High-Dimensional Covariance Estimation[*]

Varlam Kutateladze[†] and Ekaterina Seregina[‡]

Preliminary and incomplete, please do not distribute.

June 8, 2021

Abstract

Covariance matrix estimates are required in a wide range of applied problems in multivariate data analysis, including portfolio and risk management in finance, factor models and testing in economics, and graphical models and classification in machine learning. In modern applications, where often the model dimensionality is comparable or even larger than the sample size, the classical sample covariance estimator lacks desirable properties, such as consistency, and suffers from eigenvalue spreading. In recent years, improved estimators have been proposed based on the idea of regularization. Specifically, such estimators, known as rotation-equivariant estimators, shrink the sample eigenvalues, while keeping the eigenvectors of the sample covariance estimator. In high dimensions, however, the sample eigenvectors will generally be strongly inconsistent, rendering eigenvalue shrinkage estimators suboptimal. We consider an estimator that goes beyond mere eigenvalue shrinkage and aims at precise estimation of eigenvectors in sparse settings, without requiring eigenvalues to diverge. The rate of convergence is provided in terms of spectral norm and it achieves the optimal rate under reasonable assumptions. We also provide a numerical simulation demonstrating the superior performance of the proposed estimator as compared to the competition.

JEL Classification: C020

Keywords: Sparse recovery, Rotation equivariance, Random matrix theory, Large-dimensional asymptotics, Principal components

[*] This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Declarations of interest: none.

[†] Department of Economics, University of California, Riverside. Email: varlam.kutateladze@email.ucr.edu.

[‡] Department of Economics, Colby College, Waterville. Email: eseregin@colby.edu.

1 Introduction

Covariance matrix estimation is fundamental to multivariate statistical analysis. In statistics and machine learning, some of its applications include graphical modeling, clustering, classification by linear or quadratic discriminant analysis and dimensionality reduction by PCA. In finance, covariance estimates play a central role in portfolio optimization and risk management. In economics, its uses include Kalman filtering, factor analysis, hypothesis testing, GLS and GMM. Covariance estimation is also key in many application in signal processing, bioinformatics and several other fields.

With rapidly increasing availability of data, the analysis of covariance matrices in the low-dimensional (or classical) regime quickly becomes obsolete. This paper considers a large-dimensional framework, where the number of variables p is comparable or even larger than the sample size n. In such settings, the sample covariance estimator S loses desirable properties and its classical theoretical foundations break down. For instance, if $p > n$, S is not full rank, so the inverse does not exist. Even when S is invertible, its inverse is highly biased for the theoretical inverse when p and n are comparable. In portfolio optimization this may lead to imprecise and highly volatile weights. Several other major issues such as eigenvalue spreading and eigenvector inconsistency are considered in this manuscript.

On the other hand, while a generic high-dimensional analysis is complicated, sparsity may be a reasonable simplifying assumption to resort to. Furthermore, in applications ranging from genomics to finance, sparsity-inducing approach may be preferred to unrestricted estimation. For example, assets weights in eigenportfolio methodology are proportional to the corresponding eigenvector entries; hence, sparse estimation of an eigenvector leads to more parsimonious allocations with less associated transaction costs. In addition, sparse estimates are easier to interpret.

The literature on covariance estimation proposed some remedies to tackle these challenges. One approach is to shrink a high-variance sample estimator to some structured matrix which may be highly biased and thus produce a better estimator which would achieve optimal bias-variance trade-off. Another approach is to assume a low-dimensional structure in the data as in factor models and consider the implied covariance. Statisticians and mathematicians have also looked into estimation based on the behavior of random matrices, where the analysis is primarily driven by theoretical advancements in random matrix theory.

This paper proposes an estimator that is suitable for the high-dimensional regime, analyzes its theoretical properties and provides a numerical experiment comparing it with the alternative methods. The manuscript is structured as follows. Section 2 describes some of the phenomena and challenges that arise in settings where the number of dimensions is large, reviews some existing approaches and, in this context, motivates the proposed estimator. Section 3 described the model setup, introduces the estimator and point out the similarities with the factor model framework. Section 4 considers a numerical simulation and Section 5 concludes and mentions possible extensions.

Notation. For a vector $\mathrm{v} \in \mathbb{R}^d$, we write its i-th element as v_i. The corresponding ℓ_p norm is $\|\mathrm{v}\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$. For a matrix $A \in \mathbb{R}^{m \times d}$, we write its (i,j)-th entry as $\{A\}_{ij}$ and denote its i-th row (transposed) and j-th column as column vectors $A_{i\cdot}$ and $A_{\cdot j}$ respectively. Its singular values are $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_q(A)$, where $q = \min(m,d)$. The spectral norm is a matrix operator norm induced by the Euclidean norm, $\|A\|_2 = \max_{\mathrm{v} \neq 0} \frac{\|A\mathrm{v}\|_2}{\|\mathrm{v}\|_2} = \sigma_1(A)$. The max and Frobenius norms are given as $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ and $\|A\|_F = \sqrt{tr(A'A)} = \sqrt{\sum_{i=1}^q \sigma_i^2(A)}$ respectively. Finally, for a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a sequence of real nonnegative numbers $\{a_n\}_{n=1}^\infty$, denote $X_n = O_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0, \exists M, N > 0$ such that $\forall n > N,\ \mathbb{P}(|X_n/a_n| \geq M) < \epsilon$; and denote $X_n = o_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0,\ \lim_{n \to \infty} \mathbb{P}(|X_n/a_n| \geq \epsilon) = 0$. Finally, let $\mathbb{1}(\cdot)$ be an indicator function and I_d is a $d \times d$ identity matrix.

2 Background & Related Literature

We observe a data matrix $X \in \mathbb{R}^{n \times p}$, where n and p are the number of observations and the number of variables respectively. Denote the population covariance matrix by $\Sigma \in \mathbb{R}^{p \times p}$ and write its eigendecomposition as

$$\Sigma = ULU' = \sum_{j=1}^{p} \ell_j \mathrm{u}_j \mathrm{u}_j',$$

where $U = [\mathrm{u}_1 \ \cdots \ \mathrm{u}_p]$ is an orthonormal matrix of eigenvectors, and $L = diag(\ell_1, \ldots, \ell_p)$ is a diagonal matrix of eigenvalues with $\ell_1 \geq \ldots \geq \ell_p$. The sample covariance estimator is given as $S := \frac{1}{n} X'X$ (demeaned data) and we write its eigendecomposition as

$$S = V\Lambda V' = \sum_{j=1}^{p} \lambda_j \mathrm{v}_j \mathrm{v}_j',$$

with $V = [\mathrm{v}_1 \ \cdots \ \mathrm{v}_p]$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$.

2.1 Sample Eigenvalues

It is well-known that S is unbiased and consistent in a classical regime, i.e. when p is fixed and n diverges. Furthermore, it is generally invertible and has an asymptotically normal spectral distribution centered around the true value (Anderson [1963]), $\sqrt{n}(\lambda_i - \ell_i) \xrightarrow{d} \mathcal{N}(0, 2\ell_i^2), \ j \leq p$.

