Factor-Based Large Covariance Matrix Estimation: Nonconvex Optimization and Statistical Guarantees

Shanshan Zou, *Student Member, IEEE* and Ziping Zhao, *Member, IEEE*

***Abstract*—In this paper, we study the problem of large covariance matrix estimation based on factor analysis models. In this setting, the covariance matrix is decomposed into a low-rank component and a sparse component. We formulate the estimation problem as a nonconvex optimization problem and propose an iterative algorithm to solve it. Our algorithm starts with the principal orthogonal complement thresholding (POET) estimator, a widely used optimization-free method for factor-based covariance estimation. We then refine this initial estimator through alternating optimization of the low-rank and sparse components. The resulting nonconvex estimator is referred to as POET with refining iteratively (POETRY). We theoretically prove that the algorithm converges linearly to the ground truth with an improved statistical rate compared to POET, achieving the minimax rate of convergence for factor-based covariance matrix estimation. This matches the statistical rate achieved by existing convex relaxation methods but with significantly reduced per-iteration computational complexity. We further establish the statistical rate of the inverse covariance matrix estimator. Additionally, by introducing the notion of effective rank, we are able to further reduce the sample size requirement and enhance the statistical rate. Numerical experiments validate our theoretical findings and demonstrate the superiority of the proposed algorithm compared to state-of-the-art methods.**

***Index Terms*—Covariance matrix estimation, factor models, DOA estimation, correlated noise, low-rank plus sparse, minimax rate, nonconvex statistical optimization.**

I. Introduction

Covariance matrix estimation is a fundamental problem with applications in various fields related to data analytics, including statistics [2], economics [3], [4], finance [5], [6], biology [7], signal processing [8], [9], and machine learning [10]. The sample covariance matrix (SCM) is widely used in practice due to its simplicity; however, it tends to perform poorly in high-dimensional settings where the number of variables is comparable to or larger than the sample size [11]. Recognizing the criticality of this problem, recent studies have developed diverse regularization techniques for consistent estimation of large covariance matrices. Prominent approaches exploit structural assumptions, including sparsity [4], [12], [13], low-rankness [14], [15], bandedness [16], Toeplitz [17], [18], and Hankel [19] structures.

In this paper, we focus on the estimation of large covariance matrices under the factor model assumption. Covariance matrix estimation based on factor models has wide-ranging applications, including return modeling in finance [20], [21], gene expression studies in genetics [22], spectral estimation in signal processing [23], [24], and dimensionality reduction in machine learning [10], [25].

A substantial body of literature has emerged on estimating factor-based covariance matrices. A classical assumption in this field is that the covariance matrix follows a strict factor model [26], [27], where it is decomposed into a low-rank component plus a diagonal component [23], [28], [29]. In [30], the estimation problem was formulated as a least squares problem and solved using principal component analysis (PCA) followed by diagonal thresholding, resulting in a direct method that does not require iterative optimization. Furthermore, an alternating projection algorithm was proposed in [31], which iteratively updates both the low-rank and diagonal components of the covariance matrix.

Although simple, the covariance matrix based on the strict factor model can be inefficient because the diagonal component modeling assumes uncorrelated errors. However, this condition is often violated in practical applications, where the noise is weakly or partially correlated. For example, in finance, asset returns are often modeled as a combination of systematic factors and asset-specific residuals. The corresponding return covariance matrix can be expressed as the sum of a low-rank component, which captures the common factors, and a sparse component, which reflects the residual covariances between assets [32], [33]. It is well documented that a small set of common factors, such as macroeconomic shocks, interest rates, or the market factor itself, drives the observed dynamics of a large number of time series. Conditional on these pervasive factors, the idiosyncratic errors are influenced by specific factors, such as sectors, countries, or asset classes, that affect only a smaller number of variables. This structure results in a sparse covariance component that aligns with the frameworks for sparse covariance estimation considered in [12], [16], [34]. Similarly, in array processing, the received signals can be effectively modeled using a factor model that combines a low-dimensional subspace representation with a noise component. This low-dimensional representation arises because the signal component originates from only a few distinct sources. The corresponding covariance matrix, therefore, exhibits a low-rank plus sparse structure, where the sparse noise covariance captures the spatially correlated noise. A sparse noise covariance matrix is a common model for an array of sensors composed of several widely separated subarrays. When the interelement spacing within a subarray is small, the noise among sensors within the subarray tends to be spatially correlated. By contrast, the noise between different subarrays can often be assumed uncorrelated due to the large distances separating them. As a result, the noise covariance matrix of such an

This paper was presented in part at the 49th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seoul, Korea, April 14–19, 2024 [1].

array has a block diagonal structure, which is indeed sparse. Even in a typical nonsparse array, the small spacing between adjacent sensors generally leads to noise coupling between neighboring sensors. Nevertheless, it is often reasonable to assume that nonadjacent sensors have spatially uncorrelated noise, leading to a banded structure that is also sparse [35], [36], [37]. To overcome the limitation of the strict factor model in modeling weakly correlated errors, researchers have turned to the approximate factor model as an alternative [32], [38], [39], [40], [41], [42], [43], [44]. In this case, the covariance matrix can be decomposed into two components: a low-rank matrix and a sparse matrix. Specifically, the true covariance matrix $\boldsymbol{\Sigma}_\star \in \mathbb{R}^{d\times d}$ is expressed as $\boldsymbol{\Sigma}_\star = \boldsymbol{L}_\star + \boldsymbol{\Psi}_\star$, where $\boldsymbol{L}_\star$ represents the low-rank component and $\boldsymbol{\Psi}_\star$ denotes the sparse component. These models utilize sparse components to represent the covariance of weakly corrected noise. This broader model subsumes the strict factor model, where the diagonal component is a special case of a sparse structure.

To obtain a covariance estimator under the approximate factor model, one direct approach is the principal orthogonal complement thresholding (POET) estimator proposed by [41], which combines PCA for recovering the low-rank component with a thresholding operation for recovering the sparse component. In addition to POET, researchers have also explored convex optimization techniques based on penalized least squares. These methods leverage the nuclear norm for recovering the low-rank component and the ℓ_1 norm for recovering the sparse component. The error analysis in [45] reveals an estimation error bound in terms of the squared Frobenius norm as $O_p\left(\frac{rd}{n}+\frac{s\log d}{n}+\frac{s}{d^2}\right)$, where n is the number of samples, r corresponds to the rank of $\boldsymbol{L}_\star$, and s represents the number of nonzero elements of $\boldsymbol{\Psi}_\star$. However, this error bound encompasses an additional error component $\frac{s}{d^2}$, indicating that exact recovery is unattainable even in noiseless conditions due to an additional error term. In [39], [42], [46], a tighter rate was obtained as $O_p\left(\frac{rd}{n}+\frac{s\log d}{n}\right)$, which matches the minimax optimal rate and demonstrates exact recovery. Despite this, the computational complexity of the convex relaxation method is notable, with a spectral decomposition being required to compute the proximal mapping of the nuclear norm at each iteration, resulting in a high per-iteration time complexity of $O\left(d^3\right)$. This complexity becomes particularly challenging when dealing with high-dimensional problems where d is large. Additionally, there is another line of work on factor-based covariance matrix estimation that focuses on likelihood-based loss functions [47], [48], [49], [50], [51], [52], although this topic is beyond the scope of this paper.

In this paper, we focus on estimating large covariance matrices under the assumption of the approximate factor model. We formulate the estimation problem as a nonconvex optimization problem and propose an iterative algorithm to compute the estimator. The algorithm begins with the POET estimator and then updates the low-rank and sparse components in an iterative manner. After certain iterations, we obtain an improved estimator of the covariance matrix. We name the finally obtained nonconvex estimator POET with refining iteratively (POETRY). Our theoretical analysis demonstrates that the sequence generated by the algorithm converges linearly to the true covariance matrix $\boldsymbol{\Sigma}_\star$ up to a statistical rate of $O_p\left(\frac{rd}{n}+\frac{s\log d}{n}\right)$, which matches the minimax rate [42], [45], [46]. Moreover, we establish the statistical rate of the inverse covariance matrix estimate based on the covariance estimator. By introducing the notion of effective rank r_{e} [53], we reduce the required sample size from $n \geq d$ (as in convex relaxation methods [42], [45], [46]) to $n \gtrsim r_{\mathrm{e}} r \log 2d$ and enhance the statistical rate to $O_p\left(\frac{r_{\mathrm{e}} r\log 2d + s\log d}{n}\right)$. Additionally, our algorithm has a lower per-iteration time complexity of $O\left(rd^2\right)$ compared to the $O\left(d^3\right)$ per-iteration complexity of convex relaxation-based methods [42], [45], [46].

The remainder of this paper is organized as follows. In Section II, we provide our problem formulation for large covariance matrix estimation based on factor models and discuss relevant existing estimators. The detailed algorithm is elaborated in Section III. Section IV contains our theoretical results, including the statistical convergence rate of the proposed estimator and the analysis of iteration complexity. In Section V, we conduct numerical experiments to compare our proposed algorithm with state-of-the-art methods on both synthetic and real-world data. The numerical experiments demonstrate the effectiveness of our algorithm over existing approaches and validate the obtained theoretical claims regarding the statistical rate of the POETRY estimator. Finally, we conclude our paper in Section VI. The proofs of all theoretical results are provided in the Appendices.

Notation. Standard lower-case and upper-case letters, such as x and X, represent scalars. Boldface lower-case and boldface upper-case letters, such as $\boldsymbol{x}$ and $\boldsymbol{X}$, denote vectors and matrices, respectively. $\boldsymbol{X}_{ij}$ or $(\boldsymbol{X})_{ij}$ refers to the (i,j)-th entry of $\boldsymbol{X}$. $|\boldsymbol{X}|_{ij}$ denotes the absolute value of $\boldsymbol{X}_{ij}$. $\mathbf{0}$ stands for the all-zero vector or all-zero matrix. $\mathbf{1}$ stands for the all-one vector. $\boldsymbol{I}$ stands for the identity matrix. $\mathbb{R}^m$ denotes the set of real $m\times 1$ vectors. $\mathbb{R}^{m\times n}$ denotes the set of real $m\times n$ matrices. Calligraphic letters, like $\mathcal{S}$, are used to represent sets.

$\operatorname{diag}(\boldsymbol{X})$ represents a diagonal matrix containing the diagonal elements of $\boldsymbol{X}$. $\boldsymbol{X}^\top$ and $\boldsymbol{X}^{-1}$ denote the transpose and inverse of $\boldsymbol{X}$, respectively. We use $\sigma_i(\boldsymbol{X})$ to denote the i-th largest singular value of $\boldsymbol{X}$. In addition, $\|\boldsymbol{X}\|_{\mathrm{F}}$, $\|\boldsymbol{X}\|_*$, and $\|\boldsymbol{X}\|_2$ denote the Frobenius norm, the nuclear norm, and the spectral norm of matrix $\boldsymbol{X}$, respectively. $\|\boldsymbol{X}\|_1$ and $\|\boldsymbol{X}\|_\infty$ denote the entrywise ℓ_1 norm and the entrywise ℓ_∞ norm of matrix $\boldsymbol{X}$, respectively. $\|\boldsymbol{X}\|_0$ denotes the number of nonzero elements in the matrix $\boldsymbol{X}$.

$\nabla f(\cdot)$ represents the gradient of a function f. $a\vee b = \max\{a,b\}$. When comparing functionals $f(n)$ and $g(n)$ for large values of n, we denote $f(n)\gtrsim g(n)$ if $f(n)\geq cg(n)$, $f(n)\lesssim g(n)$ if $f(n)\leq Cg(n)$, and $f(n)\asymp g(n)$ if $cg(n)\leq f(n)\leq Cg(n)$ for some positive constants c and C. For sequences of numbers a_m and b_m, we denote $b_m = O(a_m)$ if $\limsup_{m\to\infty}\frac{|b_m|}{a_m}<\infty$, and $b_m = O_p(a_m)$ if, for any $\varepsilon>0$, there exists C_ε such that $P(|b_m|\leq C_\varepsilon a_m)>1-\varepsilon$ for all m. We use C and c, with appropriate subscripts, to denote universal positive constants whose values may differ from line to line.

II. Covariance Matrix Estimation based on Factor Models

The factor model describes a random vector $\boldsymbol{x} \in \mathbb{R}^d$ as

$$\boldsymbol{x} = \boldsymbol{B}_\star \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{B}_\star \in \mathbb{R}^{d\times r}$ with $r \ll d$ is a loading matrix, $\boldsymbol{f} \in \mathbb{R}^r$ represents the low-dimensional unobservable common factors, and $\boldsymbol{\epsilon} \in \mathbb{R}^d$ are idiosyncratic errors. The factors $\boldsymbol{f}$ and the errors $\boldsymbol{\epsilon}$ are assumed to be uncorrelated. Additionally, it is commonly assumed that both $\boldsymbol{f}$ and $\boldsymbol{\epsilon}$ follow normal distributions with zero mean. The factor $\boldsymbol{f}$ is assumed to have an identity covariance matrix, while the error $\boldsymbol{\epsilon}$ has a covariance matrix $\boldsymbol{\Psi}_\star$. The covariance matrix of $\boldsymbol{x}$, denoted by $\boldsymbol{\Sigma}_\star$, is then computed as

$$\boldsymbol{\Sigma}_\star = \boldsymbol{B}_\star \boldsymbol{B}_\star^\top + \boldsymbol{\Psi}_\star = \boldsymbol{L}_\star + \boldsymbol{\Psi}_\star, \tag{2}$$

where $\boldsymbol{L}_\star = \boldsymbol{B}_\star \boldsymbol{B}_\star^\top$ is low-rank, and $\boldsymbol{\Psi}_\star$ is sparse due to the weak correlation among errors in the approximate factor structure [32]. In [45], the decomposition in (2) is also referred to as factor analysis with sparse noise.

In practice, the true covariance matrix $\boldsymbol{\Sigma}_\star$ is unknown. Instead, we can compute the SCM as $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$, where $\boldsymbol{x}_i$ for $i = 1, 2, \ldots, n$ represents n independent and identically distributed (i.i.d.) samples from the model (1). Due to the limited number of available samples, $\boldsymbol{S}$ is considered a "noisy" representation of $\boldsymbol{\Sigma}_\star$ [45], [54]. This relationship is expressed as

$$\boldsymbol{S} = \boldsymbol{\Sigma}_\star + \boldsymbol{N}, \tag{3}$$

where $\boldsymbol{N}$ denotes a re-centered Wishart noise [55] arising from the normal distribution of each $\boldsymbol{x}_i$. Given $\boldsymbol{S}$, our goal is to estimate $\boldsymbol{\Sigma}_\star$ within the context of factor modeling.

