Submitted to the Annals of Statistics

STATISTICAL INFERENCE FOR PRINCIPAL COMPONENTS OF SPIKED COVARIANCE MATRICES

BY ZHIGANG BAO[1], XIUCAI DING[2,‡], JINGMING WANG[1,*] AND KE WANG[1,†]

[1]*Department of Mathematics, Hong Kong University of Science and Technology, mazgbao@ust.hk;* [*]*jwangdm@connect.ust.hk;* [†]*kewang@ust.hk*

[2]*Department of Statistics, University of California, Davis,* [‡]*xcading@ucdavis.edu*

In this paper, we study the asymptotic behavior of the extreme eigenvalues and eigenvectors of the high dimensional spiked sample covariance matrices, in the supercritical case when a reliable detection of spikes is possible. In particular, we derive the joint distribution of the extreme eigenvalues and the generalized components of the associated eigenvectors, i.e., the projections of the eigenvectors onto arbitrary given direction, assuming that the dimension and sample size are comparably large. In general, the joint distribution is given in terms of linear combinations of finitely many Gaussian and Chi-square variables, with parameters depending on the projection direction and the spikes. Our assumption on the spikes is fully general. First, the strengths of spikes are only required to be slightly above the critical threshold and no upper bound on the strengths is needed. Second, multiple spikes, i.e., spikes with the same strength, are allowed. Third, no structural assumption is imposed on the spikes. Thanks to the general setting, we can then apply the results to various high dimensional statistical hypothesis testing problems involving both the eigenvalues and eigenvectors. Specifically, we propose accurate and powerful statistics to conduct hypothesis testing on the principal components. These statistics are data-dependent and adaptive to the underlying true spikes. Numerical simulations also confirm the accuracy and powerfulness of our proposed statistics and illustrate significantly better performance compared to the existing methods in the literature. In particular, our methods are accurate and powerful even when either the spikes are small or the dimension is large.

1. Introduction. Covariance matrices play an important role in multivariate analysis and high dimensional statistics, and find applications in many scientific fields. Moreover, many statistical methodologies and techniques rely on the knowledge of the structure of the covariance matrix, to name but a few, Principal Component Analysis, Discriminant Analysis and Cluster Analysis. For detailed discussions of the applications and methodologies, we refer the readers to the monographs [1, 46, 50] for a review. It is well-known in the high dimensional setting when the dimension is comparable with or much larger than the sample size, a direct application of the sample covariance matrix for hypothesis testing may result in untrustful conclusions. Consequently, a thorough understanding of the distributions of the eigenvalues and eigenvectors of sample covariance matrices is in demand for high dimensional statistical inference.

In the literature of high dimensional statistics, a popular and sophisticated model is the spiked covariance matrix model proposed by Johnstone in [47], where a finite number of spikes (i.e., eigenvalues detached from the bulk of the spectrum) are added to the spectrum of the population covariance matrix; see (1.2) and (1.3) below. Throughout the paper, with certain abuse of terminology, we use the word "spike" to represent either a detached eigenvalue

MSC2020 subject classifications: Primary 60B20, 62G10; secondary 62H10, 15B52, 62H25.

Keywords and phrases: random matrix, sample covariance matrix, eigenvector, spiked model, principal component, adaptive estimator.

$1+d_i$ (c.f. (1.2), (1.3)) or the whole rank one matrix corresponding to a detached eigenvalue $(1+d_i)\boldsymbol{v}_i\boldsymbol{v}_i^*$ (c.f. (1.2), (1.3)). These spikes can have various practical meanings in different fields. For instance, they correspond to the first few important factors in factor models arising from financial economics [39, 70], the important patterns in genetic variation across the globe [34], the clusters in gene expression data [51] and the signals in signal detection [68, 72]. In this paper, we investigate the distributions of the principal components of the spiked sample covariance matrix, i.e, the sample counterparts of the extreme eigenvalues and eigenvectors (especially those spikes) of the population covariance matrices. The principal components of spiked sample covariance matrices play important roles in Principal Component Analysis for high dimensional data. A lot of work has been devoted to estimating the principal components in various settings. For instance, sparse principal component analysis [20, 48] is proposed to estimate the spiked eigenvalues and eigenvectors assuming some sparsity structure in the population eigenvectors; factor-model based estimators [3, 4, 73] for the eigenvectors are constructed if the population covariance matrix is of approximate factor-model type; and some regularization-based methods [62, 81, 89] have been proposed under various structural assumptions.

Despite the wide applications of the principal components in high dimensional statistics, most of the literature focus on the estimation part. Much less is known about their distributions, especially for the leading eigenvectors. As a consequence, a thorough study of the statistical inference for the population covariance matrix in the high dimensional setting is still missing, especially for hypothesis testing problems involving both eigenvalues and eigenvectors. For instance, eigenvectors and eigenspaces play an important role in statistical learning. However, the existing literature has only been able to test whether the eigenvectors or eigenspaces of the population covariance matrix are equal to some given ones under the assumption that the dimension is much smaller than the sample size [42, 56, 69, 82, 83]. For another example, in Principal Component Analysis, the loadings are transformations of the original variables to the eigenvectors. They describe how much each variable contributes to a particular eigenvector and researchers are interested in hypothesis testing and constructing confidence intervals for them [57, 74, 87]. The loadings are scaled eigenvectors using their corresponding eigenvalues and therefore, the joint distribution of the extreme eigenvalues and eigenvectors of the sample covariance matrices will be needed to conduct inference.

Driven by these challenges, we study the joint distributions of the extreme eigenvalues and the generalized components of their associated eigenvectors for the spiked sample covariance matrices, in the high dimensional setting. Based on these results, we will be able to perform hypothesis testings with statistics constructed from both eigenvalues and eigenvectors.

Specifically, in this paper, we consider the sample covariance matrices of the form

$$Q = TXX^*T^*, \tag{1.1}$$

where T is a $M \times M$ deterministic matrix and X is a $M \times N$ random matrix with independent entries and $\mathbb{E}XX^* = I_M$. Further, we assume that the population covariance matrix $\Sigma := TT^*$ admits the following form

$$\Sigma = I_M + S, \tag{1.2}$$

where S is a fixed-rank deterministic positive semi-definite matrix. Here we refer to Section 1.2 of [21] for several examples which boil down to this setting. Moreover, we denote the spectral decomposition of S by

$$S = \sum_{i=1}^{r} d_i \boldsymbol{v}_i \boldsymbol{v}_i^*, \tag{1.3}$$

where $r \geq 1$ is a fixed integer. Here $d_1 \geq \cdots \geq d_r > 0$ are the ordered eigenvalues of S, and $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iM})^*$'s are the associated unit eigenvectors. All $d_i \equiv d_i(N)$ may be N-dependent. Throughout the paper, for simplicity, we will mainly work with the setting

$$T = \Sigma^{\frac{1}{2}}. \tag{1.4}$$

We remark that our results hold for much more general T satisfying $\Sigma = TT^*$. We refer to Section A of [12] for more discussions on the extension along this direction.

1.1. *Summary of previous related theoretical results.* In this section, we summarize the results related to the spiked sample covariance matrix from the Random Matrix Theory literature.

We denote by $\mu_1 \geq \cdots \geq \mu_{M \wedge N}, M \wedge N := \min\{M, N\}$, the nontrivial eigenvalues of Q and $\boldsymbol{\xi}_i$ the unit eigenvector associated with μ_i. The primary interest of the sample covariance matrix Q lies in the asymptotic behavior of a few largest μ_i's and the associated $\boldsymbol{\xi}_i$'s when N is large, under various assumptions of d_i's and $\boldsymbol{v}_i$'s. Significant progress has been made on this topic in the last few years. It has been well-known since the seminal work of Baik, Ben Arous and Péché [9] that the largest eigenvalues μ_i's undergo a phase transition (BBP transition) w.r.t. the size of d_i's. On the level of the first order limit, when $d_i > \sqrt{y}$, the eigenvalue μ_i jumps out of the support of the Marchenko-Pastur law (MP law) and converges to a limit determined by d_i, while in the case of $d_i \leq \sqrt{y}$, it sticks to the right end of the Marchenko-Pastur (MP) law $(1 + \sqrt{y})^2$. In the former case, we call μ_i an *outlier* or *outlying eigenvalue*, while in the latter case we call μ_i a *sticking eigenvalue*. On the level of the second order fluctuation, it was revealed in [9] that a phase transition for μ_i takes place in the regime $d_i - \sqrt{y} \sim N^{-\frac{1}{3}}$. Specifically, if $d_i - \sqrt{y} \ll N^{-\frac{1}{3}}$ (subcritical regime), the eigenvalue μ_i still admits the Tracy-Widom type distribution; if $d_i - \sqrt{y} \gg N^{-\frac{1}{3}}$ (supercritical regime), the eigenvalue μ_i is asymptotically Gaussian; while if $d_i - \sqrt{y} \sim N^{-\frac{1}{3}}$ (critical regime), the limiting distribution of the eigenvalue μ_i is some interpolation between Tracy-Widom and Gaussian. On extreme eigenvalues, further study for more generally distributed covariance matrices can be found in [10, 16, 76, 7, 8, 21, 31, 59]. The limiting behavior of the extreme eigenvalues has also been studied for various related models, such as the finite-rank deformation of Wigner matrices [16, 28, 29, 41, 53, 54, 77, 80], the signal-plus-noise model [17, 60, 30], the general spiked β ensemble [22, 23], and also the finite-rank deformation of general unitary/orthogonal invariant matrices [18, 14, 15, 32].

In contrast, the study on the limiting behavior of the eigenvectors associated with the extreme eigenvalues is much less. On the level of the first order limit, it is known that the $\boldsymbol{\xi}_i$'s are delocalized and purely noisy in the subcritical regime, but has a bias on the direction of $\boldsymbol{v}_i$ in the supercritical regime. We refer to [18, 17, 26, 30, 76, 21, 31] for more details of such a phenomenon. It was recently noticed in [21] that a d_i close to the critical point can cause a bias even for the non-outlier eigenvectors. On the level of the second order fluctuation, it was proved in [21] that the eigenvectors are asymptotically Gaussian in the subcritical regime, for the spiked covariance matrices. For a related model, spiked GUE, the eigenvector distribution in the critical regime was recently obtained in the work [13]. In the supercritical regime, a non-universality phenomenon was shown in [27] and [11] for the eigenvector distribution for the finite-rank deformation of Wigner matrices and the signal-plus-noise model, respectively. The non-universality phenomenon in the supercritical regime has been previously observed in [28, 53, 54] for the extreme eigenvalues of the finite-rank deformation of Wigner matrices. Here we also refer to [64, 49, 38] for related study on the extreme eigenstructures of various finite-rank deformed models from more statistical perspective.

1.2. *An overview of our results.* In the theoretical part of this paper, we will primarily focus on the distribution of the eigenvectors $\boldsymbol{\xi}_i$'s associated with the outlying eigenvalues μ_i's. That means, we will focus on the supercritical regime, in contrast to the work [21] and [13] where the eigenvector distributions in the subcritical and critical regimes were obtained. The results in the supercritical regime are particularly important for the statistical applications, since it is well-known that a reliable detection of spikes based on eigenvalues is only possible in this regime in general; see [65, 66, 67, 78] for instance. Our assumption on the spikes is fully general (c.f. Assumption 2.4). In particular, we do allow d_i's to be divergent and multiple (i.e. some d_i's are identical). In case that the spikes are simple (i.e. d_i's are all distinct), we also establish the joint distribution of the outlying eigenvalues and the associated eigenvectors for the spiked covariance matrices. More specifically, in this paper, we are interested in the distribution of the largest μ_i's and the *generalized component* of the top eigenvectors, i.e., the projections of those eigenvectors onto a general direction. Let $\boldsymbol{w} \in S_{\mathbb{R}}^{M-1}$ be any deterministic unit vector. We will study the limiting distribution of $|\langle \boldsymbol{w}, \boldsymbol{\xi}_i \rangle|^2$ in the supercritical regime under general assumption of the spikes, and also state the joint distribution of $|\langle \boldsymbol{w}, \boldsymbol{\xi}_i \rangle|^2$ and μ_i's in case that the spikes are simple. We emphasize here that in case that a spike is multiple, one can also describe the joint distribution of eigenvalues and eigenvectors using the approach in this paper. But the result does not have a succinct form so we omit it from the statements of our main theorems; see Remark 2.11 for more details. Nevertheless, we will describe (in certain equivalent form) and prove an extension of the joint eigenvalue-eigenvector distribution to the multiple case and present applications of this result in the supplement [12].

In the application part of this paper, we construct statistics to infer the principal components. We mainly focus on two hypothesis testing problems regarding the eigenspaces, (3.2) and (3.3). To our best knowledge, it is the first time that these problems are tackled for spiked covariance matrices in the high dimensional regime (2.2) without imposing any structural assumptions on the spikes. Our proposed statistics make use of some plug-in estimators and are adaptive to the information of unknown spikes, for instance, their values and multiplicity. Thanks to the joint distribution of the eigenvalues and eigenvectors, we can easily establish the asymptotic distributions of our test statistics; see Section 3.1 for more details. Our methodology is simple, computationally cheap and easily implemented. Extensive numerical simulations lend strong support to our test statistics. In particular, our proposed statistics are accurate and powerful regardless of the value of y and magnitude of the spikes. Moreover, for testing (3.2), our statistic shows better performance compared to the existing methods in the literature both in terms of accuracy and power. We point out that our methodology can be used to study other hypothesis testing problems regarding Principal Component Analysis and this will be discussed in Section 3.

In the sequel, we further highlight some novelties, in contrast to previous works. We first point out that a related problem has been previously studied in [11] for the so-called matrix denoising model, where the distribution of the leading singular vectors of this model was studied. Due to the additive structure of this model, the distribution of the singular vector may depend on the structure of the deformation and the entire distribution of entries of the noise matrix (rather than their first 4 moments only), and may not be Gaussian or Chi-square or linear combinations of them. Such a phenomenon is called non-universality, which exists in the additive models [11, 27]. However, such a phenomenon does not show up for the spiked covariance matrix, as one can see from Theorem 2.7 in the sequel, where the distribution has a *Gaussian nature* in the sense that it is a polynomial of Gaussian variables. This is essentially due to the multiplicative structure of the spiked covariance matrix where the structure of the spikes are smoothed out by the random matrix X.

In addition, we emphasize here that in [11] the assumption of the strengths of the deformation, counterpart of d_i's, is much more limited than the assumption here, and the results in

[11] are stated in much more restricted forms. Our main assumption for this paper is Assumption 2.4 below. In [11], the strengths are assumed to be bounded and thus cannot grow with N, and also the strengths are away from the critical threshold by a constant order distance. Further, the strengths in [11] are assumed to be simple, and thus no multiplicity is allowed, and also distinct strengths are away from each other by a constant order. In Assumption 2.4, we remove all these restrictions on d_i's. In addition, in [11], only the projection of the random singular vector onto the directions of the deformations are discussed. Here we consider the projection onto arbitrary given direction. Finally, no joint distribution of eigenvalue and eigenvectors is obtained in [11]. Here we establish the joint distribution of eigenvalues and eigenvectors.

