miranda Cardoso, and Daniel P. Palomar. Adaptive estimation of graphical models under total positivity. In *International Conference on Machine Learning*, pages 40054–40074. PMLR, 2023.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94 (1):19–35, 2007.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.

Ziping Zhao and Daniel P. Palomar. Mean-reverting portfolio with budget constraint. *IEEE Transactions on Signal Processing*, 66(9):2342–2357, 2018.

Ziping Zhao, Rui Zhou, and Daniel P. Palomar. Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance. *IEEE Transactions on Signal Processing*, 67(7):1681–1695, 2019.

Rui Zhou, Jiaxi Ying, and Daniel P. Palomar. Covariance matrix estimation under low-rank factor model with nonnegative correlations. *IEEE Transactions on Signal Processing*, 70:4020–4030, 2022.

Shuheng Zhou, Philipp Rütimann, Min Xu, and Peter Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research*, 12:2975–3026, 2011.

Shanshan Zou and Ziping Zhao. Large covariance matrix estimation based on factor models via nonconvex optimization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9656–9660. IEEE, 2024.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials to Large Covariance Matrix Estimation with Nonnegative Correlations

We present further experimental results in Section A in this supplementary materials. Following that, we provide the proof of Theorem 1 in Section B. Lastly, we establish the statistical properties of the estimator, as outlined in Theorem 2 and Corollaries 3-4, in Section C.

A ADDITIONAL EXPERIMENTAL RESULTS

In the main part of the paper, we compared the estimation performance in the context of a banded matrix. Here, we present additional experiments in Tables 2 and 3 to evaluate estimation performance across various methods for both the block and Toeplitz settings.[4] Our proposed method, PDTE_NN_FC, achieves the lowest estimation error in both the Frobenius and spectral norms, consistent with the results in the banded matrix setting.

Table 2: Quantitative comparison among different methods for the block matrix structure.

	SCM	PDTE_L1	PDTE_FC	MLE_L1	MLE_FC	MLE_MTP2	MLE_MTP2_L1	MLE_MTP2_FC	MLE_NN	PDTE_NN (prop.)	PDTE_NN_FC (prop.)
						$d = 100, n = 50$					
$\|\cdot\|_F$	14.3226 (0.0652)	8.4551 (0.0586)	7.8275 (0.0713)	8.7543 (0.1567)	8.1234 (0.0712)	23.5956 (0.0024)	18.1121 (0.0037)	17.7123 (0.0014)	8.3456 (0.0653)	10.5321 (0.0529)	**7.1571 (0.0652)**
$\|\cdot\|_2$	7.4117 (0.0585)	3.4647 (0.0661)	3.4112 (0.0508)	3.5648 (0.0686)	3.4321 (0.0509)	20.0329 (0.0016)	13.7859 (0.026)	12.2766 (0.0019)	3.4567 (0.0586)	4.2661 (0.0502)	**3.3321 (0.0585)**
FPR	NA	0.1288 (0.0018)	0.0389 (0.0021)	0.1329 (0.0051)	0.0387 (0.0022)	0.9999 (0.0001)	0.9999 (0.0001)	0.9999 (0.0001)	0.3327 (0.0053)	0.5037 (0.0028)	**0.0269 (0.0022)**
TPR	NA	0.9846 (0.0132)	0.9861 (0.0005)	0.9647 (0.0158)	0.9862 (0.0006)	**0.9999 (0.0001)**	**0.9999 (0.0001)**	**0.9999 (0.0001)**	0.9649 (0.0160)	0.9998 (0.0007)	0.9992 (0.0001)
Time	–	3.09	8.89	28.63	39.12	34.85	0.78	8.86	58.78	**0.92**	0.97
						$d = 200, n = 50$					
$\|\cdot\|_F$	28.6502 (0.1564)	17.7789 (0.0451)	15.1874 (0.1601)	17.8794 (0.1958)	15.1234 (0.1602)	33.9022 (0.1565)	26.8044 (0.0452)	23.2931 (0.1603)	19.3456 (0.0566)	22.3415 (0.0291)	**13.7436 (0.1563)**
$\|\cdot\|_2$	13.5083 (0.1390)	7.2523 (0.0818)	6.8697 (0.1237)	7.6431 (0.1031)	6.8321 (0.1238)	31.5678 (0.1391)	17.4321 (0.0819)	16.8123 (0.1239)	7.4567 (0.0392)	9.9059 (0.0746)	**6.5083 (0.1389)**
FPR	NA	0.1614 (0.0014)	0.1171 (0.0021)	0.1847 (0.0019)	0.1172 (0.0022)	0.9999 (0.0001)	0.9999 (0.0001)	0.9999 (0.0001)	0.2849 (0.0021)	0.5469 (0.0023)	**0.0215 (0.0023)**
TPR	NA	0.9967 (0.0127)	0.9969 (0.0004)	0.9891 (0.0030)	0.9970 (0.0005)	**0.9999 (0.0001)**	**0.9999 (0.0001)**	**0.9999 (0.0001)**	0.9893 (0.0022)	0.9996 (0.0001)	0.9994 (0.0002)
Time	–	11.06	24.12	110.39	224.56	988.74	39.41	180.16	1010.67	**7.89**	13.37

Table 3: Quantitative comparison among different methods for the Toeplitz matrix structure.