A. Existing Estimators

1) Direct Method: We present the principal orthogonal complement thresholding (POET) estimator [41], a well-established direct method in the field of factor-based covariance matrix estimation. POET involves extracting the top r principal components of $\boldsymbol{S}$ and subsequently applying thresholding to the remaining components. The thresholding methods that can be employed include soft thresholding, hard thresholding, smoothly clipped absolute deviations, and adaptive thresholding [41]. Specifically, let $\phi_1 \geq \phi_2 \geq \ldots \geq \phi_d$ denote the eigenvalues of $\boldsymbol{S}$ in descending order, and let $\{\boldsymbol{\xi}_i\}_{i=1}^d$ denote the corresponding eigenvectors. The spectral decomposition of $\boldsymbol{S}$ can be written as

$$\boldsymbol{S} = \sum_{i=1}^r \phi_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top + \boldsymbol{R},$$

where the residual component $\boldsymbol{R} = \sum_{i=r+1}^d \phi_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top$ is referred to as the "principal orthogonal complement." We define the hard-thresholding operator, which retains only the τ largest entries in magnitude of the input matrix, as follows:

$$\left(\boldsymbol{H}_\tau(\boldsymbol{M})\right)_{ij} = \begin{cases} \boldsymbol{M}_{ij} & \text{if } |\boldsymbol{M}|_{ij} \text{ is among the largest } \tau \\ & \text{entries in magnitude of } \boldsymbol{M} \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 1: Algorithm for POET

Input: $\boldsymbol{S}$, r, τ;

let $\boldsymbol{\Xi}\boldsymbol{\Phi}\boldsymbol{\Xi}^\top$ be the top-r spectral decomposition of $\boldsymbol{S}$;

$\boldsymbol{B}_{\mathrm{POET}} = \boldsymbol{\Xi}\boldsymbol{\Phi}^{1/2}$;

$\boldsymbol{\Psi}_{\mathrm{POET}} = \boldsymbol{H}_\tau\left(\boldsymbol{S} - \boldsymbol{B}_{\mathrm{POET}}\boldsymbol{B}_{\mathrm{POET}}^\top\right)$;

Output: $\boldsymbol{\Sigma}_{\mathrm{POET}} = \boldsymbol{B}_{\mathrm{POET}}\boldsymbol{B}_{\mathrm{POET}}^\top + \boldsymbol{\Psi}_{\mathrm{POET}}$.

The POET estimator is then given by

$$\boldsymbol{\Sigma}_{\mathrm{POET}} = \sum_{i=1}^r \phi_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top + \boldsymbol{H}_\tau(\boldsymbol{R}). \tag{4}$$

This method is optimization-free, making it computationally efficient. A detailed algorithm for computing POET is provided in Algorithm 1.

2) Convex Optimization Method: Convex relaxation-based methods have been proposed for factor-based covariance matrix estimation [39], [42], [45]. This approach formulates the optimization problem as a least squares problem that incorporates the nuclear norm and the ℓ_1 norm to simultaneously promote a low-rank component and a sparse component, respectively. The corresponding problem is given by

$$\underset{\boldsymbol{L},\,\boldsymbol{\Psi}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{L} + \boldsymbol{\Psi} - \boldsymbol{S}\|_{\mathrm{F}}^2 + \kappa_1 \|\boldsymbol{L}\|_* + \kappa_2 \|\boldsymbol{\Psi}\|_1, \tag{5}$$

where $\kappa_1, \kappa_2 > 0$ are tuning parameters, and the resulting estimator is referred to as the low rank and sparse covariance (LOREC) estimator [45], [42]. The problem is convex and can be solved to global optimality. Many standard algorithms, such as the accelerated proximal gradient method [42] and alternating minimization, can be used to solve it. However, it is important to note that one limitation of the LOREC estimator is that the low-rank component $\boldsymbol{L}$ does not necessarily guarantee positive semidefiniteness.

B. Comparison with Robust PCA

In this section, we compare the factor-based covariance matrix estimation problem with the well-established robust PCA problem [56], [57], [58], [59]. We begin by briefly reviewing the robust PCA formulation. Given an observation matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, the robust PCA model assumes the following decomposition:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{Y} + \boldsymbol{E} + \boldsymbol{O},$$

where $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ is a basis matrix with $r \ll d$, $\boldsymbol{Y} \in \mathbb{R}^{r\times n}$ contains the low-dimensional representations of $\boldsymbol{X}$, $\boldsymbol{E} \in \mathbb{R}^{d\times n}$ is a noise matrix, and $\boldsymbol{O} \in \mathbb{R}^{d\times n}$ is a sparse outlier matrix. The goal of robust PCA is to obtain a low-rank representation $\boldsymbol{U}\boldsymbol{Y}$ of the possibly noisy measurement $\boldsymbol{X}$ by effectively separating and removing the sparse outliers $\boldsymbol{O}$. Although both robust PCA and factor-based covariance estimation involve a low-rank plus sparse decomposition, existing methods and theoretical results for robust PCA cannot be directly applied to our setting due to several key differences, which we summarize below:

(i) Robust PCA typically focuses on decomposing an observed data matrix directly into low-rank and sparse

components, often without a probabilistic interpretation. In contrast, the factor-based covariance estimation problem is rooted in the statistical factor model, where the decomposition is applied to the covariance matrix, motivated by a generative latent variable model.

(ii) The low-rank component in robust PCA is generally asymmetric, while the low-rank component in covariance estimation is symmetric and positive semidefinite due to the factor modeling.

(iii) The sparse component in robust PCA reflects a small number of corrupted entries (i.e., sparse outliers) at the data level. Two types of sparsity are typically considered: element-wise sparsity and column-wise sparsity, corresponding to different corruption models. In contrast, factor-based covariance estimation assumes sparsity in the residual noise covariance matrix, where the sparse component captures weak correlations among residual errors. This sparse component is required to be symmetric and may exhibit additional structure, such as a block-diagonal pattern.

(iv) In robust PCA, the noise term represents the reconstruction or representation error. It can also be absent, leading to the so-called noiseless robust PCA formulation. By contrast, in factor-based covariance estimation, the noise arises from sample variability under the finite-sample regime and is always present.

(v) In covariance estimation, it is often desirable for the combined low-rank and sparse estimator to be positive semidefinite. Robust PCA, on the other hand, does not impose such a requirement.

These distinctions lead to fundamentally different algorithmic designs, assumptions, and statistical guarantees between the two problems.

III. The Proposed Estimator and Algorithm

A. Proposed Estimator

In this paper, we introduce a nonconvex optimization method for estimating covariance matrices under the factor model assumption (2). In contrast to the regularization technique used for estimating the low-rank part of the covariance matrix, $\boldsymbol{L}$ [45], [42], we choose a nonconvex matrix factorization technique to address the computational challenges associated with rank regularization. We represent $\boldsymbol{L}$ as the product of smaller matrices, denoted as $\boldsymbol{L} = \boldsymbol{B}\boldsymbol{B}^\top$, where $\boldsymbol{B} \in \mathbb{R}^{d\times r}$. Furthermore, we apply a cardinality constraint to the sparse component of the covariance matrix, $\boldsymbol{\Psi}$.

The basic idea behind our approach is to find the "best fit" between the estimated covariance $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Psi}$ and the sample covariance $\boldsymbol{S}$. To achieve this, we consider the following nonconvex optimization problem:

$$\begin{aligned} \underset{\boldsymbol{B},\,\boldsymbol{\Psi}}{\text{minimize}} \quad & \frac{1}{2}\left\|\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Psi} - \boldsymbol{S}\right\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \|\boldsymbol{\Psi}\|_0 \le \tau, \end{aligned} \tag{6}$$

where $\tau > 0$ is a tuning parameter that controls the number of nonzero elements in $\boldsymbol{\Psi}$. When $r = 0$, the estimator simplifies to the sparse covariance matrix estimator [12]. When $\tau = 0$, the estimator reduces to the low-rank covariance matrix estimator [60]. When either r or τ is sufficiently large, the estimator becomes the SCM. Additionally, the POET estimator presented in (4) can be viewed as a heuristic solution to (6).

Algorithm 2: Algorithm for POETRY

Input: $\boldsymbol{S}$, r, τ, η, T;

let $\boldsymbol{\Xi}\boldsymbol{\Phi}\boldsymbol{\Xi}^\top$ be the top-r spectral decomposition of $\boldsymbol{S}$;

$\boldsymbol{B}_0 = \boldsymbol{\Xi}\boldsymbol{\Phi}^{1/2}$;

$\boldsymbol{\Psi}_0 = \boldsymbol{H}_\tau\left(\boldsymbol{S} - \boldsymbol{B}_0\boldsymbol{B}_0^\top\right)$;

for $t = 0, 1, \ldots, T-1$ **do**

  $\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta\nabla_{\boldsymbol{B}} L\left(\boldsymbol{B}_t, \boldsymbol{\Psi}_t\right)\left(\boldsymbol{B}_t^\top\boldsymbol{B}_t\right)^{-1}$;

  $\boldsymbol{\Psi}_{t+1} = \boldsymbol{H}_\tau\left(\boldsymbol{S} - \boldsymbol{B}_{t+1}\boldsymbol{B}_{t+1}^\top\right)$;

end

Output: $\boldsymbol{\Sigma}_T = \boldsymbol{B}_T\boldsymbol{B}_T^\top + \boldsymbol{\Psi}_T$.

B. Proposed Algorithm

We propose to solve problem (6) through an iterative optimization procedure, which is outlined in Algorithm 2. In the initialization phase, we utilize POET as the initial estimate, where we set $\boldsymbol{B}_0 = \boldsymbol{B}_{\mathrm{POET}}$, $\boldsymbol{\Psi}_0 = \boldsymbol{\Psi}_{\mathrm{POET}}$, and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\mathrm{POET}}$ based on Algorithm 1. Subsequently, we iteratively optimize the variables $\boldsymbol{B}$ and $\boldsymbol{\Psi}$ over sufficient iterations, resulting in an estimator which we name POET with refining iteratively (POETRY). This acronym highlights the iterative refinement nature of the algorithm. During the iterative refining stage, we update $\boldsymbol{B}$ using one step of preconditioned gradient descent with a step size of η [61]. The preconditioner $\left(\boldsymbol{B}_t^\top\boldsymbol{B}_t\right)^{-1}$ is right-multiplied to the gradient

$$\nabla_{\boldsymbol{B}} L\left(\boldsymbol{B}, \boldsymbol{\Psi}\right) = \left(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Psi} - \boldsymbol{S}\right)\boldsymbol{B}, \tag{7}$$

where L denotes the objective function in (6). The preconditioner is employed to alleviate the effect of the potentially large condition number of the low-rank component of the covariance matrix. This adaptive preconditioner varies with each iteration and has a lower computational complexity than computing the gradient. Therefore, it does not introduce additional high-order complexity compared to standard gradient descent. To update $\boldsymbol{\Psi}$, we employ one step of hard thresholding.

Algorithm 2 is notably efficient, as it eliminates the need for spectral decomposition in each iteration, making it substantially faster than algorithms depending on convex relaxations [45], [42], [46]. The per-iteration computational overhead of Algorithm 2 mainly arises from computing the gradient, which has a time complexity of $O\left(rd^2\right)$. This time complexity provides a significant advantage over the per-iteration complexity of $O\left(d^3\right)$ associated with the convex methods [45], [42], [46].

IV. Main Theory

In this section, we present our assumptions, definitions, and main theoretical results. We begin with the necessary assumptions and definitions essential for our subsequent analysis.

A. Preliminaries

Assumption 1. $0 < \sigma_d(\boldsymbol{\Sigma}_\star) \leq \sigma_1(\boldsymbol{\Sigma}_\star) \leq \mu < \infty$.

This assumption requires that the eigenvalues of the true covariance matrix $\boldsymbol{\Sigma}_\star$ are finite and bounded below by a positive number. It is commonly adopted in the context of factor-based covariance matrix estimation [41].

The covariance matrix estimation problem with the decomposition $\boldsymbol{\Sigma}_\star = \boldsymbol{L}_\star + \boldsymbol{\Psi}_\star$ can be ill-posed, with an identifiability issue arising when the low-rank matrix $\boldsymbol{L}_\star$ is also sparse [62]. Therefore, certain conditions are required to ensure that the decomposition is well-posed. For example, the well-known incoherence condition can be adopted to prevent the low-rank matrix from being excessively sparse by restricting the degree of coherence between singular vectors and the standard basis [63]. However, this condition is only applicable in noiseless models [45], making it unsuitable for our problem, which is inherently "noisy by nature." In this paper, we introduce the spikiness condition as an alternative, which is milder than the incoherence condition and can be applied to noisy and noiseless models [45]. The spikiness condition involves constraining the spikiness ratio, defined as $\alpha \coloneqq d\|\boldsymbol{L}_\star\|_\infty / \|\boldsymbol{L}_\star\|_{\mathrm{F}}$ for the matrix $\boldsymbol{L}_\star \in \mathbb{R}^{d\times d}$ [64].

Assumption 2 (Spikiness condition [45])**.** *For a low-rank matrix $\boldsymbol{L}_\star$, we assume that there exists a constant k such that*

$$\|\boldsymbol{L}_\star\|_\infty = \frac{\alpha\|\boldsymbol{L}_\star\|_{\mathrm{F}}}{d} \leq \frac{k}{d}.$$

Next, we introduce the concept of the effective rank of a matrix, which is instrumental in deriving the error bounds. The covariance matrices based on factor models have effectively reduced rank, thus we make use of the notation of the effective rank, which is first suggested by [53].

Definition 3 ([53])**.** The effective rank of the covariance matrix $\boldsymbol{\Sigma}_\star$ is defined as

$$r_{\mathrm{e}} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\star)}{\|\boldsymbol{\Sigma}_\star\|_2}.$$

The effective rank can be considered as a measure of the concentration level of the spectrum of $\boldsymbol{\Sigma}_\star$. As we will show in Section IV-B, when the effective rank of the covariance matrix is much smaller than d, our theoretical results provide a tight squared Frobenius norm estimation error bound.

B. Theoretical Analysis

The following theorems characterize the theoretical properties of the sequence $\{\boldsymbol{\Sigma}_t\}_{t\geq 0}$, where the composite estimator $\boldsymbol{\Sigma}_t = \boldsymbol{B}_t\boldsymbol{B}_t^\top + \boldsymbol{\Psi}_t$, generated by Algorithm 2. We first derive the statistical error of the POET estimator, i.e., $\boldsymbol{\Sigma}_0$, in Theorem 4, and then we derive the error bound for $\boldsymbol{\Sigma}_t$ when $t = 1, 2, \ldots$, which encompasses both the statistical error and the optimization error, in Theorem 5. By taking into account the effective rank, we obtain an improved statistical error characterization in Theorem 7.