All the theoretical novelties are well motivated by our applications. First, in most of mathematical work on spiked models, d_i is assumed to be bounded. However, in many popular statistical models such as the factor model [3, 4, 5, 70], $d_i \equiv d_i(N)$ could be diverging. We provide a unified result in the whole supercritical regime, no matter d_i is close to the threshold or diverging. Practically, that means our results can be applied no matter the spike is weak or strong. Second, in the application part, we consider two hypothesis testing problems. The first is to test whether an eigenspace formed by any part of the spikes is equal to some given subspace, while the second is to test whether it is orthogonal to certain given subspace. Both questions are significant in the statistics literature. Let I be an index set of certain (possibly) multiple d_i's; see Assumption 2.4 for detailed definition. Our test statistics for both testing problems are constructed from $\mu_i, i \in \mathsf{I}$ and the projection $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$ or its variants, where $\mathrm{P}_{\mathsf{I}} := \sum_{t \in \mathsf{I}} \boldsymbol{\xi}_t \boldsymbol{\xi}_t^*$ and the choice of $\boldsymbol{w}$ depends on the testing problems. The limiting distribution of the first statistic relies on our joint eigenvalue-eigenvector distribution in case $\boldsymbol{w} \in \mathrm{Span}\{\boldsymbol{v}_t\}_{t \in \mathsf{I}}$, while the limiting distribution of the second statistic relies on the joint eigenvalue-eigenvector distribution in case $\boldsymbol{w} \in \mathrm{Span}\{\boldsymbol{v}_j\}_{j \in [\![1,M]\!] \setminus \mathsf{I}}$. This explains the necessity for us to derive the distribution of the projection onto general directions. Third, in two applications, if we construct the test statistics, using the result in Theorem 2.7 solely, the limiting distribution of the statistics will contain the parameters d_i's, which are normally unknown in real application. Hence, in order to construct *adaptive* statistics which do not depend on the unknown parameters d_i's, we use a plug-in estimator of d_i which is given in terms of μ_i. Then, in order to derive the distribution of these adaptive statistics, we have to establish the joint eigenvalue-eigenvector distribution, as what we have in Theorem 2.10, and its multiple extension in Proposition I.4 of [12].

Organization: The paper is organized as the following: In Section 2, we state our main results and proof strategy. In Section 3, we discuss several applications of our results and present the simulation results. In Section 4, we provide a sketch for our proof strategy. Technical proofs and additional simulation results are deferred to supplement [12].

Notation: Throughout the paper, the sample size N will be the fundamental parameter which goes to ∞. The symbol $o_N(\cdot)$ stands for any quantity going to 0 as N goes to ∞. We use c and C to denote positive finite constants that do not depend on N. Their values may change from line to line. For two positive quantities A_N, B_N depending on N we use the notation $A_N \asymp B_N$ to denote the relation $C^{-1} A_N \leq B_N \leq C A_N$ for some constant $C > 1$. Further, we write $A_N \doteq B_N$ if $A_N = B_N(1 + o_N(1))$.

For vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^N$, we write $\boldsymbol{v}^* \boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{w} \rangle$ for their scalar product. We emphasize here, unless otherwise specified, the vectors in this paper are real vectors and thus $\boldsymbol{v}^* \boldsymbol{w} = \boldsymbol{v}^\top \boldsymbol{w}$. Further, for a matrix A, we denote by $\|A\|_{\mathrm{op}}$ its operator norm, while we use $\|\boldsymbol{v}\|$ to represent the ℓ^2 norm for a vector $\boldsymbol{v}$.

We use double brackets to denote index sets, i.e. for $n_1, n_2 \in \mathbb{R}$, $[\![n_1, n_2]\!] := [n_1, n_2] \cap \mathbb{Z}$. In addition, we use $\mathbf{1}_n = \frac{1}{\sqrt{n}}(1, \ldots, 1)^*$ to denote the n-dimensional normalized all-1 vector. Further, we denote by $\mathbb{1}(E)$ or $\mathbb{1}_E$ the indicator function of an event E.

2. Main results. In this section, we state our main results.

2.1. *Notations and assumptions.* In this subsection, we introduce some necessary notations and technical assumptions. For any vectors $\boldsymbol{a}_l = (a_l(i)) \in \mathbb{R}^M, l \in \mathbb{Z}^+$, we set

$$\boldsymbol{s}_{k_1,\cdots,k_t}(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_t) = \sum_{j=1}^{M} a_1(j)^{k_1}\cdots a_t(j)^{k_t}. \tag{2.1}$$

For instance, $\boldsymbol{s}_{1,3}(\boldsymbol{a}_1,\boldsymbol{a}_2) = \sum_{j=1}^{M} a_1(j)a_2(j)^3$. Rewrite the spectral decomposition of Σ as $\Sigma = \sum_{i=1}^{M}\sigma_i\boldsymbol{v}_i\boldsymbol{v}_i^* = I_M + \sum_{i=1}^{M} d_i\boldsymbol{v}_i\boldsymbol{v}_i^*$, where $d_{r+1} = \cdots = d_M = 0$. Further, we emphasize here that the specific choice of the orthonormal $\boldsymbol{v}_i$'s for $i = r+1,\ldots,M$ is irrelevant to our discussion since only $\sum_{j=r+1}^{M}\boldsymbol{v}_i\boldsymbol{v}_i^*$ will be involved. In the sequel, we fix an i and consider a (possibly) multiple d_i.

We will need the following notion of *stochastic domination* introduced in [37], which provides a precise statement of the form "X_N is bounded by Y_N up to a small power of N with high probability".

DEFINITION 2.1. *(Stochastic domination) Let*

$$X = \big(X_N(u) : N \in \mathbb{N}, u \in U_N\big), \quad Y = \big(Y_N(u) : N \in \mathbb{N}, u \in U_N\big)$$

be two families of random variables, where Y is nonnegative, and U_N is a possibly N-dependent parameter set. We say that X is bounded by Y, uniformly in u, if for all small $\varrho > 0$ and large $\phi > 0$, we have

$$\sup_{u\in U_N}\mathbb{P}\big(|X_N(u)| > N^{\varrho}Y_N(u)\big) \le N^{-\phi}$$

for large $N \ge N_0(\varrho,\phi)$. Throughout the paper, we use the notation $X = O_{\prec}(Y)$ or $X \prec Y$ when X is stochastically bounded by Y uniformly in u. Note that in the special case when X and Y are deterministic, $X \prec Y$ means for any given $\varrho > 0$, $|X_N(u)| \le N^{\varrho}Y_N(u)$ uniformly in u, for all sufficiently large $N \ge N_0(\varrho)$. In addition, we also say that an N-dependent event $\mathcal{E} \equiv \mathcal{E}(N)$ holds with high probability if, for any large $\varphi > 0$,

$$\mathbb{P}(\mathcal{E}) \ge 1 - N^{-\varphi},$$

for sufficiently large $N \ge N_0(\varphi)$.

To ease our statements, we will need the following definition.

DEFINITION 2.2. *Two sequences of random vectors, $\mathsf{X}_N \in \mathbb{R}^k$ and $\mathsf{Y}_N \in \mathbb{R}^k$, $N \ge 1$, are* asymptotically equal in distribution*, denoted by $\mathsf{X}_N \simeq \mathsf{Y}_N$, if they are tight (i.e., for any $\epsilon > 0$, there exists a $D > 0$ such that $\sup_N \mathbb{P}(\|X_N\| \ge D) \le \epsilon$) and satisfy*

$$\lim_{N\to\infty}\big(\mathbb{E}f(\mathsf{X}_N) - \mathbb{E}f(\mathsf{Y}_N)\big) = 0$$

for any bounded continuous function $f : \mathbb{R}^k \to \mathbb{R}$.

Next, we impose the necessary technical assumptions.

ASSUMPTION 2.3. *Throughout the paper, we suppose the following assumptions hold. (i)(On dimensionality): We assume that $M \equiv M(N)$ and N are comparable and there exist constants $\tau_2 > \tau_1 > 0$ such that*

$$y \equiv y_N = M/N \in (\tau_1,\tau_2). \tag{2.2}$$

(ii)(On X): For the matrix $X = (x_{ij})$, we assume that the entries $x_{ij} \equiv x_{ij}(N)$ are real random variables satisfying

$$\mathbb{E}x_{ij} = 0, \qquad \mathbb{E}x_{ij}^2 = 1/N.$$

Moreover, we assume the existence of large moments, i.e., for any integer $p \geq 3$, there exists a constant $C_p > 0$, such that

$$\mathbb{E}|\sqrt{N}x_{ij}|^p \leq C_p < \infty. \tag{2.3}$$

We further assume that all $\sqrt{N}x_{ij}$'s possess the same 3rd and 4th cumulants, which are denoted by κ_3 and κ_4 respectively.

We mention that the moment assumption (2.3) can be relaxed using some truncation techniques. Moreover, we can extend our results to allow different third and fourth cumulants. We do not pursue these generalizations in the current paper. For more discussions on these directions, we refer to Section A of our supplement [12]. Further, since we will focus on the supercritical regime, we make the following assumption.

ASSUMPTION 2.4. *Let $\epsilon > 0$ be any small but fixed constant. Let $d_i \equiv d_i(N), i \in [\![1, r]\!]$ be the eigenvalues of S in (1.3). There exists a maximum integer $r_0 \equiv r_0(\epsilon) \in [\![1, r]\!]$, such that for any $i \in [\![1, r_0]\!]$,*

$$d_i - y^{1/2} > N^{-\frac{1}{3}+\epsilon} \tag{2.4}$$

for all sufficiently large $N \geq N_0(\epsilon)$. Moreover, for a fixed $i \in [\![1, r_0]\!]$, there exists a (unique) index set $\mathsf{I} \equiv \mathsf{I}(i) \subset [\![1, r_0]\!]$ such that $i \in \mathsf{I}$ and for any $t \in \mathsf{I}$,

$$d_t = d_i, \quad \delta_i := \min_{j \in \mathsf{I}^c} |d_i - d_j| > d_i^{3/2}(d_i - y^{1/2})^{-\frac{1}{2}} N^{-\frac{1}{2}+\epsilon}, \tag{2.5}$$

where we denote $\mathsf{I}^c := [\![1, r]\!] \setminus \mathsf{I}$. By definition, δ_t (or $\mathsf{I}(t)$) is the same for all $t \in \mathsf{I}(i)$. Finally, in case $d_i \equiv d_i(N) \to \infty$ as $N \to \infty$ for some i, we additionally assume that $|y - 1| \geq \tau_0$ for some small but fixed $\tau_0 > 0$.

REMARK 2.5. It is known that the BBP phase transition takes place in the regime $d_i - y^{1/2} \sim N^{-\frac{1}{3}}$; see for instance [9, 53, 21]. Hence, (2.4) ensures that we are in the supercritical regime. Further, note that we do not assume the spikes $d_i \equiv d_i(N)$ to be bounded in N. That means, we do allow $d_i \sim N^c$ for any $c > 0$, say. In (2.5), the first identity means that we allow d_i to be multiple. And the second inequality is the so-called *non-overlapping condition* which guarantees that the distinct (possibly multiple) d_i's are well-separated such that the eigenvalues μ_i's corresponding to distinct d_i's do not have essential overlap on the scale of fluctuation; see detailed explanation in [21] for instance. Note that the prefactor $d_i^{3/2}$ is not included in the non-overlapping condition in [21]. But this factor is needed to cover the case when the N-dependent d_i is large. We emphasize here that in reality, it can certainly happen that two distinct d_i's are close enough to violate the non-overlapping condition. However, in this case, since the fluctuation of their sample counterparts, μ_i's, have essential overlap, effective inference of d_i's based on μ_i's is believed to be impossible in general. Also, since eigenvectors are sensitive to the eigenvalue gap, in this case, inference of $\boldsymbol{v}_i$'s based on $\boldsymbol{\xi}_i$'s will also be unreliable. Therefore, the non-overlapping condition together with (2.4) can be regarded as a nearly minimal condition for a reliable detection of spikes. Finally, the restriction $|y - 1| \geq \tau_0$ when d_i diverges is purely technical and we conjecture that our result shall hold without this restriction, as illustrated in Figure B.1 of our supplementary file [12, Section B.2]. But this extension will be left as future work.

Let $\mathsf{I} \equiv \mathsf{I}(i)$ be the index set of this multiple d_i in Assumption 2.4. In order to study the generalized components of the eigenvectors of the $|\mathsf{I}|$-fold multiple d_i, we introduce

$$Z_{\mathsf{I}} := \sum_{t\in\mathsf{I}} \boldsymbol{v}_t \boldsymbol{v}_t^*, \tag{2.6}$$

the orthogonal projection onto $\mathrm{Span}\{\boldsymbol{v}_t\}_{t\in\mathsf{I}}$ and the corresponding random projection

$$\mathrm{P}_{\mathsf{I}} := \sum_{t\in\mathsf{I}} \boldsymbol{\xi}_t \boldsymbol{\xi}_t^*, \tag{2.7}$$

which is the sample counterpart of Z_{I}. Note that in case $|\mathsf{I}| > 1$, it is meaningless to do statistical inference for an individual $d_i \boldsymbol{v}_i \boldsymbol{v}_i^*$, since there is an arbitrariness in the choice of $\{\boldsymbol{v}_t\}_{t\in\mathsf{I}}$ as a basis for certain subspace. Hence, it is more natural to study Z_{I} and its sample counterpart. For any unit $\boldsymbol{w} \in \mathbb{R}^M$, denote its projection onto $\mathrm{Span}\{\boldsymbol{v}_t\}_{t\in\mathsf{I}}$ by

$$\boldsymbol{w}_{\mathsf{I}} := Z_{\mathsf{I}} \boldsymbol{w}, \tag{2.8}$$

and its weighted projection onto $\mathrm{Span}\{\boldsymbol{v}_j\}_{j\in[\![1,M]\!]\setminus\mathsf{I}}$ by

$$\boldsymbol{\varsigma}_{\mathsf{I}} := \sum_{j\in[\![1,M]\!]\setminus\mathsf{I}} \frac{d_i\sqrt{d_j+1}}{d_i-d_j} \langle \boldsymbol{w}, \boldsymbol{v}_j\rangle \boldsymbol{v}_j \tag{2.9}$$

with its normalized version

$$\boldsymbol{\varsigma}_{\mathsf{I}}^0 := \begin{cases} \boldsymbol{\varsigma}_{\mathsf{I}}/\|\boldsymbol{\varsigma}_{\mathsf{I}}\|, & \text{if } \boldsymbol{\varsigma}_{\mathsf{I}} \neq \mathbf{0}; \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{2.10}$$

2.2. *Main theorems.* In this subsection, we state the main theorems regarding the generalized components. The results involve two symmetric $(r+2)\times(r+2)$ matrices $A_{\mathsf{I}}^{\boldsymbol{w}}$ and $B_{\mathsf{I}}^{\boldsymbol{w}}$ which are used to construct the covariance matrix of a random vector in the main result (c.f. (2.12)). The expressions of $A_{\mathsf{I}}^{\boldsymbol{w}}$ and $B_{\mathsf{I}}^{\boldsymbol{w}}$ are rather involved, and thus we state their definition in (E.1) and (E.2) of [12]. In Example 2.6 below, we consider a simple rank one spiked model, for which the matrices $A_{\mathsf{I}}^{\boldsymbol{w}}$ and $B_{\mathsf{I}}^{\boldsymbol{w}}$ admit simple forms.

