Large Sparse Covariance Matrix Estimation via Dual Proximal Gradient Method

Fengpei Li and Ziping Zhao
ShanghaiTech University, Shanghai, China
{fengpeili, zipingzhao}@shanghaitech.edu.cn

***Abstract*—Covariance matrix estimation in high dimensions is a central problem in data science, signal processing, and machine learning, yet it remains challenging due to the need to ensure both statistical accuracy and computational efficiency. In this paper, we revisit the positive definite covariance estimation framework of [1] and develop an efficient primal–dual algorithm that alternately updates primal and dual variables. We demonstrate that the proposed method can be interpreted equivalently as a proximal gradient scheme in the dual domain. The algorithm inherently preserves positive definiteness throughout the iterations and is provably globally linearly convergent. Numerical experiments establish the superiority of the proposed algorithm over state-of-the-art methods, highlighting its effectiveness for high-dimensional covariance estimation.**

I. Introduction

The estimation of covariance matrices lies at the core of numerous fundamental problems in modern multivariate data analysis, with broad applications across statistics [2], biology [3], finance [4], signal processing [5], and machine learning [6]. For example, in finance, covariance matrices are essential for portfolio optimization to manage risk [7]–[9]; in signal processing, they enable adaptive beamforming for optimizing antenna array reception [10]; and in machine learning, they are prerequisites for dimensionality reduction methods such as principal component analysis (PCA) [11] and for classification techniques like linear and quadratic discriminant analysis [12]. However, covariance estimation becomes particularly challenging in high-dimensional regimes where the problem dimension far exceeds the sample size. In this setting, the commonly used sample covariance matrix is inconsistent, and relying on it can severely degrade downstream tasks. For example, in PCA, inaccurate eigenvalue estimates may exaggerate the importance of certain components [13]. These challenges have motivated extensive research on high-dimensional covariance estimation in recent years [14]–[16].

To effectively estimate large covariance matrices, a widely adopted strategy is to impose structural assumptions such as sparsity, wherein many of the entries are assumed to be zero [17]. This reduces the effective number of parameters and improves statistical convergence rates. For example, in longitudinal data analysis, it is often reasonable to assume weak correlations between temporally distant observations [17]. A common method for sparse covariance estimation is thresholding [17]–[19], where small entries of the sample covariance matrix are set to zero. While these estimators possess strong theoretical guarantees, such as minimax optimality and fast convergence rates, they generally do not ensure positive definiteness. To simultaneously enforce positive definiteness and sparsity, Rothman [1] proposed the following regularized optimization problem:

$$\min_{\boldsymbol{\Sigma}} \quad \frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{S}\right\|_{\mathrm{F}}^{2}-\tau\log\det\boldsymbol{\Sigma}+\left\|\boldsymbol{W}\circ\boldsymbol{\Sigma}\right\|_{1}, \tag{1}$$

where $\boldsymbol{S}$ is the sample covariance matrix, $-\tau\log\det(\cdot)$ with parameter $\tau\geq 0$ is the logarithmic barrier, and $\|\boldsymbol{W}\circ\cdot\|_1$ is the weighted 1-norm with $\boldsymbol{W}$ being a non-negative weight matrix and $\circ$ denoting the elementwise product.

Problem (1) is convex; however, due to the coexistence of the logarithmic barrier and the weighted 1-norm, it does not admit a closed-form solution. Hence, we need to use numerical solving procedures to solve this problem. In [1], a row-by-row block coordinate descent (BCD) method was proposed. The algorithm leverages the structure of the symmetric positive definite matrices and updates one row (correspondingly one column) at each time, where the subproblem can be cast as a LASSO problem [20]. However, in high-dimensional settings, the BCD approach can be computationally prohibitive due to its double-loop nature and the fact that the number of subproblems scales with the dimension. Furthermore, no theoretical convergence analysis is provided for this approach. In addition, the proximal gradient (PG) method [21], [22] has also been proposed to solve this problem. Nonetheless, the PG method requires careful tuning of the step size to ensure that all iterates remain symmetric positive definite. In practice, the admissible step sizes are often small, which results in slow convergence.

Note that the numerical difficulty of Problem (1) comes from the coexistence of the two penalty functions; a natural idea is to decouple them. Introducing an auxiliary variable $\boldsymbol{\Psi}$, we can reformulate Problem (1) into the following linearly constrained convex programming:

$$\begin{aligned}\min_{\boldsymbol{\Sigma},\boldsymbol{\Psi}} \quad & \frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{S}\right\|_{\mathrm{F}}^{2}-\tau\log\det\boldsymbol{\Sigma}+\left\|\boldsymbol{W}\circ\boldsymbol{\Psi}\right\|_{1}\\ \text{s.t.} \quad & \boldsymbol{\Sigma}=\boldsymbol{\Psi}.\end{aligned} \tag{2}$$

Obviously, based on the augmented Lagrangian function, the classical augmented Lagrangian method [23]–[25] is applicable for solving Problem (2). The direct application of the augmented Lagrangian method, however, treats Problem (2) as a generic linearly constrained convex programming, and it ignores completely the separable structure of Problem

979-8-3315-7206-8/25/$31.00

(2). Therefore, the augmented Lagrangian-based alternating direction method (AL-ADM) [26], [27], also known as the alternating direction method of multipliers, can be applied. In AL-ADM, we split the minimization task of Problem (2) into optimizing the variables $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ in an alternating order.