Theorem 4. *Suppose Assumptions 1 and 2 hold. Assume that the sparsity parameter satisfies $\tau \asymp s$. Then, with probability at least $\min\{1-2\exp(-C_1 d), 1-2\exp(-C_2\log d)\}$, the initialization of Algorithm 2, which is also the POET estimator, satisfies*

$$\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2 \leq c_{1,4}\frac{rd}{n} + c_{2,4}\frac{s\log d}{n} + c_3\frac{rsd}{n} + c_4 rs + c_5 r.$$

Theorem 4 demonstrates that the POET estimator achieves a statistical rate of $O_p\left(\frac{rd}{n} + \frac{s\log d}{n} + \frac{rsd}{n}\right)$, in which we neglect the last two terms in the error bound of Theorem 4 since we focus on the high-dimensional case, i.e., $n \ll d$. This rate includes an additional error term, $\frac{rsd}{n}$, compared to the minimax optimal rate $O_p\left(\frac{rd}{n} + \frac{s\log d}{n}\right)$ [42], [46]. This discrepancy suggests that POET can hardly achieve a good rate of convergence under the squared Frobenius norm, as also noted by [41].

Theorem 5. *Suppose Assumptions 1 and 2 hold. Assume that the sample size satisfies $n > 48rd/\sigma_r^2(\boldsymbol{L}_\star)$ and the sparsity parameter satisfies $\tau \asymp s$. Suppose the step size obeys $\frac{2-\sqrt{2}}{20} < \eta < \frac{1}{2}$. Then, with probability at least $\min\{1-2\exp(-C_1 d), 1-2\exp(-C_2\log d)\}$, the output of Algorithm 2 for $t = 1, 2, \ldots$ satisfies*

$$\|\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2 \leq \underbrace{c_{1,5}\frac{rd}{n} + c_{2,5}\frac{s\log d}{n}}_{\text{statistical error}} + \underbrace{c_6\rho^T\sigma_r^2(\boldsymbol{L}_\star)}_{\text{optimization error}},$$

where $\rho = 2(1-10\eta)^2$.

Theorem 5 suggests that the estimation error of the output from Algorithm 2 comprises two components: the statistical error and the optimization error. The statistical error scales as $O_p\left(\frac{rd}{n} + \frac{s\log d}{n}\right)$, which matches the minimax rate [42], [46]. The term $\frac{s\log d}{n}$ corresponds to the statistical error of the sparse component $\boldsymbol{\Psi}_\star$, while the term $\frac{rd}{n}$ represents the statistical error of the low-rank component $\boldsymbol{L}_\star$. Regarding the optimization error, the contraction parameter ρ depends only on the step size η. It is possible to choose an appropriate step size such that ρ remains strictly between 0 and 1, which ensures a linear convergence rate for Algorithm 2. The assumption $n > 48rd/\sigma_r^2(\boldsymbol{L}_\star)$ is required to ensure that the initial estimator is sufficiently close to the true covariance matrix, and similar conditions have been adopted in previous works, e.g., [45], [42], [65], [66]. We conjecture that this requirement primarily arises from employing the sample covariance matrix as the input to our estimation procedure. If a more accurate initial estimator were used (as suggested in [42]), the sample size requirement could potentially be further relaxed.

It can be shown that by choosing $T \geq O\left(\log\left(\frac{n}{rd+s\log d}\right)\right)$, the total estimation error achieves the same order as the statistical error. From Theorem 5, we also have that $\boldsymbol{\Sigma}_T$ satisfies the following spectral norm bound:

$$\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_2^2 \leq c_1\frac{rd}{n} + c_2\frac{s\log d}{n}. \tag{8}$$

Similar to POET in [45] and LOREC in [42], the POETRY estimator $\boldsymbol{\Sigma}_T$ from Algorithm 2 is not necessarily positive definite. However, the above spectral norm bound implies that $\boldsymbol{\Sigma}_T$ will be positive definite if $\sigma_d(\boldsymbol{\Sigma}_\star)$ is larger than $\sqrt{c_1\frac{rd}{n} + c_2\frac{s\log d}{n}}$. Furthermore, this result asymptotically ensures the positive definiteness of $\boldsymbol{\Sigma}_T$ under the regularity

condition that $\sigma_d(\boldsymbol{\Sigma}_\star)$ is bounded away from zero. Consequently, we conclude that $\boldsymbol{\Sigma}_T$ will be positive definite with an overwhelming probability tending to one.

C. Statistical Error Bound for the Inverse Covariance Matrix

In this section, we provide the statistical estimation result for the inverse covariance matrix, also known as the precision matrix. This error bound is particularly useful in scenarios where the inverse covariance serves as an input to the statistical methods, including linear and quadratic discriminant analysis, Mahalanobis distance computation, weighted least squares, and Markowitz portfolio selection.

Corollary 6. *Under the conditions in Theorem 5 and assuming that* $\sigma_d(\boldsymbol{\Sigma}_\star) \geq c_0\sqrt{c_1\frac{rd}{n}+c_2\frac{s\log d}{n}}$ *with* $c_0>1$*, then, with probability at least* $\min\{1-2\exp(-C_2\log d), 1-1/(2d)\}$*, the output of Algorithm 2 with* $T \geq O\left(\log\left(\frac{n}{rd+s\log d}\right)\right)$ *satisfies*

$$\left\|\boldsymbol{\Sigma}_T^{-1}-\boldsymbol{\Sigma}_\star^{-1}\right\|_{\mathrm{F}}^2 \leq c_{1,6}\frac{rd}{n}+c_{2,6}\frac{s\log d}{n}.$$

This result shows that the statistical rate for covariance estimation also holds for the inverse covariance estimation. The statistical error matches the minimax rate for the low-rank plus sparse inverse covariance matrix estimation problem [45], [65], [67].

D. Improved Statistical Error Bound

In real-world scenarios, the effective rank r_{e} of the true covariance matrix $\boldsymbol{\Sigma}_\star$ is often much smaller than d. Then, leveraging recent progress on the asymptotic properties of the sample covariance matrix [68], we can derive a much tighter error bound which only depends on d logarithmically, as stated in the subsequent theorem.

Theorem 7. *Suppose Assumptions 1 and 2 hold. Assume that the sample size satisfies* $n > 12r_{\mathrm{e}}r\log 2d/\sigma_r^2(\boldsymbol{L}_\star)$ *and the sparsity parameter satisfies* $\tau \asymp s$*. Suppose the step size obeys* $\frac{2-\sqrt{2}}{20} < \eta < \frac{1}{2}$*. Then with probability at least* $\min\{1-2\exp(-C_2\log d), 1-1/(2d)\}$*, the output of Algorithm 2 for* $t=1,2,\ldots$ *satisfies*

$$\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2 \leq \underbrace{c_{1,7}\frac{r_{\mathrm{e}}r\log 2d}{n}+c_{2,7}\frac{s\log d}{n}}_{\textit{statistical error}}+\underbrace{c_6\rho^T\sigma_r^2(\boldsymbol{L}_\star)}_{\textit{optimization error}},$$

where $\rho=2(1-10\eta)^2$.

The statistical error in Theorem 7 scales as $O_p\left(\frac{r_{\mathrm{e}}r\log 2d+s\log d}{n}\right)$. When $r_{\mathrm{e}} \ll d$, it is significantly tighter compared to the error $O_p\left(\frac{rd}{n}+\frac{s\log d}{n}\right)$ in Theorem 5. Moreover, the condition for the sample size, $n \gtrsim r_{\mathrm{e}}r\log 2d$, is much less stringent. This result implies the efficiency of factor-based covariance matrix estimation when the true covariance model has a low effective rank.

V. Numerical Experiments

In this section, we present numerical experiments on both synthetic and real datasets to validate the properties of our proposed estimator, POETRY. We compare its performance with several methods, including the sample covariance matrix (SCM) estimator and three estimators for approximate factor models: POET (principal orthogonal complement thresholding) [41], LOREC (low rank and sparse covariance estimator) [42], which is based on convex relaxation, and the alternating projection method, AltProj [59], which employs spectral decomposition for low-rank matrix projection and hard thresholding for sparse matrix projection. Additionally, we compare POETRY with two estimators for strict factor models: POET-D (principal orthogonal complement thresholding diagonally) [30], [41], which thresholds the diagonal of the residual components of the SCM after extracting the first r principal components, and AltProj-D [31], [69], which uses spectral decomposition to obtain the low-rank matrix component followed by diagonal thresholding. The simulations are performed in MATLAB with a 2.9 GHz Intel Core i7 CPU.

A. Synthetic Data

We first generate a loading matrix $\boldsymbol{B}_\star=\boldsymbol{V}_\star\boldsymbol{\Lambda}_\star^{1/2}$, where $\boldsymbol{V}_\star \in \mathbb{R}^{d\times r}$ and $\boldsymbol{\Lambda}_\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix. To obtain $\boldsymbol{V}_\star$, we generate a $d\times r$ matrix with i.i.d. random signs and then take its r left singular vectors. The diagonal of $\boldsymbol{\Lambda}_\star$ are set to values that are linearly distributed from $\frac{1}{10}$ to 1. Next, we sample $\boldsymbol{f}_1,\ldots,\boldsymbol{f}_n$ from a multivariate normal distribution with zero mean and identity covariance matrix. We sample $\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_n$ from a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Psi}_\star$, which is a $d\times d$ positive definite matrix with a sparsity level $s=0.02\times d^2$. Finally, we generate the samples $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ based on model (1).

1) Estimation Performance: First, we compare the estimation performance of different methods based on the estimation errors in Frobenius norm for the low-rank matrix (i.e., $\|\boldsymbol{L}_T-\boldsymbol{L}_\star\|_{\mathrm{F}}$), the sparse matrix (i.e., $\|\boldsymbol{\Psi}_T-\boldsymbol{\Psi}_\star\|_{\mathrm{F}}$), and the covariance matrix (i.e., $\|\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}$). Furthermore, we assess the selection performance of the sparse matrix by analyzing both the false positive rate (FPR) and the true positive rate (TPR). FPR refers to the proportion of non-significant elements that are incorrectly identified as significant in the estimated sparse matrix. In contrast, TPR measures the proportion of actual significant elements that are correctly identified in the estimated sparse matrix. These metrics together provide a comprehensive evaluation of both estimation precision and the effectiveness of identifying significant elements within the sparse matrices. The definitions of FPR and TPR are given as follows:

$$\mathrm{FPR}=\frac{\#\left\{(i,j):\boldsymbol{\Psi}_{ij}\neq 0\,\&\,(\boldsymbol{\Psi}_\star)_{ij}=0\right\}}{\#\left\{(i,j):(\boldsymbol{\Psi}_\star)_{ij}=0\right\}},$$

$$\mathrm{TPR}=\frac{\#\left\{(i,j):\boldsymbol{\Psi}_{ij}\neq 0\,\&\,(\boldsymbol{\Psi}_\star)_{ij}\neq 0\right\}}{\#\left\{(i,j):(\boldsymbol{\Psi}_\star)_{ij}\neq 0\right\}}.$$

TABLE I
QUANTITATIVE COMPARISON AMONG SEVEN DIFFERENT METHODS

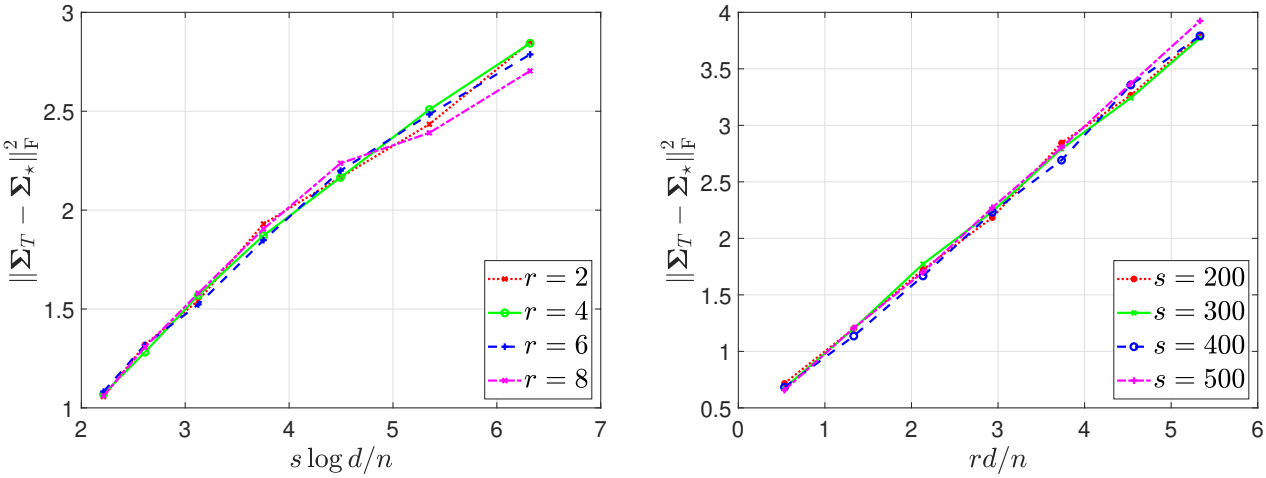
		Strict factor models		Approximate factor models			
	SCM	POET-D	AltProj-D	POET	LOREC	AltProj	POETRY (prop.)
$n = 50,\ d = 100,\ r = 2,\ s = 200$							
$\|\boldsymbol{L}_T - \boldsymbol{L}_\star\|_{\mathrm{F}}$	–	1.6242±0.1307	1.3632±0.1096	1.6242±0.1307	1.4283±0.0869	1.3653±0.1314	**1.3584±0.1281**
$\|\boldsymbol{\Psi}_T - \boldsymbol{\Psi}_\star\|_{\mathrm{F}}$	–	2.4714±0.0409	2.3842±0.0455	2.3845±0.0262	2.6151±0.0298	2.2983±0.0267	**2.2921±0.0289**
$\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}$	2.7091±0.0423	2.6586±0.0502	2.5374±0.0502	2.5418±0.0479	2.5447±0.0354	2.4397±0.0453	**2.4378±0.0462**
FPR	–	–	–	0.0220±0.0011	0.0369±0.0150	0.0343±0.0010	**0.0200±0.0011**
TPR	–	0.8981±0.3156	0.9434±0.0001	0.9472±0.0080	0.9377±0.0150	0.7953±0.0278	**0.9491±0.0091**
Time	–	–	0.2281	–	1.1719	0.3266	**0.1516**
$n = 80,\ d = 200,\ r = 5,\ s = 800$							
$\|\boldsymbol{L}_T - \boldsymbol{L}_\star\|_{\mathrm{F}}$	–	2.2082±0.0553	1.9106±0.0487	2.2082±0.0553	2.1194±0.0362	1.9247±0.0841	**1.9072±0.0814**
$\|\boldsymbol{\Psi}_T - \boldsymbol{\Psi}_\star\|_{\mathrm{F}}$	–	3.8206±0.0400	3.7043±0.0552	3.7159±0.0319	4.1742±0.0347	3.6637±0.0393	**3.6526±0.0426**
$\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}$	4.0526±0.0264	4.0300±0.0381	3.9011±0.0424	3.9144±0.0296	3.9279±0.0301	3.8254±0.0362	**3.8249±0.0361**
FPR	–	–	–	0.0110±0.0004	0.0141±0.0027	0.0175±0.0008	**0.0100±0.0004**
TPR	–	0.8971±0.5127	0.9524±0.0002	0.9571±0.0081	0.9543±0.0149	0.8505±0.0265	**0.9581±0.0080**
Time	–	–	0.6578	–	3.2141	1.4719	**0.5828**
$n = 100,\ d = 500,\ r = 10,\ s = 5000$							
$\|\boldsymbol{L}_T - \boldsymbol{L}_\star\|_{\mathrm{F}}$	–	2.6486±0.0284	2.5134±0.1010	2.6486±0.0284	2.6562±0.0246	2.4847±0.0346	**2.4800±0.0333**
$\|\boldsymbol{\Psi}_T - \boldsymbol{\Psi}_\star\|_{\mathrm{F}}$	–	7.0320±0.0205	6.9668±0.0311	6.9375±0.0107	7.5082±0.0141	6.8706±0.0134	**6.8668±0.0131**
$\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}$	7.2616±0.0113	7.2427±0.0235	7.1926±0.0246	7.1378±0.0147	7.2592±0.0159	7.0933±0.0145	**7.0924±0.0149**
FPR	–	–	–	0.0041±0.0006	0.0045±0.0002	0.0044±0.0006	**0.0040±0.0003**
TPR	–	0.9315±0.0057	0.9545±0.0003	0.9630±0.0057	0.9363±0.0058	0.8795±0.0085	**0.9637±0.0060**
Time	–	–	9.6219	–	16.5750	8.1828	**4.5719**
$n = 400,\ d = 1000,\ r = 20,\ s = 20000$							
$\|\boldsymbol{L}_T - \boldsymbol{L}_\star\|_{\mathrm{F}}$	–	4.5096±0.0245	3.8785±0.0396	4.5096±0.0245	5.6019±0.0293	3.7858±0.0374	**3.7697±0.0360**
$\|\boldsymbol{\Psi}_T - \boldsymbol{\Psi}_\star\|_{\mathrm{F}}$	–	8.3495±0.0300	8.1280±0.0346	7.9495±0.0220	9.8759±0.0230	7.6904±0.0254	**7.6815±0.0242**
$\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}$	8.8933±0.0138	8.7787±0.0284	8.5077±0.0325	8.3367±0.0208	8.6932±0.0164	8.0873±0.0243	**8.0869±0.0243**
FPR	–	–	–	0.0022±0.0003	0.0033±0.0002	0.0036±0.0001	**0.0021±0.0002**
TPR	–	0.9015±0.0057	0.9557±0.0008	0.9714±0.0060	0.9708±0.0036	0.9315±0.0069	**0.9715±0.0059**
Time	–	–	38.8047	–	207.6719	35.4063	**27.0766**

