Article

High-Dimensional Covariance Estimation via Constrained L_q-Type Regularization

Xin Wang [1,*], Lingchen Kong [1], Liqun Wang [2] and Zhaoqilin Yang [3]

1. School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China
2. Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
3. Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
* Correspondence: 18118020@bjtu.edu.cn

Abstract: High-dimensional covariance matrix estimation is one of the fundamental and important problems in multivariate analysis and has a wide range of applications in many fields. In practice, it is common that a covariance matrix is composed of a low-rank matrix and a sparse matrix. In this paper we estimate the covariance matrix by solving a constrained L_q-type regularized optimization problem. We establish the first-order optimality conditions for this problem by using proximal mapping and the subspace method. The proposed stationary point degenerates to the first-order stationary points of the unconstrained L_q regularized sparse or low-rank optimization problems. A smoothing alternating updating method is proposed to find an estimator for the covariance matrix. We establish the convergence of the proposed calculation method. The numerical simulation results show the effectiveness of the proposed approach for high-dimensional covariance estimation.

Keywords: high-dimensional covariance matrix; constrained L_q-type regularized optimization problem; smoothing alternating updating method

MSC: 62-08; 65K05

1. Introduction

High-dimensional covariance matrix estimation is an important topic in statistical inference and learning. The objective is to estimate the covariance matrix of p variables based on sample data of size n. Much statistical analysis of high-dimensional data involves covariance estimation. In a high-dimensional setting (usually $p \gg n$), researchers usually assume that the covariance matrix is sparse and have proposed various regularization techniques [1–9] to consistently estimate covariance matrix Σ. However, this assumption is not appropriate in many applications such as financial analysis, depending on the equity market risks, gene expressions stimulated by cytokines, etc. A more reasonable assumption based on the factor model [10–14] is that the high-dimensional covariance matrix is the sum of a low-rank matrix and a sparse matrix. Specifically, a factor model can be formulated as

$$Y_i = Bf_i + u_i, i = 1, \cdots, n, \tag{1}$$

where $Y_i = (Y_{1i}, \cdots, Y_{pi})^T$ is the observed variable, $B = (b_1, \cdots, b_p)^T$ is the matrix of factor loadings, f_i is a $K \times 1$ vector of common factors, and $u_i = (u_{1i}, \cdots, u_{pi})^T$ is the idiosyncratic error component, uncorrelated with f_i. Under this model, the covariance matrix of Y_i is given by

$$\begin{aligned} \Sigma &= B\text{cov}(f_i)B^T + \Sigma_u \\ &:= L + S, \end{aligned} \tag{2}$$

where L is the covariance matrix of Bf_i and S is the covariance matrix of u_i. Since K is usually less than p, L is a low-rank matrix. More generally, there is a large class of problems where one is interested in estimating the covariance matrix of X_i from the noisy observations $Y_i = X_i + u_i$, where the noise u_i is uncorrelated with the signal X_i. In addition, if X_i is generated from a linear system so that some of its elements are linear functions of the others, then the covariance matrix of X_i has low-rank. Based on the structural decomposition (2), this paper proposes a novel approach to estimate the covariance matrix by solving a constrained low-rank and sparse regularized optimization problem.

Under structural decomposition (2), researchers have proposed some approaches to estimate the covariance matrix. When the common factors in Model (1) are known and observable and S is diagonal, Reference [10] proposed a least-squares-based covariance estimation (LS-CE). The core idea of LS-CE is first to estimate B by the least-squares method and, then, obtain the final estimator assuming that S is diagonal. Compared to the sample covariance matrix estimator, the LS-CE estimator is always invertible, even in a high-dimensional setting. Moreover, this estimator is shown to have substantial gains when estimating the inverse of the covariance matrix, however, it does not have much advantage when estimating the covariance matrix. Instead of assuming diagonal S, Reference [11] assumed that S is sparse and applied the adaptive thresholding technique of [4] to obtain an estimator for S. Then, a least-squares-based covariance estimator with conditional sparsity (LS-CS-CE) under Model (1) was studied in [11]. More recently, a robust approach based on Huber loss minimization was proposed in [13]. This approach estimates the initial joint covariance matrix of the observed data and the factors and, then, uses it to recover the covariance matrix of the observed data. Moreover, the proposed estimator in [13] only requires the condition that the fourth moment of the data exists. This method is applicable to a wider range of data, including sub-Gaussian and elliptical distributions. When the common factors are unknown and unobservable, Reference [12] proposed a non-parametric estimator via the thresholding principal orthogonal complements, which is called the principal orthogonal complement thresholding (POET) method. The POET uses the K principal components to estimate L and applies the adaptive thresholding technique [4] on the rest of the $(p - K)$ components to estimate S. However, as stated in [9], using the simple thresholding approaches in [1–5,9] to estimate S may lead to the covariance estimators in [11–13], which are necessarily positive semi-definite.

Moreover, as suggested in [14], an intuitive optimization-type approach is to solve the problem:

$$\min_{L,S\in\mathbb{R}^{p\times p}} \frac{1}{2}\|L + S - \widetilde{\Sigma}\|_F^2 + \lambda_L \mathrm{rank}(L) + \lambda_S \sum_{(u,v)\in D^c} I(|S_{uv}|), \qquad (3)$$

where $\widetilde{\Sigma}$ is the sample covariance matrix, $\mathrm{rank}(L)$ denotes the rank of matrix L, $I(|S_{uv}|) = 1$ if $|S_{uv}| \neq 0$ and $I(|S_{uv}|) = 0$ if $|S_{uv}| = 0$, $D^c := \{(u,v) \in \mathbb{N} \times \mathbb{N} : 1 \leq u \leq p, 1 \leq v \leq p, u \neq v\}$ is the index set of non-diagonal elements of S, and λ_L and λ_S are tuning parameters. To facilitate the calculation, Reference [14] proposed the low-rank and sparse covariance estimator (LOREC) via solving the following convex relaxation optimization problem:

$$\min_{L,S\in\mathbb{R}^{p\times p}} \frac{1}{2}\|L + S - \widetilde{\Sigma}\|_F^2 + \lambda_R\|L\|_* + \lambda_S \sum_{(u,v)\in D^c} |S_{uv}|, \qquad (4)$$

where $\|L\|_* := \sum_{i=1}^{p} \sigma_i(L)$ denotes the nuclear norm of L and $\sigma_i(L)$ is the ith largest singular value of L. However, the estimators for L and S proposed in [14] are not guaranteed to be positive semi-definite. Clearly, the optimization problems (3) and (4) can be considered as special cases of the robust principal component analysis (RPCA) in [15] and its convex relaxation, respectively.

In this paper, we propose an optimization model similar to some models of RPCA, which is formulated as

$$\min_{L,S \in \mathbb{R}^{p \times z}} \text{rank}(L) + \lambda_S \|S\|_0, \text{ s.t. } L + S = \widetilde{\Sigma},$$

where $\|S\|_0$ denotes the number of the non-zero elements of S. Reference [15] showed that the principal component pursuit estimate solving

$$\min_{L,S \in \mathbb{R}^{p \times z}} \|L\|_* + \lambda_S \sum_{u=1}^{p} \sum_{v=1}^{p} |S_{uv}|, \text{ s.t. } L + S = \widetilde{\Sigma}, \tag{5}$$

exactly recovers the low-rank L and the sparse S. Furthermore, Reference [16] suggested that L and S can be estimated by solving the following unconstrained optimization problem:

$$\min_{L,S \in \mathbb{R}^{p \times z}} \|L\|_* + \lambda_S \sum_{u=1}^{p} \sum_{v=1}^{p} |S_{uv}| + \frac{\lambda_L}{2} \|L + S - \widetilde{\Sigma}\|_F^2. \tag{6}$$

More references about the RPCA and related works based on convex and non-convex regularization can be found in [17–21].

The main contributions of this paper can be summarized as follows:

- We construct a constrained adaptive L_q-type regularized optimization problem for covariance estimation, which is a non-smooth and non-convex problem. This optimization problem is based on the structural decomposition (2) and the idea of RPCA in [16], but is not limited to the factor model (1).
- We study the first-order optimality condition of this problem and define a class of hybrid first-order stationary points. We establish the relationship between the first-order stationary point and the global minimizer of this problem.
- We propose a smoothing alternating updating method to solve the constructed non-smooth, non-convex optimization problem and established its convergence. The simulation results show that the proposed smoothing alternating updating method is efficacious for the constrained adaptive L_q-type regularized optimization problem.

The rest of this paper is organized as follows. In Section 2, we give some notations and preliminaries. In Section 3, we construct a constrained adaptive L_q-type regularized optimization problem for covariance estimation. In Section 4, we study the first-order optimality condition of this problem and define a class of hybrid first-order stationary points. In Section 5, we propose a smoothing alternating updating method to solve the non-smooth, non-convex optimization problem and establish its convergence. The simulation results are given in Section 6. The conclusions and discussion are given in Section 7, while the mathematical proofs are given in Section 8.

2. Notations and Preliminaries

Throughout this paper, we use the following notations. The sets of $p \times p$ positive semi-definite symmetric and positive definite symmetric matrices are denoted by $\mathbb{S}_+^p$ and $\mathbb{S}_{++}^p$, respectively. It is known that $\mathbb{S}_+^p$ and $\mathbb{S}_{++}^p$ are cones.

Given matrices $X := (X_{uv})_{u,v=1}^{p} \in \mathbb{R}^{p \times p}$ and $Y := (Y_{uv})_{u,v=1}^{p} \in \mathbb{R}^{p \times p}$, where X_{uv} is the (u, v)th element of X, $\langle X, Y \rangle := \sum_{u=1}^{p} \sum_{v=1}^{p} X_{uv} Y_{uv}$, and $X \circ Y$ denotes the Hadamard product $(X \circ Y)_{uv} =: X_{uv} Y_{uv}$ for all $u, v = 1, \ldots, p$. Define the elementwise maximum norm (or max norm) $|X|_{\max} := \max_{u,v} |X_{uv}|$, the Frobenius norm $\|X\|_F := (\sum_{u=1}^{p} \sum_{v=1}^{p} X_{uv}^2)^{1/2}$, and the spectral norm $\|X\|_2 := \sup_{\|x\|_2 \leq 1} \|Xx\|_2$, where $\|x\|_2 := \sqrt{\sum_{i=1}^{p} x_i^2}$ is 2-norm of vector x. Moreover, given a matrix $X \in \mathbb{R}^{p \times p}$ and a index set $\Gamma \subseteq \{(u, v) : 1 \leq u \leq p, 1 \leq v \leq p\}$, X_Γ denotes the array of the same dimension such that $[X_\Gamma]_{uv} = X_{uv}$ for $(u, v) \in \Gamma$ and $[X_\Gamma]_{uv}$ are undefined for $(u, v) \in \Gamma^c$. For convenience, we will write the (u, v)th element of X_Γ as X_{uv} regardless if it is defined or not.