While the AL-ADM algorithm leverages the separability of Problem (2), we argue that it does not fully exploit its structural properties. In particular, when both components of a problem are strongly convex, the iterative Lagrangian method can guarantee convergence without requiring an additional augmented term [28]. In Problem (2), the subproblem with respect to $\boldsymbol{\Sigma}$ is already strongly convex, suggesting that the augmented term may be unnecessary. Motivated by this observation and inspired by [29], we propose a new algorithm, termed the Lagrangian and augmented Lagrangian-based alternating direction method (LAL-ADM). The key idea is to update $\boldsymbol{\Sigma}$ based on the Lagrangian function while optimizing $\boldsymbol{\Psi}$ via the augmented Lagrangian. Furthermore, we derive the dual formulation of Problem (2) and show that LAL-ADM admits an equivalent interpretation as a dual proximal gradient method. This dual perspective allows us to establish a global linear convergence rate. Finally, numerical experiments demonstrate that LAL-ADM significantly outperforms BCD, PG, and AL-ADM in terms of computational efficiency.

II. Proposed Algorithm

In this section, we present the LAL-ADM algorithm for solving Problem (2). The key idea of LAL-ADM is to alternately update $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ using the standard Lagrangian and the augmented Lagrangian, respectively, together with the dual variable. The detailed update steps are provided below.

a) Update of $\boldsymbol{\Sigma}$: Problem (2) leads to the following Lagrangian function

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}\right) =& \frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{S}\right\|_{\mathrm{F}}^{2}-\tau\log\det\boldsymbol{\Sigma}+\left\|\boldsymbol{W}\circ\boldsymbol{\Psi}\right\|_{1}\\ &+\left\langle\boldsymbol{\Gamma},\boldsymbol{\Sigma}-\boldsymbol{\Psi}\right\rangle,\end{aligned} \tag{3}$$

where $\boldsymbol{\Gamma}$ is the dual variable. The update for $\boldsymbol{\Sigma}$ is through the partial minimization of $\mathcal{L}\left(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}\right)$, which is given by

$$\boldsymbol{\Sigma}_{+}=\arg\min_{\boldsymbol{\Sigma}}\left\{\frac{1}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{S}+\boldsymbol{\Gamma}\right\|_{\mathrm{F}}^{2}-\tau\log\det\boldsymbol{\Sigma}\right\}. \tag{4}$$

Define the eigendecomposition of $\boldsymbol{S}-\boldsymbol{\Gamma}$ as $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with diagonal entries corresponding to the eigenvalues. Problem (4) admits a closed-form solution, given by

$$\boldsymbol{\Sigma}_{+}=\boldsymbol{V}\mathcal{J}_{\tau}\left(\boldsymbol{\Lambda}\right)\boldsymbol{V}^{\top}, \tag{5}$$

where $\mathcal{J}_{\tau}\left(\cdot\right)$ is the proximal operator of $-\tau\log\det(\cdot)$ with

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

b) Update of $\boldsymbol{\Psi}$: The augmented Lagrangian for (2) is

$$\mathcal{L}_{\alpha}\left(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}\right)=\mathcal{L}\left(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}\right)+\frac{\alpha}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{\Psi}\right\|_{\mathrm{F}}^{2},$$

where $\alpha>0$ is the penalty parameter for the violation of the linear constraint. The update for $\boldsymbol{\Psi}$ is through the partial minimization of $\mathcal{L}_{\alpha}\left(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}\right)$, given by

$$\begin{aligned}\boldsymbol{\Psi}_{+}&=\arg\min_{\boldsymbol{\Psi}}\left\{\frac{1}{2}\left\|\boldsymbol{\Psi}-\boldsymbol{\Sigma}-\frac{1}{\alpha}\boldsymbol{\Gamma}\right\|_{\mathrm{F}}^{2}+\left\|\frac{1}{\alpha}\boldsymbol{W}\circ\boldsymbol{\Psi}\right\|_{1}\right\}\\ &=\mathcal{S}_{\frac{1}{\alpha}\boldsymbol{W}}\left(\boldsymbol{\Sigma}+\frac{1}{\alpha}\boldsymbol{\Gamma}\right),\end{aligned} \tag{6}$$

where $\mathcal{S}_{\frac{1}{\alpha}\boldsymbol{W}}\left(\cdot\right)$ is the soft-thresholding operator with

$$\begin{aligned}&\left[\mathcal{S}_{\frac{1}{\alpha}\boldsymbol{W}}\left(\boldsymbol{\Sigma}+\frac{1}{\alpha}\boldsymbol{\Gamma}\right)\right]_{ij}\\ &=\operatorname{sgn}(\Sigma_{ij}+\frac{1}{\alpha}\Gamma_{ij})\max\left\{\left|\Sigma_{ij}+\frac{1}{\alpha}\Gamma_{ij}\right|-\frac{1}{\alpha}W_{ij},0\right\}.\end{aligned}$$

c) Update of $\boldsymbol{\Gamma}$: The update for the dual variable is

$$\boldsymbol{\Gamma}_{+}=\boldsymbol{\Gamma}+\alpha\left(\boldsymbol{\Sigma}-\boldsymbol{\Psi}\right). \tag{7}$$

Remark 1 (Comparison of LAL-ADM and AL-ADM). Both LAL-ADM and AL-ADM are primal-dual algorithms with similar structures. The key difference lies in the update of $\boldsymbol{\Sigma}$: LAL-ADM employs the standard Lagrangian, whereas AL-ADM relies on the augmented Lagrangian. This seemingly minor difference renders LAL-ADM more advantageous, as the performance of AL-ADM is highly sensitive to the tuning of α, for which no definitive selection rule is available.