EXAMPLE 2.6. Consider a rank one spiked model such that $S = d\boldsymbol{v}\boldsymbol{v}^*$ in (1.3). As we will see from the statistical applications in Section 3, we are mainly interested in understanding the distribution of $\boldsymbol{w}^*\boldsymbol{\xi}_1$ with $\boldsymbol{w} = \boldsymbol{v}$ or $\boldsymbol{w} \in \{\boldsymbol{v}\}^\perp$. For these choices of $\boldsymbol{w}$, the matrices $A_{\mathsf{I}}^{\boldsymbol{w}}$ and $B_{\mathsf{I}}^{\boldsymbol{w}}$ are described below. (1): $\boldsymbol{w} = \boldsymbol{v}$. In this case, $\boldsymbol{w}_1 = \boldsymbol{v}$ and $\boldsymbol{\varsigma}_1 = \varsigma_1^0 = \mathbf{0}$. $A_1 \equiv A_1^{\boldsymbol{w}}$ is 3×3 symmetric matrix whose non-zeros entries are $A_1(1,1), A_1(2,2), A_1(3,3)$ and $A_1(2,3)$. As we will see later in Remark 2.9, only $A_1(1,1)$ will appear in the results. Finally, for $B_1 \equiv B_1^{\boldsymbol{w}}$, the only non-zero entry is $B_1(1,1)$. (2): $\boldsymbol{w} \in \{\boldsymbol{v}\}^\perp$. In this case, $\boldsymbol{w}_1 = \mathbf{0}, \boldsymbol{\varsigma}_1 = \boldsymbol{\varsigma}_1^0$. The only non-zero entries for A_1 and B_1 are $A_1(3,3)$ and $B_1(3,3)$, respectively.

THEOREM 2.7. *Suppose that Assumptions 2.3, 2.4 and the setting (1.4) hold. Fix an $i \in [\![1, r_0]\!]$ and let $\boldsymbol{w} \in S_{\mathbb{R}}^{M-1}$ be any deterministic unit vector. Then there exist random variables $\Theta_{\boldsymbol{w}_{\mathsf{I}}}^{\boldsymbol{w}}, \Lambda_{\boldsymbol{\varsigma}_{\mathsf{I}}}^{\boldsymbol{w}}, \{\Delta_{\boldsymbol{v}_t}^{\boldsymbol{w}}\}_{t\in\mathsf{I}}, \{\Pi_{\boldsymbol{v}_j}^{\boldsymbol{w}}\}_{j\in\mathsf{I}^c}$ such that $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle$ admits the following expansion*

$$\begin{aligned} \langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle = {} & \frac{d_i^2 - y}{d_i(d_i+y)} \langle \boldsymbol{w}, \mathrm{Z}_{\mathsf{I}}\boldsymbol{w}\rangle + \frac{1}{\sqrt{N(d_i^2-y)}} \Theta_{\boldsymbol{w}_{\mathsf{I}}}^{\boldsymbol{w}} + \frac{\|\boldsymbol{\varsigma}_{\mathsf{I}}\|\sqrt{d_i - y^{1/2}}}{\sqrt{N} d_i} \Lambda_{\boldsymbol{\varsigma}_{\mathsf{I}}}^{\boldsymbol{w}} \\ & + \frac{\|\boldsymbol{\varsigma}_{\mathsf{I}}\|^2}{N d_i} \sum_{t\in\mathsf{I}} (\Delta_{\boldsymbol{v}_t}^{\boldsymbol{w}})^2 - \frac{1}{N} \sum_{j\in\mathsf{I}^c} \frac{d_i d_j}{(d_i - d_j)^2} (\Pi_{\boldsymbol{v}_j}^{\boldsymbol{w}})^2 \end{aligned}$$

$$+O_{\prec}\Big(\frac{1}{N^{\frac{1}{2}+\varepsilon}}\Big(\frac{\|\boldsymbol{w}_{\mathsf{I}}\|^2}{\sqrt{d_i^2-y}}+\|\boldsymbol{w}_{\mathsf{I}}\|\|\boldsymbol{\varsigma}_{\mathsf{I}}\|\frac{\sqrt{d_i-y^{1/2}}}{d_i}\Big)\Big)$$

$$+O_{\prec}\Big(\frac{1}{N^{1+\varepsilon}}\Big(\frac{\|\boldsymbol{\varsigma}_{\mathsf{I}}\|^2}{d_i}+\|\boldsymbol{w}_{\mathsf{I}}\|^2\sum_{j\in\mathsf{I}^c}\frac{d_id_j}{(d_i-d_j)^2}\Big)\Big) \tag{2.11}$$

for some small constant $\varepsilon>0$, and

$$\Big(\Theta^{\boldsymbol{w}}_{\boldsymbol{w}_{\mathsf{I}}},\Lambda^{\boldsymbol{w}}_{\boldsymbol{\varsigma}_{\mathsf{I}}},\{\Delta^{\boldsymbol{w}}_{\boldsymbol{v}_t}\}_{t\in\mathsf{I}},\{\Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j}\}_{j\in\mathsf{I}^c}\Big)\simeq\mathcal{N}\Big(\mathbf{0},A^{\boldsymbol{w}}_{\mathsf{I}}+\kappa_4\frac{d_i^2-y}{d_i^2}B^{\boldsymbol{w}}_{\mathsf{I}}\Big). \tag{2.12}$$

Here $\mathcal{N}(\mathbf{0},A^{\boldsymbol{w}}_{\mathsf{I}}+\kappa_4\frac{d_i^2-y}{d_i^2}B^{\boldsymbol{w}}_{\mathsf{I}})$ represents a Gaussian vector with mean $\mathbf{0}$ and covariance matrix $A^{\boldsymbol{w}}_{\mathsf{I}}+\kappa_4\frac{d_i^2-y}{d_i^2}B^{\boldsymbol{w}}_{\mathsf{I}}$ with $A^{\boldsymbol{w}}_{\mathsf{I}}$ and $B^{\boldsymbol{w}}_{\mathsf{I}}$ defined in (E.1) and (E.2) respectively.

REMARK 2.8. Here we further explain how to read off the information of the limiting behaviour of $\langle\boldsymbol{w},\mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle$ from the expansion in (2.11). The first term in the RHS of (2.11) is the first order deterministic estimator of $\langle\boldsymbol{w},\mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle$ which can be biased in the high-dimensional case, especially when d_i is a fixed constant independent of N. The second and the third terms in the RHS of (2.11) are asymptotically normal, and the fourth and fifth terms are (asymptotically) linear combinations of χ^2, according to (2.12). These four terms together describe the limiting distribution of $\langle\boldsymbol{w},\mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle-\frac{d_i^2-y}{d_i(d_i+y)}\langle\boldsymbol{w},\mathrm{Z}_{\mathsf{I}}\boldsymbol{w}\rangle$, after appropriate scaling. We further emphasize here that the sizes of the first five terms may not be comparable and it is not uniformly determined which one is the leading term in all cases. Under different choices of d_i's and $\boldsymbol{w}$, say, the leading term may change. However, in any case, the two error terms in (2.11) are always smaller than the sum of the second to the fifth terms with high probability. This can be checked easily from the sizes of the entries in the covariance matrix $A^{\boldsymbol{w}}_{\mathsf{I}}+\kappa_4\frac{d_i^2-y}{d_i^2}B^{\boldsymbol{w}}_{\mathsf{I}}$. Hence, from the expansion (2.11), one can get the limiting distribution of $\langle\boldsymbol{w},\mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle-\frac{d_i^2-y}{d_i(d_i+y)}\langle\boldsymbol{w},\mathrm{Z}_{\mathsf{I}}\boldsymbol{w}\rangle$ in all cases.

In the next remark, we show how to get the limiting distribution for some specific examples. For brevity, we introduce the following auxiliary functions for $d>0$

$$\mathtt{f}(d):=\frac{y(1+d)}{d(d+y)}\Big(1+\frac{d(1+d)}{d+y}\Big),\qquad \mathtt{g}(d):=\frac{2\sqrt{(d+1)(d+\sqrt{y})}}{d+y}, \tag{2.13}$$

$$\mathtt{h}(d):=\frac{d+1}{d+y},\qquad \mathtt{l}(d):=\frac{1+d}{\sqrt{d(d+y)}}. \tag{2.14}$$

REMARK 2.9. If $\boldsymbol{w}\in\mathrm{Span}\{\boldsymbol{v}_t\}_{t\in\mathsf{I}}$, then $\boldsymbol{w}_{\mathsf{I}}=\boldsymbol{w}$ and $\boldsymbol{\varsigma}_{\mathsf{I}}=\mathbf{0}$ (c.f. (2.8), (2.9)). Hence, the $\Lambda^{\boldsymbol{w}}_{\boldsymbol{\varsigma}_{\mathsf{I}}}$ and $\Delta^{\boldsymbol{w}}_{\boldsymbol{v}_t}$ terms vanish for all $t\in\mathsf{I}$. The conclusion of Theorem 2.7 is reduced to

$$\begin{aligned}\langle\boldsymbol{w},\mathrm{P}_{\mathsf{I}}\boldsymbol{w}\rangle=&\frac{d_i^2-y}{d_i(d_i+y)}+\frac{1}{\sqrt{N(d_i^2-y)}}\Theta^{\boldsymbol{w}}_{\boldsymbol{w}}-\frac{1}{N}\sum_{j\in\mathsf{I}^c}\frac{d_id_j}{(d_i-d_j)^2}(\Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j})^2\\&+O_{\prec}\Big(\frac{N^{-\varepsilon}}{\sqrt{N(d_i^2-y)}}+\frac{N^{-\varepsilon}}{N}\sum_{j\in\mathsf{I}^c}\frac{d_id_j}{(d_i-d_j)^2}\Big),\end{aligned} \tag{2.15}$$

for some small $\varepsilon>0$, and $(\Theta^{\boldsymbol{w}}_{\boldsymbol{w}},\{\Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j}\}_{j\in\mathsf{I}^c})$ is asymptotically Gaussian with mean $\mathbf{0}$ and covariance matrix with entries given by the RHS of the following equations

$$\mathrm{var}(\Theta^{\boldsymbol{w}}_{\boldsymbol{w}})\doteq 2y\mathtt{h}(d_i)^2(1+y\mathtt{h}(d_i)^2)+\kappa_4\frac{d_i^2-y}{d_i^2}\mathtt{f}(d_i)^2s_4(\boldsymbol{w}),$$

$$\mathrm{var}(\Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j}) \doteq \mathtt{l}(d_i)^2 + \kappa_4 \frac{d_i^2 - y}{d_i^2} \mathtt{l}(d_i)^2 s_{2,2}(\boldsymbol{v}_j, \boldsymbol{w}),$$

$$\mathrm{cov}(\Theta^{\boldsymbol{w}}_{\boldsymbol{w}}, \Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j}) \doteq \kappa_4 \frac{d_i^2 - y}{d_i^2} \mathtt{f}(d_i)\mathtt{l}(d_i) s_{1,3}(\boldsymbol{v}_j, \boldsymbol{w}),$$

$$\mathrm{cov}(\Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j}, \Pi^{\boldsymbol{w}}_{\boldsymbol{v}_{\bar{j}}}) \doteq \kappa_4 \frac{d_i^2 - y}{d_i^2} \mathtt{l}(d_i)^2 s_{1,1,2}(\boldsymbol{v}_j, \boldsymbol{v}_{\bar{j}}, \boldsymbol{w}),$$

for $j, \bar{j} \in \mathsf{I}^c$. In particular, if if $\kappa_4 = 0$, the limiting distribution of $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$ does not depend on the specific choice of $\boldsymbol{w} \in \mathrm{Span}\{\boldsymbol{v}_t\}_{t \in \mathsf{I}}$. Here we recall the notation $A_N \doteq B_N$ for $A_N = B_N(1 + o_N(1))$.

If $\boldsymbol{w} \in \mathrm{Span}\{\boldsymbol{v}_j\}_{j \in [\![1,M]\!] \setminus \mathsf{I}}$, then $\boldsymbol{w}_{\mathsf{I}} = 0$ and thus (2.11) becomes

$$\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle = \frac{\|\varsigma_{\mathsf{I}}\|^2}{N d_i} \sum_{t \in \mathsf{I}} (\Delta^{\boldsymbol{w}}_{\boldsymbol{v}_t})^2 + O_{\prec}\Big(\frac{\|\varsigma_{\mathsf{I}}\|^2}{N^{1+\varepsilon} d_i} \Big). \tag{2.16}$$

Finally, for the rank one spiked model considered in Example 2.6, we have more compact formulas. In particular, we have that for $\boldsymbol{w} = \boldsymbol{v}$,

$$\frac{\sqrt{N}}{\sqrt{\mathbb{V}}} \left(|\langle \boldsymbol{v}, \boldsymbol{\xi}_1 \rangle|^2 - \frac{d^2 - y}{d(d+y)} \right) \simeq \mathcal{N}(0, 1), \tag{2.17}$$

where $\mathbb{V}$ is defined by

$$\mathbb{V} = \left(2y\mathtt{h}(d)^2(1 + y\mathtt{h}(d)^2) + \kappa_4 \frac{d^2 - y}{d^2} \mathtt{f}(d)^2 s_4(\boldsymbol{v}) \right) \Big/ (d^2 - y).$$

Moreover, when $\boldsymbol{w} \in \{\boldsymbol{v}\}^{\perp}$, we have

$$\frac{N d |\langle \boldsymbol{w}, \boldsymbol{\xi}_1 \rangle|^2}{\ell_1} \simeq \chi^2_1, \quad \ell_1 = \frac{d+1}{d+y} (1 + \frac{d^2 - y}{d^2} \kappa_4 s_{2,2}(\boldsymbol{v}, \boldsymbol{w})). \tag{2.18}$$

Our second result is the joint eigenvalue-eigenvector distribution, i.e., joint distribution of the outlying eigenvalues and the generalized components of the associated eigenvectors. We state it for the case when d_i is simple, i.e. $\mathsf{I} = \{i\}$. For simplicity, we abbreviate the notation $\mathsf{A}_{\{i\}}$ to A_i for $\mathsf{A} = \boldsymbol{w}, \mathrm{P}, \Phi$, etc. Further, we use $\{i\}^c$ to represent $[\![1, r]\!] \setminus \{i\}$.

