The Annals of Statistics
2021, Vol. 49, No. 2, 1113–1138
https://doi.org/10.1214/20-AOS1995

SPIKED SEPARABLE COVARIANCE MATRICES AND PRINCIPAL COMPONENTS

BY XIUCAI DING[1] AND FAN YANG[2]

[1]*Department of Mathematics, Duke University, xiuca.ding@duke.edu*
[2]*Department of Statistics, University of Pennsylvania, fyang75@wharton.upenn.edu*

We study a class of separable sample covariance matrices of the form $\widetilde{\mathcal{Q}}_1 := \widetilde{A}^{1/2} X \widetilde{B} X^* \widetilde{A}^{1/2}$. Here, $\widetilde{A}$ and $\widetilde{B}$ are positive definite matrices whose spectrums consist of bulk spectrums plus several spikes, that is, larger eigenvalues that are separated from the bulks. Conceptually, we call $\widetilde{\mathcal{Q}}_1$ a *spiked separable covariance matrix model*. On the one hand, this model includes the spiked covariance matrix as a special case with $\widetilde{B} = I$. On the other hand, it allows for more general correlations of datasets. In particular, for spatiotemporal dataset, $\widetilde{A}$ and $\widetilde{B}$ represent the spatial and temporal correlations, respectively.

In this paper, we study the outlier eigenvalues and eigenvectors, that is, the principal components, of the spiked separable covariance model $\widetilde{\mathcal{Q}}_1$. We prove the convergence of the outlier eigenvalues $\widetilde{\lambda}_i$ and the generalized components (i.e., $\langle \mathbf{v}, \widetilde{\boldsymbol{\xi}}_i \rangle$ for any deterministic vector $\mathbf{v}$) of the outlier eigenvectors $\widetilde{\boldsymbol{\xi}}_i$ with optimal convergence rates. Moreover, we also prove the delocalization of the nonoutlier eigenvectors. We state our results in full generality, in the sense that they also hold near the so-called BBP transition and for degenerate outliers. Our results highlight both the similarity and difference between the spiked separable covariance matrix model and the spiked covariance matrix model in (*Probab. Theory Related Fields* **164** (2016) 459–552). In particular, we show that the spikes of both $\widetilde{A}$ and $\widetilde{B}$ will cause outliers of the eigenvalue spectrum, and the eigenvectors can help to select the outliers that correspond to the spikes of $\widetilde{A}$ (or $\widetilde{B}$).

1. Introduction. High-dimensional data obtained at space-time points has been increasingly employed in various scientific fields, such as geophysical and environmental sciences [30, 36], wireless communications [26, 53, 55], medical imaging [50] and financial economics [43, 44, 59]. The structural assumption of separability is a popular assumption in the analysis of spatiotemporal data. Although this assumption does not allow for space-time interactions in the covariance matrix, in many real data applications (e.g., the study of Irish wind speed [22]), the covariance matrix can be well approximated using separable covariance matrices by solving a nearest Kronecker product for a space-time covariance matrix problem (NKPST) [21].

Consider a $p \times n$ data matrix Y of the form

$$Y = \widetilde{A}^{1/2} X \widetilde{B}^{1/2}, \tag{1.1}$$

where $X = (x_{ij})$ is a $p \times n$ random matrix with independent entries such that $\mathbb{E}x_{ij} = 0$ and $\mathbb{E}|x_{ij}|^2 = n^{-1}$, and $\widetilde{A}$ and $\widetilde{B}$ are respectively $p \times p$ and $n \times n$ deterministic positive-definite matrices. We say Y has a separable covariance structure because the joint spatiotemporal covariance of Y, viewed as an (np)-dimensional vector consisting of the columns of Y

Received July 2019; revised June 2020.
MSC2020 subject classifications. Primary 15B52, 62E20; secondary 62H99.
Key words and phrases. Spiked separable covariance matrices, principal components, BBP transition, local laws.

stacked on top of one another, is given by a separable form $\widetilde{A} \otimes \widetilde{B}$, where $\otimes$ denotes the Kronecker product. This model has different names and meanings in different fields. For example, in wireless communications [26, 53, 55], especially for the multiple-input-multiple-output (MIMO) systems, the $\widetilde{A}$ and $\widetilde{B}$ represent the covariances between the receiver antennas and between the transmitter antennas, respectively. Also, Y is called the doubly-heteroscedastic noise in [35] for matrix denoising and the separable idiosyncratic part in factor model [43]. However, as a convention, in this paper we always say that the row indices of Y correspond to spatial locations while the column indices correspond to time points. Moreover, we shall call $\widetilde{A}$ and $\widetilde{B}$ as spatial and temporal covariance matrices, respectively. In this paper, we are mainly interested in the so-called *separable sample covariance matrix* $\widetilde{Q}_1 := YY^*$ for the above separable data model Y.

One special case is the classic sample covariance matrix when $\widetilde{B} = I_n$, which has been a central object of study in multivariate statistics. In the null case with $\widetilde{A} = I_p$, it is well known that the empirical spectral distribution (ESD) of $\widetilde{Q}_1$ converges to the celebrated Marchenko–Pastur (MP) law [38]. Later on the convergence result of the ESD is extended to various settings with general positive definite covariance matrices $\widetilde{A}$; we refer the readers to the monograph [1] and the review paper [47]. For the extremal eigenvalues, the Tracy–Widom distribution [51, 52] of the extremal eigenvalue was first proved in [24] for sample covariance matrices with $\widetilde{A} = I_p$ and Gaussian X (i.e., the entries of X are i.i.d. Gaussian), and later proved for X with generally distributed entries in [49]. When $\widetilde{A}$ is a general nonscalar matrix, the Tracy–Widom distribution was first proved for the case with i.i.d. Gaussian X in [16, 42] and later proved under various moment assumptions on the entries x_{ij} [3, 13, 28, 34]. Finally, for the (nonoutlier) sample eigenvectors, the completely delocalization [28, 41], quantum unique ergodicity [6], distribution of the eigenvector components [9] and convergence of eigenvector empirical spectral distribution [56] have been constructed.

In the statistical study of sample covariance matrices, a popular model is the *Johnstone's spiked covariance matrix model* [24]. In this model, a few spikes, that is, eigenvalues detached from the bulk eigenvalue spectrum, are added to $\widetilde{A}$. Since the seminal work of Baik, Ben Arous and Péché [2], it is now well understood that the extremal eigenvalues undergo a so-called BBP transition along with the change of the strength of the spikes. Roughly speaking, there is a critical value such that the following properties hold: if the strength of the spike is smaller than the critical value, then the extremal eigenvalue of the spiked sample covariance matrix will stick to the right endpoint of the bulk eigenvalue spectrum (and hence is not an outlier), and the corresponding sample eigenvector will be delocalized; otherwise, if the strength of the spike is larger than the critical value, then the associated eigenvalue will jump out of the bulk eigenvalue spectrum, and the outlier sample eigenvector will be concentrated on a cone with axis parallel to the population eigenvector with an (almost) deterministic aperture. For an extensive overview of such results, we refer the reader to [6, 11, 46].

One purpose of this paper is to generalize some important results for sample and spiked covariance matrices to the more general separable and spiked separable covariance matrices. The convergence of the ESD of separable covariance matrices to a limiting law were shown in [48, 54, 60]. The edge universality and delocalization of eigenvectors have been proved by the second author [58] for separable covariance matrices *without* spikes on $\widetilde{A}$ and $\widetilde{B}$. The convergence of VESD of separable covariance matrices was proved in [57], which is an extension of the result in [56]. Then the main goal of this paper is to study the outlier eigenvalues and eigenvectors of separable covariance matrices with spikes on both the spatial and temporal covariance matrices $\widetilde{A}$ and $\widetilde{B}$, which we shall refer to as the *spiked separable covariance matrices*. The precise definition is given in Section 2.

In this paper, we derive precise large deviation estimates on the outlier eigenvalues and the generalized components of the outlier eigenvectors. In particular, our results give both the

first-order limits and the (almost) optimal rates of convergence of the relevant quantities. We now describe them briefly. Let $\widetilde{A} = \sum_{i=1}^{p} \widetilde{\sigma}_i^a \mathbf{v}_i^a (\mathbf{v}_i^a)^*$ and $\widetilde{B} = \sum_{\mu=1}^{n} \widetilde{\sigma}_\mu^b \mathbf{v}_\mu^b (\mathbf{v}_\mu^b)^*$ be the eigendecomposition of $\widetilde{A}$ and $\widetilde{B}$, respectively, where we label the eigenvalues in descending order. We assume that the spiked eigenvalues are $\{\widetilde{\sigma}_i^a\}_{i=1}^r$ and $\{\widetilde{\sigma}_\mu^b\}_{\mu=1}^s$, where r and s are some fixed integers. Then there exists a threshold ℓ_a (or ℓ_b) such that $\widetilde{\sigma}_i^a$ (or $\widetilde{\sigma}_\mu^b$) gives rise to outliers of $\widetilde{\mathcal{Q}}_1$ if and only if $\widetilde{\sigma}_i^a > \ell_a$ (or $\widetilde{\sigma}_\mu^b > \ell_b$). Moreover, the outlier lies around a fixed location determined by the spike $\widetilde{\sigma}_i^a$ (or $\widetilde{\sigma}_\mu^b$); see Theorem 3.6. If $\widetilde{\sigma}_i^a - \ell_a \gg n^{-1/3}$ or $\widetilde{\sigma}_j^b - \ell_b \gg n^{-1/3}$, that is, the spike is supercritical, then the outlier will be well separated from the bulk spectrum and can be detected readily. For $0 < \widetilde{\sigma}_i^a - \ell_a \ll n^{-1/3}$ or $0 < \widetilde{\sigma}_j^b - \ell_b \ll n^{-1/3}$, that is, the spike is subcritical, the corresponding "outlier" cannot be distinguished from the bulk spectrum and will instead stick to the right-most edge of the bulk spectrum up to some random fluctuation of order $O(n^{-2/3})$. Next, for the sample eigenvector of $\widetilde{\mathcal{Q}}_1$ that is associated with the outlier caused by a supercritical spike $\widetilde{\sigma}_i^a$, we show that it is concentrated on a cone with axis parallel to the population eigenvector $\mathbf{v}_i^a$ with an explicit aperture determined by $\widetilde{\sigma}_i^a$. On the other hand, the sample eigenvector of $\widetilde{\mathcal{Q}}_1$ that is associated with a supercritical spike $\widetilde{\sigma}_\mu^b$ is delocalized. Similar results hold for the right singular vectors of Y, that is, the eigenvectors of $\widetilde{\mathcal{Q}}_2 := \widetilde{B}^{1/2} X^* \widetilde{A} X \widetilde{B}^{1/2}$, by switching the roles of $\widetilde{A}$ and $\widetilde{B}$. Finally, for the nonoutlier singular vectors, that is, singular vectors associated with subcritical and bulk eigenvalues, we proved that they are delocalized. We point out that our results are in the same spirit as the ones for deformed Wigner matrix [27], deformed rectangular matrix [4, 10] and spiked covariance matrices [6, 11, 46].

The information from sample singular vectors is very important in the estimation of spiked separable covariance matrices. For example, one important parameter to estimate is the number of spikes. For spiked separable covariance matrices, the outliers have two different origins from either $\widetilde{A}$ or $\widetilde{B}$. Hence we need to estimate the number of spikes for each of them. In the literature of spiked covariance matrices [45], the number of spikes is estimated using statistic constructed from eigenvalues only. However, this only gives an estimation of the total number of spikes. To distinguish the two types of spikes, we also need to utilize the information from singular vectors. This will be discussed in detail in Section 4.

Before concluding the introduction, we summarize the main contributions of our work.

- We introduce the spiked separable covariance matrix model; see (2.12). It allows for more general covariance structure and is suitable for spatiotemporal data analysis with spikes in both space and time.
- For both supercritical and subcritical spikes, we obtain the first-order limits of the corresponding eigenvalue outliers and the generalized components of the associated eigenvectors. Moreover, our results provide a precise rate of convergence, which we believe to be optimal up to some n^ε factor. They are presented in Theorems 3.6 and 3.10.
- We prove large deviation bounds for the nonoutlier eigenvalues and eigenvectors. In particular, we prove that the nonoutlier eigenvalues will stick with those of the reference matrix. Moreover, the nonoutlier eigenvectors near the spectrum edge will be biased in the direction of the population eigenvectors of the subcritical spikes. These results are presented in Theorems 3.7 and 3.14.
- We address two important issues in the estimation of spiked separable covariance matrices. First, we provide statistics to estimate the number of spikes for $\widetilde{A}$ and $\widetilde{B}$. In particular, we will show that the eigenvectors are important for us to separate the outliers from the spikes of $\widetilde{A}$ and those from the spikes of $\widetilde{B}$. Second, we obtain the optimal shrinkage for the eigenvalues, which is adaptive to the data matrix only. These are discussed in Section 4.

This paper is organized as follows. In Section 2, we define the spiked separable covariance matrix. In Section 3, we state our main results. In Section 4, we address two important issues

regarding the statistical estimation of the proposed spiked separable covariance matrices. We present the technical proofs in the Supplementary Material.

2. Definition of spiked separable covariance matrices.

2.1. *The model*. We first consider a class of separable sample covariance matrices of the form $\mathcal{Q}_1 := A^{1/2}XBX^*A^{1/2}$, where A and B are deterministic nonnegative definite symmetric (or Hermitian) matrices. Note that A and B are not necessarily diagonal. We assume that $X = (x_{ij})$ is a $p \times n$ random matrix, where the entries x_{ij}, $1 \le i \le p$, $1 \le j \le n$, are real or complex independent random variables satisfying

$$\mathbb{E}x_{ij} = 0, \qquad \mathbb{E}|x_{ij}|^2 = n^{-1}. \tag{2.1}$$

For definiteness, in this paper we focus on the real case, that is, the random variables x_{ij} are real. However, our proof can be applied to the complex case after minor modifications if we assume in addition that $\operatorname{Re} x_{ij}$ and $\operatorname{Im} x_{ij}$ are independent centered random variables with variance $(2n)^{-1}$. We assume that the entries $\sqrt{n}x_{ij}$ have bounded fourth moment:

$$\max_{i,j} \mathbb{E}|\sqrt{n}x_{ij}|^4 \le C_4, \tag{2.2}$$

for some constant $C_4 > 0$. We will also use the $n \times n$ matrix $\mathcal{Q}_2 := B^{1/2}X^*AXB^{1/2}$. We denote the eigenvalues of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ in descending order by $\lambda_1(\mathcal{Q}_1) \ge \cdots \ge \lambda_p(\mathcal{Q}_1)$ and $\lambda_1(\mathcal{Q}_2) \ge \cdots \ge \lambda_n(\mathcal{Q}_2)$. Since $\mathcal{Q}_1$ and $\mathcal{Q}_2$ share the same nonzero eigenvalues, we will simply write λ_j, $1 \le j \le p \wedge n$, to denote the jth eigenvalue of both $\mathcal{Q}_1$ and $\mathcal{Q}_2$ without causing any confusion.