III. A Dual Proximal Gradient Interpretation

In this section, we interpret the LAL-ADM algorithm from the perspective of a dual proximal gradient method. By substituting $\boldsymbol{\Sigma}_{+}$ from (5) and $\boldsymbol{\Psi}_{+}$ from (6) into (7), we obtain the following update rule for the dual variable $\boldsymbol{\Gamma}$:

$$\begin{aligned}\boldsymbol{\Gamma}_{+}&=\boldsymbol{\Gamma}+\alpha\boldsymbol{\Sigma}_{+}-\alpha\mathcal{S}_{\frac{1}{\alpha}\boldsymbol{W}}\left(\boldsymbol{\Sigma}_{+}+\frac{1}{\alpha}\boldsymbol{\Gamma}\right)\\ &=\boldsymbol{\Gamma}+\alpha\boldsymbol{\Sigma}_{+}-\mathcal{S}_{\boldsymbol{W}}\left(\boldsymbol{\Gamma}+\alpha\boldsymbol{\Sigma}_{+}\right)\\ &=\mathcal{P}_{\boldsymbol{W}}\left(\boldsymbol{\Gamma}+\alpha\boldsymbol{\Sigma}_{+}\right)\\ &=\mathcal{P}_{\boldsymbol{W}}(\boldsymbol{\Gamma}+\alpha\boldsymbol{V}\mathcal{J}_{\tau}(\boldsymbol{\Lambda})\boldsymbol{V}^{\top}),\end{aligned} \tag{8}$$

where $\mathcal{P}_{\boldsymbol{W}}\left(\cdot\right)$ is the projection onto the box constraint set $\{\boldsymbol{X}\mid|X_{ij}|\leq W_{ij}\}$, defined as

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

The third line in (8) follows from the Moreau decomposition theorem $\boldsymbol{X}=\mathcal{S}_{\boldsymbol{W}}\left(\boldsymbol{X}\right)+\mathcal{P}_{\boldsymbol{W}}\left(\boldsymbol{X}\right)$.

In the following, we show that the last step in (8) can be interpreted as a proximal gradient step on the dual objective. We derive the Lagrange dual of Problem (1). For the Lagrangian in (3), we have

$$\inf_{\boldsymbol{\Psi}}\mathcal{L}\left(\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Gamma}\right)=\begin{cases}0 & \text{if }|\Gamma_{ij}|\leq W_{ij},\\ -\infty & \text{otherwise,}\end{cases}$$

and

$$\inf_{\boldsymbol{\Sigma}} \ \mathcal{L}\left(\boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\Gamma}\right) = \frac{1}{2}\left\|\boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top - \boldsymbol{S}\right\|_{\mathrm{F}}^2 - \tau \log\det \mathcal{J}_\tau(\boldsymbol{\Lambda}) + \left\langle \boldsymbol{\Gamma}, \boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top \right\rangle.$$

Combining these results, we obtain the dual problem for (1) as follows:

$$\begin{aligned} \max_{\boldsymbol{\Gamma}} \quad & \frac{1}{2}\left\|\boldsymbol{V}\mathcal{J}_\tau\left(\boldsymbol{\Lambda}\right)\boldsymbol{V}^\top - \boldsymbol{S}\right\|_{\mathrm{F}}^2 \\ & - \tau \log\det \mathcal{J}_\tau\left(\boldsymbol{\Lambda}\right) + \left\langle \boldsymbol{\Gamma}, \boldsymbol{V}\mathcal{J}_\tau\left(\boldsymbol{\Lambda}\right)\boldsymbol{V}^\top \right\rangle \\ \text{s.t.} \quad & |\Gamma_{ij}| \le W_{ij}. \end{aligned} \tag{9}$$

Let $g(\boldsymbol{\Gamma})$ be the objective of Problem (9), the gradient of $g(\boldsymbol{\Gamma})$ is computed as

$$\nabla g\left(\boldsymbol{\Gamma}\right) = \boldsymbol{V}\mathcal{J}_\tau\left(\boldsymbol{\Lambda}\right)\boldsymbol{V}^\top.$$

This shows that the update of $\boldsymbol{\Gamma}$ in (8) is essentially a proximal gradient ascent step with step size α for solving the dual problem (9), i.e.,

$$\boldsymbol{\Gamma}_+ = \mathcal{P}_{\boldsymbol{W}}(\boldsymbol{\Gamma} + \alpha\nabla g(\boldsymbol{\Gamma})). \tag{10}$$

We summarize the proposed algorithm in Algorithm 1. Unlike in AL-ADM, where α serves as an augmented Lagrangian parameter, here α acts as a step size analogous to that in projected gradient descent.

