Adaptive Estimation of Graphical Models under Total Positivity

Jiaxi Ying [1 2] **José Vinícius de M. Cardoso** [1] **Daniel P. Palomar** [1]

Abstract

We consider the problem of estimating (diagonally dominant) M-matrices as precision matrices in Gaussian graphical models. These models exhibit intriguing properties, such as the existence of the maximum likelihood estimator with merely two observations for M-matrices (Lauritzen et al., 2019; Slawski & Hein, 2015) and even one observation for diagonally dominant M-matrices (Truell et al., 2021). We propose an adaptive multiple-stage estimation method that refines the estimate by solving a weighted ℓ_1-regularized problem at each stage. Furthermore, we develop a unified framework based on the gradient projection method to solve the regularized problem, incorporating distinct projections to handle the constraints of M-matrices and diagonally dominant M-matrices. A theoretical analysis of the estimation error is provided. The proposed method outperforms state-of-the-art methods in precision matrix estimation and graph edge identification, as evidenced by synthetic and financial time-series data sets.

1. Introduction

Total positivity, as a strong form of positive dependence, has found its applications in various fields such as factor analysis in psychometrics (Lauritzen et al., 2019), taxonomic reasoning (Lake & Tenenbaum, 2010), graph signal processing (Egilmez et al., 2017), and financial markets (Agrawal et al., 2020). For instance, stock markets exemplify total positivity, as stocks generally exhibit positive dependence due to the influence of market factors (Agrawal et al., 2020). A distribution that satisfies total positivity is also known as multivariate totally positive of order two (MTP_2). For a more in-depth understanding, we refer the reader to (Fallat et al., 2017). In the Gaussian context, the concept of MTP_2 becomes more straightforward: a Gaussian distribution is MTP_2 if and only if the precision matrix (*i.e.*, inverse covariance matrix) is an M-matrix (Karlin & Rinott, 1983). In this paper, our focus lies on estimating (diagonally dominant) M-matrices as precision matrices in Gaussian graphical models.

Estimation of precision matrices in general Gaussian graphical models have been extensively studied in the literature. A well-known method, known as graphical lasso (Banerjee et al., 2008; d'Aspremont et al., 2008), is formulated as an ℓ_1-regularized Gaussian maximum likelihood estimation. Numerous extensions of graphical lasso have been explored (Friedman et al., 2008; Ravikumar et al., 2011; Lam & Fan, 2009; Fan et al., 2014; Loh & Wainwright, 2015; 2017; Sun et al., 2018; Hsieh et al., 2014). Other notable, though not exhaustive, works include graphical Dantzig (Yuan, 2010), CLIME (Cai et al., 2011), and TIGER (Liu & Wang, 2017). Recent research has illuminated the intriguing properties of incorporating additional M-matrix constraint, which proves to be advantageous in the realms of high-dimensional statistics and signal processing on graphs.

Recent studies (Lauritzen et al., 2019; Slawski & Hein, 2015; Lauritzen et al., 2021) show that the M-matrix constraint significantly reduces the sample size required for the maximum likelihood estimator (MLE) to exist. Specifically, Lauritzen et al. (2019); Slawski & Hein (2015) established that the MLE under the M-matrix constraint exists if the sample size $n \geq 2$, regardless of the underlying dimension p, substantially reducing the $n \geq p$ requirement for general Gaussian distributions. Soloff et al. (2020) revealed that the M-matrix constraint, serving as implicit regularization, is vital for achieving the minimax rate of $\sqrt{\frac{\log p}{n}}$; without this constraint, the rate cannot exceed $\sqrt{\frac{p}{n}}$. Remarkably, Lauritzen et al. (2021) proved that under the MTP_2 constraint, the MLE in binary distributions may exist with only $n = p$ observations, in contrast to the 2^p observations required in the absence of this constraint.

[1]The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong [2]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China. Correspondence to: Jiaxi Ying <jx.ying@connect.ust.hk>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

The diagonally dominant M-matrix constraint is increasingly drawing interest in Gaussian models. In these distributions, diagonally dominant M-matrices, as precision matrices, exhibit a property called log-$L^\natural$-concave (LLC) (Robeva et al., 2021) or strong MTP_2 (Röttger et al., 2021). This property is essential for analyzing positive dependence in multivariate Pareto distributions. A recent study (Truell et al., 2021) revealed that Gaussian models under LLC/strong MTP_2 act as a convex relaxation for Brownian motion tree models—a class of Gaussian models on trees—and showed that its MLE exists almost surely when $n = 1$. This finding was established by connecting Laplacian constrained Gaussian Markov random fields (LGMRF) (Ying et al., 2020b; Cardoso et al., 2021; 2022; Kumar et al., 2020) and leveraging results on MLE existence in LGMRF with a single observation (Ying et al., 2021a).

The estimation of MTP_2 graphical models has become a growing area of interest in the field of signal processing on graphs (Shuman et al., 2013; Ortega et al., 2018; Dong et al., 2019), which focuses on handling data defined on irregular graph domains. Precision matrices in MTP_2 graphical models belong to the class of generalized Laplacian matrices (Biyikoglu et al., 2007), each of which is a symmetric matrix with non-positive off-diagonal entries that are associated with a graph. The eigenvectors of a generalized Laplacian define a graph Fourier transform (Shuman et al., 2013), supported by the discrete nodal domain theorem (Davies et al., 2001) that an eigenvector corresponding to a larger eigenvalue is associated with a higher frequency. We note that such a spectral property does not hold for general positive definite matrices.

One approach to estimating MTP_2 graphical models is MLE (Lauritzen et al., 2019; Slawski & Hein, 2015), which implicitly promotes sparsity using the M-matrix constraint. However, it cannot adjust sparsity levels to specific values, a feature often desired in practice. The ℓ_1-norm regularized MLE (Egilmez et al., 2017; Pavez & Ortega, 2016; Cai et al., 2021; Deng & So, 2020) offers improved sparsity control and can be solved using techniques like block coordinate descent (Egilmez et al., 2017; Pavez & Ortega, 2016), proximal point algorithm (Deng & So, 2020), and projected Newton-like methods (Cai et al., 2021). Notably, Pavez et al. (2018) found that some zero patterns in estimated matrix and graph components can be determined through thresholded sample covariance matrices. Estimating diagonally dominant M-matrices as precision matrices is explored in (Egilmez et al., 2017). Additionally, Wang et al. (2020) proposed a graph structure learning algorithm based on conditional independence testing, eliminating the need for adjustment of tuning parameters.

This paper focuses on estimating (diagonally dominant) M-matrices as precision matrices in MTP_2 Gaussian graphical models. The main contributions of this paper are threefold:

- We propose an adaptive multiple-stage estimation method that refines the estimate by solving a weighted ℓ_1-regularized problem at each stage. Then we develop a unified framework based on the gradient projection method to solve the regularized problem, equipped with distinct projections to handle the constraints of M-matrices and diagonally dominant M-matrices.
- We provide a thorough theoretical analysis of the estimation error for our method, which comprises two components: optimization error and statistical error. The optimization error decays at a linear rate, highlighting the progressive refinement of the estimate across iterative stages, whereas the statistical error captures the intrinsic statistical rate.
- Experiments on synthetic and financial time-series data demonstrate that our method outperforms state-of-the-art methods, concurrently achieving lower precision matrix estimation error and higher graph edge selection accuracy. Additionally, we observe that incorporating M-matrix constraint enhances the robustness regarding the selection of the regularization parameter.

Lower and upper case bold letters denote vectors and matrices, respectively. $[p]$ denotes the set $\{1, \ldots, p\}$. $\|\boldsymbol{x}\|_{\max} = \max_i |x_i|$. $\mathbb{S}^p_{++}$ denotes the set of $p \times p$ positive definite matrices. $\lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X})$ denote the maximum and minimum eigenvalues of $\boldsymbol{X}$, respectively. For matrix $\boldsymbol{X}$ and set S, $\boldsymbol{X}_S$ denotes a vector with dimension $|S|$, containing the entries of $\boldsymbol{X}$ indexed by S.

2. Preliminaries and Problem Formulation

We first introduce preliminaries about MTP_2 Gaussian graphical models, then present the problem formulation.

2.1. MTP_2 Gaussian Graphical Models

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with the set of nodes $\mathcal{V}$ and the set of edges $\mathcal{E}$. Associating a random vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^\star)$ with graph $\mathcal{G}$ forms a Gaussian graphical model that satisfies the following properties:

$$\Theta^\star_{ij} \neq 0 \iff \{i,j\} \in \mathcal{E} \; \forall \, i \neq j,$$
$$\Theta^\star_{ij} = 0 \iff x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{[p]\setminus\{i,j\}},$$

where $x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{[p]\setminus\{i,j\}}$ indicates that x_i is conditionally independent of x_j given the other random variables. Consequently, graph $\mathcal{G}$ characterizes the sparsity pattern of the precision matrix $\boldsymbol{\Theta}^\star$, where missing edges represent conditional independence between random variables.

We concentrate on estimating precision matrices for random variables following an MTP_2 Gaussian distribution. A multivariate Gaussian distribution with a positive definite covariance matrix $\mathbf{\Sigma}^\star$ is MTP_2 if and only if its precision matrix $\mathbf{\Theta}^\star$ is a symmetric M-matrix (Karlin & Rinott, 1983), *i.e.*, $\Theta^\star_{ij} \leq 0$ for all $i \neq j$. This is equivalent to stating that all partial correlations are nonnegative, where the partial correlation between any two variables x_i and x_j conditioned on all other variables equals $-\Theta^\star_{ij}/\sqrt{\Theta^\star_{ii}\Theta^\star_{jj}}$.

2.2. Problem Formulation

Consider a zero-mean random vector $\boldsymbol{x} = (x_1, \ldots, x_p)$ following a Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^\star)$. Our goal is to estimate the precision matrix $\mathbf{\Theta}^\star := (\mathbf{\Sigma}^\star)^{-1}$, which is a (diagonally dominant) M-matrix, given n independent and identically distributed observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathbb{R}^p$.

To estimate the precision matrix under MTP_2 Gaussian distributions, we can solve the penalized Gaussian maximum likelihood estimation problem as follows:

$$\underset{\mathbf{\Theta}\in\mathcal{S}}{\text{minimize}} \; -\log\det(\mathbf{\Theta}) + \operatorname{tr}\left(\mathbf{\Theta}\widehat{\mathbf{\Sigma}}\right) + \sum_{i\neq j} h_\lambda\left(|\Theta_{ij}|\right), \quad (1)$$

where h_λ is a sparsity penalty function, such as the ℓ_1-norm considered in (Egilmez et al., 2017), $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^{(i)}(\boldsymbol{x}^{(i)})^\top$ is the sample covariance matrix, and $\mathcal{S}$ is a feasible set. We consider two cases for $\mathcal{S}$:

$$\mathcal{M}^p = \left\{\mathbf{\Theta} \in \mathbb{S}^p_{++} \,|\, \Theta_{ij} = \Theta_{ji} \leq 0, \forall i \neq j\right\}, \quad (2)$$

and

$$\mathcal{M}^p_D = \left\{\mathbf{\Theta} \in \mathbb{S}^p_{++} | \Theta_{ij} = \Theta_{ji} \leq 0, \forall i \neq j, \mathbf{\Theta}\mathbf{1} \geq \mathbf{0}\right\}, \quad (3)$$

where $\mathcal{M}^p$ represents the set of all symmetric positive definite M-matrices, and $\mathcal{M}^p_D$ adds constraint $\mathbf{\Theta}\mathbf{1} \geq \mathbf{0}$ compared to $\mathcal{M}^p$, leading $\mathbf{\Theta}$ to be a diagonally dominant M-matrix. We propose an adaptive method for estimating M-matrices and diagonally dominant M-matrices.

3. Proposed Method

We first propose an adaptive multiple-stage method, then develop a unified framework for solving each stage.

3.1. Adaptive Estimation

Our method comprises multiple stages, where each stage refines the previous stage's estimate by solving a weighted ℓ_1-regularized problem. Specifically, at the k-th stage, we obtain the estimate $\widehat{\mathbf{\Theta}}^{(k)}$ by solving the following problem:

$$\underset{\mathbf{\Theta}\in\mathcal{S}}{\text{minimize}} \; -\log\det(\mathbf{\Theta}) + \operatorname{tr}\left(\mathbf{\Theta}\widehat{\mathbf{\Sigma}}\right) + \sum_{i\neq j} \lambda^{(k)}_{ij}\left|\Theta_{ij}\right|, \quad (4)$$

where $\lambda^{(k)}_{ij} = p_\lambda(|\widehat{\Theta}^{(k-1)}_{ij}|)$ with p_λ the weight-updating function, and $\widehat{\mathbf{\Theta}}^{(k-1)}$ is obtained at the $(k-1)$-th stage.

Algorithm 1 Solve Problem (5)

Input: Sample covariance matrix $\widehat{\mathbf{\Sigma}}, \mathbf{\Lambda}^{(k)}, \sigma, \alpha, \beta$.
for $t = 0, 1, 2, \ldots$ **do**
  Compute $\nabla f(\mathbf{\Theta}_t) = -\mathbf{\Theta}_t^{-1} + \widehat{\mathbf{\Sigma}} - \mathbf{\Lambda}^{(k)}$;
  $m \leftarrow 0$;
  repeat
    Update $\mathbf{\Theta}_{t+1} = \mathcal{P}_{\mathcal{S}_d}\left(\mathbf{\Theta}_t - \sigma\beta^m \nabla f(\mathbf{\Theta}_t)\right)$;
    $m \leftarrow m + 1$;
  until $\mathbf{\Theta}_{t+1} \in \mathbb{S}^p_{++}$ and
    $f(\mathbf{\Theta}_{t+1}) \leq f(\mathbf{\Theta}_t) - \alpha\sigma\beta^m \|G_{\frac{1}{\eta_t}}(\mathbf{\Theta}_t)\|^2_{\text{F}}$;
end for
Output: $\widehat{\mathbf{\Theta}}^{(k)}$.