Given a vector $x \in \mathbb{R}^p$, $\mathrm{Diag}(x)$ denotes a square matrix with vector x on its diagonal and zeros elsewhere. Given a matrix $X \in \mathbb{S}_+^p$ and letting $D_i(X)$ denote the ith largest eigenvalue of X, $i = 1, \ldots, p$, write $D(X) = (D_1(X), \ldots, D_p(X))^T \in \mathbb{R}^p$. Define sign operator:

$$\mathrm{sign}(t) = \begin{cases} 1, & \text{if } t > 0 \\ 0, & \text{if } t = 0 \\ -1, & \text{if } t < 0, \end{cases}$$

and set $D := \{(u,v) \in \mathbb{N} \times \mathbb{N} : 1 \leq u \leq p,\ 1 \leq v \leq p, u = v\}$.

Let $f : \mathbb{R}^{p \times p} \to \overline{\mathbb{R}}$ be a convex function, and let $\bar{X} \in \mathrm{dom} f$. The subdifferential of f at $\bar{X}$ is defined by

$$\partial f(\bar{X}) := \left\{ \xi \in \mathbb{R}^{p \times p} : \lim_{t \downarrow 0} \frac{f(\bar{X} + tZ) - f(\bar{X})}{t} \geq \langle \xi, Z \rangle \text{ for all } Z \in \mathbb{R}^{p \times p} \right\},$$

where $\mathrm{dom} f := \{X \in \mathbb{R}^{p \times p} : f(X) < \infty\}$ denotes the domain of f, $\overline{\mathbb{R}} := (-\infty, \infty]$. In other words, $\partial f(\bar{X})$ is the closed convex hull of all points of the form $\lim_{i \to \infty} \nabla f(X_i)$, where $\{X_i\}$ is any sequence that converges to $\bar{X}$ while avoiding the set for which f fails to be differentiable. Moreover, the singular subdifferential of f at $\bar{X}$ is defined by

$$\partial^\infty f(\bar{X}) := \left\{ \xi \in \mathbb{R}^{p \times p} : (\xi, 0) \in N\big((\bar{X}, f(\bar{X})); \mathrm{epi} f\big) \right\},$$

where $\mathrm{epi} f := \{(X, t) \in \mathbb{R}^{p \times p} \times \mathbb{R} : X \in \mathrm{dom} f, t \geq f(X)\}$ denotes the epigraph of f.

Finally, we give the definition of the proximal mapping in this paper. Given a closed set Ω and a function $f : \mathbb{R}^{p \times p} \to \overline{\mathbb{R}}$, the proximal mapping on Ω of f is the operator given by

$$\mathrm{Prox}_{f,\Omega}(X) = \underset{U \in \Omega}{\mathrm{argmin}} \left\{ f(U) + \frac{1}{2} \|U - X\|_F^2 \right\} \text{ for any } X \in \mathbb{R}^{p \times p}.$$

3. Covariance Estimation with Adaptive L_q-Type Regularization

Considering that sparse covariance matrix estimation is intrinsically a heteroscedastic problem ([4]) and L and S are all positive semi-definite matrices, we estimate the covariance matrix by solving the following adaptive L_q-type regularized optimization problem:

$$\min_{L, S \in \mathbb{S}_+^p} F(L, S) := \frac{1}{2} \|L + S - \widetilde{\Sigma}\|_F^2 + \lambda_L \|L\|_q^q + \sum_{(u,v) \in D^c} \lambda_{uv}^S |S_{uv}|^q, \tag{7}$$

where $q \in (0, 1)$, $\|L\|_q^q := \sum_{i=1}^p \sigma_i(L)^q$ is the Schatten-q quasi-norm of matrix L and λ_L and λ_{uv}^S are tuning parameters. Note that although we use the same q for L and S here for simplicity, it can be different in general. Our approach is called L_q regularized covariance (L_q-CSC) estimation with conditional sparsity. The final estimator $\widehat{\Sigma} := \widehat{L} + \widehat{S}$ is called the L_q regularized covariance estimator with conditional sparsity (L_q-CSCE).

To deal with heteroscedasticity in sparse covariance matrix estimation, we adopt the adaptive thresholding approach of Reference [4] and [12] to calculate the entry-dependent threshold λ_{uv}^S in (7) for S_{uv} as

$$\lambda_{uv}^S := \lambda \left(\frac{\widetilde{\sigma}_{uu} \widetilde{\sigma}_{vv} \log p}{n} \right)^{1/2}, \tag{8}$$

where $\lambda > 0$ is a constant. Note that these entry-dependent tuning parameters for S in Problem (7) depend on the variance of the corresponding variables. The parameter $\widetilde{\sigma}_{uu}$ can be taken as the u-th diagonal elements of the estimator of the POET for S. It is worth mentioning that, when the data matrix is normalized, the covariance matrix is equivalent to

its correlation matrix, in which case, all tuning parameters are $\lambda\sqrt{\log p/n}$, and the adaptive L_q regularization function acting on S becomes the ordinary L_q regularization function.

We now discuss the selection of $\widetilde{\Sigma}$. The common one is the ordinary sample covariance matrix

$$\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^T,$$

where $\bar{\boldsymbol{Y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{Y}_i$, $\boldsymbol{Y}_i$ is the i-th sample vector, $i = 1, \cdots, n$. The ordinary sample covariance is suitable for the data with exponential-type tails or polynomial-type tails; see [1,4,8]. However, in many practical problems data tend to be heavy-tailed. Reference [9,13,22] defined a Huber minimization sample covariance matrix, which is suitable for the data with finite fourth moment. Moreover, Reference [9] also defined a rank-based covariance matrix and a median of means covariance matrix for data with other distribution conditions.

Finally, we discuss some basic properties of Problem (7).

Proposition 1. *The following statements of $F(\boldsymbol{L}, \boldsymbol{S}) : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}_+$ in Problem (7) hold:*

(i) $F(\boldsymbol{L}, \boldsymbol{S})$ is proper, closed, and coercive, i.e., $\lim_{\|(\boldsymbol{R},\boldsymbol{S})\|_F \to \infty} F(\boldsymbol{L}, \boldsymbol{S}) = \infty$.

(ii) $F(\boldsymbol{L}, \boldsymbol{S})$ attains its minimal value over $\mathbb{S}_+^p \times \mathbb{S}_+^p$.

(iii) Let $f(\boldsymbol{L}, \boldsymbol{S}) = \frac{1}{2}\|\boldsymbol{L} + \boldsymbol{S} - \widetilde{\Sigma}\|_F^2$; the gradient $\nabla f(\boldsymbol{L}, \boldsymbol{S})$ is Lipschitz continuous with a constant $\ell = 2$, that is, for any $(\boldsymbol{L}_1, \boldsymbol{S}_1), (\boldsymbol{L}_2, \boldsymbol{S}_2) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, it holds that

$$\|\nabla f(\boldsymbol{L}_1, \boldsymbol{S}_1) - \nabla f(\boldsymbol{L}_2, \boldsymbol{S}_2)\|_F \le 2\|(\boldsymbol{L}_1, \boldsymbol{S}_1) - (\boldsymbol{L}_2, \boldsymbol{S}_2)\|_F.$$

(iv) Given a matrix $\boldsymbol{L}^* \in \mathbb{R}^{p\times p}$, the partial gradient function $\nabla_S f(\boldsymbol{L}^*, \boldsymbol{S})$ is Lipschitz continuous with a constant $\ell = 1$, that is, for any $\boldsymbol{S}_1, \boldsymbol{S}_2 \in \mathbb{R}^{p\times p}$, it holds that

$$\|\nabla_S f(\boldsymbol{L}^*, \boldsymbol{S}_1) - \nabla_S f(\boldsymbol{L}^*, \boldsymbol{S}_2)\|_F \le \|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_F.$$

Moreover, given a matrix $\boldsymbol{S}^* \in \mathbb{R}^{p\times p}$, the partial gradient function $\nabla_L f(\boldsymbol{L}, \boldsymbol{S}^*)$ is Lipschitz continuous with a constant $\ell = 1$.

4. Necessary Optimality Conditions

In this section, we define a class of first-order stationary points for Problem (7), which is a hybridization of the first-order stationary points in [23–25] and the fixed point in [25]. Furthermore, we prove that any global minimizer of Problem (7) is a first-order stationary point of Problem (7). We also define an approximate variant of this stationary point. It is worth mentioning that the constraints in Problem (7) are treated as ordinary closed convex set constraints in this paper.

For the simple unconstrained L_q regularized sparse optimization problem [23,24]:

$$\min_{\boldsymbol{x}\in\mathbb{R}^p} f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_q^q,$$

where $\|\boldsymbol{x}\|_q^q := \sum_{i=1}^{p}|x_i|^q$, its the first-order stationary point is defined as the point satisfying

$$\text{Diag}(\boldsymbol{x}^*)\nabla f(\boldsymbol{x}^*) + \lambda q|\boldsymbol{x}^*|^q = \boldsymbol{0}.$$

For the simple unconstrained L_q regularized rank optimization problem [25]:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} f(\boldsymbol{X}) + \lambda\|\boldsymbol{X}\|_q^q,$$

its first-order stationary point (or fixed point) is defined as the point satisfying

$$\boldsymbol{X}^* \in \text{Prox}_{\lambda\|\boldsymbol{X}\|_q^q/\ell, \mathbb{R}^{p\times p}}\left(\boldsymbol{X}^* - \frac{1}{\ell}\nabla f(\boldsymbol{X}^*)\right),$$

where ℓ is a constant greater than the Lipschitz constant of $\nabla f(\boldsymbol{X})$. Inspired by these works, we can define a class of hybrid first-order stationary points of Problem (7).

Definition 1. $(\boldsymbol{L}^*, \boldsymbol{S}^*) \in \mathbb{S}_+^p \times \mathbb{S}_+^p$ *is said to be the first-order stationary point of Problem (7) if* $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ *satisfies*

$$\boldsymbol{L}^* \in \mathrm{Proxi}_{(\lambda_L/\ell)\|\boldsymbol{L}\|_q^q, \mathbb{S}_+^p}\left(\boldsymbol{L}^* - \frac{1}{\ell}(\boldsymbol{L}^* + \boldsymbol{S}^* - \widetilde{\boldsymbol{\Sigma}})\right), \tag{9}$$

$$\boldsymbol{0} \in \left\{\boldsymbol{S}^* \circ (\boldsymbol{L}^* + \boldsymbol{S}^* - \widetilde{\boldsymbol{\Sigma}}) + \boldsymbol{Q}^* + \boldsymbol{S}^* \circ \boldsymbol{Y} : \boldsymbol{Y}_{\Gamma^* \cup D} \in \partial h_{\Gamma^* \cup D}(\boldsymbol{S}^*_{\Gamma^* \cup D})\right\}, \tag{10}$$

where $\boldsymbol{Q}^* := (Q_{uv}^*)_{u,v=1}^p$ *and*

$$Q_{uv}^* = \begin{cases} q\lambda_{uv}^S|S_{uv}^*|^q, & \text{if } (u,v) \in D^c, \\ 0, & \text{if } (u,v) \in D, \end{cases}$$

$\Gamma^* = \{(u,v) : S_{uv}^* \neq 0\}$, $h_{\Gamma^* \cup D}(\boldsymbol{S}_{\Gamma^* \cup D}) := \delta_{\mathbb{S}_+^p}([\boldsymbol{S}_{\Gamma^* \cup D} : \boldsymbol{0}])$, $\boldsymbol{Y} = [\boldsymbol{Y}_{\Gamma^* \cup D} : \boldsymbol{Y}_{(\Gamma^* \cup D)^c}]$, $\boldsymbol{Y}_{(\Gamma^* \cup D)^c}$ *can be any finite array.*

Next, we investigate the relationship between the global minimizer of Problem (7) and the first-order stationary point of Problem (7) without any qualification conditions.