However, it was observed that many of the desirable properties cease to hold once p also grows, specifically when $\gamma := \lim \frac{p}{n} \in (0, \infty)$. In fact, consistent estimation of the entire spectrum becomes a lot more problematic as both sample eigenvalues and eigenvectors tend to concentrate beyond their true destination.

Specifically, Marčenko and Pastur [1967] derived the empirical distribution of sample eigenvalues, which became known as Marčenko-Pastur distribution. In its simple formulation when $\Sigma = I_p$, the empirical distribution of sample eigenvalues of a random matrix $F_p(x) := \frac{1}{p}\#\{\lambda_j \leq x\}$ approaches a limiting distribution for which the density is given as

$$f^{MP}(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi x\gamma}, \quad \lambda_- = (1-\sqrt{\gamma})^2, \quad \lambda_+ = (1+\sqrt{\gamma})^2,$$

where λ_-, λ_+ are the lower and upper bounds of the support. This result illustrates how sample eigenvalues spread out away from their true values, in this case $\ell_i = 1, \ \forall i$. Moreover, there is a positive bias in the largest sample eigenvalues and a negative bias in the smallest eigenvalues. Furthermore, the magnitude of the bias increases with p.

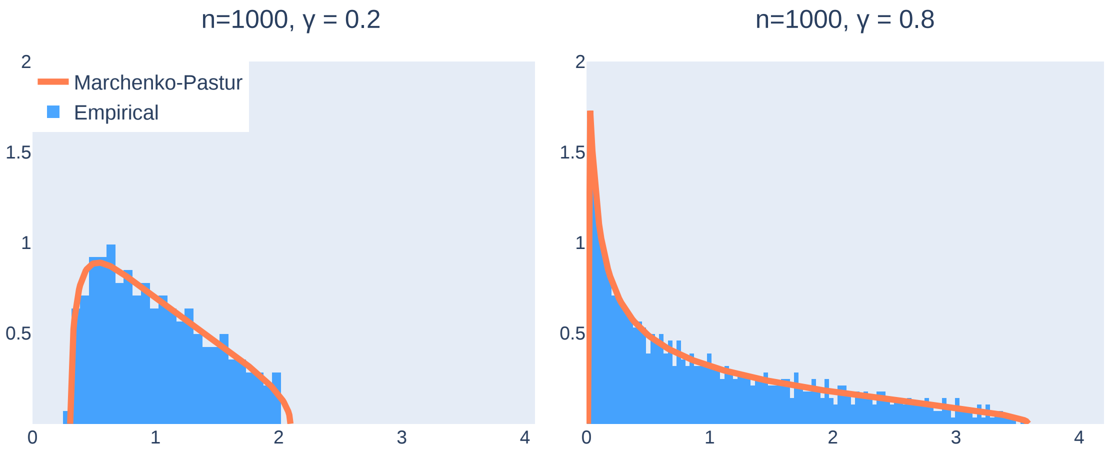

Figure 1: Marčenko-Pastur and empirical sample eigenvalue distributions

For empirical distribution both panels have $n = 1000$, while $\gamma := p/n$ is different. *Left*: $\gamma = .2$; *Right*: $\gamma = .8$.

Figure 1 plots the theoretical density on top of the empirical sample eigendistribution for n datapoints sampled from $\mathcal{N}_p(0, I_p)$ for different values of $\gamma := p/n$. It demonstrates two phenomena. First, the sample eigenvalues are spread out asymmetrically around the true value of 1. Second, the smallest and largest eigenvalues concentrate around λ_- and λ_+ respectively. In fact, Bai-Yin's law (Bai and Yin [1993]) states that for matrices with bounded fourth moments the extreme sample eigenvalues land almost surely on these edges, i.e. $\lambda_p \overset{a.s.}{\to} \lambda_-$ and $\lambda_1 \overset{a.s.}{\to} \lambda_+$. Excellent treatment is given in Bai and Silverstein [2010].

This paper focuses on the case when a few population eigenvalues are larger than the bulk, in other words top eigenvalues are spiked (Johnstone [2001]),

$$\Sigma = diag(\ell_1, \ldots, \ell_r, 1, \ldots, 1),\ \ell_r > 1.$$

The convergence of the sample eigenvalues in this case depends on the magnitude of the true spikes in comparison to the so-called BBP transition point $\lambda_+^{1/2}$, named after its discoverers Baik et al. [2005]. Specifically, we have for $j \leq r$,

$$\lambda_j \overset{a.s.}{\to} \begin{cases} \lambda_+, & \ell_j < \lambda_+^{1/2}, \\ \ell_j + \gamma \frac{\ell_j}{\ell_j - 1}, & \ell_j > \lambda_+^{1/2}, \end{cases}$$

as $n, p \to \infty$. That is, there is an upward bias in leading sample eigenvalues and the amount of bias is asymptotically known. Baik and Silverstein [2006] establish the almost sure limits of the eigenvalues of large sample covariance matrices in a spiked population model framework. Moreover, the exact asymptotic distribution of the largest and smallest eigenvalues is also known (Tracy and Widom [1996]).

Hence, if the true spikes are not large enough, the sample eigendistribution will follow the Marčenko-Pastur distribution. In the opposite case, the spiked sample eigenvalues will overshoot the true counterparts and lie above the Marčenko-Pastur sea. In general this knowledge can be used as a heuristic for inferring the number of principal components or factors.

In view of the above phenomena, statisticians have proposed to construct rotationally invariant estimators (RIE), which would correct the sample eigenvalues while assuming the sample and true eigenvectors coincide. This includes many popular estimators, e.g. linear shrinkage of the sample covariance with a structured matrix (typically, an identity) proposed in Ledoit and Wolf [2004] or nonlinear extensions as in Ledoit and Wolf [2012], which in essence seek to pull upward and downward biased sample estimates towards the center. Another approach is to set all eigenvalues inside the Marčenko-Pastur sea to some constant, as these are deemed as noise, while keeping the spikes unaltered; this strategy is known as eigenvalue clipping (Bouchaud and Potters [2009]). Donoho et al. [2018] do not address eigenvector inconsistency but partially address fix this issue by proposing an RIE that accounts for the non-vanishing angle between population and sample eigenvectors. They find a univariate function ν that when applied to eigenvalues would optimally (for a given loss) shrink it such that eigenvector estimation inaccuracy is taken into consideration.

2.2 Sample Eigenvectors

However, the assumption that sample and population eigenvectors coincide in high dimensions is unrealistic as sample eigenvectors are generally also inconsistent in high dimensions. This motivates us to develop an estimator that primarily aims at accurate eigenvector estimation. We seek to carefully characterize the assumptions this will require and the trade-offs between different sets of assumptions. Before that, we briefly review some of the related work without aiming to be exhaustive.

As described in the previous section, there has been a lot of work done examining the features of sample eigenvalues. Bai et al. [2007] points out that this may partly be explained by the quantum mechanics origins of the random matrix theory, where the sample eigenvalues are associated with energy levels of particles. Many applications, however, require precise estimates of eigenvectors and the research in this direction is gradually being recognized.