When the estimate $\widehat{\mathbf{\Theta}}^{(k-1)}$ from the previous stage exhibits a large coefficient for its (i, j) entry, it is reasonable to assign a small $\lambda^{(k)}_{ij}$ in Problem (4). Therefore, p_λ should be monotonically non-increasing. In the initial stage, we set each $\lambda^{(1)}_{ij} = \lambda$, reducing Problem (4) to the well-studied ℓ_1-regularized estimation (Egilmez et al., 2017; Ying et al., 2021b). Our method refines the ℓ_1-norm estimate in subsequent stages, offering an improvement over the standard ℓ_1-norm approach.

By choosing p_λ as the derivative of the sparsity penalty function h_λ in Problem (1), the sequence $\left\{\widehat{\mathbf{\Theta}}^{(k)}\right\}_{k\geq 1}$ converges to a stationary point of Problem (1) with a nonconvex penalty. In this context, our method aligns with the local linear approximation method (Zou & Li, 2008) and the broader framework of multi-stage convex relaxation (Zhang, 2010b). Nonconvex approaches may encounter issues with sensitivity to the regularization parameter selection, as illustrated in Figure 6. However, we demonstrate that incorporating the M-matrix constraint considerably enhances the robustness regarding the selection of regularization parameter.

3.2. Gradient Projection Method

For each $\mathbf{\Theta} \in \mathcal{S}$, where $\mathcal{S} = \mathcal{M}^p$ or $\mathcal{M}^p_D$, all off-diagonal entries are nonpositive, leading to the following representation of Problem (4):

$$\underset{\mathbf{\Theta}\in\mathcal{S}}{\text{minimize}} \; -\log\det(\mathbf{\Theta}) + \operatorname{tr}\left(\mathbf{\Theta}\widehat{\mathbf{\Sigma}}\right) - \sum_{i\neq j} \lambda^{(k)}_{ij}\Theta_{ij}. \quad (5)$$

We design a gradient projection method to solve Problem (5). Initially, we perform a gradient descent step, $\mathbf{\Theta}_t - \eta_t \nabla f(\mathbf{\Theta}_t)$, with ∇f representing the gradient of the objective function f. As this step may exceed the constraint set $\mathcal{S}$, we subsequently project it back onto $\mathcal{S}$. However, such projection onto $\mathcal{S}$ does not exist due to its nonclosedness, resulting from the nonclosedness of $\mathbb{S}^p_{++}$.

Instead, we express the set $\mathcal{S}$ as the intersection of a closed set $\mathcal{S}_d$ and the open set $\mathbb{S}^p_{++}$:

$$\mathcal{S} = \mathcal{S}_d \cap \mathbb{S}^p_{++}. \tag{6}$$

We handle the constraints of $\mathcal{S}_d$ and $\mathbb{S}^p_{++}$ separately. The closed set $\mathcal{S}_d$ constraint is manageable through projection, leading to the following projected gradient descent:

$$\mathbf{\Theta}_{t+1} = \mathcal{P}_{\mathcal{S}_d}\big(\mathbf{\Theta}_t - \eta_t \nabla f(\mathbf{\Theta}_t)\big), \tag{7}$$

where $\mathcal{P}_{\mathcal{S}_d}$ denotes the projection onto the set $\mathcal{S}_d$ with respect to the Frobenius norm.

The constraint of $\mathbb{S}_{++}$ cannot be managed by projection due to its nonclosedness. One efficient approach to enforce positive definiteness is using a line search procedure (Hsieh et al., 2014). Specifically, at the t-th iteration, we try the step size $\eta_t \in \sigma\left\{\beta^0, \beta^1, \beta^2, \ldots\right\}$, with $\beta \in (0,1)$ and $\sigma > 0$, until we find the smallest $m \in \mathbb{N}$ such that the iterate $\mathbf{\Theta}_{t+1}$ in (7), with $\eta_t = \sigma\beta^m$, ensures $\mathbf{\Theta}_{t+1} \in \mathbb{S}^p_{++}$ and satisfies:

$$f(\mathbf{\Theta}_{t+1}) \leq f(\mathbf{\Theta}_t) - \alpha\eta_t \|G_{\frac{1}{\eta_t}}(\mathbf{\Theta}_t)\|_{\mathrm{F}}^2, \tag{8}$$

where $\alpha \in (0,1)$, and $G_{\frac{1}{\eta_t}}(\mathbf{\Theta}_t) = \frac{1}{\eta_t}(\mathbf{\Theta}_t - \mathbf{\Theta}_{t+1})$. We present our algorithm in Algorithm 1, where $\mathbf{\Lambda}^{(k)}$ contains all weights $\lambda_{ij}^{(k)}$, and diagonal entries are set to zero.

Theorem 3.1. *The sequence $\left\{\mathbf{\Theta}_t\right\}_{t\geq 0}$ established by Algorithm 1 converges to the optimal solution of Problem (5).*

The proof is provided in Appendix B.2. It is important to note that the existence of the minimizer for Problem (5) is assumed throughout this paper, which can be guaranteed almost surely if the sample size $n \geq 2$ for the case of $\mathcal{S} = \mathcal{M}^p$ (Lauritzen et al., 2019; Slawski & Hein, 2015) and $n \geq 1$ for the case of $\mathcal{S} = \mathcal{M}^p_D$ (Truell et al., 2021).

Although our method needs to solve a sequence of log-determinant programs (5), numerical results indicate a rapid decrease in the number of iterations required for solving (5) as k increases (see Figure 8). This accelerated convergence can be attributed to the use of the estimate from the previous stage as an initial point, thereby providing a warm start for faster convergence.

We now present notable extensions of the proposed gradient projection algorithm. First, the algorithm can be readily extended to solve Problem (1) with a nonconvex penalty. Second, the algorithm can be further adapted to solve the following problem for the case of multivariate t-distribution:

$$\underset{\mathbf{\Theta}\in\mathcal{S}}{\text{minimize}} -\log\det(\mathbf{\Theta}) + \frac{p+\nu}{n}\sum_{i=1}^{n}\log\big(1+\frac{1}{\nu}(\boldsymbol{x}^{(i)})^\top\mathbf{\Theta}\boldsymbol{x}^{(i)}\big),$$

where ν denotes the number of degrees of freedom, and each $\boldsymbol{x}^{(i)}$ is an observation. This formulation can be employed to learn positive partial correlation graphs. Notably, elliptical MTP_2 distributions are highly restrictive, and the total positivity property is not applicable to t-distributions (Rossell & Zwiernik, 2021). It is worth mentioning that, unlike block coordinate descent (Egilmez et al., 2017) and projected Newton-like methods (Cai et al., 2021), the proposed algorithm can handle these extensions, resulting in a more flexible and versatile approach.

3.3. Computation of Projections

We present the computation of projection $\mathcal{P}_{\mathcal{S}_d}$ in (7) for both cases of estimating M-matrices and diagonally dominant M-matrices, *i.e.*, $\mathcal{S} = \mathcal{M}^p$ and $\mathcal{M}^p_D$ in Problem (5).

3.3.1. Estimation of M-matrices

For the case of estimating an M-matrix as the precision matrix, the constraint set of Problem (5) is set as $\mathcal{S} = \mathcal{M}^p$, defined in (2). The closed set $\mathcal{S}_d$ in (6) then becomes

$$\mathcal{S}_d = \left\{\mathbf{\Theta} \in \mathbb{R}^{p\times p} \,|\, \Theta_{ij} = \Theta_{ji} \leq 0, \forall i \neq j\right\}, \tag{9}$$

which can be written as the intersection of two sets:

$$\mathcal{S}_d = \mathcal{S}_A \cap \mathcal{S}_B, \tag{10}$$

where $\mathcal{S}_A := \{\mathbf{\Theta} \in \mathbb{R}^{p\times p} \,|\, \Theta_{ij} = \Theta_{ji}, \forall i \neq j\}$ and $\mathcal{S}_B := \{\mathbf{\Theta} \in \mathbb{R}^{p\times p} \,|\, \Theta_{ij} \leq 0, \forall i \neq j\}$. If $\mathbf{\Theta}_t$ is symmetric, then the update $\mathcal{P}_{\mathcal{S}_B}\big(\mathbf{\Theta}_t - \eta_t\nabla f(\mathbf{\Theta}_t)\big)$ preserves symmetry. Thus, we only need to project $\mathbf{\Theta}_t - \eta_t\nabla f(\mathbf{\Theta}_t)$ onto $\mathcal{S}_B$. As a result, the iterate (7) can be simplified to

$$\mathbf{\Theta}_{t+1} = \mathcal{P}_{\mathcal{S}_B}\big(\mathbf{\Theta}_t - \eta_t\nabla f(\mathbf{\Theta}_t)\big), \tag{11}$$

where the projection $\mathcal{P}_{\mathcal{S}_B}$ can be computed as follows:

$$[\mathcal{P}_{\mathcal{S}_B}(\boldsymbol{X})]_{ij} = \begin{cases} \min(X_{ij}, 0) & \text{if } i \neq j, \\ X_{ij} & \text{if } i = j. \end{cases} \tag{12}$$

By initializing with a symmetric $\mathbf{\Theta}_0$, every point in the sequence $\{\mathbf{\Theta}_t\}_{t\geq 0}$ generated by (11) maintains symmetry.

3.3.2. Estimation of Diagonally Dominant M-matrices

For the case of estimating a diagonally dominant M-matrix as the precision matrix, we set $\mathcal{S} = \mathcal{M}^p_D$ in (5), with $\mathcal{M}^p_D$ defined in (3). The closed set $\mathcal{S}_d$ in (6) then becomes

$$\mathcal{S}_d = \{\mathbf{\Theta} \in \mathbb{R}^{p\times p} | \Theta_{ij} = \Theta_{ji} \leq 0, \forall i \neq j, \mathbf{\Theta}\mathbf{1} \geq \mathbf{0}\}, \tag{13}$$

which can be written as the intersection of two sets:

$$\mathcal{S}_d = \mathcal{S}_A \cap \mathcal{S}_C, \tag{14}$$

where $\mathcal{S}_C := \{\mathbf{\Theta} \in \mathbb{R}^{p\times p} \,|\, \mathbf{\Theta}\mathbf{1} \geq \mathbf{0}, \Theta_{ij} \leq 0, \forall i \neq j\}$.

The iterate (7) is then written as:

$$\boldsymbol{\Theta}_{t+1} = \mathcal{P}_{\mathcal{S}_A \cap \mathcal{S}_C}\big(\boldsymbol{\Theta}_t - \eta_t \nabla f(\boldsymbol{\Theta}_t)\big). \tag{15}$$

In contrast to (11), iterate (15) cannot be simplified, as $\mathcal{P}_{\mathcal{S}_C}\big(\boldsymbol{\Theta}_t - \eta_t \nabla f(\boldsymbol{\Theta}_t)\big)$ does not guarantee symmetry.

We now discuss computing the projection $\mathcal{P}_{\mathcal{S}_A \cap \mathcal{S}_C}$ in (15), defined as the minimizer of the following problem:

$$\mathcal{P}_{\mathcal{S}_A \cap \mathcal{S}_C}(\boldsymbol{Y}) := \underset{\boldsymbol{X} \in \mathcal{S}_A \cap \mathcal{S}_C}{\operatorname{arg\,min}} \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{16}$$

Problem (16) is a projection problem of a given point $\boldsymbol{Y}$ onto the intersection of the sets $\mathcal{S}_A$ and $\mathcal{S}_C$. To solve Problem (16), we design an algorithm based on Dykstra's projection (Boyle & Dykstra, 1986), which aims to find the nearest point projection onto the intersection of closed convex sets. The algorithm is summarized in Algorithm 2, with $\boldsymbol{Q}_k$ denoting the increment associated with the set $\mathcal{S}_C$. We do not need to introduce the increment associated with $\mathcal{S}_A$ for convergence, since $\mathcal{S}_A$ is a subspace.

Theorem 3.2. *The sequences $\{\boldsymbol{A}_k\}$ and $\{\boldsymbol{C}_k\}$ generated by Algorithm 2 converge to the minimizer of Problem (16).*

The convergence established in Theorem 3.2 is based on the convergence results of Dykstra's projection algorithm, as presented in (Boyle & Dykstra, 1986).

Algorithm 2 Compute $\mathcal{P}_{\mathcal{S}_A \cap \mathcal{S}_C}(\boldsymbol{Y})$

Input: $\boldsymbol{C}_0 = \boldsymbol{Y}$, $\boldsymbol{Q}_0 = \boldsymbol{0}$, ϵ;
$k = 1$;
repeat
  $\boldsymbol{A}_k = \mathcal{P}_{\mathcal{S}_A}(\boldsymbol{C}_{k-1})$;
  $\boldsymbol{C}_k = \mathcal{P}_{\mathcal{S}_C}(\boldsymbol{A}_k + \boldsymbol{Q}_{k-1})$;
  $\boldsymbol{Q}_k = \boldsymbol{A}_k - \boldsymbol{C}_k + \boldsymbol{Q}_{k-1}$;
  $k = k + 1$;
until $\|\boldsymbol{A}_k - \boldsymbol{C}_k\|_{\mathrm{F}} < \epsilon$
Output: $\mathcal{P}_{\mathcal{S}_A \cap \mathcal{S}_C}(\boldsymbol{Y}) = \boldsymbol{A}_k$.