	SCM	PDTE_L1	PDTE_FC	MLE_L1	MLE_FC	MLE_MTP2	MLE_MTP2_L1	MLE_MTP2_FC	MLE_NN	PDTE_NN (prop.)	PDTE_NN_FC (prop.)
						$d = 100, n = 50$					
$\|\cdot\|_F$	14.1171 (0.0630)	8.9433 (0.0556)	8.6599 (0.0427)	9.7471 (0.1044)	8.1234 (0.0428)	20.5959 (0.0631)	15.4053 (0.0557)	14.5655 (0.0428)	9.3456 (0.0632)	10.1081 (0.0608)	**8.2385 (0.0569)**
$\|\cdot\|_2$	6.2553 (0.0851)	3.7761 (0.0693)	3.6271 (0.0382)	3.8367 (0.0390)	3.4321 (0.0383)	19.4093 (0.0852)	14.0043 (0.0694)	13.1192 (0.0384)	4.1267 (0.0853)	4.3249 (0.0803)	**3.4224 (0.0541)**
Time	–	10.34	34.97	49.84	94.56	58.46	2.50	13.78	89.67	**0.68**	8.19
						$d = 200, n = 50$					
$\|\cdot\|_F$	28.2106 (0.0928)	14.0822 (0.0455)	13.4672 (0.0606)	15.3108 (0.1923)	13.1234 (0.0607)	33.5678 (0.0929)	24.4321 (0.0456)	23.4123 (0.0608)	14.3456 (0.0930)	19.0706 (0.0434)	**13.1233 (0.0654)**
$\|\cdot\|_2$	10.6794 (0.0965)	4.2644 (0.0342)	4.1737 (0.0543)	4.3073 (0.0371)	4.1123 (0.0544)	32.5678 (0.0966)	18.0321 (0.0343)	16.4123 (0.0545)	4.4567 (0.0967)	8.9058 (0.0306)	**4.1129 (0.0794)**
Time	–	37.0484	44.12	121.86	44.56	534.23	12.12	49.78	121.67	**7.39**	36.07

Figure 3 presents a correlation graph representation of the estimation performance of various methods in the block matrix setting. Similar to the banded case, the proposed method, PDTE_NN_FC, achieves better estimation performance than the others.

[4] FPR and TPR for the Toeplitz matrix structure are not reported because the matrix is not sparse.

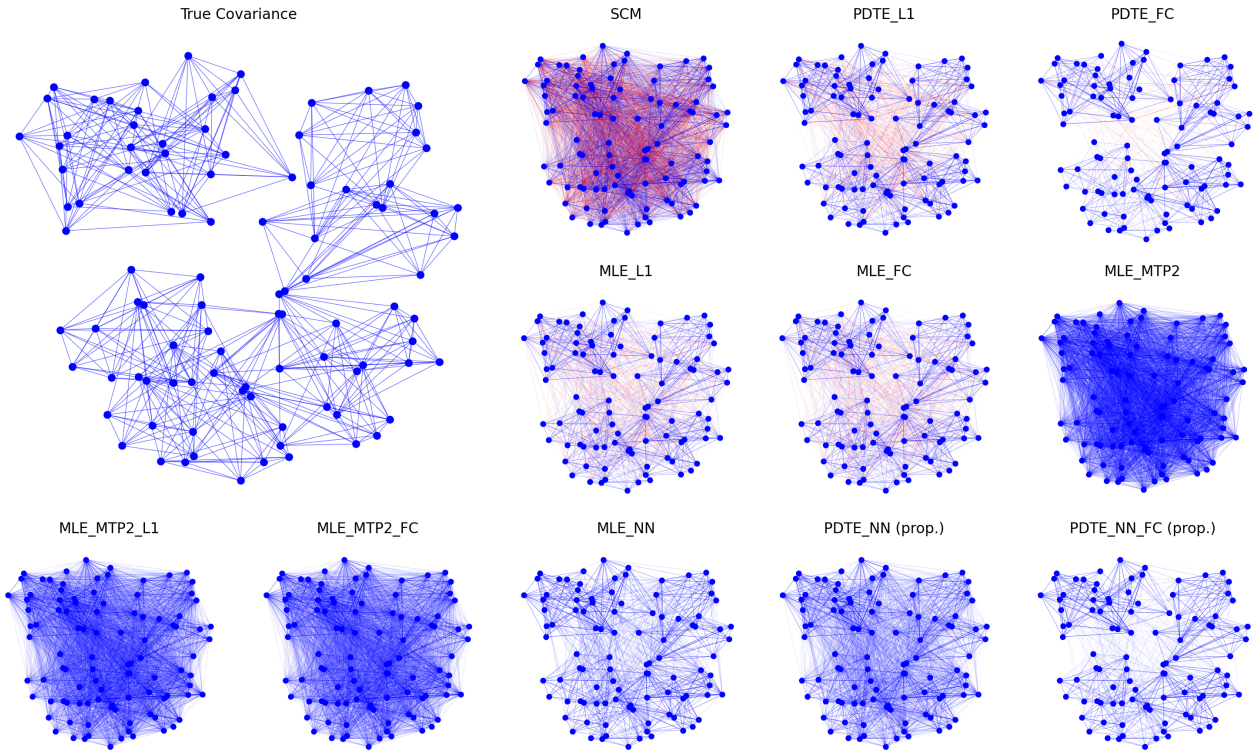

Figure 3: Correlation graph for the block matrix structure with $d = 100$ and $n = 50$.

B PROOF OF THEOREM 1

In this section, we prove the convergence result of Algorithm 2.

Define $\mathcal{C} = \left\{\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d} \mid \boldsymbol{\Sigma} \succ \mathbf{0}, \boldsymbol{\Sigma} \geq \mathbf{0}\right\}$. Given $\boldsymbol{\Sigma}_0 \in \mathcal{C}$,[5] we define a sublevel set of the objective function f_k in problem (2) as follows:

$$\mathcal{B}_k = \left\{\boldsymbol{\Sigma} \in \mathcal{C} \mid f_k\left(\boldsymbol{\Sigma}\right) \leq f_k\left(\boldsymbol{\Sigma}_0\right)\right\}.$$

For the set of $\mathcal{C}$, we can express it by the intersection of a closed set $\mathcal{N}$, with $\mathcal{N}$ the set of nonnegative matrices, and $\mathcal{P}$, with $\mathcal{P}$ the set of positive definite matrices.