We conduct a series of experiments by varying parameters n, d, r, and s. For each parameter setting, we generate ten datasets to get ten replication results and calculate the mean and standard error of the evaluation metrics. In implementing our algorithm, we set the thresholding parameter τ as $2s$, and choose a step size of $\eta = 0.25$. In both our algorithm and the benchmarks, where the rank needs to be specified, we use the ground truth rank r. For the other parameters in the benchmarks, we select them using 4-fold cross-validation.

Table I summarizes the results. In terms of Frobenius norm estimation errors, POETRY consistently outperforms the other methods. In terms of selection performance, POETRY achieves a higher TPR and a lower FPR than the benchmark methods, which highlights its effectiveness in accurately identifying significant elements. We do not report the FPR for POET-D and AltProj-D, as both methods are designed for the low-rank plus diagonal structure, which naturally yields a zero FPR. Additionally, to underscore the efficiency of our algorithm, we have included the average CPU time in Table I. Since the SCM, POET-D, and POET estimators are direct methods, we do not provide their average CPU time. We observe that POETRY offers considerable computational advantages, being nearly three times faster than the convex method. When the dimension d gets larger, the average CPU time for estimators rises, with our estimator increasing at the slowest rate. Particularly, as d scales up to several hundred, the repetitive computation of spectral decomposition in LOREC becomes extremely time-consuming, leading to a dramatic increase in computational time.

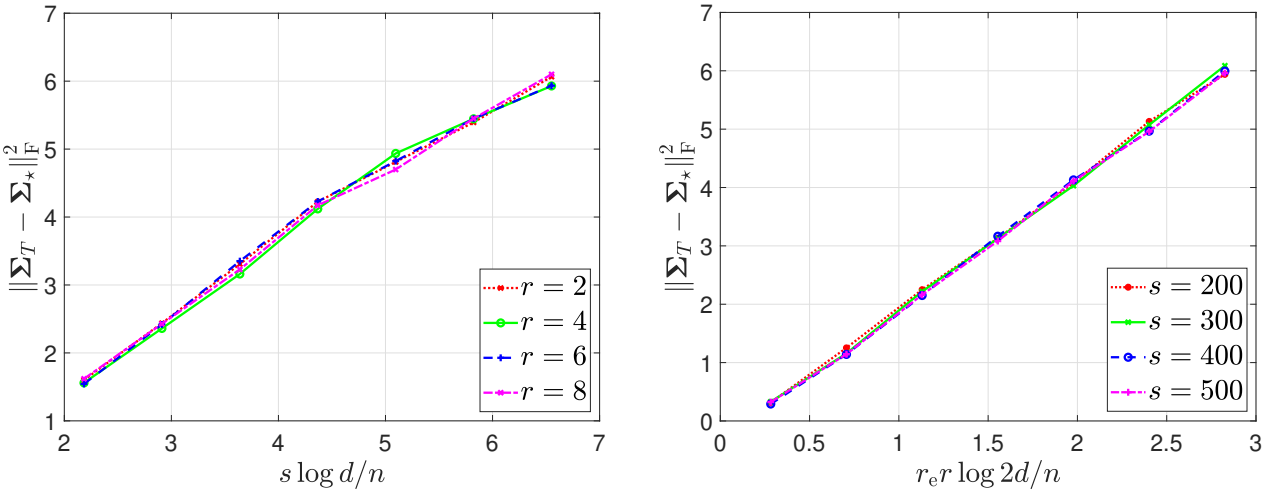
2) Statistical Convergence Rate: In this section, we verify the theoretical results presented in Theorems 5 and 7. We first analyze the squared estimation error, $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$, with the sample size n set to the magnitude of rd, as prescribed by Theorem 5. We plot this estimation error against the theoretical statistical error $s \log d/n$ for $r = 2, 4, 6, 8$ in Fig. 1(a). This figure shows that $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$ increases linearly with the statistical error $s \log d/n$. Similarly, Fig. 1(b) illustrates the estimation error against the statistical error rd/n for $s = 200, 300, 400, 500$. We observe that $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$ grows linearly with the statistical error rd/n. These results corroborate the theoretical statistical rate stated in Theorem 5, validating our guarantee of achieving the minimax optimal statistical rate. Next, we calculate the squared estimation error $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$ with the sample size n set to the magnitude of $r_{\mathrm{e}} r \log 2d$, as required in Theorem 7. The estimation error is plotted against the statistical error $s \log d/n$ for $r = 2, 4, 6, 8$ in Fig. 2(a). We observe that $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$ grows linearly with the statistical error $s \log d/n$. Furthermore, we also plot the estimation error against the statistical error $r \log (2d)/n$ for $s = 200, 300, 400, 500$ in Fig. 2(b). This plot shows that the estimation error grows linearly with the statistical error $r \log (2d)/n$. These numerical results align with the

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2025.3626934



(a) r fixed and varying n, d, s (b) s fixed and varying n, d, r

Fig. 1. Estimation errors $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$ versus theoretical statistical errors.



(a) r fixed and varying n, d, s (b) s fixed and varying n, d, r

Fig. 2. Estimation errors $\|\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2$ versus theoretical statistical errors.

theoretical findings of Theorem 7.

B. Real Datasets

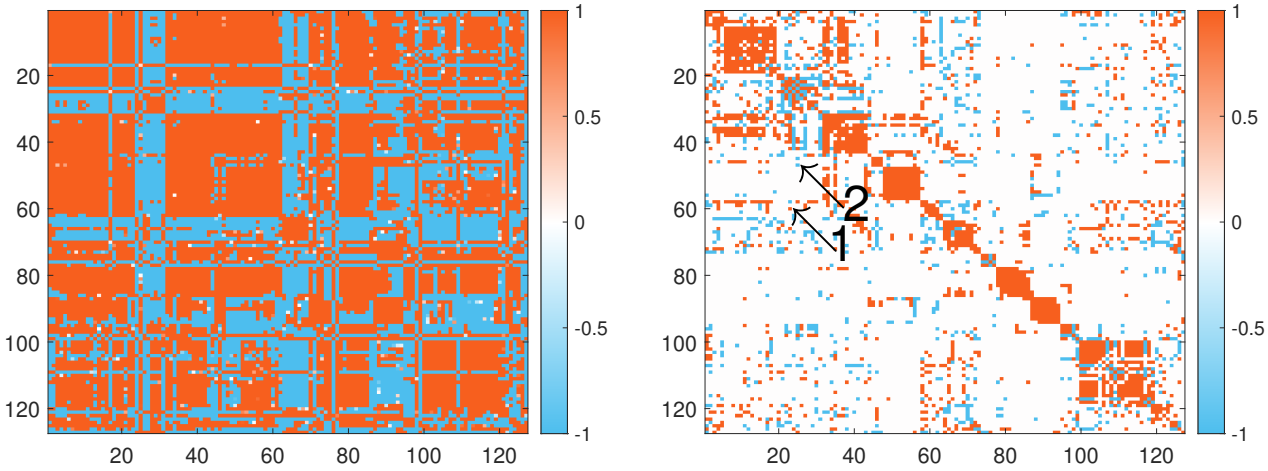
1) Macroeconomic Factors: The federal reserve economic data–monthly database (FRED-MD)[1] is an open-source dataset containing monthly U.S. macroeconomic indicators. It is maintained by the Federal Reserve Bank of St. Louis [70] and serves as a common benchmark for evaluating model performance and enhancing research reproducibility. FRED-MD covers a wide range of macroeconomic time series that cover key aspects of the U.S. economy, including employment, production, income, and prices. The dataset includes 127 standard U.S. macroeconomic indicators, with data available from January 1, 1959. These 127 indicators are categorized into seven groups based on their characteristics. To prepare the data for statistical modeling, various transformations are applied to the time series to make them stationary. The transformations include: (1) no transformation; (2) first-order difference; (3) second-order difference; (4) logarithm; (5) first-order logarithmic difference; (6) second-order logarithmic difference; and (7) percentage change. In [70], eight factors from the FRED-MD dataset were estimated using the POET method [41]. In this section, we conduct an analysis based on the proposed method, POETRY. Our analysis begins with a sample starting on January 1, 1960. After accounting for two lost observations due to data transformation, the dataset used for analysis spans from March 1, 1960, to April 1, 2015, resulting in a total of 658 observations.

First, we will verify the number of factors identified by [70] and then analyze the estimated sparse component. We define the correlation matrix estimator as follows:

$$\widehat{\boldsymbol{\Gamma}}_T = \left[\mathrm{diag}\left(\widehat{\boldsymbol{\Sigma}}_T\right)\right]^{-1/2} \widehat{\boldsymbol{\Sigma}}_T \left[\mathrm{diag}\left(\widehat{\boldsymbol{\Sigma}}_T\right)\right]^{-1/2}.$$

[1]The dataset is publicly available and can be downloaded from the federal reserve economic data website at https://fred.stlouisfed.org/.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2025.3626934



(a) low-rank component (b) sparse component

Fig. 3. Heatmaps of the low-rank component and sparse component.

According to [71], the number of eigenvalues greater than 1 in the correlation matrix is often used to determine the number of factors present. When we apply the POETRY method and set $r \geq 8$, we consistently observe 8 eigenvalues greater than 1. This suggests that the rank of the low-rank component should be chosen as 8, aligning with the results presented in [70]. Setting $r = 8$, we obtain eight factors from the correlation matrix, with eigenvalues greater than 1 being 18.70, 9.58, 8.59, 6.66, 5.65, 4.43, 3.94, and 2.90. For the parameter τ, we select its value from $0.05 \times d^2, 0.25 \times d^2, 0.45 \times d^2$, and $0.65 \times d^2$ by choosing the value that yields the lowest objective function value. We visualize both the low-rank component and the sparse component of the covariance matrix using heatmaps in Fig. 3. Both components are normalized by the largest element in magnitude within their respective matrices. This visualization helps us understand the underlying structure and correlations within the dataset. Next, we examine the estimated sparse component, which is illustrated in Fig. 3(b). The sparse component provides intriguing insights. For example, the element indicated by the first arrow shows a positive correlation between the civilian labor force and orders for consumer goods. An expanding labor force typically corresponds to increased consumer goods orders, reflecting overall economic growth and consumer confidence. In addition, the element pointed to by the second arrow reveals a negative correlation between the average duration of unemployment and average weekly hours worked. Generally, these two variables are inversely related: a longer duration of unemployment is associated with fewer average weekly hours worked, indicating a weaker labor market. We also examine other significant nonzero correlations that seem reasonable within this context.

2) Portfolio Design: We further apply our covariance estimation methods for portfolio design in financial stock markets[2]. We collect the historical daily stock prices of the components listed in the S&P 500 Index from January 1st, 2010, to December 31st, 2020. After removing stocks with missing values, we obtain daily returns of 415 stocks ($d = 415$). We construct 100 daily return time series datasets with random starting dates, each containing daily returns over 252 consecutive trading days. We first estimate the covariance matrix of each dataset. For each method, we select the parameters that yield the lowest portfolio risk. We then compute the risk of the portfolio return designed based on these estimated covariance matrices. In this context, the quality of the estimated covariance matrix is directly related to portfolio risk; a higher quality estimation leads to lower portfolio risk. This allows us to assess and compare the effectiveness of different estimation methods in minimizing portfolio risk. In this paper, we adopt the global minimum variance portfolio (GMVP) [72]. Let $\boldsymbol{\Sigma}$ be a generic estimator of the return covariance matrix, and let $\boldsymbol{w}$ be the allocation vector of a portfolio consisting of the corresponding d stocks. The GMVP is defined as the solution to the following convex optimization problem:

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \quad & \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^\top \mathbf{1} = 1, \quad \boldsymbol{w} \geq \mathbf{0}, \end{aligned}$$

where the objective represents the risk of a portfolio. The solution to this optimization problem is obtained using CVX, a MATLAB-based modeling system for convex optimization [73].