Johnstone and Lu [2004] propose an (adaptive) sparse PCA for settings when $p, n \to \infty$ in a single factor model framework. They show that plain PCA leads to consistent estimates if and only if $p/n \to 0$, however it is possible to recover consistency even when $p \gg n$ if some preselection of variables, possibly in an alternative sparse basis, is made in advance. Their Theorems 1,2 and 3 characterize the inconsistency of PCA when $\lim p/n = \gamma > 0$ in terms of an angle between the true leading eigenvector and its estimate. Theorem 5 suggest a solution for cases when the true principal eigenvector satisfies ℓ_q-ball sparsity assumption.

In a closely related work, Johnstone and Lu [2009] provide a similar characterization of inconsistency in Theorem 1 but in terms of the normalized inner product between the true and estimated principal eigenvectors, while Theorem 2 proves that consistency is recovered as long as the PCA is performed after the proposed variable selection algorithm.

Paul [2007] characterize the asymptotic behavior of sample eigenvectors and its dependence on the eigenvalue phase transition. Specifically, under mild conditions on a spiked covariance model, as $p/n \to \gamma \in (0,1)$ we have

$$\langle \mathrm{v}_j, \mathrm{u}_j \rangle^2 \overset{a.s.}{\to} \begin{cases} 0, & \ell_j < \lambda_+^{1/2}, \\ \frac{1-\gamma/(\ell_j-1)^2}{1+\gamma/(\ell_j-1)}, & \ell_j > \lambda_+^{1/2}. \end{cases}$$

That is, the sample eigenvector v_j is asymptotically orthogonal to the true vector u_j when the corresponding eigenvalue is small. On the other hand, consistent estimation of eigenvectors with sufficiently strong signals requires $\gamma \to 0,$, i.e. n grows faster than p. The conventional PCA becomes confused in the presence of large number of variables. Sparsity, either in the original or some transformed domain, becomes crucial for consistent estimation of principal component directions.

Shen et al. [2016] further investigate consistency and asymptotic behavior of sample eigenvalues and eigenvectors in a more general multiple-component spike covariance framework with r spikes, $\ell_1 > \ldots > \ell_r \gg \ell_{r+1} \to \ldots \to \ell_p \to 1$. They also consider a more general asymptotic framework with $\frac{p}{n\ell_j} \to c_j > 0,\ j \leq r$, where $0 < c_1 < \ldots < c_r < \infty$; allowing ℓ_j to potentially diverge turns out to be crucial. Their Theorem 3 states that

$$\begin{cases} \frac{\lambda_j}{\ell_j} \overset{a.s.}{\to} 1 + c_j, \quad 1 \leq j \leq r, \\ \frac{n\lambda_j}{p\ell_j} \overset{a.s.}{\to} 1, \quad r+1 \leq j \leq n \wedge p, \end{cases} \tag{1}$$

and

$$\begin{cases} |\langle v_j, u_j \rangle| \overset{a.s.}{\to} (1+c_j)^{-1/2}, \quad 1 \leq j \leq r, \\ |\langle v_j, u_j \rangle| \overset{a.s.}{\to} O_{a.s.}\big((n/p)^{1/2}\big), \quad r+1 \leq j \leq n \wedge p, \\ \angle\left(v_j, \text{span}(u_k : k = r+1, \ldots, p)\right) \overset{a.s.}{\to} 0, \quad r+1 \leq j \leq n \wedge p. \end{cases} \tag{2}$$

Equation (1) formalizes the idea that sample eigenvalues with stronger signals will be less biased, but still almost surely biased when $c_j \neq 0$. Notice that there is no bias if the eigenvalues grow linearly with the dimension. Equation (2) reveals that leading sample eigenvectors lie in a cone along the true eigendirections as long as the ratio of the dimension to the product of the sample size and the spike size, $\frac{p}{n\lambda_j} \to c_j > 0,\ j \leq r$. This shows that sample eigenvectors might still be consistent in the high-dimensional regime when $\gamma > 0$ as long as their corresponding eigenvalues are large. Intuitively, a strong signal helps identify the direction of most variation.

Observe that $\frac{p}{n\ell_j} \to c_j > 0$ can contain three cases: (i) $p, n, \ell_j \to \infty$, (ii) $p, \ell_j \to \infty$, while $n < \infty$, (iii) $p, n \to \infty$ and $\ell_j < \infty$. The result stated in equation (2) refers to the first case, Shen et al. [2016] also covers the second case. We focus on the third case, where leading eigenvalues are bounded.

A natural extension to the work of Paul [2007] and Shen et al. [2016] is the study by Wang and Fan [2017] who analyze the asymptotic distributions of the sample eigenstructure and derive the precise rates of convergence in a similar setup with a high-dimensional spiked covariance and $p, n, \ell_j \to \infty$ with $\frac{p}{n\ell_j} < \infty,\ j \leq r$. Although this comes at the cost of a sub-Gaussianity assumption on the data. In particular, they establish that the normalized spiked part of the sample eigenvector converges to a vector of ones. Fan et al. [2013] also consider a diverging eigenvalue setup.

Instead of assuming increasing eigenvalues, a simple alternative approach which permits efficient estimation in high dimensions is sparsity. Bickel and Levina [2008] propose regularizing a large covariance matrix by hard thresholding, which leads to a consistent (in the operator norm) estimator as long as the true covariance matrix is sparse. However, imposing sparsity on a high-dimensional covariance directly may be unjustified in certain applications, e.g. in portfolio theory covariance of asset returns is not sparse. Fan et al. [2013] consider conditional sparsity, i.e. sparsity after estimating and accounting for the factor structure. To distinguish the signal and noise components, they assume diverging (with p) signal eigenvalues. This assumption may be "misleading in many economic and financial applications", as pointed out some researchers (Fan et al. [2013], discussion on the paper).

On the other hand, the eigenvectors themselves are often sparse in many high-dimensional applications or may be required to be sparse in certain scenarios, e.g. in capital allocation problems. This also leads to better interpretability since eigenvectors are only linear combinations of a subset of variables. Most importantly, sparsity can help even in cases when $\gamma > 0$ and the signal is bounded, $\ell_j = O(1)$. A similar idea is considered in Amini and Wainwright [2009], however they focus on sparse eigenvector support recovery.

Besides, precise estimation of eigenvectors in high dimensions has its own benefit. For example, the top eigenvectors of a covariance identify the directions of most variation, while the bottom eigenvectors of a graph's Laplacian provide insights into its cluster structure.

