The Annals of Statistics
2022, Vol. 50, No. 1, 407–429
https://doi.org/10.1214/21-AOS2113

TESTING EQUIVALENCE OF CLUSTERING

BY CHAO GAO[1] AND ZONGMING MA[2]

[1]*Department of Statistics, University of Chicago, chaogao@galton.uchicago.edu*
[2]*Department of Statistics and Data Science, University of Pennsylvania, zongming@wharton.upenn.edu*

In this paper, we test whether two data sets measured on the same set of subjects share a common clustering structure. As a leading example, we focus on comparing clustering structures in two independent random samples from two deterministic two-component mixtures of multivariate Gaussian distributions. Mean parameters of these Gaussian distributions are treated as potentially unknown nuisance parameters and are allowed to differ. Assuming knowledge of mean parameters, we first determine the phase diagram of the testing problem over the entire range of signal-to-noise ratios by providing both lower bounds and tests that achieve them. When nuisance parameters are unknown, we propose tests that achieve the detection boundary adaptively as long as ambient dimensions of the data sets grow at a sublinear rate with the sample size.

1. Introduction. Clustering is one of the most widely used unsupervised learning techniques. As the simplest nontrivial model for clustering, suppose we have independent observations $X_i \overset{\text{ind}}{\sim} N(z_i\theta, I_p)$, $i = 1, \ldots, n$. Here, $\{z_i\}_{1\le i\le n}$ take their values in $\{-1, 1\}$ and are *deterministic* cluster labels, and $\theta \in \mathbb{R}^p$. The n subjects thus form two clusters according to their cluster labels. Note that the X_i's can only be generated from two candidate multivariate Gaussian distributions. We assume that the two means are θ and $-\theta$ for convenience. In practice, as long as both cluster sizes grow linearly with n, one can estimate the means under all settings considered in this manuscript and recenter the data at the average of these two estimated means. All conclusions in this paper then follow. In what follows, we call this generative model a *deterministic two-component Gaussian mixture model*. This is to be differentiated from the usual two-component Gaussian mixture model where the z_i's are random.

Under the foregoing model, uncovering clustering structure is equivalent to estimating the unknown deterministic label vector $z \in \{-1, 1\}^n$. Let $a \wedge b = \min(a, b)$ for any real numbers a and b. We measure the distance between two clustering structures by

$$\ell(\widehat{z}, z) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\widehat{z}_i \neq z_i\}}\right) \wedge \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\widehat{z}_i \neq -z_i\}}\right),$$

the normalized Hamming distance up to a label switching. In the last display, taking minimum over label switching is necessary since switching labels does not change the clustering structure. Under this model setting, if $\|\theta\|^2 \to \infty$ as $n \to \infty$, then the asymptotic minimax risk of estimating z satisfies

$$\inf_{\widehat{z}} \sup_{z\in\{-1,1\}^n} \mathbb{E}_{(z,\theta)}\ell(\widehat{z}, z) = \exp\left(-\big(1 + o(1)\big)\frac{\|\theta\|^2}{2}\right). \tag{1}$$

See, for instance, [22].

Received February 2020; revised June 2021.
MSC2020 subject classifications. Primary 62C20, 62H15; secondary 62H30.
Key words and phrases. Minimax testing error, sparse mixture, phase transition, discrete structure inference, high-dimensional statistics.

Nowadays, practitioners in a number of fields are able to collect multiple data sets measured on the same subjects, and it has become popular to integrate these different data sets to find a common clustering structure. For example, in cancer genomics, researchers combine different molecular features such as copy number variations and gene expressions measured on the same patients to reveal novel tumor subtypes [9, 26, 28]. In collaborative filtering, combining ratings on different types of items (e.g., movies and songs) from the same users helps to better identify user types, and thus leads to better recommendations [23, 24, 31]. In covariate-assisted network clustering [3, 10], additional nodal features are collected to facilitate the clustering of nodes in a social network. From a theoretical viewpoint, suppose that we collect observations on q additional covariates on the same n subjects as a second data set. Let $Y_1, \dots, Y_n$ denote the random vectors formed by these q covariates. Similar to our generative model on the X_i's, we assume that $Y_i \overset{\text{ind}}{\sim} N(z_i\eta, I_q)$ for some $\eta \in \mathbb{R}^q$, where the cluster labels $\{z_i\}$ are the same as those for the X_i's. Further assume that the two data sets are mutually independent. Now that they have the same clustering structure, using them together improves the estimation error rate for z from (1) to

$$\exp\biggl(-(1+o(1))\frac{\|\theta\|^2+\|\eta\|^2}{2}\biggr). \tag{2}$$

In both theory and practice that we have mentioned here, the underlying assumption is that these different data sets share the same clustering structure.