The boxplot results presented in Fig. 4 illustrate the risk of the portfolios generated by different covariance estimators. This figure indicates that the portfolio risk associated with SCM is higher than other methods, suggesting its inferior performance. Additionally, we find that POET-D and AltProj-D, which are based on strict factor models for covariance estimation, perform less effectively compared to those utilizing approximate factor models. This outcome highlights the advantages of the approximate factor model assumption. Most notably, the GMVP generated by POETRY exhibits the lowest risk among the methods analyzed. This result underscores the effectiveness of POETRY in constructing a robust and low-risk portfolio, thereby validating the superiority of our method in practical financial applications.

[2]The dataset can be downloaded from https://finance.yahoo.com/.

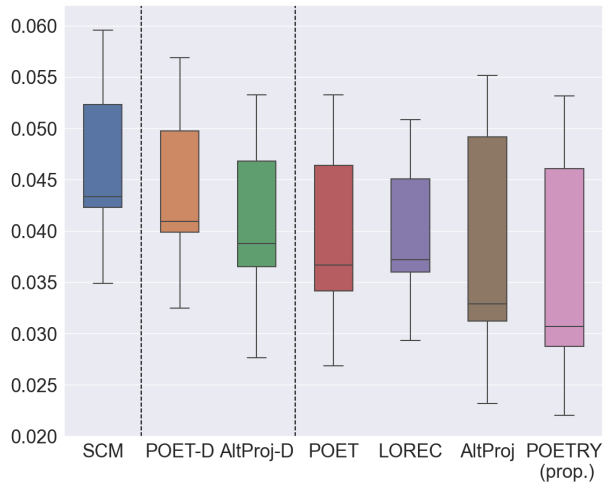


Fig. 4. Boxplots of daily volatility under different estimators on stock market data.

VI. Conclusions

In this paper, we have studied the problem of estimating large covariance matrices based on factor models, where the covariance matrix is decomposed into a low-rank matrix and a sparse matrix. We have proposed a nonconvex optimization model, together with an efficient numerical algorithm. Furthermore, we have established that the sequence generated by our algorithm converges linearly to the ground truth at a minimax statistical rate. We have also derived the statistical rate of the inverse covariance matrix estimator. By introducing the concept of effective rank, we have reduced the required sample size and further improved the statistical rate. Numerical experiments support our theoretical findings and demonstrate the superiority of our algorithm.

Appendix

A. Auxiliary Lemmata

Lemma 8 ([45]). *For the noise matrix $\boldsymbol{N} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^\top - \boldsymbol{\Sigma}_\star$, where each $\boldsymbol{y}_i \in \mathbb{R}^d$ is a Gaussian random variable with mean zero and covariance $\boldsymbol{\Sigma}_\star$, with probability at least $1 - 2\exp(-C_1 d)$, it holds that*

$$\|\boldsymbol{N}\|_2 \le \alpha\sqrt{\frac{d}{n}},$$

where $\alpha = 4\|\boldsymbol{\Sigma}_\star\|_2$. Moreover, with probability at least $1 - 2\exp(-C_2 \log d)$, we have

$$\|\boldsymbol{N}\|_\infty \le \beta\sqrt{\frac{\log d}{n}},$$

and $\beta = 8 \times \max_j (\boldsymbol{\Sigma}_\star)_{jj}$.

Lemma 9 ([68]). *For the noise matrix $\boldsymbol{N} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^\top - \boldsymbol{\Sigma}_\star$, where each $\boldsymbol{y}_i \in \mathbb{R}^d$ is a Gaussian random variable with mean zero and covariance $\boldsymbol{\Sigma}_\star$, with probability at least $1 - 1/(2d)$, it holds that*

$$\|\boldsymbol{N}\|_2 \le \gamma\left(\sqrt{\frac{2r_{\mathrm{e}}\log(2d)}{n}} \vee \frac{2r_{\mathrm{e}}\log(2d)\,(3/8 + \log n)}{n}\right),$$

where γ is a positive constant related to $\boldsymbol{\Sigma}_\star$.

Given the spectral decomposition $\boldsymbol{B}_\star \boldsymbol{B}_\star^\top = \boldsymbol{V}_\star \boldsymbol{\Lambda}_\star \boldsymbol{V}_\star^\top$, we define $\boldsymbol{B}_\star = \boldsymbol{V}_\star \boldsymbol{\Lambda}_\star^{1/2}$.

Definition 10. Define the distance between $\boldsymbol{B}$ and $\boldsymbol{B}_\star$ as

$$\operatorname{dist}(\boldsymbol{B}, \boldsymbol{B}_\star) := \min_{\boldsymbol{Q}\in\mathbb{O}^r} \left\|(\boldsymbol{B}\boldsymbol{Q} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}, \tag{9}$$

where $\mathbb{O}^r = \left\{\boldsymbol{Q} \in \mathbb{R}^{r\times r} \mid \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}\right\}$.

The factorization of the low-rank component $\boldsymbol{L}_\star = \boldsymbol{B}_\star \boldsymbol{B}_\star^\top$ is not unique, as $\boldsymbol{L}_\star = (\boldsymbol{B}_\star\boldsymbol{Q})(\boldsymbol{B}_\star\boldsymbol{Q})^\top$ holds for any orthogonal matrix $\boldsymbol{Q}$. Therefore, in Definition 10, we define the distance error metric within the ambiguity class considering orthonormal transformations. Additionally, we take into account diagonal scaling in the metric to properly address the impact of preconditioning [63]. We denote the optimal alignment matrix $\boldsymbol{Q}$ between $\boldsymbol{B}$ and $\boldsymbol{B}_\star$ as

$$\boldsymbol{Q} := \arg\min_{\boldsymbol{Q}\in\mathbb{O}^r} \left\|(\boldsymbol{B}\boldsymbol{Q} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}.$$

This implies that $\boldsymbol{Q}$ satisfies the following first-order necessary optimality condition:

$$(\boldsymbol{B}\boldsymbol{Q})^\top (\boldsymbol{B}\boldsymbol{Q} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star = \boldsymbol{0}. \tag{10}$$

Based on Definition 10, we can obtain the following results.

Lemma 11 ([74], [75]). *For any loading matrix $\boldsymbol{B}$, the distance between $\boldsymbol{B}$ and $\boldsymbol{B}_\star$ satisfies*

$$\operatorname{dist}(\boldsymbol{B}, \boldsymbol{B}_\star) \le \left(\frac{\sqrt{2}+1}{2}\right)^{1/2} \left\|\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{L}_\star\right\|_{\mathrm{F}}.$$

Lemma 12. *For any loading matrix $\boldsymbol{B}$ that satisfies the condition $\left\|(\boldsymbol{B} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2 < 1$, the following holds:*

$$\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\right\|_2 \le \frac{1}{1 - \left\|(\boldsymbol{B} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2},$$

Proof: Notice that

$$\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\right\|_2 = \frac{1}{\sigma_r\left(\boldsymbol{B}\boldsymbol{\Lambda}_\star^{-1/2}\right)}.$$

Based on Weyl's inequality, we obtain

$$\begin{aligned}\sigma_r\left(\boldsymbol{B}\boldsymbol{\Lambda}_\star^{-1/2}\right) &\ge \sigma_r\left(\boldsymbol{B}_\star\boldsymbol{\Lambda}_\star^{-1/2}\right) - \left\|(\boldsymbol{B} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2 \\ &= 1 - \left\|(\boldsymbol{B} - \boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2,\end{aligned}$$

where we have used $\sigma_r\left(\boldsymbol{B}_\star\boldsymbol{\Lambda}_\star^{-1/2}\right) = \sigma_r(\boldsymbol{V}_\star) = 1$. The result is proved by combining the preceding two relations. ■

Lemma 13. *Let $\boldsymbol{\Psi} = \boldsymbol{H}_\tau\left(\boldsymbol{S} - \boldsymbol{B}\boldsymbol{B}^\top\right)$, where $s < \tau \le \pi s$ with $\pi \ge 1$, we have*

$$\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_\star\|_\infty \le 2\left\|\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{L}_\star\right\|_\infty + 2\|\boldsymbol{N}\|_\infty, \tag{11}$$

and

$$\begin{aligned}&\left|\operatorname{tr}\left((\boldsymbol{\Psi} - \boldsymbol{\Psi}_\star)\boldsymbol{Z}^\top\right)\right| \\ &\le \left(2\left\|\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{L}_\star\right\|_{\mathrm{F}} + \left(1 + \sqrt{\pi+1}\right)\sqrt{s}\,\|\boldsymbol{N}\|_\infty\right)\|\boldsymbol{Z}\|_{\mathrm{F}},\end{aligned} \tag{12}$$

for any matrix $\boldsymbol{Z}$.

Proof: Let $\mathcal{S}$ and $\mathcal{S}_\star$ represent the support set of $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}_\star$, respectively. Consequently, $\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star$ is supported on $\mathcal{S}\cup\mathcal{S}_\star$. Define $\boldsymbol{\Delta}_L = \boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{L}_\star$. For any $(i,j)\in\mathcal{S}$, based on the definition of $\boldsymbol{H}_\tau(\cdot)$, we have $(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star)_{ij} = \boldsymbol{N}_{ij} - (\boldsymbol{\Delta}_L)_{ij}$. For any $(i,j)\in\mathcal{S}_\star\setminus\mathcal{S}$, one necessarily has $\boldsymbol{\Psi}_{ij}=0$, thus leading to $(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star)_{ij} = -(\boldsymbol{\Psi}_\star)_{ij}$. Again, by the definition of $\boldsymbol{H}_\tau(\cdot)$, we know that $|\boldsymbol{\Psi}_\star+\boldsymbol{N}-\boldsymbol{\Delta}_L|_{ij} \le |\boldsymbol{\Psi}_\star+\boldsymbol{N}-\boldsymbol{\Delta}_L|_\tau$, where $|\boldsymbol{A}|_\tau$ is the τ-th largest absolute entry of matrix $\boldsymbol{A}$. Furthermore, we know that $\boldsymbol{\Psi}_\star$ contains at most s nonzero entries and $\tau > s$. Consequently, one has $|\boldsymbol{\Psi}_\star+\boldsymbol{N}-\boldsymbol{\Delta}_L|_{ij} \le |\boldsymbol{N}-\boldsymbol{\Delta}_L|_\tau$. Combining the two cases above, we conclude that

$$|\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star|_{ij} \le \begin{cases} |\boldsymbol{N}|_{ij} + |\boldsymbol{\Delta}_L|_{ij}, & (i,j)\in\mathcal{S}, \\ |\boldsymbol{N}|_{ij} + |\boldsymbol{\Delta}_L|_{ij} + |\boldsymbol{N}|_\tau + |\boldsymbol{\Delta}_L|_\tau, & (i,j)\in\mathcal{S}_\star\setminus\mathcal{S}, \end{cases} \tag{13}$$

which immediately implies the result in (11).

Recall that $\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star$ is supported on $\mathcal{S}\cup\mathcal{S}_\star$. Therefore, we have

$$\begin{aligned} &\left|\mathrm{tr}\left((\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star)\boldsymbol{Z}^\top\right)\right| \\ &\le \sum_{(i,j)\in\mathcal{S}} |\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star|_{ij}\,|\boldsymbol{Z}|_{ij} + \sum_{(i,j)\in\mathcal{S}_\star\setminus\mathcal{S}} |\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star|_{ij}\,|\boldsymbol{Z}|_{ij} \\ &\le \underbrace{\sum_{(i,j)\in\mathcal{S}\cup\mathcal{S}_\star} |\boldsymbol{N}|_{ij}\,|\boldsymbol{Z}|_{ij}}_{I_1} + \underbrace{\sum_{(i,j)\in\mathcal{S}\cup\mathcal{S}_\star} |\boldsymbol{\Delta}_L|_{ij}\,|\boldsymbol{Z}|_{ij}}_{I_2} \\ &\quad + \underbrace{\sum_{(i,j)\in\mathcal{S}_\star\setminus\mathcal{S}} |\boldsymbol{N}|_\tau\,|\boldsymbol{Z}|_{ij}}_{I_3} + \underbrace{\sum_{(i,j)\in\mathcal{S}_\star\setminus\mathcal{S}} |\boldsymbol{\Delta}_L|_\tau\,|\boldsymbol{Z}|_{ij}}_{I_4}, \end{aligned}$$

where the second inequality follows from (13). We can bound I_1 as follows:

$$I_1 \le \sqrt{\sum_{(i,j)\in\mathcal{S}\cup\mathcal{S}_\star} |\boldsymbol{N}|_{ij}^2}\,\|\boldsymbol{Z}\|_{\mathrm{F}} \le \sqrt{(\pi+1)\,s}\,\|\boldsymbol{N}\|_\infty\,\|\boldsymbol{Z}\|_{\mathrm{F}},$$

where the first inequality follows from the Cauchy inequality, and the second inequality follows from the fact that $\mathcal{S}\cup\mathcal{S}_\star$ contains at most $\tau+s$ nonzero entries with $s<\tau\le\pi s$. For I_2, we invoke the Cauchy inequality to obtain

$$I_2 \le \|\boldsymbol{\Delta}_L\|_{\mathrm{F}}\,\|\boldsymbol{Z}\|_{\mathrm{F}}.$$

We can further upperbound I_3 as

$$I_3 \le \|\boldsymbol{N}\|_\infty \sqrt{s\sum_{(i,j)\in\mathcal{S}_\star} |\boldsymbol{Z}|_{ij}^2} \le \sqrt{s}\,\|\boldsymbol{N}\|_\infty\,\|\boldsymbol{Z}\|_{\mathrm{F}},$$

where the first inequality follows from the Jensen's inequality $\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 \le \frac{1}{n}\sum_{i=1}^n x_i^2$.