We shall consider the high-dimensional setting in this paper. More precisely, we assume that there exists a constant $0 < \tau < 1$ such that the aspect ratio $d_n := p/n$ satisfies

$$\tau \le d_n \le \tau^{-1} \quad \text{for all } n. \tag{2.3}$$

We assume that A and B have eigendecompositions

$$A = V^a\Sigma^a(V^a)^*, \qquad B = V^b\Sigma^b(V^b)^*, \tag{2.4}$$

where

$$\Sigma^a = \operatorname{diag}(\sigma_1^a, \ldots, \sigma_p^a), \qquad \Sigma^b = \operatorname{diag}(\sigma_1^b, \ldots, \sigma_n^b),$$

and

$$V^a = (\mathbf{v}_1^a, \ldots, \mathbf{v}_p^a), \qquad V^b = (\mathbf{v}_1^b, \ldots, \mathbf{v}_n^b).$$

We denote the empirical spectral distributions (ESD) of A and B by

$$\pi_A \equiv \pi_A^{(p)} := \frac{1}{p}\sum_{i=1}^{p} \delta_{\sigma_i^a}, \qquad \pi_B \equiv \pi_B^{(n)} := \frac{1}{n}\sum_{i=1}^{n} \delta_{\sigma_i^b}. \tag{2.5}$$

We assume that there exists a small constant $0 < \tau < 1$ such that, for all n large enough,

$$\max\{\sigma_1^a, \sigma_1^b\} \le \tau^{-1}, \qquad \max\{\pi_A^{(p)}([0,\tau]), \pi_B^{(n)}([0,\tau])\} \le 1 - \tau. \tag{2.6}$$

Note the first condition means that the operator norms of A and B are bounded by τ^{-1}, and the second condition means that the spectrums of A and B cannot concentrate at zero.

In this paper, we study spiked separable sample covariance matrices, which can be realized through a low rank perturbation of the nonspiked version. We shall assume that $\mathcal{Q}_1$ is a separable sample covariance matrix without spikes (see Assumption 2.6 below). To add

spikes, we follow the setup in [11] and assume that there exist some fixed intergers $r, s \in \mathbb{N}$ and constants d_i^a, $1 \le i \le r$ and d_μ^b, $1 \le \mu \le s$, such that

$$\begin{aligned} \widetilde{A} &= V^a \widetilde{\Sigma}^a (V^a)^*, \qquad \widetilde{B} = V^b \widetilde{\Sigma}^b (V^b)^*, \\ \widetilde{\Sigma}^a &= \operatorname{diag}(\widetilde{\sigma}_1^a, \ldots, \widetilde{\sigma}_p^a), \qquad \widetilde{\Sigma}^b = \operatorname{diag}(\widetilde{\sigma}_1^b, \ldots, \widetilde{\sigma}_n^b), \end{aligned} \tag{2.7}$$

where

$$\widetilde{\sigma}_i^a = \begin{cases} \sigma_i^a(1+d_i^a), & 1 \le i \le r, \\ \sigma_i^a, & \text{otherwise}, \end{cases} \qquad \widetilde{\sigma}_\mu^b = \begin{cases} \sigma_\mu^b(1+d_\mu^b), & 1 \le \mu \le s, \\ \sigma_\mu^b, & \text{otherwise}. \end{cases} \tag{2.8}$$

Without loss of generality, we assume that we have reordered indices such that

$$\widetilde{\sigma}_1^a \ge \widetilde{\sigma}_2^a \ge \ldots \ge \widetilde{\sigma}_p^a \ge 0, \qquad \widetilde{\sigma}_1^b \ge \widetilde{\sigma}_2^b \ge \ldots \ge \widetilde{\sigma}_n^b \ge 0. \tag{2.9}$$

Moreover, we assume that

$$\max\{\widetilde{\sigma}_1^a, \widetilde{\sigma}_1^b\} \le \tau^{-1}. \tag{2.10}$$

With (2.7) and (2.8), we can write

$$\begin{aligned} \widetilde{A} &= A(I_p + V_o^a D^a (V_o^a)^*) = (I_p + V_o^a D^a (V_o^a)^*)A, \\ \widetilde{B} &= B(I_n + V_o^b D^b (V_o^b)^*) = (I_n + V_o^b D^b (V_o^b)^*)B, \end{aligned} \tag{2.11}$$

where

$$D^a = \operatorname{diag}(d_1^a, \ldots, d_r^a), \qquad V_o^a = (\mathbf{v}_1^a, \ldots, \mathbf{v}_r^a),$$

and

$$D^b = \operatorname{diag}(d_1^b, \ldots, d_s^b), \qquad V_o^b = (\mathbf{v}_1^b, \ldots, \mathbf{v}_s^b).$$

Then we define the spiked separable sample covariance matrices as

$$\widetilde{\mathcal{Q}}_1 = \widetilde{A}^{1/2} X \widetilde{B} X^* \widetilde{A}^{1/2}, \qquad \widetilde{\mathcal{Q}}_2 = \widetilde{B}^{1/2} X^* \widetilde{A} X \widetilde{B}^{1/2}. \tag{2.12}$$

REMARK 2.1. In the above definition, we have assumed that the nonspiked covariance matrix A (or B) and the spiked one $\widetilde{A}$ (or $\widetilde{B}$) share the same eigenvectors. Theoretically, the more general additive model actually can be reduced to our case as following: consider the following model:

$$\widetilde{A} = A + \Delta_A,$$

where A is the nonspiked part as above, and Δ_A is a finite rank perturbation. We can perform the eigendecomposition of $\widetilde{A}$ as

$$\widetilde{A} = \sum_{i=1}^{p} \widetilde{\sigma}_i^a \widetilde{\mathbf{v}}_i^a (\widetilde{\mathbf{v}}_i^a)^*,$$

where $\widetilde{\mathbf{v}}_i^a$ are not necessarily the eigenvectors of A. Then we can decompose $\widetilde{A}$ in its eigenbasis as

$$\widetilde{A} = A' + \Delta_A', \qquad A' = \sum_{i=1}^{p} \sigma_i' \widetilde{\mathbf{v}}_i^a (\widetilde{\mathbf{v}}_i^a)^*, \tag{2.13}$$

such that A' is a nonspiked matrix and Δ_A' is a finite rank perturbation. This is reduced to our setting again. Similar discussion also applies to $\widetilde{B}$.

In general, how the eigenvalues and eigenvectors of $\widetilde{A}$ are related to those of A and Δ_A is unknown—we even do not know whether Δ_A' has the same rank as Δ_A. One possible assumption is that the eigenvalues (i.e., the signal strengths) of Δ_A are relatively large compared to those of A, then the largest few eigenvalues and the corresponding eigenvectors should be well approximated by those of Δ_A, and our results can be applied again. However, the behaviors of the smaller eigenvalues can still be very interesting. For example, in Section A.2 of the Supplementary Material [12], we construct an example such that $B = I$ and Δ_A is a rank-1 matrix with a large signal, but $\widetilde{Q}_1$ has two outlier eigenvalues. The behavior of the decomposition (2.13) should depend strongly on the assumptions on A and Δ_A, and we will not pursue this direction in the current paper—it will be a subject for future study.

We summarize our basic assumptions here for future reference. For our purpose, we shall relax the assumption (2.1) a little bit.

ASSUMPTION 2.2. We assume that X is a $p \times n$ random matrix with real entries satisfying (2.2) and that

$$\max_{i,j} |\mathbb{E}x_{ij}| \le n^{-2-\tau}, \qquad \max_{i,j} |\mathbb{E}|x_{ij}|^2 - n^{-1}| \le n^{-2-\tau}, \tag{2.14}$$

for some constant $\tau > 0$. Note that (2.14) is slightly more general than (2.1). Moreover, we assume that both A and B are deterministic nonnegative definite symmetric matrices satisfying (2.4) and (2.6), $\widetilde{A}$ and $\widetilde{B}$ are deterministic nonnegative definite symmetric matrices satisfying (2.7), (2.8), (2.9) and (2.10) and d_n satisfies (2.3).

2.2. *Resolvents and limiting laws.* In this paper, we study the eigenvalue statistics of $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\widetilde{\mathcal{Q}}_1$, $\widetilde{\mathcal{Q}}_2$ through their *resolvents* (or *Green's functions*). Throughout the paper, we shall denote the upper half complex plane and the right-half real line by

$$\mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}, \qquad \mathbb{R}^+ := [0, \infty).$$

DEFINITION 2.3 (Resolvents). For $z = E + \mathrm{i}\eta \in \mathbb{C}_+$, we define the following resolvents for $\alpha = 1, 2$:

$$\mathcal{G}_\alpha(X, z) := \big(\mathcal{Q}_\alpha(X) - z\big)^{-1}, \qquad \widetilde{\mathcal{G}}_\alpha(X, z) := \big(\widetilde{\mathcal{Q}}_\alpha(X) - z\big)^{-1}. \tag{2.15}$$

We denote the ESD $\rho^{(p)}$ of $\mathcal{Q}_1$ and its Stieltjes transform as

$$\rho^{(p)} := \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\mathcal{Q}_1)}, \qquad m^{(n)}(z) := \int \frac{\rho^{(p)}(\mathrm{d}x)}{x - z} = \frac{1}{p}\operatorname{Tr}\mathcal{G}_1(z). \tag{2.16}$$

It was shown in [48] that if $d_n \to d \in (0, \infty)$ and $\pi_A^{(p)}, \pi_B^{(n)}$ converge to certain probability distributions, then almost surely $\rho^{(p)}$ converges to a deterministic distributions ρ_∞. We now give its definition. For any finite n, $p = nd_n$ and $z \in \mathbb{C}_+$, we define $(m_{1c}^{(n)}(z), m_{2c}^{(n)}(z)) \in \mathbb{C}_+^2$ as the unique solution to the following system of self-consistent equations

$$\begin{aligned} m_{1c}^{(n)}(z) &= d_n \int \frac{x}{-z[1 + x m_{2c}^{(n)}(z)]} \pi_A^{(p)}(\mathrm{d}x), \\ m_{2c}^{(n)}(z) &= \int \frac{x}{-z[1 + x m_{1c}^{(n)}(z)]} \pi_B^{(n)}(\mathrm{d}x). \end{aligned} \tag{2.17}$$

Then we define

$$m_c(z) \equiv m_c^{(n)}(z) := \int \frac{1}{-z[1 + x m_{2c}^{(n)}(z)]} \pi_A^{(p)}(\mathrm{d}x). \tag{2.18}$$

It is easy to verify that $m_c(z) \in \mathbb{C}_+$ for $z \in \mathbb{C}_+$. Letting $\eta \downarrow 0$, we can obtain a probability measure $\rho_c^{(n)}$ with the inverse formula

$$\rho_c^{(n)}(E) = \lim_{\eta \downarrow 0} \frac{1}{\pi} \operatorname{Im} m_c^{(n)}(E + \mathrm{i}\eta). \tag{2.19}$$

If $d_n \to d \in (0, \infty)$ and $\pi_A^{(p)}$, $\pi_B^{(n)}$ converge to certain probability distributions, then $\rho_c^{(n)}$ converges weakly as $n \to \infty$, and its weak limit is ρ_∞.

The above definitions of $m_c^{(n)}$, $\rho_c^{(n)}$ and ρ_∞ make sense due to the following theorem. Throughout the rest of this paper, we often omit the super-indices (p) and (n) from our notations for simplicity.

THEOREM 2.4 (Existence, uniqueness and continuous density). *For any $z \in \mathbb{C}_+$, there exists a unique solution $(m_{1c}, m_{2c}) \in \mathbb{C}_+^2$ to the systems of equations in (2.17). The function m_c in (2.18) is the Stieltjes transform of a probability measure μ_c supported on $\mathbb{R}^+$. Moreover, μ_c has a continuous derivative $\rho_c(x)$ on $(0, \infty)$.*

PROOF. See [60], Theorem 1.2.1, [23], Theorem 2.4, and [8], Theorem 3.1. □

From (2.17), it is easy to see that if we define the function

$$f(z, m) := -m + \int \frac{x}{-z + x d_n \int \frac{t}{1+tm} \pi_A(\mathrm{d}t)} \pi_B(\mathrm{d}x), \tag{2.20}$$

then $m_{2c}(z)$ can be characterized as the unique solution to the equation $f(z, m) = 0$ that satisfies $\operatorname{Im} m > 0$ for $z \in \mathbb{C}_+$, and $m_{1c}(z)$ can be defined using the first equation in (2.17). Moreover, $m_{1c}(z)$ and $m_{2c}(z)$ are the Stieltjes transforms of the densities ρ_{1c} and ρ_{2c}:

$$\rho_{\alpha c}(E) = \lim_{\eta \downarrow 0} \frac{1}{\pi} \operatorname{Im} m_{\alpha c}(E + \mathrm{i}\eta), \quad \alpha = 1, 2. \tag{2.21}$$

Then we have the following result.