Theorem 1. *Suppose that* $(\boldsymbol{L}^*, \boldsymbol{S}^*) \in \mathbb{S}_+^p \times \mathbb{S}_+^p$ *is a global minimizer of Problem (7), and let* $\Gamma^* = \{(u,v) \in \mathbb{N} \times \mathbb{N} : S_{uv}^* \neq 0\}$. *Then,* $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ *is a first-order stationary point with* $\ell > 1$ *of Problem (7).*

Again, under some qualification conditions, the form of the first-order stationary point of Problem (7) can be further analyzed. We now discuss the specific form of the subdifferential of $h_{\Gamma^* \cup D}$ at $\boldsymbol{S}^*_{\Gamma^* \cup D}$. Note that

$$h_{\Gamma^* \cup D}(\boldsymbol{X}_{\Gamma^* \cup D}) = \delta_{\mathbb{S}_+^p}([\boldsymbol{X}_{\Gamma^* \cup D} : \boldsymbol{0}]) = \delta_{\mathbb{S}_+^p}(\mathcal{L}_{\Gamma^* \cup D}(\boldsymbol{X}_{\Gamma^* \cup D})),$$

where $\mathcal{L}_{\Gamma^* \cup D}(\boldsymbol{X}_{\Gamma^* \cup D}) := [\boldsymbol{X}_{\Gamma^* \cup D} : \boldsymbol{0}]$ is the mapping acting on array $\boldsymbol{X}_{\Gamma^*}$. The adjoint mapping $\mathcal{L}_{\Gamma^* \cup D}^* : \mathbb{R}^{p \times p} \to \mathbb{R}^{p \times p}$ is defined by

$$\langle \mathcal{L}_{\Gamma^* \cup D}(\boldsymbol{X}_{\Gamma^* \cup D}), \boldsymbol{Y} \rangle = \langle \boldsymbol{X}_{\Gamma^* \cup D}, \mathcal{L}_{\Gamma^* \cup D}^*(\boldsymbol{Y}) \rangle \text{ for all } \boldsymbol{X}_{\Gamma^* \cup D} \, \boldsymbol{Y} \in \mathbb{R}^{p \times p}.$$

Specifically, $\mathcal{L}_{\Gamma^* \cup D}^*(\boldsymbol{Y}) = \boldsymbol{Y}_{\Gamma^* \cup D}$. Clearly, $\delta_{\mathbb{S}_+^p}$ is convex, and $\mathcal{L}_{\Gamma^* \cup D}$ and $\mathcal{L}_{\Gamma^* \cup D}^*$ are linear mappings. It follows from Theorem 23.9 in [26] and Theorem 2.51 in [27] that, if $\boldsymbol{S}^* \in \mathbb{S}_+^p$, then

$$\{\mathcal{L}_{\Gamma^* \cup D}^*(\boldsymbol{Y}) : \boldsymbol{Y} \in N_{\mathbb{S}_+^p}(\boldsymbol{S}^*)\} \subseteq \partial h_{\Gamma^* \cup D}(\boldsymbol{S}^*_{\Gamma^* \cup D}). \tag{11}$$

Moreover, suppose that the qualification condition $\mathrm{Ker}\mathcal{L}_{\Gamma^* \cup D}^* \cap \partial^\infty \delta_{\mathbb{S}_+^p}(\boldsymbol{S}^*) = \{\boldsymbol{0}\}$ holds, where $\mathrm{Ker}\mathcal{L}_{\Gamma^* \cup D}^*$ is the kernel of linear mapping $\mathcal{L}_{\Gamma^* \cup D}^*$, then

$$\{\mathcal{L}_{\Gamma^* \cup D}^*(\boldsymbol{Y}) : \boldsymbol{Y} \in N_{\mathbb{S}_+^p}(\boldsymbol{S}^*)\} = \partial h_{\Gamma^* \cup D}(\boldsymbol{S}^*_{\Gamma^* \cup D}). \tag{12}$$

Based on these discussions, the following statement holds: If $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ satisfies

$$\boldsymbol{L}^* \in \mathrm{Proxi}_{\lambda_L\|\boldsymbol{L}\|_q^q, \mathbb{S}_+^p}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{S}^*),$$

$$\boldsymbol{0} \in \left\{\boldsymbol{S}^* \circ (\boldsymbol{L}^* + \boldsymbol{S}^* - \widetilde{\boldsymbol{\Sigma}}) + \boldsymbol{Q}^* + \boldsymbol{S}^* \circ \boldsymbol{Y} : \boldsymbol{Y} \in N_{\mathbb{S}_+^p}(\boldsymbol{S}^*)\right\},$$

where Q^* is defined in Definition 1 and the qualification condition $\text{Ker}\,\mathcal{L}^*_{\Gamma^* \cup D} \bigcap \partial^\infty \delta_{\mathbb{S}^p_+}(S^*) = \{\mathbf{0}\}$ holds, then (L^*, S^*) is a first-order stationary point of Problem (7).

Obviously, the qualification condition $\text{Ker}\,\mathcal{L}^*_{\Gamma^* \cup D} \bigcap \partial^\infty \delta_{\mathbb{S}^p_+}(S^*) = \{\mathbf{0}\}$ holds automatically if $\delta_{\mathbb{S}^p_+}$ is locally Lipschitz continuous at S^*. In other words, if $S^* \in \text{int}(\mathbb{S}^p_+) = \mathbb{S}^p_{++}$, this qualification condition holds. Indeed, the true covariance matrix S^o is often assumed to be positive definite [7,8]. When $S^* \in \mathbb{S}^p_{++}$, we can establish the lower bound theory of the first-order stationary point of Problem (7).

Theorem 2. *Let (L^*, S^*) be a first-order stationary point of Problem (7) satisfying $F(L^*, S^*) \leq F(L^0, S^0) + \epsilon$ for some L^0, S^0 and $\epsilon \geq 0$, $\Gamma^* := \{(u, v) : S^*_{uv} \neq 0\}$. If $S^* \in \mathbb{S}^p_{++}$, then*

$$|S^*_{uv}| \geq \left(\frac{q \min\{\lambda^S_{uv}\}}{\sqrt{2[F(L^0, S^0) + \epsilon]}} \right)^{\frac{1}{1-q}} \forall\, (u, v) \in D^c \bigcap \Gamma^*.$$

Here, we only give the lower bound theory of the first-order stationary point of Problem (7). With reference to the definitions of the first- and second-order stationary point established by [23,24] for the general unconstrained L_q regularized optimization, a class of second-order necessary optimality conditions of Problem (7) and its lower bound theory can be considered.

Note that Problem (7) is actually a very challenging optimization problem due to the positive definite constraints and the non-Lipschitz regularization terms, even finding a first-order stationary point. Here, we define a class of approximate first-order stationary points of Problem (7).

Definition 2. *We say that $(L^*, S^*) \in \mathbb{S}^p_+ \times \mathbb{S}^p_+$ is an ε-approximate first-order stationary point ($\varepsilon \geq 0$) of Problem (7) if there exist a set $\Gamma^* \subseteq \{(u, v) : 1 \leq u, v \leq p,\ S^*_{uv} \neq 0\}$ such that*

$$\text{dist}\left(L^*, \text{Prox}_{(\lambda_L/\ell)\|L\|^q_q, \mathbb{S}^p_+}\left(L^* - \frac{1}{\ell}(L^* + S^* - \widetilde{\Sigma}) \right) \right) \leq \varepsilon, \tag{13}$$

$$\text{dist}\left(0, L^*_{\Gamma^* \cup D} + S^*_{\Gamma^* \cup D} - \widetilde{\Sigma}_{\Gamma^* \cup D} + Q^*_{\Gamma^* \cup D} + \Xi_{\Gamma^* \cup D} : \Xi_{\Gamma^* \cup D} \in \partial h_{\Gamma^* \cup D}(S^*_{\Gamma^* \cup D}) \right) \leq \varepsilon, \tag{14}$$

$$\|S^*_{(\Gamma^* \cup D)^c}\|_F \leq \varepsilon. \tag{15}$$

where Q^* is defined in Definition 1.

Moreover, if (L^*, S^*) is a first-order stationary point of Problem (7), it is an ε-approximate first-order stationary point of Problem (7) with $\varepsilon = 0$. If (L^*, S^*) is an ε-approximate first-order stationary point of Problem (7) with $\varepsilon = 0$, it is a first-order stationary point of Problem (7) with $\varepsilon = 0$. In practice, the stopping point of any algorithm for an optimization problem is usually an approximate stationary point, rather than a stationary point.

5. Numerical Optimization

In this section, we propose a smoothing alternating updating (SAU) method to solve Problem (7), which is a combination of the smoothing projection gradient step in [28,29] and the proximal gradient step in [30] under the framework of the ordinary alternating minimization method. The convergence of the SAU method is established.

5.1. A Smooth Approximation Optimization Problem

Referring to the smoothing technique in [23,29], a class of smooth approximation optimizations of Problem (7) can be given by

$$\min_{L, S \in \mathbb{S}^p_+} F_\mu(L, S) := \frac{1}{2}\|L + S - \widetilde{\Sigma}\|^2_F + \lambda_L \|L\|^q_q + \sum_{(u,v) \in D^c} \lambda^S_{uv} \psi_\mu(S_{uv})^q, \tag{16}$$

where $\mu \in (0, \infty)$ and

$$\psi_\mu(S_{uv}) = \begin{cases} |S_{uv}|, & |S_{uv}| > \mu, \\ \frac{S_{uv}^2}{2\mu} + \frac{\mu}{2}, & |S_{uv}| \le \mu. \end{cases}$$

Obviously, $\psi_\mu(S_{uv})$ is first-order continuously differential and

$$(\psi_\mu(S_{uv})^q)' = \begin{cases} q|S_{uv}|^{q-1}\text{sign}(S_{uv}), & |S_{uv}| > \mu, \\ q\left(\frac{S_{uv}^2}{2\mu} + \frac{\mu}{2}\right)^{q-1}\frac{S_{uv}}{\mu}, & |S_{uv}| \le \mu. \end{cases} \tag{17}$$

Furthermore,

$$\nabla_S F_\mu(\boldsymbol{L}, \boldsymbol{S}) = \boldsymbol{L} + \boldsymbol{S} - \widetilde{\boldsymbol{\Sigma}} + \Psi(\boldsymbol{S}),$$

where

$$(\Psi(\boldsymbol{S}))_{uv} = \begin{cases} \lambda_{uv}^S(\psi_\mu(S_{uv})^q)', & (u,v) \in D^c, \\ 0, & (u,v) \in D. \end{cases} \tag{18}$$

Next, we define a class of hybrid first-order stationary points of Problem (16).