We can also further bound I_4 as

$$I_4 \le |\boldsymbol{\Delta}_L|_\tau \sum_{(i,j)\in\mathcal{S}_\star} |\boldsymbol{Z}|_{ij} \le \sqrt{s}\,|\boldsymbol{\Delta}_L|_\tau\,\|\boldsymbol{Z}\|_{\mathrm{F}} \le \|\boldsymbol{\Delta}_L\|_{\mathrm{F}}\,\|\boldsymbol{Z}\|_{\mathrm{F}},$$

where the last inequality is due to $\sqrt{s}\,|\boldsymbol{\Delta}_L|_\tau \le \|\boldsymbol{\Delta}_L\|_{\mathrm{F}}$. Combine I_1, I_2, I_3, and I_4 together to obtain the result in (12). ∎

B. Propositions

Proposition 14. *Suppose Assumptions 1 and 2 hold. Assume that the sample size satisfies $n \gtrsim rd$. Choose the thresholding parameter as $s<\tau\le\pi s$ with $\pi\ge 1$. Then, with probability at least $1-2\exp(-C_1 d)$, the initialization $\boldsymbol{B}_0$ of Algorithm 2 satisfies*

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) < \sigma_r(\boldsymbol{L}_\star),$$

where C_1 is defined in Lemma 8.

Proof: From Lemma 11, we have

$$\begin{aligned} \mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) &\le \left(\frac{\sqrt{2}+1}{2}\right)^{1/2} \left\|\boldsymbol{B}_0\boldsymbol{B}_0^\top - \boldsymbol{L}_\star\right\|_{\mathrm{F}} \\ &\le \sqrt{\left(\sqrt{2}+1\right)r}\left\|\boldsymbol{B}_0\boldsymbol{B}_0^\top - \boldsymbol{L}_\star\right\|_2, \end{aligned}$$

where the last inequality is because $\boldsymbol{B}_0\boldsymbol{B}_0^\top - \boldsymbol{L}_\star$ has a rank of at most $2r$. By applying the triangle inequality, we can derive

$$\begin{aligned} \left\|\boldsymbol{B}_0\boldsymbol{B}_0^\top - \boldsymbol{L}_\star\right\|_2 &\le \left\|\boldsymbol{S}-\boldsymbol{B}_0\boldsymbol{B}_0^\top\right\|_2 + \|\boldsymbol{S}-\boldsymbol{L}_\star\|_2 \\ &\le 2\,\|\boldsymbol{S}-\boldsymbol{L}_\star\|_2 \\ &= 2\,\|\boldsymbol{N}+\boldsymbol{\Sigma}_\star-\boldsymbol{L}_\star\|_2 \\ &\le 2\left(\|\boldsymbol{N}\|_2 + \|\boldsymbol{\Sigma}_\star\|_2 + \|\boldsymbol{L}_\star\|_2\right), \end{aligned} \tag{14}$$

where the second inequality is based on the fact that $\boldsymbol{B}_0\boldsymbol{B}_0^\top$ is the best rank-r approximation of $\boldsymbol{S}$. Hence, we have

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) \le 2\sqrt{\left(\sqrt{2}+1\right)r}\left(\|\boldsymbol{N}\|_2 + \mu + k\right), \tag{15}$$

where the inequality is due to $\|\boldsymbol{\Sigma}_\star\|_2 \le \mu$ from Assumption 1 and $\|\boldsymbol{L}_\star\|_2 \le \|\boldsymbol{L}_\star\|_{\mathrm{F}} \le d\,\|\boldsymbol{L}_\star\|_\infty \le k$ from Assumption 2, we can apply Lemma 8 to obtain the following result

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) \le 2\sqrt{\left(\sqrt{2}+1\right)r}\left(\beta\sqrt{\frac{d}{n}} + \mu + k\right),$$

which holds with probability at least $1-2\exp(-C_1 d)$. With $n > 48rd/\sigma_r^2(\boldsymbol{L}_\star)$, we can then conclude that

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) < \sigma_r(\boldsymbol{L}_\star).$$

∎

Proposition 15. *Suppose Assumptions 1 and 2 hold. Assume that the sample size satisfies $n \gtrsim r_{\mathrm{e}} r\log 2d$. Choose the thresholding parameter as $s<\tau\le\pi s$ with $\pi\ge 1$. Then, with probability at least $1-1/(2d)$, the initialization $\boldsymbol{B}_0$ of Algorithm 2 satisfies*

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) < \sigma_r(\boldsymbol{L}_\star).$$

Proof: From (15), we have

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) \le 2\sqrt{\left(\sqrt{2}+1\right)r}\left(\|\boldsymbol{N}\|_2 + \mu + k\right).$$

Based on Lemma 9 and $r_{\mathrm{e}} \ll d$, we have that with probability at least $1-1/(2d)$,

$$\mathrm{dist}(\boldsymbol{B}_0,\boldsymbol{B}_\star) \le 2\sqrt{\left(\sqrt{2}+1\right)r}\left(\gamma\sqrt{\frac{2r_{\mathrm{e}}\log(2d)}{n}} + \mu + k\right).$$

Then, with $n > 12r_{\mathrm{e}}r\log 2d/\sigma_r^2\left(\boldsymbol{L}_\star\right)$, we have

$$\operatorname{dist}\left(\boldsymbol{B}_0,\boldsymbol{B}_\star\right) < \sigma_r\left(\boldsymbol{L}_\star\right).$$

■

Proposition 16. *Given the update of $\boldsymbol{B}_{t+1}$ in Algorithm 2, if the step size satisfies $\frac{2-\sqrt{2}}{20} < \eta < \frac{1}{2}$, then for any $t = 0, 1, \ldots$, the following inequality holds*

$$\operatorname{dist}^2\left(\boldsymbol{B}_{t+1},\boldsymbol{B}_\star\right) \le \rho^{t+1}\sigma_r^2\left(\boldsymbol{L}_\star\right) + \frac{\omega}{1-\rho},$$

where $\rho = 2\left(1-10\eta\right)^2$ and $\omega = \frac{4\eta^2 s\left(1+\sqrt{\pi+1}\right)^2}{(1-\delta)^2}\left\|\boldsymbol{N}\right\|_\infty^2 + \frac{2\eta^2 r}{(1-\delta)^2}\left\|\boldsymbol{N}\right\|_2^2$.

Proof: By the definition of $\operatorname{dist}\left(\boldsymbol{B}_{t+1},\boldsymbol{B}_\star\right)$, it can be shown that

$$\operatorname{dist}^2\left(\boldsymbol{B}_{t+1},\boldsymbol{B}_\star\right) \le \left\|\left(\boldsymbol{B}_{t+1}\boldsymbol{Q}_t - \boldsymbol{B}_\star\right)\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2,$$

where $\boldsymbol{Q}_t$ is the optimal alignment matrix between $\boldsymbol{B}_t$ and $\boldsymbol{B}_\star$. Denote $\boldsymbol{B} = \boldsymbol{B}_t\boldsymbol{Q}_t$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}_t$, $\boldsymbol{\Delta}_B = \boldsymbol{B} - \boldsymbol{B}_\star$, and $\boldsymbol{\Delta}_L = \boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{L}_\star$. Given the update rule $\boldsymbol{B}_{t+1} = \boldsymbol{B}_t - \eta\left(\boldsymbol{B}_t\boldsymbol{B}_t^\top + \boldsymbol{\Psi}_t - \boldsymbol{S}\right)\boldsymbol{B}_t\left(\boldsymbol{B}_t^\top\boldsymbol{B}_t\right)^{-1}$, one has

$$\begin{aligned}
&\left(\boldsymbol{B}_{t+1}\boldsymbol{Q}_t - \boldsymbol{B}_\star\right)\boldsymbol{\Lambda}_\star^{1/2}\\
&= \left(\boldsymbol{B} - \eta\left(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Psi} - \boldsymbol{S}\right)\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1} - \boldsymbol{B}_\star\right)\boldsymbol{\Lambda}_\star^{1/2}\\
&= \boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2} - \eta\left(\boldsymbol{\Psi} - \boldsymbol{\Psi}_\star\right)\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\\
&\quad - \eta\boldsymbol{\Delta}_L\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2} + \eta\boldsymbol{N}\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\\
&= (1-\eta)\,\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2} - \eta\left(\boldsymbol{\Psi} - \boldsymbol{\Psi}_\star\right)\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\\
&\quad - \eta\boldsymbol{B}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2} + \eta\boldsymbol{N}\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2},
\end{aligned} \tag{16}$$

where the second equality is due to $\boldsymbol{S} = \boldsymbol{L}_\star + \boldsymbol{\Psi}_\star + \boldsymbol{N}$ and the last equality is due to the decomposition $\boldsymbol{\Delta}_L = \boldsymbol{\Delta}_B\boldsymbol{B}^\top + \boldsymbol{B}_\star\boldsymbol{\Delta}_B^\top$. Take the Frobenius norm of both sides of (16) and use $(a+b)^2 \le 2a^2 + 2b^2$ to obtain (17), where $\boldsymbol{Z} = \boldsymbol{B}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top$. For the first term of (17), it is easy to verify that $\operatorname{tr}\left(\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\right) = \operatorname{dist}^2\left(\boldsymbol{B}_t,\boldsymbol{B}_\star\right)$. In the following proof, we will focus on controlling J_1, J_2, J_3, J_4, J_5, and J_6 separately.

We start with J_1. Since $\boldsymbol{B}_\star = \boldsymbol{B} - \boldsymbol{\Delta}_B$ and $\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{B} = \boldsymbol{0}$ from (10), we have

$$\begin{aligned}
J_1 &= \operatorname{tr}\left(\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{B}_\star\boldsymbol{\Delta}_B^\top\right)\\
&= \operatorname{tr}\left(\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{B}\boldsymbol{\Delta}_B^\top\right)\\
&\quad - \operatorname{tr}\left(\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{\Delta}_B\boldsymbol{\Delta}_B^\top\right)\\
&= -\operatorname{tr}\left(\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{\Delta}_B\boldsymbol{\Delta}_B^\top\right).
\end{aligned}$$

We intend to show that for any $t = 0, 1, \ldots$, it holds that

$$\operatorname{dist}\left(\boldsymbol{B}_t,\boldsymbol{B}_\star\right) \le \delta\sigma_r\left(\boldsymbol{L}_\star\right), \tag{18}$$

where $0 < \delta < 1$. We prove this result by induction. For $t = 0$, the induction hypothesis (18) holds trivially, as established by Propositions 14 and 15. We therefore concentrate on the induction step from t to $t+1$, for $t \ge 0$. Specifically, suppose

$$\operatorname{dist}\left(\boldsymbol{B}_t,\boldsymbol{B}_\star\right) = \left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{-1/2}\boldsymbol{\Lambda}_\star\right\|_{\mathrm{F}} \le \delta\sigma_r\left(\boldsymbol{L}_\star\right).$$

Since $\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{-1/2}\boldsymbol{\Lambda}_\star\right\|_{\mathrm{F}} \ge \sigma_r\left(\boldsymbol{L}_\star\right)\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{-1/2}\right\|_{\mathrm{F}} \ge \sigma_r\left(\boldsymbol{L}_\star\right)\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2$, we have

$$\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2 \le \delta. \tag{19}$$

Combining (19) with Lemma 12 leads to

$$\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\right\|_2 \le \frac{1}{1-\delta}. \tag{20}$$

Then, we have

$$\begin{aligned}
|J_1| &= \left|\operatorname{tr}\left(\boldsymbol{\Lambda}_\star^{-1/2}\boldsymbol{\Delta}_B^\top\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2}\right)\right|\\
&\le \left\|\boldsymbol{\Lambda}_\star^{-1/2}\boldsymbol{\Delta}_B^\top\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\right\|_2\operatorname{tr}\left(\boldsymbol{\Lambda}_\star^{1/2}\boldsymbol{\Delta}_B^\top\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2}\right)\\
&\le \left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{-1/2}\right\|_2\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\right\|_2\operatorname{tr}\left(\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\right)\\
&\le \frac{\delta}{1-\delta}\operatorname{tr}\left(\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\right),
\end{aligned}$$

where the last inequality is due to (19) and (20).

Moving on to bound J_2, we can utilize the fact $\boldsymbol{B}^\top\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star = \boldsymbol{0}$ from (10) to obtain

$$\begin{aligned}
J_2 &= \operatorname{tr}\left(\boldsymbol{B}_\star\boldsymbol{\Delta}_B^\top\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top\boldsymbol{\Delta}_B\boldsymbol{B}_\star^\top\right)\\
&= \operatorname{tr}\left(\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star\boldsymbol{\Delta}_B^\top\right) = 0.
\end{aligned}$$

For J_3, we have

$$\begin{aligned}
|J_3| &\le \left(2\left\|\boldsymbol{\Delta}_L\right\|_{\mathrm{F}} + \left(1+\sqrt{\pi+1}\right)\sqrt{s}\left\|\boldsymbol{N}\right\|_\infty\right)\\
&\quad\cdot\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top\right\|_{\mathrm{F}}\\
&\le \left(2\left(2+\delta\right)\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}} + \left(1+\sqrt{\pi+1}\right)\sqrt{s}\left\|\boldsymbol{N}\right\|_\infty\right)\\
&\quad\cdot\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{B}\right)^{-1}\boldsymbol{\Lambda}_\star^{1/2}\right\|_2\\
&\le \frac{2\left(2+\delta\right)}{1-\delta}\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2\\
&\quad + \frac{\left(1+\sqrt{\pi+1}\right)\sqrt{s}}{1-\delta}\left\|\boldsymbol{N}\right\|_\infty\left\|\boldsymbol{\Delta}_B\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}},
\end{aligned}$$

where the first inequality is due to (12) in Lemma 13, and in the second inequality, we have used the relation $\left\|\boldsymbol{A}\boldsymbol{X}\right\|_{\mathrm{F}} \le$

$$
\begin{aligned}
\operatorname{dist}^2\left(\boldsymbol{B}_{t+1}, \boldsymbol{B}_\star\right) \leq & 2\left\|(1-\eta) \boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}-\eta\left(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star\right) \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}-\eta \boldsymbol{B}_\star \boldsymbol{\Delta}_B^\top \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2 \\
& +2\left\|\eta \boldsymbol{N} \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2 \\
\leq & 2(1-\eta)^2 \operatorname{tr}\left(\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star \boldsymbol{\Delta}_B^\top\right)-4 \eta(1-\eta) \underbrace{\operatorname{tr}\left(\boldsymbol{B}_\star \boldsymbol{\Delta}_B^\top \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star \boldsymbol{\Delta}_B^\top\right)}_{J_1} \\
& +2 \eta^2 \underbrace{\left\|\boldsymbol{B}_\star \boldsymbol{\Delta}_B^\top \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2}_{J_2}-4 \eta(1-\eta) \underbrace{\operatorname{tr}\left(\left(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star\right) \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star \boldsymbol{\Delta}_B^\top\right)}_{J_3} \\
& +4 \eta^2 \underbrace{\operatorname{tr}\left(\left(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star\right) \boldsymbol{Z}^\top\right)}_{J_4}+2 \eta^2 \underbrace{\left\|\left(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star\right) \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2}_{J_5}+2 \eta^2 \underbrace{\left\|\boldsymbol{N} \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2}_{J_6}
\end{aligned} \tag{17}
$$

$\|\boldsymbol{A}\|_2\|\boldsymbol{X}\|_{\mathrm{F}}$ and the following result

$$
\begin{aligned}
\left\|\boldsymbol{\Delta}_L\right\|_{\mathrm{F}} & =\left\|\boldsymbol{\Delta}_B \boldsymbol{B}_\star^\top+\boldsymbol{B}_\star \boldsymbol{\Delta}_B^\top+\boldsymbol{\Delta}_B \boldsymbol{\Delta}_B^\top\right\|_{\mathrm{F}} \\
& \leq\left\|\boldsymbol{\Delta}_B \boldsymbol{B}_\star^\top\right\|_{\mathrm{F}}+\left\|\boldsymbol{B}_\star \boldsymbol{\Delta}_B^\top\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Delta}_B \boldsymbol{\Delta}_B^\top\right\|_{\mathrm{F}} \\
& =2\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\left(\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{-1/2}\right)^\top\right\|_{\mathrm{F}} \\
& \leq 2\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{-1/2}\right\|_2\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}} \\
& \leq(2+\delta)\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}},
\end{aligned} \tag{21}
$$

and the last inequality is due to (20).