We now present the computation of $\mathcal{P}_{\mathcal{S}_A}$ and $\mathcal{P}_{\mathcal{S}_C}$ from Algorithm 2. The computation of $\mathcal{P}_{\mathcal{S}_A}$ is straightforward:

$$\mathcal{P}_{\mathcal{S}_A}(\boldsymbol{Y}) = \big(\boldsymbol{Y} + \boldsymbol{Y}^\top\big)/2.$$

Projection $\mathcal{P}_{\mathcal{S}_C}$ is defined as follows:

$$\mathcal{P}_{\mathcal{S}_C}(\boldsymbol{Y}) := \underset{\boldsymbol{X} \in \mathcal{S}_C}{\operatorname{arg\,min}} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{17}$$

To compute $\mathcal{P}_{\mathcal{S}_C}$, we solve Problem (17) row by row. For the r-th row, we address the following problem:

$$\begin{array}{ll} \underset{\boldsymbol{x} \in \mathbb{R}^p}{\text{minimize}} & \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2, \\ \text{subject to} & \boldsymbol{x}^\top \boldsymbol{1} \geq 0,\ \boldsymbol{x}_{\setminus r} \leq \boldsymbol{0}, \end{array} \tag{18}$$

where $\boldsymbol{y} \in \mathbb{R}^p$ contains all entries of the r-th row of $\boldsymbol{Y}$, and $\boldsymbol{x}_{\setminus r} \in \mathbb{R}^{p-1}$ includes all entries of $\boldsymbol{x}$ except the r-th one.

Proposition 3.3 below, proven in Appendix B.3, provides the optimal solution of Problem (18), which is a variant of the projection onto the simplex (Condat, 2016).

Proposition 3.3. *Let $\hat{\boldsymbol{x}}$ be the optimal solution of Problem (18), which can be obtained as follows:*

- *If $y_r \geq -\sum_{i \in [p] \setminus r} \min(y_i, 0)$, then $\hat{x}_r = y_r$, and $\hat{x}_i = \min(y_i, 0)$ for $i \neq r$.*
- *If $y_r < -\sum_{i \in [p] \setminus r} \min(y_i, 0)$, then $\hat{x}_r = y_r + \rho$, and $\hat{x}_i = \min(y_i + \rho, 0)$ for $i \neq r$, where ρ satisfies $\rho + \sum_{i \in [p] \setminus r} \min(y_i + \rho, 0) + y_r = 0$.*

Remark 3.4. Let $g(\rho) := \rho + \sum_{i \in [p] \setminus r} \min(y_i + \rho, 0) + y_r$. Note that there exists a unique solution to the piecewise linear equation $g(\rho) = 0$. On one hand, $g(0) < 0$, and $g(a) > 0$ for any $a > \max_i |y_i|$. Therefore, there exits at least one solution to $g(\rho) = 0$ since $g(\rho)$ is continuous. On the other hand, the solution is unique because $g(\rho)$ is a monotone function that is strictly increasing.

Let $\boldsymbol{y}_{\setminus r} \in \mathbb{R}^{p-1}$ represent the vector that contains all entries of $\boldsymbol{y}$ except the r-th one. To solve the equation $g(\rho) = 0$, we follow the procedures outlined in (Palomar & Fonollosa, 2005), which consists of three steps:

1. Sort $\boldsymbol{y}_{\setminus r}$ and obtain the sorted version $\tilde{\boldsymbol{y}}_{\setminus r}$, where $[\tilde{\boldsymbol{y}}_{\setminus r}]_1 \leq [\tilde{\boldsymbol{y}}_{\setminus r}]_2 \leq \cdots \leq [\tilde{\boldsymbol{y}}_{\setminus r}]_{p-1}$.
2. Find $M := \max\limits_{1 \leq m \leq p-1} \left\{ m \,\middle|\, \frac{y_r + \sum_{i=1}^m [\tilde{\boldsymbol{y}}_{\setminus r}]_i}{m+1} > [\tilde{\boldsymbol{y}}_{\setminus r}]_m \right\}$.
3. Compute $\rho = \frac{-y_r - \sum_{m=1}^M [\tilde{\boldsymbol{y}}_{\setminus r}]_m}{M+1}$.

We note that sorting the vector $\boldsymbol{y}_{\setminus r}$ is the most computationally demanding step in the above procedures, typically necessitating $\mathcal{O}(p \log p)$ operations.

4. Analysis of Estimation Error

In this section, we provide theoretical analysis of the estimation error for the proposed method.

4.1. Theoretical Results

Let $\boldsymbol{\Sigma}^\star$ and $\boldsymbol{\Theta}^\star$ denote the underlying covariance and precision matrices, respectively, where $\boldsymbol{\Theta}^\star \in \mathcal{S}$ ($\mathcal{S} = \mathcal{M}^p$ or $\mathcal{M}_D^p$). Define $S^\star$ as the support set of the underlying precision matrix $\boldsymbol{\Theta}^\star$, excluding the diagonal entries, *i.e.*,

$$S^\star := \big\{(i,j) \in [p]^2 \,\big|\, \Theta_{ij}^\star \neq 0,\ i \neq j\big\}. \tag{19}$$

We require some mild conditions for the weight-updating function p_λ in Assumption 4.1 and the underlying precision matrix $\boldsymbol{\Theta}^\star$ in Assumption 4.2.

Assumption 4.1. The function $p_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies the following conditions:

1. $p_\lambda(x)$ is monotonically nonincreasing, and $p_\lambda(0) = \lambda$;
2. There exists a $\gamma > 0$ such that $p_\lambda(x) = 0$ for $x \geq \gamma\lambda$, and $p_\lambda(c_0\lambda) \geq \frac{\sqrt{2}}{2}\lambda$, where $c_0 = 12\lambda_{\max}^2(\mathbf{\Theta}^\star)$ is a constant.

Assumption 4.2. The underlying precision matrix $\mathbf{\Theta}^\star$ has at most s nonzero entries, and the off-diagonal nonzero entries satisfy $\min_{(i,j)\in S^\star} |\Theta_{ij}^\star| \geq (c_0 + \gamma)\lambda$, with c_0 and γ defined in Assumption 4.1. There exists a constant δ such that $0 < \delta \leq \lambda_{\min}(\mathbf{\Theta}^\star) \leq \lambda_{\max}(\mathbf{\Theta}^\star) \leq 1/\delta < \infty$.

Remark 4.3. Assumption 4.1 fundamentally covers folded concave penalties, such as smooth clipped absolute deviation (SCAD) (Fan & Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010a). In Assumption 4.2, the conditions on the underlying precision matrix are relatively mild, as the regularization parameter λ in our theorems takes the order of $\sqrt{\log p/n}$, which could be small when the sample size n increases. The assumption regarding minimal magnitude of signals is often utilized in nonconvex optimization analysis (Wang et al., 2016; Ying et al., 2020a).

The following theorem, proven in Appendix B.4, provides a non-asymptotic guarantee for estimation error.

Theorem 4.4. *Under Assumptions 4.1 and 4.2, take the regularization parameter $\lambda = \sqrt{c_1\tau\log p/n}$ with $\tau > 2$. If the sample size satisfies $n \geq 8c_0c_1\tau s\log p$, then the sequence $\widehat{\mathbf{\Theta}}^{(k)}$ generated by the proposed method satisfies*

$$\left\|\widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star\right\|_{\mathrm{F}} \leq \underbrace{4c_0\left\|\left(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\right)_{S^\star\cup I_p}\right\|}_{\text{Statistical error}} + \underbrace{\rho^k\left\|\widehat{\mathbf{\Theta}}^{(0)} - \mathbf{\Theta}^\star\right\|_{\mathrm{F}}}_{\text{Optimization error}},$$

with probability at least $1 - 4/p^{\tau-2}$, where $\rho = \frac{2+\sqrt{2}}{4}$, $c_0 = 12\lambda_{\max}^2(\mathbf{\Theta}^\star)$, and $c_1 = \left(80\max_i \Sigma_{ii}^\star\right)^2$.

Theorem 4.4 establishes that the estimation error between the estimated precision matrix and the underlying precision matrix can be upper bounded by two terms: optimization error and statistical error. The optimization error decreases at a linear rate, indicating the progressive refinement of the estimate across iterative stages. In contrast, the statistical error remains independent of k and does not decrease during iterations, capturing the intrinsic statistical rate. Consequently, the estimation error is primarily dominated by the statistical error.

We note that the statements in Theorem 4.4 are applicable to estimate both M-matrices and diagonally dominant M-matrices, *i.e.*, $\mathcal{S} = \mathcal{M}^p$ or $\mathcal{M}_D^p$ in Algorithm 1. The parameter τ is user-defined; a larger τ increases the probability of the claims being true but imposes a stricter requirement on the sample size.

Corollary 4.5. *Under the same assumptions and conditions as stated in Theorem 4.4, the sequence $\widehat{\mathbf{\Theta}}^{(k)}$ generated by the proposed method satisfies*

$$\left\|\widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star\right\|_{\mathrm{F}} \leq \underbrace{c\sqrt{s\log p/n}}_{\text{Statistical error}} + \underbrace{\rho^k\left\|\widehat{\mathbf{\Theta}}^{(0)} - \mathbf{\Theta}^\star\right\|_{\mathrm{F}}}_{\text{Optimization error}},$$

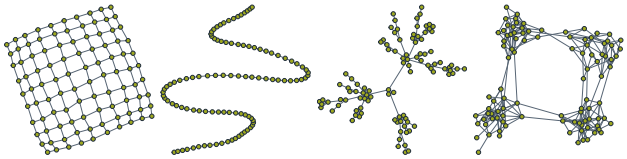
with probability at least $1 - 4/p^{\tau-2}$, where $c = 2c_0\sqrt{2c_1\tau}$.

Corollary 4.5 reveals that the statistical error follows the order of $\sqrt{s\log p/n}$, aligning with the minimax rate achieved in unconstrained graphical models (Ravikumar et al., 2011; Loh & Wainwright, 2015). However, the impact of the M-matrix constraint on the minimax rate for estimating sparse precision matrices remains uncertain. We highlight a recent study (Soloff et al., 2020) that investigated estimating precision matrices under the nonpositivity constraint without sparsity regularization. The minimax optimal rate under symmetrized Stein loss is established to be $\sqrt{\log p/n}$, while removing the nonpositivity constraint results in a reduced rate of $\sqrt{p/n}$. This finding suggests that the nonpositivity constraint affects the minimax rate, particularly in non-sparse situations.

5. Experimental Results

We conduct experiments on synthetic and financial time-series data. The code of our method is publicly available at `https://github.com/jxying/ddmtp2`.

State-of-the-art methods for comparison include GLASSO (Friedman et al., 2008), CLIME (Cai et al., 2011), GSCAD (Loh & Wainwright, 2015), GGL (Egilmez et al., 2017), DDGL (Egilmez et al., 2017), SLTP (Wang et al., 2020), and GOLAZO (Lauritzen & Zwiernik, 2022). GLASSO, CLIME and GSCAD focus on estimating precision matrices for general graphical models. The first two employ the ℓ_1-penalty, while the latter utilizes the SCAD penalty. In contrast, GGL, DDGL, and SLTP are specifically designed for MTP_2 graphical models. Both GGL and DDGL use the ℓ_1-norm regularization, aiming for estimating M-matrices and diagonally dominant M-matrices, respectively. Meanwhile, SLTP concentrates on learning graph structures rather than estimating precision matrices. Lastly, GOLAZO is tailored for estimating positive graphical models.



(a) (b) (c) (d)

Figure 1: Illustration of graph structures: (a) grid, (b) line, (c) Barabasi-Albert model, and (d) stochastic block model.

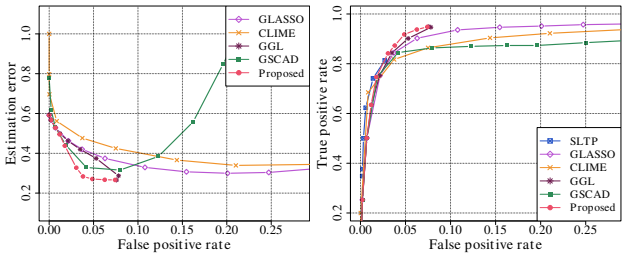


Figure 2: Estimation errors and true positive rates versus false positive rates on the grid graph with $\mathbf{\Theta}^\star \in \mathcal{M}^p$. The number of nodes $p = 100$, and the sample size $n = 100$.

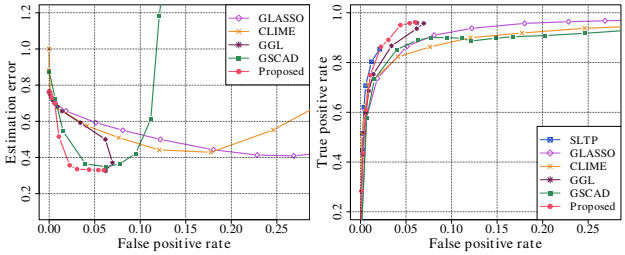


Figure 3: Estimation errors and true positive rates versus false positive rates on the line graph with $\mathbf{\Theta}^\star \in \mathcal{M}^p$. The number of nodes $p = 100$, and the sample size $n = 40$.

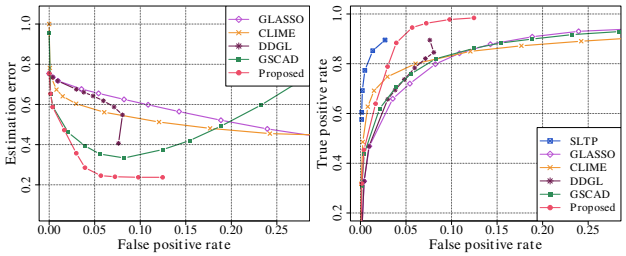


Figure 4: Estimation errors and true positive rates versus false positive rates for various methods on Barabasi-Albert model with $\mathbf{\Theta}^\star \in \mathcal{M}_D^p$, where $p = 100$ and $n = 100$.

