# DETECTING RARE AND WEAK SPIKES IN LARGE COVARIANCE MATRICES

By Zheng Tracy Ke[*]

*Harvard University*[*]

Given $p$-dimensional Gaussian vectors $X_i \overset{iid}{\sim} N(0, \Sigma)$, $1 \le i \le n$, where $p \ge n$, we are interested in testing a null hypothesis where $\Sigma = I_p$ against an alternative hypothesis where all eigenvalues of $\Sigma$ are 1, except for $r$ of them are larger than 1 (i.e., spiked eigenvalues).

We consider a Rare/Weak setting where the spikes are sparse (i.e., $1 \ll r \ll p$) and individually weak (i.e., each spiked eigenvalue is only slightly larger than 1), and discover a phase transition: the two-dimensional phase space that calibrates the spike sparsity and strengths partitions into the *Region of Impossibility* and the *Region of Possibility*. In Region of Impossibility, all tests are (asymptotically) powerless in separating the alternative from the null. In Region of Possibility, there are tests that have (asymptotically) full power.

We consider a CuSum test, a trace-based test, an eigenvalue-based Higher Criticism test, and a Tracy-Widom test [32], and show that the first two tests have asymptotically full power in Region of Possibility.

To use our results from a different angle, we derive new bounds for (a) empirical eigenvalues, and (b) cumulative sums of the empirical eigenvalues, both under the alternative hypothesis. Part (a) is related to those in [4, 40], but both the settings and results are different.

The study requires careful analysis of the $L^1$-distance of our testing problem and delicate Radom Matrix Theory. Our technical devises include (a) a Gaussian proxy model, (b) Le Cam's comparison of experiments, and (c) large deviation bounds on empirical eigenvalues.

**1. Introduction.** Suppose we have $n$ normal vectors $X_i \in \mathbb{R}^p$:

$$(1.1) \qquad X_i \overset{iid}{\sim} N(0, \Sigma), \qquad \Sigma \in \mathbb{R}^{p,p}, \qquad 1 \le i \le n.$$

We are interested in testing the null hypothesis that $\Sigma$ is the $p$ by $p$ identity matrix $I_p$ against an alternative hypothesis that $\Sigma$ is a so-called spike matrix [32]: for an integer $r$ such that $1 \le r \ll p$, all eigenvalues of $\Sigma$ are 1, except for the top $r$ of them are larger than 1 (i.e., each of them is spiked):

$$(1.2) \qquad \ell_1 \ge \ell_2 \ge \ldots \ge \ell_r > 1.$$

Motivated by the recent interest of "$p > n$", we assume $p \geq n$, but this should not be taken as a constraint. Let $X \in \mathbb{R}^{n,p}$ be the data matrix so that $X' = [X_1, X_2, \ldots, X_n]$, and denote the empirical covariance matrix by

$$\hat{\Sigma} = (1/n)X'X.$$

Since $p \geq n$, with probability 1 [48], $\hat{\Sigma}$ has $n$ distinct positive eigenvalues

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0.$$

This testing problem is of interest in many application areas.

- *Covert communication.* In computer security and privacy, covert channels are widely used. Consider a channel with $r$ antennas at transmitter and $p$ antennas at receiver. The output vectors satisfy $X_i = Hy_i + Z_i$, $1 \leq i \leq n$, where $y_i \overset{iid}{\sim} N(0, \delta \cdot I_r)$ are the input vectors, $Z_i \overset{iid}{\sim} N(0, I_p)$ is noise, and $H \in \mathbb{R}^{p,r}$ is a confidential "channel matrix" [47]. When using a covert channel, we would like to know whether our "enemy" can notice that we are sending signals. Note that $X_i$'s follow Model (1.1) with $\Sigma = I_p + \delta HH'$. From the perspective of our enemy, $H$ is unavailable, so it becomes a problem of detecting spiked eigenvalues in $\Sigma$. We will revisit this application in Section 1.6.

- *Inference of genetic population structure.* One of main challenges in analyzing genetic data is to explore whether the samples contain sub-populations that are genetically distinct [43, 39]. Let $Y \in \{0, 1, 2\}^{n,p}$ be the matrix of SNP counts for $p$ markers and $n$ subjects. When there are no sub-populations, $\{Y(i, j)\}_{i=1}^n$ are modeled as *iid* samples from $Binomial(2, p_j)$, where $p_j \in (0, 1)$ is the Minor Allele Frequency of marker $j$. Consider the normalized data matrix $X \in \mathbb{R}^{n,p}$, where $X(i, j) = [Y(i, j) - 2\hat{p}_j]/\sqrt{2\hat{p}_j(1 - \hat{p}_j)}$ and $\hat{p}_j = (2n)^{-1}\sum_{i=1}^n Y(i, j)$. Then, entries of $X$ are independent, with (approximately) zero mean and unit variance. When there are subpopulations, rows of $X$ are confounded by latent ancestry variables, and its covariance matrix has spiked eigenvalues [39]. Inference of genetic sub-populations reduces to detecting spikes in the covariance matrix.

- *Multiple testing.* How to account for data dependence is a central challenge in large-scale multiple testing [23, 21]. Recent works [35, 27, 26] used Model (1.1)-(1.2) to model test statistics, where each spiked eigenvalue comes from an unobserved confounding variable. Under this framework, various factor-adjusted multiple testing procedures have been proposed. Then, a fundamental problem is to detect the existence of confounding factors, so that we know when to use factor-adjusted methods instead of classical ones.

- *Network community detection.* Given a large social network, we are often interested in testing whether it contains only one community or multiple communities. Let $A \in \{0,1\}^{n,n}$ be the adjacency matrix of a symmetrical network with $n$ nodes. Consider the centered and rescaled adjacency matrix $\widetilde{A} \in \mathbb{R}^{n,n}$ by $\widetilde{A}(i,j) = [A(i,j) - \hat{p}]/\sqrt{\hat{p}(1-\hat{p})}$, where $\hat{p} = \sum_{i<j} A(i,j)/[n(n-1)/2]$. Under the stochastic block model, it is well-known that $\widetilde{A}$ is approximately a Wigner matrix when there is only one community and has a few spiked eigenvalues when there are multiple communities [10, 36].

Other applications include low-rank matrix recovery [13], PCA and sparse PCA [52, 40], and high-dimensional clustering [30, 31].

In Model (1.1)-(1.2), we assume all eigenvalues of $\Sigma$ are 1 under the null. This is a mild assumption since the data matrix is always pre-normalized in above applications. The setting that $\hat{\Sigma}$ is a Gaussian covariance matrix is not exactly the same as those in genetics (where data are sub-Gaussian) and in social networks (where we deal with Wigner matrices). But due to eigenvalue universality, the asymptotic behavior of many eigenvalue-based tests remains the same. Our testing framework can be viewed as an idealization of these applications, but it still captures the essential features.

Recently, encouraging progresses have been made to understanding this problem, such as Johnstone [32], Onatski et al. [37, 38] and Johnstone and Nadler [33]. However, these works have been largely focused on the case where the number of spikes $r$ is small. In fact, in the asymptotic framework used in these papers, $r$ is fixed as $n \to \infty$.

In this paper, the primary interest is to consider the testing problem in the case where we have *many weak spikes*. For covert communication, this means the transmitter has a number of antennas, each with a relatively small capacity. For genetical data, it is the case where the subjects come from a number of sub-populations whose mutual genetic distinction is so weak that to separate them is subtle or even impossible; this can happen for studies conducted on European populations (say). Similar scenarios also happen in social networks: a large network often contains many "weak communities" [44] that are hard to identify. Motivated by these applications, we adopt a Rare and Weak Spike (RWS) model where as $n \to \infty$,

$r \to \infty$ in an algebraic rate, each spiked eigenvalue is *slightly* larger than 1.

The main contribution of the paper is three-fold.

- (*Phase transition*). We discover that the two-dimensional phase space calibrating the spike sparsity and spike strengths partitions into the

Region of Impossibility and the Region of Possibility. In the former, the spikes are so rare/weak that it is impossible to separate the alternative from the null. In the latter, the spikes are sufficiently strong and it is possible to separate the alternative from the null.

- (*Methods*). We propose a CuSum test and an (eigenvalue-based) Higher Criticism (HC) test[1] as new approaches to the testing problem. In particular, we show that the CuSum test is optimal.
- (*Bounds on eigenvalues*). Knowledge/innovation flow is rarely a one-way street: while Random Matrix Theory (RMT) helps us establish the lower/upper bounds of our testing problem, the latter also shed lights on some problems in RMT in return: we obtain new bounds on $\lambda_i$ and on the cumulative sums of $\lambda_i$, a topic of great interest in RMT.

Our study requires delicate analysis, and the following are some noteworthy points of our technical contributions.

- The RWS model is similar to that in [33, 38] and is known to be hard to analyze. We overcome the technical hurdle by using a Gaussian proxy model, which is comparably easier to analyze. Our result on the intimacy of two models is of interest for its own sake, and can be used as a technical device to study problems other than spike detection.
- Le Cam's comparison of experiments provides a useful tool for analyzing complicated models. Using Le Cam's idea, we extend our main results on the RWS model to more complicated settings.
- Our analysis needs many recent results in RMT. These results scatter across the literature and are presented in forms that are not always easy to access. With substantial efforts, we adapt such results to our testing problem and make them more accessible for us (see Section 3).

1.1. *Four test statistics.* We propose CuSum (CS) and (eigenvalue-based) Higher Criticism (HC) as two new tests. We also investigate a trace-based test and a Tracy-Widom (TW) test.

The trace-based test uses the empirical moments of $\lambda_i$ for testing. While there are many tests of this kind, we focus our study on the following test:

$$S_n = \sum_{i=1}^{n} \lambda_i = \text{tr}(\hat{\Sigma}).[2]$$

In Section 1.3 we discuss other versions of trace-based statistics.

---

[1]The HC test is based on $\lambda_i$ and is very different from those in literature.

[2]For any square matrix $A$, $\text{tr}(A)$ stands for the trace.

The trace statistic uses the total sum of all eigenvalues. A natural alternative is to use the cumulative sum (CuSum)

$$S_k = \sum_{i=1}^{k} \lambda_i, \qquad 1 \le k \le n,$$

where we hope for some $k < n$, $S_k$ contains stronger evidence against the null than $S_n$. Of course, we don't know how to pick the "best" $k$. To address this problem, we propose the following test statistic which we call the *CuSum*:

$$CS_n^* = \max_{1 \le k \le n} \{CS_{n,k}\}, \qquad \text{where } CS_{n,k} = \frac{S_k - E_0[S_k]}{SD_0(S_k)}.$$

Here, $E_0[S_k]$ and $SD_0[S_k]$ are the mean and standard deviation of $S_k$ under the null, respectively. Such quantities do not have a closed-form expression, but can be conveniently simulated. In some cases, especially when the spikes are very rare, CuSum improves the trace-based test; see Section 2.

In a similar spirit, we have the eigenvalue-based *Higher Criticism (HC)* test.[3] The HC statistic is defined by

$$(1.3) \qquad HC_n^* = \max_{1 \le k \le n} \{HC_{n,k}\}, \qquad \text{where } HC_{n,k} = \frac{\lambda_k - E_0[\lambda_k]}{SD_0(\lambda_k)}.$$

The current HC is very different from existing versions of HC (e.g., [19, 20]).

The HC test is an extension of the well-known Tracy-Widom (TW) test:

$$(1.4) \qquad TW_n \equiv HC_{n,1} = \frac{\lambda_1 - E_0[\lambda_1]}{SD_0(\lambda_1)}.$$

The TW test is motivated by recent works on the Tracy-Widom law [46] and has been studied in several recent papers [32, 33, 37, 38]. HC has advantages over TW: it is known that the empirical eigenvalues are much noisier at the edge than in the bulk, so it is possible that $HC_{n,1}$ is less powerful than $HC_{n,k}$ for some $k > 1$. Numerical results confirm this point; see Section 2.

We compare all four tests in Table 1. Note that these tests are not tied to the Gaussian model (1.1) and can be implemented in much broader settings.

The CS and HC tests are convenient to use in practice, but are *difficult* to analyze *theoretically*: it is hard to pin down the magnitudes of $SD_0(S_k)$ and $SD_0(\lambda_k)$ analytically. Fortunately, we have the following upper bounds [6, 42]. With probability at least $1 - o(n^{-1})$, for all $1 \le k \le n$,

$$(1.5) \ \ SD_0(S_k) \le (L_n)^c (k/n)^{2/3}, \quad SD_0(\lambda_k) \le (L_n)^c n^{-\frac{2}{3}} [k \wedge (n+1-k)]^{-\frac{1}{3}},$$

---

[3]The test was briefly mentioned in a survey paper by Donoho and Jin [21], but has not yet been formally studied.

| Trace $(S_n)$ | CuSum $(CS_n^*)$ | HC $(HC_n^*)$ | TW $(TW_n)$ |
|---|---|---|---|
| $\sum_{i=1}^n \lambda_i$ | $\max_{1 \le k \le n}\left\{\frac{S_k - E_0[S_k]}{SD_0(S_k)}\right\}$ | $\max_{1 \le k \le n}\left\{\frac{\lambda_k - E_0[\lambda_k]}{SD_0(\lambda_k)}\right\}$ | $\frac{\lambda_1 - E_0[\lambda_1]}{SD_0(\lambda_1)}$ |

where $L_n \equiv [\log(n)]^{\log(\log(n))}$ and $c > 0$ is a universal constant. These motivate the following *variants* of $CS_n^*$ and $HC_n^*$:

(1.6)
$$CS_n^+ = \max_{1 \le k \le n}\left\{\frac{S_k - E_0[S_k]}{(k/n)^{2/3}}\right\}, \quad HC_n^+ = \max_{1 \le k \le n}\left\{\frac{\lambda_k - E_0[\lambda_k]}{n^{-2/3}[k \wedge (n+1-k)]^{-1/3}}\right\}.$$

Due to the explicit forms in the denominators, these variants are more convenient for theoretical analysis than the original version of the statistics.

1.2. *Rare and Weak Spike (RWS) model.* Let $X_i \overset{iid}{\sim} N(0, \Sigma)$ be as in (1.1). Our interest is to test whether the null

(1.7) $$H_0^{(n)} : \qquad \Sigma = I_p$$

holds or not. Fixing an integer $1 \le r \le p$ and a parameter $\delta > 0$ and letting $\mathbb{S}(p, r)$ be the Stiefel manifold [15] (consisting all matrices $Q \in \mathbb{R}^{p,r}$ such that $Q'Q = I_r$), we consider a specific alternative hypothesis:

(1.8) $$H_1^{(n)} : \quad \Sigma = (I_p - \delta QQ')^{-1} = I_p + \frac{\delta}{1 - \delta} \cdot QQ',$$

where $Q$ is uniformly generated from the Stiefel manifold $\mathbb{S}(p, r)$.

We use $n$ as the driving asymptotic parameter, and tie $(p, r, \delta)$ to $n$ by fixed parameters. In detail, fixing $\alpha, \beta \in (0, 1)$ and $\gamma \ge 1$, we assume

(1.9) $$p_n/n \to \gamma, \qquad r = r_n = n^{1-\beta}, \qquad \delta = \delta_n = n^{-\alpha}.$$

**Definition 1.1** *We call (1.8)-(1.9) the Rare and Weak Spike (RWS) model.*

Similar Rare/Weak models have been used in many recent works but for different problems [21], and the Rare/Weak Spike model here is new.

**Remark**. (*A Gaussian proxy model to RWS*). The RWS model is hard to analyze: in the likelihood ratio associated with the testing problem, it is hard to integrate $Q$ out with the law of $Q \sim \mathbb{S}(p, r)$. In the case where $r$ is fixed while $n \to \infty$, [37, 38] attacked the problem with the so-called Laplace's method and careful large deviation analysis of the spherical integrals, but

how to extend their techniques and results to our case (where $r$ grows to $\infty$ at an algebraic rate as $n \to \infty$) remains unclear.

We propose a new technical device by introducing a proxy model that is very close to RWS but is easier to analyze. In detail, we consider a proxy testing problem where we replace the alternative hypothesis $H_1^{(n)}$ by

$$\widetilde{H}_1^{(n)}: \qquad \widetilde{\Sigma} = [I_p - (\delta/p)YY']^{-1}, \quad Y = Z \cdot 1\{\|Z\| \le \frac{1}{2}\sqrt{p/\delta}\},$$

where $Z \in \mathbb{R}^{p,r}$ is the matrix that has *iid* $N(0,1)$ entries. Using Le Cam's "comparison of experiments", we are able to prove that two models are close to each other, for a wide region in the parameter space (see Figure 1 and Lemma 4.1). Note that Gaussian proxy model is comparably easier to analyze than the RWS: the analysis of the former relies on the properties of Wishart matrices where many results exist, while that of the latter relies on properties of spherical integrals, a topic that is comparably less studied.

1.3. *Main results.* Our result has two parts: phase transition for the testing problem, and bounds for empirical eigenvalues under $H_1^{(n)}$.

In our results on phase transition, especially Theorem 1.1, recent development in Random Matrix Theory (RMT) has played an important role. However, *knowledge flow is not a one-way street*: our understanding of the testing problem in turn sheds lights on some of the problems in RMT.

In detail, under $H_0^{(n)}$, recent works in RMT have shed interesting lights on the bounds of the empirical eigenvalues (e.g., [25]), but under $H_1^{(n)}$, such bounds are much less studied and remain largely unknown. Interestingly, Theorem 1.1 provides an approach to studying the bounds under $H_1^{(n)}$.

The idea is that, for parameters in the *Region of Impossibility* (see below), any test is asymptotically powerless in distinguishing $H_1^{(n)}$ from $H_0^{(n)}$. We can therefore use existing bounds on the empirical eigenvalues under $H_0^{(n)}$ to derive similar bounds under $H_1^{(n)}$; the resultant bounds are non-trivial to derive using RMT, at least for some of the parameter ranges.