Definition 3. *We say that $(\boldsymbol{L}^*, \boldsymbol{S}^*) \in \mathbb{S}_+^p \times \mathbb{S}_+^p$ is a first-order stationary point of Problem (16) if $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ satisfies*

$$\boldsymbol{L}^* \in \text{Proxi}_{(\lambda_L/\ell)\|\boldsymbol{L}\|_q^q, \mathbb{S}_+^p}\left(\boldsymbol{L}^* - \frac{1}{\ell}(\boldsymbol{L}^* + \boldsymbol{S}^* - \widetilde{\boldsymbol{\Sigma}})\right), \tag{19}$$

$$\boldsymbol{0} \in \left\{\nabla_S F_\mu(\boldsymbol{L}^*, \boldsymbol{S}^*) + \boldsymbol{Y}, \ \boldsymbol{Y} \in N_{\mathbb{S}_+^p}(\boldsymbol{S}^*)\right\}. \tag{20}$$

Similar to the proof of Theorem 1, it is easy to prove that any global solution of Problem (16) is a first-order stationary point of Problem (16) with $\ell > 1$. The following conclusions establish the relationship between the first-order stationary points of Problems (7) and (16), which is the core of the convergence of the SAU method in next subsection.

Theorem 3. *Let $\{\mu_k\}$ denote a sequence with $\mu_k > 0$, $k = 1, 2, \cdots$, and $\mu_k \to 0$ as $k \to 0$ and $(\boldsymbol{L}_{\mu_k}^*, \boldsymbol{S}_{\mu_k}^*)$ be a first-order stationary point of Problem (16) with $\mu = \mu_k$. Then, the following statements holds:*

(i) If $\mu_k \le \varepsilon/p^2$, then $(\boldsymbol{L}_{\mu_k}^*, \boldsymbol{S}_{\mu_k}^*)$ is an ε-approximate first-order stationary point of Problem (7).
(ii) Any accumulation point of $(\boldsymbol{L}_{\mu_k}^*, \boldsymbol{S}_{\mu_k}^*)$ is a first-order stationary point of Problem (7).

5.2. Smoothing Alternating Updating Method

Theorem 3 provides a feasible procedure to obtain an approximate solution of Problem (7) by solving Problem (16). This subsection gives the framework of the SAU method for Problem (7) and establishes its convergence.

First, we give the framework of the alternating updating (AU) method for solving Problem (16).

Algorithm 1. *Given $\alpha_0 > 0$, $\beta \in (0,1)$, $\ell > 1$, choose an the initial point $(\boldsymbol{L}^0, \boldsymbol{S}^0) \in \mathbb{S}_+^p \times \mathbb{S}_+^p$. For $m = 0, 1, 2, \cdots$, repeat the following two steps until convergence:*
Step 1. *Update $\boldsymbol{L}$ by solving the following optimization problem:*

$$\boldsymbol{L}^{m+1} = \underset{\boldsymbol{L} \in \mathbb{S}_+^p}{\text{argmin}} \ \langle \nabla_L f(\boldsymbol{L}^m, \boldsymbol{S}^m), \boldsymbol{L} - \boldsymbol{L}^m \rangle + \frac{\ell}{2}\|\boldsymbol{L} - \boldsymbol{L}^m\|_F^2 + \lambda_L\|\boldsymbol{L}\|_q^q. \tag{21}$$

Step 2. *Compute*

$$S^{m+1} = \underset{S \in \mathbb{S}_+^p}{\operatorname{argmin}} \ \|S - (S^m - \alpha_m \nabla_S F_\mu(L^{m+1}, S^m))\|_F^2, \tag{22}$$

where $\alpha_m = \alpha_0 \beta^{\ell_m}$ and ℓ_m is the smallest non-negative integer ℓ such that

$$F_\mu(L^{m+1}, S^m(\alpha_0 \beta^\ell))) \le F_\mu(L^{m+1}, S^m)) - \frac{c}{2} \|S^m(\alpha_0 \beta^\ell)) - S^m\|_F^2,$$

and $S^m(\alpha) = \min_{S \in \mathbb{S}_+^p} \ \|S - (S^m - \alpha \nabla_S F_\mu(L^{m+1}, S^m))\|_F^2$.

The AU method proposed in this paper applies the proximal gradient step [30] to update L and uses the smoothing projection gradient step [28,29] to update S. A key component at each iteration of the proposed AU method is to solve a singular-value minimization Problem (21), which can be rewritten as follows:

$$\min_{L \in \mathbb{S}_+^p} \frac{1}{2} \left\| L - \left(L^m - \frac{1}{\ell} \nabla_L f(L^m, S^m) \right) \right\|_F^2 + \frac{\lambda_L}{\ell} \|L\|_q^q. \tag{23}$$

Theorem 3.4 of [30] provides a solution to Problem (23). To facilitate the expression, define vector q-thresholding operator $\mathcal{H}(d) := (h_v(d_1), \cdots, h_v(d_p))^T$ with $v = 2\lambda_L / \ell$, where

$$h(d_i) = \begin{cases} h_{v,q}(d_i) & d_i > d^*, \\ (v(1-q))^{1/(2-q)} \text{ or } 0 & d_i = d^*, \\ 0, & d_i < d^*, \end{cases} \tag{24}$$

$d^* = \frac{2-q}{2(1-q)}(v(1-q))^{1/(2-q)}$, and $h_{v,q}(d_i)$ is the unique root of the single-variable minimization problem in $(\bar{d}, +\infty)$:

$$\min_{x \ge 0} g_{d_i}(x) := x^2 - 2d_i x + vx^q,$$

where $\bar{d} = (vq(1-q)/2)^{1/(2-q)}$. Moreover, $h_{v,q}(d_i)$ is differentiable and strictly increasing on $[d^*, +\infty)$. As per the suggestion in [25], $h_{v,q}(d_i)$ can be simply obtained by the numerical methods, for example the Newton method:

$$x^{z+1} = x^z - \frac{qv(x^z)^{q-1}/2 + x^z - d_i}{q(q-1)v(x^z)^{q-2}/2 + 1}$$

with the initial point $d_0 = 1.5\bar{d}$.

Theorem 4 (Theorem 3.4 in [30]). *Let $U\mathrm{Diag}(d)U^T$ be the eigenvalue decomposition of $L^m - (1/\ell)\nabla_L f(L^m, S^m)$). Then, $L^* = U\mathrm{Diag}(\mathcal{H}(d))U^T$ is an optimal solution to Problem (23), where $\mathcal{H}(d)$ is defined as (24) with $v = 2\lambda_L / \ell$.*

When $q = 1/2$, $\mathcal{H}(d)$ has a closed-form expression, that is $h_{v,q}(d_i) = \frac{2}{3}d_i(1 + \cos(\frac{2\pi}{3} - \frac{2\varphi(d_i)}{3}))$, $\varphi(d_i) = \arccos(\frac{v}{8}(\frac{d_i}{3})^{-3/2})$, $d^* = \frac{\sqrt[3]{54}}{4}v^{2/3}$. Refer to [30] for the details.

Another key component of at each iteration of the proposed AU method is a projection problem over $\mathbb{S}_+^p$, which can be written as follows:

$$\min_{S \in \mathbb{S}_+^p} \|S - C^m\|_F^2, \tag{25}$$

where $C^m := S^m - \alpha_m \nabla_S F_\mu(L^{m+1}, S^m)$. It is well known that, if C^m has the eigen-decomposition $C^m := \sum_{i=1}^p \lambda_i v_i v_i^T$, then $S^* = \sum_{i=1}^p \max\{\lambda_i, 0\}v_i v_i^T$ is the optimal solution of Problem (25).

By (23) and (25), it is easy to find that, when the initial matrices L^0 and S^0 in the AU method are symmetric, the matrices for L and S output by the AU method are also symmetric. The next theorem gives the convergence of the AU method.

Theorem 5. *Let* $\{(L^m, S^m)\}$ *be the sequence generated by the AU method. Then, the following statements hold:*

    (i) $F_\mu(L^{m+1}, S^{m+1}) \leq F_\mu(L^m, S^m) - \frac{c}{2}\|S^{m+1} - S^m\|_F^2 - \frac{\ell-1}{2}\|L^{m+1} - L^m\|_F^2$.

    (ii) $\{(L^m, S^m)\}$ *is bounded.*

    (iii) $\lim_{m\to\infty}\|L^{m+1} - L^m\|_F = 0$, $\lim_{m\to\infty}\|S^{m+1} - S^m\|_F = 0$.

    (iv) Any accumulation point of $\{(L^m, S^m)\}$ *is a first-order stationary point of Problem (16).*

We now give the framework of the SAU method for solving Problem (16).

Algorithm 2. *Give* $\tau \in (0, 1)$. *Choose an initial smoothing parameter* μ_0. *For* $k = 0, 1, 2, \cdots$, *repeat the following two steps until convergence:*

    Step 1. *Solve Problem (16) with* μ_k *by the AU method, and obtain* $(L^*_{\mu_k}, S^*_{\mu_k})$.

    Step 2. *Set* $\mu_{k+1} \leq \tau\mu_k$, *and return to* **Step 1**.

Theorem 3 shows that any accumulation point of $(L^*_{\mu_k}, S^*_{\mu_k})$ is a first-order stationary point of Problem (7), which provides the guarantee of the convergence of the SAU method. In practice, k cannot be taken as infinite, and μ_k cannot be taken as 0. Then, the L_q-CSCE we obtained by the SAU method is often an ε-approximate first-order stationary point of Problem (7). Indeed, we stop the SAU method if it reaches a maximum iteration $k_{\max}$ or $\mu_k < \varepsilon$, where $\varepsilon > 0$ is a small constant. Moreover, in order to improve the efficiency of the SAU method, a trick is to take the output point $(L^*_{\mu_k}, S^*_{\mu_k})$ as the initial point of the AU method for Problem (16) with μ_{k+1}. Thus, if the initial point of the AU method for solving Problem (16) with μ_0, then the final matrices for L and S generated by the SAU method are also symmetric.

6. Simulations

In this section, we study the performance of the L_q-CSC method and the associated algorithms. The following models are considered in this section:

- **Model 1** (banded matrix with ordering). $\Sigma^* = \text{blockdiag}(A_1, A_2)$, where $A_1 = (\sigma_{uv})$, $1 \leq u, v \leq p/2$, $\sigma_{uv} = (1 - \frac{|i-j|}{10})_+$, $A_2 = 4I_{p/2 \times p/2}$. Σ^* is a two-block diagonal matrix, A_1 is a banded and sparse covariance matrix, and A_2 is a diagonal matrix with four along the diagonal.
- **Model 2** (simple factor). $\Sigma^* = UDU^T + I$, where $U \in R^{p\times 3}$ with orthonormal columns uniformly random and $D = 8I$.
- **Model 3** (random factor). $\Sigma^* = UDU^T + \Sigma_{\text{Sparse}}$, where $U \in R^{p\times 3}$ with orthonormal columns uniformly random and $D = 8I$, where Σ_{Sparse} is the sparse matrix generated by Model 1.

The matrix Σ^* in Model 1 is sparse, while the matrix in Model 2 and Model 3 satisfies structural decomposition (2).

6.1. Comparison with Representative Methods

This subsection studies the advantages and disadvantages of the L_q-CSCE and existing methods. Four representative estimators (the adaptive thresholding estimator (ATE) with the hard thresholding rule in [4], the optimization-type estimator (OTE) in [8], the POET in [12], and the LOREC in [14]) were used for the comparison. The first two estimators are based on the assumption of sparse covariance. While the ATE is a good simple thresholding estimator, the OTE is a good optimization-type estimator. Our estimators POET and LOREC are based on the structural decomposition (2).