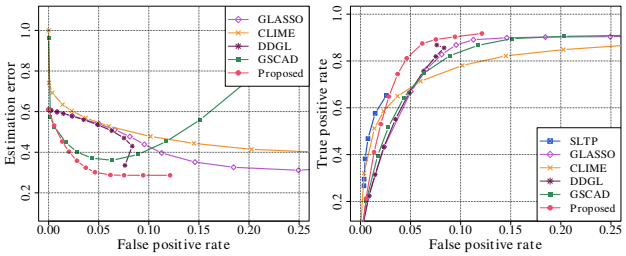


Figure 5: Estimation errors and true positive rates versus false positive rates for different methods on the stochastic block model with $\mathbf{\Theta}^\star \in \mathcal{M}_D^p$, where $p = 100$ and $n = 100$.

5.1. Synthetic Data

We examine four prevalent graph structures in synthetic experiments: grid, line, Barabasi-Albert model (Barabási & Albert, 1999), and stochastic block model (Holland et al., 1983), as depicted in Figure 1. Estimating these graph structures is essential in various applications, including social network analysis, image segmentation, and community detection. The graph structures are generated randomly, with each edge assigned a random weight. To assess the performance of precision matrix estimation, we employ estimation error as a metric. For evaluating graph edge selection accuracy, we utilize true/false positive rate and F-score. Further details about experimental settings, including performance measure definitions and synthetic data generation, can be found in Appendix A.

Figure 2 presents estimation errors and true positive rates of various methods as a function of false positive rates on a grid graph, given the number of nodes ($p = 100$), sample size ($n = 100$), and underlying precision matrix $\mathbf{\Theta}^\star \in \mathcal{M}^p$. The curve depicting true positive rates against false positive rates in Figure 2 is known as Receiver Operating Characteristic (ROC) curve, which is generated by setting a range of regularization parameter values from 10^{-3} to 1 for each method, except for SLTP. Instead, SLTP uses a parameter γ to control the sparsity of the learned graph, with values ranging from 0.75 to 0.98. The curves in both Figures 2 and 3 are averaged over 50 realizations.

Figure 2 shows that the proposed method attains a low estimation error and a high true positive rate simultaneously while maintaining a small false positive rate. In comparison, GLASSO and CLIME exhibit the best performance in estimation error and true positive rate within a region of high false positive rates. A higher true positive rate coupled with a lower false positive rate signifies better performance in identifying underlying graph edges. Our method surpasses GSCAD in achieving a smaller estimation error and a higher true positive rate, as we incorporate more prior knowledge of the constraint set $\mathcal{M}^p$. Although GGL also employs $\mathcal{M}^p$, our method achieves a better performance because it refines the GGL estimate in subsequent stages. The comparison with SLTP regarding estimation errors is excluded. It is noteworthy that GGL, SLTP, and our method all possess an upper bound on false positive rates.

Figure 3 displays the estimation errors and true positive rates of different methods as a function of false positive rates on a line graph. Our proposed method consistently achieves smaller estimation errors and higher true positive rates compared to state-of-the-art methods in the region of low false positive rates. This region is of primary interest, as an ideal estimate should exhibit low estimation error while maintaining high true positive rate and low false positive rate. Consequently, our method is capable of simultaneously achieving a reduced precision matrix estimation error and enhanced graph edge selection accuracy.

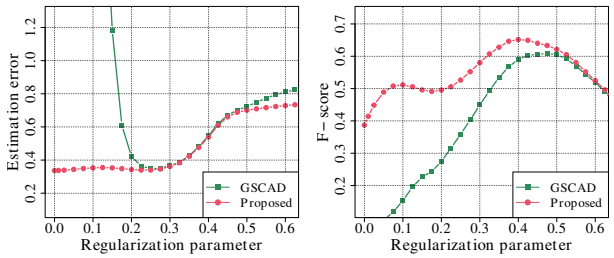


Figure 6: Estimation errors and F-scores as a function of varying regularization parameter values on a line graph.

Figure 4 presents comparisons on the Barabasi-Albert model with the underlying precision matrix $\boldsymbol{\Theta}^\star \in \mathcal{M}_D^p$. Our method consistently achieves lower estimation errors and higher true positive rates compared to state-of-the-art methods, highlighting its superior performance in both precision matrix estimation and graph edge identification. SLTP attains a higher true positive rate than all other methods when the false positive rate is below 0.05; however, its highest true positive rate is still lower than that of our method. Similar observations can be made for the stochastic block model, as displayed in Figure 5. The curves in Figures 4 and 5 are averaged over 50 realizations.

Figure 6 illustrates the robustness of GSCAD and our method regarding regularization parameter choices in terms of estimation error and F-score. Experiments are conducted on a line graph with $p = 100$ and $n = 40$. Figure 6 reveals that both estimation errors and F-scores of our method are more stable across varying regularization parameter values than those of GSCAD. Notably, the estimation error of GSCAD grows rapidly as the regularization parameter decreases, while that of our method remains steady. Thus, our method exhibits less sensitivity to the regularization parameter selection than GSCAD.

The key difference between our method and GSCAD is the former imposes an additional M-matrix constraint, while the latter does not. Thus, our method searches for a solution within a smaller space than GSCAD, excluding some stationary points that yield large estimation errors. This is why our method is more robust to regularization parameter selection. In fact, an extremely small regularization parameter can render the optimization problem for GSCAD ill-defined, meaning the minimum is not finite.

Table 1 displays the estimation errors for various methods under varying sample size ratios n/p, maintaining a low false positive rate (FPR < 0.05). The proposed method consistently achieves lower estimation errors compared to state-of-the-art methods across a range of sample size ratios. Notably, the superior performance of our method becomes more evident in cases with smaller sample size ratios, which are indicative of high-dimensional scenarios.

n/p	GLASSO	CLIME	GSCAD	GGL	Proposed
1	0.421	0.472	0.328	0.403	**0.267**
2	0.366	0.423	0.212	0.344	**0.171**
5	0.358	0.242	0.110	0.297	**0.100**

Table 1: Estimation errors for various methods across varying sample size ratios, with a consistently low false positive rate (FPR < 0.05).

5.2. Financial Time-series Data

In this subsection, we present comparisons of various methods on the financial time-series data, where the MTP_2 assumption has been well-justified due to the market factor causing positive dependence among stocks (Agrawal et al., 2020; Cardoso et al., 2020). We collect data from 204 stocks composing the S&P 500 index for the period between January 5th, 2004, and December 30th, 2006, resulting in 503 observations per stock (*i.e.*, $p = 204$ and $n = 503$). We construct the log-returns data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ as follows:

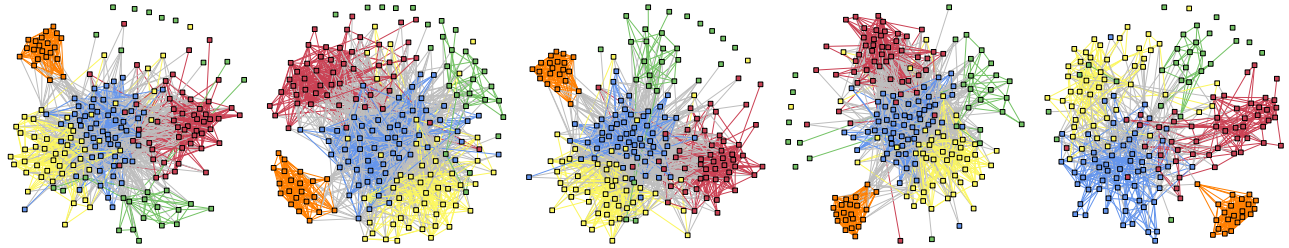
$$X_{i,j} = \log P_{i,j} - \log P_{i-1,j}, \tag{20}$$

where $P_{i,j}$ represents the closing price of the j-th stock on the i-th day. We then use the sample correlation matrix as input for each method. The 204 stocks are categorized into five sectors according to the Global Industry Classification Standard (GICS) system: Consumer Staples, Consumer Discretionary, Industrials, Energy, and Information Technology. We compare the proposed method with GLASSO, GSCAD, GGL, and GOLAZO on this financial time-series data.

In contrast to synthetic data, we cannot evaluate estimation error for each method, as the underlying precision matrix is unavailable. However, we expect stocks within the same sector to be somewhat interconnected. Consequently, we use the *modularity* (Newman, 2006) to assess the estimation performance. Define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$, where $\mathcal{V}$ is the vertex set, $\mathcal{E}$ is the edge set, and $\boldsymbol{A}$ is the adjacency matrix. The *modularity* of graph $\mathcal{G}$ is given by:

$$Q := \frac{1}{2|\mathcal{E}|} \sum_{i,j\in|\mathcal{V}|} \Big(A_{ij} - \frac{d_i d_j}{2|\mathcal{E}|}\Big)\delta(c_i, c_j),$$

where d_i represents the degree of the i-th node, c_i denotes the type of the i-th node, and $\delta(\cdot,\cdot)$ is the Kronecker delta function with $\delta(a,b) = 1$ if $a = b$ and 0 otherwise. A stock graph with high *modularity* exhibits dense connections between stocks within the same sector and sparse connections between stocks from different sectors. Therefore, a higher *modularity* value indicates a better representation of the actual stock network.



(a) GLASSO (b) GSCAD (c) GGL (d) GOLAZO (e) Proposed

Figure 7: Stock graphs characterized by the precision matrices estimated via (a) GLASSO, (b) GSCAD, (c) GGL, (d) GOLAZO, and (e) the proposed method. The *modularity* values for GLASSO, GSCAD, GGL, GOLAZO, and our method are 0.3534, 0.3433, 0.3530, 0.3578, and 0.4978, respectively. We fine-tune the regularization parameter for each method based on the *modularity* value while allowing only a few isolated nodes.

Table 2: *Modularity* values of S&P 500 stock graphs learned via various methods using a rolling-window approach, covering the period from January 3rd, 2014, to October 31st, 2017.

Methods	Jan. 2016	Apr. 2016	Jul. 2016	Oct. 2016	Jan. 2017	Apr. 2017	Jul. 2017	Oct. 2017
GLASSO	0.4038	0.4072	0.4115	0.3956	0.3982	0.3910	0.3911	0.4184
GSCAD	0.3880	0.3850	0.3935	0.3803	0.3848	0.3771	0.3822	0.3860
GGL	0.4088	0.4118	0.4193	0.3991	0.3927	0.3896	0.3845	0.4156
GOLAZO	0.4068	0.4098	0.4173	0.4007	0.4029	0.3945	0.3953	0.4199
Proposed	**0.5441**	**0.5289**	**0.5298**	**0.5139**	**0.5138**	**0.5118**	**0.5219**	**0.5019**

Figure 7 shows stock graphs based on precision matrices estimated using various methods. It is observed that the performance of the proposed method stands out, as most connections are between nodes within the same sector, and only a few connections (gray-colored edges) are between nodes from distinct sectors, which are often spurious from a practical perspective. We further compare the *modularity* value for each method. The *modularity* values for GLASSO, GSCAD, GGL, GOLAZO, and our method are 0.3534, 0.3433, 0.3530, 0.3578, and 0.4978, respectively, indicating that our method outperforms state-of-the-art methods in representing the stock network.

In the subsequent experiment, we collect data from 82 stocks in the S&P 500 index, representing three sectors: Communication Services, Utilities, and Real Estate. The data spans a timeframe from January 3rd, 2014, to October 31st, 2017, resulting in 945 observations.

Table 2 presents *modularity* values of graphs learned using a rolling-window approach. We select a window length of 504 days (roughly two years of stock market days) and shift it by 63 days (approximately three months of stock market days). Higher *modularity* signifies a more accurate representation of the actual stock network. Table 2 demonstrates that the proposed method consistently attains the highest *modularity* values throughout the entire period.

6. Conclusions

In this paper, we have proposed an adaptive multiple-stage method to estimate both M-matrices and diagonally dominant M-matrices as precision matrices in MTP_2 Gaussian graphical models. The proposed method starts with an initial estimate obtained via an ℓ_1-regularized maximum likelihood estimation and refines it in subsequent stages by solving a sequence of adaptive ℓ_1-regularized problems. We have provided theoretical analysis for estimation error. Experiments on both synthetic and real-world data have shown that our method outperforms state-of-the-art methods in estimating precision matrices and identifying graph edges. We have also demonstrated that incorporating the M-matrix constraint improves robustness regarding regularization parameter selection.

Acknowledgements

This work was supported by the Hong Kong Research Grants Council GRF 16207820, 16310620, and 16306821, the Hong Kong Innovation and Technology Fund (ITF) MHP/009/20, and the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under Grant HZQB-KCZYB-2020083. We would also like to thank the anonymous ICML reviewers for their valuable feedback on the manuscript.

References

Agrawal, R., Roy, U., and Uhler, C. Covariance Matrix Estimation under Total Positivity for Portfolio Selection. *Journal of Financial Econometrics*, 20(2):367–389, 2020.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Beck, A. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

Biyikoglu, T., Leydold, J., and Stadler, P. F. *Laplacian eigenvectors of graphs: Perron-Frobenius and Faber-Krahn type theorems*. Springer, 2007.

Boyle, J. P. and Dykstra, R. L. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in order restricted statistical inference*, pp. 28–47. Springer, 1986.

Cai, J.-F., Cardoso, J. V. D. M., Palomar, D. P., and Ying, J. Fast projected newton-like method for precision matrix estimation under total positivity. *arXiv preprint arXiv:2112.01939*, 2021.