In the present paper, we focus on checking this key assumption based on data for a pair of data sets under a stylized model. To formulate the problem, suppose the second data set $\{Y_i\}_{1\le i\le n}$ possesses a potentially different label vector $\sigma = (\sigma_1, \dots, \sigma_n)^T \in \{-1, 1\}^n$. Thus, we have two data sets $X_i \overset{\text{ind}}{\sim} N(z_i\theta, I_p)$ and $Y_i \overset{\text{ind}}{\sim} N(\sigma_i\eta, I_q)$, $i = 1, \dots, n$, measured on the same set of n subjects, and they are mutually independent. The assumption that the two data sets share a common clustering structure can then be expressed as $\ell(z, \sigma) = 0$. From a statistical viewpoint, checking whether $\ell(z, \sigma) = 0$ is equivalent to testing

$$H_0 : \ell(z, \sigma) = 0 \quad \text{vs.} \quad H_1 : \ell(z, \sigma) \ge \epsilon \tag{3}$$

for some $\epsilon > 0$. In the literature, such a testing problem is sometimes also called a detection problem. For any n and ϵ, define parameter spaces

$$\mathcal{F}_n^0 = \bigl\{(z, \sigma) : z, \sigma \in \{-1, 1\}^n, \ell(z, \sigma) = 0\bigr\}, \tag{4}$$

$$\mathcal{F}_n^1(\epsilon) = \bigl\{(z, \sigma) : z, \sigma \in \{-1, 1\}^n, \ell(z, \sigma) \ge \epsilon\bigr\}. \tag{5}$$

Here and after, we focus on the case of two clusters and leave the case of more than two clusters for future research. The parameter spaces do not put any restriction on the sizes of the two clusters under either the null or the alternative. Let $P^{(n)}_{(\theta,\eta,z,\sigma)}$ be the joint distribution of the two data sets $\{X_i\}_{1\le i\le n}$ and $\{Y_i\}_{1\le i\le n}$. For any testing procedure ψ, its worst-case testing error is

$$R_n(\psi, \theta, \eta, \epsilon) = \sup_{(z,\sigma)\in\mathcal{F}_n^0} P^{(n)}_{(\theta,\eta,z,\sigma)}\psi + \sup_{(z,\sigma)\in\mathcal{F}_n^1(\epsilon)} P^{(n)}_{(\theta,\eta,z,\sigma)}(1-\psi). \tag{6}$$

We are interested in the asymptotic setting where ψ, θ, η and ϵ all scale with the sample size n. For conciseness of notation, the dependence of these sequences on n is not expressed explicitly. For given sequences of θ, η and ϵ, we call a sequence of tests ψ consistent if $R_n(\psi, \theta, \eta, \epsilon) \to 0$ as $n \to \infty$. For every sample size n, the minimax testing error is defined as

$$R_n(\theta, \eta, \epsilon) = \inf_{\psi} R_n(\psi, \theta, \eta, \epsilon).$$

We say consistent testing of (3), or consistent detection, is possible, if the sequences of θ, η and ϵ are such that $R_n(\theta, \eta, \epsilon) \to 0$ as $n \to \infty$. Otherwise, we say consistent testing, or consistent detection, is impossible. Our goal is to find necessary and sufficient conditions (under appropriate calibrations) for θ, η and ϵ sequences such that consistent testing is possible for (3), and to find a sequence of consistent tests under these conditions.

To facilitate our discussion, let us assume for now that $\|\theta\| = \|\eta\|$, that is the two data sets have equal signal-to-noise ratios (SNRs). An intuitive approach to testing (3) is to first estimate z and σ with $\widehat{z}$ and $\widehat{\sigma}$, respectively, by separately clustering the two data sets. Then we can reject the null hypothesis when $\ell(\widehat{z}, \widehat{\sigma}) > \epsilon/2$. With the known minimax optimal estimation error rate in (1), one can show that this test is consistent as long as

$$\|\theta\|^2 = \|\eta\|^2 > 2\log\left(\frac{1}{\epsilon}\right). \tag{7}$$

However, condition (7) is far from optimal and the minimum SNR required for consistent testing is much weaker.