THEOREM 2.10. *Under the same assumptions as Theorem 2.7, with $\mathsf{I} = \{i\}$ and $\boldsymbol{w}_i = \langle \boldsymbol{w}, \boldsymbol{v}_i \rangle \boldsymbol{v}_i$, the conclusion of Theorem 2.7 for the generalized component $\langle \boldsymbol{w}, \mathrm{P}_i \boldsymbol{w} \rangle = |\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle|^2$ holds. Additionally, there exists a random variable Φ_i such that the outlying eigenvalue admits the expansion*

$$\mu_i = 1 + d_i + y + \frac{y}{d_i} + \frac{\sqrt{d_i^2 - y}}{\sqrt{N}} \Phi_i + O_{\prec}\Big(\frac{\sqrt{d_i^2 - y}}{N^{\frac{1}{2}+\varepsilon}} \Big), \tag{2.19}$$

for some small constant $\varepsilon > 0$, and

$$\Big(\Phi_i, \Theta^{\boldsymbol{w}}_{\boldsymbol{w}_i}, \Lambda^{\boldsymbol{w}}_{\varsigma_i}, \Delta^{\boldsymbol{w}}_{\boldsymbol{v}_i}, \{\Pi^{\boldsymbol{w}}_{\boldsymbol{v}_j}\}_{j \in \{i\}^c} \Big) \simeq \mathcal{N}\left(\mathbf{0}, C_i^{\boldsymbol{w}}\right).$$

Here $\mathcal{N}(\mathbf{0}, C_i^{\boldsymbol{w}})$ represents a Gaussian vector with mean $\mathbf{0}$ and covariance matrix $C_i^{\boldsymbol{w}}$ of size $r + 3$. The lower right $(r+2) \times (r+2)$ corner of $C_i^{\boldsymbol{w}}$ is given by $A_i^{\boldsymbol{w}} + \kappa_4 \frac{d_i^2 - y}{d_i^2} B_i^{\boldsymbol{w}}$ as

in (E.1) *and* (E.2)*. The entries of the first row of* $C_i^{\boldsymbol{w}}$ *is given by the RHS of the following equations*

$$\mathrm{var}(\Phi_i) \doteq (1+d_i^{-1})^2\Big(2+\kappa_4\frac{d_i^2-y}{d_i^2}s_4(\boldsymbol{v}_i)\Big),$$

$$\mathrm{cov}(\Phi_i,\Theta_{\boldsymbol{w}_i}^{\boldsymbol{w}}) \doteq 2y\mathtt{h}(d_i)^2(1+d_i^{-1})\langle \boldsymbol{w},\boldsymbol{v}_i\rangle^2+\kappa_4\frac{d_i^2-y}{d_i^2}(1+d_i^{-1})\mathtt{f}(d_i)s_{2,2}(\boldsymbol{w}_i,\boldsymbol{v}_i),$$

$$\mathrm{cov}(\Phi_i,\Lambda_{\boldsymbol{\varsigma}_i}^{\boldsymbol{w}}) \doteq \kappa_4\frac{d_i^2-y}{d_i^2}\mathtt{g}(d_i)(1+d_i^{-1})s_{1,1,2}(\boldsymbol{\varsigma}_i^0,\boldsymbol{w}_i,\boldsymbol{v}_i),$$

$$\mathrm{cov}(\Phi_i,\Delta_{\boldsymbol{v}_i}^{\boldsymbol{w}}) \doteq \kappa_4\frac{d_i^2-y}{d_i^2}\sqrt{\mathtt{h}(d_i)}(1+d_i^{-1})s_{1,3}(\boldsymbol{\varsigma}_i^0,\boldsymbol{v}_i),$$

$$\mathrm{cov}(\Phi_i,\Pi_{\boldsymbol{v}_j}^{\boldsymbol{w}}) \doteq \kappa_4\frac{d_i^2-y}{d_i^2}\mathtt{1}(d_i)(1+d_i^{-1})s_{1,1,2}(\boldsymbol{v}_j,\boldsymbol{w}_i,\boldsymbol{v}_i),\quad \textit{for } j\in\{i\}^c.$$

Here we recall the notation $A_N \doteq B_N$ *for* $A_N = B_N(1+o_N(1))$.

REMARK 2.11. Here we remark that in the supercritical regime, a generalized CLT for the eigenvalues has been established in [7] previously, for fixed d_i's which are away from $y^{1/2}$ by a constant order distance. When there is a multiple d_i in the supercritical regime with multiplicity $|\mathsf{I}|$, it is known from [7] that the corresponding eigenvalues $\{\mu_t\}_{t\in\mathsf{I}}$ will converge jointly to the eigenvalues of a $|\mathsf{I}|\times|\mathsf{I}|$ Gaussian matrix GOE. Since it is not convenient to express the distribution of the eigenvalues of this fixed-dimensional GOE and their dependence with the the generalized components of $\boldsymbol{\xi_i}$'s, we are not going to state the joint eigenvalue-eigenvector distribution in the multiple case here. Nevertheless, we will state the joint distribution of the generalized components of $\boldsymbol{\xi_i}$ and all the matrix entries of this limiting GOE in Section I of the supplement [12] which equivalently describes the joint eigenvalue-eigenvector distribution; see Proposition I.4 in [12]. Finally, we point out that since the covariance matrix of the joint eigenvalue-eigenvector distribution is provided explicitly, we can explore the independence/dependence between the eigenvalues and eigenvectors. For example, as we can conclude from Section I of [12], when the spikes are distinct and well separated, the outlier eigenvalues will be asymptotically independent. Moreover, $\{\boldsymbol{v}_i^*\boldsymbol{\xi}_i\}_{1\le i\le r}$ will also be asymptotically independent.

3. Statistical inference for principal components. In this section, we apply our results and their variants to some statistical problems. We will focus on the hypothesis testing regarding the eigenspaces of covariance matrices. Eigenspaces of covariance matrices are important in many statistical methodologies and computational algorithms. A lot of efforts have been made to infer the eigenspace of the covariance matrices in the setting $M \ll N$, for instance, see [42, 56, 69, 82, 83].

In this section, we consider a generic index set $\mathcal{I}\subset[\![1,r_0]\!]$, which may contain indices for both simple and multiple d_t's. Further, we set

$$Z_{\mathcal{I}} = \sum_{t\in\mathcal{I}}\boldsymbol{v}_t\boldsymbol{v}_t^*. \tag{3.1}$$

We remark here that $Z_{\mathcal{I}}$ shall be regarded as an extension of $Z_{\mathsf{I}(i)}$ defined in (2.6), in the sense that the former may be constituted of $\boldsymbol{v}_t$'s associated with distinct d_t's.

Specifically, in the literature [2, 6, 19, 42, 56, 69, 82, 83], researchers are particularly interested in testing the following hypothesis: for $\mathcal{I}\subset[\![1,r_0]\!]$,

$$\boldsymbol{H}_0^{(1)}:\ Z_{\mathcal{I}} = Z_0 \text{ vs } \boldsymbol{H}_a^{(1)}:\ Z_{\mathcal{I}}\neq Z_0 \tag{3.2}$$

for a given projection Z_0. For the alternative $\boldsymbol{H}_a^{(1)}$ in (3.2), we are particularly interested in testing a subset of it by considering whether Z_0 is in the complement of $Z_{\mathcal{I}}$. Specifically, the hypothesis testing problem can be formulated as

$$\boldsymbol{H}_0^{(2)} : Z_{\mathcal{I}} \perp Z_0 \text{ vs } \boldsymbol{H}_a^{(2)} : \ Z_{\mathcal{I}} \not\perp Z_0. \tag{3.3}$$

Note that (3.3) is the complement of the test considered in [84, 85] and hence it can be used to study the alternative in [84, 85]. The above hypothesis testing problems also naturally arise from applications in financial economics and biology. For instance, in [2] the authors discuss the principal eigenportfolio construction (i.e., first eigenvector) in finance and [6] is devoted to the study of some specific factor (i.e, one of the eigenvectors) in macroeconomics. Moreover, in [19], the authors propose a method to estimate the eigenvector and consider applications in gene expression. Consequently, a testing for such an estimation will be natural.

In Section 3.1, we propose accurate and powerful statistics for the aforementioned hypothesis testing problems (3.2) and (3.3) in the high dimensional regime (2.2). We construct test statistics using some plug-in estimators, which are nonlinear shrinkers of the sample eigenvalues. Consequently, the proposed statistics are adaptive to d_i's. Our test statistics for these two problems are denoted by $\mathbb{T}_1$ and $\mathbb{T}_2$, respectively, whose definitions are stated below

$$\mathbb{T}_1 = \frac{\sqrt{N}\sum_{i\in\mathcal{I}}\Big(\langle \boldsymbol{u}_i, \mathrm{P}_{\mathcal{I}}\boldsymbol{u}_i\rangle - \vartheta(\widehat{d}_i)\Big)}{\sqrt{\mathtt{V}_1(\widehat{\boldsymbol{d}}_{\mathcal{I}})}}, \qquad \mathbb{T}_2 = \frac{N\sum_{i\in\mathcal{I},j\in\mathcal{J}}\langle \boldsymbol{\xi}_i, \boldsymbol{u}_j\rangle^2}{\mathtt{q}(\widehat{\boldsymbol{d}})},$$

where we assume that $Z_0 = \sum_{i\in\mathcal{I}} \boldsymbol{u}_i\boldsymbol{u}_i^*$ in the first testing problem and $Z_0 = \sum_{j\in\mathcal{J}} \boldsymbol{u}_j\boldsymbol{u}_j^*$ for some fixed index set $\mathcal{J}$ with a family of orthonormal vectors $\{\boldsymbol{u}_j\}_{j\in\mathcal{J}}$ in the second testing problem. Here we refer to (3.6), (3.9)-(3.10) and (3.14)-(3.16) for the definitions of $\widehat{d}_i$, $\mathtt{V}_1(\widehat{\boldsymbol{d}}_{\mathcal{I}})$ and $\mathtt{q}(\widehat{\boldsymbol{d}})$, respectively. Then we derive the distributions of $\mathbb{T}_1$ and $\mathbb{T}_2$ utilizing the joint distribution of the eigenvalues and eigenvectors given in Section 2 and its extensions in Section I of [12]. It turns out $\mathbb{T}_1$ is asymptotically normal (c.f. Corollary 3.4) and the limiting distribution of $\mathbb{T}_2$ can be described by certain quadratic form of a Gaussian vector (c.f. Corollary 3.10). The rejection regions are constructed based on the distributions of the proposed statistics. In Section B.4 of the supplement [12], we also work on a real example on gene expression data to demonstrate the usefulness of our test statistic $\mathbb{T}_2$.

We mention that this methodology can be potentially applied to perform statistical inference and build up confidence intervals for other statistics related to the principal components. For instance, the loadings of principal components [50], the shrinkage of eigenvalues [36], the number of spikes [35, 75], the estimation of eigenvectors [63] and the invariant estimator for covariance matrices [25, Section 6]. These applications will be studied in the future.

3.1. *Test statistics and their asymptotic distributions.* In this section, we propose statistics to test (3.2) and (3.3). We start with (3.2). In what follows, we construct a data-dependent statistic to address the high dimensional issue. Denote $Z_0 = \sum_{i\in\mathcal{I}} \boldsymbol{u}_i\boldsymbol{u}_i^*$. We first study

$$\mathcal{T}_1 := \sum_{i\in\mathcal{I}}\Big(\langle \boldsymbol{u}_i, \mathrm{P}_{\mathcal{I}}\boldsymbol{u}_i\rangle - \vartheta(\widehat{d}_i)\Big), \tag{3.4}$$

where

$$\mathrm{P}_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^*, \quad \vartheta(d) = \frac{d^2 - y}{d(d+y)} \tag{3.5}$$

and $\widehat{d}_i$ is a nonlinear shrinkage of the sample eigenvalues defined by

$$\widehat{d}_i = \gamma(\mu_i),\ \gamma(x) = \frac{1}{2}(-y+x-1) + \frac{1}{2}\sqrt{(-y+x-1)^2 - 4y}. \tag{3.6}$$

We remark here that $\mathrm{P}_{\mathcal{I}}$ shall be regarded as an extension of $\mathrm{P}_{\mathsf{I}(i)}$ defined in (2.7), in the sense that the former may be constituted of $\boldsymbol{\xi}_t$'s associated with distinct d_t's. Further, we remark, according to the definition in (3.4), the statistic $\mathcal{T}_1$ does not depend on the specific choice of the basis $\{\boldsymbol{u}_i\}$ of Z_0. Hence, we have a freedom to choose any basis $\{\boldsymbol{u}_i\}$ of Z_0 in the sequel.

Since we are studying general $\mathcal{I}$ in this section, the indices in $\mathcal{I}$ may not belong to the same multiple d_t's. To facilitate our discussion in the sequel, we do a decomposition of $\mathcal{I}$ into subsets with each consisting of the indices for one multiple (or simple) d_t. For $\mathcal{I} = \{i_1, \cdots, i_{r_*}\} \subset [\![1, r_0]\!]$, we assume that $\mathcal{I} = \bigcup_{k=1}^{\ell} \mathcal{I}_k$ for some fixed integer ℓ such that $\mathcal{I}_k \cap \mathcal{I}_j = \varnothing$ for $k \neq j \in [\![1, \ell]\!]$. We assume that (2.5) holds for all the $d_i, i \in \mathcal{I}$. For each $i \in \mathcal{I}$, there is a $k_i \in [\![1, \ell]\!]$, such that $i \in \mathcal{I}_{k_i}$. Moreover, we suppose that for $1 \leq k \leq \ell$, $d_t, t \in \mathcal{I}_k$ are all the same, and $d_i \neq d_j$ for $i \in \mathcal{I}_{k_i}, j \notin \mathcal{I}_{k_i}$. Note that by definition $\mathcal{I}_{k_i} \equiv \mathsf{I}(i)$ (c.f. Assumption 2.4). Further note that $\ell = r_*$ corresponds to the case that all the spikes in $\mathcal{I}$ are simple , $\ell = 1$ corresponds to the case that all the spikes are equal and $1 < \ell < r_*$ corresponds to a mixture case.