The term J_4 can be bounded similarly to J_3. We invoke (12) in Lemma 13 to obtain

$$
\left|J_4\right| \leq\left(2\left\|\boldsymbol{\Delta}_L\right\|_{\mathrm{F}}+\left(1+\sqrt{\pi+1}\right) \sqrt{s}\|\boldsymbol{N}\|_\infty\right)\|\boldsymbol{Z}\|_{\mathrm{F}} .
$$

Since

$$
\begin{aligned}
\|\boldsymbol{Z}\|_{\mathrm{F}} & =\left\|\boldsymbol{B}_\star \boldsymbol{\Delta}_B^\top \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top\right\|_{\mathrm{F}} \\
& \leq\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_2^2,
\end{aligned}
$$

we have

$$
\begin{aligned}
\left|J_4\right| \leq & \left(2\left\|\boldsymbol{\Delta}_L\right\|_{\mathrm{F}}+\left(1+\sqrt{\pi+1}\right) \sqrt{s}\|\boldsymbol{N}\|_\infty\right) \\
& \cdot\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_2^2 \\
\leq & \frac{2(2+\delta)}{(1-\delta)^2}\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2 \\
& +\frac{\left(1+\sqrt{\pi+1}\right) \sqrt{s}}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}},
\end{aligned}
$$

where the last inequality is due to (20) and (21).

For the last term J_5, utilize the variational representation of the Frobenius norm to see

$$
\sqrt{J_5}=\operatorname{tr}\left(\left(\boldsymbol{\Psi}-\boldsymbol{\Psi}_\star\right) \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2} \tilde{\boldsymbol{B}}^\top\right)
$$

for some $\tilde{\boldsymbol{B}} \in \mathbb{R}^{d \times r}$ obeying $\left\|\tilde{\boldsymbol{B}}\right\|_{\mathrm{F}}=1$. Set $\boldsymbol{Z}=\tilde{\boldsymbol{B}} \boldsymbol{\Lambda}_\star^{1/2}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top$ and invoke (12) in Lemma 13 to arrive at

$$
\begin{aligned}
\sqrt{J_5} & \leq\left(2\left\|\boldsymbol{\Delta}_L\right\|_{\mathrm{F}}+\left(1+\sqrt{\pi+1}\right) \sqrt{s}\|\boldsymbol{N}\|_\infty\right) \\
& \quad \cdot\left\|\tilde{\boldsymbol{B}} \boldsymbol{\Lambda}_\star^{1/2}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top\right\|_{\mathrm{F}} \\
& \leq\left(2\left\|\boldsymbol{\Delta}_L\right\|_{\mathrm{F}}+\left(1+\sqrt{\pi+1}\right) \sqrt{s}\|\boldsymbol{N}\|_\infty\right) \\
& \quad \cdot\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_2 \\
& \leq \frac{2(2+\delta)}{1-\delta}\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}+\frac{\left(1+\sqrt{\pi+1}\right) \sqrt{s}}{1-\delta}\|\boldsymbol{N}\|_\infty,
\end{aligned} \tag{22}
$$

where the last inequality is due to (20) and (21). Square both sides of (22) and apply the elementary inequality $(a+b)^2 \leq 2a^2+2b^2$ to reach

$$
J_5 \leq \frac{8(2+\delta)^2}{(1-\delta)^2}\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2+\frac{2 s\left(1+\sqrt{\pi+1}\right)^2}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty^2 .
$$

For J_6, we have

$$
\begin{aligned}
J_6 & \leq\|\boldsymbol{N}\|_2^2\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2 \\
& \leq\|\boldsymbol{N}\|_2^2 r\left\|\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{\Lambda}_\star^{1/2}\right\|_2^2 \leq \frac{r}{(1-\delta)^2}\|\boldsymbol{N}\|_2^2,
\end{aligned}
$$

where the last inequality is due to (20). Taking collectively the bounds for J_1, J_2, J_3, J_4, J_5, and J_6, yields (23). Then identify $\operatorname{dist}^2\left(\boldsymbol{B}_t, \boldsymbol{B}_\star\right)=\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2$ and set $0<\eta \leq \frac{1-\delta}{2-\delta}$ to reach

$$
\operatorname{dist}^2\left(\boldsymbol{B}_{t+1}, \boldsymbol{B}_\star\right) \leq \rho \operatorname{dist}^2\left(\boldsymbol{B}_t, \boldsymbol{B}_\star\right)+\omega, \tag{24}
$$

$$
\begin{aligned}
\operatorname{dist}^2\left(\boldsymbol{B}_{t+1}, \boldsymbol{B}_\star\right) \leq & 2(1-\eta)^2 \operatorname{tr}\left(\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star \boldsymbol{\Delta}_B^\top\right)-4 \eta(1-\eta) \frac{\delta}{1-\delta} \operatorname{tr}\left(\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star \boldsymbol{\Delta}_B^\top\right)-8 \eta(1-\eta) \frac{(2+\delta)}{1-\delta}\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}}^2 \\
& -4 \eta(1-\eta) \frac{\left(1+\sqrt{\pi+1}\right) \sqrt{s}}{1-\delta}\|\boldsymbol{N}\|_\infty\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}}+8 \eta^2 \frac{(2+\delta)}{(1-\delta)^2}\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}}^2 \\
& +4 \eta^2 \frac{\left(1+\sqrt{\pi+1}\right) \sqrt{s}}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}}+\frac{16 \eta^2(2+\delta)^2}{(1-\delta)^2}\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}}^2 \\
& +\frac{4 \eta^2 s\left(1+\sqrt{\pi+1}\right)^2}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty^2+\frac{2 \eta^2 r}{(1-\delta)^2}\|\boldsymbol{N}\|_2^2 \\
\leq & \left(2(1-\eta)^2-4 \eta(1-\eta) \frac{\delta}{1-\delta}-8 \eta(1-\eta) \frac{(2+\delta)}{1-\delta}+8 \eta^2 \frac{(2+\delta)}{(1-\delta)^2}+\frac{16 \eta^2(2+\delta)^2}{(1-\delta)^2}\right)\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}}^2 \\
& +\left(\frac{4 \eta^2}{1-\delta}-4 \eta(1-\eta)\right) \frac{\left(1+\sqrt{\pi+1}\right) \sqrt{s}}{1-\delta}\|\boldsymbol{N}\|_\infty\left\|\boldsymbol{\Delta}_B \boldsymbol{\Lambda}_\star^{1 / 2}\right\|_{\mathrm{F}} \\
& +\frac{4 \eta^2 s\left(1+\sqrt{\pi+1}\right)^2}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty^2+\frac{2 \eta^2 r}{(1-\delta)^2}\|\boldsymbol{N}\|_2^2 \qquad (23)
\end{aligned}
$$

where $\omega=\frac{4 \eta^2 s\left(1+\sqrt{\pi+1}\right)^2}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty^2+\frac{2 \eta^2 r}{(1-\delta)^2}\|\boldsymbol{N}\|_2^2$, and the contraction rate ρ is given by

$$
\begin{aligned}
\rho & =2(1-\eta)^2-4 \eta(1-\eta) \frac{\delta}{1-\delta} \\
& \quad-8 \eta(1-\eta) \frac{(2+\delta)}{1-\delta}+8 \eta^2 \frac{(2+\delta)}{(1-\delta)^2}+\frac{16 \eta^2(2+\delta)^2}{(1-\delta)^2} \\
& =2(1-\eta)^2-4 \eta(1-\eta) \frac{\delta+2(2+\delta)}{1-\delta} \\
& \quad+8 \eta^2 \frac{(2+\delta)+2(2+\delta)^2}{(1-\delta)^2} \\
& \leq 2(1-10 \eta)^2 .
\end{aligned}
$$

With $\frac{2-\sqrt{2}}{20}<\eta<\frac{1}{2}$, one has $\rho=2(1-10 \eta)^2<1$.

Suppose $\operatorname{dist}\left(\boldsymbol{B}_t, \boldsymbol{B}_\star\right)<\sigma_r\left(\boldsymbol{L}_\star\right)$, then we have

$$
\begin{aligned}
\operatorname{dist}^2\left(\boldsymbol{B}_{t+1}, \boldsymbol{B}_\star\right) & \leq \rho \delta^2 \sigma_r^2\left(\boldsymbol{L}_\star\right)+(1-\rho) \delta^2 \sigma_r^2\left(\boldsymbol{L}_\star\right) \\
& =\delta^2 \sigma_r^2\left(\boldsymbol{L}_\star\right),
\end{aligned}
$$

where in the first inequality we have used the fact that when the sample size n is sufficiently large, we can always ensure $\omega \leq(1-\rho) \delta^2 \sigma_r^2\left(\boldsymbol{L}_\star\right)$. Therefore, we have

$$
\operatorname{dist}\left(\boldsymbol{B}_{t+1}, \boldsymbol{B}_\star\right) \leq \delta \sigma_r\left(\boldsymbol{L}_\star\right). \qquad (25)
$$

By mathematical induction, we then have $\operatorname{dist}\left(\boldsymbol{B}_t, \boldsymbol{B}_\star\right) \leq \delta \sigma_r\left(\boldsymbol{L}_\star\right)$ in (25), for any $t=0,1, \ldots$

Since (24) holds uniformly for $t=0,1, \ldots$, we further obtain that

$$
\begin{aligned}
\operatorname{dist}^2\left(\boldsymbol{B}_{t+1}, \boldsymbol{B}_\star\right) & \leq \rho \operatorname{dist}^2\left(\boldsymbol{B}_t, \boldsymbol{B}_\star\right)+\omega \\
& \leq \rho^{t+1} \sigma_r^2\left(\boldsymbol{L}_\star\right)+\frac{\omega}{1-\rho}.
\end{aligned}
$$

C. Proof of Theorem 4

Proof: We have

$$
\begin{aligned}
\left\|\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_\star\right\|_{\mathrm{F}} & =\left\|\boldsymbol{B}_0 \boldsymbol{B}_0^\top+\boldsymbol{\Psi}_0-\boldsymbol{L}_\star-\boldsymbol{\Psi}_\star\right\|_{\mathrm{F}} \\
& \leq\left\|\boldsymbol{B}_0 \boldsymbol{B}_0^\top-\boldsymbol{L}_\star\right\|_{\mathrm{F}}+\left\|\boldsymbol{\Psi}_0-\boldsymbol{\Psi}_\star\right\|_{\mathrm{F}} \\
& \leq \sqrt{2 r}\left\|\boldsymbol{B}_0 \boldsymbol{B}_0^\top-\boldsymbol{L}_\star\right\|_2+\sqrt{(1+\pi) s}\left\|\boldsymbol{\Psi}_0-\boldsymbol{\Psi}_\star\right\|_\infty,
\end{aligned}
$$

where the last inequality uses the fact that $\boldsymbol{B}_0 \boldsymbol{B}_0^\top-\boldsymbol{L}_\star$ has rank at most $2r$, and $\boldsymbol{\Psi}_0-\boldsymbol{\Psi}_\star$ has nonzero elements at most $(1+\pi) s$. Based on (11) in Lemma 13, we have

$$
\left\|\boldsymbol{\Psi}_0-\boldsymbol{\Psi}_\star\right\|_\infty \leq 2\left\|\boldsymbol{B}_0 \boldsymbol{B}_0^\top-\boldsymbol{L}_\star\right\|_\infty+2\|\boldsymbol{N}\|_\infty .
$$

Therefore, it follows that

$$
\begin{aligned}
& \left\|\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_\star\right\|_{\mathrm{F}} \\
& \leq \sqrt{2 r}\left\|\boldsymbol{B}_0 \boldsymbol{B}_0^\top-\boldsymbol{L}_\star\right\|_2+2 \sqrt{(1+\pi) s}\left\|\boldsymbol{B}_0 \boldsymbol{B}_0^\top-\boldsymbol{L}_\star\right\|_\infty \\
& \quad+2 \sqrt{(1+\pi) s}\|\boldsymbol{N}\|_\infty \\
& \leq 2\left(\sqrt{2 r}+2 \sqrt{2(1+\pi) r s}\right)\left(\|\boldsymbol{N}\|_2+\left\|\boldsymbol{\Sigma}_\star\right\|_2+\left\|\boldsymbol{L}_\star\right\|_2\right) \\
& \quad+2 \sqrt{(1+\pi) s}\|\boldsymbol{N}\|_\infty \\
& \leq 2\left(\sqrt{2 r}+2 \sqrt{2(1+\pi) r s}\right)\left(\alpha \sqrt{\frac{d}{n}}+\mu+k\right) \\
& \quad+2 \beta \sqrt{(1+\pi) s} \sqrt{\frac{\log d}{n}},
\end{aligned}
$$

where the second inequality is due to (14), and the last inequality is due to Lemma 8. Using the inequality $(a+b)^2 \leq 2a^2+2b^2$, we can conclude that with probability at least $\min\left\{1-2 \exp\left(-C_1 d\right), 1-2 \exp\left(-C_2 \log d\right)\right\}$,

$$
\left\|\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_\star\right\|_{\mathrm{F}}^2 \leq c_{1,4} \frac{r d}{n}+c_{2,4} \frac{s \log d}{n}+c_3 \frac{r s d}{n}+c_4 r s+c_5 r.
$$

■

D. Proof of Theorem 5

Proof: We have

$$\begin{aligned}\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2 &\le \left(\left\|\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{L}_\star\right\|_{\mathrm{F}}+\|\boldsymbol{\Psi}_t-\boldsymbol{\Psi}_\star\|_{\mathrm{F}}\right)^2\\ &\le 2\left\|\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{L}_\star\right\|_{\mathrm{F}}^2+2\|\boldsymbol{\Psi}_t-\boldsymbol{\Psi}_\star\|_{\mathrm{F}}^2\\ &\le 2\underbrace{\left\|\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{L}_\star\right\|_{\mathrm{F}}^2}_{K_1}+2(\pi+1)s\underbrace{\|\boldsymbol{\Psi}_t-\boldsymbol{\Psi}_\star\|_\infty^2}_{K_2},\end{aligned}$$

where the last inequality uses the fact that $\boldsymbol{\Psi}_t-\boldsymbol{\Psi}_\star$ has nonzero elements at most $(1+\pi)\,s$.