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

Cardoso, J. V. d. M., Ying, J., and Palomar, D. P. Algorithms for learning graphs in financial markets. *arXiv preprint arXiv:2012.15410*, 2020.

Cardoso, J. V. D. M., Ying, J., and Palomar, D. P. Graphical models in heavy-tailed markets. In *Advances in Neural Information Processing Systems*, 2021.

Cardoso, J. V. D. M., Ying, J., and Palomar, D. P. Learning bipartite graphs: Heavy tails and multiple components. In *Advances in Neural Information Processing Systems*, 2022.

Condat, L. Fast projection onto the simplex and the l_1 ball. *Mathematical Programming*, 158(1):575–585, 2016.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

Davies, E. B., Leydold, J., and Stadler, P. F. Discrete nodal domain theorems. *Linear Algebra and its Applications*, 336(1):51–60, 2001.

Deng, Z. and So, A. M.-C. A fast proximal point algorithm for generalized graph Laplacian learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5425–5429, 2020.

Dong, X., Thanou, D., Rabbat, M., and Frossard, P. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3): 44–63, 2019.

Drazin, M. P. and Haynsworth, E. V. Criteria for the reality of matrix eigenvalues. *Mathematische Zeitschrift*, 78(1): 449–452, 1962.

Egilmez, H. E., Pavez, E., and Ortega, A. Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, 2017.

Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., and Zwiernik, P. Total positivity in Markov structures. *The Annals of Statistics*, 45(3):1152–1184, 2017.

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

Fan, J., Xue, L., and Zou, H. Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819, 2014.

Fan, J., Liu, H., Sun, Q., and Zhang, T. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics*, 46(2):814, 2018.

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

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. QUIC: quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):2911–2947, 2014.

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

Kumar, S., Ying, J., Cardoso, J. V. d. M., and Palomar, D. P. A unified framework for structured graph learning via spectral constraints. *Journal of Machine Learning Research*, 21(22):1–60, 2020.

Lake, B. and Tenenbaum, J. Discovering structure by learning sparse graphs. In *Proceedings of the Annual Cognitive Science Conference*, pp. 778—783, 2010.

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

Lauritzen, S. and Zwiernik, P. Locally associated graphical models and mixed convex exponential families. *The Annals of Statistics*, 50(5):3009–3038, 2022.

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

Lauritzen, S., Uhler, C., and Zwiernik, P. Total positivity in exponential families with application to binary variables. *The Annals of Statistics*, 49(3):1436–1459, 2021.

Liu, H. and Wang, L. Tiger: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Electronic Journal of Statistics*, 11(1):241–294, 2017.

Loh, P.-L. and Wainwright, M. J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16 (1):559–616, 2015.

Loh, P.-L. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics*, 45(6):2455–2482, 2017.

Newman, M. E. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., and Vandergheynst, P. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

Palomar, D. P. and Fonollosa, J. R. Practical algorithms for a family of waterfilling solutions. *IEEE transactions on Signal Processing*, 53(2):686–695, 2005.

Pavez, E. and Ortega, A. Generalized Laplacian precision matrix estimation for graph signal processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6350–6354, 2016.

Pavez, E., Egilmez, H. E., and Ortega, A. Learning graphs with monotone topology properties and multiple connected components. *IEEE Transactions on Signal Processing*, 66(9):2399–2413, 2018.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Robeva, E., Sturmfels, B., Tran, N., and Uhler, C. Maximum likelihood estimation for totally positive log-concave densities. *Scandinavian Journal of Statistics*, 48(3):817–844, 2021.

Rossell, D. and Zwiernik, P. Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics*, 15(2): 4236–4263, 2021.

Röttger, F., Engelke, S., and Zwiernik, P. Total positivity in multivariate extremes. *arXiv preprint arXiv:2112.14727*, 2021.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

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

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

Sun, Q., Tan, K. M., Liu, H., and Zhang, T. Graphical nonconvex optimization via an adaptive convex relaxation. In *International Conference on Machine Learning*, pp. 4810–4817, 2018.

Truell, M., Hütter, J.-C., Squires, C., Zwiernik, P., and Uhler, C. Maximum likelihood estimation for Brownian motion tree models based on one sample. *arXiv preprint arXiv:2112.00816*, 2021.

Wang, L., Ren, X., and Gu, Q. Precision matrix estimation in high dimensional Gaussian graphical models with faster rates. In *Artificial Intelligence and Statistics*, pp. 177–185, 2016.

Wang, Y., Roy, U., and Uhler, C. Learning high-dimensional Gaussian graphical models under total positivity without adjustment of tuning parameters. In *International*

Conference on Artificial Intelligence and Statistics, pp. 2698–2708, 2020.

Xiao, L. and Zhang, T. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Ying, J., Cardoso, J. V. d. M., and Palomar, D. P. Does the ℓ_1-norm learn a sparse graph under Laplacian constrained graphical models? *arXiv preprint arXiv:2006.14925*, 2020a.

Ying, J., de Miranda Cardoso, J. V., and Palomar, D. P. Nonconvex sparse graph learning under Laplacian constrained graphical model. In *Advances in Neural Information Processing Systems*, 2020b.

Ying, J., Cardoso, J. V. D. M., and Palomar, D. Minimax estimation of Laplacian constrained precision matrices. In *International Conference on Artificial Intelligence and Statistics*, pp. 3736–3744, 2021a.

Ying, J., Cardoso, J. V. D. M., and Palomar, D. P. A fast algorithm for graph learning under attractive Gaussian Markov random fields. In *Asilomar Conference on Signals, Systems, and Computers*, pp. 1520–1524, 2021b.

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

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

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010b.

Zou, H. and Li, R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4): 1509, 2008.

A. Experimental Settings and Additional Results

We first provide a more in-depth description of experimental settings in Appendix A.1. Next, we present additional experimental results in Appendix A.2.

A.1. Experimental Settings

To generate synthetic data, we first construct a graph structure randomly, then associate each edge with a weight that is uniformly sampled from $U(2,5)$. Then we obtain a weighted adjacency matrix $\boldsymbol{A}$, where A_{ij} denotes the graph weight between node i and node j.

To generate an underlying precision matrix $\boldsymbol{\Theta}^\star \in \mathcal{M}^p$, we adopt the procedures used in (Slawski & Hein, 2015; Wang et al., 2020). We first set

$$\boldsymbol{\Theta} = \delta \boldsymbol{I} - \boldsymbol{A}, \quad \text{with} \quad \delta = 1.05\lambda_{\max}\big(\boldsymbol{A}\big), \tag{21}$$

where $\lambda_{\max}\big(\boldsymbol{A}\big)$ denotes the largest eigenvalue of $\boldsymbol{A}$. We set $\boldsymbol{\Theta}^\star = \boldsymbol{E}\boldsymbol{\Theta}\boldsymbol{E}$, where $\boldsymbol{E}$ is a diagonal matrix chosen such that the covariance matrix $(\boldsymbol{\Theta}^\star)^{-1}$ has unit diagonal elements. The matrix $\boldsymbol{\Theta}^\star$ as generated above must be an M-matrix, but usually not diagonally dominant. Given a precision matrix $\boldsymbol{\Theta}^\star$, we generate n independent and identically distributed observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \sim \mathcal{N}(\boldsymbol{0}, (\boldsymbol{\Theta}^\star)^{-1})$. The sample covariance matrix is constructed by $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^{(i)}(\boldsymbol{x}^{(i)})^\top$.

We also conduct experiments for estimating diagonally dominant M-matrices. To generate an underlying precision matrix $\boldsymbol{\Theta}^\star \in \mathcal{M}_D^p$, we construct $\boldsymbol{\Theta}^\star = \boldsymbol{D} - \boldsymbol{A} + \boldsymbol{V}$, where $\boldsymbol{D}$ and $\boldsymbol{V}$ are diagonal matrices with each $D_{ii} = \sum_{j=1}^p A_{ij}$ and each V_{ii} uniformly sampled from $U(0,1)$.

To evaluate the estimation performance across different methods, we compute the estimation error as $\big\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\big\|_{\mathrm{F}} / \big\|\boldsymbol{\Theta}^\star\big\|_{\mathrm{F}}$, where $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^\star$ denote the estimated and underlying precision matrices, respectively. The true positive rate (TPR) and false positive rate (FPR) are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \text{and} \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}},$$

where TP, FP, TN and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. The true positive rate denotes the proportion of correctly identified edges in the graph, and the false positive rate denotes the proportion of incorrectly identified edges in the graph. An estimate with a higher true positive rate and a lower false positive rate indicates a better performance of identifying the underlying graph edges.

The F-score of a graph is defined as

$$\text{F-score} := \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{22}$$

The F-score takes values in $[0,1]$, and 1 indicates perfect structure recovery. Our method uses the weight-updating function corresponding to the SCAD penalty.

A.2. Additional Experimental Results

We perform experiments to demonstrate the empirical convergence of the proposed gradient projection algorithm, as illustrated in Figure 8. We observe that the number of iterations required for convergence descends rapidly with increasing stage k, and empirically the algorithm enjoys a linear convergence rate.

When estimating diagonally dominant M-matrices, our proposed algorithm involves double loops of iterations to solve the subproblem at each stage, potentially leading to computational inefficiency due to the use of Dykstra's algorithm for handling the diagonally dominant M-matrix (DDM-matrix) constraint. Table 3 displays the running time of each algorithm for the case of estimating DDM-matrices. We observe that our algorithm is less computationally efficient in this scenario, primarily due to managing multiple constraints to ensure a DDM-matrix solution. To enhance our algorithm's efficiency, we could compute an approximate solution when addressing subproblems in earlier stages, a technique employed in previous studies (Xiao & Zhang, 2013; Fan et al., 2018). Investigating the development of numerical algorithms for handling diagonally dominant M-matrix constraints more efficiently would be a valuable future research direction.

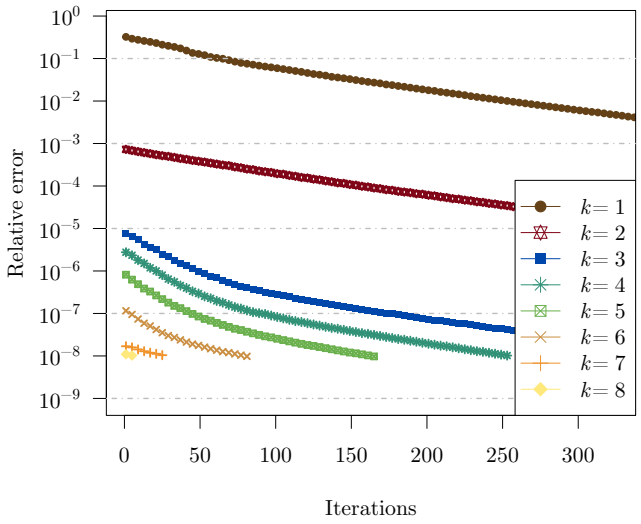


Figure 8: An empirical convergence result of the gradient projection algorithm at different stages. At the k-th stage with $k = 1, 2, \ldots, 8$, we plot the relative error $\left\|\mathbf{\Theta}_t - \widehat{\mathbf{\Theta}}^{(k)}\right\|_{\mathrm{F}} / \left\|\widehat{\mathbf{\Theta}}^{(k)}\right\|_{\mathrm{F}}$ as a function of the iteration t, where $\{\mathbf{\Theta}_t\}_{t\geq 1}$ is established by the gradient projection algorithm, and $\widehat{\mathbf{\Theta}}^{(k)}$ is the optimal solution at the k-th stage.

Table 3: Comparison of running times for methods in the case of estimating diagonally dominant M-matrices.

Methods	SLTP	GLASSO	CLIME	GSCAD	DDGL	Proposed
Running time (s)	9.48	0.05	4.06	0.14	1.17	7.85

It is essential to emphasize that our proposed algorithm is the sole method effectively addressing the Gaussian maximum likelihood estimation problem under the DDM-matrix constraints. In contrast, GLASSO, CLIME, and GSCAD tackle unconstrained problems and cannot guarantee a DDM-matrix solution. Such a solution provides desirable spectral properties for performing the graph Fourier transform and is crucial for analyzing positive dependence in multivariate Pareto distributions. Furthermore, the popular Block Coordinate Descent (BCD) algorithm cannot ensure convergence to the global minimizer, as demonstrated in Table 4.

We compare the optimality gap for solutions obtained by the BCD algorithm and our proposed algorithm. The optimality gap is evaluated using two approaches:

$$\Delta_{\Theta} = \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}} / \|\mathbf{\Theta}^*\|_{\mathrm{F}}, \qquad \text{and} \qquad \Delta_f = \hat{f} - f^*,$$

where $\widehat{\mathbf{\Theta}}$ is the output from BCD or our algorithm, $\mathbf{\Theta}^*$ is the global minimizer obtained through CVX, and $\hat{f}$ and f^* represent the objective function values at $\widehat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$, respectively. Experiments are conducted on the Barabasi-Albert model with 50 nodes ($p = 50$) and the number of samples n ranging from 50 to 5000. For both algorithms, the stopping criterion is met when successive iterations satisfy the condition $\|\mathbf{\Theta}_{t+1} - \mathbf{\Theta}_t\|_{\mathrm{F}} / \|\mathbf{\Theta}_t\|_{\mathrm{F}} < 10^{-12}$.

Table 4: Optimality gap comparison for algorithms in the case of estimating diagonally dominant M-matrices.