For brevity, in the first testing problem (3.2), we further restrict ourselves to the case satisfying the following assumption.

ASSUMPTION 3.1. *Let the index set $\mathcal{I} \subset [\![1, r_0]\!]$ be defined above. We assume that for any $i \in \mathcal{I}$, the following inequality holds*

$$\frac{1}{N} \sum_{j \in (\mathsf{I}(i))^c} \frac{d_i d_j}{(d_i - d_j)^2} \leq N^{-\varepsilon} \frac{1}{\sqrt{N}(d_i^2 - y)}. \tag{3.7}$$

for some small but fixed $\varepsilon > 0$. Here $i \in \mathcal{I}_{k_i} \equiv \mathsf{I}(i)$ for some $k_i \in [\![1, \ell]\!]$.

REMARK 3.2. We remark here that the inequality (3.7) ensures that the χ^2 terms in (2.15) are suppressed by the Gaussian term. We impose such a condition in order to simplify the discussion in the application part, thanks to the simplicity of the Gaussianity. But our result can be applied without this additional assumption. In the general case, we need to work with a linear combination of Gaussian and χ^2 variables. For brevity, we omit such a general discussion and leave it to the future work.

We record the results regarding the asymptotic distribution of (3.4) in the following theorem and postpone its proof to Section I of [12].

THEOREM 3.3. *Suppose that Assumptions 2.3 , 2.4, and the setting (1.4) hold. Suppose that $\boldsymbol{H}_0^{(1)}$ of (3.2) and Assumption 3.1 hold. For the statistic (3.4), we have that*

$$\frac{\sqrt{N}\mathcal{T}_1}{\sqrt{\mathtt{V}_1(\boldsymbol{d}_{\mathcal{I}})}} \simeq \mathcal{N}(0, 1), \tag{3.8}$$

where $\mathtt{V}_1(\boldsymbol{d}_{\mathcal{I}}), \boldsymbol{d}_{\mathcal{I}} = (d_{i_1}, \cdots, d_{i_{r_*}})$ is defined as

$$\mathtt{V}_1(\boldsymbol{d}_{\mathcal{I}}) := \boldsymbol{\alpha}^* C_{\mathcal{I}} \boldsymbol{\alpha}.$$

Here $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{2r_*})^* \in \mathbb{R}^{2r_*}$ is defined as

$$\alpha_k = \begin{cases} -y(d_{i_k}^2 + 2d_{i_k} + y)(d_{i_k} + y)^{-2}(d_{i_k}^2 - y)^{-\frac{1}{2}}, & 1 \leq k \leq r_*; \\ (d_{i_{k-r_*}}^2 - y)^{-\frac{1}{2}}, & r_* + 1 \leq k \leq 2r_*, \end{cases}$$

and $C_{\mathcal{I}}$ is a positive definite matrix of dimension $2r_*$ and explicitly defined in Proposition I.1 of [12]. Particularly, when all the spikes $d_t, t \in \mathcal{I}$ are equal to d_e, i.e., $\mathcal{I} = \mathsf{I}(i)$ for some i, we have that

$$
\begin{aligned}
\mathtt{V}_1(\boldsymbol{d}_{\mathcal{I}}) &= 2(d_e^2-y)^{-1}\Big(y\mathtt{h}(d_e)^2|\mathcal{I}| - \frac{y(d_e^2+2d_e+y)(1+d_e)}{(d_e+y)^2 d_e}\Big)^2 \\
&\quad + 2(d_e^2-y)^{-1}\Big(y\mathtt{h}(d_e)^2|\mathcal{I}| + y^2\mathtt{h}(d_e)^4(|\mathcal{I}|-|\mathcal{I}|^2)\Big) \\
&\quad + \kappa_4\Big(\frac{\mathtt{f}(d_e)}{d_e} - \frac{y(d_e^2+2d_e+y)(1+d_e)}{(d_e+y)^2 d_e^2}\Big)^2 \sum_{k,t\in\mathcal{I}} s_{2,2}(\boldsymbol{v}_k,\boldsymbol{v}_t),
\end{aligned} \tag{3.9}
$$

where $\mathtt{h}(\cdot)$ *and* $\mathtt{f}(\cdot)$ *are defined in (2.13) and (2.14).*

By Theorems 3.3 and 2.10, we can construct a pivotal statistic. Rewrite

$$
\mathbb{T}_1 = \frac{\sqrt{N}\mathcal{T}_1}{\sqrt{\mathtt{V}_1(\widehat{\boldsymbol{d}}_{\mathcal{I}})}}, \quad \widehat{\boldsymbol{d}}_{\mathcal{I}} := (\gamma(\mu_{i_1}),\cdots,\gamma(\mu_{i_{r_*}})). \tag{3.10}
$$

We mention that (3.10) is adaptive to the d_i's by utilizing their estimators (3.6). We summarize the distribution of $\mathbb{T}_1$ in the corollary below.

COROLLARY 3.4. *Under the assumptions of Theorem 3.3, we have that*

$$
\mathbb{T}_1 \simeq \mathcal{N}(0,1).
$$

Since $\mathbb{T}_1$ is asymptotically pivotal, we will use (3.10) as our statistic for the testing of (3.2). For an illustration, we record the behavior of our statistic for a single spike model (i.e., $r_0 = r_* = 1$) in Figure 1. The more general and extensive simulations will be conducted in Section 3.2. We find that under the null hypothesis of (3.2), our proposed statistic is close to $\mathcal{N}(0,1)$ for different values of d and hence it is suitable for the hypothesis testing problem (3.2). We mention that even though we have not justified the case d_i diverges under the assumption $y = 1$ theoretically, our statistic is still accurate and powerful according to empirical illustrations for this case. Hence, in the sequel, we also present the simulation results for the case $y = 1$.

In what follows, we provide a few examples with explicit formulas of $\mathtt{V}_1(\cdot)$. These will be used for the simulations in Section 3.2.

EXAMPLE 3.5. We consider that $\mathcal{I} = \{1,2\}$, both d_1 and d_2 are simple and $\boldsymbol{v}_i = \boldsymbol{e}_i, i = 1,2$. In this case, we have that

$$
\mathtt{V}_1(d_1,d_2) = \sum_{i=1}^{2} \boldsymbol{\alpha}^* \begin{bmatrix} A_{11}(d_i) & A_{12}(d_i) \\ A_{21}(d_i) & A_{22}(d_i) \end{bmatrix} \boldsymbol{\alpha},
$$

where

$$
\begin{aligned}
\boldsymbol{\alpha} &= \Big(-\frac{y(d_i^2+2d_i+y)}{(d_i+y)^2(d_i^2-y)^{\frac{1}{2}}}, (d_i^2-y)^{-\frac{1}{2}}\Big)^* \\
A_{11}(d_i) &= 2(1+d_i^{-1})^2 + \kappa_4(1-yd_i^{-2})(1+d_i^{-1})^2, \\
A_{12}(d_i) &= 2y\mathtt{h}(d_i)^2(1+d_i^{-1}) + \kappa_4(1-yd_i^{-2})\mathtt{f}(d_i)(1+d_i^{-1}), \\
A_{22}(d_i) &= 2y\mathtt{h}(d_i)^2(1+y\mathtt{h}(d_i)^2) + \kappa_4(1-yd_i^{-2})\mathtt{f}(d_i)^2,
\end{aligned}
$$

and the functions $\mathtt{h}(\cdot)$ and $\mathtt{f}(\cdot)$ are defined in (2.13) and (2.14).

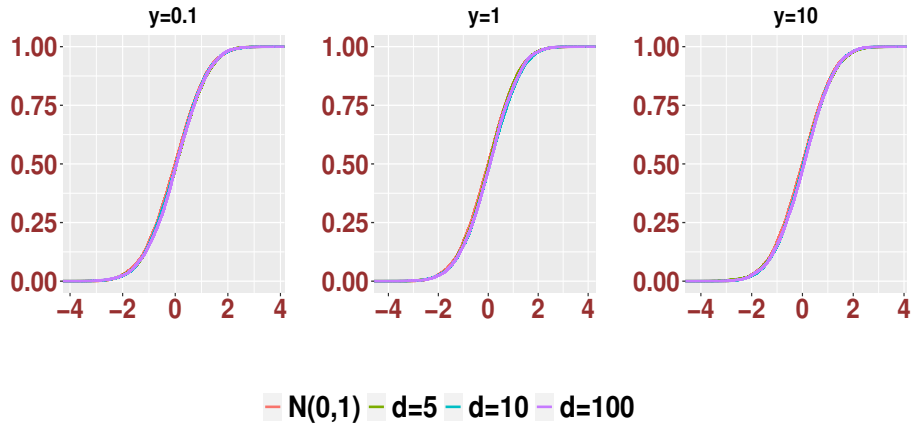

Fig 1: Simulated empirical cumulative distribution function (ECDF) for the proposed statistic (3.10) under null of (3.2) with $r_0 = r_* = 1$. Here, the spiked covariance matrix is denoted as $\Sigma = \mathrm{diag}(d+1, 1, \cdots, 1)$ and we use the statistic $\sqrt{N}(|\langle \boldsymbol{\xi}_1, \boldsymbol{e}_1\rangle|^2 - \vartheta(\widehat{d}))/\sqrt{\mathtt{V}_1(\widehat{d})}$, where $\mathtt{V}_1(d) = \frac{1}{2}\mathtt{V}_1(d,d)$; see Example 3.5 for the definition of $\mathtt{V}_1(\cdot,\cdot)$. Here $N = 500$ and we report our results based on 8,000 simulations with Gaussian random variables.

EXAMPLE 3.6. We consider the case that $\mathcal{I} = \{1,2\}$ with $d_1 = d_2 = d$ and $\boldsymbol{v}_i = \boldsymbol{e}_i, i = 1,2$. In this case, we have that

$$\begin{aligned}\mathtt{V}_1(d,d) =& 2(d^2-y)^{-1}\Big(2y\mathtt{h}(d)^2 - \frac{y(d^2+2d+y)(1+d)}{(d+y)^2 d}\Big)^2 \\ &+ 2(d^2-y)^{-1}\Big(2y\mathtt{h}(d)^2 - 2y^2\mathtt{h}(d)^4\Big) \\ &+ \kappa_4\Big(\frac{\mathtt{f}(d)}{d} - \frac{y(d^2+2d+y)(1+d)}{(d+y)^2 d^2}\Big)^2 \sum_{k,t\in\mathcal{I}} s_{2,2}(\boldsymbol{v}_k, \boldsymbol{v}_t).\end{aligned}$$

REMARK 3.7. We provide some remarks on the asymptotic power of the statistic $\mathbb{T}_1$ for a rank-one spiked model assuming that $r = r_0 = 1$ as in Assumption 2.4. Note that for any given nominal level (i.e., type I error rate) α, according to Corollary 3.4, our critical region is constructed as $\{|\mathbb{T}_1| > z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard Gaussian distribution. Recall that in this setting $w_k = \langle \boldsymbol{v}_0, \boldsymbol{v}_k\rangle, 1 \le k \le M$. Under the alternative $\boldsymbol{H}_a^{(1)}$, our test statistic can be decomposed as

$$\begin{aligned}\mathbb{T}_1 &= \frac{\sqrt{N}\left(\left\langle \boldsymbol{\xi}_1, w_1\boldsymbol{v}_1 + \sum_{k=2}^M w_2\boldsymbol{v}_k\right\rangle^2 - \vartheta(\widehat{d}_1)\right)}{\sqrt{\mathtt{V}_1(\widehat{d}_1)}} = \frac{\sqrt{N}\left(\langle \boldsymbol{\xi}_1, w_1\boldsymbol{v}_1\rangle^2 - \vartheta(\widehat{d}_1)\right)}{\sqrt{\mathtt{V}_1(\widehat{d}_1)}} + O_\prec(1) \\ &= \frac{\sqrt{N}\left(\langle \boldsymbol{\xi}_1, \boldsymbol{v}_1\rangle^2 - \vartheta(\widehat{d}_1)\right)}{\sqrt{\mathtt{V}_1(\widehat{d}_1)}} + (w_1^2 - 1)\frac{\sqrt{N}\langle \boldsymbol{\xi}_1, \boldsymbol{v}_1\rangle^2}{\sqrt{\mathtt{V}_1(\widehat{d}_1)}} + O_\prec(1) =: \mathbb{T}_{11} + \mathbb{T}_{12} + O_\prec(1).\end{aligned}$$

According to Corollary 3.4, $\mathbb{T}_{11}$ is asymptotically Gaussian. Since $\langle \boldsymbol{\xi}_1, \boldsymbol{v}_1\rangle^2 \asymp 1$ with high probability, if we assume that for some $0 < \delta \le 1/2$

$$1 - w_1^2 \ge \sqrt{\mathtt{V}_1(\widehat{d}_1)}N^{-1/2+\delta}, \tag{3.11}$$

then with high probability $|\mathbb{T}_{12}| \to \infty$ as $N \to \infty$. Consequently, under the nominal level α, we need to reject $\boldsymbol{H}_0^{(1)}$. This implies that once we have a small deviation from the null hypothesis, our statistic will be able to reject it.

Next, we consider the hypothesis testing problem (3.3). In this case, we further assume that the true model or the population matrix Σ only contains supercritical spikes, i.e.,

$$r_0 = r. \tag{3.12}$$

It will be seen that if there exist subcritical spikes, one will need to provide a plug-in estimator of the subcritical d_i's in order to raise a test statistic which is adaptive to all the spiked eigenvalues. However, it is well-known now that an effective detection of subcritical d_i's based on μ_i's is impossible in general [65, 66, 67, 78], unless one employs additional information such as the structure of $\boldsymbol{v}_i$'s [88]. And also, indeed, in many applications, d_i's are very large and even divergent, and thus are certainly supercritical. Hence, in the sequel, we will focus on the case when (3.12) is satisfied. Nevertheless, we would like to mention the problem of detecting weak spikes has also been considered in the literature. For example, assuming X is Gaussian, the power of likelihood ratio test was studied in [71, 72]. Later on, the maximal asymptotic power among tests based on linear spectral statistics was studied in [33] for a more general spiked model which recovers [71, 72] as a special case.