Lemma B.1. *For all $\boldsymbol{\Sigma}$ in $\mathcal{B}_k$, there exists $m > 0$ such that $\boldsymbol{\Sigma} \succeq m\boldsymbol{I}$.*

Proof. Because of $\log\det\left(\boldsymbol{\Sigma}\right) = \sum_{i=1}^{d}\lambda_i$, the term $-\tau\log\det\left(\boldsymbol{\Sigma}\right)$ in f_k dominates the objective as $\lambda_{\min}\left(\boldsymbol{\Sigma}\right)$ tends to 0. Meanwhile, the term $\frac{1}{2}\left\|\boldsymbol{\Sigma} - \boldsymbol{S} + \boldsymbol{\Lambda}^k\right\|_{\mathrm{F}}^2$ remains bounded below owing to its nonnegativity. Hence, $f_k\left(\boldsymbol{\Sigma}\right)$ tends to positive infinity as $\lambda_{\min}\left(\boldsymbol{\Sigma}\right)$ tends to 0. So, there must exist $m > 0$ such that $\boldsymbol{\Sigma} \succeq m\boldsymbol{I}$ for all $\boldsymbol{\Sigma}$ in $\mathcal{B}_k$. □

We define the proximal step (4) in Algorithm 2 as follows:

$$\boldsymbol{\Sigma}_t\left(\phi_t\right) = \Pi_{\mathcal{N}}\left(\boldsymbol{\Sigma}_t - \phi_t^{-1}\nabla f_k\left(\boldsymbol{\Sigma}_t\right)\right).$$

At the t-th iteration, given $\boldsymbol{\Sigma}_t \in \mathcal{B}_k$, there must exist $\sigma > 0$ such that for any $\phi_t^{-1} \in (0, \sigma)$, $\boldsymbol{\Sigma}_t\left(\phi_t\right)$ satisfies $\boldsymbol{\Sigma}_t\left(\phi_t\right) \in \mathcal{P}$. This is due to $\boldsymbol{\Sigma}_t$ is an interior point of $\mathcal{P}$, indicating that there exists $\omega > 0$, such that for any $\boldsymbol{\Sigma} \in \mathcal{P}$, it satisfies $\left\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_t\right\|_{\mathrm{F}} < \omega$.

Let $\phi_t^{-1} < \frac{\omega}{\left\|\nabla f_k\left(\boldsymbol{\Sigma}_t\right)\right\|_{\mathrm{F}}}$, then

$$\left\|\boldsymbol{\Sigma}_t\left(\phi_t\right) - \boldsymbol{\Sigma}_t\right\|_{\mathrm{F}} = \left\|\Pi_{\mathcal{N}}\left(\boldsymbol{\Sigma}_t - \phi_t^{-1}\nabla f_k\left(\boldsymbol{\Sigma}_t\right)\right) - \Pi_{\mathcal{N}}\left(\boldsymbol{\Sigma}_t\right)\right\|_{\mathrm{F}} \leq \phi_t^{-1}\left\|\nabla f_k\left(\boldsymbol{\Sigma}_t\right)\right\|_{\mathrm{F}} < \omega.$$

[5] For simplicity, we omit the superscript k from $\boldsymbol{\Sigma}_t^k$ in this section.

establishing that $\boldsymbol{\Sigma}_t(\phi_t) \in \mathcal{P}$. The projection $\Pi_{\mathcal{N}}$ ensures $\boldsymbol{\Sigma}_t(\phi_t) \in \mathcal{N}$. As a result, $\boldsymbol{\Sigma}_t(\phi_t) \succ \mathbf{0}$.

The backtracking line search ensures that $f_k(\boldsymbol{\Sigma}_t(\phi_t)) \leq f_k(\boldsymbol{\Sigma}_t)$, implying that $\boldsymbol{\Sigma}_t(\phi_t) \in \mathcal{B}_k$. The gradient of f_k is Lipschitz continuous with parameter $L = 1 + \tau m^{-2}$ over $\mathcal{B}_k$. Since its Hessian matrix is given by $\boldsymbol{I} \otimes \boldsymbol{I} + \tau \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}$, we derive

$$f_k(\boldsymbol{\Sigma}_t(\phi_t)) \leq f_k(\boldsymbol{\Sigma}_t) + \langle \nabla f_k(\boldsymbol{\Sigma}_t), \boldsymbol{\Sigma}_t(\phi_t) - \boldsymbol{\Sigma}_t \rangle + \frac{L}{2} \|\boldsymbol{\Sigma}_t(\phi_t) - \boldsymbol{\Sigma}_t\|_{\mathrm{F}}^2. \tag{6}$$

According to the projection theorem, it follows that

$$\left\langle \boldsymbol{\Sigma}_t - \phi_t^{-1} \nabla f_k(\boldsymbol{\Sigma}_t) - \boldsymbol{\Sigma}_t(\phi_t), \boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t(\phi_t) \right\rangle \leq 0. \tag{7}$$

Combining with (6), we have

$$f_k(\boldsymbol{\Sigma}_t(\phi_t)) \leq f_k(\boldsymbol{\Sigma}_t) + \left(\frac{L}{2} - \phi_t\right) \|\boldsymbol{\Sigma}_t(\phi_t) - \boldsymbol{\Sigma}_t\|_{\mathrm{F}}^2. \tag{8}$$

Given $\phi_t \geq \frac{L}{3}$, it leads to $\left(\frac{L}{2} - \phi_t\right) \leq \frac{\phi_t}{2}$. For any $\phi_t \geq \max\left(\frac{L}{3}, \frac{1}{\sigma}\right)$, $\boldsymbol{\Sigma}_t(\phi_t)$ can simultaneously satisfy $\boldsymbol{\Sigma}_t(\phi_t) \succ \mathbf{0}$ and the backtracking line search condition. Then $\boldsymbol{\Sigma}_t(\phi_t) = \boldsymbol{\Sigma}_{t+1}$. Thus the step size in line search has a lower bound $\phi_t^{-1} \geq \min\left(\frac{3}{L\gamma}, \phi_0, \sigma\right)$. We set $\boldsymbol{\Sigma}_t(\phi_t)$ satisfying this condition as $\boldsymbol{\Sigma}_{t+1}$.

By induction, $\boldsymbol{\Sigma}_t \in \mathcal{B}_k$ for any $t \geq 0$. Sequence $\{f_k(\boldsymbol{\Sigma}_t)\}$ is monotonically nonincreasing, and $f_k(\boldsymbol{\Sigma}_{t+1}) \leq f_k(\boldsymbol{\Sigma}_t)$ until $\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t$, indicating that $\boldsymbol{\Sigma}_{t+1}$ is a stationary point. The stationary point is the unique minimizer since problem (2) is a strictly convex problem. Therefore, the proposed Algorithm 2 converges to the optimal solution.