Now, first, consider the trace-based test statistic $S_n$. Under $H_0^{(n)}$, $nS_n \sim \chi^2_{np}(0)$, and so $(2np)^{-1/2}(nS_n - np) \approx N(0,1)$ for large $(n,p)$. Fixing a parameter $q > 0$, suppose we reject $H_0^{(n)}$ if and only if

$$(1.10) \qquad (2np)^{-1/2}(nS_n - np) \ge \sqrt{2q\log(n)}, \qquad q > 0.$$

The following theorem establishes the phase transition associated with the testing problem (1.7)-(1.8) and the optimality of the trace-based test $S_n$.
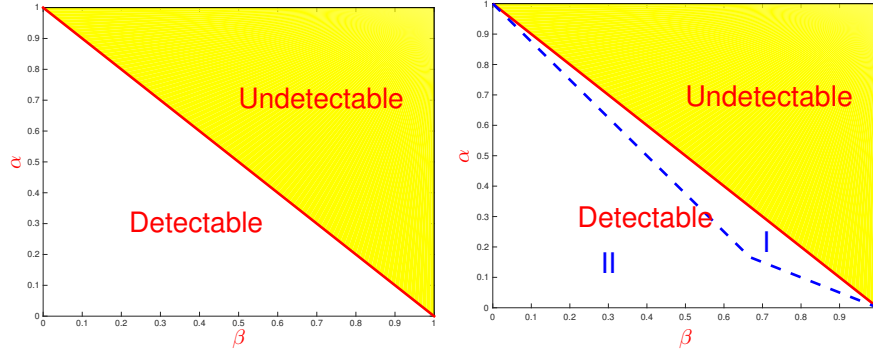
FIG 1. *Left: the curve $\alpha = (1-\beta)$ partitions the phase space into the "detectable region" and the "undetectable region". Right: the curve $\alpha = \max\{1-5\beta/4, (1-\beta)/2\}$ (see Lemma 4.1) partitions the "detectable region" into part I and II. In both the undetectable region and part I of the detectable region, the RWS model and the Gaussian proxy model have negligible differences.*

**Theorem 1.1** *(Phase transition and optimality of the trace-based test). Fix $\alpha, \beta \in (0,1)$, $\gamma \geq 1$, and $q > 0$. Suppose $X_i \overset{iid}{\sim} N(0, \Sigma)$ as in (1.1). Consider the testing problem (1.7)-(1.8) where $\Sigma$ satisfies the RWS model (1.8)-(1.9).*

- *If $\alpha + \beta > 1$, then as $n \to \infty$, for any test, the sum of Type I and Type II testing errors tends to 1.*
- *If $\alpha + \beta < 1$, then as $n \to \infty$, the level of the trace-based test in (1.10) tends to 0 while the power tends to 1.*

See Figure 1. The proof of the second bullet point is not hard so is omitted. The first bullet point is proved in Section 4, where the key ingredients include (a) justifying that the RWS is close to the Gaussian proxy model, and (b) analyzing the proxy model with delicate Random Matrix Theory. Note that the case $\alpha + \beta = 1$ is more delicate and is addressed in Section 1.5.

Second, we consider the CS test statistic, focusing on $CS_n^+$ (the variant of $CS_n^*$). By definition of $CS_n^+$ (see (1.5)-(1.6)), we reject $H_0^{(n)}$ if and only if

$$(1.11) \qquad CS_n^+ \geq \tilde{L}_n, \qquad \text{where } \tilde{L}_n = [\log(n)]^{\log^2(\log(n))}.$$

Recall that $S_k = \lambda_1 + \ldots + \lambda_k$ is the $k$-th cumulative sum of the empirical eigenvalues. The following theorem shows that $CS_n^+$ is an optimal test and provides some non-trivial bounds on $S_k$ under the alternative.

**Theorem 1.2** *(Optimality of the CuSum test and bounds for $S_k$). Fix $\alpha, \beta \in (0,1)$ and $\gamma \geq 1$. Suppose $X_i \overset{iid}{\sim} N(0, \Sigma)$ as in (1.1) and let $\tilde{L}_n = [\log(n)]^{\log^2(\log(n))}$.*

*Consider the testing problem* (1.7)-(1.8) *where* $\Sigma$ *satisfies the RWS model* (1.8)-(1.9).

- *If* $\alpha + \beta < 1$*, then as* $n \to \infty$*, the level of the CS test defined in* (1.6) *and* (1.11) *tends to* 0 *while the power tends to* 1.
- *If* $\alpha + \beta > 1$ *and* $H_1^{(n)}$ *holds, then for sufficiently large* $n$*, with probability at least* $1 - C \log(n) n^{1-(\alpha+\beta)}$*,*

$$(1.12) \qquad \left|S_k - E_0[S_k]\right| \leq C\tilde{L}_n(k/n)^{2/3}, \qquad \text{for all } 1 \leq k \leq n.$$

To the best of my knowledge, the bounds in (1.12) are new. The probability bound on the excluded event (i.e., $C \log(n) n^{1-(\alpha+\beta)}$) is derived from the relationship between the $L^1$-distance and the Neyman-Pearson Lemma [49]; the bound may be improved, using presumably a different technique.

Next, consider the HC test statistic, focusing on $HC_n^+$ (the variant of $HC_n^*$). By definitions (i.e., (1.5)-(1.6)), we reject $H_0^{(n)}$ if and only if

$$(1.13) \qquad HC_n^+ \geq \tilde{L}_n, \qquad \text{where } \tilde{L}_n = [\log(n)]^{\log^2(\log(n))}.$$

**Theorem 1.3** *(Behavior of the HC test and bounds for* $\lambda_k$*). Fix* $\alpha, \beta \in (0, 1)$ *and* $\gamma > 1$.[4] *Suppose* $X_i \overset{iid}{\sim} N(0, \Sigma)$ *as in* (1.1)*. Let* $\tilde{L}_n = [\log(n)]^{\log^2(\log(n))}$*. Consider the testing problem* (1.7)-(1.8) *where* $\Sigma$ *satisfies the RWS model* (1.8)-(1.9)*.*

- *If* $\alpha + \beta < 1$*, then as* $n \to \infty$*, the level of the HC test defined in* (1.6) *and* (1.13) *tends to* 0.
- *If* $\alpha + \beta > 1$ *and* $H_1^{(n)}$ *holds, then for sufficiently large* $n$*, with probability at least* $1 - \log(n) n^{1-(\alpha+\beta)}$*,*

$$\left|\lambda_k - E_0[\lambda_k]\right| \leq \tilde{L}_n n^{-2/3}[k \wedge (n + 1 - k)]^{-1/3}, \qquad \text{for all } 1 \leq k \leq n.$$

Compared to bounds of $\lambda_k$ in literature (e.g., [4, 5, 40]), our results are new for (a) the literature focus on the case where $r_n$ is fixed as $n \to \infty$, while our results are for the case of $r_n = n^{1-\beta}$, and (b) the literature focus on bounds for only edge eigenvalues and our bounds are for all eigenvalues.

Last, consider the Tracy-Widom test statistic $TW_n$. It is known that under $H_0^{(n)}$, $TW_n$ converges weakly to the Tracy-Widom law. Let $F_{TW}$ be the CDF

---

[4]The case of $\gamma = 1$ is more complicated, for the behavior of the smallest eigenvalues is different. To save space, we omit discussions of this case.

of the Tracy-Widom law. It is known that $1 - F_{TW}(t) \sim e^{-(2/3)t^{3/2}}$ [22] for large $t$. In light of this, we reject $H_0^{(n)}$ if and only if

$$(1.14) \qquad TW_n \geq [3\log(n)]^{2/3}.$$

**Corollary 1.1** *(Sub-optimality of the TW test). Fix $\alpha, \beta \in (0,1)$ and $\gamma \geq 1$. Suppose $X_i \overset{iid}{\sim} N(0, \Sigma)$ as in (1.1). Consider the testing problem (1.7)-(1.8) where $\Sigma$ satisfies the RWS model (1.8)-(1.9). If $\alpha + \beta < 1$, then as $n \to \infty$, the level of the TW test defined in (1.13) tends to $0$. If additionally $\alpha > 2/3$, then the power of the TW test also tends to $0$.*

The TW test is asymptotically powerless in the sub-region

$$\{(\beta, \alpha) : 0 < \beta < 1, \alpha > 2/3, \alpha + \beta < 1\};$$

In this region, both the CS test and the trace-based test have asymptotically full power. This result is consistent with our numerical study (see Section 2), where we observe that when there are multiple spikes, the TW test usually behaves unsatisfactorily.

**Remark**. (*Comparison with the testing problem of sparse normal means*). Consider a setting where we have samples $y_i \sim N(\mu_i, 1)$, $1 \leq i \leq n$. We are interested in testing whether $\mu_i = 0$ for all $1 \leq i \leq n$, or that most of $\mu_i$ are $0$ except for a small fraction of them are nonzero. In such a sparse normal means test setting, it is preferable to use a small fraction of extreme observations (instead of the bulk) for testing (e.g., [19]).

To those familiar with the normal means problem, Theorem 1.1 may strike as a surprise, for it shows the trace-based test is optimal, even when the spikes are very sparse; this is so because the current setting is very different from the normal means settings: every spike in $\Sigma$ affects the *bulk* of $\lambda_i$ in a subtle way; and even when the spikes in $\Sigma$ are very sparse, the sparsity is lost when we look at the vector

$$(1.15) \qquad \big(E_1[\lambda_1] - E_0[\lambda_1],\ E_1[\lambda_2] - E_0[\lambda_2],\ \ldots,\ E_1[\lambda_n] - E_0[\lambda_n]\big),$$

where $E_0$ and $E_1$ are the expectation under $H_0^{(n)}$ and $H_1^{(n)}$, respectively. In Figure 2, we plot the vector in (1.15) for the cases $r = 1$ and $r = 5$ ($n, p, \delta$ are the same as those of Table 2). We observe that as soon as $r$ move away from 1, the vector in (1.15) become resonably non-sparse.

**Remark**. (*Other trace-based statistics*). One possible variant of the trace-based test $S_n$ is $S_n^{(2)} = \sum_{k=1}^n (\lambda_k - 1)^2$. The test targets on the case where the off-diagonals of $\hat{\Sigma}$ contain stronger evidence against the null than the diagonals, so it is not surprising that, asymptotically, for $(\alpha, \beta)$ such that
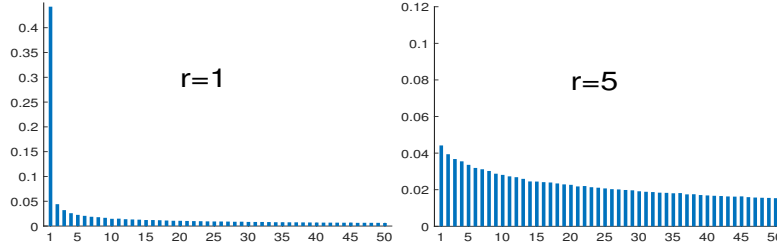
FIG 2. *Plots of $E_1[\lambda_k] - E_0[\lambda_k]$ for $k = 1, \cdots, 50$, where $(n, p) = (1000, 1200)$, $(r, \delta) = (1, 2)$ in left panel, and $(r, \delta) = (5, 0.8)$ in right panel.*

$2\alpha + \beta > 1$ while $\alpha + \beta < 1$, $S_n^{(2)}$ is powerless while $S_n$ has full power. On the other hand, $S_n^{(2)}$ may be more powerful in some other cases (e.g., [14]).

1.4. *Extensions and Le Cam's comparison of experiments.* In his work on comparison of experiments, Le Cam asserts that "adding noise always makes the inference more difficult". This allows us to compare our setting with many other settings, and generalize our lower bound argument in Theorem 1.1 (one of our major contribution in this paper) to much broader settings.

We now consider two extensions of our lower bound argument. In the first extension, for two covariance matrices $\Sigma$ and $\widetilde{\Sigma}$, we compare two experiments. In the first one, we are interested in testing

$$H_0^{(n)} : X_i \overset{iid}{\sim} N(0, \Sigma) \qquad vs. \qquad H_1^{(n)} : (X_i | Q) \overset{iid}{\sim} N(0, \Sigma + \frac{\delta}{1 - \delta} QQ'),$$

and in the second one, we are interested in testing

$$\widetilde{H}_0^{(n)} : X_i \overset{iid}{\sim} N(0, \widetilde{\Sigma}) \qquad vs. \qquad \widetilde{H}_1^{(n)} : (X_i | Q) \overset{iid}{\sim} N(0, \widetilde{\Sigma} + \frac{\delta}{1 - \delta} QQ'),$$

where as before, $Q$ is uniformly distributed over the Stiefel manifold $\mathbb{S}(p, r)$. Let $f_0, f_1, \tilde{f}_0, \tilde{f}_1$ be the joint density of $\{X_i\}_{i=1}^n$ under $H_0^{(n)}$, $H_1^{(n)}$, $\widetilde{H}_0^{(n)}$, and $\widetilde{H}_1^{(n)}$, respectively. We assume $\Sigma \preceq \widetilde{\Sigma}$, so the second experiment can be viewed as the result of adding noise to the first experiment. Lemma 1.1 solidifies the claim that "adding noise makes the testing problem harder".

**Lemma 1.1** *If $\Sigma \preceq \widetilde{\Sigma}$, then $\|\tilde{f}_1 - \tilde{f}_0\|_1 \le \|f_1 - f_0\|_1$.*

Applying Lemma 1.1 with $\Sigma = c_0 I_p$ and $\widetilde{\Sigma} = \Sigma^*$ (see below), we have the following theorem, as a direct result of Theorem 1.1.

**Theorem 1.4** *(Extension of lower bound argument, I). Fix $\alpha, \beta \in (0, 1)$, $c_0 > 0$, and $\gamma \ge 1$. Given $n$ independent vectors $X_i \in \mathbb{R}^p$, $1 \le i \le n$,*

we are interested in testing $H_0^{(n)} : X_i \overset{iid}{\sim} N(0, \Sigma^*)$ versus the alternative $H_1^{(n)} : (X_i | Q) \overset{iid}{\sim} N(0, \Sigma^* + \frac{\delta}{1-\delta} QQ')$, where $Q$ is uniformly distributed over the Stiefel manifold $\mathbb{S}(p, r)$, $\Sigma^* \in \mathbb{R}^{p,p}$ is a covariance matrix, and $(\delta, r, p)$ satisfies (1.9) as in Theorem 1.1. If $\alpha + \beta > 1$ and $\lambda_{\min}(\Sigma^*) \geq c_0$,[5] then as $n \to \infty$, the sum of Type I and Type II errors of any test tends to $1$.

For another extension of the lower bound argument, we show that "either adding more spikes or increasing the spike strengths makes the testing problem (1.7)-(1.8) easier". Consider two experiments. In the first one, we are interested in testing

$$H_0^{(n)} : X_i \overset{iid}{\sim} N(0, I_p) \qquad vs. \qquad H_1^{(n)} : (X_i | Q) \overset{iid}{\sim} N(0, I_p + QDQ').$$

In the second one, we are interested in testing

$$H_0^{(n)} : X_i \overset{iid}{\sim} N(0, I_p) \qquad vs. \qquad \widetilde{H}_1^{(n)} : (X_i | Q) \overset{iid}{\sim} N(0, I_p + Q\widetilde{D}Q'),$$

where $D, \widetilde{D} \in \mathbb{R}^{r,r}$ are positive semi-definite matrices, and $Q$ is uniformly distributed over the Stiefel manifold $\mathbb{S}(p, r)$. Let $f_0, f_1, \tilde{f}_1$ be the joint density of $\{X_i\}_{i=1}^n$ under $H_0^{(n)}$, $H_1^{(n)}$, and $\widetilde{H}_1^{(n)}$, respectively. Similarly, we have the following lemma, which is proved in Section 4 with nontrivial efforts.

**Lemma 1.2** If $D \preceq \widetilde{D}$, then $\|f_1 - f_0\|_1 \leq \|\tilde{f}_1 - f_0\|_1$.

We also have the following theorem.

**Theorem 1.5** (Extension of lower bound argument, II). Fix $\alpha, \beta \in (0, 1)$, $c_0 > 0$, and $\gamma \geq 1$. Given $n$ independent vectors $X_i \in \mathbb{R}^p$, $1 \leq i \leq n$, we are interested in testing $H_0^{(n)} : X_i \overset{iid}{\sim} N(0, I_p)$ versus the alterative $H_1^{(n)} : (X_i | Q) \overset{iid}{\sim} N(0, I_p + QD^*Q')$, where $Q$ is uniformly distributed over the Stiefel manifold $\mathbb{S}(p, r)$, $D^* \in \mathbb{R}^{r,r}$ is a positive semi-definite matrix, and $(\delta, r, p)$ satisfies (1.9) as in Theorem 1.1. If $\alpha + \beta > 1$ and $\|D^*\| \leq c_0 \frac{\delta}{1-\delta}$, then as $n \to \infty$, the sum of Type I and Type II errors of any test tends to $1$.

For a proof, assume $c_0 = 1$ without loss of generality. Applying Lemma 1.2 with $D = D^*$ and $\widetilde{D} = \frac{\delta}{1-\delta} I_r$, Theorem 1.5 follows from Theorem 1.1.

1.5. *Spike detection in the critical case of $r_n \delta_n = O(1)$.* In Theorem 1.1, we have either $r_n \delta_n \to \infty$ or $r_n \delta_n \to 0$. We consider a more subtle case where $r_n \delta_n = O(1)$ (in RWS, this is the case of $\alpha + \beta = 1$).

---

[5]For a symmetric matrix $A$, $\lambda_{\min}(A)$ stands for the minimum eigenvalue of $A$.

For analysis, we continue to use RWS, except for that $(r_n, \delta_n)$ are calibrated slightly differently. In detail, fixing the parameter $\beta \in (0,1)$, $\gamma \geq 1$ and $\theta > 0$, we assume as $n \to \infty$,

$$(1.16) \qquad p_n/n \to \gamma, \qquad r = r_n = n^{1-\beta}, \qquad \delta = \delta_n = r_n^{-1}(\theta\sqrt{2\gamma}).$$

The following Theorem is proved in Section 4.

**Theorem 1.6** *(Critical case of $\alpha + \beta = 1$). Fix $\beta \in (0,1)$, $\theta > 0$, and $\gamma \geq 1$. Consider the testing problem* (1.7)-(1.8) *where* (1.16) *holds. As $n \to \infty$, the log-likelihood ratio $\log(LR_n)$ converges weakly to $N(\mp\frac{\theta^2}{2}, \theta^2)$, under the null and under the alternative, respectively.*

Recall that $S_n$ denotes the trace-based test statistic. By elementary statistics, it is seen that (the convergence is weak convergence):

$$\theta \frac{(nS_n - np)}{\sqrt{2np}} - \frac{\theta^2}{2} \to N(\mp\frac{\theta^2}{2}, \theta^2), \quad \text{under } H_0^{(n)} \text{ and } H_1^{(n)}, \text{ respectively.}$$

This suggests that the trace-based statistic $S_n$ is asymptotically efficient.

In our case, $r_n\delta_n = O(1)$ but $r_n \to \infty$. A closely related case is that as $n \to \infty$, both $(r, \delta)$ are fixed. Such a case was studied in detail in [38], with very different techniques. Our framework can be extended to such a case. Let $\psi(\lambda) = \frac{\gamma}{\delta}\log(1 + \frac{\delta}{\gamma(1-\delta)} - \frac{\delta}{\gamma}\lambda)$. Let $\mu_{\hat{\Sigma}}$ be the empirical spectral measure associated with $\hat{\Sigma}$ and let $\mu_{n,p}$ be the Marchenko-Pastur law (see (3.17) for definition). By a simple modification of the proof of Theorem 1.6, we can show that $\log(LR_n) = -\frac{nr\delta}{2}[\int g(\lambda)\mu_{\hat{\Sigma}}(d\lambda) - \int g(\lambda)\mu_{n,p}(d\lambda)] - \frac{r^2\delta^2}{2\gamma(1-\delta)} + o_P(1)$. So $\log(LR_n)$ converges to a weak limit according to the central limit theorem for linear spectral statistics [3]. It yields that $\log(LR_n) \to N(\mp\frac{\tilde{\theta}^2}{2}, \tilde{\theta}^2)$ under $H_0^{(n)}$ and $H_1^{(n)}$, respectively, where $\tilde{\theta} = r[-\frac{1}{2}\log(1 - \frac{\delta^2}{\gamma})]^{1/2}$. This result coincides with that of [38, Proposition 4].