Suppose that in this case $Z_0 = \sum_{j\in\mathcal{J}} \boldsymbol{u}_j \boldsymbol{u}_j^*$ for some fixed index set $\mathcal{J}$ and $\{\boldsymbol{u}_j\}_{j\in\mathcal{J}}$ is a family of orthonormal vectors. We define the following test statistic

$$\mathcal{T}_2 = \sum_{i\in\mathcal{I}, j\in\mathcal{J}} \langle \boldsymbol{\xi}_i, \boldsymbol{u}_j\rangle^2. \tag{3.13}$$

The asymptotic distribution of $\mathbb{T}_2$ is recorded in the following theorem. It turns out that its asymptotic distribution coincides with linear combinations of χ^2 variables. For convenience, we first define $\boldsymbol{d} := (d_1, \ldots, d_r)$ and

$$\mathtt{q}(\boldsymbol{d}) := \max_{i\in\mathcal{I}, j\in\mathcal{J}} \sum_{k\in[\![1,M]\!]\setminus\mathcal{I}} \mathtt{h}(d_i) \frac{d_i(d_k+1)}{(d_i-d_k)^2} \langle \boldsymbol{u}_j, \boldsymbol{v}_k\rangle^2 \tag{3.14}$$

which depends on the subspace Z_0 and all the d_i's for $i \in [\![1, r]\!]$. Here $d_{r+1} = \cdots d_M = 0$. We emphasize that all d_i's for $i \in [\![1, r]\!]$ satisfy (2.4) and (2.5) in this part.

THEOREM 3.8. *Suppose that Assumptions 2.3 , 2.4, and the settings (1.4) and (3.12) hold. Suppose that $\boldsymbol{H}_0^{(2)}$ of (3.3) holds true. For the statistic $\mathcal{T}_2$ defined in (3.13), we have*

$$\frac{N\mathcal{T}_2}{\mathtt{q}(\boldsymbol{d})} \simeq \frac{\boldsymbol{g}^*\boldsymbol{U}\boldsymbol{g}}{\mathtt{q}(\boldsymbol{d})}, \tag{3.15}$$

where $\boldsymbol{g} \in \mathbb{R}^{|\mathcal{I}||\mathcal{J}|}$, $\boldsymbol{g} \sim \mathcal{N}(0, I_{|\mathcal{I}||\mathcal{J}|})$, and $\boldsymbol{U} \equiv \boldsymbol{U}(\boldsymbol{d})$ is a symmetric matrix of dimension $|\mathcal{I}||\mathcal{J}|$ defined explicitly in Proposition I.1 of [12].

REMARK 3.9. We remark that in (3.15), the factor $1/\mathtt{q}(\boldsymbol{d})$ on both sides is used to scale the quantities to order one, since the notation " $\simeq$ " (c.f. Definition 2.2) requires the tightness.

The results of Theorem 3.8, especially $\mathtt{q}(\boldsymbol{d})$ and $\boldsymbol{U}$, still contain the values of $d_i, i \in \mathcal{I}$ and also the other nonzero spikes $d_j, j \in \mathcal{I}^c$ which are all supercritical (c.f. (3.12)) so that we can use (3.6) to estimate them all. To construct a data-dependent statistic, we can use the plug-in estimator (3.6) to generate critical values of the hypothesis testing (3.3) using the samples.

COROLLARY 3.10. *Under the assumptions of Theorem 3.8, we have that*

$$\mathbb{T}_2 := \frac{N\mathcal{T}_2}{\mathtt{q}(\widehat{\boldsymbol{d}})} \simeq \frac{\boldsymbol{g}^*\widehat{\boldsymbol{U}}\boldsymbol{g}}{\mathtt{q}(\widehat{\boldsymbol{d}})}, \qquad \textit{where} \quad \widehat{\boldsymbol{U}} := \boldsymbol{U}(\widehat{\boldsymbol{d}}), \quad \textit{and} \quad \widehat{\boldsymbol{d}} := (\widehat{d}_1, \ldots, \widehat{d}_r). \tag{3.16}$$

We can use our statistic $\mathbb{T}_2$ with the critical values generated from Corollary 3.10 to study the hypothesis testing problem (3.3). Here we shall point out that although our statistic $\mathbb{T}_2$ is adaptive to the d_i's, it is nevertheless dependent on $\{\boldsymbol{v}_i\}_{i\in[\![1,r]\!]\setminus\mathcal{I}}$ and also a κ_4 term which involves some $\{\boldsymbol{v}_i\}_{i\in\mathcal{I}}$-dependent parameters of the form $s_{1,1,1,1}(\cdot,\cdot,\cdot,\cdot)$; see the definition of $\boldsymbol{U}$ in Proposition I.1 in [12]. Hence, first of all, we shall only apply our statistic $\mathbb{T}_2$ in case either $\{\boldsymbol{v}_i\}_{i\in[\![1,r]\!]\setminus\mathcal{I}}$ is known a priori, or $\mathcal{I} = [\![1,r]\!]$ such that the set $\{\boldsymbol{v}_i\}_{i\in[\![1,r]\!]\setminus\mathcal{I}}$ is empty. In practice, this restriction is mild and fits the following real scenario: if some of the $\boldsymbol{v}_i$'s are already known, we only need to do inference for those unknown $\boldsymbol{v}_i$'s, while if none of the $\boldsymbol{v}_i$ is known a priori, we consider the inference for all $\{\boldsymbol{v}_i\}_{i\in[\![1,r]\!]}$ together. Certainly, it is also natural to consider a part of $\boldsymbol{v}_i$'s even if none of them is known, as what we did in the test (3.2). Nevertheless, due to the restriction of the theoretical result, we focus on the aforementioned scenario, which is more restricted but still very natural. Second, the unknown κ_4 term will be absent in case we consider the Gaussian matrix X which is often the case in reality. Hence, in the Gaussian case, we can apply our statistic directly if we restrict ourself to the aforementioned scenario of $\boldsymbol{v}_i$'s. Nevertheless, for the reader's reference, we also present our simulation study in our supplementary file [12] for the two-point case, as if the additional parameter, the κ_4 term, was known a priori. Here, we remark that the restriction of scenario to apply our theoretical result for the test (3.3), which is not necessary for (3.2), is actually quite reasonable. Note that, in (3.2), the necessary parameters from $\{\boldsymbol{v}_i\}_{i\in\mathcal{I}}$ are completely fixed by the null hypothesis, and the distribution of our statistic (under the null hypothesis) can be expressed in terms of these given parameters. However, in the test (3.3), our null hypothesis is $Z_{\mathcal{I}} \perp Z_0$. In case that the rank of the projection of the given Z_0 is small, as a low-rank subspace living in the complement of Z_0, $Z_{\mathcal{I}}$ can have many choices and thus there is a big uncertainty on the unknown parameters of $Z_{\mathcal{I}}$ which cannot be fixed by Z_0.

For an illustration, we record the behavior and power of our statistic for a single spiked model (i.e., $r_0 = r = 1$) in Figure 2. We find that our proposed statistic is close to Chi-square distribution with one degree of freedom for various values of d and hence it can be applied for testing (3.3) for this single spike model. For more general case, the asymptotic distribution of (3.13) is a linear combination of Chi-square distributions. We will conduct extensive simulations in Section 3.2.

We next consider a few examples to specify the asymptotic distribution stated in Theorem 3.8 and the results will be used in Section 3.2.

EXAMPLE 3.11. We consider that $r_0 = 3$, $\mathcal{I} = \{1,2\}$, $d_i, i = 1,2,3$ are simple and satisfy (2.4), (2.5) and $\boldsymbol{v}_i = \boldsymbol{e}_i, i = 1,2,3$ and $Z_0 = \boldsymbol{e}_3\boldsymbol{e}_3^* + \boldsymbol{e}_4\boldsymbol{e}_4^*$. In this case, since $\boldsymbol{\varsigma}_{\{1\}}^{\boldsymbol{e}_3} = \frac{d_1\sqrt{d_3+1}}{d_1-d_3}\boldsymbol{e}_3$, $\boldsymbol{\varsigma}_{\{2\}}^{\boldsymbol{e}_3} = \frac{d_2\sqrt{d_3+1}}{d_2-d_3}\boldsymbol{e}_3$ and $\boldsymbol{\varsigma}_{\{1\}}^{\boldsymbol{e}_4} = \boldsymbol{\varsigma}_{\{2\}}^{\boldsymbol{e}_4} = \boldsymbol{e}_4$ and $s_{1,1,1,1}(\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}) = 0$ for $i_{1,2} = 1,2, j_{1,2} = 3,4$. Further, $\mathtt{q}(\boldsymbol{d}) = \mathtt{q}(d_1,d_2,d_3) = \max\{\frac{(d_3+1)d_1\mathtt{h}(d_1)}{(d_1-d_3)^2}, \frac{\mathtt{h}(d_1)}{d_1}, \frac{(d_3+1)d_2\mathtt{h}(d_2)}{(d_2-d_3)^2}, \frac{\mathtt{h}(d_2)}{d_2}\}$. We have that the statistic $N\mathcal{T}_2/\mathtt{q}(\boldsymbol{d})$ will be asymptotically distributed as

$$\frac{1}{\mathtt{q}(\boldsymbol{d})}\boldsymbol{g}^* \operatorname{diag}\left(\frac{(d_3+1)d_1\mathtt{h}(d_1)}{(d_1-d_3)^2}, \frac{\mathtt{h}(d_1)}{d_1}, \frac{(d_3+1)d_2\mathtt{h}(d_2)}{(d_2-d_3)^2}, \frac{\mathtt{h}(d_2)}{d_2}\right)\boldsymbol{g},$$

where $\boldsymbol{g} \in \mathbb{R}^4$ is a standard Gaussian random vector, i.e., $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_4)$.

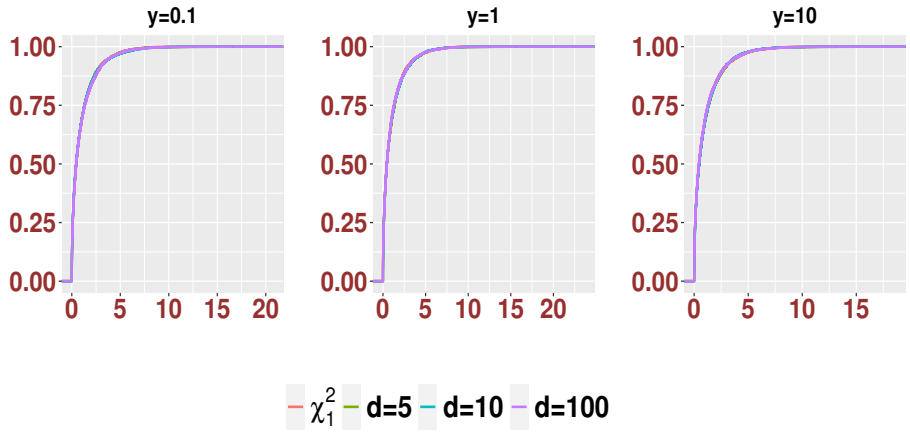

Fig 2: Simulated empirical cumulative distribution function (ECDF) for the proposed statistic under null of (3.3) with $r_0 = 1$. Here, the spiked covaraince matrix is denoted as $\Sigma = \mathrm{diag}\{d+1, 1, \cdots, 1\}$ and $Z_0 = \boldsymbol{e}_3$ in (3.3). We use the statistic $N\widehat{d}(\widehat{d}+y)|\langle \boldsymbol{\xi}_1, \boldsymbol{e}_3\rangle|^2/(\widehat{d}+1)$. Here $\widehat{d} = \gamma(\mu_1), N = 500$ and we report our results based on 8,000 simulations with Gaussian random variables.

EXAMPLE 3.12. We consider that $r_0 = 3$, $\mathcal{I} = \{1, 2\}, d_1 = d_2 = d, d_3$ is distinct from d by a distance of order 1, and $\boldsymbol{v}_i = \boldsymbol{e}_i, i = 1, 2, 3$. In this case, $\mathtt{q}(\boldsymbol{d}) = \max\{\frac{(d_3+1)d\mathtt{h}(d)}{(d-d_3)^2}, \frac{\mathtt{h}(d)}{d}\}$. We have that the statistic $N\mathcal{T}_2/\mathtt{q}(\boldsymbol{d})$ will be asymptotically distributed as

$$\frac{1}{\mathtt{q}(\boldsymbol{d})}\boldsymbol{g}^* \,\mathrm{diag}\left(\frac{(d_3+1)d\mathtt{h}(d)}{(d-d_3)^2}, \frac{\mathtt{h}(d)}{d}, \frac{(d_3+1)d\mathtt{h}(d)}{(d-d_3)^2}, \frac{\mathtt{h}(d)}{d}\right)\boldsymbol{g},$$

where $\boldsymbol{g} \in \mathbb{R}^4$ is a Gaussian random vector such that $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_4)$.

3.2. *Simulation studies.* In this subsection, we perform extensive Monte Carlo simulations to study the finite-sample accuracy and power of our proposed statistics. We focus on (3.2). The discussion on (3.3) can be found in Section B.5-B.7 of [12]. For (3.2), we not only report our results but also compare them with the existing statistics in the literature. We will call our statistics as `Fr-Adaptive`. Moreover, we use `Fr-boostrap` to represent the bootstrapping method using the Frobenius norm proposed in [69], `Fr-Bayes` to represent the frequentist Bayes using the Frobenius norm proposed in [83], `Fr-DataDriven` to represent the sample splitting method using the Frobenius norm proposed in [56], `HPV-LeCam` to represent the Le Cam optimal test proposed in [42], `En-bootstrap` and `En-Bayes` to represent the bootstrapping method and the frequentist Bayes method respectively using the power-enhanced norm introduced in [82] with $s_1 = s_2 = 1$ (see Definition 3.1 of [82]), and `Sp-bootstrap` and `Sp-Bayes` the bootstrapping method and the frequentist Bayes method respectively using the spectral norm. In the following discussion, we compare the performance of our `Fr-Adaptive` with all the aforementioned statistics.

In all the following simulations, we conduct 2,000 Monte-Carlo repetitions for the bootstrapping and frequentist Bayes procedure. For the accuracy of the tests, we focus on reporting the results with the type I error rate 0.1 under different values of $y = 0.1, 1, 10$ and various choices of the spikes. Moreover, we consider the following Scenario I to illustrate the usefulness and generality of our results. In Section B.6 of our supplement [12], we conduct more simulations when some of the spikes are equal, which will be called Scenario II.

Scenario I: We consider the case $r_0 = 3$ with $d_1 = d + 7$, $d_2 = 7$ and $d_3 = 5$, where d takes a variety of values. We consider the hypothesis testing for the eigenspace of $\Sigma =$

$I + \sum_{i=1}^{3} d_i \boldsymbol{e}_i \boldsymbol{e}_i^*$ with $\mathcal{I} = \{1, 2\}$, where the null is

$$Z_0 = \boldsymbol{e}_1 \boldsymbol{e}_1^* + \boldsymbol{e}_2 \boldsymbol{e}_2^*. \tag{3.17}$$

In this scenario, the spiked eigenvalues are simple. We will consider the standard Gaussian distribution and the two-point distribution $\frac{1}{3}\delta_{\sqrt{2}} + \frac{2}{3}\delta_{-\frac{1}{\sqrt{2}}}$ as the distribution of entries of X. We mention that the asymptotic distribution of our statistic (3.10) under the null hypothesis of (3.2) has been established in Examples 3.5 and 3.6 for Scenarios I and II, respectively.