Remark 2. Given the interpretation above, at each iteration, α_t in practice can be chosen to satisfy the following sufficient descent condition:

$$\begin{aligned} & g\left(\boldsymbol{\Gamma}_t\right) + \left\langle \nabla g\left(\boldsymbol{\Gamma}_t\right), \boldsymbol{\Gamma}_{t+1} - \boldsymbol{\Gamma}_t \right\rangle \\ & \le g\left(\boldsymbol{\Gamma}_{t+1}\right) + \frac{1}{2\alpha_t}\left\|\boldsymbol{\Gamma}_{t+1} - \boldsymbol{\Gamma}_t\right\|_{\mathrm{F}}^2, \end{aligned} \tag{11}$$

via backtracking line search.

IV. Convergence Analysis

In this section, we analyze the theoretical properties of Algorithm 1. We first examine the strong convexity and smoothness of the function g, which are key to determining the theoretical step size α and establishing the convergence guarantees of the algorithm.

Lemma 1. *The function $g\left(\boldsymbol{\Gamma}\right)$ is $(\frac{a^2}{a^2+\tau})$-strongly concave and $(\frac{b^2}{b^2+\tau})$-smooth over the set $\mathcal{C} = \{\boldsymbol{\Gamma} \mid a\mathbf{I} \preceq \boldsymbol{V}\mathcal{J}_\tau(\boldsymbol{\Lambda})\boldsymbol{V}^\top \preceq b\mathbf{I},\ \boldsymbol{S} - \boldsymbol{\Gamma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\}$, where $0 < a < b$. Specifically, for any matrices $\boldsymbol{\Gamma}_1,\ \boldsymbol{\Gamma}_2 \in \mathcal{C}$, the following relations hold,*

$$\begin{aligned} \left\|\nabla g\left(\boldsymbol{\Gamma}_1\right) - \nabla g\left(\boldsymbol{\Gamma}_2\right)\right\|_{\mathrm{F}} &\ge \frac{a^2}{a^2+\tau} \cdot \left\|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\right\|_{\mathrm{F}} \\ \left\|\nabla g\left(\boldsymbol{\Gamma}_1\right) - \nabla g\left(\boldsymbol{\Gamma}_2\right)\right\|_{\mathrm{F}} &\le \frac{b^2}{b^2+\tau} \cdot \left\|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\right\|_{\mathrm{F}}. \end{aligned}$$

Proof: We prove the smoothness and strong convexity of the dual objective g by analyzing the primal objective in (1). Denote

$$f(\boldsymbol{\Sigma}) = \frac{1}{2}\left\|\boldsymbol{\Sigma} - \boldsymbol{S}\right\|_{\mathrm{F}}^2 - \tau\log\det(\boldsymbol{\Sigma}).$$

Algorithm 1: Dual Proximal Gradient Algorithm

Input: $\boldsymbol{S}$, $\boldsymbol{W}$, τ.
Initialize $\boldsymbol{\Gamma}_0 = \mathrm{diag}(\boldsymbol{S}) - \boldsymbol{S} - \tau\mathrm{diag}(\boldsymbol{S})^{-1}$, $t = 0$.
while *not converged* **do**
  Compute $\nabla g(\boldsymbol{\Gamma}) = \boldsymbol{V}_t\mathcal{J}_\tau\left(\boldsymbol{\Lambda}_t\right)\boldsymbol{V}_t^\top$, where $\boldsymbol{V}_t\boldsymbol{\Lambda}_t\boldsymbol{V}_t^\top$ is the eigendecomposition of $\boldsymbol{S} - \boldsymbol{\Gamma}_t$;
  $\boldsymbol{\Gamma}_{t+1} = \mathcal{P}_{\boldsymbol{W}}\left(\boldsymbol{\Gamma}_t + \alpha\nabla g\left(\boldsymbol{\Gamma}\right)\right)$;
end
Output: $\boldsymbol{\Gamma}_{t+1}$.

By duality, if f is μ-strongly convex and L-smooth, the dual function g is $\frac{1}{L}$-strongly concave and $\frac{1}{\mu}$-smooth.

Note that $\nabla f\left(\boldsymbol{\Sigma}\right) = \boldsymbol{\Sigma} - \boldsymbol{S} - \tau\boldsymbol{\Sigma}^{-1}$. Applying the mean value theorem to $\mathrm{vec}\left(\nabla f\left(\boldsymbol{\Sigma}\right)\right)$ gives

$$\begin{aligned} & \mathrm{vec}\left(\nabla f(\boldsymbol{\Sigma}_1)\right) - \mathrm{vec}\left(\nabla f(\boldsymbol{\Sigma}_2)\right) \\ & = \left(\boldsymbol{I}\otimes\boldsymbol{I} + \tau\boldsymbol{\Sigma}_c^{-1}\otimes\boldsymbol{\Sigma}_c^{-1}\right)\left(\mathrm{vec}\left(\boldsymbol{\Sigma}_1\right) - \mathrm{vec}\left(\boldsymbol{\Sigma}_2\right)\right), \end{aligned} \tag{12}$$

where $\boldsymbol{\Sigma}_c = c\boldsymbol{\Sigma}_1 + (1-c)\boldsymbol{\Sigma}_2$ with $c \in [0,1]$. Denote the eigenvalues of $\boldsymbol{\Sigma}_c$ by $\lambda_d\left(\boldsymbol{\Sigma}_c\right) \le \cdots \le \lambda_1\left(\boldsymbol{\Sigma}_c\right)$ for a given value of c. By Weyl's inequality, we have

$$\begin{aligned} \lambda_d\left(\boldsymbol{\Sigma}_c\right) &\ge \min\left\{\lambda_d\left(\boldsymbol{\Sigma}_1\right), \lambda_d\left(\boldsymbol{\Sigma}_2\right)\right\} \ge a, \\ \lambda_1\left(\boldsymbol{\Sigma}_c\right) &\le \max\left\{\lambda_1\left(\boldsymbol{\Sigma}_1\right), \lambda_1\left(\boldsymbol{\Sigma}_2\right)\right\} \le b. \end{aligned}$$