LEMMA 2.5. *The densities ρ_c, ρ_{1c} and ρ_{2c} all have the same support on $(0, \infty)$, which is a union of intervals: for $\alpha = 1, 2$,*

$$\operatorname{supp} \rho_c \cap (0, \infty) = \operatorname{supp} \rho_{\alpha c} \cap (0, \infty) = \bigcup_{k=1}^{L} [e_{2k}, e_{2k-1}] \cap (0, \infty), \tag{2.22}$$

where $L \in \mathbb{N}$ depends only on $\pi_{A,B}$. Moreover, $(x, m) = (e_k, m_{2c}(e_k))$ are the real solutions to the equations

$$f(x, m) = 0, \quad \text{and} \quad \frac{\partial f}{\partial m}(x, m) = 0. \tag{2.23}$$

Finally, we have $e_1 = \mathrm{O}(1)$, $m_{1c}(e_1) \in (-(\max_\mu \sigma_\mu^b)^{-1}, 0)$ and $m_{2c}(e_1) \in (-(\max_i \sigma_i^a)^{-1}, 0)$.

PROOF. See Section 3 of [8]. □

We shall call e_k the spectral edges. In particular, we focus on the rightmost edge $\lambda_+ := e_1$. Now we make the following assumption. It guarantees a regular square-root behavior of the spectral densities ρ_{1c} and ρ_{2c} near λ_+ and rules out the existence of outliers.

ASSUMPTION 2.6. There exists a constant $\tau > 0$ such that

$$1 + m_{1c}(\lambda_+) \max_\mu \sigma_\mu^b \geq \tau, \qquad 1 + m_{2c}(\lambda_+) \max_i \sigma_i^a \geq \tau. \tag{2.24}$$

3. Main results. In this section, we state the main results on the eigenvalues and eigenvectors of $\widetilde{\mathcal{Q}}_1$ and $\widetilde{\mathcal{Q}}_2$, together with some interpretations of these results. Their proof will be presented in the Supplementary Material.

Throughout this paper, we use the words *spikes* and *spiked eigenvectors* for those of the population matrices $\widetilde{A}$ and $\widetilde{B}$. Meanwhile, we shall use the words *outlier eigenvalues* and *outlier eigenvectors* for those of the sample separable covariance matrices $\widetilde{\mathcal{Q}}_1$ and $\widetilde{\mathcal{Q}}_2$.

We will see that a spike $\widetilde{\sigma}_i^a$, $1 \le i \le r$, or $\widetilde{\sigma}_\mu^b$, $1 \le \mu \le s$, causes an outlier eigenvalue beyond λ_+, if

$$\widetilde{\sigma}_i^a > -m_{2c}^{-1}(\lambda_+) \quad \text{or} \quad \widetilde{\sigma}_\mu^b > -m_{1c}^{-1}(\lambda_+), \tag{3.1}$$

where $m_{1c}(\cdot)$ and $m_{2c}(\cdot)$ are defined in (2.17). Moreover, such an outlier is around a deterministic location

$$\theta_1(\widetilde{\sigma}_i^a) := g_{2c}\big(-(\widetilde{\sigma}_i^a)^{-1}\big) \quad \text{or} \quad \theta_2(\widetilde{\sigma}_\mu^b) := g_{1c}\big(-(\widetilde{\sigma}_\mu^b)^{-1}\big), \tag{3.2}$$

where g_{1c} and g_{2c} are the inverse functions of $m_{1c} : (\lambda_+, \infty) \to (m_{1c}(\lambda_+), 0)$ and $m_{2c} : (\lambda_+, \infty) \to (m_{2c}(\lambda_+), 0)$, respectively. Note that the inverse functions exist because

$$m_{\alpha c}(x) = \int_0^{\lambda_+} \frac{\rho_{\alpha c}(t)}{t - x}\,\mathrm{d}t, \quad x > \lambda_+, \alpha = 1, 2, \tag{3.3}$$

are monotonically increasing functions of x for $x > \lambda_+$.

For X, we introduce the following bounded support condition.

DEFINITION 3.1 (Bounded support condition). We say a random matrix X satisfies the *bounded support condition* with ϕ_n if

$$\max_{i,j} |x_{ij}| \le \phi_n, \tag{3.4}$$

where ϕ_n is a deterministic parameter and usually satisfies $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some (small) constant $c_\phi > 0$. Whenever (3.4) holds, we say that X has support ϕ_n.

The main reason for introducing this notation is as following: for a random matrix X whose entries have at least $(4 + \varepsilon)$-moments, it can be reduced to a random matrix with bounded support with probability $1 - o(1)$ using a standard cut-off argument; see Corollary 3.19 below.

ASSUMPTION 3.2. We assume that (3.1) holds for all $1 \le i \le r$ and $1 \le \mu \le s$. Otherwise, if (3.1) fails for some $\widetilde{\sigma}_i^a$ or $\widetilde{\sigma}_\mu^b$, we can simply redefine it as the unperturbed version σ_i^a or σ_μ^b. Moreover, we define the integers $0 \le r^+ \le r$ and $0 \le s^+ \le s$ such that

$$\widetilde{\sigma}_i^a \ge -m_{2c}^{-1}(\lambda_+) + n^{-1/3} + \phi_n \quad \text{if and only if} \quad 1 \le i \le r^+, \tag{3.5}$$

and

$$\widetilde{\sigma}_\mu^b \ge -m_{1c}^{-1}(\lambda_+) + n^{-1/3} + \phi_n \quad \text{if and only if} \quad 1 \le \mu \le s^+. \tag{3.6}$$

The lower bound $n^{-1/3} + \phi_n$ is chosen for definiteness, and it can be replaced with any n-dependent parameter that is of the same order.

REMARK 3.3. Consider the case where $\phi_n \le n^{-1/3}$ (this holds if we assume the existence of 12th moment). A spike $\widetilde{\sigma}_i^a$ or $\widetilde{\sigma}_\mu^b$ that does not satisfy (3.5) or (3.6) will give an outlier that lies within an $O(n^{-2/3})$ neighborhood of the rightmost edge λ_+. It is essentially indistinguishable from the extremal eigenvalue of $\mathcal{Q}_1$, which has typical fluctuation of order $n^{-2/3}$ around λ_+. Hence in (3.5) and (3.6), we simply choose the "real" spikes of $\widetilde{A}$ and $\widetilde{B}$.

We will use the following notion of stochastic domination, which was first introduced in [18] and subsequently used in many works on random matrix theory, such as [5–7, 19, 20, 28]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded by ζ with high probability up to a small power of n."

DEFINITION 3.4 (Stochastic domination).

(i) Let

$$\xi = \big(\xi^{(n)}(u) : n \in \mathbb{N}, u \in U^{(n)}\big), \qquad \zeta = \big(\zeta^{(n)}(u) : n \in \mathbb{N}, u \in U^{(n)}\big)$$

be two families of nonnegative random variables, where $U^{(n)}$ is a possibly n-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any fixed (small) $\varepsilon > 0$ and (large) $D > 0$,

$$\sup_{u \in U^{(n)}} \mathbb{P}\big(\xi^{(n)}(u) > n^{\varepsilon}\zeta^{(n)}(u)\big) \le n^{-D}$$

for large enough $n \ge n_0(\varepsilon, D)$, and we shall use the notation $\xi \prec \zeta$. Throughout this paper, the stochastic domination will always be uniform in all parameters that are not explicitly fixed (such as matrix indices, and z that takes values in some compact set). Note that $n_0(\varepsilon, D)$ may depend on quantities that are explicitly constant, such as τ in Assumption 2.2 and (2.24). If for some complex family ξ we have $|\xi| \prec \zeta$, then we will also write $\xi \prec \zeta$ or $\xi = O_{\prec}(\zeta)$.

(ii) We extend the definition of $O_{\prec}(\cdot)$ to matrices in the weak operator norm sense as follows. Let A be a family of random matrices and ζ be a family of nonnegative random variables. Then $A = O_{\prec}(\zeta)$ means that $|\langle \mathbf{v}, A\mathbf{w}\rangle| \prec \zeta \|\mathbf{v}\|_2 \|\mathbf{w}\|_2$ uniformly in any deterministic vectors $\mathbf{v}$ and $\mathbf{w}$. Here and throughout the following, whenever we say "uniformly in any deterministic vectors," we mean that "uniformly in any deterministic vectors belonging to a set of cardinality $n^{O(1)}$."

(iii) We say an event Ξ holds with high probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \ge 1 - n^{-D}$ for large enough n.

3.1. *Eigenvalue statistics.* In this subsection, we describe the results on the sample eigenvalues. To state our result on the outlier eigenvalues, we first introduce the following labeling of such outliers.

DEFINITION 3.5. We define the labelling functions $\alpha : \{1, \dots, p\} \to \mathbb{N}$ and $\beta : \{1, \dots, n\} \to \mathbb{N}$ as follows. For any $1 \le i \le r$, we assign to it a label $\alpha(i) \in \{1, \dots, r+s\}$ if $\theta_1(\widetilde{\sigma}_i^a)$ is the $\alpha(i)$th largest element in $\{\theta_1(\widetilde{\sigma}_i^a)\}_{i=1}^r \cup \{\theta_2(\widetilde{\sigma}_\mu^b)\}_{\mu=1}^s$. We also assign to any $1 \le \mu \le s$ a label $\beta(\mu) \in \{1, \dots, r+s\}$ in a similar way. Moreover, we define $\alpha(i) = i + s$ if $i > r$ and $\beta(\mu) = \mu + r$ if $\mu > s$. We define the following sets of outlier indices:

$$\mathcal{O} := \{\alpha(i) : 1 \le i \le r\} \cup \{\beta(\mu) : 1 \le \mu \le s\},$$

and

$$\mathcal{O}^+ := \{\alpha(i) : 1 \le i \le r^+\} \cup \{\beta(\mu) : 1 \le \mu \le s^+\}.$$

We first state the results on the locations of the outlier and the first few nonoutlier eigenvalues. Denote the nontrivial eigenvalues of $\widetilde{\mathcal{Q}}_{1,2}$ by $\widetilde{\lambda}_1 \ge \widetilde{\lambda}_2 \ge \cdots \ge \widetilde{\lambda}_{n\wedge p}$. For $1 \le i \le r$ and $1 \le \mu \le s$, we define

$$\Delta_1(\widetilde{\sigma}_i^a) := \big(\widetilde{\sigma}_i^a + m_{2c}^{-1}(\lambda_+)\big)^{1/2}, \qquad \Delta_2(\widetilde{\sigma}_\mu^b) := \big(\widetilde{\sigma}_\mu^b + m_{1c}^{-1}(\lambda_+)\big)^{1/2}. \tag{3.7}$$

THEOREM 3.6. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that Assumptions 2.2, 2.6 and 3.2 hold. Then we have*

$$|\widetilde{\lambda}_{\alpha(i)} - \theta_1(\widetilde{\sigma}_i^a)| \prec n^{-1/2}\Delta_1(\widetilde{\sigma}_i^a) + \phi_n\Delta_1^2(\widetilde{\sigma}_i^a), \quad 1 \le i \le r^+, \tag{3.8}$$

and

$$|\widetilde{\lambda}_{\beta(\mu)} - \theta_2(\widetilde{\sigma}_\mu^b)| \prec n^{-1/2}\Delta_2(\widetilde{\sigma}_\mu^b) + \phi_n\Delta_2^2(\widetilde{\sigma}_\mu^b), \quad 1 \le \mu \le s^+. \tag{3.9}$$

Furthermore, for any fixed integer $\varpi > r + s$, we have

$$|\widetilde{\lambda}_i - \lambda_+| \prec n^{-2/3} + \phi_n^2, \quad \text{for } i \notin \mathcal{O}^+ \text{ and } i \le \varpi. \tag{3.10}$$

The above theorem gives the large deviation bounds for the locations of the outliers and the first few extremal nonoutlier eigenvalues. Again consider the case with $\phi_n \le n^{-1/3}$. Then Theorem 3.6 shows that the fluctuation of the outlier changes from the order $n^{-1/2}\Delta_1(\widetilde{\sigma}_i^a)$ to $n^{-2/3}$ when $\Delta_1(\widetilde{\sigma}_i^a)$ or $\Delta_2(\widetilde{\sigma}_\mu^b)$ crosses the scale $n^{-1/6}$. This implies the occurrence of the BBP transition [2]. In a future work, we will show that under certain assumptions, the outlier eigenvalues are normally distributed, whereas the extremal nonoutlier eigenvalues follow the Tracy–Widom law.

Next, we study the nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$. We prove that the eigenvalues of $\widetilde{\mathcal{Q}}_1$ for $i > r^+ + s^+$ are governed by *eigenvalue sticking*, which states that the nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$ "stick" with high probability to the eigenvalues of the reference matrix $\mathcal{Q}_1$. Recall that we denote the eigenvalues of $\mathcal{Q}_1$ as $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{p\wedge n}$.

THEOREM 3.7. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that Assumptions 2.2, 2.6 and 3.2 hold. We define*

$$\alpha_+ := \min\Big\{\min_i|\widetilde{\sigma}_i^a + m_{2c}^{-1}(\lambda_+)|, \min_\mu|\widetilde{\sigma}_\mu^b + m_{1c}^{-1}(\lambda_+)|\Big\}. \tag{3.11}$$

Assume that $\alpha_+ \ge n^{c_0}\phi_n$ for some constant $c_0 > 0$. Fix any sufficiently small constant $\tau > 0$. We have that for $1 \le i \le \tau n$,

$$|\widetilde{\lambda}_{i+r^++s^+} - \lambda_i| \prec \frac{1}{n\alpha_+} + n^{-3/4} + i^{1/3}n^{-5/6} + n^{-1/2}\phi_n + i^{-2/3}n^{-1/3}\phi_n^2. \tag{3.12}$$

If either (a) the third moments of the entries of X vanish in the sense that

$$\mathbb{E}x_{ij}^3 = 0, \quad 1 \le i \le p, 1 \le j \le n, \tag{3.13}$$

or (b) either A or B is diagonal, then we have the stronger estimate

$$|\widetilde{\lambda}_{i+r^++s^+} - \lambda_i| \prec \frac{1}{n\alpha_+}, \quad 1 \le i \le \tau n. \tag{3.14}$$

Theorem 3.7 establishes the large deviation bounds for the nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$ with respect to the eigenvalues of $\mathcal{Q}_1$. In particular, when $\alpha_+ \gg n^{-1/3}$ and $\phi_n \ll n^{-1/6}$, the right-hand side of (3.12) or (3.14) is much smaller than $n^{-2/3}$ for $i = O(1)$. In fact, it was proved in [58] that the limiting joint distribution of the first few eigenvalues $\{\lambda_i\}_{1\le i\le k}$ of $\mathcal{Q}_1$ is universal under an $n^{2/3}$ scaling for any fixed $k \in \mathbb{N}$. Together with (3.12), this implies that the limiting distribution of the largest nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$ is also universal under an $n^{2/3}$ scaling as long as $\alpha_+ \gg n^{-1/3}$ and $\phi_n \ll n^{-1/6}$. In a future paper, we will prove that $\{n^{2/3}(\lambda_i - \lambda_+)\}_{1\le i\le k}$ converges to the Tracy–Widom law for any fixed $k \in \mathbb{N}$, which immediately implies that the largest nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$ also satisfy the Tracy–Widom law.