C PROOF OF STATISTICAL THEORY

We begin by introducing some notations. Recall $\mathcal{S}$ is the support set of the true covariance matrix $\boldsymbol{\Sigma}^\star$. We further define two sets $\mathcal{T}^\star$ and $\mathcal{I}$ relative to $\boldsymbol{\Sigma}^\star$, where

$$\mathcal{T}^\star = \left\{(i,j) \mid \Sigma_{ij}^\star = 0, i \neq j\right\}, \tag{9}$$

and

$$\mathcal{I} = \{(i,i) \mid i \in \{1, 2, \ldots, d\}\}. \tag{10}$$

Based on Assumption 1, we define the set

$$\mathcal{L}_\alpha(\boldsymbol{\Sigma}) = \{(i,j) \mid |\Sigma_{ij}| \geq \alpha\lambda\}. \tag{11}$$

We define $\mathcal{T}_\alpha(\boldsymbol{\Sigma}) = \mathcal{L}_\alpha(\boldsymbol{\Sigma}) \cup \mathcal{S}$, $\bar{\mathcal{T}}_\alpha(\boldsymbol{\Sigma}) = \mathcal{T}_\alpha(\boldsymbol{\Sigma}) \cup \mathcal{I}$, and $\bar{\mathcal{T}}_\alpha^c(\boldsymbol{\Sigma}) = \mathcal{T}_\alpha^c(\boldsymbol{\Sigma}) \setminus \mathcal{I}$.

C.1 Proof of Lemma C.1

Lemma C.1. *Let $g(\boldsymbol{\Sigma}) = \frac{1}{2}||\boldsymbol{\Sigma} - \boldsymbol{S}||_{\mathrm{F}}^2 - \tau \log\det(\boldsymbol{\Sigma})$. Then we have the following inequality:*

$$\langle \nabla g(\boldsymbol{\Sigma}_2) - \nabla g(\boldsymbol{\Sigma}_1), \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 \rangle \geq ||\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1||_{\mathrm{F}}^2. \tag{12}$$

Proof. By mean value theorem, there exist $\eta \in [0,1]$, such that

$$g(\boldsymbol{\Sigma}_2) = g(\boldsymbol{\Sigma}_1) + \langle \nabla g(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1) \rangle + \frac{1}{2} \mathrm{vec}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^\top \nabla^2 g(\boldsymbol{\Sigma}_t) \mathrm{vec}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1), \tag{13}$$

where $\boldsymbol{\Sigma}_t = \eta\boldsymbol{\Sigma}_1 + (1-\eta)\boldsymbol{\Sigma}_2$. Note that $\nabla^2 g(\boldsymbol{\Sigma}) = \boldsymbol{I} \otimes \boldsymbol{I} + \tau \boldsymbol{\Sigma}_t^{-1} \otimes \boldsymbol{\Sigma}_t^{-1}$, one has

$$\lambda_{\min}\left(\nabla^2 g(\boldsymbol{\Sigma})\right) = \lambda\left(\boldsymbol{I} \otimes \boldsymbol{I} + \tau \boldsymbol{\Sigma}_t^{-1} \otimes \boldsymbol{\Sigma}_t^{-1}\right) \geq 1. \tag{14}$$

According to

$$\frac{1}{2}\mathrm{vec}\left(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right)^{\top}\nabla^2 g\left(\boldsymbol{\Sigma}_t\right)\mathrm{vec}\left(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right)\geq\frac{1}{2}\lambda_{\min}\left(\nabla^2 g\left(\boldsymbol{\Sigma}\right)\right)\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right\|_{\mathrm{F}}^2,$$

we have

$$g\left(\boldsymbol{\Sigma}_2\right)\geq g\left(\boldsymbol{\Sigma}_1\right)+\left\langle\nabla g\left(\boldsymbol{\Sigma}_1\right),\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right\rangle+\frac{1}{2}\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right\|_{\mathrm{F}}^2, \tag{15}$$

and

$$g\left(\boldsymbol{\Sigma}_1\right)\geq g\left(\boldsymbol{\Sigma}_2\right)+\left\langle\nabla g\left(\boldsymbol{\Sigma}_2\right),\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\right\rangle+\frac{1}{2}\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right\|_{\mathrm{F}}^2. \tag{16}$$

Combining (15) and (16), we obtain

$$\left\langle\nabla g\left(\boldsymbol{\Sigma}_2\right)-\nabla g\left(\boldsymbol{\Sigma}_1\right),\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right\rangle\geq\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1\right\|_{\mathrm{F}}^2. \tag{17}$$

□

C.2 Proof of Lemma C.2

Lemma C.2. *Suppose the event* $\left\|\boldsymbol{\Sigma}^{\star}-\boldsymbol{S}\right\|_{\max}\leq\frac{\sqrt{2}}{2}\lambda$ *holds, with* $\alpha=3$*, the parameter* $\tau\leq\sqrt{\frac{s}{n}}\left\|\left(\boldsymbol{\Sigma}^{\star}\right)^{-1}\right\|_{\mathrm{F}}^{-1}$*, and* $\lambda\asymp\sqrt{\frac{\log d}{n}}$*. Under Assumptions 1 and 2, if* $\mathrm{card}\left(\mathcal{T}_{\alpha}\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)\leq 2s-d$ *holds for some* $\widetilde{\boldsymbol{\Sigma}}$*, then*

$$\left\|G_{\lambda,\boldsymbol{S}}\left(\widetilde{\boldsymbol{\Sigma}}\right)-\boldsymbol{\Sigma}^{\star}\right\|_{\mathrm{F}}\leq\left\|p_{\lambda}'\left(\widetilde{\boldsymbol{\Sigma}}_{\mathcal{S}}\right)\right\|_{\mathrm{F}}+\left\|\left(\boldsymbol{\Sigma}^{\star}-\boldsymbol{S}\right)_{\overline{\mathcal{T}}_{\alpha}\left(\widetilde{\boldsymbol{\Sigma}}\right)}\right\|_{\mathrm{F}}, \tag{18}$$

and

$$\mathrm{card}\left(\mathcal{T}_{\alpha}\left(G_{\lambda,\boldsymbol{S}}\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)\right)\leq 2s-d \tag{19}$$

where

$$G_{\lambda,S}\left(\widetilde{\boldsymbol{\Sigma}}\right)=\arg\min_{\boldsymbol{\Sigma}'\succ\mathbf{0},\boldsymbol{\Sigma}'\geq\mathbf{0}}\frac{1}{2}\left\|\boldsymbol{\Sigma}'-\boldsymbol{S}\right\|_{\mathrm{F}}^2-\tau\log\det\boldsymbol{\Sigma}'+\sum_{i\neq j}p_{\lambda}'\left(\widetilde{\Sigma}_{ij}\right)\Sigma_{ij}'.$$

