Positive Definite Sparse Covariance Estimation via Dual Space Optimization

Fengpei Li*, Wenfu Xia*, Ziping Zhao†

ShanghaiTech University
{fengpeili, wenfuxia, zipingzhao}@shanghaitech.edu.cn

Abstract

Covariance matrix estimation in high dimensions is a fundamental problem in machine learning and signal processing. A common structural assumption used to mitigate the challenges posed by high dimensionality is sparsity, which posits that most variable pairs exhibit negligible correlations. In this paper, we revisit the classical problem of positive definite sparse covariance estimation (PDSCE) introduced by Rothman (2012). Unlike many earlier approaches, this formulation incorporates a logarithmic barrier, which guarantees that the resulting covariance estimator is positive definite and thereby ensures the well-posedness of the estimation problem. However, the inclusion of the logarithmic barrier also leads to nontrivial optimization difficulties. To overcome these difficulties, we propose a dual proximal gradient method (DPGM) for solving the PDSCE problem. In contrast to existing primal-space approaches, DPGM operates directly in the dual space. This dual perspective provides several key advantages. First, DPGM significantly reduces computational costs, because positive definiteness is preserved automatically and no iterative subproblem solvers are required. Second, compared with primal optimization algorithms, DPGM offers stronger theoretical guarantees, including principled step size selection and improved iteration complexity. Extensive numerical experiments demonstrate that DPGM consistently outperforms existing methods, which confirms its effectiveness and scalability for high-dimensional sparse covariance estimation.

1 Introduction

The estimation of covariance matrices lies at the core of numerous fundamental problems in modern multivariate data analysis, with broad applications in machine learning (Jolliffe 2002), finance (Ledoit and Wolf 2003), and biology (Schäfer and Strimmer 2005). For example, in machine learning, they are prerequisites for dimensionality reduction methods such as principal component analysis (Jolliffe 2002) and for classification techniques like linear and quadratic discriminant analysis (Witten and Tibshirani 2009; Xiong et al. 2018); in finance, covariance matrices are essential for portfolio optimization to manage risk and allocate assets (Markowitz 1952; Zhao and Palomar 2018; Zhao, Zhou, and Palomar 2019); in biology, covariance matrices are used to infer large-scale gene association networks (Faust and Raes 2012). However, in high-dimensional settings where the problem dimension far exceeds the sample size, covariance matrix estimation becomes particularly challenging. A commonly used estimator is the sample covariance matrix. However, when the problem dimension and the sample size grow proportionally, or even when the dimensionality exceeds the number of samples, the sample covariance matrix is no longer a consistent estimator of the population covariance matrix. As a result, reliance on the sample covariance matrix can severely degrade the performance of downstream tasks. For instance, in principal component analysis, it may overestimate the importance of certain components due to inaccurate eigenvalue estimates (Karoui 2008b). These limitations have spurred intense research interest in high-dimensional covariance estimation in recent years (Zhao and Liu 2013; Fan, Liao, and Liu 2016; Donoho, Gavish, and Johnstone 2018; Yan, Yang, and Zhao 2025).

To effectively estimate high-dimensional covariance matrices, a widely adopted strategy is to impose structural assumptions. Sparsity is one of the most commonly used approaches, wherein many of the entries are assumed to be zero (Bickel and Levina 2008). This reduces the effective number of parameters and improves statistical convergence rates. A common method for sparse covariance estimation is thresholding, where small entries of the sample covariance matrix are set to zero (Karoui 2008a; Rothman, Levina, and Zhu 2009). Despite possessing desirable theoretical properties, including minimax optimality and fast statistical convergence rates, these estimators generally lack guaranteed positive definiteness. To simultaneously enforce positive definiteness and sparsity, many estimation problems have been proposed. In this paper, we revisit the positive definite sparse covariance estimation (PDSCE) problem in Rothman (2012):

$$\min_{\boldsymbol{\Sigma}\in\mathbb{S}^d_{++}} \quad \frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{S}\right\|_{\mathsf{F}}^2 - \tau\log\det\boldsymbol{\Sigma} + \left\|\boldsymbol{W}\circ\boldsymbol{\Sigma}\right\|_1, \quad (1)$$

where $\boldsymbol{S}$ is the sample covariance matrix, $-\tau\log\det(\cdot)$ is the logarithmic barrier with parameter $\tau\geq 0$, and $\|\boldsymbol{W}\circ\cdot\|_1$ is the weighted ℓ_1-norm with $\boldsymbol{W}$ being a nonnegative weight

*These authors contributed equally.
†Corresponding author.

matrix and $\circ$ denoting the element-wise product. Alternative approaches, such as those proposed by Xue, Ma, and Zou (2012) and Liu, Wang, and Zhao (2014), enforce positive definiteness through explicit eigenvalue constraints. However, these formulations introduce two nonsmooth terms, making them less compelling than Problem (1) from an optimization standpoint.

To solve the PDSCE problem, several algorithms have been proposed. Rothman (2012) introduced a row-by-row block coordinate minimization method (BCM) that exploits the structure of symmetric positive definite matrices by updating one row (and the corresponding column) at a time. However, each subproblem requires solving a Lasso problem (Tibshirani 1996) via an inner loop, resulting in a computationally expensive double-loop algorithm. A further limitation of BCM is that it does not necessarily guarantee the iterates remain positive definite unless the algorithm converges. In addition, the iteration complexity of BCM scales linearly with the problem dimension d, requiring $\mathcal{O}(d \log 1/\epsilon)$ iterations to reach an ϵ-stationary point, which becomes prohibitively slow in high-dimensional settings (Li et al. 2018). Kyrillidis et al. (2014) proposed an inexact proximal Newton method (PNM) that leverages local Hessian information and achieves a local quadratic convergence rate. While theoretically appealing, PNM suffers from two major practical drawbacks. First, it requires damped proximal Newton steps to ensure positive definiteness and to compensate for the lack of global Lipschitz smoothness, which often leads to overly conservative updates. Second, like BCM, it adopts a double-loop structure that necessitates solving a Lasso problem. These computational bottlenecks significantly limit its practical efficiency.