1.6. *A stylized application: Covert Communications.* We wish to communicate with our "friends" through a covert channel, and our "enemy" is trying to intercept it. We encode the desired information by a length-$p$ string of "0" and "1", denoted by $\eta$ (note $\eta$ has $r$ nonzero entries). Fix $\delta > 0$ and a $p \times p$ matrix $\widetilde{Q} = [q_1, q_2, \ldots, q_p]$. We send our friends vectors $X_i \in \mathbb{R}^p$

$$X_i = \delta \sum_{\{k:\eta(k)=1\}} z_k \cdot q_k + Z_i, \qquad 1 \leq i \leq n,$$

where $z_k \overset{iid}{\sim} N(0,1)$, $Z_i \overset{iid}{\sim} N(0, I_p)$, and they are independent. Let $Q$ be the $p \times r$ matrix $\{q_k : \eta(k) = 1, 1 \leq k \leq p\}$. Note that $X_i$ can be equivalently viewed as samples from $N(0, \Sigma)$, where $\Sigma = I_p + \delta QQ'$.

Assume that the matrix $\widetilde{Q}$ is available to our "friends", but not to our "enemy". When our "friends" receive $X_i$, first, they can obtain $\widetilde{X}_i$:

$$\widetilde{X}_i = \delta \sum_{\{k:\eta(k)=1\}} z_k \cdot e_k + \widetilde{Z}_i, \qquad \widetilde{X}_i = \widetilde{Q}'X_i, \qquad \widetilde{Z}_i = \widetilde{Q}'Z_i.$$

Next, from $\widetilde{X}_i$, they are able to retrieve the vector $\eta$, provided with some mild conditions on $(r, \delta)$. For our "enemy", $\widetilde{Q}$ is not available, so the problem reduces to our previous setting of (1.8). Note also that if $\widetilde{Q}$ is randomly generated, then $Q$ is uniformly distributed in the Stiefel manifold $\mathbb{S}(p, r)$. Applying Theorem 1.1 and conventional results of sparse normal-means problems, we have the following theorem:

**Theorem 1.7** *(Covert communitcation). Fix $\alpha, \beta \in (0, 1)$ and $\gamma \geq 1$. Suppose $(\delta, r, Q)$ satisfy model (1.8)-(1.9), and where $1/2 < \beta < 1$ and $(1-\beta) < \alpha < 1/2$. Then with high probability, our "friends" are able to exactly decode the vector $\eta$, while our "enemy" is not able to even distinguish whether we are sending some signals or we are merely sending white noise.*

1.7. *Summary.* Our results provide both a better understanding of the problem of detecting rare/weak spikes, and better eigenvalue bounds for the empirical covariance matrix $\hat{\Sigma}$ (based on *iid* samples from $N(0, \Sigma)$).

For the testing problem, we adopt a Rare and Weak Spike (RWS) model and discover an interesting phase transition. We study four different tests, where the Higher Criticism test and the CuSum test are new. We show that both the trace test and the CuSum test are optimal. We find that even when the spikes are very rare, it is still necessary to use *many* eigenvalues (instead of a few extreme eigenvalues) for testing against the null; such a finding is very different from those in testing settings regarding sparse normal means.

Motivated by the interest of "$p > n$" in modern applications, we assume $\gamma \geq 1$, but this should not be taken as a constraint. When $p/n \to \gamma < 1$, the detection boundary is the same, and the asymptotic behavior of all four tests can be studied similarly.

Our study is connected to the large body of literature on testing of sphericity [2, 9, 12, 14, 18, 32, 33, 34, 37, 38], but most of these works focus on specific tests, not on the phase transition. A few exceptions are [9, 12, 37, 38], but the alternative hypotheses considered there are very different from ours: in our setting, the number of spikes grow rapidly with $n$; such a case has not been studied in the literature.

In [37, 38], they considered a testing framework equivalent to our RWS model with both $(r, \delta)$ being constants. The test statistic they proposed is a function of all empirical eigenvalues, resulting from a Laplace approximation

of the log-likelihood ratio. They showed that the asymptotic power of their test is better than that of the Tracy-Widom test. In a high level, both their works and our work point out the advantage of using bulk eigenvalues in the test, even for a small $r$. However, there are several major differences between the two works: We focus on the setting of Rare/Weak Spikes where $r \to \infty$ and $\delta \to 0$; there is no straightforward extension of their method/theory to our setting. Their test is essentially the likelihood ratio test, which requires knowing $(r, \delta)$, but all four tests considered in our work are adaptive, whose construction doesn't depend on parameters of the alternative. The technical approaches are also different. Their main tools are the Laplace's method and large-deviation analysis of spherical integrals; how to adapt these techniques to our setting is unclear. We use the Gaussian proxy model and Le Cam's comparison of experiments. The Gaussian proxy model and the theory of its intimacy to RWS (Lemma 4.1), as a new technical device, will be useful for studying other problems associated with RWS.

For bounds on the eigenvalues of $\hat{\Sigma}$, while most of the literature have been focused on the case of $\Sigma = I_p$, our results focus on the case where $\Sigma$ is a spike matrix. For the latter, a few works exist (e.g., [4, 5, 40]), but the focus there is on the bounds for the extreme eigenvalues *only*, and is for the case of *finitely* many spikes. In comparison, our bounds are for all eigenvalues and are for the case where the number of spikes grows rapidly with $n$ as $n \to \infty$. See Section 5 for more discussion.

Our philosophy is that, knowledge flow is a two-way street: while RMT may help us obtain better results in statistical inference, statistical inference can also provide better results in RMT, in return.

1.8. *Content and notations.* The remaining part of the paper is organized as follows. Section 2 contains a small-scale numerical study. In Section 3, we present a list of useful results on Random Matrix Theory (RMT), including some new results; this section can be read independently. In Section 4, we prove the main theorems and lemmas. Section 5 contains discussions. Proofs of the secondary lemmas are relegated to the appendix.

For two real numbers $a$ and $b$, $a \wedge b$ and $a \vee b$ denote the minimum and maximum of them, respectively. We say two positive sequences $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$, $a_n \asymp b_n$ if $|a_n/b_n|$ are uniformly upper and lower bounded by constants. When $\xi$ is a vector, $\|\xi\|$ denotes the vector $L^2$-norm; when $\xi$ is a matrix, $\|\xi\|$ denotes the matrix spectral norm and $\|\xi\|_F$ denotes the matrix Frobenius norm. For two square matrices $A, B$, we say $A \preceq B$ if $B - A$ is positive semidefinite. For two probability densities $f$ and $g$, $\|f - g\|_1 \equiv \int |f(x) - g(x)|dx$ denotes the $L^1$-distance between them.

**2. Numerical study.** We investigate the four tests with a small-scale experiment. Fix $(n, p) = (1000, 1200)$. For each $(r, \delta)$, we generate 100 data sets of $\{X_i\}_{i=1}^n$ from (1.7) to represent the null, and generate 100 data sets of $\{X_i\}_{i=1}^n$ from (1.8) to represent the alternative, and apply all four tests. We measure the "ideal testing error" for each test which corresponds to the rejection threshold that minimizes the sum of type I and type II errors over 200 data sets. For each setting, we run 20 repetitions.

First, we compare the HC test with the TW test.[6] The HC statistic is the maximum of $HC_{n,k}$ over $k$, and each $HC_{n,k}$ can be used as a test statistic, including TW as a special case with $k = 1$. We record the "ideal testing error" of $HC_{n,k}$ for all $1 \leq k \leq n$, and let $\hat{k}_{hc}$ minimize this error. We also record the "ideal testing error" of the HC test, where $k$ is chosen adaptively by data. The results are summarized in Table 2.

TABLE 2
*Comparison of TW and HC tests (standard deviations are in brackets).*

| $(r, \delta)$ | $\hat{k}_{hc}$ | TW | HC |
|:---:|:---:|:---:|:---:|
| $(1, 2)$ | $1(0)$ | $0(0)$ | $.001(.003)$ |
| $(5, .8)$ | $22(9.7)$ | $.48(.05)$ | $.28(.04)$ |
| $(10, .5)$ | $53(25)$ | $.56(.05)$ | $.19(.03)$ |

The results suggest that (a) when $r$ is very small, $\hat{k}_{hc}$ is very close to 1, and two tests have very similar behaviors, with TW being slightly better, and (b) when $r$ gets larger, $\hat{k}_{hc}$ is bounded away from 1, and the HC test is usually better, and significantly so when $r$ increases.

Next, we compare the trace test, CuSum test, and TW test. Each cumulative sum $CS_{n,k}$ can be used as a test statistic, with TW and trace being two special cases of $k = 1$ and $k = n$, respectively. Similarly, we let $\hat{k}_{cs}$ be the $k$ such that the "ideal testing error" of $CS_{n,k}$ is minimized. The results are summarized in Table 3.

TABLE 3
*Comparison of TW, Trace, and CuSum (standard deviations are in brackets).*

| $(r, \delta)$ | $\hat{k}_{cs}$ | TW | Trace | CuSum |
|:---:|:---:|:---:|:---:|:---:|
| $(1, 2)$ | $1(0)$ | $0(0)$ | $.48(.05)$ | $0(0)$ |
| $(5, .8)$ | $125(65)$ | $.48(.05)$ | $.18(.03)$ | $.16(.03)$ |
| $(10, .5)$ | $321(207)$ | $.56(.05)$ | $.08(.03)$ | $.12(.03)$ |

It suggests that (a) when $r$ is very small, $\hat{k}_{cs}$ is very close to 1, and TW and CuSum have similar performances, and (b) when $r$ gets larger, $\hat{k}_{cs}$ is

---

[6]For HC, $E_0[\lambda_k]$ and $SD_0(\lambda_k)$ are computed by simulating the null for $10{,}000$ times. Same for CuSum.
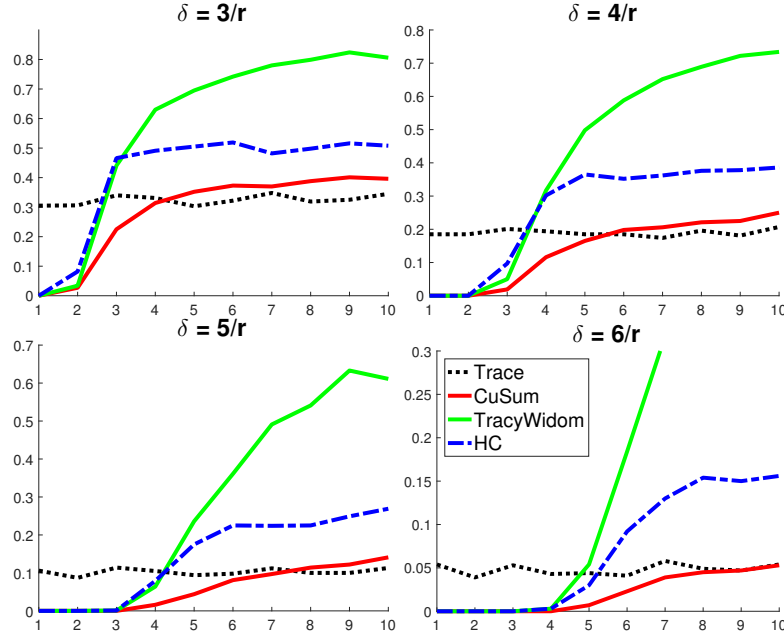
FIG 3. *Comparison of the ideal testing errors ($p = 1200$, $n = 1000$). The x-axis represents the number of spikes $r$, and for each $r = 1, \cdots, 10$, the strength of spikes is $\delta = c/r$, for a constant $c > 0$.*

much larger than 1, but it is also smaller than $n$; CuSum performs better than the trace test, and the trace test outperforms the TW test.

Third, we compare all four tests for a variety of $(r, \delta)$. Fixing a constant $c > 0$, we consider $r \in \{1, \cdots, 10\}$ and for each $r$ we let $\delta = c/r$. As $r$ changes, the performance of the trace test is roughly the same because

$$\frac{S_n - np}{\sqrt{2np}} \approx N(0, 1) \text{ and } \approx N(\frac{n(\delta r)}{\sqrt{2np}}, 1), \text{ under } H_0^{(n)} \text{ and } H_1^{(n)}, \text{ respectively.}$$

So the trace test can be used as a benchmark. The results for $c \in \{3, 4, 5, 6\}$ are displayed in Figure 3. When $r = 1$, the TW test usually performs the best, and the trace test behaves unsatisfactorily. When $r$ is slightly larger, TW becomes less satisfactory and is inferior to all other three tests; in this setting, trace and CuSum have the best performance.

Recall that CuSum can be viewed as a hybrid of the trace and TW tests, we expect that CuSum has both the advantage of TW and trace. This is confirmed by the numerical results: when $r$ is very small, CuSum behaves similarly to TW, and when $r$ is slightly larger, CuSum behaves similarly to trace. Overall, it seems CuSum has the best performance, especially when
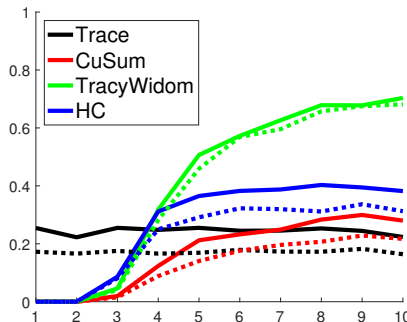
FIG 4. *The case where some non-spiked eigenvalues are smaller than* 1 *under the alternative (p = 1500, n = 800). Dotted lines: all non-spiked eigenvalues are* 1*. Solid lines: most non-spiked eigenvalues are* 1*, except that* 30 *of them are drawn uniformly from* [0.95, 1].

the spikes are relatively strong (i.e., $c$ is large).

Fourth, we investigate a case where a small number of non-spiked eigenvalues are strictly smaller than 1 under the alternative. Fix $(n, p) = (800, 1500)$. For each of $r \in \{1, 2, \ldots, 10\}$, let $\delta = 5/r$. The null hypothesis is the same as before. For the alternative hypothesis, we generate $X_i \overset{iid}{\sim} N(0, \Sigma)$, where $\Sigma$ is a $p \times p$ diagonal matrix such that $\Sigma(j, j) = 1 + \delta$ for $1 \le j \le r$, $\Sigma(j, j)$ is drawn uniformly from $[0.95, 1]$ for $p - 29 \le j \le p$, and $\Sigma(j, j) = 1$ for the remaining. The ideal testing errors of all four tests are displayed in Figure 4 (solid lines). As a benchmark, we also consider a similar setting where the only difference is that $\Sigma(j, j) = 1$ for $p - 29 \le j \le p$ under the alternative. The results for the bench mark setting are also in Figure 4 (dotted lines).

Compared with the benchmark setting, when some non-spiked eigenvalues are strictly smaller than 1, the performance of all four tests become worse. The reason is that each empirical eigenvalue $\lambda_k$ is affected by all population eigenvalues (spiked and non-spiked ones). Decreasing of non-spiked eigenvalues will make $\lambda_k$'s smaller, so the powers of all four tests deteriorate. Such an effect is more mild for smaller $k$, which explains why the performance of TW test remains almost the same but the performance of trace test changes more significantly. By looking at the solid curves only, we can see that the CuSum test still has the best overall performance.

Last, we consider the broader settings in Section 1.4 where $\Sigma$ is not necessarily the identity matrix under the null. Fix $(n, p) = (800, 1500)$. Given $b \in [0, 1)$, we generate a $p \times p$ diagonal matrix $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ where $\sigma_j$'s are *iid* drawn from the uniform distribution on $[1 - b, 1 + b]$. Then, for each $(r, \delta)$, letting $\Sigma^* \in \mathbb{R}^{p,p}$ be the diagonal matrix with $\Sigma^*(j, j) = \Sigma(j, j) + \delta \cdot 1\{1 \le j \le r\}$, we consider the testing problem where $X_i \overset{iid}{\sim} N(0, \Sigma)$ under

the null and $X_i \overset{iid}{\sim} N(0, \Sigma^*)$ under the alternative. Although the null model is different from (1.7), we still use the same implementation for all tests; for example, when computing the CuSum test statisitc, we still use $E_0(S_k)$ and $SD_0(S_k)$ simulated from model (1.7). We consider $b \in \{0, 0.1, 0.2\}$, and for each choice of $b$, we run experiments for $r \in \{1, 2, \ldots, 10\}$ with $\delta = 5/r$. The results are displayed in Figure 5.

For $b = 0.1$, results are similar to those for $\Sigma = I_p$ (i.e., $b = 0$). For $b = 0.2$, when $r$ is small, results are similar to those for $\Sigma = I_p$; when $r$ gets larger, the performance of all tests becomes worse than that in the case of $\Sigma = I_p$ but is still reasonably good. Note that we haven't used any knowledge of $\Sigma$ in these tests. It is hopeful to further improve the performance by incorporating $\Sigma$, say, to compute the mean and standard deviation of $\lambda_k$ and $S_k$.

**3. Some results on Random Matrix Theory (RMT).** In this section, we present some Random Matrix Theory that are useful for our proofs. Some of them come from existing literature and some are newly developed.

We write $Z \in \mathcal{G}(n, p)$ if $Z$ is an $n \times p$ random matrix with independent entries of $N(0, 1)$. With probability 1, the matrix $(1/n)Z'Z$ has $\min\{n, p\}$ distinct positive eigenvalues [48], denoted as $\lambda_1 > \lambda_2 > \cdots > \lambda_{n \wedge p} > 0$. We are interested in both cases of $p/n \to 0$ and $p/n \to \gamma > 0$.