3 Model

Given a $n \times p$ data matrix X of p i.i.d. mean-zero variables with the population covariance

$$\Sigma = \mathbb{E}(X'X) = \sum_{i=1}^{r} \ell_i \mathrm{u}_i \mathrm{u}_i' + \sum_{i=r+1}^{p} \ell_i \mathrm{u}_i \mathrm{u}_i', \tag{3}$$

and the sample covariance estimator

$$S = \frac{1}{n} X'X = \sum_{i=1}^{r} \lambda_i \mathrm{v}_i \mathrm{v}_i' + \sum_{i=r+1}^{p} \lambda_i \mathrm{v}_i \mathrm{v}_i',$$

where r is the number of signal eigenvalues (assumed to be known and fixed), our primary goal is accurate estimation of Σ in a high-dimensional setting under a spiked covariance framework (Johnstone [2001]). Throughout this paper we measure the estimation error in terms of spectral (operator) norm $\|\cdot\|$. By Weyl's and Davis-Kahan Theorems (see Appendix A.4) the $\|\widehat{\Sigma}-\Sigma\| \to 0$ implies the convergence of the corresponding eigenvalues and eigenvectors as well as the convergence of PCA loadings.

Assumption 1 (Spiked covariance)**.** *There are $r \ll p \wedge n$ spikes in eigenvalues $\ell_1 > \ldots > \ell_r > 1$, independent of p and n, with $\Delta := \ell_r - \ell_{r+1} \gg 0$. All spiked eigenvalues are distinct.*

Remark. In particular, while we need not have diverging signals, the eigengap $\ell_r - \ell_{r+1}$ should be large enough for identification purposes. This also inherently relates to the eigenvector instability demonstrated in the following example.

Example. (**Wainwright [2019]**) Consider a perturbation of a diagonal A by another diagonal matrix ϵP,

$$A_\epsilon = A + \epsilon P = \begin{pmatrix} 1 & 0 \\ 0 & 1.01 \end{pmatrix} + \epsilon \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Clearly, the eigenvalues of unperturbed A are $\{1,\, 1.01\}$ and the eigenvalues of the perturbed A_n are

$$\left\{ \frac{1}{2}(2.01 + \sqrt{.0001 + 4\epsilon^2}),\ \frac{1}{2}(2.01 - \sqrt{.0001 + 4\epsilon^2}) \right\},$$

satisfying Weyl's theorem

$$\max_{i=1,2} |\ell_i(A) - \ell_i(A_\epsilon)| = \frac{1}{2}|.01 - \sqrt{.0001 + 4\epsilon^2}| \leq \|\epsilon P\|_2 = \epsilon,$$

and thus displaying resilience to small perturbations. On the other hand, the maximal eigenvector of A changes its direction substantially from $\mathrm{u}_1(A) = (0\ ,1)'$ to $\mathrm{u}_1(A_\epsilon) \approx (.53\ ,.85)'$, so that $\|\mathrm{u}_1(A) - \mathrm{u}_1(A_\epsilon)\|_2 \gg \epsilon$. The problem arises due to small eigengap, and hence a large enough eigengap is needed to ensure the stability.

A simple example of a spiked model that was widely considered in the literature is of the form,

$$\Sigma = \ell_1 \mathrm{u}_1 \mathrm{u}_1' + I_p,$$

with $\ell_1 > 0, \|\mathrm{u}_1\|_2 = 1$, where u_1 is a unique maximal eigenvector with eigenvalue $1+\ell_1$ and all other eigenvalues are 1. That is, we have a low-rank perturbation of a sparse matrix. Berthet and Rigollet [2013] examine the possibility of detection of the low-rank component and propose a minimax optimal test based on an eigenvalue statistic.

Spiked models are also inherently related to factor models considered in financial econometrics and we examine the similarity in Section 3.2. The differences mainly arise due to different assumptions placed on the behavior of the eigenstructure.

Assumption 2 (High-dimensional asymptotics)**.** *$n, p \to \infty$ and $\ell_j = O(1),\ j = 1, \ldots, p.$*

Remark. In particular, we need not have strong (pervasive) signals, so it is not necessary that $\ell_i = O(p),\ i \leq r$. In fact, the pervasiveness assumption (Fan et al. [2013]) can make consistent estimation impossible in terms of spectral norm, as discussed in the following example.

Example. Suppose we know the entire spectrum of Σ except for the first eigenvector, for which suppose we have a good estimator with $\|\mathrm{v}_1 - \mathrm{u}_1\| = O_p(n^{-1/2})$. Then we can construct a sample covariance S^* with this population information in its spectrum, then

$$\|S^* - \Sigma\| = \|\ell_1(\mathrm{v}_1\mathrm{v}_1' - \mathrm{u}_1\mathrm{u}_1')\| = \ell_1 O_p(\|\mathrm{v}_1 - \mathrm{u}_1\|) = O_p(\ell_1 n^{-1/2}),$$

which does not converge if ℓ_1 grows linearly in p and $n = O(p^2)$, so Σ could not be estimated consistently in terms of spectral norm in the presence of diverging spiked eigenvalues.

In accordance with Equation (3) we can decompose the true covariance into two parts,

$$\Sigma = \Sigma_s + \Sigma_e, \tag{4}$$

where the two matrices on the right-hand side represent signal and noise (error) components. In particular, one can view the above equation as low-rank plus sparse matrix structure. This idea is formalized in the following assumption.

Assumption 3 (Low-rank plus sparse). *In equation* (4)*, rank*$(\Sigma_s) = r$ *and* Σ_e *is* (*approximately*) *sparse with bounded eigenvalues. Moreover,* Σ_s *has a fixed number* r *sparse unit norm eigenvectors,* $\|\mathrm{u}_j\|_0 = s,\ \ j \leq r, \|\mathrm{u}_j\|_2 = 1,\ \forall j.$

Remark. The low-rank plus sparse structure has been thoroughly studied, see e.g. Wright et al. [2009] or Candès et al. [2011] on the possibility of identification of the two matrices, low-rank and sparse, only from the sum alone. This structure is also implied by approximate factor models (Chamberlain and Rothschild [1983]) where $\Sigma = \Lambda\Lambda' + \Omega$.

Remark. In general, the results throughout this paper can be adapted to cases with approximate sparsity. For example, if $\widetilde{\mathrm{u}}_1$ is an approximation of exactly s-sparse vector u_1 of Σ_s, we can instead analyze a slightly different perturbation, $S = \Sigma + E = \widetilde{\Sigma} + (E - \widetilde{\Sigma} + \Sigma) = \widetilde{\Sigma} + \widetilde{E}$, where $\widetilde{E} := E - \widetilde{\Sigma} + \Sigma$.

Remark. Notice that this formulation with sparse eigenvectors does not necessarily imply that Σ is sparse. A similar framework with sparse eigenstructure was analyzed by Amini and Wainwright [2009] with the focus on support recovery.

Notice that Σ_e is approximately sparse; we characterize sparsity as in Bickel and Levina [2008],

$$m := \max_{i \leq p} \sum_{j \leq p} |\{\Sigma_e\}_{i,j}|^q,$$

so that m stays bounded for some $0 \leq q < 1$, although for simplicity we will focus on the case with $q = 0$ corresponding to exact sparsity, $m = \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\{\Sigma_e\}_{i,j} \neq 0)$. The fact that eigenvalues are thus bounded can be seen from

$$\|\Sigma_e\| \leq \|\Sigma_e\|_1 \leq \max_{i \leq p} \sum_{j \leq p} |\{\Sigma_e\}_{i,j}|^q (\{\Sigma_e\}_{i,i} \{\Sigma_e\}_{j,j})^{(1-q)/2} = O(m),$$

when $\{\Sigma_e\}_{i,i}$ are bounded. This assumption is not very restrictive and corresponds to weak correlation between idiosyncratic components in factor models.