Recently, Wei and Zhao (2023) developed a proximal gradient method (PGM) for the PDSCE problem, which requires careful step-size selection to ensure positive definiteness. However, checking positive definiteness in each line search relies on the Cholesky decomposition, which incurs substantial computational overhead. Furthermore, this work does not provide further convergence analysis. To bridge this gap, in this paper, we first provide a complete analysis of the PGM algorithm, offering guidelines for step-size selection to ensure positive definiteness and establishing its iteration complexity. However, the theoretical step size is overly conservative, which significantly hinders the practical efficiency of PGM.

Motivated by the limitations of the algorithms discussed above, we develop an efficient algorithm for the PDSCE problem. The core innovation of our approach lies in applying the proximal gradient method to the dual formulation of Problem (1). The proposed algorithm offers several distinct advantages over existing methods. Its computational cost is lower than that of current algorithms, as it naturally generates positive definite iterates of $\boldsymbol{\Sigma}$ at each iteration without additional overhead, and no subproblems need to be solved iteratively. In addition, the method provides theoretical benefits through an effective step size that eliminates the need for backtracking line searches. We conduct extensive experiments on both synthetic and real-world datasets, which demonstrate the superior efficiency of our algorithm compared to state-of-the-art methods and confirm its effectiveness for high-dimensional problems.

2 Preliminaries

Notations

Given a d-dimensional symmetric matrix $\boldsymbol{X}$, its eigendecomposition is $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with ordered eigenvalues $\lambda_1(\boldsymbol{X}) = \Lambda_{1,1} \geq \cdots \geq \lambda_d(\boldsymbol{X}) = \Lambda_{d,d}$ and $\boldsymbol{V}$ is an orthogonal matrix of eigenvectors. We define $\|\boldsymbol{X}\|_{\mathsf{F}} = \sqrt{\sum_{i,j} X_{ij}^2}$ as the Frobenius norm, $\|\boldsymbol{X}\|_1 = \sum_{ij} |X_{ij}|$ as the element-wise ℓ_1-norm, and $\|\boldsymbol{X}\|_\infty = \max_{ij} |X_{ij}|$ as the element-wise ℓ_∞-norm. We denote by $\boldsymbol{I}$ the identity matrix, by $\otimes$ the Kronecker product, and by $\circ$ the Hadamard product. $\mathbb{S}^d_{++}$ denotes the set of symmetric positive definite matrices of dimension d, $\mathbb{S}^d$ denotes the set of symmetric matrices of dimension d, and $\mathbb{R}$ denotes the set of real numbers.

Lipschitz smoothness and strong convexity

A differentiable convex function f is called L-Lipschitz smooth if, for all $\boldsymbol{X}, \boldsymbol{Y}$,

$$f(\boldsymbol{Y}) \leq f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{Y} - \boldsymbol{X} \rangle + \frac{L}{2} \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathsf{F}}^2,$$

and is called μ-strongly convex if

$$f(\boldsymbol{Y}) \geq f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{Y} - \boldsymbol{X} \rangle + \frac{\mu}{2} \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathsf{F}}^2.$$

3 Eigenvalue bounds for the PDSCE solution

We first establish eigenvalue bounds for the solution to the PDSCE problem (1) in the following theorem.

Theorem 1. *The solution to the PDSCE problem* (1)*, $\widehat{\boldsymbol{\Sigma}}$, is unique and satisfies $\alpha \boldsymbol{I} \preceq \widehat{\boldsymbol{\Sigma}} \preceq \beta \boldsymbol{I}$, where the universal constants α and β are defined as*

$$\alpha = \frac{\lambda_d(\boldsymbol{S}) - \|\boldsymbol{W}\|_\infty d + \sqrt{(\lambda_d(\boldsymbol{S}) - \|\boldsymbol{W}\|_\infty d)^2 + 4\tau}}{2},$$

$$\beta = \frac{\lambda_1(\boldsymbol{S}) + \|\boldsymbol{W}\|_\infty d + \sqrt{(\lambda_1(\boldsymbol{S}) + \|\boldsymbol{W}\|_\infty d)^2 + 4\tau}}{2}.$$

To the best of our knowledge, Theorem 1 presents a new spectral characterization that has not been documented in existing literature, and it will serve as a cornerstone of the following algorithm analysis.

4 Existing Algorithms for PDSCE

Before introducing our proposed method, we review three representative algorithms, namely BCM, PNM, and PGM, for solving the PDSCE problem (1), with a focus on their algorithmic structures and limitations. BCM was introduced in the seminal work (Rothman 2012) and exhibits several critical limitations that make it unsuitable for large-scale problems. First, it does not guarantee the positive definiteness of the iterates, which is essential for producing a valid covariance matrix. Second, its iteration complexity scales linearly

Method	Reference	Iteration Complexity	Single Loop	Positive Definiteness Guarantee
BCM	Rothman (2012)	$\mathcal{O}(dL^2\mu^{-2}\log(1/\epsilon))$	✗	✗ prior to convergence
PNM	Kyrillidis et al. (2014)	$\mathcal{O}(\log\log(1/\epsilon))$ (local)	✗	✓ via damped step size
PGM	Wei and Zhao (2023)	$\mathcal{O}\left(L\frac{(\beta+\sqrt{d}(\alpha+\beta))^2}{(\beta+\sqrt{d}(\alpha+\beta))^2+\tau}\log(1/\epsilon)\right)$	✓	✓ via line-search step size
DPGM	This Paper	$\mathcal{O}(L\mu^{-1}\log(1/\epsilon))$	✓	✓ naturally

We define $L=\frac{\alpha^2+\tau}{\alpha^2}$ and $\mu=\frac{\beta^2+\tau}{\beta^2}$ as the locally smooth and strongly convex parameters of Problem (1).

Table 1: Comparison of existing optimization methods for the PDSCE problem.

with the ambient dimension d. Third, its double-loop architecture incurs substantial computational overhead in high-dimensional settings. A summary of these properties is provided in Table 1.