Let $\mu_n^Z$ be the empirical spectral measure of $(1/n)Z'Z$: for any real value $x$, $\mu_n^Z((-\infty, x]) = p^{-1} \sum_{j=1}^p 1\{\lambda_j \le x\}$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ are all the eigenvalues of $(1/n)Z'Z$. Let $\mu_{n,p}$ be the Marchenko-Pastur (MP) measure which has a point mass of $\max\{1 - n/p, 0\}$ at 0 and the probability density

$$(3.17) \qquad \mu_{n,p}(x) = \frac{n}{2\pi x p} \sqrt{(x - a_{n,p}^-)(a_{n,p}^+ - x)}, \qquad a_{n,p}^- < x < a_{n,p}^+,$$

where $a_{n,p}^{\mp} = (1 \mp \sqrt{p/n})^2$. For any $1 \le k \le (n \wedge p)$, let $q_k = q_k^{(n,p)}$ be such that

$$(3.18) \qquad \int_{q_k}^{a_{n,p}^+} \mu_{n,p}(x) dx = k/n.$$

We also define $\mu_\gamma$, the Marchenko-Pastur (MP) measure associated with $\gamma$, which has a point mass of $\max\{1 - 1/\gamma, 0\}$ at 0 and the probability density

$$(3.19) \qquad \mu_\gamma(x) = \frac{1}{2\pi x \gamma} \sqrt{(x - a_\gamma^-)(a_\gamma^+ - x)}, \qquad a_\gamma^- < x < a_\gamma^+,$$

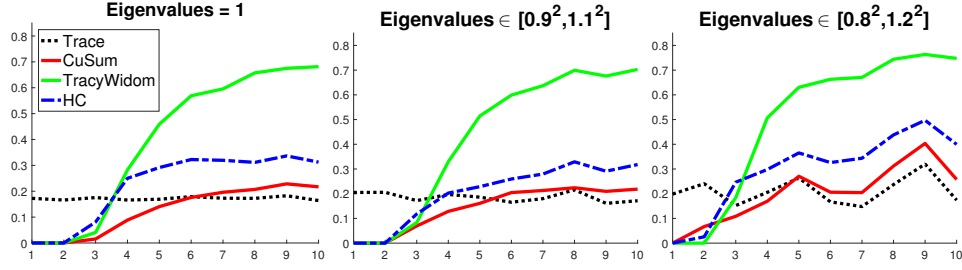where $a_\gamma^{\mp} = (1 \mp \sqrt{\gamma})^2$. Note that $\mu_{n,p}$ is a special case of $\mu_\gamma$ with $\gamma = p/n$.

FIG 5. *The case where $\Sigma$ is not the identity matrix under the null ($p = 1500$, $n = 800$). The eigenvalues of $\Sigma$ are between $(1 - b)^2$ and $(1 + b)^2$, for $b = 0$ (left), $0.1$ (middle) and $0.2$ (right).*

3.1. *Deviation of linear eigenvalue statistics.* In this section, we assume $p \leq n$ without loss of generality. For any function $f : \mathbb{R} \to \mathbb{R}$, we call

$$\mu_n^Z(f) = \frac{1}{p} \sum_{i=1}^{p} f(\lambda_i)$$

the *linear eigenvalue statistic* associated with $f$. The following theorem establishes a deviation inequality for such linear statistics when $f'$ is uniformly bounded. It is proved in the appendix.

**Theorem 3.1** *Suppose $Z \in \mathcal{G}(n, p)$, $p \leq n$ and that $f$ is continuously differentiable satisfying $\|f'\|_\infty \equiv \sup_x |f'(x)| < \infty$. For any $t > 0$,*

$$P \left( |\mu_n^Z(f) - E[\mu_n^Z(f)]| > t \|f'\|_\infty / \sqrt{np} \right) \leq 2 \inf_{y \geq \sqrt{2 \log(2)}} \left\{ e^{-\frac{t^2}{c_0 y^2}} + 2 e^{-y^2} \right\},$$

*where $c_0 > 0$ is an absolute constant.*

A similar large-deviation inequality is given in [29, Corollary 1.8(b)], but Theorem 3.1 improves it in two folds. First, Theorem 3.1 gives a sharper rate in the case $p/n \to 0$. Second, Theorem 3.1 only requires that $f$ is Lipschitz, while [29, Corollary 1.8(b)] requires that $f(x^2)$ is Lipschitz. However, we pay a price of the extra term $2 e^{-y^2}$ on the right hand side.

A direct application of Theorem 3.1 is to derive a large deviation result for $\sum_{i=1}^{p} (\lambda_i - 1)^m$ in the case $p/n \to 0$, where $m$ is a positive integer. The following lemma is proved in the appendix.

**Lemma 3.1** *Suppose $Z \in \mathcal{G}(n, p)$ and $p/n \to 0$ as $n \to \infty$. For any fixed integer $m \geq 1$, as $p \to \infty$, with probability at least $1 - o(n^{-1})$,*

$$\left| \sum_{i=1}^{p} \{ (\lambda_i - 1)^m - E[(\lambda_i - 1)^m] \} \right| \leq C \log(n) \left[ \frac{p \vee \log(n)}{n} \right]^{m/2}.$$

Although large-deviation inequalities for $\sum_{i=1}^{p}(\lambda_i-1)^m$ could be obtained from existing results on polynomials of independent normals (e.g., [1, 51]), those inequalities are expressed in terms of tensor norms or subgraph counts, from which it is hard to get the explicit rate. We obtain Lemma 3.1 almost for free by applying Theorem 3.1 to $f(x) = (x-1)^m$ (with a proper truncation).

3.2. *Concentration of empirical spectral measure.* In this section, we focus on the case that $p/n \to \gamma$, where $\gamma > 0$. It is well known that $\mu_n^Z \to \mu_\gamma$ weakly. The following theorem, which is a special case of [28, Theorem 1.1], gives its convergence rate in terms of the Kolmogorov-Smirnov distance between distribution functions of $\mu_n^Z$ and $\mu_{n,p}$.

**Theorem 3.2** *Suppose $Z \in \mathcal{G}(n,p)$ and $p/n \to \gamma$ for a constant $\gamma > 0$. Let $F_n^Z$ and $F_{n,p}$ be the CDF of $\mu_n^Z$ and $\mu_{n,p}$, respectively. For any $\zeta > 0$, there exist positive constants $C$ and $c$ which depend on $(\zeta, \gamma)$, such that*

$$P\left(\sup_{x \in \mathbb{R}} |F_n^Z(x) - F_{n,p}(x)| > n^{-1}\omega_{n,\zeta}^6\right) \le C_\zeta \exp(-c_\zeta \omega_{n,\zeta}),$$

*where $\omega_{n,\zeta} = \log(n)[\log(\log(n))]^\zeta$.*

The next lemma gives an explicit expression of $\int \log(t-x)\mu_\gamma(x)dx$, for $t > a_\gamma^+$, and is proved in the appendix. We have not found such a result in the literature (although a formula for $\int \log(t+x)\mu_\gamma(x)dx$ is given in [47]).

**Lemma 3.2** *For any $\gamma > 0$ and $t > (1+\sqrt{\gamma})^2$,*

$$\int \log(t-x)\mu_\gamma(x)dx = \frac{\gamma-1}{2\gamma}\log\left((\gamma+1)A_t - (\gamma-1)B_t + 4\gamma\right)$$

(3.20) $$+ \frac{A_t - B_t}{2\gamma} - \frac{\gamma+1}{2\gamma}\log(A_t - B_t) + \frac{1}{\gamma}\log(2) + \frac{\gamma+1}{2\gamma}\log(\gamma),$$

*where $A_t = t - \gamma - 1$ and $B_t = \sqrt{(t-\gamma-1)^2 - 4\gamma}$.*

3.3. *Behavior of eigenvalues and their cumulative sums.* In this section, we study $\lambda_k$ and $S_k = \sum_{j=1}^{k} \lambda_j$. Again, we focus on the case that $p/n \to \gamma > 0$. Consider $\lambda_k$ first. The following theorem is a special case of [42, Theorem 3.3].

**Theorem 3.3** *Suppose $Z \in \mathcal{G}(n,p)$ and $p/n \to \gamma$ for a constant $\gamma > 0$ and $\gamma \neq 1$. Write $L_n = [\log(n)]^{\log(\log(n))}$. For any $\zeta > 0$, there exists a constant $C_\zeta > 0$ such that*

$$P\left(\cup_{1 \le k \le n}\left\{|\lambda_k - q_k| > L_n^{C_\zeta}n^{-2/3}[k \wedge (n+1-k)]^{-1/3}\right\}\right) \le n^{C_\zeta}\exp(-L_n^\zeta).$$

In the case $\gamma = 1$, Theorem 3.3 continues to hold, except that $1 \le k \le n$ is now replaced with $1 \le k \le (1-\eta)n$, for a positive constant $\eta \in (0,1)$; see the remarks in the end of [42, Section 4]. We note that a direct corollary of Theorem 3.2 gives similar results but only for $\omega_{n,\zeta}^6 \le k \le n - \omega_{n,\zeta}^6$, however, Theorem 3.3 holds for the whole range $1 \le k \le n$.

As a corollary of Theorem 3.3,

$$(3.21) \qquad |E(\lambda_k) - q_k| \le L_n^c n^{-2/3} \tilde{k}^{-1/3}, \quad SD(\lambda_k) \le L_n^c n^{-2/3} \tilde{k}^{-1/3},$$

where $\tilde{k} = k \wedge (n+1-k)$ and $c > 0$ is a properly large constant. The null behavior of $HC_n^+$ follows immediately. Now, we consider $HC_n^*$. To access its null behavior, we additionally need a lower bound for $SD(\lambda_k)$. Unfortunately, we have not found a desirable lower bound of $SD(\lambda_k)$ in the existing literature. According to [16, 45], we conjecture that

$$|E(\lambda_k) - q_k| \asymp n^{-2/3} \tilde{k}^{-1/3}, \quad SD(\lambda_k) \asymp \sqrt{\log(n)} n^{-2/3} \tilde{k}^{-1/3}.$$

The following theorem gives a similar result for the cumulative sum $S_k = \sum_{j=1}^k \lambda_j$. It is a generalization of the result in [8] and proved in [6].

**Theorem 3.4** *Suppose $Z \in \mathcal{G}(n,p)$ and $p/n \to \gamma$ for a constant $\gamma > 0$. Write $L_n = [\log(n)]^{\log(\log(n))}$. For any $\zeta > 0$, there exists a constant $C_\zeta > 0$ such that*

$$P\left( \cup_{1 \le k \le n} \left\{ |S_k - \sum_{i=1}^k q_i| > L_n^{C_\zeta} (k/n)^{2/3} \right\} \right) \le n^{C_\zeta} \exp(-L_n^\zeta).$$

The discussions of the null behavior of $CS_n^+$ and $CS_n^*$ are similar to those of $HC_n^+$ and $HC_n^*$, so we omit them.

**4. Proof of main theorems.** We prove the main results in Section 1. For Theorem 1.2, the first claim is a direct result of Theorem 3.4 (so the level of the test tends to 0) and that $CS_n^+ \ge S_n - E_0[S_n]$ (so the power of the test is no smaller than the power of the trace test), and the second claim follows from the lower bound argument and the null behavior of eigenvalues (Theorem 3.4). Similarly, Theorem 1.3 follows from Theorem 3.3 and the lower bound argument. Theorems 1.4-1.5 are direct results of Lemmas 1.1-1.2, respectively, and Theorem 1.7 follows directly from Theorem 1.1.

What remains is to prove Theorem 1.1, Corollary 1.1, Lemmas 1.1-1.2, and Theorem 1.6.

4.1. *Preliminary I: Gaussian proxy model and its intimacy to RWS.* At the heart of our analysis is to consider a proxy testing problem where we replace the alternative hypothesis $H_1^{(n)}$ by

$$(4.22) \quad \widetilde{H}_1^{(n)}: \qquad \widetilde{\Sigma} = [I_p - (\delta/p)YY']^{-1}, \quad Y = Z \cdot 1\{\|Z\| \le \frac{1}{2}\sqrt{p/\delta}\},^7$$

where $Z \in \mathbb{R}^{p,r}$ is the matrix that has *iid* $N(0,1)$ entries.

**Definition 4.1** *We call (1.9) and (4.22) the Gaussian proxy model of RWS.*

The next lemma says that two models are close to each other, for a wide region in the parameter space. Let $LR_n$ and $\widehat{LR}_n$ be the likelihood ratios associated with the RWS and the proxy model, respectively. Let $E_0$ stand for the expectation under $H_0^{(n)}$.

**Lemma 4.1** *(Intimacy of two models). Consider the testing problem* (1.7)-(1.8) *where* (1.9) *holds. If* $\alpha > \max\{1 - 5\beta/4, (1-\beta)/2\}$, *then as* $n \to \infty$, $E_0|LR_n - 1| - E_0|\widetilde{LR}_n - 1| \to 0$.

See Figure 1 for an illustration.

This lemma is proved in Section 4.8. The proof uses Le Cam's "comparison of experiments", an idea explained in Section 1.4. We conjecture a stronger result of $E_0|LR_n - \widetilde{LR}_n| \to 0$ holds; see Section 4.8 for more discussions.

Our motivation of introducing the proxy model is that it is much easier to analyze than the RWS. Let $F(Q)$ and $F(Y)$ be the CDF of $Q$ in RWS and $Y$ in proxy model. It will be shown in Section 4.2 that

$$(4.23) \qquad LR_n = (1-\delta)^{\frac{nr}{2}} \int \exp(\frac{\delta}{2}\mathrm{tr}(\hat{\Sigma}QQ'))dF(Q),$$

$$(4.24) \qquad \widetilde{LR}_n \approx (1-\delta)^{\frac{nr}{2}} e^{\frac{nr\delta}{2(1-\delta)}(1-\frac{r\delta}{p(1-\delta)})} \int e^{\frac{\delta n}{2p}\mathrm{tr}((\hat{\Sigma}-\frac{1}{1-\delta}I_p)YY')}dF(Y).$$

The main technical hurdle we face in RWS is that, for the integral in (4.23), it is hard to integrate $Q$ out. However, such a hurdle is removed by using a proxy model: for the integral in (4.24), we can integrate $Y$ out and derive a (relatively) simple formula, because the entries of $Y$ are (approximately) *iid* $N(0,1)$ and the integral can be expressed using linear spectral statistics of a standard Wishart matrix.

---

[7]Requiring that $\|Z\| \le \frac{1}{2}\sqrt{p/\delta}$ seems to be a nuisance but is necessary: otherwise the matrix $[I_p - (\delta/p)YY']$ may not be invertible; note that $P(\|Z\| > (1/2)\sqrt{p/\delta})$ is negligible.

4.2. *Preliminary II: Likelihood ratio of the proxy model.* We justify the approximation in (4.24) and calculate the integral over $Y$. The justification of (4.24) is relatively short and conventional, however, for $LR_n$, it is unclear how to derive such an approximation.

Let $Y \in \mathbb{R}^{p,r}$ be a random matrix of *iid* $N(0,1)$ entries. Since $r \ll p$, with probability 1, $p^{-1}YY'$ has $r$ distinct positive eigenvalues [48], denoted as $\eta_1 > \eta_2 \ldots > \eta_r > 0$. Then, the proxy model (4.22) is equivalent to

$$(4.25) \qquad \widetilde{H}_1^{(n)} : \quad \widetilde{\Sigma} = \begin{cases} [I_p - \delta(p^{-1}YY')]^{-1}, & \text{if } \delta\eta_1 \leq 1/2, \\ I_p, & \text{otherwise.} \end{cases}$$

Let $F(Y)$ be the joint distribution function of $Y$. The likelihood ratio $\widetilde{LR}_n$ associated with (4.25) is

$$\widetilde{LR}_n = (1-\delta)^{\frac{nr}{2}} \int_{\delta\eta_1 \leq 1/2} \exp\left\{\frac{\delta n}{2p}\mathrm{tr}(\hat{\Sigma}YY') + \frac{n}{2}g_n(Y,\delta,r)\right\} dF(Y)$$

$$(4.26) \qquad + P(\delta\eta_1 > 1/2),$$

where $g_n(Y,\delta,r) = \sum_{k=1}^{r} \log(1 - \frac{\delta}{1-\delta}(\eta_k - 1))$.

We then show that $\widetilde{LR}_n$ can be approximated by $\widetilde{R}_n$, to be introduced below, which has a relatively simple form. Introduce $g_n^*(Y,\delta,r) = -\frac{\delta}{1-\delta}\sum_{k=1}^{r}(\eta_k - 1) - \frac{\delta^2}{2(1-\delta)^2}\sum_{k=1}^{r} E[(\eta_k - 1)^2]$; it is the Taylor expansion of $g_n(Y,\delta,r)$ up to the second term, except for $(\eta_k - 1)^2$ is replaced by $E[(\eta_k - 1)^2]$. Substituting $g_n$ with $g_n^*$ gives a proxy of $\widetilde{LR}_n$:[8]

$$\widetilde{R}_n = (1-\delta)^{\frac{nr}{2}} \int \exp\left\{\frac{\delta n}{2p}\mathrm{tr}(\hat{\Sigma}YY') + \frac{n}{2}g_n^*(Y,\delta,r)\right\} dF(Y)$$

$$(4.27) \qquad = (1-\delta)^{\frac{nr}{2}} b_n(\delta,r) \int \exp\left\{\frac{\delta n}{2p}\mathrm{tr}\left((\hat{\Sigma} - \frac{1}{1-\delta}I_p)YY'\right)\right\} dF(Y),$$

where $b_n(\delta,r) = e^{\frac{nr\delta}{2(1-\delta)} - \frac{nr^2\delta^2}{4p(1-\delta)^2}}$. Here the second equality follows from that $\sum_{k=1}^{r} E[(\mu_k-1)^2] = r^2/p$ [47, Lemma 2.9]. Noting that the above $\widetilde{R}_n$ is finite only when $\frac{\delta n}{p}(\lambda_1 - \frac{1}{1-\delta}) < 1$, we simply set $\widetilde{R}_n = 1$ when $\frac{\delta n}{p}(\lambda_1 - \frac{1}{1-\delta}) \geq 1$.

The following lemma is proved in the appendix.

**Lemma 4.2** *Consider the testing problem* (1.7) *and* (4.22) *with parameters as in* (1.9). *If* $\alpha > \max\{(4 - 5\beta)/6, (1 - \beta)/2\}$, *then* $E_0|\widetilde{LR}_n - \widetilde{R}_n| \leq Cn^{1-\min\{2\alpha+\beta,\, 3\alpha+5\beta/2-1\}}$.

---

[8]We also remove $\delta\eta_1 < 1$ in the integral of (4.27) so that it has a closed form.

A nice property of $\widetilde{R}_n$ is that we can integrate out $Y$ explicitly in (4.27). Denote by $\mu_n^X$ the empirical spectral measure of $\hat{\Sigma}$ under the null, i.e., $\mu_n^X = p^{-1} \sum_{j=1}^p \delta_{\lambda_j}$, where $\delta_a$ represents the point mass at $a$ and $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$ are all eigenvalues of $\hat{\Sigma}$ (including zero ones). Let $\mu_{n,p}$ be the Marchenko-Pastur (MP) measure with parameter $\gamma = p/n$ as in (3.17). The next lemma gives two equivalent expressions of $\widetilde{R}_n$.