Proof. For ease of presentation, we denote $G_{\lambda,\boldsymbol{S}}\left(\widetilde{\boldsymbol{\Sigma}}\right)$ by $\boldsymbol{\Sigma}$. Applying Lemma C.1, and let $\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}^{\star}$, then one has

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\|_{\mathrm{F}}^2\leq\left\langle\boldsymbol{\Sigma}-\tau\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{\star}+\tau\left(\boldsymbol{\Sigma}^{\star}\right)^{-1},\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\rangle. \tag{20}$$

Recall the formulation of the problem (2), then we use its Lagrangian function

$$\mathcal{L}\left(\boldsymbol{\Sigma}',\boldsymbol{\Gamma}'\right)=\frac{1}{2}\left\|\boldsymbol{\Sigma}'-\boldsymbol{S}\right\|_{\mathrm{F}}^2-\tau\log\det\left(\boldsymbol{\Sigma}'\right)+\sum_{i\neq j}\Lambda_{ij}\Sigma_{ij}'-\left\langle\boldsymbol{\Gamma}',\boldsymbol{\Sigma}'\right\rangle, \tag{21}$$

where $\boldsymbol{\Gamma}$ is a Karush-Kuhn-Tucker (KKT) multiplier with $\Gamma_{ii}=0$ for $i\in\{1,2,\ldots,d\}$. Let $(\boldsymbol{\Sigma},\boldsymbol{\Gamma})$ be the primal and dual optimal point, then the KKT conditions are as follows:

$$\boldsymbol{\Sigma}-\boldsymbol{S}-\tau\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Lambda}-\boldsymbol{\Gamma}=\mathbf{0}, \tag{22}$$

$$\Sigma_{ij}\Gamma_{ij}=0,\quad\Sigma_{ij}\geq 0,\quad\Gamma_{ij}\geq 0,\ \forall i\neq j, \tag{23}$$

$$\Gamma_{ii}=0,\quad\Lambda_{ij}\geq 0, \tag{24}$$

By the equation (22), we are able to get $\boldsymbol{\Sigma}-\tau\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Gamma}-\boldsymbol{\Lambda}+\boldsymbol{S}$, then (20) can be rewritten as

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\|_{\mathrm{F}}^2\leq\underbrace{\left\langle\boldsymbol{\Gamma},\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\rangle}_{\text{term I}}+\underbrace{\left\langle\boldsymbol{S}-\boldsymbol{\Sigma}^{\star},\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\rangle}_{\text{term II}}+\underbrace{\left\langle-\boldsymbol{\Lambda},\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\rangle}_{\text{term III}}+\underbrace{\tau\left\langle\left(\boldsymbol{\Sigma}^{\star}\right)^{-1},\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^{\star}\right\rangle}_{\text{term IV}}. \tag{25}$$

The term I can be bounded by

$$\langle \boldsymbol{\Gamma}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star \rangle = -\sum_{i \neq j} \Gamma_{ij} \Sigma_{ij}^\star \leq 0, \tag{26}$$

where the equality follows from (23), and the inequality follows from the $\Gamma_{ij} \geq 0$ and $\Sigma_{ij}^\star \geq 0$, $\forall i \neq j$.

Consider the event $\mathcal{J}$

$$\|\boldsymbol{\Sigma}^\star - \boldsymbol{S}\|_{\max} \leq \frac{\sqrt{2}}{2}\lambda. \tag{27}$$

For term II, we separate its support set into parts $\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)$ and $\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)$. Then one has

$$\begin{aligned}
\langle \boldsymbol{S} - \boldsymbol{\Sigma}^\star, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star \rangle &= \left\langle (\boldsymbol{S} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle + \left\langle (\boldsymbol{S} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle \\
&\leq \left\langle (\boldsymbol{S} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle + \left\| (\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\mathrm{F}} \left\| (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\mathrm{F}}.
\end{aligned} \tag{28}$$

In order to bound term III, separating the support set of it into $\mathcal{S}$ and $\mathcal{T}^\star$, thus we derive

$$\begin{aligned}
\langle -\boldsymbol{\Lambda}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star \rangle &\leq \langle -\boldsymbol{\Lambda}_{\mathcal{S}}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\mathcal{S}} \rangle + \langle -\boldsymbol{\Lambda}_{\mathcal{T}^\star}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\mathcal{T}^\star} \rangle \\
&\leq \|\boldsymbol{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} \|(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\mathcal{S}}\|_{\mathrm{F}} + \langle -\boldsymbol{\Lambda}_{\mathcal{T}^\star}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\mathcal{T}^\star} \rangle \\
&\leq \|\boldsymbol{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} \|(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\mathcal{S}}\|_{\mathrm{F}} + \left\langle -\boldsymbol{\Lambda}_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle,
\end{aligned} \tag{29}$$

where the second inequality can be acquired by the definition of $\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)$ and $\mathcal{T}^\star$, we can get $\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right) \subseteq \mathcal{T}^\star$ and combine with $\Lambda_{ij} \geq 0$.