In what follows, we primarily introduce the PNM (Kyrillidis et al. 2014) and PGM (Tran-Dinh, Kyrillidis, and Cevher 2015; Wei and Zhao 2023) methods for the PDSCE problem. Both methods exploit the composite structure of the objective function, which consists of a smooth term f and a non-smooth term g. We define the functions $f, g : \mathbb{S}^d_{++} \to \mathbb{R}$ by $f(\boldsymbol{\Sigma}) \coloneqq \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_{\mathsf{F}}^2 - \tau\log\det\boldsymbol{\Sigma}$ and $g(\boldsymbol{\Sigma}) = \|\boldsymbol{W}\circ\boldsymbol{\Sigma}\|_1$. Specifically, at the $(t+1)$-th iteration, the updates of both methods can be formulated as the solution to the following subproblem:

$$\boldsymbol{\Sigma}_{t+1} = \arg\min_{\boldsymbol{\Sigma}\in\mathbb{S}^d_{++}} \{u(\boldsymbol{\Sigma},\boldsymbol{\Sigma}_t)+g(\boldsymbol{\Sigma})\}, \tag{2}$$

where $u(\boldsymbol{\Sigma},\boldsymbol{\Sigma}_t)$ is the quadratic approximation of $f(\boldsymbol{\Sigma})$ at $\boldsymbol{\Sigma}_t$, defined by

$$u(\boldsymbol{\Sigma},\boldsymbol{\Sigma}_t) \coloneqq f(\boldsymbol{\Sigma}_t) + \langle\nabla f(\boldsymbol{\Sigma}_t), \boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t\rangle + \frac{1}{2}\left\langle \operatorname{vec}(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t), \boldsymbol{M}_t \operatorname{vec}(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_t)\right\rangle,$$

where $\nabla f(\boldsymbol{\Sigma}_t) = \boldsymbol{\Sigma}_t - \boldsymbol{S} - \tau\boldsymbol{\Sigma}_t^{-1}$ and $\boldsymbol{M}_t$ is a symmetric positive semidefinite matrix that determines the second-order information of the surrogate.

Proximal Newton Method

PNM (Kyrillidis et al. 2014) incorporates second-order information through the Hessian matrix:

$$\boldsymbol{M}_t = \nabla^2 f(\boldsymbol{\Sigma}_t) = \boldsymbol{I} + \tau\boldsymbol{\Sigma}_t^{-1}\otimes\boldsymbol{\Sigma}_t^{-1}.$$

Since f is not globally Lipschitz smooth, and to ensure positive definiteness of $\boldsymbol{\Sigma}_{t+1}$, Kyrillidis et al. (2014) proposed a damped update scheme consisting of two phases. First, an intermediate solution is computed by solving the following problem:

$$\boldsymbol{\Sigma}_{t+\frac{1}{2}} = \arg\min_{\boldsymbol{\Sigma}\in\mathbb{S}^d_{++}} \{u(\boldsymbol{\Sigma},\boldsymbol{\Sigma}_t)+g(\boldsymbol{\Sigma})\}. \tag{3}$$

Then, a damped update is performed using the rule

$$\boldsymbol{\Sigma}_{t+1} = (1-\gamma_t)\boldsymbol{\Sigma}_t + \gamma_t\boldsymbol{\Sigma}_{t+\frac{1}{2}}, \tag{4}$$

where $\gamma_t\in(0,1)$ is a relaxation parameter adaptively determined based on both $\boldsymbol{\Sigma}_{t+\frac{1}{2}}$ and $\boldsymbol{\Sigma}_t$.

In practice, the overall performance often falls short of expectations due to several key factors. First, the surrogate problem (3) lacks a closed-form solution and must be solved using iterative methods such as FISTA (Beck and Teboulle 2009), resulting in a double-loop algorithmic structure. Second, the damped update often leads to overly conservative step sizes: when $\boldsymbol{\Sigma}_{t+\frac{1}{2}}$ deviates significantly from $\boldsymbol{\Sigma}_t$, the relaxation parameter γ_t becomes very small, causing $\boldsymbol{\Sigma}_{t+1}$ to remain close to $\boldsymbol{\Sigma}_t$. Third, although Kyrillidis et al. (2014) established the convergence rate of PNM without assuming global Lipschitz smoothness by utilizing the self-concordance property of f, the locally quadratic convergence rate is typically attained only within a small unknown region of the optimum.

Proximal Gradient Method

The earliest use of the PGM to solve the PDSCE problem appears in Tran-Dinh, Kyrillidis, and Cevher (2015), where the two updates in (3) and (4) are retained, except that the matrix $\boldsymbol{M}_t$ is set to $\boldsymbol{M}_t = \frac{1}{\eta_t}\boldsymbol{I}$, with η_t being a step size. However, the damped-type update in (4) often leads to performance degradation in practice. Recently, Wei and Zhao (2023) proposed an alternative PGM method that eliminates the relaxation step (4). The step size η_t is selected via backtracking line search to ensure both the sufficient decrease condition

$$f(\boldsymbol{\Sigma}_{t+1}) \le u(\boldsymbol{\Sigma}_{t+1},\boldsymbol{\Sigma}_t),$$

and the positive definiteness of $\boldsymbol{\Sigma}_{t+1}$. This algorithm has been shown to achieve better empirical performance compared to the method in Tran-Dinh, Kyrillidis, and Cevher (2015). In this paper, we establish the convergence of the PGM algorithm proposed in Wei and Zhao (2023). As a preliminary step, the following lemma shows that while f is neither globally Lipschitz smooth nor strongly convex over $\mathbb{S}^d_{++}$, both properties hold locally over any compact subset.