**Lemma 4.3** *Consider the testing problem* (1.7) *and* (4.22) *with parameters as in* (1.9). *Let* $\psi_n(\lambda) = \frac{p}{n\delta} \log(1 - \frac{\delta n}{p}(\lambda - \frac{1}{1-\delta}))$. *When* $\frac{\delta n}{p}(\lambda_1 - \frac{1}{1-\delta}) < 1$,

$$(4.28) \quad \widetilde{R}_n = (1 - \delta)^{\frac{nr}{2}} b_n(\delta, r) \exp\Big\{ -\frac{r}{2} \sum_{j=1}^p \log\Big(1 - \frac{\delta n}{p}(\lambda_j - \frac{1}{1-\delta})\Big)\Big\}$$

$$(4.29) \quad = e^{-\frac{nr^2\delta^2}{4p(1-\delta)^2}} \exp\Big\{ -\frac{nr\delta}{2}\Big[ \int \psi_n(\lambda)\mu_n^X(d\lambda) - \int \psi_n(\lambda)\mu_{n,p}(d\lambda)\Big]\Big\},$$

*where* $b_n(\delta, r)$ *is the same as in* (4.27).

We have seen that

$$LR_n \approx \widetilde{LR}_n \approx \widetilde{R}_n,$$

and according to (4.29), the key of analyzing $\widetilde{R}_n$ is to characterize the convergence of empirical spectral measure $\mu_n^X$ to the Marchenko-Pastur measure $\mu_{n,p}$. This is well-studied in Random Matrix Theory.

4.3. *Proof of Theorem* 1.1. We only prove the first claim. By Neyman-Pearson lemma, it suffices to show that when $\alpha + \beta > 1$, $E_0|LR_n - 1| \to 0$. Recall that $\widetilde{LR}_n$ is the likelihood ratio associated with the Gaussian proxy model, and $\widetilde{R}_n$ is a proxy of $\widetilde{LR}_n$ introduced in Section 4.2. By Lemma 4.1 and Lemma 4.2, $E_0|LR_n - 1| = E_0|\widetilde{R}_n - 1| + o(1)$. So it suffices to show that

$$(4.30) \qquad\qquad E_0|\widetilde{R}_n - 1| \to 0.$$

We now show (4.30). Write $\psi_n(\lambda) = \frac{p}{n\delta} \log(1 - \frac{\delta n}{p}(\lambda - \frac{1}{1-\delta}))$. Let $\mu_n^X$ be the empirical spectral measure associated with $\hat{\Sigma}$, and let $\mu_{n,p}$ be the Marchenko-Pastur (MP) measure as in Section 3. Using (4.29) of Lemma 4.3, we have
(4.31)
$$\log(\widetilde{R}_n) = -\frac{nr\delta}{2}\Big[ \int \psi_n(\lambda)\mu_n^X(d\lambda) - \int \psi_n(\lambda)\mu_{n,p}(d\lambda)\Big] - \frac{nr^2\delta^2}{4p(1-\delta)^2},$$

Let $F_n^X$ and $F_{n,p}$ be the CDF associated with $\mu_n^X$ and $\mu_{n,p}$, respectively. Introduce $E_n = \{X \in \mathbb{R}^{n,p} : \sup_{x \in \mathbb{R}} |F_n^X(x) - F_{n,p}(x)| \leq n^{-1} \log^6(n) \log^6(\log(n))\}$, and let $I_{E_n}$ be the indicator that $X \in E_n$, similarly for $I_{E_n^c}$. By Theorem 3.2,

as $n \to \infty$, we have $P_0(X \in E_n^c) = O(e^{-C \log(n) \log(\log(n))}) = o(n^{1-\alpha-\beta})$. Since $E_0(\widetilde{R}_n) \le E_0(\widetilde{LR}_n) + o(1)$, which is bounded, we have $E_0(\widetilde{R}_n \cdot I_{E_n^c}) = o(1)$. As a result,

$$(4.32) \qquad E_0(|\widetilde{R}_n - 1| \cdot I_{E_n^c}) \le E_0(\widetilde{R}_n \cdot I_{E_n^c}) + P_0(X \in E_n^c) \to 0.$$

Below, we consider $E_0(|\widetilde{R}_n - 1| \cdot I_{E_n})$. Over the event $E_n$, both the supports of $\mu_n^X$ and $\mu_{n,p}$ are strictly contained in $A_\epsilon = [(1 - \sqrt{\gamma})^2 - \epsilon, (1 + \sqrt{\gamma})^2 + \epsilon]$, for any $\epsilon \in (0, 1)$. Furthermore, $\psi_n'(\lambda) = -[1 - \frac{\delta n}{p}(\lambda - \frac{1}{1-\delta})]^{-1}$; it follows that $\sup_{\lambda \in A_\epsilon} |\psi_n'(\lambda)| \le 2$. Combining the above results, for $X \in E_n$,

$$\left| \int \psi_n(\lambda) \mu_n^X(d\lambda) - \int \psi_n(\lambda) \mu_{n,p}(d\lambda) \right| \le \sup_{\lambda \in A_\epsilon} |\psi_n'(\lambda)| \cdot \sup_\lambda |F_n^X(\lambda) - F_{n,p}(\lambda)|$$
$$\le 2n^{-1} \log^6(n) \log^6(\log(n)).$$

Plugging it into (4.31) gives $|\log(\widetilde{R}_n)| \le L_n \delta r + C\delta^2 r^2 \le L_n \delta r$, where $L_n$ is a multi-log$(n)$ term. This implies that for $X \in E_n$, $|\widetilde{R}_n - 1| \le \max\{|e^{L_n \delta r} - 1|, |e^{-L_n \delta r} - 1|\} \le L_n \delta r$. As a result,

$$(4.33) \qquad E_0(|\widetilde{R}_n - 1| \cdot I_{E_n}) \le L_n \delta r = L_n n^{1-\alpha-\beta}.$$

Combining (4.32)-(4.33) gives (4.30). $\qquad \square$

4.4. *Proof of Corollary 1.1.* The first claim follows immediately from the null behavior of $\lambda_1$. We only prove the second claim, that is, if $\alpha > 2/3$, the power of the TW test tends to 0.

We first show that, if we fix all other parameters in the RWS (1.8) but increase $r$, the power of the TW test always increases. For $r_1, r_2$ such that $r_2 > r_1$, let $Q_1$ and $Q_2$ be uniformly distributed over $\mathbb{S}(p, r_1)$ and $\mathbb{S}(p, r_2)$, respectively, and write $\Sigma_1 = I_p + \frac{\delta}{1-\delta} Q_1 Q_1'$ and $\Sigma_2 = I_p + \frac{\delta}{1-\delta} Q_2 Q_2'$. Consider the two RWS models $X_i | Q_1 \overset{iid}{\sim} N(0, \Sigma_1)$ and $X_i | Q_2 \overset{iid}{\sim} N(0, \Sigma_2)$. Write $Q_2 = [\tilde{Q}_2, R]$, where $\tilde{Q}_2$ is the submatrix of $Q_2$ formed by its first $r_1$ columns. It is known that $\tilde{Q}_2$ is uniformly distributed over $\mathbb{S}(p, r_1)$. As a result,

$$\Sigma_1 \overset{(d)}{=} I_p + \frac{\delta}{1-\delta} \tilde{Q}_2 \tilde{Q}_2', \qquad \Sigma_2 = I_p + \frac{\delta}{1-\delta} (\tilde{Q}_2 \tilde{Q}_2' + RR').$$

For each realization of $Q_2 = [\tilde{Q}_2, R]$, $\Sigma_2 - \Sigma_1$ is positive semi-definite. So, for any value $a > 0$, the probability that $\lambda_1 > a$ under the first RWS model is no larger than the probability that $\lambda_1 > a$ under the second RWS model. This implies that the power of the TW test is larger in the second model.

Then, it suffices to consider the case that $r = n$. In this case, $\Sigma = \frac{1}{1-\delta} I_p$ under the alternative. By [32], if $X_i \overset{iid}{\sim} N(0, I_p)$, then $\frac{n^{2/3}(\lambda_1 - a_{n,p}^+)}{\sigma_{n,p}}$ converges weakly to the Tracy-Widom distribution, where $a_{n,p}^+ = (1 + \sqrt{p/n})^2$ and $\sigma_{n,p} = (1 + \sqrt{p/n})(1 + \sqrt{n/p})^{1/3}$, both converging to a constant as $n \to \infty$. Therefore, when $\Sigma = \frac{1}{1-\delta} I_p$,

$$W \equiv \frac{n^{2/3}[(1-\delta)\lambda_1 - a_{n,p}^+]}{\sigma_{n,p}} \text{ converges weakly to the TW law.}$$

The rejection region is that $\lambda_1 > a_{n,p}^+ + n^{-2/3}\sigma_{n,p}[3\log(n)]^{2/3}$. Hence, under the alternative, the rejection probability is equal to

$$P\left(W > -(n^{2/3}\delta) \cdot \sigma_{n,p}^{-1} a_{n,p}^+ + (1-\delta)[3\log(n)]^{2/3}\right).$$

Since $\alpha > 2/3$, the term $n^{2/3}\delta = n^{2/3-\alpha}$ is dominated by the term $[\log(n)]^{2/3}$, so the above probability tends to 0. □

4.5. *Proof of Lemma 1.1.* Let $A = \widetilde{\Sigma} - \Sigma$ and $Z_i \overset{iid}{\sim} N(0, A)$, $1 \leq i \leq n$, independent of $X_i$'s. Write $X = [X_1, \cdots, X_n]'$ and $Z = [Z_1, \cdots, Z_n]'$. It is seen that for $s \in \{0, 1\}$, if $X$ follows the model in $H_s^{(n)}$, then $\widetilde{X} = X + Z$ follows the model in $\widetilde{H}_s^{(n)}$. Therefore, letting $h(Z)$ be the joint density of $Z$, we have

$$\tilde{f}_s(\widetilde{X}) = \int f_s(\widetilde{X} - Z) h(Z) dZ, \qquad s = 0, 1.$$

It follows that

$$\|\tilde{f}_0 - \tilde{f}_1\| = \int \left| \int f_0(\widetilde{X} - Z) h(Z) dZ - \int f_0(\widetilde{X} - Z) h(Z) dZ \right| d\widetilde{X}$$

$$\leq \int \int |f_0(\widetilde{X} - Z) - f_1(\widetilde{X} - Z)| d\widetilde{X} h(Z) dZ$$

$$= \int \|f_0 - f_1\| h(Z) dZ = \|f_0 - f_1\|,$$

where the third line is due to a variable change $X = \widetilde{X} - Z$. □

4.6. *Proof of Lemma 1.2.* We prove the following lemma, and Lemma 1.2 is a special case of it.

**Lemma 4.4** (*Monotonicity of $L^1$-distance*). *Let $\Sigma_0(Q)$, $\Sigma_1(Q)$, and $\Sigma_2(Q)$ be three covariance matrix models indexed by a random quantity $Q$ such that*

- *For any non-random orthogonal matrix $U$, $U'\Sigma_\ell(Q)U$ follows the same distribution as $\Sigma_\ell(Q)$, $\ell = 0, 1, 2$.*
- *Given any realization $Q$, $\Sigma_1(Q) \preceq \Sigma_2(Q)$.*
- *Given any two realizations $Q$ and $\widetilde{Q}$, $\Sigma_0(Q) \preceq \Sigma_1(\widetilde{Q})$.*

*For each $\ell \in \{0, 1, 2\}$, consider a model where $(X_i|Q)$ are iid samples from $N(0, \Sigma_\ell(Q))$, $1 \leq i \leq n$, and let $f_\ell$ be the joint density of $\{X_i\}_{i=1}^n$. We have $\|f_1 - f_0\|_1 \leq \|f_2 - f_0\|_1$.*

*Proof.* Consider the problem of testing $H_1 : X_i|Q \overset{iid}{\sim} N(0, \Sigma_1(Q))$ against $H_0 : X_i|Q \overset{iid}{\sim} N(0, \Sigma_0(Q))$. By Neyman-Pearson lemma, $1 - (1/2)\|f_1 - f_0\|_1$ is a lower bound for the sum of type I and type II errors of any test, and this lower bound is achieved by the likelihood ratio test which rejects $H_0$ if and only if

$$(4.34) \qquad T(X) > 1, \qquad \text{where} \quad T(X) \equiv f_1(X)/f_0(X).$$

We then use this test for testing $H_2 : X_i|Q \overset{iid}{\sim} N(0, \Sigma_2(Q))$ against $H_0$. The type I error remains the same as before. Let $P_\ell$ be the probability measure associated with the model $H_\ell$, $\ell = 1, 2$. If we can show that the type II error does not increase, i.e.,

$$(4.35) \qquad P_2(T(X) \leq 1) \leq P_1(T(X) \leq 1),$$

then the sum of type I and type II errors does not increase either, i.e., it is no larger than $1 - (1/2)\|f_1 - f_0\|_1$. On the other hand, by Neyman-Pearson lemma again, with the test (4.34) for testing $H_2$ against $H_0$, the sum of type I and type II errors is no smaller than $1 - (1/2)\|f_2 - f_0\|_1$. This proves that $\|f_1 - f_0\|_1 \leq \|f_2 - f_0\|_1$.

It remains to show (4.35). Without loss of generality, we assume $\Sigma_0(Q)$ is non-degenerate almost surely. Write $\hat{\Sigma} = (1/n)\sum_{i=1}^n X_i X_i'$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be its eigenvalues. We first show that $T(X)$ depends on $X$ only through $\lambda_1, \cdots, \lambda_p$, i.e.,

$$(4.36) \qquad T(X) = T(\lambda_1, \cdots, \lambda_p).$$

Let $\hat{\Sigma} = U\Lambda U'$ be the eigen-decomposition of $\hat{\Sigma}$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$ and $U$ is an orthogonal matrix. For $\ell = 0, 1, 2$, write $\Omega_\ell(Q) = [\Sigma_\ell(Q)]^{-1}$ and $\tilde{\Omega}_\ell(Q) = U'\Omega_\ell(Q)U$. By our assumption, for any $U$, $\Omega_\ell(Q)$ and $\tilde{\Omega}_\ell(Q)$ have

the same distribution. It follows that

$$
\begin{aligned}
f_\ell(X) &= \frac{1}{(2\pi)^{np/2}} \int \exp\Big( -\frac{n}{2}\mathrm{tr}(U\Lambda U'\Omega_\ell(Q)) + \frac{n}{2}\log(|\Omega_\ell(Q)|)\Big) dF(Q) \\
&= \frac{1}{(2\pi)^{np/2}} \int \exp\Big( -\frac{n}{2}\mathrm{tr}(\Lambda\tilde\Omega_\ell(Q)) + \frac{n}{2}\log(|\tilde\Omega_\ell(Q)|)\Big) dF(Q) \\
&= \frac{1}{(2\pi)^{np/2}} \int \exp\Big( -\frac{n}{2}\mathrm{tr}(\Lambda\Omega_\ell(Q)) + \frac{n}{2}\log(|\Omega_\ell(Q)|)\Big) dF(Q),
\end{aligned}
$$

where the second equality is because $|\Omega_\ell(Q)| = |U\tilde\Omega_\ell(Q)U'| = |\tilde\Omega_\ell(Q)|$. This implies that $f_\ell(X)$ depends on $X$ only through $\Lambda$, i.e., $f_\ell(X) = f_\ell(\lambda_1, \cdots, \lambda_p)$, for $\ell = 0, 1, 2$. Combining it with the definition of $T(X)$ gives (4.36).

From the above expression of $f_\ell(X)$, we see that $\frac{\partial}{\partial \lambda_j} f_\ell$ exists for $\ell = 0, 1, 2$, so $\frac{\partial}{\partial \lambda_j} T$ is well defined. We next show that

(4.37) $$\frac{\partial T}{\partial \lambda_j} \geq 0, \qquad \text{for all } 1 \leq j \leq p.$$

Fix $1 \leq j \leq p$. We note that $\frac{\partial T}{\partial \lambda_j}$ has the same sign as $f_0 \frac{\partial f_1}{\partial \lambda_j} - f_1 \frac{\partial f_0}{\partial \lambda_j}$. Define $h(B; \Lambda) = (2\pi)^{-np/2} e^{-\frac{n}{2}\mathrm{tr}(\Lambda B) + \frac{n}{2}\log(|B|)}$ . Then,

$$
\begin{aligned}
f_\ell(\lambda_1, \cdots, \lambda_p) &= E_Q\big[h\big(\Omega_\ell(Q); \Lambda\big)\big], \\
\frac{\partial f_\ell}{\partial \lambda_j}(\lambda_1, \cdots, \lambda_p) &= -\frac{n}{2} E_Q\big[\Omega_\ell^{jj}(Q) \cdot h\big(\Omega_\ell(Q); \Lambda\big)\big],
\end{aligned}
$$

where $\Omega_\ell^{jj}(Q)$ is the $j$-th diagonal of $\Omega_\ell(Q)$ and $E_Q[\cdot]$ denotes the expectation with respect to the randomness of $Q$ only. Let $Q$ and $\widetilde Q$ be two independent copies of $Q$. Write $h(B; \Lambda) = h(B)$ for short. It is seen that

$$
\begin{aligned}
f_0 \frac{\partial f_1}{\partial \lambda_j} &= E_Q\big[h(\Omega_0(Q))\big] \cdot -\frac{n}{2} E_{\tilde Q}\big[\Omega_1^{jj}(\widetilde Q) \cdot h(\Omega_1(\widetilde Q))\big] \\
&= -\frac{n}{2} E_{Q,\tilde Q}\big[\Omega_1^{jj}(\widetilde Q) \cdot h(\Omega_1(\widetilde Q)) \cdot h(\Omega_0(Q))\big] \\
&\geq -\frac{n}{2} E_{Q,\tilde Q}\big[\Omega_0^{jj}(Q) \cdot h(\Omega_1(\widetilde Q)) \cdot h(\Omega_0(Q))\big] \\
&= E_{\tilde Q}\big[h(\Omega_1(\widetilde Q))\big] \cdot -\frac{n}{2} E_Q\big[\Omega_0^{jj}(Q) \cdot h(\Omega_0(Q))\big] = f_1 \frac{\partial f_0}{\partial \lambda_j},
\end{aligned}
$$

where the inequality is due to the assumption that $\Sigma_1(\tilde Q) \succeq \Sigma_0(Q)$ for any realized $Q, \tilde Q$. This proves (4.37).