For the sake of bounding term IV, one has

$$\tau \left\langle (\boldsymbol{\Sigma}^\star)^{-1}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star \right\rangle \leq \tau \left\| (\boldsymbol{\Sigma}^\star)^{-1} \right\|_{\mathrm{F}} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}. \tag{30}$$

Noting that for any $(i,j) \in \bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)$, $\Sigma_{ij} \geq 0$, and $\Lambda_{ij} \geq \frac{\sqrt{2}}{2}\lambda$ according to Assumption 1. Combining it with the event $\mathcal{J}$ mentioned above, we can secure

$$\begin{aligned}
&\left\langle (\boldsymbol{S} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle + \left\langle -\boldsymbol{\Lambda}_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle \\
=&\left\langle (\boldsymbol{S} - \boldsymbol{\Sigma}^\star - \boldsymbol{\Lambda})_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)}, (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha^c\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\rangle \leq 0.
\end{aligned} \tag{31}$$

By bounding the four different terms in (25), we obtain the following result

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 \leq \left\| (\boldsymbol{S} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\mathrm{F}} \left\| (\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} \|(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star)_{\mathcal{S}}\|_{\mathrm{F}} + \tau \left\| (\boldsymbol{\Sigma}^\star)^{-1} \right\|_{\mathrm{F}} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}} \quad . \tag{32}$$

The equality can both be divided by $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}$, then we have

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}} \leq \left\| (\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} + \tau \left\| (\boldsymbol{\Sigma}^\star)^{-1} \right\|_{\mathrm{F}}. \tag{33}$$

Due to the event $\mathcal{J}$, we know

$$\begin{aligned}
\left\| (\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\mathrm{F}} &\leq \sqrt{\mathrm{card}\left(\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)} \left\| (\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)} \right\|_{\max} \\
&\leq \sqrt{\mathrm{card}\left(\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)} \frac{\sqrt{2}}{2}\lambda \\
&\leq \sqrt{s}\lambda,
\end{aligned} \tag{34}$$

where the last inequality is from

$$\operatorname{card}\left(\bar{\mathcal{T}}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)=\operatorname{card}\left(\mathcal{T}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)\cup\mathcal{I}\right)\leq\operatorname{card}\left(\mathcal{T}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)+\operatorname{card}\left(\mathcal{I}\right)\leq 2s-d+d=2s. \tag{35}$$

Since the condition $\operatorname{card}\left(\mathcal{T}_\alpha\left(\widetilde{\boldsymbol{\Sigma}}\right)\right)\leq 2s-d$ is given in Lemma C.2.

According to Assumption 1, it is known that $\Lambda_{ij}=p'_\lambda\left(\widetilde{\Sigma}_{ij}\right)\leq\lambda,\forall i\neq j$, together with $\operatorname{card}\left(\mathcal{S}\right)\leq s-d\leq s$. Thus one has

$$\left\|\boldsymbol{\Lambda}_{\mathcal{S}}\right\|_{\mathrm{F}}\leq\sqrt{\operatorname{card}\left(\mathcal{S}\right)}\lambda\leq\sqrt{s}\lambda. \tag{36}$$

Recall it that $\tau\leq\sqrt{\frac{s}{n}}\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}^{-1}$ and $\lambda\asymp\sqrt{\frac{\log d}{n}}$, then we take

$$\tau\leq\sqrt{\frac{s}{n}}\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}^{-1}\leq\lambda\sqrt{s}\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}^{-1}.$$

Combining the results above and substituting equations (34) and (36) into equation (33), we obtain

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}\leq 3\sqrt{s}\lambda. \tag{37}$$

Putting $\boldsymbol{\Lambda}_{\mathcal{S}}=p'_\lambda\left(\widetilde{\boldsymbol{\Sigma}}_{\mathcal{S}}\right)$ and $\boldsymbol{\Sigma}=G_{\lambda,S}\left(\widetilde{\boldsymbol{\Sigma}}\right)$ into (33), we obtain

$$\left\|G_{\lambda,\boldsymbol{S}}\left(\widetilde{\boldsymbol{\Sigma}}\right)-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}\leq\left\|\left(\boldsymbol{\Sigma}^\star-\boldsymbol{S}\right)_{\bar{\mathcal{T}}_\alpha\left(\tilde{\boldsymbol{\Sigma}}\right)}\right\|_{\mathrm{F}}+\left\|p'_\lambda\left(\widetilde{\boldsymbol{\Sigma}}_{\mathcal{S}}\right)\right\|_{\mathrm{F}}+\tau\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}. \tag{38}$$

In next step, we need to prove $\operatorname{card}\left(\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}\right)\right)\leq 2s-d$. We separate the set $\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}\right)$ from $\mathcal{S}$ and $\mathcal{L}_\alpha\left(\boldsymbol{\Sigma}\right)\setminus\mathcal{S}$. For any $(i,j)\in\mathcal{L}_\alpha\left(\boldsymbol{\Sigma}\right)\setminus\mathcal{S}$, one has $|\Sigma_{ij}|\geq\alpha\lambda$, then one further obtains

$$\begin{aligned}\operatorname{card}\left(\mathcal{L}_\alpha\left(\boldsymbol{\Sigma}\right)\setminus\mathcal{S}\right)&\leq\frac{\left\|\boldsymbol{\Sigma}_{\mathcal{L}_\alpha(\boldsymbol{\Sigma})\setminus\mathcal{S}}\right\|_{\mathrm{F}}^2}{\left(\alpha\lambda\right)^2}\\&\leq\frac{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2}{\left(\alpha\lambda\right)^2}\\&\leq s.\end{aligned} \tag{39}$$

Here, the last inequality follows from (37) and $\alpha=3$, which implies that the above condition is established. Then we obtain

$$\begin{aligned}\operatorname{card}\left(\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}\right)\right)&=\operatorname{card}\left\{\left(\mathcal{L}_\alpha\left(\boldsymbol{\Sigma}\right)\setminus\mathcal{S}\right)\cup\mathcal{S}\right\}\\&\leq\operatorname{card}\left(\mathcal{L}_\alpha\left(\boldsymbol{\Sigma}\right)\setminus\mathcal{S}\right)+\operatorname{card}\left(\mathcal{S}\right)\\&\leq 2s-d,\end{aligned} \tag{40}$$

completing the proof.