Lemma 2. *The function $f(\boldsymbol{\Sigma})$ is $(1+\frac{\tau}{a^2})$-smooth and $(1+\frac{\tau}{b^2})$-strongly convex over the set*

$$\mathcal{C}_\Sigma(a,b) = \left\{\boldsymbol{\Sigma}\in\mathbb{S}^d_{++} \,\middle|\, a\boldsymbol{I}\preceq\boldsymbol{\Sigma}\preceq b\boldsymbol{I}\right\},$$

where $0<a<b$. Specifically, for any $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\in\mathcal{C}_\Sigma(a,b)$, the following inequalities hold:

$$\|\nabla f(\boldsymbol{\Sigma}_1)-\nabla f(\boldsymbol{\Sigma}_2)\|_{\mathsf{F}} \le \left(1+\frac{\tau}{a^2}\right)\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_{\mathsf{F}},$$
$$\|\nabla f(\boldsymbol{\Sigma}_1)-\nabla f(\boldsymbol{\Sigma}_2)\|_{\mathsf{F}} \ge \left(1+\frac{\tau}{b^2}\right)\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_{\mathsf{F}}.$$

To establish the iteration complexity of the PGM algorithm, we select parameters a and b such that $\boldsymbol{\Sigma}_t \in \mathcal{C}_\Sigma(a,b)$ for all iterations t. This ensures that the Lipschitz smoothness and strong convexity properties of f remain well-controlled throughout the optimization process. Furthermore, Theorem 1 guarantees that the optimal solution satisfies $\widehat{\boldsymbol{\Sigma}} \in \mathcal{C}_\Sigma(\alpha,\beta)$, which necessitates the condition $\mathcal{C}_\Sigma(\alpha,\beta) \subseteq \mathcal{C}_\Sigma(a,b)$ to ensure consistency between the iterates and the optimal solution. We are now ready to present the convergence result.

Theorem 3. *Let α and β be the constants defined in Theorem 1. Suppose the step size satisfies $\eta_t \leq \frac{\alpha^2}{\tau+\alpha^2}$. Then all iterates generated by PGM remain within the set $\mathcal{C}_\Sigma(\alpha,\ \beta+\sqrt{d}(\alpha+\beta))$. Moreover, for any $t \geq 0$, let $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_{t+1}$ be two successive iterates generated by the update in* (2)*, we have:*

$$\left\|\boldsymbol{\Sigma}_{t+1}-\widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{F}} \leq \sqrt{1-\frac{\alpha^2}{\alpha^2+\tau}\cdot\frac{(\beta+\sqrt{d}(\alpha+\beta))^2+\tau}{(\beta+\sqrt{d}(\alpha+\beta))^2}} \cdot \left\|\boldsymbol{\Sigma}_t-\widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{F}},$$

with $\eta_t = \frac{\alpha^2}{\tau+\alpha^2}$, leading to an iteration complexity of

$$\mathcal{O}\left(\frac{\alpha^2+\tau}{\alpha^2}\cdot\frac{(\beta+\sqrt{d}(\alpha+\beta))^2}{(\beta+\sqrt{d}(\alpha+\beta))^2+\tau}\cdot\log\frac{1}{\epsilon}\right).$$

Theorem 3 establishes the existence of a step size that guarantees positive definiteness and allows the derivation of the corresponding iteration complexity. In practice, however, this step size is overly conservative. Using this step size in PGM results in very slow convergence.

5 Proposed Method: DPGM

In this section, we propose a new method for solving the PDSCE problem (1), which is based on applying the PGM in the dual space. We begin by deriving the dual formulation of Problem (1). To this end, we introduce an auxiliary variable $\boldsymbol{\Psi}$ and reformulate the original problem into the following linearly constrained convex program:

$$\begin{aligned} \min_{\boldsymbol{\Sigma}\in\mathbb{S}^d_{++},\boldsymbol{\Psi}\in\mathbb{S}^d} \quad & f(\boldsymbol{\Sigma})+g(\boldsymbol{\Psi}) \\ \text{s.t.} \quad & \boldsymbol{\Sigma}=\boldsymbol{\Psi}. \end{aligned} \tag{5}$$

The Lagrangian associated with the reformulated problem is given by

$$\mathcal{L}(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}) = f(\boldsymbol{\Sigma})+g(\boldsymbol{\Psi})+\langle\boldsymbol{\Gamma},\boldsymbol{\Sigma}-\boldsymbol{\Psi}\rangle.$$

where $\boldsymbol{\Gamma}\in\mathbb{S}^d$ denotes the Lagrange multiplier associated with the constraint $\boldsymbol{\Sigma}=\boldsymbol{\Psi}$.

To derive the dual function, we minimize $\mathcal{L}$ with respect to the primal variables $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$. The minimization over $\boldsymbol{\Sigma}$ admits the following solution

$$\arg\min_{\boldsymbol{\Sigma}\in\mathbb{S}^d_{++}}\{f(\boldsymbol{\Sigma})+\langle\boldsymbol{\Gamma},\boldsymbol{\Sigma}\rangle\} = \boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top,$$

Algorithm 1: DPGM for PDSCE Problem

Input: $\boldsymbol{S}, \boldsymbol{W}$
Initialize $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\boldsymbol{S})$, $\boldsymbol{\Gamma}_0 = \boldsymbol{\Sigma}_0 - \boldsymbol{S} - \tau\boldsymbol{\Sigma}_0^{-1}$
while *not converged* **do**
  Compute gradient $\nabla h(\boldsymbol{\Gamma}_t)$ via (7)
  Select η_t via backtracking line search or as given in Theorem 5
  Update dual variable $\boldsymbol{\Gamma}_{t+1}$ via (8)
end
Output: $\boldsymbol{\Gamma}_{t+1}$, $\boldsymbol{\Sigma}_{t+1} = \nabla h(\boldsymbol{\Gamma}_{t+1})$

where $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ come from the eigendecomposition $\boldsymbol{S}-\boldsymbol{\Gamma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ and $\mathcal{J}_\tau(\cdot)$ is the proximal operator of $-\tau\log\det(\cdot)$ with

$$[\mathcal{J}_\tau(\boldsymbol{\Lambda})]_{i,j} = \begin{cases} \frac{\Lambda_{ij}+\sqrt{\Lambda_{ij}^2+4\tau}}{2} & i=j, \\ 0 & i\neq j. \end{cases}$$