We now show (4.35). For $\ell = 0, 1, 2$, let $\lambda_k^{(\ell)}(Q)$ be the $k$-th leading eigenvalue of $\hat\Sigma$ under the model $H_\ell$. By our assumption, $\Sigma_1(Q) \preceq \Sigma_2(Q)$ for any

realized $Q$, which implies $\lambda_k^{(1)}(Q) \le \lambda_k^{(2)}(Q)$, $1 \le k \le p$. Combining this with (4.36)-(4.37), we have $T(\lambda_1^{(1)}(Q), \cdots, \lambda_p^{(1)}(Q)) \le T(\lambda_1^{(2)}(Q), \cdots, \lambda_p^{(2)}(Q))$. It follows that

$$P_2(T(X) > 1|Q) \ge P_1(T(X) > 1|Q), \qquad \text{for any realized } Q.$$

Taking expectation with respect to $Q$ on both sides gives (4.35). $\qquad\square$

4.7. *Proof of Theorem 1.6.* By Lemmas 4.1, 4.2 and 4.3,

$$\log(LR_n) = \log(\widetilde{R}_n) + o_P(1)$$
$$= -\frac{nr^2\delta^2}{4p(1-\delta)^2} - \frac{nr\delta}{2}\left[\int \psi_n(\lambda)\mu_n^X(d\lambda) - \int \psi_n(\lambda)\mu_{n,p}(d\lambda)\right] + o_P(1),$$

where $\psi_n(\lambda) = \frac{p}{n\delta}\log(1 - \frac{\delta n}{p}(\lambda - \frac{1}{1-\delta}))$. With high probability, the support of $\mu_n^X$ and the support of $\mu_{n,p}$ are both strictly contained in a bounded interval $[(1-\sqrt{\gamma})^2 - 0.1, (1+\sqrt{\gamma})^2 + 0.1]$. Write $\gamma_n = p/n$. Since $\delta = o(1)$, the Taylor expansion yields $\psi_n(\lambda) = \frac{\gamma_n}{\delta}[-\frac{\delta}{\gamma_n}(\lambda - \frac{1}{1-\delta}) + O(\delta^2)] = -(\lambda - \frac{1}{1-\delta}) + O(\delta)$. Moreover, since $\int \mu_n^X(d\lambda) = \int \mu_{n,p}(d\lambda) = 1$, the term $\frac{1}{1-\delta}$ does not affect the integral. As a result,

$$\log(LR_n) = -\frac{nr^2\delta^2}{4p(1-\delta)} + \frac{nr\delta}{2}\Big[\int \lambda\mu_n^X(d\lambda) - \int \lambda\mu_{n,p}(d\lambda)\Big] + o_P(1)$$
$$= -\frac{(r\delta)^2}{4\gamma_n} + \frac{(r\delta)}{2\gamma_n}[S_n - p] + o_P(1),$$

where $S_n$ is the trace statistic, and the second line is because $\int \lambda\mu_n^X(d\lambda) = p^{-1}S_n$ and $\int \lambda\mu_{n,p}(d\lambda) = 1$. In other words,

$$\log(LR_n) = -\frac{\theta_n^2}{2} - \theta_n \cdot \left(\frac{S_n - p}{\sqrt{2\gamma_n}}\right) + o_P(1), \qquad \text{where } \theta_n \equiv \frac{r\delta}{\sqrt{2\gamma_n}} \to \theta.$$

Note that $\frac{S_n - p}{\sqrt{2\gamma_n}} = \frac{(nS_n - np)}{\sqrt{2np}}$. By elementary statistics, this term converges to $N(0,1)$ under the null, and converges to $N(\frac{nr\delta}{\sqrt{2np}}, 1) = N(\theta_n, 1)$ under the alternative. The claim then follows. $\qquad\square$

4.8. *Proof of Lemma 4.1.* Let $Y \in \mathbb{R}^{p,r}$ be a matrix such that $Y(i,j) \overset{iid}{\sim} N(0,1)$. It is known that when $Q$ is uniformly distributed over the Stiefel manifold $\mathbb{S}(p,r)$, $QQ'$ has the same distribution as $Y(Y'Y)^{-1}Y'$. So, the RWS model (1.8) is equivalent to

$$(4.38) \qquad H_1^{(n)}: \quad \Sigma = [I_p - \delta Y(Y'Y)Y']^{-1}.$$

Denote by $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_r > 0$ the nonzero eigenvalues of $(1/p)YY'$. Let $B_n = \{Y : \max_{1 \leq k \leq r} |\eta_k - 1| \leq 2\sqrt{r/p}\}$.

We first show that $Y \in B_n^c$ has a negligible effect on the likelihood ratio. Denote by $\overline{\Sigma}(Y)$ a general covariance matrix model indexed by $Y$. Introduce

$$(4.39) \qquad \overline{\Sigma}^*(Y) = \overline{\Sigma}(Y) \cdot 1\{Y \in B_n\} + I_p \cdot 1\{Y \in B_n^c\}.$$

Consider the model $X_i|Y \overset{iid}{\sim} N(0, \overline{\Sigma}(Y))$ and let $\overline{LR}_n$ be the corresponding likelihood ratio with respect to $H_0^{(n)} : X_i \overset{iid}{\sim} N(0, I_p)$. Define $\overline{LR}_n^*$ similarly by replacing $\overline{\Sigma}(Y)$ with $\overline{\Sigma}^*(Y)$. Below, we show that

$$(4.40) \qquad E_0|\overline{LR}_n - \overline{LR}_n^*| \to 0.$$

To see this, let $f_0(X)$ be the joint density of $X$ under $H_0^{(n)}$, $f(Y)$ be the joint density of $Y$ and $f_1(X|Y)$ be the conditional density of $X$ under $X_i|Y \overset{iid}{\sim} N(0, \overline{\Sigma}(Y))$. Then, $\overline{LR}_n = \int \frac{f_1(X|Y)f(Y)}{f_0(X)}dY$ and $\overline{LR}_n^* = \int_{B_n} \frac{f_1(X|Y)f(Y)}{f_0(X)}dY + \int_{B_n^c} f(Y)dY$. So $E_0|LR_n - LR_n^*| \leq \int f_0(X) \int_{B_n^c} \left[\frac{f_1(X|Y)f(Y)}{f_0(X)} + f(Y)\right]dYdX = \int_{B_n^c}[\int f_1(X|Y)dX]f(Y)dY + \int_{B_n^c}[\int f_0(X)dX]f(Y)dY = 2\int_{B_n^c} f(Y)dY = 2 \cdot P(Y \in B_n^c)$. By [50, Corollary 5.35], $P(Y \in B_n^c) \leq 2e^{-r}$. So (4.40) follows.

Regarding (4.40), for all the covariance models considered here, switching between $Y$ and $Y \cdot 1\{Y \in B_n\}$ only affects the $L^1$-distance by an $o(1)$ term, and we omit such a difference for notation simplicity.

Introduce $\delta^{\pm} = \delta/(1 \mp 2\sqrt{p/n})$ and let $\widetilde{LR}_n^{\pm}$ be the likelihood ratio associated with the Gaussian proxy model with $\delta = \delta^{\pm}$. For $Y \in B_n$,

$$(4.41) \qquad \delta^-(p^{-1}YY') \preceq \delta[Y(Y'Y)^{-1}Y'] \preceq \delta^+(p^{-1}YY').$$

We apply Lemma 4.4 with $\Sigma_0(Y) = I_p$, $\Sigma_1(Y) = [I_p - \delta Y(Y'Y)^{-1}Y']^{-1}$ and $\Sigma_2(Y) = [I_p - \delta^+ p^{-1}YY']^{-1}$.[9] It yields that

$$E_0|1 - \widetilde{LR}_n^-| \leq E_0|1 - LR_n| \leq E_0|1 - \widetilde{LR}_n^+|.$$

Similarly, since $\delta^-(p^{-1}YY') \preceq \delta(p^{-1}YY') \preceq \delta^+(p^{-1}YY')$,

$$E_0|1 - \widetilde{LR}_n^-| \leq E_0|1 - \widetilde{LR}_n| \leq E_0|1 - \widetilde{LR}_n^+|.$$

---

[9]When applying Lemma 4.4, we consider $Y \cdot 1\{Y \in B_n\}$ in all the covariance models. So, we need to show that for any nonradom orthogonal matrix $U \in \mathbb{R}^{p,p}$, $UY \cdot 1\{UY \in B_n\}$ has the same distribution as $Y \cdot 1\{Y \in B_n\}$. We note that $Y \overset{(d)}{=} P_1 \Gamma P_2$, where $P_1$ and $P_2$ are uniformly distributed over $\mathbb{S}(p,r)$ and $\mathbb{S}(r,r)$, respectively, and $\Gamma \in \mathbb{R}^{r,r}$ is a random diagonal matrix, and $(P_1, \Gamma, P_2)$ are independent. So $UY \cdot 1\{UY \in B_n\} \overset{(d)}{=} UP_1 \Gamma P_2 \cdot 1\{\|\Gamma - I_r\| \leq 2\sqrt{r/p}\}$. The claim follows by noting that $UP_1$ has the same distribution as $P_1$.

As a result,

$$|E_0|1-\widetilde{LR_n}|-E_0|1-LR_n|| \le E_0|1-\widetilde{LR}_n^+|-E_0|1-\widetilde{LR}_n^-| \le E_0|\widetilde{LR}_n^- - \widetilde{LR}_n^+|.$$

To show the claim, it suffices to show that

$$(4.42) \qquad\qquad E_0|\widetilde{LR}_n^- - \widetilde{LR}_n^+| \to 0.$$

We now show (4.42). Denote by $\mu_n^X$ the empirical spectral measure of $\hat{\Sigma}$, and by $\mu_{n,p}$ the Marchenko-Pastur (MP) measure with parameter $\gamma = p/n$ (see Section 3). Let $F_n^X$ and $F_{n,p}$ be the corresponding CDF's. Introduce $E_n = \{X \in \mathbb{R}^{n,p} : \sup_{t\in\mathbb{R}} |F_n^X(t) - F_{n,p}(t)| \le n^{-1}\log^6(n)\log^6(\log(n))\}$. By Theorem 3.2, $P_0(X \in E_n^c) = O(e^{-c\log(n)\log(\log(n))})$, as $n \to \infty$. Note that

$$E_0\big(|\widetilde{LR}_n^- - \widetilde{LR}_n^+| \cdot I_{E_n^c}\big) \le E_0(\widetilde{LR}_n^- \cdot I_{E_n^c}) + E_0(\widetilde{LR}_n^+ \cdot I_{E_n^c}).$$

Since $E_0(\widetilde{LR}^\pm) = 1$ is finite, $P_0(E_n^c) = o(1)$ implies $E_0(\widetilde{LR}_n^- \cdot I_{E_n^c}) = o(1)$. Therefore, to show (4.42), it suffices to show that

$$(4.43) \qquad\qquad E_0\big(|\widetilde{LR}_n^- - \widetilde{LR}_n^+| \cdot I_{E_n}\big) \to 0.$$

Below, we show (4.43). The assumption on $(\alpha, \beta)$ ensures that $e_n \le n^{-\epsilon}$ for some $\epsilon > 0$, where

$$e_n \equiv \delta^2 r^{5/2} n^{-1/2} + \delta r^{3/2} n^{-1/2} + \delta^2 r.^{[10]}$$

In Section 4.2, we define $\widetilde{R}_n$, a proxy of $\widetilde{LR}_n$. Let $\widetilde{R}_n^\pm$ be the proxy associated with $\delta = \delta^\pm$. By Lemma 4.2, $E_0|\widetilde{LR}_n^\pm - \widetilde{R}_n^\pm| = o(1)$. So it suffices to show

$$(4.44) \qquad\qquad E_0\big(|\widetilde{R}_n^- - \widetilde{R}_n^+| \cdot I_{E_n}\big) \to 0.$$

Let $g_n^\pm(\lambda) = \frac{p}{n}\log(1 - \frac{\delta^\pm n}{p}(\lambda - \frac{1}{1-\delta^\pm}))$. By Lemma 4.3,

$$\log(\widetilde{R}_n^\pm) = -\frac{nr^2(\delta^\pm)^2}{4p(1-\delta^\pm)^2} - \frac{nr}{2}\left[\int g_n^\pm(\lambda)\mu_n^X(d\lambda) - \int g_n^\pm(\lambda)\mu_{n,p}(d\lambda)\right].$$

Let $\Delta g_n = g_n^+ - g_n^-$. It follows that

$$\log\big(\widetilde{R}_n^-/\widetilde{R}_n^+\big) = \frac{nr}{2}\left[\int \Delta g_n(\lambda)\mu_n^X(d\lambda) - \int \Delta g_n(\lambda)\mu_{n,p}(d\lambda)\right]$$
$$+ \frac{nr^2(\delta^+)^2}{4p(1-\delta^+)^2} - \frac{nr^2(\delta^-)^2}{4p(1-\delta^-)^2} \equiv (I) + (II).$$

---

[10]The first two terms are from the calculation of $(I)$ and $(II)$ below, and the last term is due to $E_0|\widetilde{LR}_n - \widetilde{R}_n| \le C(\delta^2 r + \delta^3 r^{5/2} n^{-1/2})$, as seen in the proof of Lemma 4.2.

On the event $E_n$,

$$|(I)| \leq \frac{nr}{2} \cdot \sup_{\lambda} |(\Delta g_n)'(\lambda)| \cdot \sup_{\lambda \in \mathbb{R}} |F_n^X(\lambda) - F_{n,p}(\lambda)| \leq L_n r \sup_{\lambda} |(\Delta g_n)'(\lambda)|,$$

where $L_n = \log^6(n) \log^6(\log(n))/2$. For $X \in E_n$, the supports of $\mu_{n,p}$ and $\mu_n^X$ are both strict subsets of $A = [\frac{1}{2}(1 - \sqrt{\gamma})^2, 2(1 + \sqrt{\gamma})^2]$. So it suffices to consider $\sup_{\lambda \in A} |(\Delta g_n)'(\lambda)|$. By direct calculations, $(\Delta g_p)'(\lambda) = -(\delta^+ - \delta^-) \cdot [1 - \frac{n\delta^+\delta^-}{p(1-\delta^+)(1-\delta^-)}][1 - \frac{\delta^+ n}{p}(\lambda - \frac{1}{1-\delta^+})]^{-1}[1 - \frac{\delta^- n}{p}(\lambda - \frac{1}{1-\delta^-})]^{-1} = -(\delta^+ - \delta^-)[1 + o(1)]$, for $\lambda \in A$. It follows that

$$\sup_{\lambda \in A} |(\Delta g_n)'(\lambda)| \leq C(\delta^+ - \delta^-) \leq C\delta\sqrt{r/p}.$$

Combining the above gives

(4.45) $$|(I)| \leq CL_n \cdot \delta r^{3/2} n^{-1/2}.$$

By direct calculation, $(II) = \frac{nr^2}{4p} \frac{(\delta^+ - \delta^-)(\delta^+ + \delta^- - 2\delta^+\delta^-)}{(1-\delta^+)^2(1-\delta^-)^2}$. It follows that

(4.46) $$|(II)| \leq Cr^2(\delta^+ - \delta^-)\delta \leq C\delta^2 r^{5/2} n^{-1/2}.$$

Both the right hands of (4.45)-(4.46) are $O(e_n)$. As a result, for $X \in E_n$,

$$|\widetilde{R}_n^- / \widetilde{R}_n^+ - 1| \leq \max\{|e^{Ce_n} - 1|, |e^{-Ce_n} - 1|\} \leq Ce_n.$$

So $E_0(|\widetilde{LR_n^-} - \widetilde{LR_n^+}| \cdot I_{E_n}) \leq Ce_n E_0(\widetilde{R}_n^+)$. Since $E_0|\widetilde{R}_n^+ - \widetilde{LR_n^+}| = o(1)$ and $E_0(\widetilde{LR_n^+}) = 1$, $E_0(\widetilde{R}_n^+) \leq 2$. Combining these gives (4.44). $\qquad\square$

**Remark**. (*Conjecture on $E_0|LR_n - \widetilde{LR_n}|$*). We conjecture that a stronger result holds:

$$E_0|LR_n - \widetilde{LR_n}| \to 0.$$

To see this, let $\Sigma(Y)$ be the covariance matrix (4.38) of RWS, and let $\widetilde{\Sigma}(Y)$, $\widetilde{\Sigma}^\pm(Y)$ be the covariance matrices of proxy models associated with $\delta$, $\delta^\pm$, respectively. By (4.41), $\widetilde{\Sigma}^-(Y) \preceq \Sigma(Y) \preceq \widetilde{\Sigma}^+(Y)$, so the RWS is "sandwiched" by two proxy models. We conjecture that

(4.47) $$E_0|LR_n - \widetilde{LR_n^-}| \leq E_0|\widetilde{LR_n^+} - \widetilde{LR_n^-}|.$$

Similarly, since $\widetilde{\Sigma}^-(Y) \preceq \widetilde{\Sigma}(Y) \preceq \widetilde{\Sigma}^+(Y)$, we conjecture $E_0|\widetilde{LR_n} - \widetilde{LR_n^-}| \leq E_0|\widetilde{LR_n^+} - \widetilde{LR_n^-}|$. By the triangular inequality, $E_0|LR_n - \widetilde{LR_n}| \leq E_0|LR_n - \widetilde{LR_n^-}| + E_0|\widetilde{LR_n} - \widetilde{LR_n^-}| \leq 2E_0|\widetilde{LR_n^+} - \widetilde{LR_n^-}|$. It then follows from (4.42) that $E_0|LR_n - \widetilde{LR_n}| \to 0$.

To rigorously prove (4.47), we need a slightly stronger result than Lemma 4.4 (monotonicity of $L^1$-distance). In Lemma 4.4, one assumption is that $\Sigma_0(Q) \preceq \Sigma_1(\widetilde{Q})$ for any realizations $Q, \widetilde{Q}$. If this assumption is relaxed to requiring $\Sigma_0(Q) \preceq \Sigma_1(Q)$ for any $Q$, then we can apply the result to $(\Sigma_0, \Sigma_1, \Sigma_2) = (\widetilde{\Sigma}^-(Y), \Sigma(Y), \widetilde{\Sigma}^+(Y))$ and obtain (4.47).

**5. Discussions.** Testing of sphericity, i.e., whether the population covariance matrix $\Sigma$ is the identity, has been well-studied. While many tests were proposed for high-dimensional settings [2, 14, 32, 34], there are relatively few results about the phase transition: [12] studied the testing limit for a class of alternatives that are characterized by the Frobenius norm of $\Sigma - I_p$, and [9] investigated a class of alternatives which are rank-1 sparse perturbations of the null. We are interested in a different class of alternatives, the Rare and Weak Spike model. When the number of spikes is finite, [37, 38] studied the asymptotic power envelopes of eigenvalue-based tests, and their results implicitly gave the phase transition of this testing problem. Compared with [37, 38], our results cover broader situations where the number of spikes may tend to infinity as $n, p \to \infty$. We also extend our results to more complicated settings where $\Sigma$ is not necessarily the identity in the null, by exploring the idea of "comparison of experiments."

Various tests were proposed for testing $\Sigma = I_p$; some are based on extreme eigenvalues [32, 33], and some are based on bulk eigenvalues [2, 18, 34, 37]. It has been recognized that the extreme-eigenvalue tests are more powerful for testing Rare and Strong spikes, while for Rare and Weak spikes, it is better to use more than just a few top eigenvalues. In the spirit of "letting the data decide which eigenvalues to use", we proposed two new tests, the Higher Criticism (HC) test, and the Cumulative Sum (CuSum) test. Both tests can be viewed as extensions of the Higher Criticism test for testing sparse normal means [19]. In contrast to bulk-eigenvalue tests [18, 37], our proposed tests are purely adaptive. Moreover, we find that the CuSum test is both theoretically optimal and has impressive numerical performance.