REMARK 3.8. The Theorems 3.6 and 3.7 can be combined to potentially estimate the spikes of $\widetilde{A}$ and $\widetilde{B}$ if they are low-rank perturbations of identity matrices. By Theorem 3.6, the spike $\widetilde{\sigma}_i^a$ or $\widetilde{\sigma}_\mu^b$ can be effectively estimated using $-m_{2c}^{-1}(\widetilde{\lambda}_{\alpha(i)})$ or $-m_{1c}^{-1}(\widetilde{\lambda}_{\beta(\mu)})$. Although calculating m_{1c} and m_{2c} needs the knowledge of the spectrums of A and B, we will see that m_{1c} and m_{2c} can be well approximated using the eigenvalues of $\widetilde{\mathcal{Q}}_1$ and $\widetilde{\mathcal{Q}}_2$ only. We record such result in Theorem 4.5.

On the other hand, for the nonspiked eigenvalues, to our best knowledge there does not exist any literature on the estimation of the spectrums of general A and B using the eigenvalues of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ only. However, for sample covariance matrices with $B = I$, the spectrum of A can be estimated using the eigenvalues of $A^{1/2}XX^*A^{1/2}$ by solving a convex optimization problem involving the self-consistent equation for m_{2c} in [17, 29]. In the future work, we will try to generalize their results to the separable covariance matrices with more general B. Note that although we cannot observe the eigenvalues of $\mathcal{Q}_1$, Theorem 3.7 implies that the nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$ are close to those of $\mathcal{Q}_1$.

REMARK 3.9. We have seen from Theorem 3.6 that the locations of the outlier eigenvalues depend on the spikes and the spectrums of both A and B. Consider the case with $r = s = 1$ and supercritical spikes (c.f. Assumption 4.1). By (3.8), we see that the outlier locations depend on the 4-tuple $(\widetilde{\sigma}^a, \widetilde{\sigma}^b, \boldsymbol{\sigma}(A), \boldsymbol{\sigma}(B))$, where $\widetilde{\sigma}^a$ and $\widetilde{\sigma}^b$ are the spikes associated with A and B, respectively, and $\boldsymbol{\sigma}(A)$ and $\boldsymbol{\sigma}(B)$ denote the spectrums of A and B. In general, the 4-tuple is not jointly identifiable. Indeed, even the pair $(\boldsymbol{\sigma}(A), \boldsymbol{\sigma}(B))$ is not jointly identifiable [37].

To handle this issue, one needs to impose some constraints. For instance, when $B = I_n$, $\widetilde{\sigma}^a$ can be efficiently estimated using the eigenvalues of $\widetilde{\mathcal{Q}}_1$ by Theorem 4.5. Moreover, as mentioned in Remark 3.8, the spectrum of A can be estimated using the methods mentioned in [17, 29, 32]. In this situation, $(\widetilde{\sigma}^a, \boldsymbol{\sigma}(A))$ is identifiable. More generally, assume we know that the two triplets $(\widetilde{\sigma}_\alpha^a, \boldsymbol{\sigma}(A_\alpha), \boldsymbol{\sigma}(B))$ and $(\widetilde{\sigma}_\beta^a, \boldsymbol{\sigma}(A_\beta), \boldsymbol{\sigma}(B))$ share the same temporal covariance matrix B. Then using their sample eigenvalues $\{\widetilde{\lambda}_k^\alpha\}$ and $\{\widetilde{\lambda}_k^\beta\}$, we can employ the following two-step procedure to check whether they are identifiable.

Step (i): Checking whether they have the same number of outliers and whether the outliers share the same values. More precisely given a threshold $\omega \to 0$, we need to check whether $|\widetilde{\lambda}_k^\alpha - \widetilde{\lambda}_k^\beta| \le \omega$, $1 \le k \le r$, where r is the number of outliers. If this does not hold true, then the two triples are different according to Theorem 3.6. Otherwise, we continue with the second step.

Step (ii): Checking whether the spectrums of A_α and A_β are the same. In fact, the eigenvalues of $\mathcal{Q}_1$ are determined by the spectrums of A and B; see the eigenvalues rigidity result, Theorem C.11, in the Supplementary Material [12]. Then with Theorem 3.7, if $\boldsymbol{\sigma}(A_\alpha) = \boldsymbol{\sigma}(A_\beta)$, we should have $|\widetilde{\lambda}_k^\alpha - \widetilde{\lambda}_k^\beta| \le \omega$, $k \ge r + 1$, for the nonoutliers. If this does not hold true, we claim that these two triplets are different.

Finally, we mention that for a rigorous statement of the above hypothesis testing on whether $(\widetilde{\sigma}_\alpha^a, \boldsymbol{\sigma}(A_\alpha), \boldsymbol{\sigma}(B))$ and $(\widetilde{\sigma}_\beta^a, \boldsymbol{\sigma}(A_\beta), \boldsymbol{\sigma}(B))$ are the same, we need to derive the second order asymptotics of the eigenvalues. This will be our future work.

3.2. *Eigenvector statistics*. In this subsection, we state the results on the eigenvectors of $\widetilde{\mathcal{Q}}_1$ and $\widetilde{\mathcal{Q}}_2$. We denote the eigenvectors of $\widetilde{\mathcal{Q}}_1$ by $\widetilde{\boldsymbol{\xi}}_k$, $1 \le k \le p$, and the eigenvectors of $\widetilde{\mathcal{Q}}_2$ by $\widetilde{\boldsymbol{\zeta}}_\mu$, $1 \le \mu \le n$. To remove the arbitrariness in the definitions of eigenvectors, we shall consider instead the products of generalized components

$$\langle \mathbf{v}, \widetilde{\boldsymbol{\xi}}_k \rangle \langle \widetilde{\boldsymbol{\xi}}_k, \mathbf{w} \rangle, \qquad \langle \mathbf{v}', \widetilde{\boldsymbol{\zeta}}_k \rangle \langle \widetilde{\boldsymbol{\zeta}}_k, \mathbf{w}' \rangle,$$

where $\mathbf{v}$, $\mathbf{w}$, $\mathbf{v}'$ and $\mathbf{w}'$ are some given deterministic vectors. Note that these products characterize the eigenvectors $\widetilde{\boldsymbol{\xi}}_k$ and $\widetilde{\boldsymbol{\zeta}}_k$ completely up to the ambiguity of a phase. More generally, if we consider degenerate or near-degenerate outliers, then only eigenspace matters. Here, the degenerate (or near-degenerate) outliers refer to the outliers corresponding to identical (or near-degenerate) population spikes. As in [6], we shall consider the generalized components $\langle \mathbf{v}, \mathcal{P}_S \mathbf{w}\rangle$ of the random projection

$$\mathcal{P}_S := \sum_{k\in S} \widetilde{\boldsymbol{\xi}}_k \widetilde{\boldsymbol{\xi}}_k^*, \quad \text{for } S \subset \mathcal{O}^+.$$

In particular, in the nondegenerate case $S = \{k\}$, the generalized components of $\mathcal{P}_S$ are the products of the generalized components of $\widetilde{\boldsymbol{\xi}}_k$.

For $1 \le i \le r^+$, $1 \le j \le p$ and $1 \le \nu \le n$, we define

$$\delta^a_{\alpha(i),\alpha(j)} := |\widetilde{\sigma}^a_j - \widetilde{\sigma}^a_i|, \qquad \delta^a_{\alpha(i),\beta(\nu)} := |\widetilde{\sigma}^b_\nu + m_{1c}^{-1}(\theta_1(\widetilde{\sigma}^a_i)|. \tag{3.15}$$

Similarly, for $1 \le \mu \le s^+$, $1 \le j \le p$ and $1 \le \nu \le n$, we define

$$\delta^b_{\beta(\mu),\alpha(j)} := |\widetilde{\sigma}^a_j + m_{2c}^{-1}(\theta_2(\widetilde{\sigma}^b_\mu))|, \qquad \delta^b_{\beta(\mu),\beta(\nu)} := |\widetilde{\sigma}^b_\nu - \widetilde{\sigma}^b_\mu|. \tag{3.16}$$

Given any $S \subset \mathcal{O}^+$, if $\mathfrak{a} \in S$, then we define

$$\delta_{\mathfrak{a}}(S) := \begin{cases} \Big(\min\limits_{k:\alpha(k)\notin S} \delta^a_{\mathfrak{a},\alpha(k)}\Big) \wedge \Big(\min\limits_{\mu:\beta(\mu)\notin S} \delta^a_{a,\beta(\mu)}\Big), & \text{if } \mathfrak{a} = \alpha(i) \in S, \\ \Big(\min\limits_{k:\alpha(k)\notin S} \delta^b_{\mathfrak{a},\alpha(k)}\Big) \wedge \Big(\min\limits_{\mu:\beta(\mu)\notin S} \delta^b_{\mathfrak{a},\beta(\mu)}\Big), & \text{if } \mathfrak{a} = \beta(\mu) \in S; \end{cases}$$

if $\mathfrak{a} \notin S$, then we define

$$\delta_{\mathfrak{a}}(S) := \Big(\min_{k:\alpha(k)\in S} \delta^a_{\alpha(k),\mathfrak{a}}\Big) \wedge \Big(\min_{\mu:\beta(\mu)\in S} \delta^b_{\beta(\mu),\mathfrak{a}}\Big).$$

We now state the results on the left outlier singular vectors of $\widetilde{A}^{1/2} X \widetilde{B}^{1/2}$, that is, the outlier eigenvectors of $\widetilde{\mathcal{Q}}_1$.

THEOREM 3.10. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that Assumptions 2.2, 2.6 and 3.2 hold. Fix any $S \subset \mathcal{O}^+$, we define the following deterministic positive quadratic form*

$$\langle \mathbf{v}, \mathcal{Z}_S \mathbf{v}\rangle := \sum_{i:\alpha(i)\in S} \frac{|v_i|^2}{\widetilde{\sigma}^a_i} \frac{g'_{2c}(-(\widetilde{\sigma}^a_i)^{-1})}{g_{2c}(-(\widetilde{\sigma}^a_i)^{-1})}, \quad \text{for } \mathbf{v} \in \mathbb{C}^p,\ v_i := \langle \mathbf{v}^a_i, \mathbf{v}\rangle. \tag{3.17}$$

Then for any deterministic vector $\mathbf{v} \in \mathbb{C}^p$, we have that

$$\begin{aligned} &\big|\langle \mathbf{v}, \mathcal{P}_S \mathbf{v}\rangle - \langle \mathbf{v}, \mathcal{Z}_S \mathbf{v}\rangle\big| \\ &\prec \sum_{1\le i\le r:\alpha(i)\in S} |v_i|^2 \psi_1(\widetilde{\sigma}^a_i) \\ &\quad + \sum_{1\le i\le r:\alpha(i)\notin S} |v_i|^2 \frac{\phi_n^2}{\delta_{\alpha(i)}(S)} + \sum_{i=1}^p |v_i|^2 \Big(\frac{\psi_1^2(\widetilde{\sigma}^a_i)\Delta_1^2(\widetilde{\sigma}^a_i)}{\delta^2_{\alpha(i)}(S)} + \frac{\kappa_i}{n^{1/2}} \Big) \\ &\quad + \langle \mathbf{v}, \mathcal{Z}_S \mathbf{v}\rangle^{1/2} \Big[\sum_{1\le i\le r:\alpha(i)\notin S} |v_i|^2 \frac{\phi_n^2}{\delta_{\alpha(i)}(S)} \\ &\quad + \sum_{1\le i\le p:\alpha(i)\notin S} |v_i|^2 \Big(\frac{\psi_1^2(\widetilde{\sigma}^a_i)\Delta_1^2(\widetilde{\sigma}^a_i)}{\delta^2_{\alpha(i)}(S)} + \frac{\kappa_i}{n^{1/2}} \Big) \Big]^{1/2}, \end{aligned} \tag{3.18}$$

where we denote

$$\psi_1(\widetilde{\sigma}_i^a) := \phi_n + n^{-1/2}\Delta_1^{-1}(\widetilde{\sigma}_i^a).$$

If we have (a) (3.13) *holds, or* (b) *either A or B is diagonal, then the above estimate holds without the* $n^{-1/2}\kappa_i$ *terms.*

REMARK 3.11. For any deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^p$, we can state Theorem 3.10 for more general quantities of the form $\langle \mathbf{v}, \mathcal{Z}_S \mathbf{w} \rangle$ using the polarization identity. Moreover, $\mathcal{Z}_S$ is a matrix that is uniquely determined by the quadratic form in (3.17). It can be written as

$$\mathcal{Z}_S = \sum_{i:\alpha(i)\in S} \mathbf{v}_i^a (\mathbf{v}_i^a)^* \frac{1}{\widetilde{\sigma}_i^a} \frac{g_{2c}'(-(\widetilde{\sigma}_i^a)^{-1})}{g_{2c}(-(\widetilde{\sigma}_i^a)^{-1})}.$$

The index set S in Theorem 3.10 can be chosen according to user's goal. We now consider two typical cases to illustrate the idea.