The infimum over $\boldsymbol{\Psi}$ can be computed in closed form as

$$\inf_{\boldsymbol{\Psi}\in\mathbb{S}^d}\{g(\boldsymbol{\Psi})-\langle\boldsymbol{\Gamma},\boldsymbol{\Psi}\rangle\} = \begin{cases} 0 & \text{if } |\Gamma_{ij}|\leq W_{ij}, \\ -\infty & \text{otherwise.} \end{cases}$$

which leads to the constraint $|\Gamma_{ij}| \leq W_{ij}$ for all i,j in the dual space. Combining the above results, the dual formulation of the PDSCE problem (1) becomes

$$\begin{aligned} \max_{\boldsymbol{\Gamma}} \quad & \frac{1}{2}\left\|\boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top-\boldsymbol{S}\right\|_{\mathsf{F}}^2 - \tau\log\det\mathcal{J}_\tau(\boldsymbol{\Lambda}) \\ & + \left\langle\boldsymbol{\Gamma},\boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top\right\rangle \\ \text{s.t.} \quad & |\Gamma_{ij}|\leq W_{ij}. \end{aligned} \tag{6}$$

We now present our proposed DPGM approach for solving the dual PDSCE problem (6). Let $h(\boldsymbol{\Gamma}):\mathbb{S}^d\to\mathbb{R}$ denote the objective function in the dual problem (6). Given the current iterate $\boldsymbol{\Gamma}_t$ at iteration t, the gradient is computed as

$$\nabla h(\boldsymbol{\Gamma}_t) = \boldsymbol{V}_t\mathcal{J}_\tau(\boldsymbol{\Lambda}_t)\boldsymbol{V}_t^\top, \tag{7}$$

where $\boldsymbol{S}-\boldsymbol{\Gamma}_t = \boldsymbol{V}_t\boldsymbol{\Lambda}_t\boldsymbol{V}_t^\top$ is the eigendecomposition. The dual variable is then updated via the proximal gradient step:

$$\boldsymbol{\Gamma}_{t+1} = \mathcal{P}_{\boldsymbol{W}}\left(\boldsymbol{\Gamma}_t+\eta_t\nabla h(\boldsymbol{\Gamma}_t)\right), \tag{8}$$

where η_t is the step size, and $\mathcal{P}_{\boldsymbol{W}}(\cdot)$ denotes the projection onto the box constraint set $\{\boldsymbol{X} \mid |X_{ij}|\leq W_{ij}\}$, defined elementwise by

$$[\mathcal{P}_{\boldsymbol{W}}(\boldsymbol{X})]_{ij} = \min\{\max\{X_{ij},-W_{ij}\},W_{ij}\}.$$

The step size η_t is selected through a backtracking line search until the following condition is satisfied:

$$\begin{aligned} & h(\boldsymbol{\Gamma}_{t+1})-h(\boldsymbol{\Gamma}_t)+\frac{1}{2\eta_t}\|\boldsymbol{\Gamma}_{t+1}-\boldsymbol{\Gamma}_t\|_{\mathsf{F}}^2 \\ & \geq \langle\nabla h(\boldsymbol{\Gamma}_t),\boldsymbol{\Gamma}_{t+1}-\boldsymbol{\Gamma}_t\rangle, \end{aligned} \tag{9}$$

or alternatively, the theoretically optimal choice of η_t given in Theorem 5 can be adopted.

Notably, the primal solution $\widehat{\boldsymbol{\Sigma}}$ can be recovered from the dual optimum $\widehat{\boldsymbol{\Gamma}}$ via the strong duality: $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{V}}\mathcal{J}_\tau(\widehat{\boldsymbol{\Lambda}})\widehat{\boldsymbol{V}}^\top$,

where $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{\Lambda}}$ come from the eigendecomposition of $\boldsymbol{S}-\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ is the solution to Problem (6). Consequently, the proposed method simultaneously generates a sequence of primal covariance matrix iterates:

$$\boldsymbol{\Sigma}_t = \boldsymbol{V}_t \mathcal{J}_\tau(\boldsymbol{\Lambda}_t) \boldsymbol{V}_t^\top. \tag{10}$$

Crucially, the positive definiteness of all iterates $\boldsymbol{\Sigma}_t$ is naturally preserved by the proximal operator $\mathcal{J}_\tau(\cdot)$, independent of the step size η_t. The complete dual proximal gradient method is summarized in Algorithm 1.

Computation Costs We briefly outline the practical computational advantages of DPGM. The inner subproblems of BCM, PNM, and PGM all entail multiple iterative operations, such as solving Lasso problems and performing Cholesky decompositions to verify positive definiteness, each incurring $\mathcal{O}(d^3)$ time complexity. In contrast, DPGM requires only a single $\mathcal{O}(d^3)$ operation for eigendecomposition.

6 Convergence Analysis

To demonstrate that solving Problem (1) in the dual space offers advantages beyond the guaranteed positive definiteness of $\boldsymbol{\Sigma}$, we conduct a rigorous analysis of Algorithm 1. As a first step, we characterize the strong concavity and smoothness properties of the dual objective function h.

Lemma 4. *The dual objective function $h(\boldsymbol{\Gamma})$ is $(\frac{b^2}{b^2+\tau})$-smooth and $(\frac{a^2}{a^2+\tau})$-strongly concave over the set $\mathcal{C}_\Gamma(a,b) = \{\boldsymbol{\Gamma} \,|\, a\boldsymbol{I} \preceq \boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top \preceq b\boldsymbol{I},\ \boldsymbol{S}-\boldsymbol{\Gamma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\}$. Specifically, for any $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2 \in \mathcal{C}_\Gamma(a,b)$, the following inequalities hold:*

$$\|\nabla h(\boldsymbol{\Gamma}_1) - \nabla h(\boldsymbol{\Gamma}_2)\|_{\mathsf{F}} \leq \frac{b^2}{b^2+\tau} \|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\|_{\mathsf{F}},$$

$$\|\nabla h(\boldsymbol{\Gamma}_1) - \nabla h(\boldsymbol{\Gamma}_2)\|_{\mathsf{F}} \geq \frac{a^2}{a^2+\tau} \|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\|_{\mathsf{F}}.$$

Moreover, Algorithm 1 generates iterates satisfying $\boldsymbol{\Gamma}_t \in \mathcal{C}_\Gamma(\alpha,\beta)$ and $\boldsymbol{\Sigma}_t \in \mathcal{C}_\Sigma(\alpha,\beta)$ for all $t \geq 0$.