Technically, we introduce a Gaussian proxy model for studying the RWS model. We show that the two models are close enough for a wide range of parameters. At the same time, the likelihood of the proxy model is easier to analyze, where we can use the available results on Wishart matrices instead of analyzing any spherical integrals. Such an approach can be useful for other problems related to the RWS model.

In Random Matrix Theory, it is an interesting topic to study the empirical eigenvalues of a spiked covariance model. In the seminal work [4], they derived the limiting joint distribution of top $k$ eigenvalues of a complex

Gaussian Wishart matrix, with the population covariance a rank-$r$ perturbation of the identity, where $r$ is finite and $k \leq r$. Later, [5, 40] proved similar results for real covariance matrices. However, there is little understanding on the case that $r$ tends to infinity as $n, p$ grows. Moreover, existing works mainly focus on a few top eigenvalues. The bulk eigenvalues of a spiked model are also of great interest in many applications, but such results are only found for the identify covariance case [16, 42, 45]. Our results fill in the gap by both allowing $r \to \infty$ (in a algebraic rate) and concerning all eigenvalues.

A somewhat relevant setting to the spiked covariance model is the spiked Wigner model, where we consider an $n \times n$ Wigner matrix plus a non-random, rank-$r$ diagonal matrix. Such a setting with $r \to \infty$ was studied in [41], and they derived the limiting behavior of top eigenvalues. Besides that the two settings are very different, [41] assumes the magnitude of entries in the diagonal perturbation is independent of $n$, so it is not about rare and "weak" spikes.

Our work also brings a new angle of using statistical theory to shed lights on random matrix theory. We recognize that, for parameters where all tests are asymptotically powerless, any bounds of empirical eigenvalues that hold in the null should also hold in the alternative. Hence, as long as we know the statistical limits, we can take advantage of results in the identity covariance case for studying the spiked covariance models. Alternatively, if we wish to derive such results directly from random matrix theory, we have to pay non-trivial efforts, noting that how to extend [4, 5, 40] to the setting of rare and weak spikes is largely unclear.

It is possible to extend our results to nonGaussian settings: $X = Y\Sigma^{1/2}$, where $Y$ is an $n \times p$ random matrix with *iid* entries of mean 0 and variance 1. To test $\Sigma = I_p$, we can still apply the four tests considered in this paper, and their performance will be similar to the Gaussian case, due to universality of eigenvalue statistics (e.g., [42]). The lower bound arguments may be slightly different, depending on the distribution of the entries of $Y$. However, the key technique of this paper continues to work: we can approximate the RWS model (on $\Sigma$) by a Gaussian proxy model, to overcome the hurdle of analyzing likelihood ratios. We leave this for future work.

Another extension of our work is to consider the same testing problem for population *correlation* matrices. For this setting, it is natural to apply the tests to eigenvalues of the sample correlation matrix. The existing results about the null behavior of these eigenvalues [7, 24] will be helpful. We also leave this for future study.

## 6. Appendix.

6.1. *Proof of Theorem 3.1.* Define the function $h : \mathbb{R}^{n,p} \to \mathbb{R}$ by $h(Z) = p\mu_n^Z(f)$, i.e., we view $p\mu_n^Z(f) = \sum_{j=1}^p f(\lambda_j)$ as a function of the entries of $Z$. For a differentiable $f$ such that $\|f'\|_\infty < \infty$, Delyon [17] shows that $\nabla h$ exists. According to the last equation on [17, Page 554], for each $1 \leq i \leq n$, $\sum_{1 \leq j \leq p} \left(\frac{\partial h(Z)}{\partial Z_{ij}}\right)^2 \leq 4n^{-2}\|f'\|_\infty^2 \sum_{1 \leq j \leq p} Z_{ij}^2$. It follows that

$$(6.48) \qquad \|\nabla h(Z)\|^2 \leq 4n^{-2}\|f'\|_\infty^2 \|Z\|_F^2.$$

Introduce

$$\tilde{h}(Z) = \frac{\sqrt{n \log(2)}}{2\sqrt{2p}\|f'\|_\infty} h(Z).$$

Since $\|Z\|_F^2$ has a $\chi_{np}^2(0)$ distribution, $E[e^{t\|Z\|_F^2}] = (1-2t)^{-np/2}$ for all $t \geq 0$. From (6.48), $\|\nabla \tilde{h}(Z)\|^2 \leq \frac{\log(2)}{2np}\|Z\|_F^2$. It follows that

$$(6.49) \qquad E[e^{\|\nabla \tilde{h}(Z)\|^2}] \leq \left[1 - (2np)^{-1}\log(2)\right]^{-np/2} \leq 2.$$

We now apply [11, Theorem 6.1], a concentration inequality for non-Lipschitz functions, to $\tilde{h}$. The probability measure $\mu$ there corresponds to the standard Gaussian measure of dimension $np$ in our settings, so the subGaussian constant $\sigma(\mu) = 1$. It yields that for $L_0$ such that $P(|\nabla \tilde{h}(Z)| > L_0) \leq 1/2$ and all $t > 0$,

$$P\big(|\tilde{h}(Z) - E[\tilde{h}(Z)]| > t\big) \leq 2 \inf_{y \leq L_0} \left\{ e^{-\frac{t^2}{cy^2}} + P(\|\nabla \tilde{h}(Z)\| > y) \right\},$$

where $c > 0$ is an absolute constant. By (6.49) and the Chebyshev inequality, $P(\|\nabla \tilde{h}(Z)\| > L) \leq 2e^{-L^2}$. Therefore, for any $y \geq L_0 \equiv \sqrt{2\log(2)}$,

$$(6.50) \qquad P\big(|\tilde{h}(Z) - E[\tilde{h}(Z)]| > t\big) \leq 2e^{-\frac{t^2}{cy^2}} + 4e^{-y^2}.$$

The claim follows by noticing that $\mu_n^Z(f) = \frac{2\sqrt{2}\|f'\|_\infty}{\sqrt{np\log(2)}}\tilde{h}(Z)$. □

6.2. *Proof of Lemma 3.1.* Fix $m \geq 1$. Write for short $H = \sum_{k=1}^p (\lambda_j - 1)^m$ and $e_n = \log(n)[(p \vee \log(n))/n]^{m/2}$. We aim to show that for some constant $C > 0$,

$$P(|H - E(H)| > Ce_n) = o(n^{-1}).$$

Note that $p \leq n$. By [50, Corollary 5.35], for every $t \geq 0$,

$$(6.51) \qquad P\left(\max_{1 \leq j \leq p} |\lambda_j - 1| > \sqrt{p/n} + t/\sqrt{n}\right) \leq 2e^{-t^2/2}.$$

Let $a_n = \sqrt{p/n} + 2\sqrt{\log(n)/n}$ and $B_n$ be the event $\max_{1 \leq j \leq p} |\lambda_j - 1| \leq a_n$. The above implies that $P(B_n^c) = o(n^{-1})$. So it suffices to show that

$$(6.52) \qquad P(|H - E(H)| > Ce_n, \ B_n) = o(n^{-1}).$$

Define a function $f_n(x) = h_n(x - 1)$, where

$$h_n(x) = \begin{cases} x^m, & \text{if } 0 \leq x \leq a_n, \\ (2a_n)^m, & \text{if } x \geq 2a_n, \\ a_n^m + m \int_0^{x-a_n} (2a_n - x)^{m-1}, & \text{if } a_n < x < 2a_n, \\ (-1)^m h_n(-x), & \text{if } x < 0. \end{cases}$$

Introduce $H^* = \sum_{j=1}^p f_n(\lambda_j)$. Then, $H = H^*$ over the event $B_n$. Therefore, to show (6.52), it suffices to show that for some constants $C_1, C_2 > 0$,

$$(6.53) \qquad P(|H^* - E(H^*)| > C_1 e_n) = o(n^{-1}),$$

and

$$(6.54) \qquad |E(H^*) - E(H)| < C_2 e_n.$$

First, consider (6.53). We write $H^* = p\mu_n^Z(f_n)$, and note that $f_n$ is continuously differentiable and $\|f_n'\|_\infty = ma_n^{m-1}$. By Theorem 3.1 where we set $y = \sqrt{2}\log(n)$ and $t = 2\sqrt{c_0}\log(n)$, with probability at least $1 - o(n^{-1})$,

$$|H^* - E(H^*)| \leq \frac{2\sqrt{c_0 p}\log(n)}{\sqrt{n}} \|f_n'\|_\infty \leq Cma_n^{m-1}\sqrt{p/n}\log(n).$$

The right hand side is no larger than $Ce_n$. This proves (6.53).

Next, consider (6.54). Let $\xi = \max_{1 \leq j \leq p} |\lambda_j - 1|$. We have

$$(6.55) \qquad |E(H) - E(H^*)| \leq \sum_{j=1}^p E(|\lambda_j - 1|^m \cdot I_{B_n^c}) \leq pE(\xi^m \cdot I_{\{\xi > a_n\}}).$$

We use the simple fact that for any continuous random variable $W$ and real number $a$, $E(W \cdot I_{\{W > a\}}) = aP(W > a) + \int_a^\infty P(W > t)dt$. It follows that

$$E(\xi^m \cdot I_{\{\xi > a_n\}}) = a_n^m P(\xi > a_n) + \int_{a_n^m}^\infty P(\xi^m > x)dx$$

$$= o(n^{-1}a_n^m) + \frac{m}{\sqrt{n}} \int_{2\sqrt{\log(n)}}^\infty (\sqrt{p/n} + t/\sqrt{n})^{m-1} P(\xi > \sqrt{p/n} + t/\sqrt{n})dt$$

$$= o(n^{-1}a_n^m) + \frac{m}{(\sqrt{n})^m} \int_{2\sqrt{\log(n)}}^\infty 2^{m-1}[(\sqrt{p})^{m-1} + t^{m-1}]e^{-t^2/2}dt$$

$$\leq o(n^{-1}a_n^m) + \frac{2^{m-1}m}{(\sqrt{n})^m} \cdot [(\sqrt{p})^{m-1} + (2\sqrt{\log(p)})^{m-1}] \cdot Ce^{-2\log(n)}.$$

Here the second line is due to a variable change $t = \sqrt{n}(x^{1/m} - \sqrt{p/n})$, the third line follows from (6.51) and that $(a+b)^k \leq 2^k(a^k+b^k)$ for $a, b > 0$ and $k$ being a positive integer, and the last line is from elementary calculation. It is easy to see that the right hand side is no larger than $Cp^{-1}e_n$. Combining it with (6.55) gives (6.54). $\square$

6.3. *Proof of Lemma 3.2.* Write $L(t;\gamma) = \int \log(t - \lambda)dF_\gamma^{mp}(\lambda)$ for any $\gamma \geq 1$ and $t > (1 + \sqrt{\gamma})^2$. It is known that (e.g., equation (3.6) of [37]) the Hilbert transformation of $F_\gamma^{mp}$ is

$$H(t;\gamma) \equiv \int (t - \lambda)^{-1}dF_\gamma^{mp}(\lambda) = \frac{t + \gamma - 1 - \sqrt{(t - \gamma - 1)^2 - 4\gamma}}{2\gamma t},$$

for $t > (1 + \sqrt{\gamma})^2$; the sign of the square root is the same as that of $(t - \gamma - 1)$. Since $\frac{\partial}{\partial t} \log(t - \lambda) = (t - \lambda)^{-1}$, we have

$$(6.56) \qquad L(t;\gamma) = \int H(t;\gamma)dt + C_0,$$

where $\int H(t;\gamma)dt$ is any indefinite integral of $H(t;\gamma)$. We take the following choice (it can be verified by checking the derivative):

$$\int H(t;\gamma)dt = \frac{1}{2\gamma}\Big\{(A_t - B_t) - (\gamma + 1)\log(A_t - B_t)$$

$$(6.57) \qquad\qquad + (\gamma - 1)\log\big((\gamma + 1)A_t - (\gamma - 1)B_t + 4\gamma\big)\Big\},$$

where $A_t = t - \gamma - 1$ and $B_t = \sqrt{(t - \gamma - 1)^2 - 4\gamma}$ are the same as those in Lemma 3.2. To get the constant $C_0$ in (6.56), we note that $\lim_{t\to\infty}[L(t;\gamma) - \log(t)] = \lim_{t\to\infty} \int [\log(t - \lambda) - \log(t)]dF_\gamma^{mp}(\lambda) = 0$. As a result,

$$C_0 = -\lim_{t\to\infty}\big[\int H(t;\gamma)dt - \log(t)\big].$$

As $t \to \infty$, we have $A_t/t \to 1$ and $B_t/t \to 1$. Also, $A_t - B_t = 4\gamma/(A_t + B_t) \to 0$ and $\log(A_t - B_t) + \log(2t) \to \log(4\gamma)$. Moreover, $(2t)^{-1}[(\gamma + 1)A_t - (\gamma - 1)B_t + 4\gamma] = (2t)^{-1}[A_t + B_t + \gamma(A_t - B_t) + 4\gamma] \to 1$; it follows that $\log((\gamma + 1)A_t - (\gamma - 1)B_t + 4\gamma) - \log(2t) \to 0$. Plugging these results into (6.57), we find that

$$\lim_{t\to\infty}\Big[\int H(t;\gamma)dt - \frac{\gamma + 1}{2\gamma}\log(2t) - \frac{\gamma - 1}{2\gamma}\log(2t)\Big] \to -\frac{\gamma + 1}{2\gamma}\log(4\gamma).$$

It follows that

$$(6.58) \qquad C_0 = -\log(2) + \frac{\gamma + 1}{2\gamma}\log(4\gamma).$$

Combining (6.56)-(6.58) gives the explicit expression of $L(t;\gamma)$. $\square$

6.4. *Proof of Lemma 4.2.* Let $e_n = \delta^3 r^{5/2} n^{-1/2} + \delta^2 r$. It suffices to show that when $e_n = o(1)$, $E_0|\widetilde{LR}_n - \widetilde{R}_n| \le C e_n$.

Let $h(X, Y)$ be the integrand in (4.26). Define the events $A_n = \{Y : \delta\eta_1 < 1/2\}$ and $D_n = \{X : p^{-1}\delta n(\lambda_1 - \frac{1}{1-\delta}) < 1\}$. Introduce

$$\Delta_n(Y) = \Delta_n(Y, \delta, r) \equiv g_n^*(Y, \delta, r) - g_n(Y, \delta, r).$$

By (4.26)-(4.27), $\widetilde{LR}_n = \int_{A_n} h(X, Y) dF(Y) + \int_{A_n^c} dF(Y)$ and $\widetilde{R}_n = I_{D_n^c}(X) + I_{D_n}(X) \cdot \int h(X, Y) e^{\frac{n}{2}\Delta_n(Y)} dF(Y)$. Let $F_0(X)$ be the joint CDF of $X$ under $H_0^{(n)}$. Then,

$$
\begin{aligned}
E_0|\widetilde{LR}_n - \widetilde{R}_n| \le & \int_{A_n \times D_n} h(X, Y) |1 - e^{\frac{n}{2}\Delta_n(Y)}| dF(Y) dF_0(X) \\
& + \int_{A_n^c \times D_n} |1 - h(X, Y) e^{\frac{n}{2}\Delta_n(Y)}| dF(Y) dF_0(X) \\
& + \int_{A_n \times D_n^c} |h(X, Y) - 1| dF(Y) dF_0(X).
\end{aligned}
$$
(6.59)

By definition of the Gaussian model, conditioning on any $Y \in A_n$, $h(X, Y) dF_0(X)$ is the joint density of $X$ under $\widetilde{H}_1^{(n)}$. Consequently,

$$\int h(X, Y) dF_0(X) = 1, \qquad \text{for } Y \in A_n.$$
(6.60)

We plug (6.60) into (6.59), and note that $|1 - h(X, Y)| \le 1 + h(X, Y)$ and $|1 - h(X, Y) e^{\frac{n}{2}\Delta_n(Y, \delta, r)}| \le 1 + h(X, Y) e^{\frac{n}{2}\Delta_n(Y, \delta, r)}$. It follows that

$$
\begin{aligned}
E_0|\widetilde{LR}_n - \widetilde{R}_n| \le & \int_{A_n} |1 - e^{\frac{n}{2}\Delta_n(Y)}| dF(Y) \\
& + P_0(X \in D_n, Y \in A_n^c) + \int_{A_n^c \times D_n} h(X, Y) e^{\frac{n}{2}\Delta_n(Y)} dF_0(X) dF(Y) \\
& + P_0(X \in D_n^c, Y \in A_n) + \widetilde{P}_1(X \in D_n^c, Y \in A_n),
\end{aligned}
$$
(6.61)

where $P_0$ and $\widetilde{P}_1$ are the probability measures under $H_0^{(n)}$ and $\widetilde{H}_1^{(n)}$, respectively. By (6.51), $P_0(X \in D_n^c) = P_0(\lambda_1 - 1 > \frac{\delta}{1-\delta} + \frac{p}{\delta n}) \le 2e^{-\frac{n}{2}(\frac{\delta}{1-\delta} + \frac{p}{\delta n} - \sqrt{p/n})^2} \le 2e^{-p/2}$, where the last inequality follows from $\frac{\delta}{1-\delta} + \frac{p}{\delta n} \ge 2\sqrt{p/n}$ by Cauchy-Schwartz inequality. Again, by (6.51), $P_0(Y \in A_n^c) \le 2e^{-\frac{p}{2}[(2\delta)^{-1} - 1 - \sqrt{r/p}]^2} \le 2e^{-cn}$. Combining these with (6.61), to show the claim, it suffices to show

that

$$J_1 \equiv \int_{A_n} |1 - e^{\frac{n}{2}\Delta_n(Y)}| dF(Y) \leq Ce_n$$

$$J_2 \equiv \int_{A_n^c \times D_n} h(X,Y) e^{\frac{n}{2}\Delta_n(Y)} dF_0(X) dF(Y) = o(e_n)$$

(6.62)     $$J_3 \equiv \widetilde{P}_1(X \in D_n^c, Y \in A_n) = o(e_n).$$

Consider $J_2$. By (4.27), $h(X,Y)e^{\frac{n}{2}\Delta_n(Y)} = (1-\delta)^{\frac{nr}{2}} b_n(\delta,r) \exp\{\frac{\delta n}{2p}\mathrm{tr}((\hat{\Sigma} - \frac{1}{1-\delta})YY')$. We note that $(1-\delta)^{\frac{nr}{2}} \leq 1$ and $b_n(\delta,r) \leq e^{\frac{nr\delta}{2p(1-\delta)}}$. Moreover, since $YY'$ is positive semi-definite, $\mathrm{tr}((\hat{\Sigma} - \frac{1}{1-\delta})YY') \leq (\lambda_1 - \frac{1}{1-\delta})\mathrm{tr}(YY')$. For $X \in D_n$, $(\lambda_1 - \frac{1}{1-\delta}) < \delta^{-1}p/n$. As a result,

$$J_2 \leq e^{\frac{nr\delta}{2p(1-\delta)}} \int_{\delta\eta_1 > 1/2} \exp\{\frac{1}{2}\mathrm{tr}(YY')\} dF(Y).$$

Let $W = (2pr)^{-1/2}(\mathrm{tr}(YY') - pr)$. Note that $\eta_1 \geq (pr)^{-1}\mathrm{tr}(YY')$. So $W > \frac{1-\delta}{2\delta}\sqrt{(pr)/2}$ implies that $\delta\eta_1 > 1/2$. It follows that

$$J_2 \leq e^{\frac{nr\delta}{2p(1-\delta)} + \frac{pr}{2}} \int_{W > \frac{1-\delta}{2\delta}\sqrt{\frac{pr}{2}}} e^{\sqrt{\frac{pr}{2}}W} dF(W).$$

Since $\mathrm{tr}(YY')$ has a chi-square distribution, $W$ weakly converges to $N(0,1)$. So $e^{\sqrt{\frac{pr}{2}}W} dF(W) = e^{\frac{pr}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}(W - \sqrt{\frac{pr}{2}})^2}[1+o(1)]dW$, and the above integral is equal to $e^{pr/2}[1 + o(1)] \cdot P(N(0,1) > (\frac{1-\delta}{2\delta} - 1)\sqrt{pr/2})$. It follows that $J_2 \leq e^{\frac{nr\delta}{2p(1-\delta)} + \frac{pr}{2}} \cdot 2e^{-Cpr/\delta^2} \leq 2e^{-Cpr/(2\delta^2)} = o(e_n)$.