3.1 Estimator

In what follows we consider a simpler version of Equation (4) with $r = 1$, namely

$$\Sigma = \ell_1 \mathrm{u}_1 \mathrm{u}_1' + \Sigma_e, \tag{5}$$

where the single signal eigenvector is s-sparse, i.e. $\|\mathrm{u}_1\|_0 = s$, and Σ_e is approximately sparse in a sense that $\|\Sigma_e\|$ has bounded eigenvalues as $p \to \infty$. In other words, the covariance is a rank-1 perturbation of a sparse matrix.

As a first step, we seek to estimate the sparse eigenvectors of Σ_s. To recover the first eigenvector, one could proceed by explicitly imposing the sparsity restriction in a rank-one approximation problem by solving

$$\begin{gathered} \min_{\nu, \xi} \|X'X - \nu \xi \xi'\|_F^2 \\ \nu \geq 0, \ \|\xi\|_0 \leq s, \ \|\xi\|_2 = 1, \end{gathered} \tag{6}$$

which is equivalent to solving

$$\begin{gathered} \min_{\xi} \xi' X' X \xi \\ \|\xi\|_0 \leq s, \ \|\xi\|_2 = 1. \end{gathered} \tag{7}$$

Clearly, both are NP-hard problems due to the presence of $\| \cdot \|_0$-norm. Efficient convex relaxations with $\|\cdot\|_1$-norm substitution have been studied (d'Aspremont et al. [2007]) as well as alternative formulations imposing sparse structure (e.g. Jolliffe et al. [2003],Zou et al. [2006], Witten et al. [2009]). The theoretical underpinnings for the above algorithms and formulations are less developed.

Luckily, it is possible to directly approximate the solution via the principle of power iteration. Yuan and Zhang [2011] propose a truncated power method which tackles the optimization problem in Equation (7) and analyze its theoretical properties. This approach achieves an optimal bound (Cai and Zhou [2012]) and is guaranteed to converge under mild technical conditions. The goal is to recover a sparse eigenvector given a perturbation of an original matrix.

The truncated power iteration procedure is described in Algorithm 1. The only difference with a conventional power method is in the truncation step, which forces the $p - r$ smallest entries of a vector to zero in each iteration thus naturally inducing sparse estimates.

The following perturbation formulation is useful,

$$S = \Sigma + E, \tag{8}$$

where S is a sample covariance matrix, and $E := S - \Sigma$ is an error. The theorem of Yuan and Zhang [2011] is adapted in 1 and assures the recovery of s-sparse eigenvectors so long as the spectral norm of $\hat{s} \times \hat{s}$ principal submatrix of E, denoted as $\underline{\mathrm{E}}_{\hat{s}}$, is sufficiently small for some initial estimate of sparsity $\hat{s}$. Notice that this norm can be a substantially smaller than the norm of the full matrix E.

Theorem 1 (Sparse recovery). *Given Assumptions 1, 3 and the initial vector* $\widehat{\mathrm{v}}_1^{(0)}$ *with* $\|\widehat{\mathrm{v}}_1^{(0)}\|_0 \leq \widehat{s},\ \|\widehat{\mathrm{v}}_1^{(0)}\|_2 = 1,\ \widehat{s} \geq s,\ |\widehat{\mathrm{v}}_1^{(0)\prime}\mathrm{u}_1| - \delta \geq \theta$, *where* $0 < \theta < 1$ *and*

$$\delta := \frac{\sqrt{2}\|\underline{E}_{\hat{s}}\|}{\sqrt{\|\underline{E}_{\hat{s}}\|^2 + (\Delta - 2\|\underline{E}_{\hat{s}}\|)^2}},$$

we have

$$\sqrt{1 - |\widehat{\mathrm{v}}_1^{(t)\prime}\mathrm{u}_1|} \leq c_1^t \sqrt{1 - |\widehat{\mathrm{v}}_1^{(0)\prime}\mathrm{u}_1|} + \sqrt{10}\delta(1-c_1)^{-1}, \quad \forall t \geq 0, \tag{9}$$

where c_1 can be chosen to be less than 1.

Proof. See Yuan and Zhang [2011] Theorem 4. □

Since the convergence of the algorithm is guaranteed, let us denote $\lim_{t\to\infty} \widehat{\mathrm{v}}_1^{(t)} = \widehat{\mathrm{v}}_1$.

Corollary 1.1. *Given the assumptions of Theorem 1, we have*

$$\|\widehat{\mathrm{v}}_1 - \mathrm{u}_1\| = O_{\mathbb{P}}(\|\underline{E}_{\hat{s}}\|).$$

Corollary 1.2. *Given the assumptions of Theorem 1 and assuming entries in E are Gaussian iid and $\hat{s} = O(s)$, we have*

$$\|\widehat{\mathrm{v}}_1 - \mathrm{u}_1\| = O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right).$$

Theorem 1 bounds the angle between the t-th iteration of sparse eigenvector estimate $\widehat{\mathrm{v}}_1^{(t)}$ and its population counterpart u_1. Corollary 1.2 is an immediate consequence and states that the error depends on the norm of s-dimensional principal submatrix which could be much smaller than the norm of the entire perturbation implied by standard perturbation inequalities. Corollary 1.2 follows when entries of the perturbation are normally distributed; this is a standard result in random matrix theory. The arguments of the proof are based on eigenvector perturbation inequalities provided in the appendix. Initialization is also discussed.

Hence, this iterative approach can recover the leading sparse eigenvector even from noisy observations. The remaining sparse eigenvector estimates could be obtained greedily, i.e. for the second iteration one would optimize over an unexplained component $X'X - (\hat{\xi}'X'X\hat{\xi})\hat{\xi}\hat{\xi}'$, where $\hat{\xi}$ solves equation (7). This procedure is known as iterative deflation.

Suppose that we have obtained eigenvector estimates $\{\widehat{\mathrm{v}}\}_{i=1}^r$. The corresponding weight estimates for r top matrices can be estimated consistently by least squares under the standard assumptions with usual parametric rates, i.e. $|\widehat{\lambda}_j - \ell_j| = O(n^{-1/2})$. This completes the estimation of the signal part of the covariance matrix.