A better test can be based on a reduction to a well-studied sparse signal detection problem. Recall that $\|\theta\| = \|\eta\|$ is assumed. To further avoid technicalities, suppose that z and σ have already been aligned so that $\ell(z, \sigma) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i \neq \sigma_i\}}$, and so there is no ambiguity due to label switching. In this case, for the ith subject, the pairwise difference of projected data is

$$D_i = \frac{\theta^T X_i - \eta^T Y_i}{\sqrt{\|\theta\|^2 + \|\eta\|^2}} \overset{\text{ind}}{\sim} N\left(\frac{1}{\sqrt{2}}(z_i - \sigma_i), 1\right),$$

and the D_i's are mutually independent. Under H_0 in (3), $D_i \overset{\text{iid}}{\sim} N(0, 1)$, while under H_1, there are at least an ϵ fraction of coordinates distributed either as $N(\sqrt{2}\|\theta\|, 1)$ or $N(-\sqrt{2}\|\theta\|, 1)$. This is the sparse signal detection problem studied by [11, 16, 17] under the asymptotic setting where $\epsilon = n^{-\beta}$ and $\sqrt{2}\|\theta\| = \sqrt{2r\log n}$ for some constants $\beta, r > 0$. It was shown in [11] and [8] that the higher criticism test is consistent as long as the pair (r, β) satisfy

$$\beta < \beta^*_{\text{IDJ}}(r) = \begin{cases} \dfrac{1}{2} + r, & 0 < r \le \dfrac{1}{4}, \\ 1 - (1 - \sqrt{r})_+^2, & r > \dfrac{1}{4}. \end{cases} \tag{8}$$

Following [8], we call $\beta^*_{\text{IDJ}}(r)$ the Ingster–Donoho–Jin threshold. In addition, it was shown in [16] that the threshold on the right-hand side of (8) cannot be improved. As we shall show, it is also the information theoretic limit (to the first order) for testing (3) if we only observe $\{D_i\}_{1 \le i \le n}$. This approach improves upon the plug-in procedure in the previous paragraph as the SNR condition (8) is always weaker than (7). We shall show that we can further improve condition (8) by fully utilizing the original data $\{X_i\}_{1 \le i \le n}$ and $\{Y_i\}_{1 \le i \le n}$ since one actually loses information by working only with the difference sequence $\{D_i\}_{1 \le i \le n}$. The threshold function on the right-hand side of (8) partitions all (r, β) values of interest into two disjoint regions. One corresponds to (r, β) values where consistent testing is possible, and the other where consistent testing is impossible.[1] Thus, the (r, β) plane with such a detection boundary function (or threshold function) is called a phase diagram in the literature.

Main contributions. The main result of the present paper is twofold. First, we determine the phase diagram of the testing problem (3) under a natural asymptotic setting comparable to that used in [11], assuming knowledge of θ and η. In addition, we derive an asymptotically

[1] Throughout the paper, we do not focus on parameter values on the graph of the detection boundary function.

optimal test that achieves the detection boundary adaptively when θ and η are unknown while p and q grow at sublinear rates with n.

In the equal SNR case where we continue to use the foregoing (r, β) parametrization, we shall show that the detection boundary for (3) by using all information in data can be improved from (8) to

$$\beta^*(r) = \begin{cases} \frac{1}{2}(1+3r), & 0 < r \le \frac{1}{5}, \\ \sqrt{1-(1-2r)_+^2}, & r > \frac{1}{5}. \end{cases} \tag{9}$$

This detection boundary is uniformly better than that in (8) as $\beta^*(r) \ge \beta^*_{\mathrm{IDJ}}(r)$ for all $r > 0$. See Figure 1 for a graphical comparison. The determination of the new detection boundary relies crucially on the investigation of a closely related bivariate sparse normal mixture detection problem, which takes into account not only the information in D_i but also that in $\frac{\theta^T X_i + \eta^T Y_i}{\sqrt{\|\theta\|^2 + \|\eta\|^2}}$.

Turn to the general case. We use three quantities to parametrize the testing problem (3): parameter β related to distance between two clustering structures, parameter r characterizing average SNR and an additional parameter s characterizing difference in SNRs between the two data sets. So the pair (r, s) jointly characterize the SNRs in data. The foregoing case of equal SNRs then reduces to a special case of the present parametrization with $s = 0$. In the general case, our first contribution is the phase diagram in the (r, s, β) space for testing (3). The detection boundary function that partitions the space into two disjoint regions (one for parameter combinations where consistent testing is possible and the other for impossible cases) can be expressed as a bivariate function $\beta^*(r, s)$. As we shall show, this function has five pieces defined on five disjoint subsets of the (r, s) domain. Our second contribution in the general case is the proposal of an asymptotically optimal test that achieves the detection boundary. The test that provably achieves the detection boundary is motivated by a higher criticism test for the related bivariate sparse mixture detection problem. At the heart of our proposal is a precise likelihood ratio approximation. This leads to a sequence of "asymptotically sufficient" statistics, based on which a relatively simple higher criticism type test can be shown to be optimal. When θ and η are unknown, the optimal test can be made adaptive if p and q grow at some sublinear rates with n. When $p, q \gg n$, one may still achieve the detection boundary adaptively if additional structural assumptions on θ and η are imposed so that they can be estimated with sufficiently high accuracy.