Consider $J_3$. First, $X \in D_n^c$ implies $\lambda_1 > \frac{p}{n\delta} + \frac{1}{1-\delta} > \frac{p}{2n\delta}$, noting that $\delta \to 0$. Second, under $\widetilde{H}_1^{(n)}$, $X_i|Y \overset{iid}{\sim} N(0, \widetilde{\Sigma}(Y))$, $1 \leq i \leq n$, where $[\widetilde{\Sigma}(Y)]^{-1} = I_p - \delta p^{-1}YY'$. Since for $Y \in A_n$, the maximum eigenvalue of $\widetilde{\Sigma}(Y)$ is $(1 - \delta\eta_1)^{-1} \leq 2$, we find that $\widetilde{P}_1(\lambda_1 > t) \leq P_0(2\lambda_1 > t)$ for any $t > 0$, where $P_0$ is the probability measure under $H_0^{(n)}$. Combining the above results and applying (6.51),

$$J_3 \leq P_0\left(2\lambda_1 > \frac{p}{2n\delta}\right) \leq 2e^{-\frac{n}{2}(\frac{p}{4n\delta} - 1 - \sqrt{\frac{p}{n}})^2} \leq 2e^{-C\delta^{-1}n} = o(e_n).$$

Last, consider $J_1$. Let $q(x) = -\log(1-x) - x - x^2/2$ and recall that $\eta_1 > \eta_2 > \cdots > \eta_r > 0$ are the nonzero eigenvalues of $(1/p)YY'$. Introduce

$$M_1(Y) = \sum_{k=1}^r q\left(\frac{\delta(\eta_k - 1)}{1 - \delta}\right), \quad M_2(Y) = \frac{\delta^2}{2(1-\delta)^2} \sum_{k=1}^r \{(\eta_k-1)^2 - E[(\eta_k-1)^2]\}.$$

Then, $\Delta_n(Y) = M_1(Y) + M_2(Y)$. Let $E^*[f(Y)] = \int_{Y \in A_n} f(Y) dF(Y)$ for any function $f$. We then have

$$J_1 \le E^* |1 - e^{\frac{n}{2} M_1(Y)}| + E^* \big[ e^{\frac{n}{2} M_1(Y)} |1 - e^{\frac{n}{2} M_2(Y)}| \big]$$

$$(6.63) \qquad \le E^* |1 - e^{\frac{n}{2} M_1(Y)}| + \big( E^* [e^{n M_1(Y)}] \cdot E^* |1 - e^{\frac{n}{2} M_2(Y)}|^2 \big)^{1/2}.$$

We claim that

$$(6.64) \qquad E^* |1 - e^{\frac{n}{2} M_1(Y)}| \le C e_n, \qquad E^* |1 - e^{n M_1(Y)}| \le C e_n,$$

and

$$(6.65) \qquad E^* |1 - e^{\frac{n}{2} M_2(Y)}| \le C e_n, \qquad E^* |1 - e^{n M_2(Y)}| \le C e_n.$$

Note that (6.65) implies $E^* |1 - e^{\frac{n}{2} M_2(Y)}|^2 \le C e_n$. Combining it with (6.63)-(6.64) gives $J_1 \le C e_n$.

It remains to show (6.64)-(6.65). We first prove (6.64). It suffices to consider the first inequality, as the second one is similar. Define $U = \frac{\delta}{1-\delta}(\eta_1 - 1)$ if $|\eta_1 - 1| \ge |\eta_r - 1|$, and $U = \frac{\delta}{1-\delta}(\eta_r - 1)$ if $|\eta_1 - 1| < |\eta_r - 1|$. Note that $U \le 1/2$ for $Y \in A_n$. Since $q(x) = \sum_{\ell=3}^{\infty} x^\ell / \ell$, we have $|M_1(Y)| \le r q(|U|)$. Therefore, to show (6.64), it suffices to show that

$$(6.66) \qquad E\big( |1 - e^{\frac{nr}{2} q(|U|)}| \cdot 1\{U \le 1/2\} \big) \le C e_n.$$

Let $b_n = 3\delta \sqrt{r/p}$. By Taylor expansion, $q(|x|) \le C_1 |x|^3$ for $x \in [-b_n, b_n]$, $q(|x|) \le C_2 x^2$, for $x \in (b_n, 1/2] \cup (-b_n, \infty)$, where $C_1, C_2 > 0$ are absolute constants. Note that $n r b_n^3 \le C e_n \to 0$. Hence, when $|U| \le b_n$, $|1 - e^{\frac{nr}{2} q(|U|)}| \le C n r b_n^3$; when $|U| > b_n$, $|1 - e^{\frac{nr}{2} q(|U|)}| \le 1 + e^{\frac{nr}{2} q(|U|)} \le 1 + e^{C n r U^2}$. It follows that the left hand side of (6.66) is upper bounded by

$$(6.67) \qquad C n r b_n^3 + P(|U| > b_n) + E\big[ e^{C n r U^2} 1\{|U| > b_n\} \big].$$

The first term is $O(e_n)$. To bound the second term, by (6.51), write

$$|U| = \frac{\delta}{1 - \delta} (\sqrt{r/p} + U_1/\sqrt{p}), \qquad \text{where} \quad P(U_1 > t) \le 2 e^{-t^2/2} \quad \text{for } t > 0.$$

It follows that $P(|U| > b_n) \le P(U_1 > \sqrt{r}) \le 2 e^{-r^2/2} = o(e_n)$. Last, consider the third term. Since $n r U^2 \le C \delta^2 r^2 + C \delta^2 r U_1^2$, this term is bounded by $e^{C \delta^2 r^2} \cdot E[e^{C \delta^2 r U_1^2} 1\{U_1 > \sqrt{r}\}] \le e^{C \delta^2 r^2 - r/4}$, provided that $\sqrt{r} \gg \delta^2 r$ (which is obviously true here). Since $\delta^2 r \le e_n \to 0$, this term is $O(e^{-r/8}) = o(e_n)$. Combing the above gives (6.66), and (6.64) follows.

Next, we prove (6.65). We only consider the first inequality, and the second one is similar. Let $U$ and $b_n$ be the same as above, and let $V = \sum_{k=1}^r \{(\eta_k - 1)^2 - E[(\eta_k - 1)^2]\}$. It is seen that $n|M_2(Y)| = n\frac{\delta^2}{2(1-\delta)^2}|V|$. Moreover, since $\sum_{k=1}^r (\eta_k - 1)^2 \le r \max\{|\eta_1 - 1|^2, |\eta_r - 1|^2\}$ and $\sum_{k=1}^r E[(\eta_k - 1)^2] = r/p$, we have a naive upper bound: $n|M_2(Y)| \le CnrU^2 + C\delta^2 r$. As a result,

$$E^*|1 - e^{\frac{n}{2}M_2(Y)}| \le E\big(|1 - e^{Cn\delta^2 V}| \cdot 1\{|U| \le b_n\}\big) + P(|U| > b_n)$$

(6.68)
$$+ e^{C\delta^2 r} E\big[e^{CnrU^2} 1\{|U| > b_n\}\big].$$

In the analysis of (6.67), we have seen that the second term is $o(e_n)$ and the last term is $O(e^{C\delta^2 r - r/8}) = o(e_n)$. Consider the first term. We adapt the proof of Lemma 3.1 by constructing a function $\tilde{h}_n$, similar to $h_n$ in the proof of Lemma 3.1, except that $a_n$ is now replaced with $3\sqrt{r/p}$. Then, $\|\tilde{h}_n'\|_\infty \le 6\sqrt{r/p}$; in addition, when $|U| \le b_n$, $V = \sum_{k=1}^r \tilde{h}_n(\eta_k) - \sum_{k=1}^r E[(\eta_k - 1)^2]$. Similar to (6.55), we can show that $\sum_{k=1}^r E[(\eta_k - 1)^2] = \sum_{k=1}^r E[\tilde{h}_n(\eta_k)] + o(r/p)$. Therefore, by Theorem 3.1 (with $m = 2$),

$$P\big(|V| > 6p^{-1}rt, |U| \le b_n\big) \le 2 \inf_{y \ge \sqrt{2\log(2)}} \big\{e^{-\frac{t^2}{c_0 y^2}} + 2e^{-y^2}\big\} \le 4e^{-t/\sqrt{c_0}}.$$

With $V_1 = r^{-1}pV \cdot 1\{|U| \le b_n\}$, the above implies $P(|V_1| > 6t) \le 4e^{-c_0^{-1/2}t}$, so $V_1$ has an exponential tail. The first term in (6.68) is upper bounded by $E|1 - e^{C\delta^2 rV_1}| \le C\delta^2 r = O(e_n)$. Combining the above gives (6.65). $\square$

6.5. *Proof of Lemma 4.3.* We first show (4.28). Let $\hat{\Sigma} = \hat{Q}\Lambda\hat{Q}'$ be the eigen decomposition of $\hat{\Sigma}$, where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$ and $\hat{Q}$ is an orthogonal matrix. Let $Z = \hat{Q}'Y$. Write $b_n^*(\delta, r) = (1 - \delta)^{\frac{nr}{2}} b_n(\delta, r)$. We can rewrite

$$\widetilde{R}_n = b_n^*(\delta, r) \int \exp\left\{\frac{\delta n}{2p}\text{tr}\big(Z'(\Lambda - \frac{1}{1-\delta}I_n)Z\big)\right\} dF(Z)$$

$$= b_n^*(\delta, r) \int \exp\left\{\frac{\delta n}{2p}\sum_{k=1}^r \sum_{j=1}^p (\lambda_j - \frac{1}{1-\delta})Z^2(j, k)\right\} dF(Z).$$

Conditioning on $X$, $Z$ has *iid* entries of $N(0, 1)$. For $z \sim N(0, 1)$ and $c < 1$, $E[e^{cz^2/2}] = (1 - c)^{-1/2} = e^{-\log(1-c)/2}$. Therefore,

$$\widetilde{R}_n = b_n^*(\delta, r)\exp\left\{-\frac{r}{2}\sum_{j=1}^p \log\big(1 - \frac{\delta n}{p}(\lambda_j - \frac{1}{1-\delta})\big)\right\}.$$

We then show (4.29). By definition, we have $\sum_{j=1}^p \log(1 - \frac{\delta n}{p}(\lambda_j - \frac{1}{1-\delta})) = \frac{n\delta}{p}\sum_{j=1}^p \psi_n(\lambda_j) = n\delta \int \psi_n(\lambda)\mu_n^X(d\lambda)$. Hence, given (4.28), it suffices to show

that

$$(6.69) \qquad \int \psi_n(\lambda)\mu_{n,p}(d\lambda) = \frac{1}{\delta}\log(1-\delta) + \frac{1}{1-\delta}.$$

Write $\gamma_n = p/n$. Then $\psi_n(\lambda) = \frac{\gamma_n}{\delta}\log(\frac{\delta}{\gamma_n}) + \frac{\gamma_n}{\delta}\log(\frac{\gamma_n}{\delta} + \frac{1}{1-\delta} - \lambda)$. Let $H_\gamma(x) = \int \log(x-\lambda)\mu_\gamma(d\lambda)$ for any $\gamma \geq 1$ and $x > (1+\sqrt{\gamma})^2$. It is seen that

$$(6.70) \qquad \int \psi_n(\lambda)\mu_{n,p}(d\lambda) = \frac{\gamma_n}{\delta}H_{\gamma_n}\left(\frac{\gamma_n}{\delta} + \frac{1}{1-\delta}\right) + \frac{\gamma_n}{\delta}\log(\delta/\gamma_n).$$

We apply Lemma 3.2 with $t = \frac{\gamma_n}{\delta} + \frac{1}{1-\delta}$ and $\gamma = \gamma_n$. By direct calculations, $A_t = \frac{1-\delta}{\delta}\gamma_n + \frac{\delta}{1-\delta}$, $B_t = \frac{1-\delta}{\delta}\gamma_n - \frac{\delta}{1-\delta}$ and $(\gamma_n + 1)A_t - (\gamma_n - 1)B_t + 4\gamma_n = \frac{2\gamma_n}{\delta(1-\delta)}$. Plugging them into Lemma 3.2 and inserting the result into (6.70) gives (6.69). $\qquad \square$

**7. Acknowlegement.** The author thanks Jiashun Jin for pointing out the topic and numerous thoughtful pointers and comments on the manuscript. She thanks Zhigang Bao for personal communications on Random Matrix Theory. She thanks Mark Low, Zhonggen Su and Ofer Zeitouni for helpful pointers.

**References.**

[1] ADAMCZAK, R. and WOLFF, P. (2015). Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probab. Theory Related Fields* **162** 531–586.

[2] BAI, Z., JIANG, D., YAO, J.-F. and ZHENG, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37** 3822–3840.

[3] BAI, Z. D. and SILVERSTEIN, J. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Prob.* **32** 553–605.

[4] BAIK, J., BEN AROUS, G. and PECHE, S. (2005). Phase transition of the largest eigenvalue for none complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697.

[5] BAIK, J. and SILVERSTEIN, J. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivar. Anal.* **97** 1382–1408.

[6] BAO, Z. (2015). A large deviation inequality on the partial sums of eigenvalues of large-dimensional sample covariance matrices. *Manuscript.*

[7] BAO, Z., PAN, G. and ZHOU, W. (2012). Tracy-Widom law for the extreme eigenvalues of sample correlation matrices. *Electron. J. Probab.* **17** 1–32.

[8] BAO, Z., PAN, G. and ZHOU, W. (2013). Central limit theorem for partial linear eigenvalue statistics of Wigner matrices. *J. Stat. Phys.* **150** 88–129.

[9] BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815.

[10] BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **78** 253–273.

[11] BOBKOV, S. G., NAYAR, P. and TETALI, P. (2017). Concentration properties of restricted measures with applications to non-Lipschitz functions. In *Geometric Aspects of Functional Analysis* 25–53. Springer.

44

[12] CAI, T. T. and MA, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli* **19** 2359–2388.

[13] CANDES, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory* **57** 2342–2359.

[14] CHEN, S. X., ZHANG, L.-X. and ZHONG, P.-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105** 810–819.

[15] CHIKUSE, Y. (2003). *Statistics on special manifolds*. Springer Science & Business Media.

[16] DALLAPORTA, S. (2013). Eigenvalue variance bounds for covariance matrices. *arXiv:1309.6265*.

[17] DELYON, B. (2010). Concentration inequalities for the spectral measure of random matrices. *Electron. Comm. Probab.* **15** 549–562.

[18] DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: all eigenvalues matter. *Ann. Statist.* **45** 1810–1833.

[19] DONOHO, D. and JIN, J. (2004). Higher Criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* 962–994.

[20] DONOHO, D. and JIN, J. (2008). Higher Criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **105** 14790–14795.

[21] DONOHO, D. and JIN, J. (2015). Higher Criticism for large-scale inference: especially for rare and weak effects. *Statist. Sci.* **30** 1–25.

[22] DUMAZ, L. and VIRG, B. (2013). The right tail exponent of the Tracy-Widom $\beta$ distribution. *Ann. Inst. H. Poincar Probab. Statist.* **49** 915–933.

[23] EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103.

[24] EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405.

[25] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515.

[26] FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Soc.* **107** 1019–1035.

[27] FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Soc.* **104** 1406–1415.

[28] GÖTZE, F. and TIKHOMIROV, A. (2011). On the rate of convergence to the Marchenko-Pastur distribution. *arXiv:1110.1284*.

[29] GUIONNET, A. and ZEITOUNI, O. (2000). Concentration of the spectral measure for large matrices. *Electron. Comm. Probab.* **5** 119-136.

[30] JIN, J., KE, Z. T. and WANG, W. (2017). Phase transitions for high dimensional clustering and related problems. *Ann. Statist.* **45** 2151–2189.

[31] JIN, J. and WANG, W. (2016). Influential Features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359.

[32] JOHNSTONE, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* 295–327.

[33] JOHNSTONE, I. and NADLER, B. (2017). Roy's largest root test under rank-one alternatives. *Biometrika* **104** 181–193.

[34] LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102.

[35] LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3** e161.

[36] Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424.

[37] Onatski, A., Moreira, M. and Hallin, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1206–1231.

[38] Onatski, A., Moreira, M. and Hallin, M. (2014). Signal detection in high dimension: the multispiked case. *Ann. Statist.* **42** 225–254.

[39] Patterson, N., Price, A. L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics* **2** e190.

[40] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617.

[41] Péché, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173.

[42] Pillai, N. S. and Yin, J. (2014). Universality of covariance matrices. *Ann. Appl. Probab.* **24** 935–1001.

[43] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38** 904.

[44] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci.* **101** 2658–2663.

[45] Su, Z. (2006). Gaussian fluctuations in complex sample covariance matrices. *Electron. J. Probab.* **11** 1284–1320.

[46] Tracy, C. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177** 727–754.

[47] Tulino, A. and Verdu, S. (2004). *Radom matrix theory and wireless communications* **1**. Now Publisher Inc.

[48] Uhlig, H. (1994). On singular Wishart and singular multivariate beta distributions. *Ann. Statist.* **22** 395–405.

[49] Van der Vaart, A. (2000). *Asymptotic statistics.* Cambridge University Press.

[50] Vershynin, R. (2012). *Compressed Sensing* Introduction to the non-asymptotic analysis of random matrices, 210-268. Cambridge Univ. Press.

[51] Vu, V. (2002). Concentration of non-Lipschitz functions and applications. *Random Structures Algorithms* **20** 262–316.

[52] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comp. Graph. Stat.* **15** 265–286.

Z. Ke
Department of Statistics
Harvard University
Cambridge, Massachusetts, 02138
USA
E-mail: zke@fas.harvard.edu