Once the signal component is estimated we turn to the error part Σ_e. Consistent with the assumption of conditionally sparse covariance, after removing the estimated low-rank part we threshold the remainder. In a general case one can obtain an estimate of the remainder as

$$\widehat{S}_e = S - \sum_{j=1}^{r} \widehat{\lambda}_j \widehat{\mathrm{v}}_j \widehat{\mathrm{v}}_j{}', \tag{10}$$

where we subtract the estimated low-rank component from the sample covariance. Next we apply entry-wise adaptive hard thresholding similar to Cai and Liu [2011] to obtain $\widehat{S}_e^{\tau}$, where each entry is set as

$$\{\widehat{S}_e^{\tau}\}_{i,j} = \begin{cases} \{\widehat{S}_e\}_{i,i}, \quad i = j, \\ \{\widehat{S}_e\}_{i,j} \mathbb{1}(|\{\widehat{S}_e\}_{i,j}| \geq \tau \sqrt{\{\widehat{S}_e\}_{i,i} \{\widehat{S}_e\}_{i,j}}), \quad i \neq j, \end{cases} \tag{11}$$

for a given $\tau > 0$, which amounts to thresholding the corresponding correlation matrix. This approach yields an optimal rate of convergence (Cai and Zhou [2012]) for Σ_e.

Theorem 2 (Error component). *Under assumptions of Theorem 1, Assumption 2 and given* $\|S - \Sigma\|_{max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right)$*, for a large enough* $\tau > 0$ *we have*

$$\|\widehat{S}_e^{\tau} - \Sigma_e\| = O_p\left(m\sqrt{\frac{s \log p}{n}}\right).$$

Proof. See Appendix A.2. □

Optimal estimation of a sparse component is discussed in detail in Cai and Zhou [2012], Fan et al. [2013]. The assumptions are not unusual and are easy to verify.

Thus the final estimator is given as

$$\widehat{S} = \sum_{j=1}^{r} \widehat{\ell}_j \widehat{\mathrm{v}}_j \widehat{\mathrm{v}}_j{}' + \widehat{S}_e^{\tau}, \tag{12}$$

The estimation involves two stages, one for estimating each of the components. The full algorithm is provided in Algorithm 2.

Theorem 3. *Suppose the assumptions of Theorem 2 hold. Then*

$$\|\widehat{S} - \Sigma\| = O_p\left(m\sqrt{\frac{s\log p}{n}}\right).$$

Proof. See Appendix A.3. □

Notice that this estimator achieves the optimal bound Cai and Zhou [2012] for high-dimensional covariance estimators in sparse settings. The proof is provided for rank-1 perturbations as in Equation (5), however it can be generalized to multi-spike covariances.

3.2 Factor Model framework

This subsection demonstrates that the covariance structure considered in the previous sections is implied by weak (non-pervasive) factors with approximately sparse loading matrices.

A latent factor model is given as

$$\underset{p\times 1}{X_i} = \Lambda \underset{r\times 1}{F_i} + e_i, \tag{13}$$

Λ is an approximately sparse loading matrix for the r factors in F_i, e_i is an idiosyncratic disturbance; and $i = 1, \ldots, n$. Only X_i are observable. The latter equation can be rewritten in matrix form

$$X = F\Lambda' + e, \tag{14}$$

where $X = [X_1 \ \cdots \ X_n]'$ and $F = [F_1 \ \cdots \ F_n]'$.

For the above factor model specification, we have the corresponding population covariance matrix of X_i,

$$\Sigma := \mathbb{E}(X'X) = \Lambda\Lambda' + \Omega, \tag{15}$$

where $\Omega = \mathbb{E}(e'e)$ is assumed to be approximately sparse. Hence Equation (15) admits low-rank plus sparse representation. The two components can be asymptotically identified only when the eigengap Δ is sufficiently large while Ω has bounded eigenvalues as a consequence of sparsity. This is crucial for consistent estimation since if nonzero eigenvalues of $\Lambda\Lambda'$ are smaller than $\|\Omega\|$, then it is impossible to distinguish signal from noise. In practice, factors are expected to exhibit sufficiently strong signal while the remaining part normally has weak correlation. The leading r eigenvectors of Σ should be nearly aligned with the corresponding columns of Λ. One can also view Σ in Equation (15) as a perturbation of $\Lambda\Lambda'$ by Ω. Further, denote the eigendecomposition of Σ as in Equation (3).

In vanilla factor modeling, one can obtain factors and loading estimates via PCA by solving

$$\begin{gathered}\underset{F,\Lambda}{\arg\min}\ \ \|X - F\Lambda'\|_F^2 \\ p^{-1}\Lambda'\Lambda = I_r, \quad F'F \text{ diagonal.}\end{gathered}$$

We can formulate a similar optimization problem in a penalized PCA fashion, similar to Equation (7), that would correspond to a sparse setting considered in this paper. Specifically, to obtain an approximate solution for the first loading column Λ_1 one can solve

$$\begin{gathered}\underset{\Lambda_1}{\arg\min}\ \Lambda_1' X' X \Lambda_1 \\ \|\Lambda_1\|_0 \leq s,\ \|\Lambda_1\|_2 = 1.\end{gathered} \tag{16}$$

Clearly, the truncated power iteration method described earlier can be used. Its solution $\widehat{\Lambda}$ will be the first r sparse eigenvectors of $X'X$. The corresponding factor estimate can simply be calculated as $\widehat{F}_i = (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'X_i = \widehat{\Lambda}'X_i$. Hence, the technique and the theory considered above are applicable when a sparsity assumption on the loadings is justifiable. This may also be beneficial for constructing interpretable factor models; a similar setup from the Bayesian viewpoint is considered in Pati et al. [2014].

4 Numerical experiment

Generate the covariance as follows

$$\Sigma = ULU' = U_r L_r U_r' + \Sigma_e,$$

where U_r is a $p \times r$ matrix of eigenvectors corresponding to top eigenvalues and L_r is an $r \times r$ diagonal matrix of eigenvalues in descending order. Specifically, we set $r = 2$ across all simulations are generate the two columns in $U_r = (\mathrm{u}_1 \quad \mathrm{u}_2)$ as follows,

$$\{\mathrm{u}_1\}_i = \begin{cases} \frac{1}{\sqrt{s}}, & i \in [1,\ s] \\ 0, & \text{otherwise} \end{cases}, \quad \{\mathrm{u}_2\}_i = \begin{cases} \frac{1}{\sqrt{s}}, & i \in [s+1,\ 2s] \\ 0, & \text{otherwise} \end{cases},$$

We set ℓ_1, ℓ_2 to either $(200, 100)$ or $(500, 300)$. The entries of the error component Σ_e are generated in a block-diagonal fashion with

$$\{\Sigma_e\}_{i,j} = \rho^{|i-j|} \mathbb{1}(|i-j| \leq 1),$$

and set $\rho = .5$. This design ensures sparsity and is sometimes referred to as MA(1).

The data $X \in \mathbb{R}^{n \times p}$ is generated by drawing n samples from $X \sim \mathcal{N}_p(0, \Sigma)$. We run 100 Monte Carlo simulations. The proposed method is compared against POET (Fan et al. [2013]), a hard-thresholding estimator of Bickel and Levina [2008] and a linear shrinkage estimator of Ledoit and Wolf [2004].