Under each model, we set $n = 100$, and all random sample are generated from the normal distribution with mean 0 and covariance matrix Σ, for $p = 30, 100, 200$. The losses are measured by three matrix norms: the spectral norm, the max norm, and the Frobenius norm. The standard errors are given in parentheses. It is known that $q = 0.5$ is a good choice for the variable selection problem and the low-rank problem [24,25]. For convenience, the parameter q in Problem (7) is taken as 0.5. In this subsection, we do not consider different options of q.

For Model 1, the tuning parameters λ_L in Problems (4) and (7) are taken as 10^8. Moreover, we set $K = 0$ in the POET. Under these setting, the LOREC, POET, and L_q-CSCE degenerate into the simple soft thresholding estimator [1,2], the ATE, and the constrained adaptive thresholding estimator, respectively. Table 1 reports the means and standard errors of the five estimators under the three losses. In our analysis, the performance of the POET is almost the same as the ATE. Note that compared with the ATE, OTE, POET, and L_q-CSCE, the LOREC has a large bias under the spectral norm loss and the Frobenius norm loss, which is caused by the bias of the L_1 penalty. Moreover, the L_q-CSCE performs best under the spectral norm loss and the Frobenius norm loss. All methods perform similarly under the max norm.

For Models 2 and 3, we choose the tuning parameters in the LOREC and L_q-CSCE by setting the grid and setting $K = 3$ in the POET. Tables 2 and 3 report the means and standard errors of the five estimators under the three losses. Note that the ATE and OTE performed poorly for Models 2 and 3. In contrast, the LOREC, POET, and L_q-CSCE were more effective. The L_q-CSCE performs best under the spectral norm loss and the Frobenius norm loss. The same as the result of Model 1, all methods perform similarly under the max norm. It is worth mentioning that the adaptive technology is invalid for Model 2.

Moreover, Table 4 records the number of times that the estimators of the POET, LOREC, and L_q-CSCE for $\boldsymbol{L}$ and $\boldsymbol{S}$ are positive semi-definite matrices. Clearly, all estimators of the L_q-CSC estimation for $\boldsymbol{L}$ and $\boldsymbol{S}$ are positive semi-definite. It is not guaranteed that the estimators of the LOREC and POET for $\boldsymbol{S}$ are positive semi-definite matrices.

Table 1. Simulation results for Model 1 over 100 replications. The standard errors are given in parentheses.

p	SE	ATE	OTE	POET	LOREC	L_q-CSCE
(Spectral norm)						
30	3.76 (0.58)	1.73 (0.50)	1.89 (1.04)	1.74 (0.48)	3.04 (0.43)	1.74 (0.48)
100	8.58 (0.67)	2.69 (0.44)	2.70 (0.50)	2.61 (0.49)	4.94 (0.38)	2.60(0.50)
200	14.10 (0.82)	4.73 (0.52)	3.27 (0.42)	3.02 (0.45)	6.17 (0.38)	3.02(0.46)
(Max norm)						
30	1.25 (0.23)	1.17 (0.28)	1.16 (0.28)	1.16 (0.28)	1.22 (0.22)	1.16 (0.28)
100	1.59 (0.26)	1.59 (0.26)	1.48 (0.32)	1.51 (0.33)	1.50 (0.25)	1.51 (0.33)
200	1.73 (0.21)	1.73 (0.21)	1.63 (0.25)	1.67 (0.25)	1.63 (0.19)	1.67 (0.25)
(Frobenius norm)						
30	7.69 (0.52)	3.17 (0.48)	3.32 (1.03)	3.22 (0.43)	6.35 (0.47)	3.12 (0.44)
100	25.30 (0.64)	8.10 (0.66)	7.24 (0.52)	6.73 (0.49)	16.06 (0.58)	6.41 (0.52)
200	50.50 (0.79)	21.41 (0.80)	12.11 (0.48)	9.80 (0.47)	25.59 (0.62)	9.25 (0.49)

Table 2. Simulation results for Model 2 over 100 replications. The standard errors are given in parentheses.

p	SE	ATE	OTE	POET	LOREC	L_q-CSCE
(Spectral norm)						
30	3.30 (0.75)	3.31 (0.10)	3.28 (0.03)	3.77 (0.79)	3.23 (0.73)	3.27 (0.77)
100	4.99 (0.75)	7.60 (0.08)	7.62 (0.13)	5.48 (0.85)	4.60 (0.62)	4.60 (0.63)
200	6.85 (0.59)	8.36 (0.79)	8.65 (0.49)	7.40 (0.28)	6.30 (0.25)	6.27 (0.63)
(Max norm)						
30	0.73 (0.18)	0.74 (0.17)	0.73 (0.18)	0.75 (0.19)	0.72 (0.17)	0.72 (0.16)
100	0.57 (0.08)	0.62 (0.05)	0.62 (0.05)	0.56 (0.09)	0.55 (0.07)	0.50 (0.07)
200	0.54 (0.07)	0.54 (0.07)	0.52 (0.08)	0.54 (0.07)	0.52 (0.07)	0.51 (0.07)
(Frobenius norm)						
30	5.59 (0.58)	5.70 (0.59)	5.58 (0.57)	5.47 (0.76)	5.17 (0.58)	4.98 (0.63)
100	12.62 (0.40)	13.58 (0.05)	13.49 (0.24)	10.26 (0.57)	10.06 (0.40)	7.91 (0.48)
200	22.54 (0.39)	14.21 (0.09)	13.94 (0.02)	16.04 (0.53)	18.71 (0.39)	14.37 (0.48)

Table 3. Simulation results for Model 3 over 100 replications. The standard errors are given in parentheses.

p	SE	ATE	OTE	POET	LOREC	L_q-CSCE
(Spectral norm)						
30	5.59 (1.04)	5.57 (0.85)	5.53 (1.03)	5.73 (1.09)	5.57 (1.04)	5.33 (0.99)
100	10.09 (0.77)	9.71 (0.76)	10.01 (0.77)	10.19 (0.78)	9.52 (0.76)	9.21 (0.96)
200	15.00 (0.93)	13.17 (0.84)	14.22 (0.89)	15.69 (0.95)	14.41 (0.93)	11.28 (1.11)
(Max norm)						
30	1.63 (0.30)	1.70 (0.28)	1.63 (0.30)	1.65 (0.29)	1.63 (0.30)	1.59 (0.29)
100	1.71 (0.29)	1.71 (0.29)	1.71 (0.29)	1.71 (0.29)	1.67 (0.29)	1.72 (0.37)
200	1.76 (0.22)	1.76 (0.22)	1.75 (0.22)	1.76 (0.22)	1.75 (0.21)	1.60 (0.16)
(Frobenius norm)						
30	10.44 (0.88)	11.09 (0.94)	10.34 (0.87)	10.54 (0.91)	10.39 (0.88)	10.10 (0.84)
100	27.99 (0.77)	27.48 (0.76)	27.78 (0.77)	27.76 (0.78)	25.69 (0.76)	19.66 (0.96)
200	52.86 (0.75)	49.25 (0.74)	50.16 (0.73)	49.76 (0.76)	49.09 (0.74)	32.23 (0.78)

Table 4. Record of the number of positive semi-definite matrices over 100 replications.

Method	$p = 30$		$p = 100$		$p = 200$	
	L	S	L	S	L	S
Model1						
ATE	-	-	-	-	-	-
OTE	-	-	-	-	-	-
POET	100	0	100	0	100	0
LOREC	100	86	100	0	100	0
L_q-CSCE	100	100	100	100	100	100
Model2						
ATE	-	-	-	-	-	-
OTE	-	-	-	-	-	-
POET	100	100	100	2	100	0
LOREC	0	100	53	100	100	100
L_q-CSCE	100	100	100	100	100	100
Model3						
ATE	-	-	-	-	-	-
OTE	-	-	-	-	-	-
POET	100	0	100	0	100	0
LOREC	31	0	0	25	0	0
L_q-CSCE	100	100	100	100	100	100

6.2. Choice of q

In the previous simulations, we only consider the setting where $q = 0.5$. This subsection supplements some simulations to study the influence of the choice of q on the performance of L_q-CSCE. Here, we consider the cases $q = 0.1, 0.3, 0.5\ 0.7$, and 0.9.

Tables 5–7 report the means and standard errors of the L_q-CSCE with different q under the three losses. Note that the options for q did not affect the result of the L_q-CSCE under the max norm loss. Under other losses, for different models, the performance of the L_q-CSCE depends on the choice of q deeply or shallowly. Specifically, there was little difference between $q = 0.1$, $q = 0.3$, and $q = 0.5$ under the spectral norm loss and the max norm loss. The L_q-CSCE with $q = 0.7$ and $q = 0.9$ performed poorly under the Frobenius norm, especially when $q = 0.9$. Note that, when $q = 0.1$, the L_q-CSCE is slightly worse than that of $q = 0.3$ and $q = 0.5$. We guess that a smaller q would make the minimizing functional more non-convex and, thus, more difficult to solve. The estimator induced by the L_q function with a large q had a large deviation.

Problem (23) has an explicit solution when $q = 0.5$, which makes the iteration of the SAU method simple. When $q = 0.1, 0.3, 0.7$, and 0.9, it is necessary to embed other numerical methods to find the approximate solution of Problem (23). Therefore, in practical applications, we recommend the setting where $q = 0.5$.

Table 5. Simulation results of the L_q-CSCE with different q for Model 1 over 100 replications. The standard errors are given in parentheses.

p	$q = 0.1$	$q = 0.3$	$q = 0.5$	$q = 0.7$	$q = 0.9$
(Spectral norm)					
30	1.70 (0.49)	1.71 (0.48)	1.66 (0.51)	1.68 (0.46)	2.12 (0.43)
100	2.61 (0.54)	2.61 (0.54)	2.62 (0.55)	2.80 (0.46)	3.71 (0.32)
200	3.10 (0.52)	3.11 (0.52)	3.10 (0.53)	3.42 (0.49)	4.10 (0.46)
(Max norm)					
30	1.18 (0.31)	1.18 (0.32)	1.13 (0.30)	1.21 (0.32)	1.21 (0.33)
100	1.53 (0.33)	1.53 (0.32)	1.53 (0.34)	1.54 (0.31)	1.52 (0.28)
200	1.65 (0.26)	1.65 (0.27)	1.65 (0.33)	1.66 (0.32)	1.66 (0.30)
(Frobenius norm)					
30	3.12 (0.52)	3.15 (0.51)	3.10 (0.54)	3.73 (0.53)	4.45 (0.46)
100	6.38 (0.54)	6.34 (0.54)	6.34 (0.52)	6.43 (0.64)	7.21 (0.55)
200	9.31 (0.53)	9.22 (0.55)	9.21 (0.57)	9.70 (0.51)	11.12 (0.49)

Table 6. Simulation results of the L_q-CSCE with different q for Model 2 over 100 replications. The standard errors are given in parentheses.