EXAMPLE 3.12 (Nondegenerate case). If all the outliers are well separated, then we can choose $S = \{\alpha(i)\}$ or $S = \{\beta(\mu)\}$. For example, suppose $S = \{\alpha(i)\}$ and $\mathbf{v} = \mathbf{v}_i^a$. Denote $\delta_{\alpha(i)} := \delta_{\alpha(i)}(\{\alpha(i)\})$. Then we get from (3.18) that

$$|\langle \mathbf{v}_i^a, \widetilde{\boldsymbol{\xi}}_{\alpha(i)} \rangle|^2 = \frac{1}{\widetilde{\sigma}_i^a} \frac{g_{2c}'(-(\widetilde{\sigma}_i^a)^{-1})}{g_{2c}(-(\widetilde{\sigma}_i^a)^{-1})} + O_{\prec}\bigg(\psi_1(\widetilde{\sigma}_i^a) + \frac{\psi_1^2(\widetilde{\sigma}_i^a)\Delta_1^2(\widetilde{\sigma}_i^a)}{n\delta_{\alpha(i)}^2}\bigg).$$

Note that $\widetilde{\boldsymbol{\xi}}_i$ is concentrated on a cone with axis parallel to $\mathbf{v}_i^a$ if the error term is much smaller than the first term, which is of order

$$\frac{1}{\widetilde{\sigma}_i^a} \frac{g_{2c}'(-(\widetilde{\sigma}_i^a)^{-1})}{g_{2c}(-(\widetilde{\sigma}_i^a)^{-1})} \sim \widetilde{\sigma}_i^a + m_{2c}^{-1}(\lambda_+)$$

by Lemma C.6 in the Supplementary Material. This leads to the following conditions:

$$\widetilde{\sigma}_i^a + m_{2c}^{-1}(\lambda_+) \gg \phi_n + n^{-1/3}, \qquad \delta_{\alpha(i)} \gg \phi_n + n^{-1/2}\Delta_1^{-1}(\widetilde{\sigma}_i^a). \tag{3.19}$$

The first condition means that $\widetilde{\lambda}_{\alpha(i)}$ is truly an outlier (cf. Theorem 3.6), whereas the second condition is a *nonoverlapping condition*. In fact, by (3.8), $\widetilde{\lambda}_{\alpha(i)}$ fluctuates around $\theta_1(\widetilde{\sigma}_i^a)$ on the scale of order $n^{-1/2}\Delta_1(\widetilde{\sigma}_i^a) + \phi_n\Delta_1^2(\widetilde{\sigma}_i^a)$. Therefore, $\widetilde{\lambda}_{\alpha(i)}$ is well separated from the other outlier eigenvalues if

$$\begin{aligned} &\Big(\min_{\alpha(j)\in\mathcal{O}\setminus\{\alpha(i)\}} |\theta_1(\widetilde{\sigma}_i^a) - \theta_1(\widetilde{\sigma}_j^a)|\Big) \wedge \Big(\min_{\beta(\mu)\in\mathcal{O}} |\theta_1(\widetilde{\sigma}_i^a) - \theta_2(\widetilde{\sigma}_\mu^b)|\Big) \\ &\quad \gg n^{-1/2}\Delta_1(\widetilde{\sigma}_i^a) + \phi_n\Delta_1^2(\widetilde{\sigma}_i^a). \end{aligned} \tag{3.20}$$

Moreover, by Lemma C.6 in the Supplementary Material, the left-hand side of (3.20) is of order $\delta_{\alpha(i)}\Delta_1^2(\widetilde{\sigma}_i^a)$. This gives the second condition in (3.19).

For degenerate or near-degenerate outliers, their indices should be included in the same set S. We now consider an example with multiple outliers that share exactly the same classical location.

EXAMPLE 3.13 (Degenerate case). Suppose that we have an $|S|$-fold degenerate outlier, that is, for some $\theta_0 > \lambda_+$,

$$\theta_1(\widetilde{\sigma}_i^a) = \theta_2(\widetilde{\sigma}_\mu^b) = \theta_0, \quad \text{for all } \alpha(i), \beta(\mu) \in S.$$

Suppose the outlier θ_0 is well separated from both the bulk and the other outliers (i.e., with distances of order 1). Then by (3.18), we have that

$$\mathcal{P}_S = \sum_{\alpha(i)\in S} \frac{1}{\widetilde{\sigma}_i^a} \frac{g_{2c}'(-(\widetilde{\sigma}_i^a)^{-1})}{g_{2c}(-(\widetilde{\sigma}_i^a)^{-1})} \mathbf{v}_i^a (\mathbf{v}_i^a)^* + \mathcal{E},$$

where $\mathcal{E}$ is an error that is delocalized in the basis of $\mathbf{v}_i^a$, that is, $\langle \mathbf{v}_i^a, \mathcal{E}\mathbf{v}_j^a\rangle \prec \phi_n$. This can be regarded as a generalized cone concentration for the subspace spanned by $\{\widetilde{\boldsymbol{\xi}}_{\mathfrak{a}}\}_{\mathfrak{a}\in S}$.

Then we state the delocalization results on the nonoutlier eigenvectors when $\alpha(i) \notin \mathcal{O}^+$. Denote

$$\eta_i := n^{-3/4} + n^{-5/6} i^{1/3} + n^{-1/2}\phi_n, \qquad \kappa_i := i^{2/3} n^{-2/3}.$$

THEOREM 3.14. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that Assumptions 2.2, 2.6 and 3.2 hold. Fix any sufficiently small constant $\tau > 0$. For $\alpha(i) \notin \mathcal{O}^+$, $i \le \tau p$ and any deterministic vector $\mathbf{v} \in \mathbb{C}^p$, we have*

$$|\langle \mathbf{v}, \widetilde{\boldsymbol{\xi}}_{\alpha(i)}\rangle|^2 \prec \sum_{j=1}^{p} |v_j|^2 \frac{n^{-1} + \eta_i\sqrt{\kappa_i} + \phi_n^3}{|\widetilde{\sigma}_j^a + m_{2c}^{-1}(\lambda_+)|^2 + \phi_n^2 + \kappa_i}. \tag{3.21}$$

If we have (a) (3.13) *holds, or* (b) *either A or B is diagonal, then the following stronger estimate holds*:

$$|\langle \mathbf{v}, \widetilde{\boldsymbol{\xi}}_{\alpha(i)}\rangle|^2 \prec \sum_{j=1}^{p} |v_j|^2 \frac{n^{-1} + \phi_n^3}{|\widetilde{\sigma}_j^a + m_{2c}^{-1}(\lambda_+)|^2 + \phi_n^2 + \kappa_i}. \tag{3.22}$$

REMARK 3.15. Note that for $\phi_n \le n^{-1/3}$ and $i \le n^{1/4}$, we have $\eta_i\sqrt{\kappa_i} + \phi_n^3 = O(n^{-1})$. Hence (3.21) becomes the stronger estimate (3.22) for the nonoutlier eigenvalues with indices $i \le n^{1/4}$.

EXAMPLE 3.16. Again we assume that $\phi_n \le n^{-1/3}$. If $\widetilde{\sigma}_j^a + m_{2c}^{-1}(\lambda_+) \gtrsim 1$, that is, $\widetilde{\sigma}_j^a$ is well separated from the threshold, then $\widetilde{\boldsymbol{\xi}}_{\alpha(i)}$ is completely delocalized in the direction of $\mathbf{v}_j^a$ for all $i \notin \mathcal{O}^+$ and $i \le n^{1/4}$. We next consider the outliers that are close to the threshold.

Suppose that $i \le C$, that is, $\widetilde{\lambda}_i$ is near the edge. Then (3.22) gives

$$|\langle \mathbf{v}_j^a, \widetilde{\boldsymbol{\xi}}_{\alpha(i)}\rangle|^2 \prec \frac{1}{n(|\widetilde{\sigma}_j^a + m_{2c}^{-1}(\lambda_+)|^2 + n^{-2/3})}. \tag{3.23}$$

Therefore, the delocalization bound for the generalized component $|\langle \mathbf{v}_j^a, \widetilde{\boldsymbol{\xi}}_{\alpha(i)}\rangle|$ changes from the optimal order $n^{-1/2}$ to $n^{-1/6}$ as $\widetilde{\sigma}_j^a$ approaches the transition point $m_{2c}^{-1}(\lambda_+)$. This shows that the nonoutlier eigenvectors near the edge are biased in the direction of $\mathbf{v}_j^a$ provided that $\widetilde{\sigma}_j^a$ is near the transition point $m_{2c}^{-1}(\lambda_+)$. In particular, for $|\widetilde{\sigma}_j^a + m_{2c}^{-1}(\lambda_+)| \le n^{-1/3}$, we have that

$$|\langle \mathbf{v}_j^a, \widetilde{\boldsymbol{\xi}}_{\alpha(i)}\rangle|^2 \prec n^{-1}|\widetilde{\sigma}_j^a + m_{2c}^{-1}(\lambda_+)|^{-2}. \tag{3.24}$$

In the literature, the $\widetilde{\sigma}_j^a$ in this case is called a weak spike in statistics [25] or subcritical spike in probability [6]. Thus (3.24) shows that the nonoutlier eigenvectors still retain information about the weak spikes of $\widetilde{A}$ in contrast to the nonoutlier eigenvalues as seen from (3.10).

The Theorems 3.6, 3.7, 3.10 and 3.14 give the first order limits and convergent rates of the principal eigenvalues and eigenvectors of $\widetilde{\mathcal{Q}}_1$. The second-order asymptotics of the outlier eigenvalues and eigenvectors will be studied in another paper.

Note that for separable covariance matrices, $\widetilde{A}^{1/2}X\widetilde{B}^{1/2}$ and $\widetilde{B}^{1/2}X^*\widetilde{A}^{1/2}$ take exactly the same form. Hence by exchanging the roles of $(\widetilde{A}, X)$ and $(\widetilde{B}, X^*)$, one can immediately obtain from Theorems 3.10 and 3.14 the similar results for the eigenvectors $\widetilde{\boldsymbol{\zeta}}_k$ of $\widetilde{\mathcal{Q}}_2$. For reader's convenience, we state them in the following two theorems. Denote

$$\mathcal{P}'_S := \sum_{k\in S} \widetilde{\boldsymbol{\zeta}}_k \widetilde{\boldsymbol{\zeta}}_k^*, \quad \text{for } S \subset \mathcal{O}^+.$$

THEOREM 3.17. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that Assumptions 2.2, 2.6 and 3.2 hold. Fix any $S \subset \mathcal{O}^+$, we define the following deterministic positive quadratic form*:

$$\langle \mathbf{w}, \mathcal{Z}'_S \mathbf{w}\rangle := \sum_{\mu:\beta(\mu)\in S} \frac{|w_\mu|^2}{\widetilde{\sigma}^b_\mu} \frac{g'_{1c}(-(\widetilde{\sigma}^b_\mu)^{-1})}{g_{1c}(-(\widetilde{\sigma}^b_\mu)^{-1})}, \quad \textit{for } \mathbf{w} \in \mathbb{C}^n,\ w_\mu := \langle \mathbf{v}^b_\mu, \mathbf{w}\rangle.$$

Then for any deterministic vector $\mathbf{w} \in \mathbb{C}^n$, we have that

$$\begin{aligned}
&\big|\langle \mathbf{w}, \mathcal{P}'_S \mathbf{w}\rangle - \langle \mathbf{w}, \mathcal{Z}'_S \mathbf{w}\rangle\big| \\
&\prec \sum_{1\le\mu\le s:\beta(\mu)\in S} |w_\mu|^2 \psi_2(\widetilde{\sigma}^b_\mu) \\
&\quad + \sum_{1\le\mu\le s:\beta(\mu)\notin S} |w_\mu|^2 \frac{\phi_n^2}{\delta_{\beta(\mu)}(S)} + \sum_{\mu=1}^n |w_\mu|^2 \Big(\frac{\psi_2^2(\widetilde{\sigma}^b_\mu)\Delta_2^2(\widetilde{\sigma}^b_\mu)}{\delta^2_{\beta(\mu)}(S)} + \frac{\kappa_\mu}{n^{1/2}} \Big) \\
&\quad + \langle \mathbf{w}, \mathcal{Z}'_S \mathbf{w}\rangle^{1/2} \Big[\sum_{1\le\mu\le s:\beta(\mu)\notin S} \frac{|w_\mu|^2\phi_n^2}{\delta_{\beta(\mu)}(S)} \\
&\quad + \sum_{1\le\mu\le n:\beta(\mu)\notin S} |w_\mu|^2 \Big(\frac{\psi_2^2(\widetilde{\sigma}^b_\mu)\Delta_2^2(\widetilde{\sigma}^b_\mu)}{\delta^2_{\beta(\mu)}(S)} + \frac{\kappa_\mu}{n^{1/2}} \Big) \Big]^{1/2},
\end{aligned}$$

where we denote

$$\psi_2(\widetilde{\sigma}^b_\mu) := \phi_n + n^{-1/2}\Delta_2^{-1}(\widetilde{\sigma}^b_\mu).$$

If we have (a) (3.13) *holds, or* (b) *either A or B is diagonal, then the above estimate holds without the $n^{-1/2}\kappa_\mu$ terms.*

THEOREM 3.18. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that Assumptions 2.2, 2.6 and 3.2 hold. Fix any sufficiently small constant $\tau > 0$. For $\beta(\mu) \notin \mathcal{O}^+$, $\mu \le \tau n$ and any deterministic vector $\mathbf{w} \in \mathbb{C}^n$, we have*

$$\big|\langle \mathbf{w}, \widetilde{\boldsymbol{\zeta}}_{\beta(\mu)}\rangle\big|^2 \prec \sum_{\nu=1}^n |w_\nu|^2 \frac{n^{-1} + \eta_\mu\sqrt{\kappa_\mu} + \phi_n^3}{|\widetilde{\sigma}^b_\nu + m_{1c}^{-1}(\lambda_+)|^2 + \phi_n^2 + \kappa_\mu}.$$

If we have (a) (3.13) *holds, or* (b) *either A or B is diagonal, then we have the stronger estimate*

$$\big|\langle \mathbf{w}, \widetilde{\boldsymbol{\zeta}}_{\beta(\mu)}\rangle\big|^2 \prec \sum_{\nu=1}^n |w_\nu|^2 \frac{n^{-1} + \phi_n^3}{|\widetilde{\sigma}^b_\nu + m_{1c}^{-1}(\lambda_+)|^2 + \phi_n^2 + \kappa_\mu}.$$

Using a simple cutoff argument, it is easy to obtain the following corollary under certain moment assumptions. Since we do not assume the entries of X are identically distributed, the means and variances of the truncated entries may be different. This is why we assume the slightly more general conditions in (2.14).