We vary $p \in \{100, 300, 500\}$ and the sparsity $s \in \{5, 15, 25\}$. The results report the ratios of a spectral (or Frobenius) norm of the covariance estimation error of a given method with respect to POET's corresponding norm. Tables 1 and 2 consider cases with $\ell_1 = 200, \ell_2 = 100$ and $\ell_1 = 500, \ell_2 = 300$ respectively. The proposed method is dubbed as "DSCE" for doubly sparse covariance estimator.

The tables are indicative of high estimation accuracy of the proposed method in sparse settings as compared to the competition. The precision of DSCE seems to gen-

Table 1: Ratios to POET error $n = 300, \ell_1 = 200, \ell_2 = 100$.

p	Method	Spectral			Frobenius		
		$s = 5$	$s = 15$	$s = 25$	$s = 5$	$s = 15$	$s = 25$
100	DSCE	0.7462	0.7308	0.7800	0.7587	0.7776	0.7717
100	B&L	0.8701	0.9139	0.9287	0.9072	0.9246	0.9501
100	L&W	1.0297	1.0122	0.9598	1.0317	1.0066	0.9804
300	DSCE	0.7076	0.7382	0.7551	0.7380	0.7733	0.7706
300	B&L	0.8844	0.9253	0.9331	0.9073	0.9336	0.9481
300	L&W	1.0318	0.9989	0.9887	1.0518	1.0377	0.9894
500	DSCE	0.6943	0.7100	0.7362	0.6964	0.7390	0.7477
500	B&L	0.9127	0.9286	0.9463	0.9286	0.9483	0.9623
500	L&W	1.0937	1.0626	1.0215	1.1254	1.0750	1.0532

Table 2: Ratios to POET error $n = 300, \ell_1 = 500, \ell_2 = 300$.

p	Method	Spectral			Frobenius		
		$s = 5$	$s = 15$	$s = 25$	$s = 5$	$s = 15$	$s = 25$
100	DSCE	0.7108	0.7135	0.7776	0.7292	0.7415	0.7677
100	B&L	0.8999	0.9257	0.9280	0.9183	0.9569	0.9454
100	L&W	1.0446	0.9944	0.9720	1.0603	1.0266	0.9937
300	DSCE	0.6919	0.7038	0.7395	0.7008	0.7299	0.7474
300	B&L	0.9144	0.9257	0.9239	0.9409	0.9542	0.9544
300	L&W	1.0358	1.0209	0.9789	1.0847	1.0125	0.9800
500	DSCE	0.6560	0.6886	0.6955	0.6877	0.7167	0.7115
500	B&L	0.9504	0.9685	0.9605	0.9593	0.9733	0.9714
500	L&W	1.0873	1.0594	1.0239	1.1203	1.0807	1.0673

erally increase with dimension p and decreased in sparsity s. As evident from Table 2, stronger signal makes both DSCE and POET perform better compared to the other methods.

5 Concluding Remarks

We consider estimation high-dimensional covariance estimation in sparse settings. Our approach goes beyond mere eigenvalue shrinkage by taking into consideration the behavior of eigenvectors in a large dimensional framework. Furthermore, we do not require diverging (pervasive) signals, but instead assume sparsity, which is a feature of many large datasets and a desirable characteristic to require in a number of applications. Our model consists of low-rank and sparse components, where the former also has sparse eigenvectors.

Our analysis shows that accurate estimation is possible, with the rate of convergence proportional to the optimal rate for sparse estimation as in Cai and Zhou [2012]. The numerical experiment also reveals that the proposed algorithm can accurately estimate the induced covariance. It would also be worth considering an empirical application where nonzero coefficients are associated with loss, e.g. transaction costs in finance, so that it is desirable to have sparse estimates.

Our empirical application also demonstrates that the proposed method may offer substantial advantages over other high-dimensional estimation techniques. One of the possible extensions is a closer examination of a multi-spike covariance model and the issues arising therein, e.g. whether sparsity should vary across eigenvectors and how. Another extension would be the consideration of linear shrinkage (of eigenvalues) of a sparse estimator discussed here with a structured estimator. Finally, it is valuable to explore the properties of the inverse (precision) estimator induced by the proposed method.

Appendix

A.1 Useful Lemmas

Lemma 1. *For unit vectors* u *and* v*,*

$$\|\mathrm{uu}' - \mathrm{vv}'\|_F^2 = 2 - 2(\mathrm{u}'\mathrm{v})^2$$

Proof.

$$\begin{aligned}\|\mathrm{uu}' - \mathrm{vv}'\|_F^2 &= \mathbf{tr}((\mathrm{uu}' - \mathrm{vv}')(\mathrm{uu}' - \mathrm{vv}')) \\ &= 1 - \mathbf{tr}(\mathrm{uu}'\mathrm{vv}) - \mathbf{tr}(\mathrm{vv}'\mathrm{uu}') + 1 \\ &= 2 - 2(\mathrm{u}'\mathrm{v})^2\end{aligned}$$

□

Lemma 2. *For a rank-1 matrix* $A = \mathrm{x}_1\mathrm{x}_2'$ *we have*

$$\|A\| = \|\mathrm{x}_1\|_2\|\mathrm{x}_2\|_2.$$

Proof. The equality is trivial if $\mathrm{x}_2 = 0$, so consider $\mathrm{x}_2 \neq 0$. For a vector $\mathrm{u} = \frac{\mathrm{x}_2}{\|\mathrm{x}_2\|_2}$ we have

$$\|A\| \geq \|A\mathrm{u}\|_2 = \left\|\mathrm{x}_1\mathrm{x}_2' \frac{\mathrm{x}_2}{\|\mathrm{x}_2\|_2}\right\|_2 = \frac{1}{\|\mathrm{x}_2\|_2}\|\mathrm{x}_1\mathrm{x}_2'\mathrm{x}_2\|_2 = \frac{\|\mathrm{x}_2\|_2^2}{\|\mathrm{x}_2\|_2}\|\mathrm{x}_1\|_2 = \|\mathrm{x}_1\|_2\|\mathrm{x}_2\|_2.$$

On the other hand,

$$\|A\| = \|\mathrm{x}_1\mathrm{x}_2'\|_2 \leq \|\mathrm{x}_1\|_2\|\mathrm{x}_2'\|_2 = \|\mathrm{x}_1\|_2\|\mathrm{x}_2\|_2.$$

□

A.2 Proof of Theorem 2

Proof. It suffices to prove that the max error is bounded

$$\|\widehat{S}_e - \Sigma_e\|_{\text{max}} = O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right).$$

The desired rate on adaptive threshold estimator would follow immediately as discussed in Cai and Liu [2011] and Rothman et al. [2009].