From (12), we have

$$\begin{aligned} 1 + \frac{\tau}{b^2} &\le \lambda_d\left(\boldsymbol{I}\otimes\boldsymbol{I} + \tau\boldsymbol{\Sigma}_c^{-1}\otimes\boldsymbol{\Sigma}_c^{-1}\right) \\ &\le \frac{\left\|\nabla f\left(\boldsymbol{\Sigma}_1\right) - \nabla f\left(\boldsymbol{\Sigma}_2\right)\right\|_{\mathrm{F}}}{\left\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\right\|_{\mathrm{F}}} \\ &\le \lambda_1\left(\boldsymbol{I}\otimes\boldsymbol{I} + \tau\boldsymbol{\Sigma}_c^{-1}\otimes\boldsymbol{\Sigma}_c^{-1}\right) \le 1 + \frac{\tau}{a^2}, \end{aligned}$$

which completes the proof. ∎

Based on Lemma 1, since $\frac{b^2}{b^2+\tau} < 1$ holds for all $b > 0$, any step size α with $0 < \alpha \le 1$ is admissible in Algorithm 1. The following theorem establishes the global linear convergence of the dual variable $\boldsymbol{\Gamma}$.

Theorem 2. *Suppose all iterates of $\boldsymbol{\Gamma}$ lie in the set $\mathcal{C}$ defined in Lemma 1 with parameters a and b. Let $\boldsymbol{\Gamma}_{t+1}$ and $\boldsymbol{\Gamma}_t$ denote successive iterates of Algorithm 1, and let $\boldsymbol{\Gamma}_\star$ be the optimal solution to Problem* (9)*. Then,*

$$\left\|\boldsymbol{\Gamma}_{t+1} - \boldsymbol{\Gamma}_\star\right\|_{\mathrm{F}} \le \left(1 - \alpha\frac{a^2}{b^2+\tau}\right)\left\|\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_\star\right\|_{\mathrm{F}},$$

with $0 < \alpha \le 1$.

Proof: The solution to Problem (9) is the fixed point of the following equation

$$\boldsymbol{\Gamma}_\star = \mathcal{P}_{\boldsymbol{W}}\left(\boldsymbol{\Gamma}_\star + \alpha\nabla g\left(\boldsymbol{\Gamma}_\star\right)\right). \tag{13}$$

Using the update rule $\boldsymbol{\Gamma}_{t+1}$ in (10), we obtain:

$$\begin{aligned} & \left\|\boldsymbol{\Gamma}_{t+1} - \boldsymbol{\Gamma}_\star\right\|_{\mathrm{F}} \\ & = \left\|\mathcal{P}_{\boldsymbol{W}}(\boldsymbol{\Gamma}_t + \alpha\nabla g(\boldsymbol{\Gamma}_t)) - \mathcal{P}_{\boldsymbol{W}}(\boldsymbol{\Gamma}_\star + \alpha\nabla g(\boldsymbol{\Gamma}_\star))\right\|_{\mathrm{F}} \\ & \le \left\|\boldsymbol{\Gamma}_t + \alpha\nabla g\left(\boldsymbol{\Gamma}_t\right) - \left(\boldsymbol{\Gamma}_\star + \alpha\nabla g\left(\boldsymbol{\Gamma}_\star\right)\right)\right\|_{\mathrm{F}}, \end{aligned}$$

TABLE I
COMPARISON OF FOUR DIFFERENT METHODS ON A SIMULATED DATASET WITH $d = 1000$, $n = 400$.

		BCD		PG		AL-ADM		LAL-ADM (Dual PG)	
Structure	κ	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter
Block	0.02	3607.33	34	3585.48	2543	11.84	26	**7.28**	13
	0.06	3643.19	33	2805.54	1985	27.70	58	**17.36**	31
	0.10	3700.25	63	1042.14	674	48.02	95	**31.21**	64
Band	0.02	3729.17	55	2941.03	2153	14.74	28	**6.64**	13
	0.06	3680.03	67	1957.71	1369	25.45	71	**10.82**	24
	0.10	4012.36	86	1034.52	869	43.02	99	**14.98**	30
Toeplitz	0.02	3669.72	47	3480.71	3274	11.65	24	**7.87**	12
	0.06	3622.96	68	2714.88	1963	29.84	45	**18.30**	19
	0.10	4015.70	91	923.95	737	53.26	76	**24.98**	30

where the inequality follows from the nonexpansiveness of the projection operator $\mathcal{P}_{\boldsymbol{W}}(\cdot)$.