Lemma 4 highlights the key advantages of solving the dual problem. First, the Lipschitz smoothness modulus of the dual objective h is upper bounded by $\frac{b^2}{b^2+\tau} < 1$, which implies that a dimensional independent worst-case step size choice $\eta_t = 1$ is always valid. Second, the iterates of Algorithm 1 remain within a tighter spectral set $\mathcal{C}_\Sigma(\alpha,\beta)$, in contrast to $\mathcal{C}_\Sigma(\alpha, \beta+\sqrt{d}(\alpha+\beta))$ required for PGM. This confinement enables stronger convergence guarantees and leads to improved iteration complexity, as formalized in the following theorem.

Theorem 5. *Let $\boldsymbol{\Gamma}_{t+1}$ and $\boldsymbol{\Gamma}_t$ be iterates generated by Algorithm 1 according to* (8)*, and let $\widehat{\boldsymbol{\Gamma}}$ denote the optimal solution to Problem* (6)*. Then, the following holds:*

$$\left\|\boldsymbol{\Gamma}_{t+1} - \widehat{\boldsymbol{\Gamma}}\right\|_{\mathsf{F}} \leq \max\left\{\left|1-\eta_t\frac{\alpha^2}{\alpha^2+\tau}\right|, \left|1-\eta_t\frac{\beta^2}{\beta^2+\tau}\right|\right\} \cdot \left\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\right\|_{\mathsf{F}},$$

for some step size $\eta_t > 0$. Furthermore:

1. *Algorithm 1 converges linearly if the step size η_t satisfies*
$$0 < \max\left\{\left|1-\eta_t\cdot\frac{\alpha^2}{\alpha^2+\tau}\right|, \left|1-\eta_t\cdot\frac{\beta^2}{\beta^2+\tau}\right|\right\} < 1.$$
2. *The optimal worst-case contraction is achieved when $\eta_t = 2\cdot(\frac{\alpha^2}{\alpha^2+\tau} + \frac{\beta^2}{\beta^2+\tau})^{-1}$ in which case the contraction factor becomes*
$$\max\left\{\left|1-\eta_t\cdot\frac{\alpha^2}{\alpha^2+\tau}\right|, \left|1-\eta_t\cdot\frac{\beta^2}{\beta^2+\tau}\right|\right\} = \frac{L-\mu}{L+\mu},$$
where $L = \frac{\alpha^2+\tau}{\alpha^2}$ and $\mu = \frac{\beta^2+\tau}{\beta^2}$. This leads to an iteration complexity of
$$\mathcal{O}\left(\frac{1}{4}\cdot L\mu^{-1}\cdot\log\frac{1}{\epsilon}\right).$$

The core of Theorem 5 lies in its provision of a simple and effective theoretical step size selection. Additionally, compared to PGM, the iteration complexity of DPGM is improved. Building on Theorem 5, the following corollary establishes the linear convergence of the corresponding primal sequence.

Corollary 6. *Let $\boldsymbol{\Sigma}_t$ be the primal iterate generated via* (10) *in Algorithm 1, and let $\widehat{\boldsymbol{\Sigma}}$ denote the optimal solution to Problem* (1)*. Then, for all $t \geq 0$,*

$$\left\|\boldsymbol{\Sigma}_t - \widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{F}} \leq \left(\frac{L-\mu}{L+\mu}\right)^t \left\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\right\|_{\mathsf{F}}.$$

7 Numerical Experiments

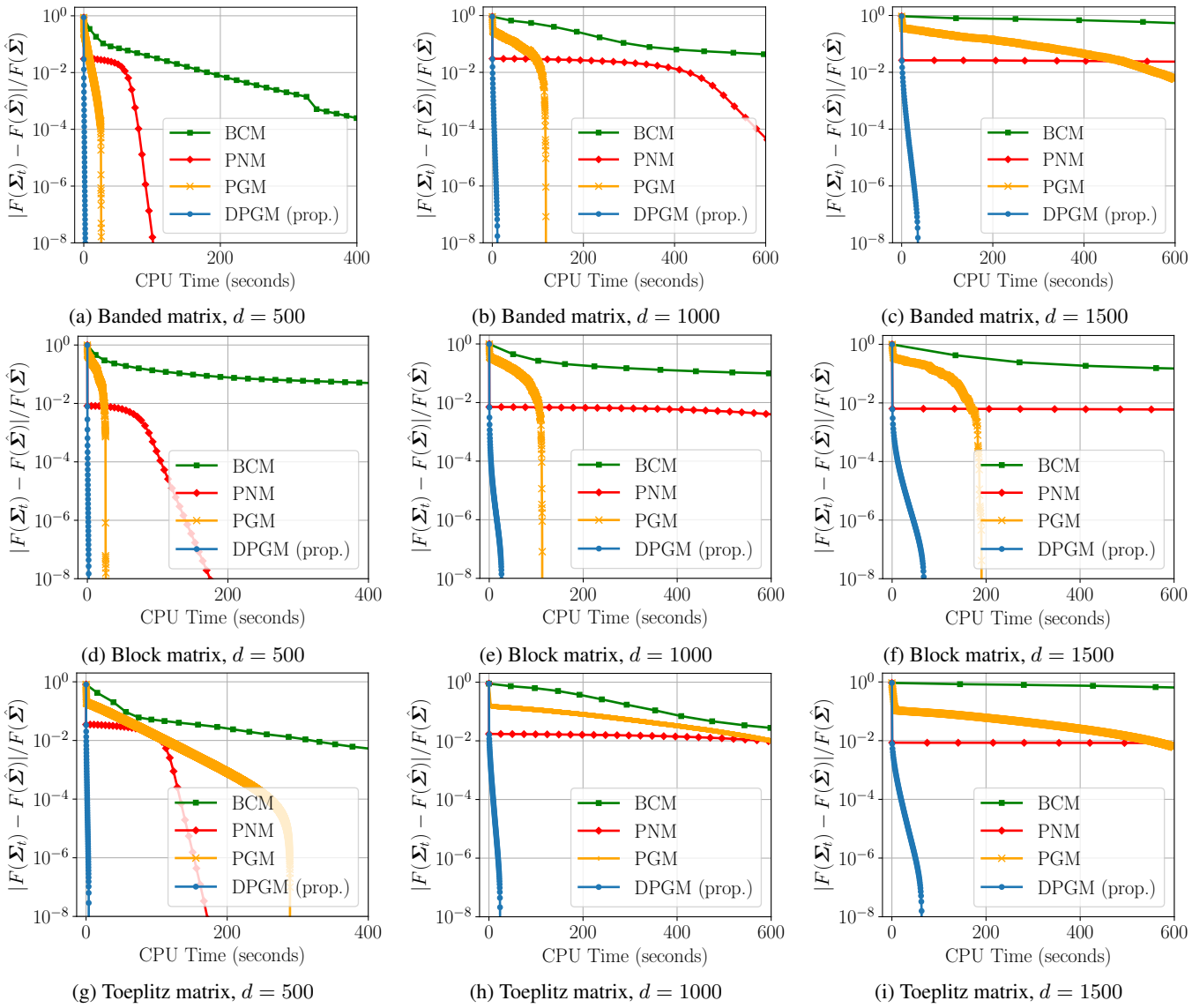
In this section, we compare the performance of the proposed DPGM algorithm with BCM, PNM, and PGM on both synthetic and real-world datasets. All methods are initialized with $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\boldsymbol{S})$. For the proposed method, we additionally set $\boldsymbol{\Gamma}_0 = \boldsymbol{\Sigma}_0 - \boldsymbol{S} - \tau\boldsymbol{\Sigma}_0^{-1}$.