□

C.3 Proof of Lemma C.3

Lemma C.3. *(Lemma D.1 in (Sun et al., 2018)) Let $\boldsymbol{x}$ be a zero mean sub-Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}^\star$ and $\{\boldsymbol{x_i}\}_{i=1}^n$ be a collection of i.i.d samples from $\boldsymbol{x}$. There exist some constants c_1,c_2 and t_0 such that for all t with $0<t<t_0$, the sample covariance matrix $\boldsymbol{S}$ satisfies the following tail bound*

$$\mathbb{P}\left(\left|\Sigma_{ij}^\star-S_{ij}\right|>t\right)\leq c_1\exp\left(-c_2nt^2\right). \tag{41}$$

C.4 Proof of Lemma C.4

Lemma C.4. *Under the same condition in Lemma C.3, if taking* $\lambda = \sqrt{\frac{6\log d}{c_2 n}} \asymp \sqrt{\frac{\log d}{n}} < t_0$, *we obtain*

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}^\star - \boldsymbol{S}\|_{\max} \leq \frac{\sqrt{2}}{2}\lambda\right) \geq 1 - \frac{c_1}{d}. \tag{42}$$

Proof. Applying Lemma C.3 and union bound, for any $\frac{\sqrt{2}}{2}\lambda$ such that $0 < \lambda < t_0$, then the following result holds

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{\Sigma}^\star - \boldsymbol{S}\|_{\max} > \frac{\sqrt{2}}{2}\lambda\right) \leq & c_1 d^2 \exp\left(\frac{-c_2 n\lambda^2}{2}\right) \\ = & c_1 \exp\left(\frac{-c_2 n\lambda^2}{2} + 2\log d\right).\end{aligned} \tag{43}$$

For the n is sufficiently large such that $n \geq \frac{4\log d}{c_2 t_0^2}$ and $\lambda = \sqrt{\frac{6\log d}{c_2 n}} \asymp \sqrt{\frac{\log d}{n}} < t_0$, there exists

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}^\star - \boldsymbol{S}\|_{\max} \leq \frac{\sqrt{2}}{2}\lambda\right) \geq 1 - c_1 \exp\left(\frac{-c_2 n\lambda^2}{2} + 2\log d\right) = 1 - \frac{c_1}{d}. \tag{44}$$

□

C.5 Proof of Lemma C.5

Lemma C.5. *Under the conditions that exist in the Lemma C.3, the following result holds*

$$\mathbb{P}\left(\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\|_{\mathrm{F}} \lesssim \sqrt{\frac{s}{n}}\right) \geq 1 - \frac{c_1}{s}. \tag{45}$$

Proof. Applying Lemma C.3 and union bound, for any M such that $0 < M\sqrt{\frac{1}{n}} < t_0$, one has

$$\mathbb{P}\left(\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\|_{\max} > M\sqrt{\frac{1}{n}}\right) \leq c_1 \mathrm{card}\left(\mathcal{S}\right) \exp\left(-c_2 M^2\right) \leq c_1 \exp\left(-c_2 M^2 + \log s\right). \tag{46}$$

where the second inequality is obtained by $\mathrm{card}\left(\mathcal{S}\cup\mathcal{I}\right) \leq s$. By taking M such that $\sqrt{\frac{2\log s}{c_2}} < M < t_0\sqrt{n}$, we have

$$\mathbb{P}\left(\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\|_{\max} \leq M\sqrt{\frac{1}{n}}\right) \geq 1 - c_1 \exp\left(-c_2 M^2 + \log s\right) \geq 1 - \frac{c_1}{s}. \tag{47}$$

Then, applying the following inequality

$$\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\|_{\mathrm{F}} \leq \sqrt{s}\,\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\|_{\max},$$

we can obtain

$$\mathbb{P}\left(\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\|_{\mathrm{F}} \leq M\sqrt{\frac{s}{n}}\right) \geq 1 - \frac{c_1}{s}, \tag{48}$$

completing the proof. □

C.6 Proof of Theorem 2

Given $\boldsymbol{\Sigma}^k = G_{\lambda,\boldsymbol{S}}\left(\boldsymbol{\Sigma}^{k-1}\right)$ for $k \geq 2$ and $p'_\lambda(0) = \lambda$ in Assumption 1, we can write $\boldsymbol{\Sigma}^1 = G_{\lambda,\boldsymbol{S}}\left(\boldsymbol{\Sigma}^0\right)$, where $\Sigma_{ij}^{(0)} = 0$ for any $i \neq j$, and $\operatorname{card}\left(\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}^0\right)\right) \leq 2s - d$. Following from Lemma C.2, we can further obtain $\operatorname{card}\left(\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}^1\right)\right) \leq 2s - d$.

By induction, we have for any $k \geq 1$

$$\operatorname{card}\left(T_\alpha\left(\boldsymbol{\Sigma}^k\right)\right) \leq 2s - d. \tag{49}$$

To simplify notation, we denote $\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}^k\right)$ by $\mathcal{T}_\alpha^k$ and $\bar{\mathcal{T}}_\alpha\left(\boldsymbol{\Sigma}^k\right)$ by $\bar{\mathcal{T}}_\alpha^k$.

Due to (38), we get

$$\left\|\boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq \left\|(\boldsymbol{\Sigma}^\star - \mathrm{S})_{\bar{\mathcal{T}}_\alpha^{k-1}}\right\|_{\mathrm{F}} + \left\|p'_\lambda\left(\boldsymbol{\Sigma}_{\mathcal{S}}^{(k-1)}\right)\right\|_{\mathrm{F}} + \tau\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}} \quad . \tag{50}$$

Let $\rho_1^{k-1} = \left\|p'_\lambda\left(\boldsymbol{\Sigma}_{\mathcal{S}}^{k-1}\right)\right\|_{\mathrm{F}}$, $\rho_2^{k-1} = \left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\bar{\mathcal{T}}_\alpha^{k-1}}\right\|_{\mathrm{F}} + \tau\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}}$ then one has:

$$\left\|\boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq \rho_1^{k-1} + \rho_2^{k-1}, \tag{51}$$

The term ρ_1^{k-1} can be bounded by $\left\|\boldsymbol{\Sigma}^{k-1} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}$, for any $(i,j) \in \mathcal{S}$, if $\left|\Sigma_{ij}^\star - \Sigma_{ij}^{k-1}\right| \geq \alpha\lambda$, then we have

$$0 \leq p'_\lambda\left(\Sigma_{ij}^{k-1}\right) \leq \lambda \leq \frac{\lambda}{\alpha\lambda}\left|\Sigma_{ij}^\star - \Sigma_{ij}^{k-1}\right| = \frac{1}{\alpha}\left|\Sigma_{ij}^\star - \Sigma_{ij}^{k-1}\right|, \tag{52}$$

where the first two inequalities are from Assumption 1.