For both of the two scenarios, we consider the alternative

$$Z_a = \boldsymbol{v}_1(\varphi)\boldsymbol{v}_1(\varphi)^* + \boldsymbol{v}_2(\varphi)\boldsymbol{v}_2(\varphi)^*, \tag{3.18}$$

where for $\varphi \in [0, \frac{\pi}{2}]$

$$\boldsymbol{v}_1(\varphi) = \cos\varphi \boldsymbol{e}_1 + \sin\varphi \boldsymbol{e}_4, \ \boldsymbol{v}_2(\varphi) = \cos\varphi \boldsymbol{e}_2 + \sin\varphi \boldsymbol{e}_5.$$

Note that $\varphi = 0$ corresponds to the null case of (3.17). It is easy to see that $\kappa_4 = -1.5$ for the two-point distribution and $\kappa_4 = 0$ for standard Gaussian random variable.

For Scenario I, from Tables 1–3, we find that our proposed statistic (3.10) is very accurate even for small values of N and d. Moreover, our statistic reaches accuracy regardless of the values of y. In contrast, for the other methods in the literature, we find that all of them lose their accuracy when y increases (i.e. M increases). Moreover, we find that most of the methods are conservative except for the Le Cam test. Finally, we find that some of the methods, especially the frequentiest Bayes method with Frobenius norm, spectral norm or the power-enhanced norm in [82] are also reasonably accurate for large values of d when y is small. For Scenario II when some spikes are equal, we can obtain similar results as reported in Section B.6 of [12]. For the two-point distribution, the results can be bound in Section B.7 of [12].

In summary, our proposed statistic (3.10) is quite accurate for different values of d satisfying Assumption 2.4, even for small and multiple ones. This accuracy is robust against different values of y. As summarized in [82, Section 7.4], all the previous methods in the literature request that $M \ll N$. Therefore, when y increases (i.e., M diverges faster), we find that our method performs better than all the other methods. Indeed, all our current results can be extended to the regime $\log M \asymp \log N$ following the discussion of [21, 24]. We will pursue this direction in the future work. Moreover, since the computational complexity of the aforementioned methods depends on a polynomial order of the dimensionality M (see [82, Section 7.4]), they can be computationally intensive as M diverges faster. In contrast, our method works faster since it only depends on the sample eigenvalues and eigenvectors.

Then we compare the power of the above statistics under the alternative (3.18) for different values of φ regarding Scenario I. First of all, we apply the above statistics directly for $d = 5, 50$, respectively, in Figures B.12 and B.13 of [12]. We find that our method is very powerful even when y is relatively large and d is relatively small, and it outperforms the other methods. When y is small and d is large, even though for larger values of φ, many methods can obtain high power, we find that our proposed statistic (3.10) is quite powerful even under a relatively weak alternative, i.e., smaller values of φ. Second, as one can see from Tables 1, 2 and 3, the existing methods in the literature are inaccurate. Consequently, if we compare the power of all the methods by constructing the rejection regions using the limiting distributions, it is not very informative. To address this issue, we do more experiments and compare the power of all the methods according to the same type I error rate 0.1. In particular, the rejection regions of all the methods are constructed using the simulated critical values based on 5,000 Monte Carlo repetitions so their type I error rates are exactly 0.1. In Figures B.2, B.3 and B.4 of [12], we reported the results for $y = 0.1, y = 1$ and $y = 10$ with various choices of d's, respectively. We can conclude that our method is still the most

	$N = 200$					$N = 500$				
Method	$d = 2$	$d = 5$	$d = 10$	$d = 50$	$d = 100$	$d = 2$	$d = 5$	$d = 10$	$d = 50$	$d = 100$
`Fr-bootstrap`	0.041	0.044	0.044	0.054	0.062	0.047	0.051	0.049	0.063	0.067
`Fr-Bayes`	0.055	0.049	**0.079**	0.089	**0.095**	0.045	0.053	0.082	0.093	0.094
`En-bootstrap`	0.047	0.053	0.068	0.067	0.077	0.052	0.049	0.063	0.069	0.075
`En-Bayes`	0.053	0.058	0.077	**0.093**	**0.096**	0.051	0.064	**0.088**	**0.098**	0.093
`Fr-Datadriven`	0.046	0.049	0.051	0.063	0.067	0.041	0.043	0.057	0.059	0.065
`HPV-LeCam`	0.381	0.374	0.383	0.391	0.373	0.376	0.368	0.365	0.342	0.373
`Sp-bootstrap`	0.047	0.054	0.062	0.073	0.079	0.042	0.053	0.063	0.069	0.076
`Sp-Bayes`	**0.066**	**0.069**	0.072	0.083	0.094	**0.072**	**0.078**	0.075	0.088	**0.095**
`Fr-Adaptive`	**0.091**	**0.108**	**0.103**	**0.107**	**0.096**	**0.11**	**0.107**	**0.095**	**0.104**	**0.102**

TABLE 1

Simulated type I error rates under the nominal level 0.1 *for* $y = 0.1$. *The results are based on 2,000 Monte-Carlo simulations with Gaussian random variables. We highlighted the two most accurate methods for each value of* d.

	$N = 200$					$N = 500$				
Method	$d = 2$	$d = 5$	$d = 10$	$d = 50$	$d = 100$	$d = 2$	$d = 5$	$d = 10$	$d = 50$	$d = 100$
`Fr-bootstrap`	0.053	0.045	0.051	0.049	0.047	0.052	0.049	0.047	0.053	0.061
`Fr-Bayes`	0.045	0.039	0.047	0.063	0.071	0.039	0.041	0.052	0.061	0.068
`En-bootstrap`	**0.057**	0.051	0.042	0.043	0.049	0.041	0.039	0.048	0.052	0.059
`En-Bayes`	**0.057**	0.053	0.061	0.067	0.075	0.048	**0.059**	**0.064**	**0.073**	**0.078**
`Fr-Datadriven`	0.026	0.023	0.034	0.037	0.039	0.041	0.04	0.048	0.042	0.047
`HPV-LeCam`	0.87	0.79	0.82	0.81	0.85	0.882	0.835	0.823	0.872	0.823
`Sp-bootstrap`	0.049	0.057	0.056	0.062	0.059	0.043	0.041	0.045	0.053	0.062
`Sp-Bayes`	**0.057**	**0.058**	**0.062**	**0.074**	**0.081**	**0.053**	0.057	0.059	0.059	0.069
`Fr-Adaptive`	**0.103**	**0.092**	**0.105**	**0.107**	**0.099**	**0.11**	**0.107**	**0.104**	**0.097**	**0.103**

TABLE 2

Simulated type I error rates under the nominal level 0.1 *for* $y = 1$.

	$N = 200$					$N = 500$				
Method	$d = 2$	$d = 5$	$d = 10$	$d = 50$	$d = 100$	$d = 2$	$d = 5$	$d = 10$	$d = 50$	$d = 100$
`Fr-bootstrap`	0.028	0.034	0.037	0.041	0.043	0.039	0.045	0.052	0.038	0.047
`Fr-Bayes`	0.038	**0.051**	0.049	**0.062**	**0.071**	**0.051**	**0.049**	0.046	0.053	**0.069**
`En-bootstrap`	0.032	0.041	0.045	0.046	0.059	0.037	0.041	0.054	0.049	0.054
`En-Bayes`	**0.046**	0.048	**0.057**	0.061	0.068	0.039	0.042	0.049	0.052	0.064
`Fr-Datadriven`	0.027	0.033	0.038	0.041	0.043	0.038	0.034	0.029	0.045	0.052
`HPV-LeCam`	0.897	0.939	0.964	0.971	0.972	0.891	0.911	0.943	0.932	0.953
`Sp-bootstrap`	**0.046**	0.048	0.045	0.054	0.052	0.039	0.047	0.049	0.053	0.058
`Sp-Bayes`	0.043	0.049	0.052	0.057	0.068	**0.051**	0.048	**0.059**	**0.063**	0.068
`Fr-Adaptive`	**0.104**	**0.102**	**0.095**	**0.098**	**0.103**	**0.091**	**0.097**	**0.104**	**0.097**	**0.103**

TABLE 3

Simulated type I error rates under the nominal level 0.1 *for* $y = 10$.

powerful one once φ is reasonably large. When φ is small, we find that the $\texttt{HP} - \texttt{LeCam}$ and $\texttt{Sp} - \texttt{bootstrap}$ are slightly more powerful than us. Third, since our proposed statistic is

accurate, we can construct the reject region using our limiting distribution without conducting extensive numerical simulations. The type I error rates for our method are recorded in the last rows of Tables 1, 2 and 3 which are fairly close to 0.1. In Figures B.5, B.6 and B.7 of [12], we report these results where the rejection regions of our methods are constructed using its limiting distribution. We can obtain similar results as in Figures B.2– B.4. We emphasize that since it is impossible to simulate the type I error rates with one matrix in practice, our method will be preferred in real applications. Finally, we investigate the receiver operating characteristic curve (ROC) which is helpful for us to understand the connection between the type I error and power. In practice, researchers will choose the test with the largest area under an ROC curve (AUC) [40]. In Figures B.9 and B.10 of [12], we provide the ROC curves for two different alternatives (i.e., two different values of φ) for $y = 0.1, 1, 10$ when $d = 10$. It can be seen that our test have the largest AUC among all the tests. This also shows that our test outperforms the other tests and will be preferred in decision making.

4. Sketch of proof strategy. In this section, we provide a sketch of the proof strategy, and state all details of the proof in the supplement [12].

The starting point of our proof is to express both $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$ and μ_i in terms of the Green function $\mathcal{G}_1(z) := (XX^* - z)^{-1}$; see Lemmas D.1 and D.5 of our supplement [12]. Both representations can be obtained by applying resolvent expansions to certain functionals of the Green function and its derivative. The error terms in the expansions can be estimated with the aid of the isotropic local laws from [24, 55]. These expressions allow us to work with the Green function instead of the eigenvalue and eigenvector statistics. We remark here that similar derivation of the Green function representation has appeared in previous work such as [53, 21, 54] using a second order resolvent expansion. But here for eigenvectors, we need to do it up to the third order (with a fourth order error), in order to capture all contributing terms for the fluctuation. For instance, when $\boldsymbol{w} \in \mathrm{Span}\{\boldsymbol{v}_j\}_{j \in [\![1,M]\!] \setminus \mathsf{I}}$, the fluctuation of $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$ is one order smaller than that of the case $\boldsymbol{w} \in \mathrm{Span}\{\boldsymbol{v}_t\}_{t \in \mathsf{I}}$. In order to cover the situation like the former case, we will need to investigate a higher order term in the expansion. It turns out that all the leading terms in the expansions in Lemmas D.1 and D.5 are given in terms of certain quadratic forms of the Green function. For example, for the eigenvector, Lemma D.1 suggests that the distribution of $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$ is ultimately governed by the joint distribution of the quadratic forms $\boldsymbol{w}_{\mathsf{I}}^* \Xi \boldsymbol{w}_{\mathsf{I}}, \boldsymbol{\varsigma}_{\mathsf{I}}^* \Xi \boldsymbol{w}_{\mathsf{I}}, \boldsymbol{w}_{\mathsf{I}}^* \Xi' \boldsymbol{w}_{\mathsf{I}}, \{\boldsymbol{v}_t^* \Xi \boldsymbol{\varsigma}_{\mathsf{I}}\}_{t \in \mathsf{I}}, \{\boldsymbol{v}_j^* \Xi \boldsymbol{w}_{\mathsf{I}}\}_{j \in \mathsf{I}^c}$, where $\Xi := \mathcal{G}_1(z) - m_1(z) I$. Here $m_1(z)$ is defined in (C.3) and $\boldsymbol{\varsigma}_{\mathsf{I}}$ is defined in (2.9).

With the expressions in terms of the quadratic forms of Green function, we then apply a recursive moment estimate to derive the distribution. For instance, from Lemma D.1 of [12], one can see that the asymptotic distribution of $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$ can be obtained from that of the random vector

$$\left(\boldsymbol{w}_{\mathsf{I}}^* \Xi'(z) \boldsymbol{w}_{\mathsf{I}}, \boldsymbol{w}_{\mathsf{I}}^* \Xi(z) \boldsymbol{w}_{\mathsf{I}}, \boldsymbol{\varsigma}_{\mathsf{I}}^* \Xi(z) \boldsymbol{w}_{\mathsf{I}}, \{\boldsymbol{v}_t^* \Xi(z) \boldsymbol{\varsigma}_{\mathsf{I}}\}_{t \in \mathsf{I}}, \{\boldsymbol{v}_j^* \Xi(z) \boldsymbol{w}_{\mathsf{I}}\}_{j \in \mathsf{I}^c} \right)^*,$$

whose components are all quadratic forms of Green function. In order to derive a multivariate CLT for the above random vector, it suffices to show the CLT for any linear combination of its components. Let $\mathcal{P}$ be a linear combination of the components of the above vectors with any appropriately scaled deterministic coefficients. Our aim is to show that the following recursive moment estimate is satisfied

$$(i) : \mathbb{E}\mathcal{P} = o(1), \qquad (ii) : \mathbb{E}\mathcal{P}^l = (l-1) V \mathbb{E}\mathcal{P}^{l-2} + o(1),$$

where V is a deterministic constant and the above holds for any given integer $l \geq 2$. We also refer to Proposition E.2 of [12] for a more precise statement. The above recursive moment estimate then leads to the Gaussianity of $\mathcal{P}$, which further implies the CLT of $\langle \boldsymbol{w}, \mathrm{P}_{\mathsf{I}} \boldsymbol{w} \rangle$. The proof of the above recursive moment estimate heavily relies on the cumulant expansion

formula and the local laws stated in Section C of the supplement [12]. The derivation for the joint eigenvalue-eigenvector distribution is similar. We emphasize that the Green function has also been used in [27] and [38] for the eigenvector distribution for the deformed Wigner matrices, but in much more limited ways. Especially, in [27], only the projections to very special directions are considered, and in [38] the strength of the spikes are assumed to be divergent in a sufficiently fast speed so that the Green function can be expanded directly around a large parameter z and eventually one only needs to deal with the quadratic forms of the Wigner matrix itself.