COROLLARY 3.19. *Assume that $X = (x_{ij})$ is a real $p \times n$ matrix, whose entries are independent random variables that satisfy (2.1) and*

$$\max_{i,j} \mathbb{E}|\sqrt{n}x_{ij}|^a \le C, \tag{3.25}$$

for some constants $C > 0$ and $a > 4$. Suppose A, B, $\widetilde{A}$, $\widetilde{B}$ and d_n satisfy Assumptions 2.2 and 2.6. Then Theorems 3.6, 3.7, 3.10, 3.14, 3.17 and 3.18 hold for $\phi_n = n^{2/a-1/2}$ on an event with probability $1 - o(1)$.

Its proof is given in Section B of the Supplementary Material.

REMARK 3.20. We remark that one can take $r = 0$ or $s = 0$ (i.e., either $\widetilde{A}$ or $\widetilde{B}$ has no spikes) in the statements of our main results, although some results will become trivial null results. As an example, we consider the case where $r \ge 1$ and $s = 0$. In this case, the outlier eigenvalues only come from $\widetilde{A}$. Consequently, in Definition 3.5, we have that $\mathcal{O} := \{\alpha(i) : 1 \le i \le r\}$ and $\mathcal{O}^+ := \{\alpha(i) : 1 \le i \le r^+\}$. Then Theorem 3.6 still holds, although (3.9) becomes a null result since there is no μ such that $1 \le \mu \le 0$; Theorem 3.7 holds true with $s^+ = 0$ and $\alpha_+ := \min_i |\widetilde{\sigma}_i^a + m_{2c}^{-1}(\lambda_+)|$; Theorems 3.10, 3.14, 3.17 and 3.18 still hold for the left and right singular vectors, although Theorem 3.17 actually can be derived from Theorem 3.18 since there is no outlier coming from $\widetilde{B}$.

If $r = s = 0$, $\widetilde{Q}_1$ reduces to the nonspiked version $\mathcal{Q}_1 = A^{1/2}XBX^*A^{1/2}$. All of our main results are still valid, but better estimates actually hold in this case as given in [58], which studied nonspiked separable covariance matrices. Some of these results are also stated in Theorem C.11 and Lemma C.13 of our Supplementary Material [12].

3.3. *Strategy for the proof.* We conclude this section by describing briefly the main ideas and mathematical tools used in our proof. Using a linearization method (cf. (C.32) of [12]), we can show that the outlier eigenvalues satisfy a master equation in terms of the resolvents in (2.15) (cf. Lemma D.1 of [12]). Moreover, the resolvents appear in the forms $(V_o^a)^*\mathcal{G}_1 V_o^a$ and $(V_o^b)^*\mathcal{G}_2 V_o^b$, where we recall the notations in (2.11). These functionals of resolvents can be estimated using the anisotropic local law in [58], which shows that they are close to certain deterministic matrices up to some small errors (cf. Theorem C.9 of [12]). By replacing $(V_o^a)^*\mathcal{G}_1 V_o^a$ and $(V_o^b)^*\mathcal{G}_2 V_o^b$ with their deterministic equivalents, we can solve the master equation to get the asymptotic locations $\theta_1(\widetilde{\sigma}_i^a)$ and $\theta_2(\widetilde{\sigma}_\mu^b)$ of the outliers. To obtain the convergence rates in Theorems 3.6 and 3.7, we need to control the errors using the anisotropic local law and a three-step proof strategy developed in [27], which is summarized at the beginning of Section D in the Supplementary Material [12].

Once we know the asymptotic locations of the outliers, we can use Cauchy's integral formula to study the eigenvectors. For example, suppose the largest outlier $\widetilde{\lambda}_1$ is well separated from all the other eigenvalues. Then using the Cauchy's integral formula, we get

$$|\langle \mathbf{v}, \widetilde{\boldsymbol{\xi}}_1 \rangle|^2 = -\frac{1}{2\pi \mathrm{i}} \oint_\Gamma \mathbf{v}^* \sum_{k=1}^p \frac{\widetilde{\boldsymbol{\xi}}_k(i)\widetilde{\boldsymbol{\xi}}_k^*(j)}{\widetilde{\lambda}_k - z} \mathbf{v} \, \mathrm{d}z = -\frac{1}{2\pi \mathrm{i}} \oint_\Gamma \sum_{k=1}^p \mathbf{v}^* \widetilde{\mathcal{G}}_1(z) \mathbf{v} \, \mathrm{d}z$$

where Γ is a small contour enclosing $\widetilde{\lambda}_1$ only. For a more general integral representation of $\langle \mathbf{v}, \mathcal{P}_S \mathbf{v} \rangle$, we refer the reader to (E.13) of [12]. Using the anisotropic local law, we can obtain

the convergence limits and rates in Theorem 3.10. The proof of Theorem 3.14 relies on the simple bound

$$|\langle \mathbf{v}, \widetilde{\boldsymbol{\xi}}_k \rangle|^2 \le \eta \cdot \left(\mathbf{v}^* \sum_{k=1}^{p} \frac{\eta \widetilde{\boldsymbol{\xi}}_k(i) \widetilde{\boldsymbol{\xi}}_k^*(j)}{|\widetilde{\lambda}_k - z_k|^2} \mathbf{v} \right) = \eta \operatorname{Im} \mathbf{v}^* \widetilde{\mathcal{G}}_1(z_k) \mathbf{v},$$

where we take $z_k = \widetilde{\lambda}_k + \mathrm{i}\eta$. Again we will use the anisotropic local law to establish the delocalization bounds.

4. Statistical estimation for spiked separable covariance matrices. In this section, we consider the estimation of $\widetilde{A}$ and $\widetilde{B}$ from the data matrix $\widetilde{A}^{1/2} X \widetilde{B}^{1/2}$. In particular, we address two fundamental issues:

(1) estimating the number of spikes in $\widetilde{A}$ and $\widetilde{B}$;
(2) adaptive optimal shrinkage of the eigenvalues of $\widetilde{A}$ and $\widetilde{B}$.

To ease our discussion, until the end of this section, we will replace Assumption 3.2 with the following stronger *supercritical* condition. It is commonly used in the statistical literature, for instance [4, 14, 15, 39].

ASSUMPTION 4.1. For some fixed constant $\tau > 0$, we assume that there are r spikes for $\widetilde{A}$ and s spikes for $\widetilde{B}$, which satisfy

$$\widetilde{\sigma}_i^a + m_{2c}^{-1}(\lambda_+) > \tau, \quad 1 \le i \le r, \quad \text{and} \quad \widetilde{\sigma}_\mu^b + m_{1c}^{-1}(\lambda_+) > \tau, \quad 1 \le \mu \le s.$$

For simplicity of presentation, we will also assume the following nonoverlapping condition.

ASSUMPTION 4.2. Recall (3.15) and (3.16). For some fixed constant $\tau > 0$, we assume that

$$\min_{1 \le j \le r} \delta^a_{\alpha(i),\alpha(j)} \wedge \min_{1 \le \mu \le s} \delta^a_{\alpha(i),\beta(\mu)} \ge \tau, \quad 1 \le i \le r,$$

and

$$\min_{1 \le \nu \le s} \delta^b_{\beta(\mu),\beta(\nu)} \wedge \min_{1 \le i \le r} \delta^b_{\beta(\mu),\alpha(i)} \ge \tau, \quad 1 \le \mu \le s.$$

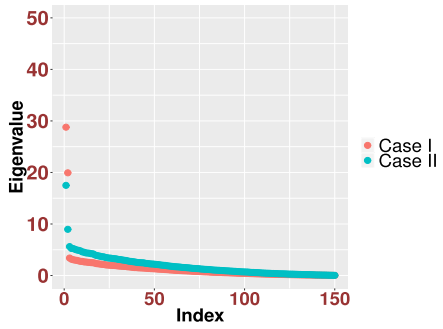
4.1. *Estimating the number of spikes.* The number of spikes has important meaning in practice. For instance, it represents the number of factors in factor model [43, 44] and number of signals in signal processing [40]. Such a problem has been studied for spiked covariance matrix; see, for example, [45]. In this section, we extend the discussion to the more general spiked separable model (2.12).

Different from the spiked covariance matrix model, we have two sources of spikes from either $\widetilde{A}$ or $\widetilde{B}$. For spiked covariance matrices, the statistic only involves sample eigenvalues. However, as we have seen from Theorem 3.6, the sample eigenvalues only contain information of the total number of spikes, that is, $r + s$. One way to deal with this issue is to use the information from the sample eigenvectors and apply Theorem 3.10. In Figure 1, we use a numerical simulation to illustrate how the eigenvectors can help us to gather information of separable covariance matrices. We consider two different settings:

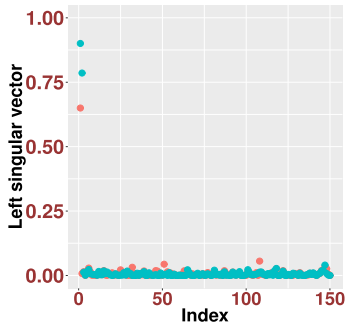
(Case I) $\qquad \widetilde{\Sigma}^a = \operatorname{diag}(5, 1, \ldots, 1), \qquad \widetilde{\Sigma}^b = \operatorname{diag}(5, 1, \ldots, 1),$

and

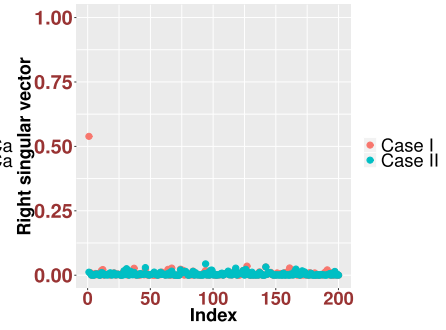
(Case II) $\qquad \widetilde{\Sigma}^a = \operatorname{diag}(3, 2, 1, \ldots, 1), \qquad \widetilde{\Sigma}^b = \operatorname{diag}(1, 1, \ldots, 1).$



(a) Eigenvalues.



(b) $|\langle \mathbf{v}_i^a, \widetilde{\boldsymbol{\xi}}_{\alpha(i)} \rangle|^2$



(c) $|\langle \mathbf{v}_\mu^b, \widetilde{\boldsymbol{\zeta}}_{\beta(\mu)} \rangle|^2$

FIG. 1. *Eigenvalues and eigenvectors for spiked separable covariance matrices.*

Figure 1(a) shows that there are two spikes in both cases. However, from Figure 1(b) and Figure 1(c), we can see that there are two parts of spikes in Case I, but only one part in Case II as expected. It shows the necessity to take into consideration the information from the eigenvectors. Here, we take $p = 150$, $n = 200$.

In the following discussion, we assume that the population eigenvectors of $\widetilde{A}$ and $\widetilde{B}$ are known. For the more general case where such information is unavailable, we will study it somewhere else (see also Remark 4.4).

We provide our statistic and start with a heuristic discussion. Under Assumptions 4.1 and 4.2, we get from Theorems 3.6, 3.10 and 3.14 that

$$\widetilde{\lambda}_{\alpha(i)} = \theta_1(\widetilde{\sigma}_i^a) + O_\prec(\phi_n),$$

and for $1 \le i \le r$,

$$|\langle \mathbf{v}_i^a, \widetilde{\boldsymbol{\xi}}_k \rangle|^2 = \mathbf{1}(k = \alpha(i)) \Big[\frac{1}{\widetilde{\sigma}_i^a} \frac{g_{2c}'(-(\widetilde{\sigma}_i^a)^{-1})}{\theta_1(\widetilde{\sigma}_i^a)} + O_\prec(\phi_n) \Big] + O_\prec(\phi_n^2).$$

Hence, if all the spiked eigenvalues are well separated, the ratio between $\widetilde{\lambda}_{\alpha(i)}$ and $\widetilde{\lambda}_{\alpha(i+1)}$ are strictly greater than 1. However, for the nonoutlier eigenvalues, these ratios will converge to 1 at a rate $O_\prec(n^{-2/3} + \phi_n^2)$ by Theorem 3.7 and eigenvalue rigidity, Theorem C.11 in the Supplementary Material. Moreover, the (cosine of) the angle $|\langle \mathbf{v}_i^a, \widetilde{\boldsymbol{\xi}}_k \rangle|$ is of order $O_\prec(\phi_n)$ except when $k = \alpha(i)$, in which case we have that $|\langle \mathbf{v}_i^a, \widetilde{\boldsymbol{\xi}}_k \rangle|$ is larger than a constant. Therefore, the ratios between consecutive eigenvalues and the angles will be used as our statistics.

Formally, for a given threshold $\omega > 0$ and a properly chosen constant $c > 0$, we define the statistic q by

$$q \equiv q(\omega) := \operatorname*{argmin}_{1 \le i \le c(p \wedge n)} \Big\{ \frac{\widetilde{\lambda}_{i+1}}{\widetilde{\lambda}_{i+2}} - 1 \le \omega \Big\}, \tag{4.1}$$

and $q_{a,b} \equiv q_{a,b}(\omega)$ by

$$q_a(\omega) := \operatorname*{argmin}_{1\le i\le c(p\wedge n)} \Big\{ \max_{1\le k\le c(p\wedge n)} |\langle \mathbf{v}_{i+1}^a, \widetilde{\boldsymbol{\xi}}_k\rangle|^2 \le \omega \Big\},$$

$$q_b(\omega) := \operatorname*{argmin}_{1\le \mu\le c(p\wedge n)} \Big\{ \max_{1\le \nu\le c(p\wedge n)} |\langle \mathbf{v}_{\mu+1}^b, \widetilde{\boldsymbol{\zeta}}_\nu\rangle|^2 \le \omega \Big\}.$$

As discussed above, q is used to estimate the total number of spikes, whereas q_a and q_b are used to estimate the number of spikes for $\widetilde{A}$ and $\widetilde{B}$, respectively. With Theorems 3.6, 3.7, 3.10, 3.14, 3.17 and 3.18, it is easy to show that they are consistent estimators for carefully chosen threshold ω. Denote the event $\Omega \equiv \Omega(\omega)$ by

$$\Omega := \{q = r + s, q_a = r, q_b = s\}.$$

THEOREM 4.3. *Suppose X has bounded support ϕ_n such that $n^{-1/2} \le \phi_n \le n^{-c_\phi}$ for some constant $c_\phi > 0$. Suppose that the Assumptions 2.2, 2.6, 4.1 and 4.2 hold. Then if ω satisfies that for some constant $\varepsilon > 0$,*

$$\omega \to 0, \quad \frac{\omega}{n^\varepsilon (n^{-2/3} + \phi_n^2)} \to \infty, \tag{4.2}$$

then we have that Ω holds with high probability for large enough n.