If $\left|\Sigma_{ij}^\star - \Sigma_{ij}^{k-1}\right| \leq \alpha\lambda$, then we have

$$0 \leq p'_\lambda\left(\Sigma_{ij}^{k-1}\right) \leq p'_\lambda\left(\left|\Sigma_{ij}^\star\right| - \alpha\lambda\right) = 0, \tag{53}$$

where the second inequality follows from Assumption 2 that $\min_{(i,j)\in\mathcal{S}}|\Sigma_{ij}^\star| \geq (\alpha+\gamma)\lambda$ and $p'_\lambda(x) = 0$ for $x \geq \gamma\lambda$.

As a result, we can obtain

$$\delta_1^{k-1} \leq \frac{\lambda}{\alpha\lambda}\left\|\left(\boldsymbol{\Sigma}^{k-1} - \boldsymbol{\Sigma}^\star\right)_{\mathcal{S}}\right\|_{\mathrm{F}} \leq \frac{1}{\alpha}\left\|\boldsymbol{\Sigma}^{k-1} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}. \tag{54}$$

In order to to bound ρ_2^{k-1}, we separate $\bar{\mathcal{T}}_\alpha^{k-1}$ into $\mathcal{L}_\alpha^{k-1} \setminus S^\star$ and $\mathcal{S} \cup \mathcal{I}$. The term $\left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}}\right\|$ can be bounded as following

$$\begin{aligned}
\left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}}\right\|_{\mathrm{F}} \leq & \sqrt{\left|\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}\right|}\left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}}\right\|_{\max} \\
\leq & \frac{\sqrt{2}\lambda}{2\alpha\lambda}\left\|\boldsymbol{\Sigma}^{k-1} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \\
= & \frac{\sqrt{2}}{2\alpha}\left\|\boldsymbol{\Sigma}^{k-1} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}},
\end{aligned} \tag{55}$$

where the last inequality is from $\left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}}\right\|_{\max} \leq \frac{\sqrt{2}}{2}\lambda$, and $\sqrt{\operatorname{card}\left(\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}\right)} \leq \frac{\left\|\boldsymbol{\Sigma}^{k-1}_{\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}}\right\|_{\mathrm{F}}}{\alpha}$. The latter is according to the definition of $\mathcal{L}_\alpha^{k-1}$.

So the ρ_2^{k-1} is bounded as following

$$\begin{aligned}
\rho_2^{k-1} \leq & \left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{L}_\alpha^{k-1}\setminus\mathcal{S}}\right\|_{\mathrm{F}} + \left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}} + \tau\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}} \\
\leq & \frac{\sqrt{2}}{2\alpha}\left\|\boldsymbol{\Sigma}^{k-1} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} + \left\|(\boldsymbol{\Sigma}^\star - \boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}} + \tau\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\mathrm{F}}.
\end{aligned} \tag{56}$$

Integrating (54) and (56) into (50), we have

$$\left\|\boldsymbol{\Sigma}^k-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq \frac{\left(2+\sqrt{2}\right)}{2\alpha}\left\|\boldsymbol{\Sigma}^{k-1}-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}+\left\|(\boldsymbol{\Sigma}^\star-\boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}}+\tau\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}} . \tag{57}$$

Given $\alpha=3$, by setting $\delta=\frac{2+\sqrt{2}}{6}$, $d_k=\left\|\boldsymbol{\Sigma}^k-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}$ and $V=\left\|(\boldsymbol{\Sigma}^\star-\boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}}+\tau\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}$, we obtain

$$d_k \leq V+\delta d_{k-1}, \quad \forall k \geq 1. \tag{58}$$

Owing to the fact of $\delta\in(0,1)$, we have

$$d_k \leq \frac{1}{1-\delta}V+\delta^{k-1}d_1. \tag{59}$$

i.e.,

$$\left\|\boldsymbol{\Sigma}^k-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq \underbrace{\delta^{k-1}\left\|\boldsymbol{\Sigma}^1-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}}_{\text{optimization error}}+\underbrace{\frac{1}{1-\delta}\left(\left\|(\boldsymbol{\Sigma}^\star-\boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}}+\tau\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}\right)}_{\text{statistical error}}. \tag{60}$$

C.7 Proof of Corollary 3

In the initial stage, we set $\boldsymbol{\Sigma}^0=\boldsymbol{I}$. This implies $\operatorname{card}\left(\mathcal{T}_\alpha\left(\boldsymbol{\Sigma}^0\right)\right) \leq 2s-d$. According to Lemma C.2 and $d_1 \leq 3\lambda\sqrt{s}$, we can apply the inequality (37) and set $\boldsymbol{\Sigma}^1=\boldsymbol{\Sigma}$, we can have the following

$$\left\|\boldsymbol{\Sigma}^1-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq 3\lambda\sqrt{s}. \tag{61}$$

Note $\lambda \asymp \sqrt{\frac{\log d}{n}}$, we obtain

$$\left\|\boldsymbol{\Sigma}^1-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{s\log d}{n}}. \tag{62}$$

C.8 Proof of Corollary 4

Note $\tau \leq \sqrt{\frac{s}{n}}\left\|\left(\boldsymbol{\Sigma}^\star\right)^{-1}\right\|_{\mathrm{F}}^{-1}$, together with Lemma C.5 providing that $\left\|(\boldsymbol{\Sigma}^\star-\boldsymbol{S})_{\mathcal{S}\cup\mathcal{I}}\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{s}{n}}$ holds with high probability, it shows that $V \lesssim \sqrt{\frac{s}{n}}$ holds with high probability, too.

Recall $d_1 \leq 3\lambda\sqrt{s}$ and plug this into (60), we can have

$$\left\|\boldsymbol{\Sigma}^k-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \leq \frac{1}{1-\delta}V+\delta^{k-1}3\lambda\sqrt{s}. \tag{63}$$

If $K \geq 1+\frac{\log\left(\lambda\sqrt{n}\right)}{\log\delta^{-1}} \gtrsim \log\left(\lambda\sqrt{n}\right) \gtrsim \log\log d$, then we have

$$\delta^{k-1}\lambda\sqrt{s} \leq \frac{1}{\lambda\sqrt{n}}\lambda\sqrt{s} \leq \sqrt{\frac{s}{n}}. \tag{64}$$

Combing above results will yield $\left\|\boldsymbol{\Sigma}^K-\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{s}{n}}$ holds for high probability, which makes our formulation exhibit oracle property.