Synthetic Data Experiments

We consider three types of covariance matrices as ground truth, all of which are guaranteed to be positive definite.

1. Block Matrix: The indices $\{1,\ldots,d\}$ are evenly partitioned into groups, with $\Sigma^\star_{ij} = 0.8$ if $i \neq j$ and i,j belong to the same group and 0 otherwise.
2. Banded Matrix: The entries are defined as $\Sigma^\star_{ij} = 1 - \frac{|i-j|}{m}$ for $|i-j| \leq m$, and 0 otherwise.
3. Toeplitz Matrix: The entries are $\Sigma^\star_{ij} = 0.75^{|i-j|}$.

We fix the group size of the block matrix at 100, yielding $\lceil\frac{d}{100}\rceil$ blocks. This setting ensures that the largest and smallest eigenvalues of $\boldsymbol{\Sigma}^\star$ are independent of the dimensionality. For banded matrix, we set the bandwidth $m = \lceil\frac{d}{100}\rceil$ to maintain a consistent sparsity level in $\boldsymbol{\Sigma}^\star$. The Toeplitz structure inherently ensures that its extremal eigenvalues are dimension-independent. Consistent with high-dimensional settings, we set $n = \lceil\frac{d}{5}\rceil$ observations for all experiments. The parameter τ was fixed at 10^{-4} as recommended by Rothman (2012) to obtain a stable solution.



(a) Banded matrix, $d = 500$ (b) Banded matrix, $d = 1000$ (c) Banded matrix, $d = 1500$

(d) Block matrix, $d = 500$ (e) Block matrix, $d = 1000$ (f) Block matrix, $d = 1500$

(g) Toeplitz matrix, $d = 500$ (h) Toeplitz matrix, $d = 1000$ (i) Toeplitz matrix, $d = 1500$

Figure 1: Relative errors of the objective values versus the computational time for dimensions ranging from 500 to 1500.

We first investigate the performance of the algorithms under different dimensions. We evaluate three scenarios with $d = 500$, $d = 1000$, and $d = 1500$. The diagonal elements of $\boldsymbol{W}$ are set to zero, and the off-diagonal elements are set to ρ, where ρ is selected to minimize $\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^{\star}\right\|_{\mathsf{F}}$ through a grid search on a logarithmic scale from 10^{-3} to 1. Each iteration, we calculate the relative errors $\frac{F(\boldsymbol{\Sigma}) - F(\widehat{\boldsymbol{\Sigma}})}{|F(\widehat{\boldsymbol{\Sigma}})|}$ of all algorithms, where $F(\boldsymbol{\Sigma}) = f(\boldsymbol{\Sigma}) + g(\boldsymbol{\Sigma})$ signifies the objective function of Problem (1).

As shown in Figure 1, our proposed method significantly outperforms all state-of-the-art approaches in terms of convergence time for. In most cases, PGM is faster than BCM and PNM, likely due to the double-loop structure of the latter two algorithms. PGM exhibits smaller per-step increments, possibly because the step size must simultaneously satisfy both the positive definiteness and descent conditions. We can observe that PNM exhibits local quadratic convergence. However, it is evident that this behavior occurs only when the algorithm is close to convergence.

Next, we fix the dimension at $d = 1000$ to examine how different regularization settings affect algorithmic efficiency. Specifically, we test regularization parameters $\rho = 0.04$, 0.08, and 0.12. Table 2 presents the average runtime and number of iterations required for each algorithm. We see that the proposed algorithm outperforms every other algorithm with respect to running time on all the high-dimensional settings. At the same time, we also observe that the proposed algorithm exhibits fewer iterations compared to other algorithms.

Structure	Algorithm	BCM		PNM		PGM		DPGM (prop.)	
	ρ	Time(s)	Iter	Time(s)	Iter	Time(s)	Iter	Time(s)	Iter
Block	0.04	3607.33	34	2871.34	162	519.14	242	**3.59**	20
	0.08*	4643.19	33	2452.51	98	151.86	91	**5.89**	33
	0.12	6700.25	63	1934.73	76	80.77	67	**12.60**	52
Banded	0.04	3729.17	55	3076.28	193	276.02	165	**5.84**	28
	0.08*	5680.03	67	2831.92	165	327.70	193	**10.17**	51
	0.12	6012.36	86	2614.66	143	361.77	231	**13.86**	63
Toeplitz	0.04	3669.72	47	3162.55	178	1183.75	561	**7.01**	27
	0.08	5622.96	68	2981.49	154	386.17	218	**12.44**	48
	0.12*	6015.70	91	2561.77	130	182.21	149	**9.46**	41
S&P 500	0.01	1753.29	32	597.70	12	46.72	21	**0.239**	3

Table 2: Runtime comparisons on synthetic data with $d = 1000$ and S&P 500 data with $d = 415$. The asterisk (*) marks the configuration achieving the minimal Frobenius norm error $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_{\mathsf{F}}$ across all three parameter settings.