p	$q = 0.1$	$q = 0.3$	$q = 0.5$	$q = 0.7$	$q = 0.9$
(Spectral norm)					
30	3.41 (0.75)	3.41 (0.74)	3.40 (0.72)	3.39 (0.68)	3.38 (0.62)
100	4.81 (0.75)	4.75 (0.10)	4.77 (0.03)	4.74 (0.79)	4.83 (0.79)
200	6.35 (0.75)	6.30 (0.10)	6.31 (0.03)	6.40 (0.79)	6.41 (0.79)
(Max norm)					
30	0.73 (0.25)	0.74 (0.18)	0.74 (0.18)	0.73 (0.18)	0.74 (0.19)
100	0.57 (0.11)	0.53 (0.10)	0.53 (0.07)	0.53 (0.04)	0.54 (0.05)
200	0.53 (0.10)	0.50 (0.07)	0.51 (0.09)	0.52 (0.07)	0.54 (0.08)
(Frobenius norm)					
30	5.58 (0.62)	5.36 (0.65)	5.33 (0.64)	5.32 (0.61)	5.52 (0.56)
100	8.15 (0.62)	8.09 (0.52)	8.11 (0.50)	8.30 (0.52)	9.14 (0.57)
200	14.81 (0.65)	14.49 (0.57)	14.62 (0.53)	14.74 (0.56)	15.55 (0.49)

Table 7. Simulation results of the L_q-CSCE with different q for Model 3 over 100 replications. The standard errors are given in parentheses.

p	$q = 0.1$	$q = 0.3$	$q = 0.5$	$q = 0.7$	$q = 0.9$
(Spectral norm)					
30	5.80 (1.01)	5.63 (0.99)	5.51 (0.98)	5.71 (0.95)	5.65 (1.00)
100	8.91 (1.04)	8.89 (0.90)	8.97 (0.93)	8.85 (0.84)	9.21 (0.79)
200	12.40 (1.13)	12.20 (1.04)	12.21 (1.01)	12.34 (1.06)	13.37 (0.99)
(Max norm)					
30	1.60 (0.35)	1.56 (0.29)	1.56 (0.28)	1.56 (0.31)	1.60 (0.29)
100	1.73 (0.43)	1.73 (0.31)	1.74 (0.34)	1.76 (0.33)	1.69 (0.31)
200	1.71 (0.29)	1.71 (0.22)	1.72 (0.19)	1.73 (0.19)	1.78 (0.21)
(Frobenius norm)					
30	10.12 (0.91)	9.98 (0.87)	10.03 (0.89)	10.14 (0.84)	10.42 (0.86)
100	20.60 (1.03)	20.01 (0.99)	20.04 (1.01)	20.45 (0.95)	21.36 (0.83)
200	31.56 (0.99)	31.40 (0.90)	31.29 (0.84)	31.49 (0.79)	34.47 (0.72)

7. Discussion and Conclusions

In this paper, we considered the problem of estimating a high-dimensional covariance matrix from noisy data and studied a class of estimators based on the structural decomposition (2). We estimated the high-dimensional covariance matrix by solving a constrained adaptive L_q-type regularized optimization Problem (7). This paper provided a systematic study for Problem (7). We first defined a class of hybrid first-order stationary points of Problem (7) and analyzed the relationship between these stationary points and the global solution of Problem (7). Secondly, we proposed an SAU method for solving Problem (7) and established its convergence. The final simulation results showed that the proposed SAU method can solve Problem (7) effectively and that the final L_q-CSCE had good numerical performance.

8. Proof

This section provides the detailed proof of the conclusions in this paper.

Proof of Proposition 1. $(i)(ii)$ The properness of $F(L, S)$ is obvious. Note that F is continuous over its domain and $\mathrm{dom}(F)$ is closed. Theorem 2.8 in [31] shows that F is closed. Furthermore, whether $\|L\|_F \to \infty$ or $\|S\|_F \to \infty$, at least one of the three terms in F tends to infinity. Thus, the coerciveness of $F(L, S)$ holds. Then, it follows from Theorem 2.14 in [31] that $F(L, S)$ attains its minimal value over $\mathbb{S}_+^p \times \mathbb{S}_+^p$.

(iii) By calculation, $\nabla f(L, S) = [\nabla_L^T f(L, S), \nabla_S^T f(L, S)]^T$, where

$$\nabla_L f(L, S) = \nabla_S f(L, S) = L + S - \widetilde{\Sigma}.$$

For any $(L_1, S_1), (L_2, S_2) \in \mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p}$, it holds that

$$
\begin{aligned}
\|\nabla f(L_1, S_1) - \nabla f(L_2, S_2)\|_F^2 &= \|\nabla_L f(L_1, S_1) - \nabla_L f(L_2, S_2)\|_F^2 + \|\nabla_S f(L_1, S_1) - \nabla_S f(L_2, S_2)\|_F^2 \\
&= 2\|L_1 - L_2 + S_1 - S_2\|_F^2 \\
&= 2\|L_1 - L_2\|_F^2 + 2\|S_1 - S_2\|_F^2 + 4\langle L_1 - L_2, S_1 - S_2 \rangle \\
&\leq 4\|L_1 - L_2\|_F^2 + 4\|S_1 - S_2\|_F^2 \\
&= 4\|(L_1, S_1) - (L_2, S_2)\|_F^2,
\end{aligned}
$$

which implies that the gradient $\nabla f(L, S)$ is Lipschitz continuous with a constant $\ell = 2$.

(iv) By calculation, $\nabla_S f(L^*, S) = L^* + S - \widetilde{\Sigma}$. For any $S_1, S_2 \in \mathbb{R}^{p \times p}$, it holds that $\nabla_S f(L^*, S_1) - \nabla_S f(L^*, S_2) = S_1 - S_2$. Then, the partial gradient function $\nabla_S f(L^*, S)$ is Lipschitz continuous with a constant $\ell = 1$. The same result holds for the partial gradient function $\nabla_L f(L, S^*)$. $\square$

Proof of Theorem 1. Since $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ is a global minimizer of Problem (7), then $\boldsymbol{L}^*$ is also a global minimizer of the problem:

$$\min_{\boldsymbol{L} \in \mathbb{S}_+^p} f(\boldsymbol{L}, \boldsymbol{S}^*) + \lambda_L \|\boldsymbol{L}\|_q^q.$$

By (iv) of Proposition 1, we have that

$$f(\boldsymbol{L}_1, \boldsymbol{S}) \le f(\boldsymbol{L}_2, \boldsymbol{S}) + \langle \nabla_L f(\boldsymbol{L}_2, \boldsymbol{S}), \boldsymbol{L}_1 - \boldsymbol{L}_2 \rangle + \frac{1}{2} \|\boldsymbol{L}_1 - \boldsymbol{L}_2\|_F^2, \ \forall \ \boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathbb{R}^{p \times p}.$$

It follows that

$$\begin{aligned}
f(\boldsymbol{L}^*, \boldsymbol{S}^*) + \lambda_L \|\boldsymbol{L}^*\|_q^q &\le f(\boldsymbol{L}, \boldsymbol{S}^*) + \lambda_L \|\boldsymbol{L}\|_q^q \\
&\le f(\boldsymbol{L}, \boldsymbol{S}^*) + \lambda_L \|\boldsymbol{L}\|_q^q + \frac{\ell - 1}{2} \|\boldsymbol{L} - \boldsymbol{L}^*\|_F^2 \\
&\le f(\boldsymbol{L}^*, \boldsymbol{S}^*) + \langle \nabla_L f(\boldsymbol{L}^*, \boldsymbol{S}^*), \boldsymbol{L} - \boldsymbol{L}^* \rangle + \frac{1}{2} \|\boldsymbol{L} - \boldsymbol{L}^*\|_F^2 + \lambda_L \|\boldsymbol{L}\|_q^q + \frac{\ell - 1}{2} \|\boldsymbol{L} - \boldsymbol{L}^*\|_F^2 \\
&= f(\boldsymbol{L}^*, \boldsymbol{S}^*) + \langle \nabla_L f(\boldsymbol{L}^*, \boldsymbol{S}^*), \boldsymbol{L} - \boldsymbol{L}^* \rangle + \frac{\ell}{2} \|\boldsymbol{L} - \boldsymbol{L}^*\|_F^2 + \lambda_L \|\boldsymbol{L}\|_q^q.
\end{aligned}$$

Thus,

$$\begin{aligned}
\boldsymbol{L}^* &\in \operatorname*{argmin}_{\boldsymbol{L} \in \mathbb{S}_+^p} \left\{ f(\boldsymbol{L}^*, \boldsymbol{S}^*) + \langle \nabla_L f(\boldsymbol{L}^*, \boldsymbol{S}^*), \boldsymbol{L} - \boldsymbol{L}^* \rangle + \frac{\ell}{2} \|\boldsymbol{L} - \boldsymbol{L}^*\|_F^2 + \lambda_L \|\boldsymbol{L}\|_q^q \right\} \\
&= \operatorname{Proxi}_{(\lambda_L/\ell)\|\boldsymbol{L}\|_q^q, \mathbb{S}_+^p} \left(\boldsymbol{L}^* - \frac{1}{\ell}(\boldsymbol{L}^* + \boldsymbol{S}^* - \widetilde{\boldsymbol{\Sigma}}) \right).
\end{aligned}$$

Hence, (9) holds.

We now prove that (10) holds for $(\boldsymbol{L}^*, \boldsymbol{S}^*)$. Observe that $\boldsymbol{S}^*$ is a global minimizer of the problem:

$$\min_{\boldsymbol{S} \in \mathbb{S}_+^p} f(\boldsymbol{L}^*, \boldsymbol{S}) + \sum_{(u,v) \in D^c} \lambda_{uv}^S |S_{uv}|^q,$$

which can be equivalently rewritten as

$$\min_{\boldsymbol{S} \in \mathbb{R}^{p \times p}} f(\boldsymbol{L}^*, \boldsymbol{S}) + \sum_{(u,v) \in D^c} \lambda_{uv}^S |S_{uv}|^q + \delta_{\mathbb{S}_+^p}(\boldsymbol{S}).$$

Then, $\boldsymbol{S}_{\Gamma^*}^*$ is a global minimizer of the problem:

$$\min_{\boldsymbol{S}_{\Gamma^*}} f(\boldsymbol{L}^*, \boldsymbol{S}_{\Gamma^*}, \boldsymbol{S}_{(\Gamma^*)^c}^*) + \sum_{(u,v) \in D^c \cap \Gamma^*} \lambda_{uv}^S |S_{uv}|^q + \delta_{\mathbb{S}_+^p}([\boldsymbol{S}_{\Gamma^*} : \boldsymbol{S}_{(\Gamma^*)^c}^*]). \tag{26}$$

It follows from the first-order optimality condition of Problem (26) and Exercise 10.10 in [32] that there exists a $\boldsymbol{Y}_{\Gamma^*} \in \partial h_{\Gamma^*}(\boldsymbol{S}_{\Gamma^*})$ such that

$$0 = L_{uv}^* + S_{uv}^* - \widetilde{\Sigma}_{uv} + \lambda_{uv}^{S*} \operatorname{sign}(S_{uv}^*) |S_{uv}^*|^{q-1} + Y_{uv}, \text{ for all } (u,v) \in \Gamma^* \cap D^c, \tag{27}$$

$$0 = L_{uv}^* + S_{uv}^* - \widetilde{\Sigma}_{uv} + Y_{uv}, \text{ for all } (u,v) \in \Gamma^* \cap D. \tag{28}$$