Finally, we highlight some difficulties and novelties in the above strategy. First, since the d_i's could be either very close to the critical threshold or diverging, and meanwhile could be equal or close to each other, the control of the sizes of the terms (especially the error terms) becomes much more delicate. One needs to keep tracking the dependence of the size of terms on $d_i - \sqrt{y}$, $d_i - d_j$ and d_i carefully to conduct a unified analysis in all cases of d_i's. In particular, in case that d_i is diverging, one needs to exploit a hidden cancellation between two quadratic forms of the Green function, which is absent in case that d_i is fixed. In order to see such a cancellation, one needs to adopt the recently established nearly optimal convergence rate of the so-called *eigenvector empirical spectral distribution (VESD)* in [86]. Second, in contrast to [11], where only the projection onto the direction of the deformation is considered, here we consider the projection onto arbitrary direction. In particular, when one considers the projection onto the orthogonal complement of the direction of the deformation, the size of the whole projection will degenerate to a smaller order. As we mentioned above, in order to study the fluctuation of the eigenvector projection onto arbitrary directions, including the direction orthogonal to the one of the spike, one needs to express the eigenvector projection in terms of the Green function up to a higher order projection onto, and involve the higher order term in the recursive moment estimate, since it could be significant. Third, the joint distribution of the eigenvalue and eigenvector statistics is obtained for the first time in the supercritical regime for the whole range of d_i, thanks to our unified method of proving CLT for both eigenvalues and eigenvectors.

Acknowledgements. The second author would like to thank Igor Silin for sharing the Python codes of [82] and providing some insights on the statistical applications. We would also like thank Alexander Aue, Jiang Hu, Zeng Li, Debashis Paul, Dong Xia, Yanrong Yang, Jeff Yao and Lin Zhang for many helpful discussions.

The first author was partially supported by Hong Kong RGC grant GRF 16301519 and NSFC 11871425. The second author is partially supported by NSF-DMS 2113489 and grateful for the AMS-Simons Travel Grants (2020-2022). The third author was partially supported by Hong Kong RGC grant ECS 26301517 and GRF 16300618. The fourth author was partially supported by Hong Kong RGC grant GRF 16301618, GRF 16308219 and ECS 26304920.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical inference for principal components of spiked covariance matrices". In [12], we provide a supplementary file which contains additional simulation results, the proofs of our main results and some auxiliary lemmas.
().

REFERENCES

[1] T. ANDERSON. (2003). An Introduction to Multivariate Statistical Analysis, 3rd edition. Wiley Series in Probability and Statistics, Wiley. 2003.

[2] M. AVELLANEDA, AND B. HEALY, AND A. PAPANICOLAOU, AND G. PAPANICOLAOU, AND T. XU. (2020). Principal Eigenportfolios for U.S. Equities. Available at SSRN.
[3] J. BAI, AND S. NG. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, **176(1)**:18–29.
[4] J. BAI, AND S. NG. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics*, **212(1)**:78–96.
[5] J. BAI, AND S. NG. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, **70(1)**:191–221.
[6] J. BAI, AND S. NG. (2006). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics*, **131(1)**:507–537.
[7] Z.D. BAI, AND J.F. YAO. (2008). Central limit theorems for eigenvalues in a spiked population model. *Annales de l'IHP Probabilités et statistiques*, **44(3)**: 447–474.
[8] Z.D. BAI, AND J.F. YAO. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, **106**:167–177.
[9] J. BAIK, G. BEN AROUS, AND S. PÉCHÉ. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, **33(5)**: 1643–1697.
[10] J. BAIK, AND J.W. SILVERSTEIN. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, **97(6)**: 1382–1408.
[11] Z.G. BAO, X.C. DING, AND K. WANG. (2021). Singular vector and singular subspace distribution for the matrix denoising model. *The Annals of Statistics*,**49(1)**: 370–392.
[12] Z.G. BAO, X.C. DING, J.M. WANG AND K. WANG. (2020). Supplement to "Statistical inference for principal components of spiked covaraince matrices".
[13] Z.G. BAO, D. WANG. (2021). Eigenvector distribution in the critical regime of BBP transition. *Probability Theory and Related Fields* (to appear)
[14] S. BELINSCHI, H. BERCOVICI, AND M. CAPITAINE. (2017). On the outlying eigenvalues of a polynomial in large independent random matrices. *arXiv:1703.08102*.
[15] S. BELINSCHI, H. BERCOVICI, M. CAPITAINE, AND M. FÉVRIER. (2017). Outliers in the spectrum of large deformed unitarily invariant models. *The Annals of Probability*, **45(6A)**:3571–3625.
[16] F. BENAYCH-GEORGES, A. GUIONNET, AND M. MAIDA. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electronic Journal of Probability*, **16**: 1621–1662.
[17] F. BENAYCH-GEORGES, AND R. R. NADAKUDITI. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, **111**: 120–135.
[18] F. BENAYCH-GEORGES, AND R. R. NADAKUDITI. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, **227(1)**:494–521.
[19] K. BENIDIS, Y. SUN, P. BABU, AND D. PALOMAR. (2016). Orthogonal Sparse PCA and Covariance Estimation via Procrustes Reformulation. *IEEE Transactions on Signal Processing*, **64(23)**:6211-6226.
[20] A. BIRNBAUM, I. JOHNSTONE, B. NADLER, AND D. PAUL. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Stat.*, **41(3)**:1055–1084.
[21] A. BLOEMENDAL, A. KNOWLES, H.-T. YAU, AND J. YIN. (2016). On the principal components of sample covariance matrices. *Probability theory and related fields*, **164(1-2)**:459–552.
[22] A. BLOEMENDAL, AND B. VIRÁG. (2013). Limits of spiked random matrices I. *Probability Theory and Related Fields*, **156(3-4)**: 795–825.
[23] A. BLOEMENDAL, AND B. VIRÁG. (2016). Limits of spiked random matrices II. *The Annals of Probability*, **44(4)**: 2726–2769.
[24] A. BLOEMENDAL, L. ERDOS, A. KNOWLES, H.-T. YAU, AND J. YIN. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab*, **19(33)**:1–53.
[25] J. BUN, J. BOUCHAUD, AND M. POTTERS. (2017). Cleaning large correlation matrices: Tools from Random Matrix Theory. *Physics Reports*, **666**:1-109.
[26] M. CAPITAINE. (2018). Limiting eigenvectors of outliers for Spiked Information-Plus-Noise type matrices. *Séminaire de Probabilités XLIX* 119–164, Springer, Cham.
[27] M. CAPITAINE, AND C. DONATI-MARTIN. (2018). Non universality of fluctuations of outlier eigenvectors for block diagonal deformations of Wigner matrices. *arXiv:1807.07773*.
[28] M. CAPITAINE, C. DONATI-MARTIN, AND D. FÉRAL. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, **37(1)**:1–47.
[29] M. CAPITAINE, C. DONATI-MARTIN, AND D. FÉRAL. (2012). Central limit theorems for eigenvalues of deformations of Wigner matrices. *Annales de l'IHP Probabilités et statistiques* **48(1)**:107–133.
[30] X.C. DING. (2020). High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli*, **26(1)**:387–417.

[31] X.C. DING, AND F. YANG. (2021). Spiked separable covariance matrices and principal components. *The Annals of Statistics*, **49(2)**: 1113–1138.

[32] X.C. DING, AND H.C. JI. (2020). Local laws for multiplication of random matrices and spiked invariant model. *arXiv:2010.16083*.

[33] E. DOBRIBAN. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *The Annals of Statistics*, **45(4)**:1810–1833.

[34] E. DOBRIBAN AND A. OWEN. (2019). Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B*, **81(1)**:163-183.

[35] E. DOBRIBAN. (2020). Permutation methods for factor analysis and PCA. *The Annals of Statistics*(to appear).

[36] D. DONOHO, M. GAVISH, AND I. JOHNSTONE. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model.*The Annals of Statistics*, **46(4)**: 1742–1778.

[37] L. ERDŐS, A. KNOWLES, AND H.-T. YAU. (2013). Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, **14(8)**:1837–1926.

[38] J. FAN, Y. FAN, X. HAN, AND J. LV. (2019). Asymptotic theory of eigenvectors for large random matrices. *arXiv:1902.06846*.

[39] J. FAN, Y. LIAO, AND M. MINCHEVA. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Stat.*, **39(6)**:3320–3356.

[40] T. FAWCETT.(2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27(8)**: 861–874.

[41] D. FÉRAL, S. PÉCHÉ. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Communications in mathematical physics*, **272(1)**:185–228.

[42] M. HALLIN, D. PAINDAVEINE AND T. VERDEBOUT. (2010). Optimal rank-based testing for principal components. *The Annals of Statistics*, **38(6)**:3245–3299.

[43] Y. HE, A. KNOWLES. (2017). Mesoscopic eigenvalue statistics of Wigner matrices. *The Annals of Applied Probability*, **27(3)**: 1510-1550.

[44] Y. HE, A. KNOWLES. (2019). Mesoscopic eigenvalue density correlations of Wigner matrices. *Probability Theory and Related Fields*, 1-70.

[45] J. HWANG, J. LEE, AND K. SCHNELLI. (2019). Local law and Tracy-Widom limit for sparse sample covariance matrices. *Ann. Appl. Probab.*, **29(5)**:3006–3036.

[46] G. JAMES, D. WITTEN, T. HASTIE, AND R. TIBSHIRANI. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics , Springer.

[47] I. JOHNSTONE. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, **29(2)**: 295–327.

[48] I. JOHNSTONE, AND A. LU. (2009). On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of American Statistical Association*, **104(486)**:682–693.

[49] I. JOHNSTONE, AND J. YANG. (2018). Notes on asymptotics of sample eigenstructure for spiked covariance models with non-Gaussian data. ARXIV:1810.10427.

[50] I. T. JOLLIFFE. (2002). Principal Component Analysis. Springer Series in Statistics, Springer-Verlag, New York.

[51] Z. KE, Y. MA, AND X. LIN. (2020). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *arXiv preprint arXiv: 2006.00436*.

[52] A.M. KHORUNZHY, B.A. KHORUZHENKO, AND L.A. PASTUR. (1996). Asymptotic properties of large random matrices with independent entries. *Journal of Mathematical Physics*, **37(10)**:5033–5060.

[53] A. KNOWLES, AND J. YIN. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, **66(11)**:1663–1749.

[54] A. KNOWLES, AND J. YIN. (2014). The outliers of a deformed Wigner matrix. *The Annals of Probability*, **42(5)**: 1980–2031.

[55] A. KNOWLES, AND J. YIN. (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, **169(1-2)**:257–352.

[56] V. KOLTCHINSKII AND K. LOUNICI. (2017). New asymptotic results in Principal Component Analysis. *Sankhya A.*, **79(2)**:254-297.

[57] Z.V. LAMBERT, A.R. WILDT, AND R. M. DURAND. (1991). Approximating Confidence Intervals for Factor Loadings. *Multivariate Behavioral Research*, **26(3)**:421-434.

[58] J. O. LEE, AND K. SCHNELLI. (2018). Local law and Tracy-Widom limit for sparse random matrices. *Probability Theory and Related Fields*, **171(1-2)**:543-616.

[59] Z. Li, F. Han, and J. Yao. (2020). Asymptotic joint distribution of extreme eigenvalues and trace of large sample covariance matrix in a generalized spiked population model. *The Annals of Statistics* (in press).

[60] P. LOUBATON, AND P. VALLET. (2011). Almost sure localization of the eigenvalues in a Gaussian information plus noise model. Application to the spiked models. *Electronic Journal of Probability*, **16**, 1934–1959.

[61] V.A. MARČENKO, AND L.A. PASTUR. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, **1(4)**:457.

[62] Z. MA. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Stat.*, **41(2)**:772–801.

[63] R. MONASSON, AND D. VILLAMAINA. (2015). Estimating the principal components of correlation matrices from all their empirical eigenvectors. *Europhysics Letters*, **112(5)**: 50001.

[64] D. MORALES-JIMENEZ, I. M. JOHNSTONE, M. R. MCKAY, AND J. YANG. (2018). Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models. *arXiv:1810.10214*.

[65] R. R. NADAKUDITI, AND A., EDELMAN. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Transactions on Signal Processing*, **56(7)**:2625-2638.

[66] R. R. NADAKUDITI, AND J. W. SILVERSTEIN. (2010). Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE Journal of Selected Topics in Signal Processing*, 4(3): 468-480.

[67] B. NADLER. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, **36(6)**:2791-2817.

[68] B. NADLER AND I. M. JOHNSTONE. (2011). Detection performance of Roy's largest root test when the noise covariance matrix is arbitrary. *2011 IEEE Statistical Signal Processing Workshop*, 681–684.

[69] A. NAUMOV, V. SPOKOINY, AND V. ULYANOV. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields*, 1091–1132.

[70] A. ONATSKI. (2009). Testing Hypotheses About the Number of Factors in Large Factor Models. *Econometrica*, **77(5)**:1447–1479.

[71] A. ONATSKI, M. MOREIRA, AND M. HALLIN. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.*, **41(3)**: 1204-1231.

[72] A. ONATSKI, M. MOREIRA, AND M. HALLIN. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.*, **42(1)**: 225-254.

[73] A. ONATSKI. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, **168(2)**:244-258.

[74] F.J. OORT. (2011). Likelihood-Based Confidence Intervals in Exploratory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, **18(3)**:383-396.

[75] D. PASSEMIER, AND J.F. YAO. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, **127**: 173-183.

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

[77] S. PÉCHÉ. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probability Theory and Related Fields*, **134(1)**:127–173.

[78] A. PERRY, A.S. WEIN, A.S. BANDEIRA, A. MOITRA. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *The Annals of Statistics*, **46(5)**:2416–2451.

[79] N. S. PILLAI, J. YIN. (2014). Universality of covariance matrices. *The Annals of Applied Probability*, **24(3)**:935–1001.

[80] A. PIZZO, D. RENFREW, AND A. SOSHNIKOV. (2013). On finite rank deformations of Wigner matrices. *Annales de l'IHP Probabilités et statistiques*, **49(1)**:64–94.

[81] H. SHEN, AND J. HUANG. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99(6)**:1015–1034.

[82] I. SILIN, AND J. FAN. (2020). Hypothesis testing for eigenspaces of covariance matrix. *arXiv: 2020.09810*.

[83] I. SILIN, AND V. SPOKOINY. (2018). Bayesian inference for spectral projectors of the covariance matrix. *Electronic Journal of Statistics*, **12**:1948-1987.

[84] D.E.TYLER. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, **9**:725–736.

[85] D.E.TYLER. (1983). A class of asymptotic tests for principal component vectors. *The Annals of Statistics*, **11**: 1243–1250.

[86] H.K. XI, F. YANG, J. YIN. (2020). Convergence of eigenvector empirical spectral distribution of sample covariance matrices. *The Annals of Statistics*, **48(2)**:953–982.

[87] H. YAMAMOTO, T. FUJIMORI, H. SATO, G. ISHIKAWA, K. KAMI, AND Y. OHASHI. (2014). Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinformatics*, **15**:51.

[88] F. YANG (2020). Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices. *arXiv: 2005.00999*.

[89] H. ZOU, T. HASTIE, AND R. TIBSHIRANI. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15(2)**: 265–286.