For K_1, use the inequality (21) to obtain

$$\begin{aligned}K_1 &\le (2+\delta)^2\left\|(\boldsymbol{B}_t\boldsymbol{Q}_t-\boldsymbol{B}_\star)\boldsymbol{\Lambda}_\star^{1/2}\right\|_{\mathrm{F}}^2\\ &= (2+\delta)^2\operatorname{dist}^2(\boldsymbol{B}_t,\boldsymbol{B}_\star).\end{aligned}$$

For K_2, we have

$$\begin{aligned}K_2 &\le \left(2\left\|\boldsymbol{B}_t\boldsymbol{B}_t^\top-\boldsymbol{L}_\star\right\|_\infty+2\|\boldsymbol{N}\|_\infty\right)^2\\ &\le 8(2+\delta)^2\operatorname{dist}^2(\boldsymbol{B}_t,\boldsymbol{B}_\star)+8\|\boldsymbol{N}\|_\infty^2,\end{aligned}$$

where the first inequality is due to (11) in Lemma 13, and the last inequality uses (21) and $(a+b)^2\le 2a^2+2b^2$. Taking collectively the bounds for K_1 and K_2, we have

$$\begin{aligned}\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2 &\le 2(2+\delta)^2(1+8(\pi+1)s)\operatorname{dist}^2(\boldsymbol{B}_t,\boldsymbol{B}_\star)\\ &\quad+16(\pi+1)s\|\boldsymbol{N}\|_\infty^2.\\ &\le 2(2+\delta)^2(1+8(\pi+1)s)\rho^t\sigma_r^2(\boldsymbol{L}_\star)\\ &\quad+2(2+\delta)^2(1+8(\pi+1)s)\frac{\omega}{1-\rho}\\ &\quad+16(\pi+1)s\|\boldsymbol{N}\|_\infty^2,\end{aligned}\tag{26}$$

where the last inequality is due to Proposition 16 with $\rho=2(1-10\eta)^2$ and $\omega=\frac{4\eta^2 s\left(1+\sqrt{\pi+1}\right)^2}{(1-\delta)^2}\|\boldsymbol{N}\|_\infty^2+\frac{2\eta^2 r}{(1-\delta)^2}\|\boldsymbol{N}\|_2^2$. Based on Lemma 8, for any $t=1,2,\ldots$, we have, with probability at least $\min\{1-2\exp(-C_1 d),1-2\exp(-C_2\log d)\}$,

$$\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}^2\le c_{1,5}\frac{rd}{n}+c_{2,5}\frac{s\log d}{n}+c_6\rho^t\sigma_r^2(\boldsymbol{L}_\star).$$

■

E. Proof of Corollary 6

Proof: From the assumption and (8), we have

$$\|\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star\|_2\le\sqrt{c_1\frac{rd}{n}+c_2\frac{s\log d}{n}}\le\frac{\sigma_d(\boldsymbol{\Sigma}_\star)}{c_0}.$$

From Weyl's inequality, we further have

$$\begin{aligned}\sigma_d(\boldsymbol{\Sigma}_T) &\ge \sigma_d(\boldsymbol{\Sigma}_\star)-\sigma_d(\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star)\\ &\ge \sigma_d(\boldsymbol{\Sigma}_\star)-\|\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star\|_2\ge\frac{c_0-1}{c_0}\sigma_d(\boldsymbol{\Sigma}_\star).\end{aligned}$$

Due to the fact that $\|\boldsymbol{AB}\|_{\mathrm{F}}\le\|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_{\mathrm{F}}$, we have

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_T^{-1}-\boldsymbol{\Sigma}_\star^{-1}\right\|_{\mathrm{F}} &= \left\|\boldsymbol{\Sigma}_T^{-1}(\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1}\right\|_{\mathrm{F}}\\ &\le \left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2\|\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}\left\|\boldsymbol{\Sigma}_\star^{-1}\right\|_2\\ &\le \frac{1}{\sigma_d(\boldsymbol{\Sigma}_T)}\|\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}\frac{1}{\sigma_d(\boldsymbol{\Sigma}_\star)}\\ &\le \frac{c_0}{c_0-1}\cdot\frac{1}{\sigma_d^2(\boldsymbol{\Sigma}_\star)}\|\boldsymbol{\Sigma}_T-\boldsymbol{\Sigma}_\star\|_{\mathrm{F}}\\ &= \sqrt{c_{1,6}\frac{rd}{n}+c_{2,6}\frac{s\log d}{n}}.\end{aligned}$$

■

F. Proof of Theorem 7

Proof: It can be proved similarly using the method in Section D. ■

REFERENCES

[1] S. Zou and Z. Zhao, "Large covariance matrix estimation based on factor models via nonconvex optimization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, pp. 9656–9660, 2024.

[2] P. M. Bentler and P. Dudgeon, "Covariance structure analysis: Statistical practice, theory, and directions," *Annu. Rev. Psychol*, vol. 47, no. 1, pp. 563–592, 1996.

[3] O. Ledoit and M. Wolf, "The power of (non-)linear shrinking: A review and guide to covariance matrix estimation," *J. Finan. Econom.*, vol. 20, no. 1, pp. 187–218, 2022.

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

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

[6] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1681–1695, 2019.

[7] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, 2005.

[8] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE.*, vol. 57, no. 8, pp. 1408–1418, 1969.

[9] S. U. Pillai, *Array Signal Processing*. Springer, 2012.

[10] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, no. 2, pp. 295–327, 2001.

[12] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Ann. Statist.*, vol. 36, no. 6, 2008.

[13] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate via majorization-minimization," *IEEE Trans. Signal Process.*, 2023.

[14] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Ann. Statist.*, vol. 39, no. 2, 2011.

[15] A. P. Shikhaliev, L. C. Potter, and Y. Chi, "Low-rank structured covariance matrix estimation," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 700–704, 2019.

[16] T. T. Cai, C.-H. Zhang, and H. H. Zhou, "Optimal rates of convergence for covariance matrix estimation," *Ann. Statist*, vol. 38, pp. 2118–2144, 2010.

[17] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, no. 9, pp. 963–974, 1982.

[18] T. T. Cai, Z. Ren, and H. H. Zhou, "Optimal rates of convergence for estimating Toeplitz covariance matrices," *Probab. Theory Relat. Fields.*, vol. 156, no. 1-2, pp. 101–143, 2013.

[19] G. Camba-Mendez and G. Kapetanios, "Testing the rank of the hankel covariance matrix: a statistical approach," *IEEE Trans. Autom. Control.*, vol. 46, no. 2, pp. 331–336, 2001.

[20] Y. Feng, D. P. Palomar *et al.*, "A signal processing perspective on financial engineering," *Found. Trends® Signal Process.*, vol. 9, no. 1–2, pp. 1–231, 2016.

[21] S. C. Ludvigson and S. Ng, "The empirical risk–return relation: A factor analysis approach," *J. Financ. Econ.*, vol. 83, no. 1, pp. 171–222, 2007.

[22] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, p. e161, 2007.

[23] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[24] R. Roy and T. Kailath, "ESPRIT-Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[25] L. Van Der Maaten, E. O. Postma, H. J. Van Den Herik *et al.*, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, vol. 10, no. 66-71, p. 13, 2009.

[26] S. A. Ross, "The arbitrage theory of capital asset pricing," *J. Econ. Theory*, vol. 13, no. 3, pp. 341–360, 1976.

[27] R. Roll and S. A. Ross, "An empirical investigation of the arbitrage pricing theory," *The Journal of Finance*, vol. 35, no. 5, pp. 1073–1103, 1980.

[28] J. Bai, "Inferential theory for factor models of large dimensions," *Econometrica*, vol. 71, no. 1, pp. 135–171, 2003.

[29] D. Passemier and J.-F. Yao, "Variance estimation and goodness-of-fit test in a high-dimensional strict factor model," *Submitted to Statist. Sinica, arXiv*, vol. 1308, 2013.

[30] J. Fan, Y. Fan, and J. Lv, "High dimensional covariance matrix estimation using a factor model," *J. Econom.*, vol. 147, no. 1, pp. 186–197, 2008.

[31] B. Liao, S.-C. Chan, L. Huang, and C. Guo, "Iterative methods for subspace and DOA estimation in nonuniform noise," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3008–3020, 2016.

[32] G. Chamberlain and M. Rothschild, "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica: J. Econom.*, pp. 1281–1304, 1983.

[33] J. Boivin and S. Ng, "Are more data always better for factor analysis?" *J. Econom.*, vol. 132, no. 1, pp. 169–194, 2006.

[34] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 104, no. 485, pp. 177–186, 2009.

[35] M. Agrawal and S. Prasad, "A modified likelihood function approach to DOA estimation in the presence of unknown spatially correlated Gaussian noise using a uniform linear array," *IEEE Trans. Signal Process.*, vol. 48, no. 10, pp. 2743–2749, 2002.

[36] M. Malek-Mohammadi, M. Jansson, A. Owrang, A. Koochakzadeh, and M. Babaie-Zadeh, "DOA estimation in partially correlated noise using low-rank/sparse matrix decomposition," in *Proc. IEEE Workshop Sensor Array Multichannel Signal Process.*, 2014, pp. 373–376.

[37] S. A. Vorobyov, A. B. Gershman, and K. M. Wong, "Maximum likelihood direction-of-arrival estimation in unknown noise fields using sparse sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 34–43, 2004.

[38] J. Bai and S. Ng, "Determining the number of factors in approximate factor models," *Econometrica*, vol. 70, no. 1, pp. 191–221, 2002.

[39] X. Luo, "High dimensional low rank and sparse covariance matrix estimation via convex minimization," *arXiv preprint arXiv:1111.1133*, vol. 199, 2011.

[40] J. Fan, Y. Liao, and M. Mincheva, "High dimensional covariance matrix estimation in approximate factor models," *Annals of statistics*, vol. 39, no. 6, p. 3320, 2011.

[41] ——, "Large covariance estimation by thresholding principal orthogonal complements," *J. R. Stat. Soc. B (Stat. Methodol.)*, vol. 75, no. 4, 2013.

[42] X. Luo, "Recovering model structures from large low rank and sparse covariance matrix estimation," *arXiv preprint arXiv:1111.1133*, 2013.

[43] M. Farnè and A. Montanari, "A large covariance matrix estimator under intermediate spikiness regimes," *J. Multivariate Anal.*, vol. 176, p. 104577, 2020.

[44] ——, "Large factor model estimation by nuclear norm plus ℓ_1 norm penalization," *J. Multivariate Anal.*, vol. 199, p. 105244, 2024.

[45] A. Agarwal, S. Negahban, and M. J. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," *Ann. Statist.*, vol. 40, no. 2, 2012.

[46] E. Yang and P. K. Ravikumar, "Dirty statistical models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[47] K. G. Jöreskog, "Some contributions to maximum likelihood factor analysis," *Psychometrika*, vol. 32, no. 4, pp. 443–482, 1967.

[48] D. B. Rubin and D. T. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, vol. 47, pp. 69–76, 1982.

[49] J. Bai and K. Li, "Statistical analysis of factor models of high dimension," *Ann. Statist.*, vol. 40, no. 1, p. 436, 2012.

[50] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1783–1795, 1990.

[51] ——, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, 2002.

[52] P. Stoica and P. Babu, "Low-rank covariance matrix estimation for factor analysis in anisotropic noise: Application to array processing and portfolio selection," *IEEE Trans. Signal Process.*, vol. 71, pp. 1699–1711, 2023.

[53] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[54] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Ann. Statist.*, vol. 37, no. 5B, 2009.

[55] T. W. Anderson, T. W. Anderson, T. W. Anderson, T. W. Anderson, and E.-U. Mathématicien, *An introduction to multivariate statistical analysis*. Wiley New York, 1958.

[56] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[57] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *in Adv. Neural Inf. Process. Syst.*, vol. 23, 2010.

[58] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *J. Struct. Biol.*, vol. 181, no. 2, pp. 116–127, 2013.

[59] P. Netrapalli, N. U N, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust PCA," in *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[60] A. Hippert-Ferrer, M. N. El Korso, A. Breloy, and G. Ginolhac, "Robust low-rank covariance matrix estimation with a general pattern of missing values," *Signal Process.*, vol. 195, p. 108460, 2022.

[61] J. Tanner and K. Wei, "Low rank matrix completion by alternating steepest descent methods," *Appl. Comput. Harmon. A.*, vol. 40, no. 2, pp. 417–429, 2016.

[62] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[63] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent." *J. Mach. Learn. Res.*, vol. 22, pp. 150–1, 2021.

[64] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1665–1697, 2012.

[65] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *Ann. Stat.*, pp. 1935–1967, 2012.

[66] P. Xu, J. Ma, and Q. Gu, "Speeding up latent variable gaussian graphical model estimation via nonconvex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[67] Z. Meng, B. Eriksson, and A. Hero, "Learning latent variable gaussian graphical models," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2014, pp. 1269–1277.

[68] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *Bernoulli*, pp. 1029–1058, 2014.

[69] A. R. Zhang, T. T. Cai, and Y. Wu, "Heteroskedastic PCA: Algorithm, optimality, and applications," *Ann. Statist.*, vol. 50, no. 1, pp. 53–80, 2022.

[70] M. W. McCracken and S. Ng, "FRED-MD: A monthly database for macroeconomic research," *J. Bus. Econ. Stat.*, vol. 34, no. 4, pp. 574–589, 2016.

[71] J. Fan, J. Guo, and S. Zheng, "Estimating number of factors by adjusted eigenvalues thresholding," *J. Am. Stat. Assoc.*, vol. 117, no. 538, pp. 852–861, 2022.

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

[73] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," 2009.

[74] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via Procrustes flow," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 964–973.

[75] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2017, pp. 1233–1242.