Algorithms	BCD		Proposed	
$n = 50$	$\Delta_{\Theta} = 0.5295,$	$\Delta_f = 0.0738$	$\Delta_{\Theta} = 2.81 \times 10^{-6},$	$\Delta_f = 1.49 \times 10^{-12}$
$n = 500$	$\Delta_{\Theta} = 0.0765,$	$\Delta_f = 0.0022$	$\Delta_{\Theta} = 1.09 \times 10^{-5},$	$\Delta_f = 2.68 \times 10^{-12}$
$n = 5000$	$\Delta_{\Theta} = 0.0166,$	$\Delta_f = 0.0004$	$\Delta_{\Theta} = 1.39 \times 10^{-5},$	$\Delta_f = 2.77 \times 10^{-12}$

Table 4 reveals that the popular BCD algorithm exhibits a substantial optimality gap, indicating its failure to converge to the global minimizer. In contrast, our proposed algorithm effectively converges to the minimizer. We note that the BCD algorithm can perform well under the M-matrix constraint without diagonal dominance.

B. Proofs

We first present two lemmas in Appendix B.1. Then we provide the proofs of Theorem3.1 in Appendix B.2, Proposition 3.3 in Appendix B.3, and Theorem 4.4 and Corollary 4.5 in Appendix B.4.

We begin by introducing the notations used in proofs. Define

$$L_\alpha(\mathbf{\Theta}) := \left\{(i,j) \,\middle|\, |\Theta_{ij}| \geq \alpha,\, i \neq j\right\}, \tag{23}$$

where $\alpha = c_0\lambda$ with c_0 defined in Assumption 4.1. Then we define the set $T_\alpha(\mathbf{\Theta})$ as follows:

$$T_\alpha(\mathbf{\Theta}) := L_\alpha(\mathbf{\Theta}) \cup S^\star, \tag{24}$$

where $S^\star$ is defined in (19). Define

$$G_{\lambda,\widehat{\mathbf{\Sigma}}}(\widetilde{\mathbf{\Theta}}) := \underset{\mathbf{\Theta}\in\mathcal{M}^p}{\text{arg min}} - \log\det(\mathbf{\Theta}) + \operatorname{tr}\left(\mathbf{\Theta}\widehat{\mathbf{\Sigma}}\right) + \sum_{i\neq j} p_\lambda(|\widetilde{\Theta}_{ij}|)\,|\Theta_{ij}|\,.$$

For ease of presentation, by a slight abuse of notation, we write $p_\lambda\left(|\mathbf{\Theta}_{S^\star}|\right)$ for $\left(p_\lambda(|\Theta_{ij}|)\right)_{(i,j)\in S^\star}$.

B.1. Lemmas

Lemma B.1. *Let $g(\mathbf{\Theta}) = -\log\det(\mathbf{\Theta})$. Define a local region of $\mathbf{\Theta}^\star$ by*

$$\mathcal{B}\left(\mathbf{\Theta}^\star, r\right) = \left\{\mathbf{\Theta}\in\mathcal{M}^p \,\middle|\, \|\mathbf{\Theta}-\mathbf{\Theta}^\star\|_{\mathrm{F}} \leq r\right\}.$$

Then for any $\mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \mathcal{B}\left(\mathbf{\Theta}^\star; \lambda_{\max}\left(\mathbf{\Theta}^\star\right)\right)$, we have

$$\left\langle \nabla g\left(\mathbf{\Theta}_1\right) - \nabla g\left(\mathbf{\Theta}_2\right), \mathbf{\Theta}_1 - \mathbf{\Theta}_2\right\rangle \geq \frac{1}{4}\lambda_{\max}^{-2}\left(\mathbf{\Theta}^\star\right)\|\mathbf{\Theta}_1-\mathbf{\Theta}_2\|_{\mathrm{F}}^2\,.$$

Proof. For any $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2 \in \mathcal{B}\left(\mathbf{\Theta}^\star; \lambda_{\max}\left(\mathbf{\Theta}^\star\right)\right)$, by Mean Value Theorem, one obtains

$$g\left(\mathbf{\Theta}_2\right) = g\left(\mathbf{\Theta}_1\right) + \left\langle \nabla g\left(\mathbf{\Theta}_1\right), \mathbf{\Theta}_2 - \mathbf{\Theta}_1\right\rangle + \frac{1}{2}\operatorname{vec}\left(\mathbf{\Theta}_2-\mathbf{\Theta}_1\right)^\top \nabla^2 g\left(\mathbf{\Theta}_t\right)\operatorname{vec}\left(\mathbf{\Theta}_2-\mathbf{\Theta}_1\right),$$

where $\mathbf{\Theta}_t = t\mathbf{\Theta}_2 + (1-t)\mathbf{\Theta}_1$ with $t\in[0,1]$. Note that $\mathbf{\Theta}_t \in \mathcal{B}\left(\mathbf{\Theta}^\star; r\right)$. One has

$$\lambda_{\min}\left(\nabla^2 g\left(\mathbf{\Theta}_t\right)\right) = \lambda_{\min}\left(\left(\mathbf{\Theta}_t\otimes\mathbf{\Theta}_t\right)^{-1}\right) = \lambda_{\max}^{-2}\left(\mathbf{\Theta}_t\right),$$

where the second equality follows from the property that the eigenvalues of $\boldsymbol{A}\otimes\boldsymbol{B}$ are $\lambda_i\mu_j$, where $\lambda_1,\ldots,\lambda_p$ and $\mu_1,\ldots,\mu_p$ are the eigenvalues of $\boldsymbol{A}\in\mathbb{R}^{p\times p}$ and $\boldsymbol{B}\in\mathbb{R}^{p\times p}$, respectively. Following from the Weyl's inequality, one obtains

$$\lambda_{\max}\left(\mathbf{\Theta}_t\right) \leq \lambda_{\max}\left(\mathbf{\Theta}_t-\mathbf{\Theta}^\star\right) + \lambda_{\max}\left(\mathbf{\Theta}^\star\right) \leq r + \lambda_{\max}\left(\mathbf{\Theta}^\star\right).$$

Therefore, We have

$$g\left(\mathbf{\Theta}_2\right) \geq g\left(\mathbf{\Theta}_1\right) + \left\langle \nabla g\left(\mathbf{\Theta}_1\right), \mathbf{\Theta}_2-\mathbf{\Theta}_1\right\rangle + \frac{1}{2}\left(r+\lambda_{\max}\left(\mathbf{\Theta}^\star\right)\right)^{-2}\|\mathbf{\Theta}_2-\mathbf{\Theta}_1\|_{\mathrm{F}}^2\,, \tag{25}$$

and

$$g\left(\mathbf{\Theta}_1\right) \geq g\left(\mathbf{\Theta}_2\right) + \left\langle \nabla g\left(\mathbf{\Theta}_2\right), \mathbf{\Theta}_1-\mathbf{\Theta}_2\right\rangle + \frac{1}{2}\left(r+\lambda_{\max}\left(\mathbf{\Theta}^\star\right)\right)^{-2}\|\mathbf{\Theta}_2-\mathbf{\Theta}_1\|_{\mathrm{F}}^2\,. \tag{26}$$

Setting $r = \lambda_{\max}\left(\mathbf{\Theta}^\star\right)$ and combining (25) and (26), we obtain

$$\left\langle \nabla g\left(\mathbf{\Theta}_1\right) - \nabla g\left(\mathbf{\Theta}_2\right), \mathbf{\Theta}_1 - \mathbf{\Theta}_2\right\rangle \geq \frac{1}{4}\lambda_{\max}^{-2}\left(\mathbf{\Theta}^\star\right)\|\mathbf{\Theta}_1-\mathbf{\Theta}_2\|_{\mathrm{F}}^2\,.$$

completing the proof. □

Lemma B.2. *Suppose the event* $\left\|\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}\right\|_{\max} \leq \frac{\sqrt{2}}{2}\lambda$ *holds and the sample size satisfies* $n \geq 8c_0c_1\tau s \log p$*, and take* $\lambda = \sqrt{c_1\tau \log p/n}$*. Under Assumptions 4.1 and 4.2, if* $|T_\alpha(\widetilde{\boldsymbol{\Theta}})| \leq 2s - p$ *holds for some* $\widetilde{\boldsymbol{\Theta}}$*, then*

$$\left\|G_{\lambda,\widehat{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\Theta}}) - \boldsymbol{\Theta}^\star\right\|_{\mathrm{F}} \leq \frac{c_0}{2}\left(\left\|p_\lambda\left(\left|\widetilde{\boldsymbol{\Theta}}_{S^\star}\right|\right)\right\| + \left\|\left(\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}\right)_{T_\alpha(\widetilde{\boldsymbol{\Theta}})\cup I_p}\right\|\right),$$

and

$$\left|T_\alpha(G_{\lambda,\widehat{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\Theta}}))\right| \leq 2s - p. \tag{27}$$

Proof. Drawing inspiration from (Sun et al., 2018), our proof employs techniques that establish upper-bounded estimation error in regularized maximum likelihood estimation problems using Karush–Kuhn–Tucker (KKT) conditions. Notably, our formulation incorporates multiple constraints, contrasting the unconstrained problem in (Sun et al., 2018), which leads to distinct KKT conditions and error-bounding techniques.

Due to the sign constraint, $G_{\lambda,\widehat{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\Theta}})$ can be expressed as the minimizer of the following problem:

$$\underset{\boldsymbol{\Theta}\in\mathcal{M}^p}{\text{minimize}} - \log\det(\boldsymbol{\Theta}) + \operatorname{tr}\left(\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}\right) - \sum_{i\neq j} W_{ij}\Theta_{ij}, \tag{28}$$

where $W_{ij} = p_\lambda\left(\left|\widetilde{\Theta}_{ij}\right|\right)$ if $i \neq j$, and zero otherwise. Define a local region

$$\mathcal{B}\left(\boldsymbol{\Theta}^\star, \lambda_{\max}\left(\boldsymbol{\Theta}^\star\right)\right) = \{\boldsymbol{\Theta} \in \mathcal{M}^p \mid \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star\|_{\mathrm{F}} \leq \lambda_{\max}(\boldsymbol{\Theta}^\star\}.$$

For ease of presentation, we denote $G_{\lambda,\widehat{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\Theta}})$ by $\widehat{\boldsymbol{\Theta}}$. We first prove that $\widehat{\boldsymbol{\Theta}} \in \mathcal{B}\left(\boldsymbol{\Theta}^\star, \lambda_{\max}\left(\boldsymbol{\Theta}^\star\right)\right)$ under event $\mathcal{J}$ defined in (53). We construct an intermediate estimate:

$$\boldsymbol{\Theta}_t := \boldsymbol{\Theta}^\star + t\left(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right), \tag{29}$$

where t is taken such that $\|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}^\star\|_{\mathrm{F}} = \lambda_{\max}(\boldsymbol{\Theta}^\star)$ if $\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\|_{\mathrm{F}} > \lambda_{\max}\left(\boldsymbol{\Theta}^\star\right)$, and $t = 1$ otherwise. Hence $\|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}^\star\|_{\mathrm{F}} \leq \lambda_{\max}(\boldsymbol{\Theta}^\star)$ holds for any $t \in [0, 1]$. Then we apply Lemma B.1 with $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_t$, $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}^\star$, and get

$$\frac{c_0}{2}t\left\langle(\boldsymbol{\Theta}^\star)^{-1} - \boldsymbol{\Theta}_t^{-1}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\rangle \geq \|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}^\star\|_{\mathrm{F}}^2. \tag{30}$$

Construct a function $q(t) := -\log\det\left(\boldsymbol{\Theta}^\star + t(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star)\right) + t\langle(\boldsymbol{\Theta}^\star)^{-1}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\rangle$ and $t \in [0, 1]$. One has

$$q'(t) = \left\langle\left(\boldsymbol{\Theta}^\star\right)^{-1} - \boldsymbol{\Theta}_t^{-1}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\rangle, \tag{31}$$

Let $\boldsymbol{A} = \boldsymbol{\Theta}_t^{-1}$, and $\boldsymbol{B} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star$. Then one also has

$$q''(t) = \operatorname{tr}\left(\boldsymbol{ABAB}\right),$$

Next, we show that $q''(t) \geq 0$. Let $\boldsymbol{C} = \boldsymbol{AB}$. Following from Theorem 1 in (Drazin & Haynsworth, 1962), all eigenvalues of a matrix $\boldsymbol{X}$ are real if there exists a symmetric and positive definite matrix $\boldsymbol{Y}$ such that $\boldsymbol{XY}$ are symmetric. Since $\boldsymbol{CA}$ is symmetric and $\boldsymbol{A}$ symmetric and positive definite, all eigenvalues of $\boldsymbol{C}$ are real. Suppose $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{C}$. Then one obtains $q''(t) = \sum_{i=1}^p \lambda_i^2 \geq 0$, implying that $q'(t)$ is monotonically non-decreasing. Then one obtains

$$\left\langle\left(\boldsymbol{\Theta}^\star\right)^{-1} - \widehat{\boldsymbol{\Theta}}^{-1}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\rangle \geq q'(t) \geq \frac{2}{c_0 t}\|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}^\star\|_{\mathrm{F}}^2. \tag{32}$$

where the first inequality follows from $q'(1) \geq q'(t)$ and $t \leq 1$, and the second inequality follows from (30).