Moreover, let $\boldsymbol{Y} = [\boldsymbol{Y}_{\Gamma^*} : \boldsymbol{Y}_{(\Gamma^*)^c}]$, where $\boldsymbol{Y}_{(\Gamma^*)^c}$ is some finite array. For any $(u,v) \in (\Gamma^*)^c \cap D^c$, we have

$$S_{uv}^*(L_{uv}^* + S_{uv}^* - \widetilde{\Sigma}_{uv}) + \lambda_{uv}^S |S_{uv}|^q + S_{uv}^* Y_{uv} = 0, \tag{29}$$

and for any $(u,v) \in (\Gamma^*)^c \bigcap D$, we have

$$S_{uv}^*(L_{uv}^* + S_{uv}^* - \widetilde{\Sigma}_{uv}) + S_{uv}^* Y_{uv} = 0, \tag{30}$$

Combining (27)-(30) with the definition of Y, (10) holds. Thus, (L^*, S^*) is a first-order stationary point of Problem (7). $\square$

Proof of Theorem 2. Since (L^*, S^*) is a first-order stationary point of Problem (7) and $S^* \in \mathbb{S}_{++}^p$, by Definition 1, it holds that

$$\mathbf{0} = S^* \circ (L^* + S^* - \widetilde{\Sigma}) + Q^*, \tag{31}$$

where $Q^* := (Q_{uv}^*)_{u,v=1}^p$ and

$$Q_{uv}^* = \begin{cases} q\lambda_{uv}^S |S_{uv}^*|^q, & \text{if } (u,v) \in D^c, \\ 0, & \text{if } (u,v) \in D. \end{cases}$$

For all $(u,v) \in D^c \bigcap \Gamma^*$, by (31), we have

$$0 = S_{uv}^*(L_{uv}^* + S_{uv}^* - \widetilde{\Sigma}_{uv}) + q\lambda_{uv}^S |S_{uv}^*|^q. \tag{32}$$

Moreover, the statement (iv) in Proposition 1 shows that $\nabla_S f(L, S)$ is Lipschitz continuous with a constant 1, and it follows that

$$f(L^*, Y) \leq f(L^*, X) + \langle \nabla_S f(L^*, X), Y - X \rangle + \frac{1}{2} \|Y - X\|_F^2, \; \forall \, X, Y \in \mathbb{R}^{p \times p}.$$

Taking $X = S^*$ and $Y = S^* - \nabla_S f(L^*, S^*)$, we obtain that

$$f(L^*, S^* - \nabla_S f(L^*, S^*)) \leq f(L^*, S^*) - \frac{1}{2} \|\nabla_S f(L^*, S^*)\|_F^2.$$

Note that

$$f(L^*, S^* - \nabla_S f(L^*, S^*)) \geq 0, \; f(L^*, S^*) \leq F(L^*, S^*) \leq F(L^0, S^0) + \epsilon.$$

It follows that

$$\|\nabla_S f(L^*, S^*)\|_F \leq \sqrt{2[F(L^0, S^0) + \epsilon]}.$$

This, together with the relation (32), implies that

$$q\lambda_{uv}^S |S_{uv}^*|^{q-1} = |(\nabla_S f(L^*, S^*))_{uv}| \leq \|\nabla_S f(L^*, S^*)\|_F \leq \sqrt{2[F(L^0, S^0) + \epsilon]}, \; \forall (u,v) \in D^c \bigcap \Gamma^*.$$

Thus, we have

$$|S_{uv}^*| \geq \left(\frac{q \min\{\lambda_{uv}^S\}}{\sqrt{2[F(L^0, S^0) + \epsilon]}} \right)^{\frac{1}{1-q}} \forall (u,v) \in D^c \bigcap \Gamma^*,$$

which completes this proof. $\square$

Proof of Theorem 3. (i) We just need to prove that Relation (20) is a sufficient condition for (14) and the inequality (15) holds. Since $(L_{\mu_k}^*, S_{\mu_k}^*)$ is a first-order stationary point of Problem (16), there exists a matrix $Y_{\mu_k}^* \in N_{\mathbb{S}_+^p}(S^*)$ such that

$$\mathbf{0} = \nabla_S F_{\mu_k}(L_{\mu_k}^*) + Y_{\mu_k}^*.$$

Let $\mathcal{N} := \{(u,v) \in D^c : (S_{\mu_k}^*)_{uv} > \mu\}$. It holds that

$$\mathbf{0} = (L_{\mu_k}^*)_{\mathcal{N} \cup D} + (S_{\mu_k}^*)_{\mathcal{N} \cup D} - \widetilde{\Sigma}_{\mathcal{N} \cup D} + [\Psi(S_{\mu_k}^*)]_{\mathcal{N} \cup D} + (Y_{\mu_k}^*)_{\mathcal{N} \cup D}. \tag{33}$$

By (11), it is known that

$$(Y^*_{\mu_k})_{\mathcal{N} \cup D} \in \mathcal{L}^*_{\mathcal{N} \cup D}(N_{\mathbb{S}^p_+}(S^*_{\mu_k})) \subseteq \partial h_{\mathcal{N} \cup D}((S^*_{\mu_k})_{\mathcal{N} \cup D}).$$

Then,

$$\text{dist}\left(\mathbf{0}, (L^*_{\mu_k})_{\mathcal{N} \cup D} + (S^*_{\mu_k})_{\mathcal{N} \cup D} - \widetilde{\Sigma}_{\mathcal{N} \cup D} + [\Psi(S^*_{\mu_k})]_{\mathcal{N} \cup D} + (Y_{\mu_k})_{\mathcal{N} \cup D} : Y_{\mathcal{N} \cup D} \in \partial h_{\mathcal{N} \cup D}((S^*_{\mu_k})_{\mathcal{N} \cup D})\right) \leq \varepsilon.$$

Moreover, it follows from the condition that $\mu \leq \varepsilon/p^2$ that

$$\|(S^*_{\mu_k})_{(\Gamma^* \cup D)^c}\|_F \leq \sum_{(u,v) \in (\Gamma^* \cup D)^c} |(S^*_{\mu_k})_{uv}| \leq \mu p^2 \leq \varepsilon.$$

Therefore, $(L^*_{\mu_k}, S^*_{\mu_k})$ is an ε-approximate first-order stationary point of Problem (7).

(ii) Let (L^*, S^*) be an accumulation point of $(L^*_{\mu_k}, S^*_{\mu_k})$. By the definition of the first-order stationary point of Problem (16), for any $(u, v) \in \mathcal{N}^c \cap D^c$, there exists a matrix $Y^*_{\mu_k} \in N_{\mathbb{S}^p_+}(S^*_{\mu_k})$ such that

$$0 = (L^*_{\mu_k})_{uv} + (S^*_{\mu_k})_{uv} - \widetilde{\Sigma}_{uv} + q\lambda_{uv}\left(\frac{(S^*_{\mu_k})^2_{uv}}{2\mu_k} + \frac{\mu_k}{2}\right)^{q-1}\frac{(S^*_{\mu_k})_{uv}}{\mu_k} + (Y^*_{\mu_k})_{uv}. \tag{34}$$

Note that

$$0 \leq q\lambda_{uv}\left(\frac{(S^*_{\mu_k})^2_{uv}}{2\mu_k} + \frac{\mu_k}{2}\right)^{q-1}\frac{(S^*_{\mu_k})^2_{uv}}{\mu_k} \leq q\lambda_{uv}\left(\frac{(S^*_{\mu_k})^2_{uv}}{\mu_k}\right)^q \leq q\lambda^S_{uv}|(S^*_{\mu_k})_{uv}|^q. \tag{35}$$

If $|(S^*_{\mu_k})_{uv}| \leq \mu_k$ for arbitrarily large k, then $S^*_{uv} = \lim_{k\to\infty}(S^*_{\mu_k})_{uv} = 0$, and by (35), we have

$$\lim_{k\to\infty} q\lambda_{uv}\left(\frac{(S^*_{\mu_k})^2_{uv}}{2\mu_k} + \frac{\mu_k}{2}\right)^{q-1}\frac{(S^*_{\mu_k})^2_{uv}}{\mu_k} = 0 \text{ and } q\lambda^S_{uv}|S^*_{uv}|^q = 0.$$

Since $Y^*_{\mu_k} \in N_{\mathbb{S}^p_+}(S^*_{\mu_k})$, $\lim_{k\to\infty} Y^*_{\mu_k} \in N_{\mathbb{S}^p_+}(S^*)$.

This, together with the relation (34), implies that, for any $(u, v) \in \mathcal{N}^c \cap D^c$,

$$0 \in \left\{S^*_{uv}(L^*_{uv} + S^*_{uv} - \widetilde{\Sigma}_{uv}) + q\lambda^S_{uv}|S^*_{uv}|^q + S^*_{uv}Y_{uv}, \mathbf{Y} \in N_{\mathbb{S}^p_+}(S^*)\right\}. \tag{36}$$

Moreover, for any $(u, v) \in \mathcal{N} \cup D$, we have

$$0 = (L^*_{\mu_k})_{uv} + (S^*_{\mu_k})_{uv} - \widetilde{\Sigma}_{uv} + Q^{\mu_k}_{uv} + (Y^*_{\mu_k})_{uv}, \tag{37}$$

where

$$Q^{\mu_k}_{uv} = \begin{cases} 0, & (u,v) \in D, \\ q\lambda^S_{uv}\text{sign}((S^*_{\mu_k})_{uv})|(S^*_{\mu_k})_{uv}|^{q-1}, & (u,v) \in (\mathcal{N} \cup D)/D. \end{cases}$$

For any $(u, v) \in \mathcal{N} \cup D$, it follows that

$$0 \in \left\{S^*_{uv}(L^*_{uv} + S^*_{uv} - \widetilde{\Sigma}_{uv}) + S^*_{uv}Q_{uv} + S^*_{uv}Y_{uv}, \mathbf{Y} \in N_{\mathbb{S}^p_+}(S^*)\right\}, \tag{38}$$

where

$$Q_{uv} = \begin{cases} 0, & (u,v) \in D, \\ q\lambda^S_{uv}\text{sign}(S^*_{uv})|S^*_{uv}|^{q-1}, & (u,v) \in (\mathcal{N} \cup D)/D. \end{cases}$$

Since the $Y^*_{\mu_k}$ in (34) coincides with that in (37), by (36) and (38), we have that (L^*, S^*) is the first-order stationary point of Problem (16). $\square$

Proof of Theorem 5. (i) By (iv) of Proposition 1, we have that

$$\frac{1}{2}\|L_1 + S - \widetilde{\Sigma}\|_F^2 \leq \frac{1}{2}\|L_2 + S - \widetilde{\Sigma}\|_F^2 + \langle L_2 + S - \widetilde{\Sigma}, L_1 - L_2 \rangle + \frac{1}{2}\|L_1 - L_2\|_F^2, \ \forall \ L_1, L_2 \in \mathbb{R}^{p \times p}.$$