PROOF. This theorem is an easy consequence of Theorems 3.6, 3.7, 3.10, 3.14, 3.17 and 3.18. □

For the practical implementation, we employ a resampling procedure to choose the threshold ω for the statistic q using a reference matrix. Such procedure has been used in estimating the number of spikes for spiked covariance matrix [45]. We consider the case where the entries of X have finite $(12+\varepsilon)$th moments, such that we can take $\phi_n \ll n^{-1/3}$ by Corollary 3.19. Then by Theorem 3.7, the extreme nonoutlier eigenvalues of $\widetilde{\mathcal{Q}}_1$ have the same limiting distribution as those of the nonspiked matrix $\mathcal{Q}_1$, which, by the edge universality result [58], Theorem 2.7, fluctuate on the scale $n^{-2/3}$. Since the edge eigenvalues of Wishart matrix satisfy the Tracy–Widom distribution up to an $n^{-2/3}$ rescaling, the edge eigenvalue ratios of $\mathcal{Q}_1$ should be close to those of the Wishart matrix. More precisely, we can use Wishart matrix as the reference matrix and take the following steps to choose ω.

Step (i): Generate a sequence of N, say $N = 10^4$, $p \times p$ Wishart matrices $X_i X_i^*$ and the associated sequence of statistics $\{\mathcal{T}_i\}_{i=1}^N$,

$$\mathcal{T}_i := \max_{1\le k\le c(p\wedge n)} \{\lambda_k^{(i)}/\lambda_{k+1}^{(i)}\},$$

where $\{\lambda_k^{(i)}\}_{k=1}^{p\wedge n}$ are the eigenvalues of $X_i X_i^*$ arranged in descending order.

Step (ii): Given the nominal level ε (say $\varepsilon = 0.05$), we choose ω such that

$$\frac{\#\{\mathcal{T}_i \le 1 + \omega\}}{N} \ge 1 - \varepsilon.$$

In Figure 2, we consider the estimation of the number of spikes of $\widetilde{B}$ and analyze the frequency (over 10^4 simulations) of misestimation as a function of the value of x under different combinations of p and n. We make use of the statistic q_b and choose ω according to the above steps (i) and (ii). Specifically, we report the frequency of misestimation of the setting

$$\widetilde{A} = \operatorname{diag}(4, 1, \ldots, 1), \qquad \widetilde{B} = \operatorname{diag}(x + 2, x, 1, \ldots, 1), \quad x \ge 1.$$

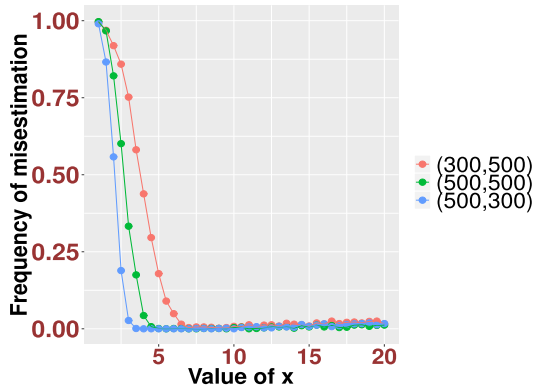


FIG. 2. *Frequency of misestimation for different values of x.*

We can see that our estimator performs quite well for x above some threshold.

Before concluding this subsection, we provide some insights on the choices of ω. In general, the choice of ω should depend on both A and B, denoted as $\omega_{A,B}$. Even though in the above procedure we have used ω_{I_p,I_n}, such a simple choice is usually sufficient for our purpose. In Section A.1 of the Supplementary Material [12], we show by simulations to verify our findings. On one hand, as illustrated in Figure A.1, the difference $|\omega_{I_p,I_n} - \omega_{A,B}|$ is already very small for $n = 200$ and the difference decreases when n increases. Moreover, empirically we see from the simulations that $|\omega_{A,B} - \omega_{I_p,I_n}| \le 0.008$ when $n \ge 300$ for a variety of d_n. On the other hand, for different choices of A and B, when the spiked eigenvalues are reasonably large, the frequency of misestimation will not be influenced if we simply use the threshold ω_{I_p,I_n}. In Section A.1 of our Supplementary Material [12], we record such simulation results in Figure A.2.

For smaller spikes, an accurate estimation of A and B can lead to more prudential choices of $\omega_{A,B}$. As discussed in Remark 3.8, there does not exist any method to estimate general A and B. Even though the construction of such estimators are out of the scope of this paper, when either A or B is identity, it reduces to estimating the spectrum of a sample covariance matrix. In this case, we can use many state-of-the-art algorithms to estimate the spectrum, for instance, [17, 29, 32]. In [12], Section A.1, assuming that $B = I_n$, we first use the numerical method as described in [33] to find an estimator of A, denoted as $\widehat{A}$, and then use $\omega_{\widehat{A},I_n}$ as our threshold. The results are recorded in Tables A.1–A.3. We see that it will reduce the frequency of misestimation for smaller spikes.

4.2. *Adaptive optimal shrinkage for spiked separable covariance matrices.* In most of the real applications, we have no a priori information on the true eigenvectors of $\widetilde{A}$ or $\widetilde{B}$. Then the natural choice for us is to use the sample eigenvectors $\{\widetilde{\boldsymbol{\xi}}_i\}_{1\le i\le p}$ and $\{\widetilde{\boldsymbol{\zeta}}_\mu\}_{1\le\mu\le n}$. Consider similar setting as in Johnstone's spiked covariance model [15, 24] with $A = I_p$ and $B = I_n$. Suppose we know the number of spikes $r + s$. Then we want to estimate

$$\widetilde{A} = \sum_{i=1}^{r} \widetilde{\sigma}_i^a \mathbf{v}_i^a (\mathbf{v}_i^a)^* + \sum_{i=r+1}^{p} \mathbf{v}_i^a (\mathbf{v}_i^a)^*, \qquad \widetilde{B} = \sum_{\mu=1}^{s} \widetilde{\sigma}_\mu^b \mathbf{v}_i^a (\mathbf{v}_i^a)^* + \sum_{\mu=s+1}^{n} \mathbf{v}_\mu^b (\mathbf{v}_\mu^b)^*,$$

using the estimators

$$\begin{aligned} \widehat{A} &= \sum_{i=1}^{r+s} \varrho_a(\widetilde{\lambda}_i) \widetilde{\boldsymbol{\xi}}_i \widetilde{\boldsymbol{\xi}}_i^* + \sum_{i=r+s+1}^{p} \widetilde{\boldsymbol{\xi}}_i \widetilde{\boldsymbol{\xi}}_i^*, \\ \widehat{B} &= \sum_{\mu=1}^{r+s} \varrho_b(\widetilde{\lambda}_i) \widetilde{\boldsymbol{\zeta}}_\mu \widetilde{\boldsymbol{\zeta}}_\mu^* + \sum_{i=r+s+1}^{n} \widetilde{\boldsymbol{\zeta}}_\mu \widetilde{\boldsymbol{\zeta}}_\mu^*, \end{aligned} \tag{4.3}$$

where $\varrho^a(\cdot)$ and $\varrho^b(\cdot)$ are some shrinkage functions characterized by the minimizers of certain loss functions:

$$\widehat{A} := \operatorname*{argmin}_{\mathcal{A}} \mathcal{L}_a(\mathcal{A}, \widetilde{A}), \qquad \widehat{B} := \operatorname*{argmin}_{\mathcal{B}} \mathcal{L}_b(\mathcal{B}, \widetilde{B}).$$

In [15], the authors consider this problem for spiked covariance matrices for a variety of loss functions assuming that r, s are known. In this section, we study this problem for spiked separable covariance matrices using the Frobenius norm as the loss functional. We will also prove the optimal convergent rate for such estimators. The other loss functions as discussed in [15] can be studied in a similar way.

We shall only consider $\varrho_a(\widetilde{\lambda}_i)$, while $\varrho_b(\widetilde{\lambda}_i)$ can be handled with the same argument by symmetry. We calculate that

$$\|\widehat{A} - \widetilde{A}\|_F^2 = \|T\|_F^2, \qquad T := \sum_{i=1}^{r+s} \bigl[\bigl(\varrho(\widetilde{\lambda}_i) - 1\bigr)\widetilde{\boldsymbol{\xi}}_i \widetilde{\boldsymbol{\xi}}_i^* - \bigl(\widetilde{\sigma}_i^a - 1\bigr)\mathbf{v}_i^a (\mathbf{v}_i^a)^*\bigr]. \tag{4.4}$$

We expand T to get

$$\begin{aligned}\|T\|_F^2 = \sum_{i=1}^{r+s} \bigl[&\bigl(\varrho_a(\widetilde{\lambda}_i) - 1\bigr)^2 + \bigl(\widetilde{\sigma}_i^a - 1\bigr)^2 - 2\bigl|\langle \mathbf{v}_i^a, \widetilde{\boldsymbol{\xi}}_i\rangle\bigr|^2 \bigl(\varrho_a(\widetilde{\lambda}_i) - 1\bigr)\bigl(\widetilde{\sigma}_i^a - 1\bigr)\bigr] \\ &- 2\sum_{i\neq j}^{r+s} \bigl(\varrho_a(\widetilde{\lambda}_i) - 1\bigr)\bigl(\widetilde{\sigma}_j^a - 1\bigr)\bigl|\langle \widetilde{\mathbf{v}}_j^a, \widetilde{\boldsymbol{\xi}}_i\rangle\bigr|^2.\end{aligned}$$

Therefore, (4.4) is minimized if

$$\varrho_a(\widetilde{\lambda}_i) = 1 + \sum_{j=1}^{r+s} \bigl(\widetilde{\sigma}_j^a - 1\bigr)\bigl|\langle \mathbf{v}_j^a, \widetilde{\boldsymbol{\xi}}_i\rangle\bigr|^2.$$

Under Assumptions 4.1 and 4.2, by Theorems 3.10 and 3.14 we find that for $\widetilde{\sigma}_k^a := d_k^a + 1$,

$$\varrho_a(\widetilde{\lambda}_i) = \mathbf{1}\bigl(i = \alpha(k) \text{ for some } k = 1, \ldots, r\bigr)\frac{d_k^a}{\widetilde{\sigma}_k^a}\frac{g_{2c}'(-(\widetilde{\sigma}_k^a)^{-1})}{g_{2c}(-(\widetilde{\sigma}_k^a)^{-1})} + O_{\prec}(\phi_n).$$

Under the setting with $A = I_p$ and $B = I_n$, $m_{2c}(z)$ is the Stieltjes transform of the standard Marchenko–Pastur (MP) law. Then it is known that g_{2c} is given by [28], Section 2.2,

$$g_{2c}(x) = -\frac{1}{x} + d_n \frac{1}{x+1},$$

where we recall that $d_n = p/n$. Therefore, we can calculate that

$$\varrho_a(\widetilde{\lambda}_i) = \frac{(d_k^a)^2 - d_n}{d_k^a + d_n} + O_{\prec}(\phi_n), \quad i = \alpha(k).$$

For d_k^a, we can use Theorem 3.6 to get that $d_k^a = -m_{2c}^{-1}(\lambda_i) - 1 + O_{\prec}(\phi_n)$ for $i = \alpha(k)$. We have the following explicit form for m_{2c} (see, e.g., (4.10) of [10]):

$$m_{2c}(x) = \frac{d_n - 1 - x + \sqrt{(x-\lambda_+)(x-\lambda_-)}}{2x}, \quad \lambda_\pm = \bigl(1 \pm d_n^{1/2}\bigr)^2,$$

when $x > \lambda_+$. Thus we can define the following shrinkage function:

$$\widehat{\varrho}_a(\widetilde{\lambda}_i) = \mathbf{1}\bigl(i = \alpha(k) \text{ for } k \in \{1, \ldots, r\}\bigr)\frac{(\widehat{d}_k^a)^2 - d_n}{\widehat{d}_k^a + d_n}, \quad \widehat{d}_k^a = -m_{2c}^{-1}(\widetilde{\lambda}_{\alpha(k)}) - 1,$$

which satisfies that

$$\varrho_a(\widetilde{\lambda}_i) = \widehat{\varrho}_a(\widetilde{\lambda}_i) + O_{\prec}(\phi_n).$$

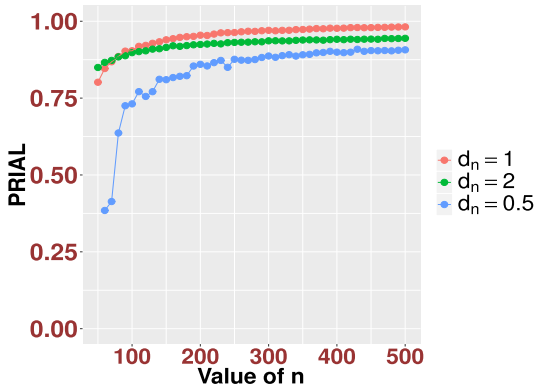


FIG. 3. *PRIAL against matrix dimension n. We consider the setting $\widetilde{A} = \mathrm{diag}(8, 5, 1, \ldots, 1)$ and $\widetilde{B} = \mathrm{diag}(3, 1, \ldots, 1)$.*

REMARK 4.4. Note that the definition of the shrinkage function depends on a priori knowledge of the indices of the outliers caused by the spikes of $\widetilde{A}$, which may not be available in applications. Moreover, the methods in Section 4.1 cannot be used since we have no information on the eigenvectors of $\widetilde{A}$ and $\widetilde{B}$. However, this kind of information is still possible to obtain by exploring the "cone condition" in Example 3.12, that is, we can project the left and right outlier-singular vectors onto some suitably chosen directions and take average over many samples. To have a rigorous theory, it is necessary to establish the second-order asymptotics of the outlier eigenvectors. Both of these topics will be explored elsewhere.