Define $\boldsymbol{\phi}(\boldsymbol{\Gamma}) \triangleq \mathrm{vec}(\boldsymbol{\Gamma}) + \alpha \mathrm{vec}(\nabla g(\boldsymbol{\Gamma}))$. The following inequality holds:

$$\|\boldsymbol{\phi}(\boldsymbol{\Gamma}_t) - \boldsymbol{\phi}(\boldsymbol{\Gamma}_\star)\|_2 \leq \left\{ \sup_{c\in[0,1]} \|\boldsymbol{J}(\boldsymbol{\Gamma}_c)\|_2 \right\} \|\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}},$$

where $\boldsymbol{J}(\cdot)$ denotes the Jacobian matrix of $\boldsymbol{\phi}(\cdot)$ and $\boldsymbol{\Gamma}_c = c\boldsymbol{\Gamma}_t + (1-c)\boldsymbol{\Gamma}_\star$. Next, we derive the explicit expression for $\boldsymbol{J}(\boldsymbol{\Gamma}_c)$ and demonstrate that when $\alpha \leq 1$, the supremum $\sup_{c\in[0,1]} \|\boldsymbol{J}(\boldsymbol{\Gamma}_c)\|_2$ is strictly less than 1 for all t. This establishes the linear convergence of $\boldsymbol{\Gamma}$.

To compute the Jacobian of $\mathrm{vec}(\nabla g(\cdot))$, note that $\nabla g(\boldsymbol{\Gamma})$ is given by the solution $\boldsymbol{A}$ to the equation $\boldsymbol{A} - \boldsymbol{S} - \tau\boldsymbol{A}^{-1} + \boldsymbol{\Gamma} = \boldsymbol{0}$. Accordingly, define $u(\boldsymbol{A}, \boldsymbol{\Gamma}) = \mathrm{vec}(\boldsymbol{A}) - \mathrm{vec}(\boldsymbol{S}) - \mathrm{vec}\left(\tau\boldsymbol{A}^{-1}\right) + \mathrm{vec}(\boldsymbol{\Gamma})$. Applying the implicit function theorem, we obtain:

$$\begin{aligned}\frac{\partial \mathrm{vec}(\nabla g(\boldsymbol{\Gamma}))}{\partial \mathrm{vec}(\boldsymbol{\Gamma})} &= \left[\frac{\partial u(\boldsymbol{A},\boldsymbol{\Gamma})}{\partial \mathrm{vec}(\boldsymbol{A})}\right]^{-1} \cdot \frac{\partial u(\boldsymbol{A},\boldsymbol{\Gamma})}{\partial \mathrm{vec}(\boldsymbol{\Gamma})} \\ &= -\left(\boldsymbol{I} + \tau\boldsymbol{A}^{-1} \otimes \boldsymbol{A}^{-1}\right)^{-1}.\end{aligned}$$

Thus, the Jacobian $\boldsymbol{J}(\boldsymbol{\Gamma}_c)$ is given by

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{\Gamma}_c) &= \frac{\partial \mathrm{vec}(\boldsymbol{\Gamma}_c)}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_c)} + \alpha \frac{\partial \mathrm{vec}(\nabla g(\boldsymbol{\Gamma}_c))}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_c)} \\ &= \boldsymbol{I} - \alpha\left(\boldsymbol{I} + \tau\boldsymbol{A}_c^{-1} \otimes \boldsymbol{A}_c^{-1}\right)^{-1},\end{aligned}$$

where $\otimes$ is the Kronecker product and $\boldsymbol{A}_c = \boldsymbol{V}_c \mathcal{J}_\tau(\boldsymbol{\Lambda}_c)\boldsymbol{V}_c^\top$, where $\boldsymbol{V}_c$ and $\boldsymbol{\Lambda}_c$ are from the eigendecomposition $\boldsymbol{S} - \boldsymbol{\Gamma}_c = \boldsymbol{V}_c\boldsymbol{\Lambda}_c\boldsymbol{V}_c^\top$.

Next, the eigenvalues of $(\boldsymbol{I} + \tau\boldsymbol{A}_c^{-1} \otimes \boldsymbol{A}_c^{-1})^{-1}$ are

$$\rho_{ij} = \frac{[\mathcal{J}_\tau(\boldsymbol{\Lambda}_c)]_{ii}[\mathcal{J}_\tau(\boldsymbol{\Lambda}_c)]_{jj}}{[\mathcal{J}_\tau(\boldsymbol{\Lambda}_c)]_{ii}[\mathcal{J}_\tau(\boldsymbol{\Lambda}_c)]_{jj} + \tau}.$$

Since $0 < \rho_{ij} < 1$ and $a \leq \lambda_d(\boldsymbol{A}_c) \leq \lambda_1(\boldsymbol{A}_c) \leq b$, for any $0 < \alpha \leq 1$, we have

$$\|\boldsymbol{\Gamma}_{t+1} - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}} \leq \left(1 - \alpha\frac{a^2}{b^2+\tau}\right)\|\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}}.$$

■

Building on Theorem 2, we can establish that the sequence $\boldsymbol{\Sigma}_t$ converges to the solution of Problem (1) at a linear rate. This result is formalized in the following corollary.