Following from (32), we have

$$\|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}^\star\|_{\mathrm{F}}^2 \leq \frac{c_0}{2}t\left\langle\left(\boldsymbol{\Theta}^\star\right)^{-1} - \widehat{\boldsymbol{\Theta}}^{-1}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\rangle. \tag{33}$$

We present the Lagrangian of the optimization (28) as follows:

$$L\left(\boldsymbol{\Theta}, \boldsymbol{\Gamma}\right) = -\log\det(\boldsymbol{\Theta}) + \operatorname{tr}\left(\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}\right) - \sum_{i\neq j} W_{ij}\Theta_{ij} + \langle\boldsymbol{\Gamma}, \boldsymbol{\Theta}\rangle,$$

where $\mathbf{\Gamma}$ is a KKT multiplier with $\Gamma_{ii} = 0$ for $i \in [p]$. Let $\big(\widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Gamma}}\big)$ be the primal and dual optimal point, which must satisfy the KKT conditions as follows:

$$-\widehat{\mathbf{\Theta}}^{-1} + \widehat{\mathbf{\Sigma}} - \boldsymbol{W} + \widehat{\mathbf{\Gamma}} = \mathbf{0}; \tag{34}$$

$$\widehat{\Theta}_{ij}\widehat{\Gamma}_{ij} = 0,\ \widehat{\Theta}_{ij} \le 0,\ \widehat{\Gamma}_{ij} \ge 0,\ \forall\, i \ne j; \tag{35}$$

$$\widehat{\Gamma}_{ii} = 0,\ \text{for } i \in [p]; \tag{36}$$

According to (34), one has

$$\big\langle -\widehat{\mathbf{\Theta}}^{-1} + \widehat{\mathbf{\Sigma}},\, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star \big\rangle = \big\langle \boldsymbol{W} - \widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star \big\rangle. \tag{37}$$

Plugging (37) into (33) yields

$$\|\mathbf{\Theta}_t - \mathbf{\Theta}^\star\|_{\mathrm{F}}^2 = \frac{c_0}{2} t \Big(\underbrace{\big\langle \widehat{\mathbf{\Gamma}}, \mathbf{\Theta}^\star - \widehat{\mathbf{\Theta}} \big\rangle}_{\text{term I}} + \underbrace{\big\langle \boldsymbol{W}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star \big\rangle}_{\text{term II}} + \underbrace{\big\langle \mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star \big\rangle}_{\text{term III}} \Big). \tag{38}$$

The term I in (38) can be bounded by

$$\big\langle \widehat{\mathbf{\Gamma}}, \mathbf{\Theta}^\star - \widehat{\mathbf{\Theta}} \big\rangle = \sum_{i \ne j} \widehat{\Gamma}_{ij} \Theta^\star_{ij} \le 0.$$

where the equality follows from (35) and (36), and the inequality follows from $\widehat{\Gamma}_{ij} \ge 0$ in (35) and the fact that $\Theta^\star_{ij} \le 0$ for any $i \ne j$.

To bound term II, we separate the support of $\boldsymbol{W}$ off the diagonal into two parts, $S^\star$ and $T^\star$ which is defined as follows:

$$T^\star = \big\{ (i,j) \,|\, \Theta^\star_{ij} = 0, i \ne j \big\}.$$

Let $I_p := \{(i,i) \,|\, i \in [p]\}$. Define

$$\bar{T}_\alpha(\widetilde{\mathbf{\Theta}}) = T_\alpha(\widetilde{\mathbf{\Theta}}) \cup I_p \quad \text{and} \quad \bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}}) = T^c_\alpha(\widetilde{\mathbf{\Theta}}) \setminus I_p. \tag{39}$$

Since $|T_\alpha(\widetilde{\mathbf{\Theta}})| \le 2s - p$, $|\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})| \le 2s$. Then one has

$$\begin{aligned} \big\langle \boldsymbol{W}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star \big\rangle &= \big\langle \boldsymbol{W}_{S^\star}, \big(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\big)_{S^\star} \big\rangle + \big\langle \boldsymbol{W}_{T^\star}, \widehat{\mathbf{\Theta}}_{T^\star} \big\rangle \\ &\le \|\boldsymbol{W}_{S^\star}\| \, \big\|\big(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\big)_{S^\star}\big\| + \big\langle \boldsymbol{W}_{\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})}, \widehat{\mathbf{\Theta}}_{\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})} \big\rangle, \end{aligned}$$

where the inequality follows from the fact that $W_{ij} \ge 0$, $\widehat{\Theta}_{ij} \le 0$ for $i \ne j$, and $\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}}) \subseteq T^\star$.

For term III, we separate the support of $\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)$ into parts, $\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})$ and $\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})$. Then one has

$$\big\langle \mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}},\, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star \big\rangle \le \big\langle \big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})}, \widehat{\mathbf{\Theta}}_{\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})} \big\rangle + \big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})}\big\| \big\|\big(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\big)_{\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})}\big\|.$$

We note that, for any $(i,j) \in \bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})$, $\widehat{\Theta}_{ij} \le 0$ and $W_{ij} \ge \frac{\sqrt{2}}{2}\lambda$, following from the definition of T_α in (24) and Assumption 4.1. Then under the event that $\big\|\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big\|_{\max} \le \frac{\sqrt{2}}{2}\lambda$, we have

$$\big\langle \big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}} + \boldsymbol{W}\big)_{\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})}, \widehat{\mathbf{\Theta}}_{\bar{T}^c_\alpha(\widetilde{\mathbf{\Theta}})} \big\rangle \le 0. \tag{40}$$

By bounding term I, term II, and term III in (38), together with (40), we obtain

$$\big\|\mathbf{\Theta}_t - \mathbf{\Theta}^\star\big\|_{\mathrm{F}}^2 \le \frac{c_0}{2} t \Big(\big\|\boldsymbol{W}_{S^\star}\big\| \big\|\big(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\big)_{S^\star}\big\| + \big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})}\big\| \big\|\big(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\big)_{\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})}\big\| \Big). \tag{41}$$

Through the definition of $\mathbf{\Theta}_t$ in (29), one has $\|\mathbf{\Theta}_t - \mathbf{\Theta}^\star\|_{\mathrm{F}} = t\big\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}}$. Then (41) can be written as follows:

$$\|\mathbf{\Theta}_t - \mathbf{\Theta}^\star\|_{\mathrm{F}} \le \frac{c_0}{2} \Big(\big\|\boldsymbol{W}_{S^\star}\big\| + \big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{\bar{T}_\alpha(\widetilde{\mathbf{\Theta}})}\big\| \Big). \tag{42}$$

Recall that $W_{ij} = p_\lambda\big(\big|\widetilde{\Theta}_{ij}\big|\big) \le \lambda$ for any $i \neq j$ according to Assumption 4.1, and $|S^\star| \le s - p$. Thus one has

$$\|\boldsymbol{W}_{S^\star}\| \le \sqrt{s}\lambda.$$

Under the event $\mathcal{J}$, one can bound

$$\big\|\big(\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}\big)_{\bar{T}_\alpha(\widetilde{\boldsymbol{\Theta}})}\big\| \le |\bar{T}_\alpha(\widetilde{\boldsymbol{\Theta}})|^{\frac{1}{2}} \big\|\big(\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}\big)_{\bar{T}_\alpha(\widetilde{\boldsymbol{\Theta}})}\big\|_{\max} \le \sqrt{s}\lambda, \tag{43}$$

where the second inequality follows from $|\bar{T}_\alpha(\widetilde{\boldsymbol{\Theta}})| \le 2s$.

Plugging (42) and (43) into (41) yields

$$\|\boldsymbol{\Theta}_t - \boldsymbol{\Theta}^\star\|_{\mathrm{F}} \le c_0\sqrt{s}\lambda \le \lambda_{\max}\left(\boldsymbol{\Theta}^\star\right),$$

which indicates that $t = 1$ in (29), *i.e.*, $\boldsymbol{\Theta}_t = \widehat{\boldsymbol{\Theta}}$. The last inequality is established by plugging $\lambda = \sqrt{c_1\tau\log p/n}$ with $n \ge 8c_0c_1\tau s\log p$. Therefore, we obtain

$$\big\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\big\|_{\mathrm{F}} \le c_0\sqrt{s}\lambda. \tag{44}$$

Putting $\boldsymbol{W}_{S^\star} = p_\lambda(|\widetilde{\boldsymbol{\Theta}}_{S^\star}|)$ and $\boldsymbol{\Theta}_t = G_{\lambda,\widehat{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\Theta}})$ into (42) gets

$$\big\|G_{\lambda,\widehat{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\Theta}}) - \boldsymbol{\Theta}^\star\big\|_{\mathrm{F}} \le \frac{c_0}{2}\big(\big\|p_\lambda\big(\big|\widetilde{\boldsymbol{\Theta}}_{S^\star}\big|\big)\big\| + \big\|\big(\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}\big)_{\bar{T}_\alpha(\widetilde{\boldsymbol{\Theta}})}\big\|\big).$$

To establish $\big|T_\alpha(\widehat{\boldsymbol{\Theta}})\big| \le 2s - p$, we separate the set $T_\alpha(\widehat{\boldsymbol{\Theta}})$ into two parts, $S^\star$ and $L_\alpha(\widehat{\boldsymbol{\Theta}})\backslash S^\star$. For any $(i,j) \in L_\alpha(\widehat{\boldsymbol{\Theta}})\backslash S^\star$, one has $\big|\widehat{\Theta}_{ij}\big| \ge \alpha$, and further obtains

$$\big|L_\alpha(\widehat{\boldsymbol{\Theta}})\backslash S^\star\big| \le \frac{\big\|\widehat{\boldsymbol{\Theta}}_{L_\alpha(\widehat{\boldsymbol{\Theta}})\backslash S^\star}\big\|^2}{\alpha^2} \le \frac{\big\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\big\|_{\mathrm{F}}^2}{\alpha^2} \le s,$$

where the last inequality follows from (44). Then we have

$$\big|T_\alpha(\widehat{\boldsymbol{\Theta}})\big| = \big|S^\star \cup L_\alpha(\widehat{\boldsymbol{\Theta}})\backslash S^\star\big| = \big|S^\star\big| + \big|L_\alpha(\widehat{\boldsymbol{\Theta}})\backslash S^\star\big| \le 2s - p,$$

completing the proof. □

B.2. Proof of Theorem 3.1

Proof. Our convergence analysis largely builds upon standard techniques for the projected gradient descent, as provided in (Beck, 2014). The main difference lies in the open constraint set in our formulation, which leads to an additional requirement regarding the positive definiteness of the matrix during stepsize selection, extending beyond the backtracking condition. Consequently, it is crucial to incorporate certain adaptations in comparison to the standard methods employed for analyzing the convergence of projected gradient descent.

Given $\boldsymbol{\Theta}^o \in \mathcal{S}$, define the lower level set of the objective f of Problem (5) as follows:

$$L_f := \{\boldsymbol{\Theta} \in \mathcal{S} \,|\, f(\boldsymbol{\Theta}) \le f(\boldsymbol{\Theta}^o)\}. \tag{45}$$

Following from (Cai et al., 2021, Lemma 3.5), there exists $m > 0$ such that $\boldsymbol{\Theta} \succeq m\boldsymbol{I}$ holds for any $\boldsymbol{\Theta} \in L_f$. The gradient of f is Lipschitz continuous with parameter $L = m^{-2}$ over L_f, since the Hessian is $\boldsymbol{\Theta}^{-1} \otimes \boldsymbol{\Theta}^{-1}$.

Suppose that the initial point of the sequence $\boldsymbol{\Theta}_0 \in L_f$. To guarantee this, we may consider a $\boldsymbol{\Theta}^o$ with $f(\boldsymbol{\Theta}^o)$ sufficiently large. Let $\boldsymbol{\Theta}_t(\eta) = \mathcal{P}_{\mathcal{S}_d}\big(\boldsymbol{\Theta}_t - \eta\nabla f(\boldsymbol{\Theta}_t)\big)$. At the t-th iteration, if $\boldsymbol{\Theta}_t \in L_f$, then there must exist $\gamma > 0$ such that for any $\eta \in (0,\gamma)$, $\boldsymbol{\Theta}_t(\eta)$ satisfies that $\boldsymbol{\Theta}_t(\eta) \in \mathbb{S}^p_{++}$. This is because $\boldsymbol{\Theta}_t$ is an interior point of $\mathbb{S}^p_{++}$, indicating that there exists $r > 0$, such that for any $\boldsymbol{\Theta}$ satisfying $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_t\|_{\mathrm{F}} < r$, $\boldsymbol{\Theta} \in \mathbb{S}^p_{++}$. Taking $\eta < \frac{r}{\|\nabla f(\boldsymbol{\Theta}_t)\|_{\mathrm{F}}}$, one has,

$$\|\boldsymbol{\Theta}_t(\eta) - \boldsymbol{\Theta}_t\|_{\mathrm{F}} = \|\mathcal{P}_{\mathcal{S}_d}\big(\boldsymbol{\Theta}_t - \eta\nabla f(\boldsymbol{\Theta}_t)\big) - \mathcal{P}_{\mathcal{S}_d}\big(\boldsymbol{\Theta}_t\big)\|_{\mathrm{F}} \le \eta\|\nabla f(\boldsymbol{\Theta}_t)\|_{\mathrm{F}} < r,$$

establishing that $\mathbf{\Theta}_t(\eta) \in \mathbb{S}^p_{++}$. The projection $\mathcal{P}_{\mathcal{S}_d}$ ensures $\mathbf{\Theta}_t(\eta) \in \mathcal{S}_d$. Totally, $\mathbf{\Theta}_t(\eta) \in \mathcal{S}$.