Recall from Equation 10 that

$$\widehat{S}_e = S - \widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r',$$

where $\widehat{V}_r$ is a $p \times r$ matrix of sparse eigenvector estimates and $\widehat{\Lambda}_r$ is an $r \times r$ diagonal matrix of the corresponding eigenvalue estimates in descending order.

$$\begin{aligned}\|\widehat{S}_e - \Sigma_e\|_{\text{max}} &= \|S - \widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - (\Sigma - U_rL_rU_r')\|_{\text{max}}\\ &\le \|S - \Sigma\|_{\text{max}} + \|\widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - U_rL_rU_r'\|_{\text{max}}\end{aligned}$$

By assumption, $\|S-\Sigma\|_{\text{max}} = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right)$, so we want to show $\|\widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - U_rL_rU_r'\|_{\text{max}} = O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right)$. Then since the eigenvalues are bounded by assumption, we have

$$\begin{aligned}&\|\widehat{V}_r\widehat{\Lambda}_r\widehat{V}_r' - U_rL_rU_r'\|_{\text{max}}\\ &\le \|\widehat{V}_r(\widehat{\Lambda}_r - L_r)\widehat{V}_r'\|_{\text{max}} + \|(\widehat{V}_r - U_r)L_r(\widehat{V}_r - U_r)'\|_{\text{max}} + 2\|U_rL_r(\widehat{V}_r - U_r)'\|_{\text{max}}\\ &= O_{\mathbb{P}}\left(\|\widehat{\Lambda}_r - L_r\|_{\text{max}} + \|\widehat{V}_r - U_r\|_{\text{max}}\right)\\ &= O_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n}}\right).\end{aligned}$$

□

A.3 Proof of Theorem 3

Proof.

$$\begin{aligned}
\|\widehat{S} - \Sigma\| &= \|\widehat{S}_s + \widehat{S}_e^{\tau} - \Sigma_s - \Sigma_e\| \\
&\leq \|\ell_1 \mathrm{u}_1 \mathrm{u}_1' - \widehat{\lambda}_1 \widehat{\mathrm{v}}_1 \widehat{\mathrm{v}}_1'\| + \|\widehat{S}_e^{\tau} - \Sigma_e\| \\
&= \|\ell_1(\mathrm{u}_1 \mathrm{u}_1' - \widehat{\mathrm{v}}_1 \widehat{\mathrm{v}}_1') + (\ell_1 - \widehat{\lambda}_1)\widehat{\mathrm{v}}_1 \widehat{\mathrm{v}}_1'\| + O_p\left(m\sqrt{n^{-1} s \log p}\right) \\
&= \ell_1 O_{\mathbb{P}}(\|\mathrm{u}_1 - \widehat{\mathrm{v}}_1\|) + O_{\mathbb{P}}(n^{-1/2})\|\widehat{\mathrm{v}}_1 \widehat{\mathrm{v}}_1'\| + O_p\left(m\sqrt{n^{-1} s \log p}\right) && (17) \\
&= O_{\mathbb{P}}(\|\underline{\mathrm{E}}_{\hat{s}}\|) + O_{\mathbb{P}}(n^{-1/2})\|\widehat{\mathrm{v}}_1 \widehat{\mathrm{v}}_1'\| + O_p\left(m\sqrt{n^{-1} s \log p}\right) && (18) \\
&= O_{\mathbb{P}}(\|\underline{\mathrm{E}}_{\hat{s}}) + O_{\mathbb{P}}(n^{-1/2})\|\widehat{\mathrm{v}}_1\|_2^2 + O_p\left(m\sqrt{n^{-1} s \log p}\right) && (19) \\
&= O_{\mathbb{P}}(\|\underline{\mathrm{E}}_{\hat{s}}) + O_{\mathbb{P}}(n^{-1/2}) + O_p\left(m\sqrt{n^{-1} s \log p}\right), && (20)
\end{aligned}$$

where (17) follows from Lemma 1 and the fact that $\|\cdot\| \leq \|\cdot\|_F$ for a square matrices; (19) follows from Corrolary 1.1; (19) holds by Lemma 2.

To complete the proof observe that in Equation 20 the first term becomes $O_{\mathbb{P}}\left(\sqrt{\frac{s \log p}{n}}\right)$ by Corrolary 1.2, while the second term is asymptotically negligible under the Assumption 2. □

A.4 Weyl and Davis-Kahan

Denote eigenvectors and eigenvalues as $\xi_i(\cdot)$ and $\lambda_i(\cdot)$ respectively. For Hermitian $p \times p$ matrices $A, \widehat{A}$,

Proposition A.1.

$$\max_{i=1,\ldots,p} |\lambda_i(\widehat{A}) - \lambda_i(A)| \leq \|\widehat{A} - A\|.$$

Proposition A.2.

$$\left\|\xi_i(\widehat{A}) - \xi_i(A)\right\| \leq \frac{\sqrt{2}\|\widehat{A} - A\|}{\min(|\lambda_{i-1}(\widehat{A}) - \lambda_i(A)|, |\lambda_{i+1}(\widehat{A}) - \lambda_i(A)|)}.$$

A.5 T-Power Algorithm

Algorithm 1: Truncated Power Method Yuan and Zhang [2011]

```
Input: $p \times p$ PSD matrix $A$, initial estimate $x_0 \in \mathbb{R}^p$, dimension $r$.
t = 1
while not converged do
    $x'_t = Ax_{t-1}/\|Ax_{t-1}\|$;                      /* Power iteration */
    $F_t = \text{supp}(x'_t, r)$ ;                       /* Support of indices */
    $\hat{x}_t = \text{Truncate}(x'_t, F_t)$;            /* Truncate */
    $x_t = \hat{x}_t/\|\hat{x}_t\|$ ;                    /* Standardize */
    $t = t + 1$
end
```

A.6 DSCE Algorithm

Algorithm 2: Proposed Algorithm

```
Input: standardized $p \times n$ data matrix $X$, dimension $r$.
$S = \frac{1}{n}X'X$ ;                                   /* Sample covariance */
$\widehat{S}_s = \text{T-Power}(S)$ ;                    /* Rank-$r$ Truncuted power method */
$\widehat{S}_e^\tau = \tau(S - \widehat{S}_s)$ ;         /* Adaptive Thresholding,(10) */
Output: $p \times p$ matrix $\widehat{S} = \widehat{S}_s + \widehat{S}_e^\tau$.
```

References

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877 – 2921.

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, 34(1):122 – 148.

Bai, Z. and Silverstein, J. (2010). *Spectral Analysis of Large Dimensional Random Matrices*.

Bai, Z. D., Miao, B. Q., and Pan, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532 – 1572.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294.

Baik, J., Ben Arous, G., and Peche, S. (2005). Phase transition of the largest eigenvalue for non-null complex covariance matrices. *Annals of Probability*, 33.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408.

Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780 – 1815.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604.

Bouchaud, J.-P. and Potters, M. (2009). Financial applications of random matrix theory: a short review. *arXiv.org, Quantitative Finance Papers*.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389 – 2420.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *J. ACM*, 58(3).

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448.

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

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295 – 327.

Johnstone, I. M. and Lu, A. Y. (2004). Sparse principal components analysis.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693. PMID: 20617121.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483.

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102 – 1130.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642.

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

Shen, D., Shen, H., Zhu, H., and Marron, J. S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26(4):1747–1770.

Tracy, C. A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727 – 754.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342 – 1374.

Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10:515–34.

Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization.

In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Yuan, X.-T. and Zhang, T. (2011). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.