Corollary 3. *Let $\boldsymbol{\Sigma}_\star$ be the solution of Problem* (1)*. The iterate $\boldsymbol{\Sigma}_t$ is computed as in equation* (5)*, i.e., $\boldsymbol{\Sigma}_t = \boldsymbol{V}_t\mathcal{J}_\tau(\boldsymbol{\Lambda}_t)\boldsymbol{V}_t^\top$. Then,*

$$\|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}} \leq \left(1 - \alpha\frac{a^2}{b^2+\tau}\right)^t \|\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}}.$$

Proof: Note that $\boldsymbol{\Sigma}_t = \arg\min_{\boldsymbol{\Sigma}}\{f(\boldsymbol{\Sigma}) - \langle\boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}\rangle\}$ and $\boldsymbol{\Sigma}_\star = \arg\min_{\boldsymbol{\Sigma}}\{f(\boldsymbol{\Sigma}) - \langle\boldsymbol{\Gamma}_\star, \boldsymbol{\Sigma}\rangle\}$. We get

$$\nabla f(\boldsymbol{\Sigma}_t) = -\boldsymbol{\Gamma}_t, \quad \nabla f(\boldsymbol{\Sigma}_\star) = -\boldsymbol{\Gamma}_\star.$$

Given that $f(\boldsymbol{\Sigma})$ is strongly convex with parameter $1 + \frac{\tau}{b^2}$ as shown in Lemma 1, we obtain

$$\begin{aligned}\|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}} &\leq \frac{b^2}{b^2+\tau}\|\nabla f(\boldsymbol{\Sigma}_t) - \nabla f(\boldsymbol{\Sigma}_\star)\|_{\mathrm{F}} \\ &\leq \left(1 - \alpha\frac{a^2}{b^2+\tau}\right)^t \|\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}}.\end{aligned}$$

■

V. NUMERICAL EXPERIMENTS

A. Linear convergence

In this section, we first demonstrate the linear convergence of the proposed LAL-ADM algorithm and examine how it depends on the choice of regularization parameter $\boldsymbol{W}$. We set all off-diagonal elements of $\boldsymbol{W}$ to κ, while the diagonal elements are set to zero. In this experiment, we set the dimension to $d = 1000$ and the number of samples to $n = 400$. The ground-truth covariance matrix is generated as a banded matrix, as described in Section V-B. Fig. 1 illustrates the convergence of $\|\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}}$ with increasing iterations, confirming the theoretical result in Theorem 2. Notably, the algorithm converges more slowly as the regularization parameter κ increases. This aligns with the dual problem formulated in Section III, where κ serves as the box-constraint parameter.

TABLE II
COMPARISON OF FOUR DIFFERENT METHODS ON A SIMULATED DATASET WITH $d = 1000$, $n = 1000$.

		BCD		PG		AL-ADM		LAL-ADM (Dual PG)	
Structure	κ	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter
Block	0.02	3768.04	46	2885.80	1648	12.37	26	**8.39**	18
	0.06	227.94	6	405.54	603	17.89	26	**9.09**	19
	0.10	186.06	5	118.55	321	20.78	25	**10.36**	21
Band	0.02	3681.51	73	3200.18	1332	13.62	29	**5.26**	23
	0.06	229.93	7	801.24	841	34.48	54	**11.68**	29
	0.10	173.16	5	171.11	381	49.93	78	**15.63**	33
Toeplitz	0.02	3601.08	71	4103.71	2816	10.82	23	**6.62**	15
	0.06	116.32	3	1093.35	1109	36.85	39	**11.71**	22
	0.10	130.96	3	512.31	216	61.88	58	**13.08**	29

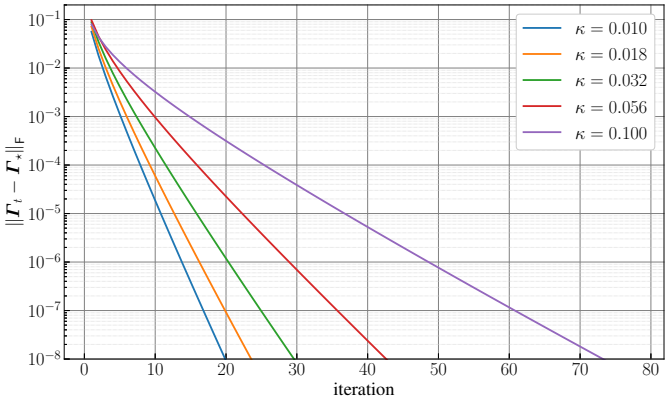

Fig. 1. Convergence of LAL-ADM with varying κ.

B. Synthetic experiments

In this section, we compare the performance of the proposed LAL-ADM to the performance of BCD [1], PG [21], and AL-ADM. All the methods are initialized as $\boldsymbol{\Sigma}_0 = \text{diag}(\boldsymbol{S})$, and for LAL-ADM and AL-ADM, we additionally initialize $\boldsymbol{\Psi}_0 = \boldsymbol{\Sigma}_0$ and $\boldsymbol{\Gamma}_0 = \boldsymbol{\Sigma}_0 - \boldsymbol{S} - \tau\boldsymbol{\Sigma}_0^{-1}$.

We consider three distinct 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 10 groups, where $\Sigma_{ij} = 0.8$ if i and j ($i \neq j$) belong to the same group, and 0 otherwise. The smallest eigenvalue satisfies $\lambda_{\min}(\boldsymbol{\Sigma}) = 0.2$.
2) Banded matrix: The entries are defined as $\Sigma_{ij} = 1 - \frac{|i-j|}{100}$ for $|i-j| \leq 100$ and 0 otherwise. For $d = 1000$, the smallest eigenvalue is approximately $\lambda_{\min}(\boldsymbol{\Sigma}) \approx 0.005$.
3) Toeplitz matrix: The entries follow $\Sigma_{ij} = 0.75^{|i-j|}$. When $d = 1000$, the smallest eigenvalue is approximately $\lambda_{\min}(\boldsymbol{\Sigma}) \approx 0.143$.