The line search condition (8) ensures that $f(\mathbf{\Theta}_t(\eta)) \leq f(\mathbf{\Theta}_t)$, implying that $\mathbf{\Theta}_t(\eta) \in L_f$. Then one has

$$f(\mathbf{\Theta}_t(\eta)) \leq f(\mathbf{\Theta}_t) + \langle \nabla f(\mathbf{\Theta}_t),\, \mathbf{\Theta}_t(\eta) - \mathbf{\Theta}_t \rangle + \frac{L}{2}\|\mathbf{\Theta}_t(\eta) - \mathbf{\Theta}_t\|_{\mathrm{F}}^2. \tag{46}$$

Following from projection theorem (Beck, 2014, Theorem 9.8), one has

$$\langle \mathbf{\Theta}_t - \eta \nabla f(\mathbf{\Theta}_t) - \mathbf{\Theta}_t(\eta),\, \mathbf{\Theta}_t - \mathbf{\Theta}_t(\eta) \rangle\,. \tag{47}$$

Together with (46), one has

$$f(\mathbf{\Theta}_t(\eta)) \leq f(\mathbf{\Theta}_t) - \eta\Big(1 - \frac{\eta L}{2}\Big)\big\|G_{\frac{1}{\eta}}(\mathbf{\Theta}_t)\big\|_{\mathrm{F}}^2. \tag{48}$$

Taking $\eta \leq \frac{2(1-\alpha)}{L}$ leads to $1 - \frac{\eta L}{2} \geq \alpha$. Therefore, for any $\eta \leq \min\big(\frac{2(1-\alpha)}{L}, \gamma\big)$, $\mathbf{\Theta}_t(\eta)$ can simultaneously satisfy $\mathbf{\Theta}_t(\eta) \in \mathcal{S}$ and the line search condition (8). Thus, the step size in line search has a lower bound $\eta_t \geq \min\big(\frac{2(1-\alpha)\beta}{L}, \gamma, \sigma\big)$, and $\mathbf{\Theta}_{t+1} \in L_f$. By induction, $\mathbf{\Theta}_t \in L_f$ for any $t \geq 0$. Sequence $\{f(\mathbf{\Theta}_t)\}$ is monotonically decreasing, and $f(\mathbf{\Theta}_{t+1}) < f(\mathbf{\Theta}_t)$ until $G_{\frac{1}{\eta_t}}(\mathbf{\Theta}_t) = \mathbf{0}$, indicating that $\mathbf{\Theta}_t$ is a stationary point (Beck, 2014, Theorem 9.10). Since Problem (5) is a strictly convex problem, the stationary point is the unique minimizer. Therefore, the proposed algorithm converges to the optimal solution. □

B.3. Proof of Proposition 3.3

Proof. The Lagrangian of the optimization (18) is

$$L(\boldsymbol{x}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 - \mu_r \boldsymbol{x}^\top \mathbf{1} + \langle \boldsymbol{\mu}_{\setminus r},\, \boldsymbol{x}_{\setminus r} \rangle,$$

where $\boldsymbol{\mu}$ is a KKT multiplier. Let $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\mu}})$ be the primal and dual optimal point. Then $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\mu}})$ must satisfy the KKT system:

$$\hat{x}_i - y_i - \hat{\mu}_r + \hat{\mu}_i = 0,\ \hat{\mu}_i \hat{x}_i = 0, \quad \text{for } i \neq r; \tag{49}$$

$$\hat{x}_i - y_i - \hat{\mu}_i = 0,\ \hat{\mu}_i \hat{\boldsymbol{x}}^\top \mathbf{1} = 0, \quad \text{for } i = r; \tag{50}$$

$$\hat{\mu}_1, \ldots, \hat{\mu}_p \geq 0; \tag{51}$$

Therefore, for any $i \neq r$, it holds that $\hat{x}_i = y_i + \hat{\mu}_r - \hat{\mu}_i$, $\hat{\mu}_i \hat{x}_i = 0$, $\hat{\mu}_i \geq 0$ and $\hat{x}_i \leq 0$. As a result, we obtain the following results:

- If $y_i + \hat{\mu}_r \leq 0$, then $\hat{\mu}_i = 0$, indicating that $\hat{x}_i = y_i + \hat{\mu}_i$.
- If $y_i + \hat{\mu}_r > 0$, then $\hat{\mu}_i = y_i + \hat{\mu}_r$, indicating that $\hat{x}_i = 0$.

Overall, we obtain that

$$\hat{x}_i = \min(y_i + \hat{\mu}_r,\, 0), \quad \forall\, i \neq r. \tag{52}$$

On the other hand, $\hat{x}_r$ and $\hat{\mu}_r$ satisfy that $\hat{x}_r = y_r + \hat{\mu}_r$, $\hat{\mu}_r \hat{\boldsymbol{x}}^\top \mathbf{1} = 0$, $\hat{\mu}_r \geq 0$, and $\hat{\boldsymbol{x}}^\top \mathbf{1} \geq 0$. Then we have the following results:

- If $y_r \geq -\sum_{i \in [p]\setminus r} \min(y_i,\, 0)$, then $\hat{\mu}_r = 0$, indicating that $\hat{x}_r = y_r$. Following from (52), $\hat{x}_i = \min(y_i,\, 0)$ for any $i \neq r$.
- If $y_r < -\sum_{i \in [p]\setminus r} \min(y_i,\, 0)$, then $\hat{\mu}_r \neq 0$. This is because $\hat{\mu}_r = 0$ will result in $\hat{\boldsymbol{x}}^\top \mathbf{1} < 0$, which does not fulfill the KKT system. Together with the KKT condition that $\hat{\mu}_r \hat{\boldsymbol{x}}^\top \mathbf{1} = 0$, one has $\hat{\boldsymbol{x}}^\top \mathbf{1} = 0$. Therefore, one obtains that $\hat{x}_r = y_r + \hat{\mu}_r$, where $\hat{\mu}_r$ satisfies $\hat{\mu}_r + y_r + \sum_{i \in [p]\setminus r} \min(y_i + \hat{\mu}_r,\, 0) = 0$.

Then the ρ in Proposition 3.3 is the dual optimal point $\hat{\mu}_r$ above. □

B.4. Proofs of Theorem 4.4 and Corollary 4.5

Proof. We prove the theorem under the case of estimating M-matrices, *i.e.*, $\mathcal{S} = \mathcal{M}^p$. The proof can be directly applied to the case of estimating diagonally dominant M-matrices, *i.e.*, $\mathcal{S} = \mathcal{M}_D^p$, since $\mathcal{M}_D^p$ is a subset of $\mathcal{M}^p$. Provided that the following event holds:

$$\mathcal{J} := \Big\{ \big\|\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big\|_{\max} \le \frac{\sqrt{2}}{2}\lambda \Big\}. \tag{53}$$

Recall that $\widehat{\mathbf{\Theta}}^{(k)} = G_{\lambda,\widehat{\mathbf{\Sigma}}}\big(\widehat{\mathbf{\Theta}}^{(k-1)}\big)$ for $k \ge 2$. According to Assumption 4.1 that $p_\lambda(0) = \lambda$, we can rewrite $\widehat{\mathbf{\Theta}}^{(1)}$ as $\widehat{\mathbf{\Theta}}^{(1)} = G_{\lambda,\widehat{\mathbf{\Sigma}}}\big(\widehat{\mathbf{\Theta}}^{(0)}\big)$, where $\widehat{\Theta}_{ij}^{(0)} = 0$ for any $i \ne j$. Thus, $\big|T_\alpha\big(\widehat{\mathbf{\Theta}}^{(0)}\big)\big| \le 2s$. Following from Lemma B.2, we can further obtain that $\big|T_\alpha\big(\widehat{\mathbf{\Theta}}^{(1)}\big)\big| \le 2s$. To simplify notation, we denote $T_\alpha\big(\widehat{\mathbf{\Theta}}^{(k)}\big)$ by T_α^k for short. By induction, we obtain that, for any $k \ge 1$,

$$\big|T_\alpha\big(\widehat{\mathbf{\Theta}}^{(k)}\big)\big| \le 2s, \tag{54}$$

and

$$\big\|\widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}} \le \frac{c_0}{2}\big(\delta_1^{k-1} + \delta_2^{k-1}\big), \tag{55}$$

where $\delta_1^{k-1} = \big\|p_\lambda\big(\big|\widehat{\mathbf{\Theta}}_{S^\star}^{(k-1)}\big|\big)\big\|$ and $\delta_2^{k-1} = \big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{T_\alpha^{k-1}\cup I_p}\big\|$.

In what follows, we show that the term δ_1^{k-1} in (55) can be bounded in terms of $\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|$. For any $(i,j) \in S^\star$ and $\mathbf{\Theta} \in \mathbb{R}^{p\times p}$, if $\big|\Theta_{ij}^\star - \Theta_{ij}\big| \ge \alpha$, then one has

$$0 \le p_\lambda\left(|\Theta_{ij}|\right) \le \lambda \le \frac{\lambda}{\alpha}\left|\Theta_{ij}^\star - \Theta_{ij}\right|,$$

where the first two inequalities follows from Assumption 4.1. If $\big|\Theta_{ij}^\star - \Theta_{ij}\big| \le \alpha$, then one has

$$0 \le p_\lambda\left(|\Theta_{ij}|\right) \le p_\lambda\left(\big|\Theta_{ij}^\star\big| - \alpha\right) = 0, \tag{56}$$

where the second inequality follows from Assumption 4.1 that p_λ is monotonically non-increasing, and the equality is established because $\min_{(i,j)\in S^\star}\big|\Theta_{ij}^\star\big| - \alpha \ge \gamma\lambda$ and $p_\lambda(x) = 0$ for any $x \ge \gamma\lambda$ following from Assumptions 4.1 and 4.2. As a result, one can obtain

$$\delta_1^{k-1} \le \frac{\lambda}{\alpha}\big\|\big(\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big)_{S^\star}\big\| \le \frac{\lambda}{\alpha}\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}}. \tag{57}$$

To bound δ_2^{k-1}, we separate the set T_α^{k-1} into two parts, $S^\star$ and $L_\alpha^{k-1}\backslash S^\star$, where L_α^k denotes the set $L_\alpha\big(\widehat{\mathbf{\Theta}}^{(k)}\big)$. The term $\big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{L_\alpha^{k-1}\backslash S^\star}\big\|$ can be bounded in terms of $\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}}$ as follows:

$$\big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{L_\alpha^{k-1}\backslash S^\star}\big\| \le \sqrt{\big|L_\alpha^{k-1}\backslash S^\star\big|}\big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{L_\alpha^{k-1}\backslash S^\star}\big\|_{\max} \le \frac{\sqrt{2}\lambda}{2\alpha}\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}},$$

where the last second equality follows from

$$\sqrt{\big|L_\alpha^{k-1}\backslash S^\star\big|} \le \frac{\big\|\widehat{\mathbf{\Theta}}^{(k-1)}_{L_\alpha^{k-1}\backslash S^\star}\big\|}{\alpha} \le \frac{\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}}}{\alpha}, \tag{58}$$

where the first inequality follows from the definition of L_α^{k-1}. Thus one has

$$\delta_2^{k-1} \le \big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{S^\star\cup I_p}\big\| + \frac{\sqrt{2}\lambda}{2\alpha}\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}}. \tag{59}$$

Substituting (57) and (59) into (55) yields

$$\big\|\widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}} \le \frac{c_0}{2}\big\|\big(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}}\big)_{S^\star\cup I_p}\big\| + \rho\big\|\widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}^\star\big\|_{\mathrm{F}},$$

where $\rho = \frac{2+\sqrt{2}}{4}$. By induction, if $d_k \leq a_0 + \rho d_{k-1}$ for any $k \geq 1$ with $\rho \in [0, 1)$, then

$$d_k \leq \frac{1-\rho^k}{1-\rho} a_0 + \rho^k d_0. \tag{60}$$

Applying (60) with $a_0 = \frac{c_0}{2} \left\| \left(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}} \right)_{S^\star \cup I_p} \right\|$, and $d_k = \left\| \widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star \right\|_{\mathrm{F}}$ yields

$$\left\| \widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star \right\|_{\mathrm{F}} \leq 4c_0 \left\| \left(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}} \right)_{S^\star \cup I_p} \right\| + \rho^k \left\| \widehat{\mathbf{\Theta}}^{(0)} - \mathbf{\Theta}^\star \right\|_{\mathrm{F}}.$$

Next, we establish Corollary 4.5. Under the event $\mathcal{J}$, one has

$$\left\| \left(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}} \right)_{S^\star \cup I_p} \right\| \leq \frac{\sqrt{2}}{2} \lambda \sqrt{s}, \tag{61}$$

where the inequality follows from $|S^\star \cup I_p| \leq s$. As a result, we plug $\lambda = \sqrt{c_1 \tau \log p / n}$ into (61) and obtain

$$\left\| \widehat{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star \right\|_{\mathrm{F}} \leq c \sqrt{s \log p / n} + \rho^k \left\| \widehat{\mathbf{\Theta}}^{(0)} - \mathbf{\Theta}^\star \right\|_{\mathrm{F}}, \tag{62}$$

where $c = 2c_0 \sqrt{2c_1 \tau}$.

Finally, we calculate the probability that event $\mathcal{J}$ holds. We apply (Ravikumar et al., 2011, Lemma 1) and union sum bound, and obtain

$$\mathbb{P}\left[\left\| \left(\mathbf{\Theta}^\star \right)^{-1} - \widehat{\mathbf{\Sigma}} \right\|_{\max} \geq \frac{\sqrt{2}}{2} \lambda \right] \leq 4p^2 \exp\left(- \frac{n\lambda^2}{\left(80 \max_i \Sigma_{ii}^\star \right)^2} \right).$$

Take $\lambda = \sqrt{c_1 \tau \log p / n}$ with $\tau > 2$ and $c_1 = \left(80 \max_i \Sigma_{ii}^\star \right)^2$. By calculation, we obtain that

$$\mathbb{P}\left[\left\| \left(\mathbf{\Theta}^\star \right)^{-1} - \widehat{\mathbf{\Sigma}} \right\|_{\max} \leq \frac{\sqrt{2}}{2} \lambda \right] \geq 1 - 4/p^{\tau - 2}, \tag{63}$$

completing the proof. □