We then present the results of some Monte-Carlo simulations designed to illustrate the finite-sample properties of the shrinkage estimator $\widehat{A}$. We study the improvement of $\widehat{A}$ over the separable covariance matrix $\widetilde{\mathcal{Q}}_1$, which also uses the sample eigenvectors. Denote $\overline{A}$ as in (4.3) but with $\varrho_a(\widetilde{\lambda}_i)$ replaced by $\widehat{\varrho}_a(\widetilde{\lambda}_i)$. In Figure 3, we report the Percentage Relative Improvement in Average Loss (PRIAL) [31], Section 1.3, for $\overline{A}$:

$$\mathrm{PRIAL} := 100 \times \left\{1 - \frac{\mathbb{E}\|\overline{A} - \widehat{A}\|_F^2}{\mathbb{E}\|\widetilde{\mathcal{Q}}_1 - \widehat{A}\|_F^2}\right\}\%, \tag{4.5}$$

where $\mathbb{E}(\cdot)$ denotes the average over 10^4 Monte Carlo simulations. We can see that our estimators perform better than sample separable covariance matrix even for "not so large" matrix dimensions.

Before concluding this section, we provide a useful result for the estimation of spikes. By Theorem 3.6, we need to know the form of m_{2c} in order to estimate the spikes of $\widetilde{A}$. However, thanks to the anisotropic local law in [58] (see also Theorem C.9 and Theorem C.12 in the Supplementary Material), it is possible to have an adaptive estimator for the spikes of $\widetilde{A}$ based only on the data matrices $\widetilde{\mathcal{Q}}_2$ if $\widetilde{B}$ is a small-rank perturbation of the identity matrix. We define

$$\widehat{\sigma}_i^a := -\left(\frac{1}{n}\sum_{\nu=r+s+1}^{n} \frac{1}{\widetilde{\lambda}_\nu(\widetilde{\mathcal{Q}}_2) - \widetilde{\lambda}_{\alpha(i)}}\right)^{-1}, \quad 1 \le i \le r+s.$$

Similarly, if A is a small-rank perturbation of the identity matrix, then we have the following estimator for the spikes of $\widetilde{B}$:

$$\widehat{\sigma}_\mu^b := -\left(\frac{1}{n}\sum_{k=r+s+1}^{p} \frac{1}{\widetilde{\lambda}_k(\widetilde{\mathcal{Q}}_1) - \widetilde{\lambda}_{\beta(\mu)}}\right)^{-1}, \quad 1 \le \mu \le r+s.$$

We claim the following result.

TABLE 1
The value of $\widetilde{\sigma}^a$. We record the average of $\widehat{\sigma}^a$ over 2000 *simulations*

$\widetilde{\sigma}^a/(p,n)$	(100, 200)	(200, 400)	(300, 400)	(400, 300)	(500, 400)
4	3.67	3.58	3.83	4.61	4.43
5	4.78	4.65	4.84	5.49	5.37
8	7.75	7.62	7.86	8.47	8.33
10	9.83	9.65	9.88	10.51	10.37
15	14.95	14.86	14.93	15.56	15.42

THEOREM 4.5. *Suppose that the Assumptions 2.2, 2.6 and 4.1 hold. Suppose $\widetilde{B} = I_n + \mathcal{M}_n$, where $\mathcal{M}_n$ is a matrix of rank l_n. Then we have that for $1 \le i \le r$,*

$$\widetilde{\sigma}_i^a = \widehat{\sigma}_i^a + O_{\prec}(n^{-1}l_n + \phi_n). \tag{4.6}$$

Similarly, if $\widetilde{A}$ is an l_n-rank perturbation of the identity matrix, then for $1 \le \mu \le s$,

$$\widetilde{\sigma}_\mu^b = \widehat{\sigma}_\mu^b + O_{\prec}(n^{-1}l_n + \phi_n). \tag{4.7}$$

The proof of Theorem 4.5 will be given in the Supplementary Material. Here, we use some Monte Carlo simulations to illustrate the accuracy of the above estimators. We set

$$\widetilde{A} = \operatorname{diag}(\widetilde{\sigma}^a, 1, \ldots, 1), \qquad \widetilde{B} = \operatorname{diag}(3, 1, \ldots, 1).$$

In Table 1, we give the estimation of $\widetilde{\sigma}^a$ using $\widehat{\sigma}^a$ for various combinations of p and n. Each value is recorded by taking an average over 2000 simulations. We find that our estimator is quite accurate even for a small sample size.

Acknowledgments. The authors would like to thank Zhou Fan and Edgar Dobriban for helpful discussions. We also want to thank the Editor, the Associated Editor and two anonymous referees for their helpful comments, which have improved the paper significantly.

SUPPLEMENTARY MATERIAL

Supplement to "Spiked separable covariance matrices and principal components" (DOI: 10.1214/20-AOS1995SUPP; .pdf). In this file, we provide auxiliary lemmas and technical proofs for the results in Sections 3 and 4. We also report some additional simulation results in this file.

REFERENCES

[1] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

[2] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

[3] BAO, Z., PAN, G. and ZHOU, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.* **43** 382–421. MR3311864 https://doi.org/10.1214/14-AOS1281

[4] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.* **111** 120–135. MR2944410 https://doi.org/10.1016/j.jmva.2012.04.019

[5] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** 33. MR3183577 https://doi.org/10.1214/ejp.v19-3054

[6] BLOEMENDAL, A., KNOWLES, A., YAU, H.-T. and YIN, J. (2016). On the principal components of sample covariance matrices. *Probab. Theory Related Fields* **164** 459–552. MR3449395 https://doi.org/10.1007/s00440-015-0616-x

[7] BOURGADE, P., YAU, H.-T. and YIN, J. (2014). Local circular law for random matrices. *Probab. Theory Related Fields* **159** 545–595. MR3230002 https://doi.org/10.1007/s00440-013-0514-z

[8] COUILLET, R. and HACHEM, W. (2014). Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices Theory Appl.* **3** 1450016. MR3279621 https://doi.org/10.1142/S2010326314500166

[9] DING, X. (2019). Singular vector distribution of sample covariance matrices. *Adv. Appl. Probab.* **51** 236–267. MR3984017 https://doi.org/10.1017/apr.2019.10

[10] DING, X. (2020). High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli* **26** 387–417. MR4036038 https://doi.org/10.3150/19-BEJ1129

[11] DING, X. (2020). Spiked sample covariance matrices with possibly multiple bulk components. *Random Matrices Theory Appl.*. (In press).

[12] DING, X. and YANG, F. (2021). Supplement to "Spiked separable covariance matrices and principal components." https://doi.org/10.1214/20-AOS1995SUPP

[13] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. MR3809475 https://doi.org/10.1214/17-AAP1341

[14] DOBRIBAN, E. and OWEN, A. B. (2019). Deterministic parallel analysis: An improved method for selecting factors and principal components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 163–183. MR3904784

[15] DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

[16] EL KAROUI, N. (2007). Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.* **35** 663–714. MR2308592 https://doi.org/10.1214/009117906000000917

[17] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012 https://doi.org/10.1214/07-AOS581

[18] ERDŐS, L., KNOWLES, A. and YAU, H.-T. (2013). Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14** 1837–1926. MR3119922 https://doi.org/10.1007/s00023-013-0235-y

[19] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). Delocalization and diffusion profile for random band matrices. *Comm. Math. Phys.* **323** 367–416. MR3085669 https://doi.org/10.1007/s00220-013-1773-3

[20] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** 59. MR3068390 https://doi.org/10.1214/EJP.v18-2473

[21] GENTON, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics* **18** 681–695. MR2408938 https://doi.org/10.1002/env.854

[22] GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475 https://doi.org/10.1198/016214502760047113

[23] HACHEM, W., LOUBATON, P. and NAJIM, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *Ann. Appl. Probab.* **17** 875–930. MR2326235 https://doi.org/10.1214/105051606000000925

[24] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

[25] KE, Z. T. (2016). Detecting rare and weak spikes in large covariance matrices. arXiv preprint. Available at arXiv:1609.00883.

[26] KERMOAL, J. P., SCHUMACHER, L., PEDERSEN, K. I., MOGENSEN, P. E. and FREDERIKSEN, F. (2002). A stochastic MIMO radio channel model with experimental validation. *IEEE J. Sel. Areas Commun.* **20** 1211–1226.

[27] KNOWLES, A. and YIN, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1750. MR3103909 https://doi.org/10.1002/cpa.21450

[28] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770 https://doi.org/10.1007/s00440-016-0730-4

[29] KONG, W. and VALIANT, G. (2017). Spectrum estimation from samples. *Ann. Statist.* **45** 2218–2247. MR3718167 https://doi.org/10.1214/16-AOS1525

[30] KYRIAKIDIS, P. C. and JOURNEL, A. G. (1999). Geostatistical space-time models: A review. *Math. Geol.* **31** 651–684. MR1694654 https://doi.org/10.1023/A:1007528426688

[31] LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718 https://doi.org/10.1007/s00440-010-0298-3

[32] LEDOIT, O. and WOLF, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *J. Multivariate Anal.* **139** 360–384. MR3349498 https://doi.org/10.1016/j.jmva.2015.04.006

[33] LEDOIT, O. and WOLF, M. (2017). Numerical implementation of the QuEST function. *Comput. Statist. Data Anal.* **115** 199–223. MR3683138 https://doi.org/10.1016/j.csda.2017.06.004

[34] LEE, J. O. and SCHNELLI, K. (2016). Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Ann. Appl. Probab.* **26** 3786–3839. MR3582818 https://doi.org/10.1214/16-AAP1193

[35] LEEB, W. (2019). Matrix denoising for weighted loss functions and heterogeneous signals. arXiv preprint. Available at arXiv:1902.09474.

[36] LI, B., GENTON, M. G. and SHERMAN, M. (2008). Testing the covariance structure of multivariate random fields. *Biometrika* **95** 813–829. MR2461213 https://doi.org/10.1093/biomet/asn053

[37] LU, N. and ZIMMERMAN, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.* **73** 449–457. MR2187860 https://doi.org/10.1016/j.spl.2005.04.020

[38] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR, Sb.* **1** 457.

[39] NADAKUDITI, R. R. (2014). OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inf. Theory* **60** 3002–3018. MR3200641 https://doi.org/10.1109/TIT.2014.2311661

[40] NADLER, B. (2010). Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator. *IEEE Trans. Signal Process.* **58** 2746–2756. MR2789420 https://doi.org/10.1109/TSP.2010.2042481

[41] O'ROURKE, S., VU, V. and WANG, K. (2016). Eigenvectors of random matrices: A survey. *J. Combin. Theory Ser. A* **144** 361–442. MR3534074 https://doi.org/10.1016/j.jcta.2016.06.008

[42] ONATSKI, A. (2008). The Tracy–Widom limit for the largest eigenvalues of singular complex Wishart matrices. *Ann. Appl. Probab.* **18** 470–490. MR2398763 https://doi.org/10.1214/07-AAP454

[43] ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* **92** 1004–1016.

[44] ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258. MR2923766 https://doi.org/10.1016/j.jeconom.2012.01.034

[45] PASSEMIER, D. and YAO, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *J. Multivariate Anal.* **127** 173–183. MR3188885 https://doi.org/10.1016/j.jmva.2014.02.017

[46] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[47] PAUL, D. and AUE, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. MR3206718 https://doi.org/10.1016/j.jspi.2013.09.005

[48] PAUL, D. and SILVERSTEIN, J. W. (2009). No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *J. Multivariate Anal.* **100** 37–57. MR2460475 https://doi.org/10.1016/j.jmva.2008.03.010

[49] PILLAI, N. S. and YIN, J. (2014). Universality of covariance matrices. *Ann. Appl. Probab.* **24** 935–1001. MR3199978 https://doi.org/10.1214/13-AAP939

[50] SKUP, M. (2010). Longitudinal fMRI analysis: A review of methods. *Stat. Interface* **3** 235–252. MR2659514 https://doi.org/10.4310/SII.2010.v3.n2.a10

[51] TRACY, C. A. and WIDOM, H. (1994). Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.* **159** 151–174. MR1257246

[52] TRACY, C. A. and WIDOM, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177** 727–754. MR1385083

[53] TULINO, A. and VERDÚ, S. (2004). *Random Matrix Theory and Wireless Communications. Foundations and Trends in Communications and Information Theory*. Publishers Inc..

[54] WANG, L. and PAUL, D. (2014). Limiting spectral distribution of renormalized separable sample covariance matrices when $p/n \to 0$. *J. Multivariate Anal.* **126** 25–52. MR3173080 https://doi.org/10.1016/j.jmva.2013.12.015

[55] WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Trans. Signal Process.* **56** 478–491. MR2445531 https://doi.org/10.1109/TSP.2007.907834

[56] XI, H., YANG, F. and YIN, J. (2020). Convergence of eigenvector empirical spectral distribution of sample covariance matrices. *Ann. Statist.* **48** 953–982. MR4102683 https://doi.org/10.1214/19-AOS1832

[57] YANG, F. Local laws of random matrices and their applications Ph.D. thesis, UCLA.
[58] YANG, F. (2019). Edge universality of separable covariance matrices. *Electron. J. Probab.* **24** 123. MR4029426 https://doi.org/10.1214/19-ejp381
[59] YEO, J. and PAPANICOLAOU, G. (2016). Random matrix approach to estimation of high-dimensional factor models. arXiv preprint. Available at arXiv:1611.05571.
[60] ZHANG, L. Spectral Analysis of Large Dimensional Random Matrices Ph.D. thesis, National University of Singapore.