Real-world Data Experiments

We conduct an experiment using real-world datasets from financial stock markets. We collect historical daily stock prices for S&P 500 Index constituents from January 1, 2010, to December 31, 2020. After excluding missing data, we obtain daily returns for 415 stocks ($d = 415$). We generate 100 daily return time series datasets, each with a randomly selected start date and spanning 252 consecutive trading days. For each dataset, we estimate the covariance matrix and calculate the portfolio return risk of the global minimum variance portfolio based on these estimated covariances. Let $\boldsymbol{\Sigma}$ denote a generic estimator of the covariance matrix of asset returns, and $\boldsymbol{w}$ represent the allocation vector for a portfolio comprising d financial securities. The risk of the estimated portfolio is defined as:

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

The quality of the estimated covariance matrix $\boldsymbol{\Sigma}$ directly affects the portfolio risk, with more accurate estimates resulting in lower risk.

We consider four types of covariance matrices as inputs: the sample covariance matrix $\boldsymbol{S}$ (SCM), and three regularized estimators based on Problem (1): the ℓ_1-norm penalty, the adaptive Lasso penalty (Zou 2006), and the minimax concave penalty (MCP) (Zhang 2010). The first two methods are equivalent to directly solving Problem (1) under different $\boldsymbol{W}$ configurations, whereas MCP involves solving a sequence of Problem (1) through iterative convexification (Fan et al. 2018). We first compare the runtime performance of the proposed method with other approaches when solving the ℓ_1-penalized estimator using S&P 500 data, as reported in the last row of Table 2. Furthermore, Figure 2 presents a violin plot analysis of portfolio risk across different covariance estimators. It can be observed that the SCM exhibits a higher median and a wider range of volatility, indicating both higher portfolio risk and greater instability, whereas using the formulation in Problem (1) significantly reduces both risk and volatility.

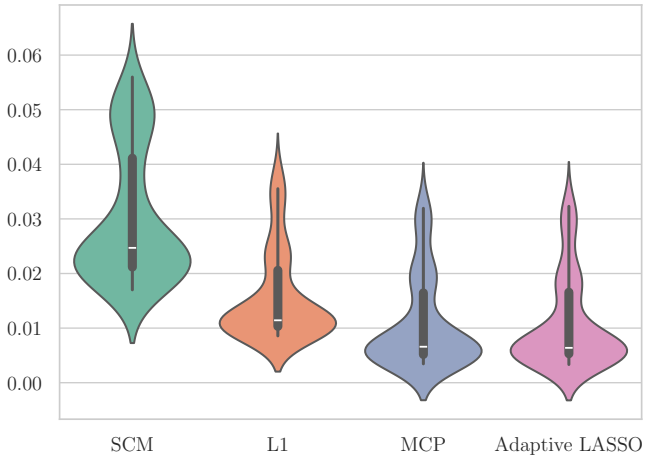


Figure 2: Violin plot of the daily volatility for different covariance estimators on stock market data.

8 Conclusion

In this paper, we have focused on the positive definite sparse covariance estimation problem. We have analyzed existing methods and identified their key limitations. To address these issues, we have proposed a novel dual proximal gradient method that inherently preserves positive definiteness through its dual-space formulation. Extensive numerical experiments have demonstrated superior computational efficiency across various high-dimensional settings.

References

Beck, A.; and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1): 183–202.

Bickel, P. J.; and Levina, E. 2008. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1): 199 – 227.

Donoho, D. L.; Gavish, M.; and Johnstone, I. M. 2018. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, 46(4): 1742–1778.

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

Fan, J.; Liu, H.; Sun, Q.; and Zhang, T. 2018. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2): 814–841.

Faust, K.; and Raes, J. 2012. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8): 538–550.

Jolliffe, I. T. 2002. *Principal Component Analysis*. Springer Series in Statistics. New York, NY, USA: Springer, 2 edition.

Karoui, N. E. 2008a. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6): 2717–2756.

Karoui, N. E. 2008b. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6): 2757–2790.

Kyrillidis, A.; Mahabadi, R. K.; Tran-Dinh, Q.; and Cevher, V. 2014. Scalable sparse covariance estimation via self-concordance. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1946–1952. Québec City, Québec, Canada: AAAI Press.

Ledoit, O.; and Wolf, M. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5): 603–621.

Li, X.; Zhao, T.; Arora, R.; Liu, H.; and Hong, M. 2018. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, 18(184): 1–24.

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

Markowitz, H. 1952. Portfolio selection. *The Journal of Finance*, 7(1): 77–91.

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

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

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

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Tran-Dinh, Q.; Kyrillidis, A.; and Cevher, V. 2015. Composite self-concordant minimization. *Journal of Machine Learning Research*, 16(12): 371–416.

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

Witten, D. M.; and Tibshirani, R. 2009. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3): 615–636.

Xiong, H.; Cheng, W.; Fu, Y.; Hu, W.; Bian, J.; and Guo, Z. 2018. De-biasing covariance-regularized discriminant analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2889–2897. International Joint Conferences on Artificial Intelligence Organization.

Xue, L.; Ma, S.; and Zou, H. 2012. Positive-definite ℓ1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500): 1480–1491.

Yan, Y.; Yang, Q.; and Zhao, Z. 2025. Large covariance matrix estimation with nonnegative correlations. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, 3502–3510. PMLR.

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

Zhao, T.; and Liu, H. 2013. Sparse inverse covariance estimation with calibration. In *Advances in Neural Information Processing Systems*, volume 26, 2274–2282. Lake Tahoe, Nevada, USA: NeurIPS.

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

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

Zou, H. 2006. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476): 1418–1429.