It follows that

$$
\begin{aligned}
F_\mu(L^{m+1}, S^m) &= \frac{1}{2}\|L^{m+1} + S^m - \widetilde{\Sigma}\|_F^2 + \lambda_L \|L^{m+1}\|_q^q + \sum_{(u,v)\in D^c} \lambda_{uv}^S \psi_\mu(S_{uv}^m)^q, \\
&\leq \frac{1}{2}\|L^m + S^m - \widetilde{\Sigma}\|_F^2 + \langle L^m + S^m - \widetilde{\Sigma}, L^{m+1} - L^m \rangle + \frac{1}{2}\|L^{m+1} - L^m\|_F^2 + \lambda_L \|L^{m+1}\|_q^q \\
&\qquad + \sum_{(u,v)\in D^c} \lambda_{uv}^S \psi_\mu(S_{uv}^m)^q \\
&\leq \frac{1}{2}\|L^m + S^m - \widetilde{\Sigma}\|_F^2 + \langle L^m + S^m - \widetilde{\Sigma}, L^{m+1} - L^m \rangle + \frac{\ell}{2}\|L^{m+1} - L^m\|_F^2 + \lambda_L \|L^{m+1}\|_q^q \\
&\qquad + \sum_{(u,v)\in D^c} \lambda_{uv}^S \psi_\mu(S_{uv}^m)^q - \frac{\ell-1}{2}\|L^{m+1} - L^m\|_F^2 \\
&\leq \frac{1}{2}\|L^m + S^m - \widetilde{\Sigma}\|_F^2 + \lambda_L \|L^m\|_q^q + \sum_{(u,v)\in D^c} \lambda_{uv}^S \psi_\mu(S_{uv}^m)^q - \frac{\ell-1}{2}\|L^{m+1} - L^m\|_F^2 \\
&= F_\mu(L^m, S^m) - \frac{\ell-1}{2}\|L^{m+1} - L^m\|_F^2,
\end{aligned}
$$

where the third inequality is due to the iteration of L in AU. Then,

$$
\begin{aligned}
F_\mu(L^{m+1}, S^{m+1}) &\leq F_\mu(L^{m+1}, S^m) - \frac{c}{2}\|S^{m+1} - S^m\|_F^2 \\
&\leq F_\mu(L^m, S^m) - \frac{c}{2}\|S^{m+1} - S^m\|_F^2 - \frac{\ell-1}{2}\|L^{m+1} - L^m\|_F^2.
\end{aligned}
$$

(ii) It is easy to show that F_μ is coercive, which implies that F_μ is level bounded. By (i) in this Theorem, we have $F_\mu(L^m, S^m) \leq F_\mu(L^0, S^0)$. Then, $\{(L^m, S^m)\}$ is bounded.

(iii) It follows from the inequality in (i) that

$$
\begin{aligned}
\sum_{m=0}^\infty \left(\frac{c}{2}\|S^{m+1} - S^m\|_F^2 + \frac{\ell-1}{2}\|L^{m+1} - L^m\|_F^2 \right) &\leq \sum_{m=0}^\infty \left(F_\mu(L^m, S^m) - F_\mu(L^{m+1}, S^{m+1}) \right) \\
&\leq F_\mu(L^0, S^0) - \lim_{m\to\infty} F_\mu(L^m, S^m) < \infty,
\end{aligned}
$$

where the last inequality is due to F_μ being bounded from below. Hence, $\lim_{m\to\infty} \|L^{m+1} - L^m\|_F = 0$, $\lim_{m\to\infty} \|S^{m+1} - S^m\|_F = 0$.

(iv) Let (L^*, S^*) be an accumulation point of (L^m, S^m). We assumed that $\lim_{j\to\infty} L^{m_j} = L^*$ and $\lim_{j\to\infty} S^{m_j} = S^*$, where $m_j \to \infty$ as $j \to \infty$. By the results in (iii),

$$\lim_{j\to\infty} L^{m_j+1} = \lim_{j\to\infty} L^{m_j} + (L^{m_j+1} - L^{m_j}) = L^*, \quad \lim_{j\to\infty} S^{m_j+1} = \lim_{j\to\infty} S^{m_j} + (S^{m_j+1} - S^{m_j}) = S^*.$$

By the AU method in this paper, we have that

$$L^{m_j+1} = \operatorname*{argmin}_{L \in \mathbb{S}_+^p} \ \langle \nabla_L f(L^{m_j}, S^{m_j}), L - L^* \rangle + \frac{\ell}{2}\|L - L^{m_j}\|_F^2 + \lambda_L \|L\|_q^q,$$

$$\langle S^{m_j+1} - (S^{m_j} - \alpha_{m_j} \nabla_S F_\mu(L^{m_j}, S^{m_j})), S - S^{m_j+1} \rangle \geq 0, \ \forall \ S \in \mathbb{S}_+^p.$$

Taking the limits on both sides, it follows that

$$\boldsymbol{L}^* = \underset{\boldsymbol{L} \in \mathbb{S}^p_+}{\text{argmin}} \ \langle \nabla_L f(\boldsymbol{L}^*, \boldsymbol{S}^*), \boldsymbol{L} - \boldsymbol{L}^* \rangle + \frac{\ell}{2} \|\boldsymbol{L} - \boldsymbol{L}^{m_j}\|_F^2 + \lambda_L \|\boldsymbol{L}\|_q^q,$$

$$\langle \nabla_S F_\mu(\boldsymbol{L}^*, \boldsymbol{S}^*), \boldsymbol{S} - \boldsymbol{S}^* \rangle \geq 0, \ \forall \ \boldsymbol{S} \in \mathbb{S}^p_+,$$

i.e., $(\boldsymbol{L}^*, \boldsymbol{S}^*)$ is a first-order stationary point of Problem (16). □

Author Contributions: Methodology, X.W., L.K., and L.W.; software, X.W. and Z.Y.; writing—original draft, X.W., L.K., and L.W.; validation, X.W., L.K., L.W., and Z.Y. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bickel, P.; Levian, E. Covariance regularization by thresholding. *Ann. Stat.* **2008**, *36*, 2577–2604. [CrossRef]
2. El Karoui, P. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Stat.* **2008**, *36*, 2717–2756. [CrossRef]
3. Rothman, A.; Levian, E.; Zhu, J. Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **2009**, *104*, 177–186. [CrossRef]
4. Cai, T.; Liu, W. Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.* **2011**, *106*, 672–684. [CrossRef]
5. Cai, T.; Zhou, H. Minimax estimation of large covariance matrices under ℓ_1 norm. *Stat. Sinica* **2008**, *36*, 2577–2604.
6. Rothman, A. Positive definite estimators of large covariance matrices. *Biometrika* **2012**, *99*, 733–740. [CrossRef]
7. Xue, L.; Ma, S.; Zou, H. Positive definite L_1 penalized estimation of large covariance matrices. *J. Am. Stat. Assoc.* **2012**, *107*, 1480–1491. [CrossRef]
8. Cui, Y.; Leng, C.; Sun, D. Sparse estimation of high-dimensional correlation matrices. *Comput. Stat. Data. An.* **2016**, *93*, 390–403. [CrossRef]
9. Avella-Medina, M.; Battey, H.; Fan, J.; Li, Q. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **2018**, *105*, 271–284. [CrossRef]
10. Fan, J.; Fan, Y.; Lv, J. High dimensional covariance matrix estimation using a factor model. *J. Econom.* **2008**, *147*, 186–197. [CrossRef]
11. Fan, J.; Liao, Y.; Mincheva, M. High dimensional covariance matrix estimation in appriximate. *Ann. Stat.* **2011**, *39*, 3320–3356. [CrossRef] [PubMed]
12. Fan, J.; Liao, Y.; Mincheva, M. Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B* **2013**, *75*, 603–680. [CrossRef] [PubMed]
13. Fan, J.; Wang, W.; Zhong, Y. Robust covariance estimation for approximate factor models. *J. Econom.* **2019**, *208*, 5–22. [CrossRef] [PubMed]
14. Luo, X. High dimensional low rank and sparse covariance matrix estimation via convex minimization. *Biometrika* **2018**, *105*, 271–284.
15. Candès, E.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM* **2011**, *58*, 1–37.
16. Feng, J.; Xu, H.; Yan, S. Online robust PCA via stochastic optimization. In Proceedings of the 26th International Conference on Neural Information Processing Systems, New York, NY, USA, 5 December 2013; pp. 404–412.
17. Kang, Z.; Peng, C.; Cheng, Q. Robust PCA via nonconvex rank approximation. In Proceedings of 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015; pp. 1550–4786.
18. Song, W.; Zhu, J.; Li, Y.; Chen, C. Image alignment by online robust PCA via stochastic gradient descent. *IEEE Trans. Circ. Syst. Vid* **2016**, *26*, 1241–1250. [CrossRef]
19. Bouwmans, T.; Jaced, S.; Zhang, H.; Lin, Z.; Ricardo, O. On the applications of robust PCA in image and video processing . *Proc. IEEE* **2018**, *106*, 1427–1457. [CrossRef]
20. Javed, S.; Mahmood, A.; Al-Maadeed, S.; Bouwmans, T; Jung, S.K. Moving object detection in complex scene using spatiotemporal structured-sparse RPCA . *IEEE Trans. Image Process* **2018**, *28*, 1007–1022. [CrossRef]

21. Liu, Q.; Li, X. Efficient low-rank matrix factorization based on $\ell_{1,\epsilon}$-norm for online background subtraction. *IEEE T Circ. Syst. Vid* **2022**, *32*, 4900–4904. [CrossRef]

22. Fan, J.; Li, Q.; Wang, Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B* **2017**, *79*, 247–265. [CrossRef]

23. Chen, X.; Xu, F.; Ye, Y. Lower bound theory of nonzero entries in solutions of L_2-L_p minimization. *SIAM J. Sci. Comput.* **2010**, *32*, 2832–2852. [CrossRef]

24. Lu, Z. Iterative reweighted minimization methods for L_p regularized unconstrained nonlinear programming. *Math. Program.* **2014**, *147*, 277–307. [CrossRef]

25. Peng, D.; Xiu, N.; Yu, J. Global optimality and fixed point continuation algorithm for non-Lipschitz ℓ_p regularized matrix minimization. *Sci. China Math.* **2018**, *61*, 171–184. [CrossRef]

26. Rockafellar, R.T. *Convex Analysis*, 2nd ed; Princeton University Press: Princeton, NJ, USA, 1970.

27. Mordukhovich, B.S.; Nguyen, M.N. *An Easy Path to Convex Analysis and Applications*; Morgan and Claypool: San Rafael, CA, USA, 2014.

28. Zhang, C.; Chen, X. Smoothing prijected gradient method and its application to stochastic linear complementarity problems. *SIAM J. Optimiz.* **2009**, *20*, 627–649. [CrossRef]

29. Chen, X.; Ng, M.K.; Zhang, C. Nonconvex L_p regularization and box constrained model for image restoration. *IEEE T Image Process* **2012**, *21*, 4709–4721. [CrossRef]

30. Chen, Y.; Xiu, N.; Peng, D. Global solutions of non-Lipschitz S_2-S_p minimization over the positive semidefinite cone. *Optim. Lett.* **2014**, *8*, 2053–2064. [CrossRef]

31. Beck, A. *First-Order Methods in Optimization*; Society for Industrial and Applied Mathematics and Mathematical Optimization Society: Philadelphia, PA, USA, 2017.

32. Rockafellar, R.T.; Wets, R. J-B. *Variational Analysis*; Springer: Berlin/Heidelberg, Germany, 1998.