The parameter τ was fixed at 10^{-4} as recommended by [1] for stable solution. We conduct two sets of experiments with a fixed dimension of $d = 1000$: one with $n = 400$ for high-dimensional setting where $n < d$, and another with $n = 1000$ for moderate-dimensional setting where $n \approx d$. In each setting, we systematically evaluate the performance of all algorithms across three different sparsity penalty levels ($\kappa = 0.02$, 0.06, and 0.10) to ensure comprehensive comparison under varying regularization conditions.

In the presented tables, we report the average computational time and number of iterations required for each algorithm to satisfy the stopping criterion. As shown in Table I, the LAL-ADM algorithm outperforms all other methods in terms of runtime across all high-dimensional settings. Additionally, LAL-ADM consistently requires fewer iterations than other algorithms, including the BCD algorithm, which is generally recognized for achieving substantial progress per iteration. Compared to the AL-ADM algorithm, LAL-ADM demonstrates both faster CPU times and fewer iterations. This observation is consistent with the discussion in Section II, where the strong convexity of the $\boldsymbol{\Sigma}$-related component in the LAL-ADM formulation eliminates the need for an additional augmented Lagrangian term. Conversely, the PG algorithm frequently requires the highest number of iterations to achieve convergence. As analyzed in Section IV, this behavior can be attributed to the necessity of an extremely small step size to ensure the identification of a feasible solution. As presented in Table II, for moderate-dimensional settings, the runtimes of other algorithms decrease, yet LAL-ADM remains the most efficient in terms of computational speed.

VI. CONCLUSIONS

This paper focuses on the estimation of large sparse positive definite covariance matrices and proposes a fast algorithm based on the alternating direction method. Each step of the algorithm admits a closed-form solution, ensuring positive definiteness at every iteration. We further interpret the proposed method as a proximal gradient ascent algorithm in the dual domain and derive the rule for step size selection. Convergence analysis demonstrates a linear convergence rate, and comprehensive experiments validate the effectiveness of the proposed algorithm.

REFERENCES

[1] A. J. Rothman, "Positive definite estimators of large covariance matrices," *Biometrika*, vol. 99, no. 3, pp. 733–740, Jun. 2012.

[2] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[3] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[4] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

[5] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5016–5029, 2010.

[6] B. Hao, W. W. Sun, Y. Liu, and G. Cheng, "Simultaneous clustering and estimation of heterogeneous graphical models," *Journal of Machine Learning Research*, vol. 18, no. 217, pp. 1–58, 2018.

[7] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[8] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio with budget constraint," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2342–2357, 2018.

[9] Z. Zhao, R. Zhou, Z. Wang, and D. P. Palomar, "Optimal portfolio design for statistical arbitrage in finance," in *2018 IEEE Statistical Signal Processing Workshop, SSP 2018, Freiburg im Breisgau, Germany, June 10-13, 2018*, IEEE, 2018, pp. 801–805.

[10] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*, 2nd. Raleigh, NC, USA: SciTech Publishing, 2004.

[11] I. T. Jolliffe, *Principal Component Analysis* (Springer Series in Statistics), 2nd ed. New York, NY, USA: Springer, 2002.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.

[13] N. E. Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *The Annals of Statistics*, vol. 36, no. 6, pp. 2757–2790, 2008.

[14] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 75, no. 4, pp. 603–680, Aug. 2013.

[15] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, 2016.

[16] D. L. Donoho, M. Gavish, and I. M. Johnstone, "Optimal shrinkage of eigenvalues in the spiked covariance model," *Annals of Statistics*, vol. 46, no. 4, pp. 1742–1778, Aug. 2018.

[17] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[18] N. E. Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *The Annals of Statistics*, vol. 36, no. 6, pp. 2717–2756, 2008.

[19] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 177–186, 2009.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[21] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate via majorization-minimization," *IEEE Transactions on Signal Processing*, vol. 71, pp. 3328–3342, 2023.

[22] W. Xia, Z. Zhao, and Y. Sun, "C-ISTA: iterative shrinkage-thresholding algorithm for sparse covariance matrix estimation," in *IEEE Statistical Signal Processing Workshop, SSP 2023, Hanoi, Vietnam, July 2-5, 2023*, IEEE, 2023, pp. 215–219.

[23] M. R. Hestenes, "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, no. 5, pp. 303–320, 1969.

[24] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," in *Optimization*, R. Fletcher, Ed., London: Academic Press, 1969, pp. 283–298.

[25] W. Sun and Y.-X. Yuan, *Optimization Theory and Methods: Nonlinear Programming* (Springer Optimization and Its Applications), 1st ed. New York, NY: Springer, 2006.

[26] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[28] H. Uzawa, "Iterative methods for concave programming," in *Studies in Linear and Nonlinear Programming*, K. J. Arrow, L. Hurwicz, and H. Uzawa, Eds., Stanford, CA: Stanford University Press, 1958, pp. 154–165.

[29] P. Tseng, "Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming," *Mathematical Programming*, vol. 48, pp. 249–263, 1990.