Related works. The testing problem (3) is related to feature selection in clustering analysis. In the literature, this has mainly been investigated in the context of sparse clustering [2, 5, 19, 20, 29], where it is assumed that only a small subset of covariates are useful for finding clusters, and so it is important to identify them. In comparison, the testing problem (3) is concerned with whether inclusion of an additional set of covariates $\{Y_i\}_{1 \le i \le n}$ can potentially lead to smaller clustering errors than using $\{X_i\}_{1 \le i \le n}$ alone. A major difference is that the additional set of covariates may admit a completely different clustering structure in our setting, while in sparse clustering, covariates that are not useful have no clustering structure. On a separate note, [19] also touched on the problem of testing existence of clustering structure for one data set. In the most comparable scenarios, their work focused on the regime where phase transition happens when $\beta \le \frac{1}{2}$ while the present paper focuses exclusively on phase transitions occurring at some $\beta > \frac{1}{2}$. We leave the investigation of phase transitions with $\beta \le \frac{1}{2}$ for future research.

In addition to testing whether clustering structures in multiple data sets are the same, it is of interest to approach the problem from a different angle. In particular, one could test

whether multiple clustering structures share anything in common. We refer the readers to [14] and [15] for studies along this line. In particular, [14] considered the case where one observes two data sets from some random mixture distributions and entries of clustering vectors z and σ follow some joint distribution. They tested whether the random vectors z and σ are independent or not, which is complementary to the problem (3) we investigate here. See [15] for generalization to network data. Both [14] and [15] used pseudo likelihood ratio tests. Zhao et al. [32] used higher criticism approach toward testing independence against positive correlation between random z and σ based on indirect observations, which is related to the independence testing problem in [14]. See also [33]. The alternative hypotheses in [32] and [33] are one-sided, while for testing independence of clustering label vectors, they are naturally two-sided due to potential label permutation. Furthermore, Arias-Castro et al. [1] studied optimality of higher criticism test of independence against positive correlation based on a random sample of bivariate normal random vectors.

As we have mentioned earlier, there has been a growing literature on integrative clustering. However, almost all of them were devoted to methodology development in different application scenarios. To the best of our limited knowledge, there is little theoretical work on statistical optimality besides [10], which focused on optimal estimation rate of a common clustering structure shared by a two-block stochastic block model and a two-component Gaussian mixture.

Paper organization. The rest of the paper is organized as the following. Section 2 studies (3) with an additional equal SNR assumption. This simplified setting demonstrates essence of the problem while reducing a lot of technicalities. The general version of the problem without equal SNR assumption is studied in Section 3. In Section 4, we consider (3) with $\epsilon = 1/n$, which is testing for exact equality. Optimal adaptive tests with unknown parameters are discussed in Section 5. Section 6 presents some simulation results. Finally, technical proofs are given in the Supplementary Material [13].

Notation. For $d \in \mathbb{N}$, we write $[d] = \{1, \dots, d\}$. Given $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$ and $a_+ = \max(a, 0)$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ when there exists a constant $C > 0$ independent of n such that $a_n \le C b_n$ for all n. Moreover, $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a set S, we use $\mathbf{1}_{\{S\}}$ and $|S|$ to denote its indicator function and cardinality, respectively. For any matrix A, A^T stands for its transpose. Any vector $v \in \mathbb{R}^d$ is by default a $d \times 1$ matrix. For a vector $v = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, we define $\|v\|^2 = \sum_{\ell=1}^d v_\ell^2$. The trace inner product between two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ is defined as $\langle A, B \rangle = \sum_{\ell=1}^{d_1} \sum_{\ell'=1}^{d_2} A_{\ell\ell'} B_{\ell\ell'}$, while the Frobenius and operator norms of A are $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A \rangle}$ and $\|A\|_{\mathrm{op}} = s_{\max}(A)$, respectively, where $s_{\max}(\cdot)$ denotes the largest singular value. We use $\mathbb{P}$ and $\mathbb{E}$ for generic probability and expectation operators whose distribution is determined by the context.

2. Testing with equal signal-to-noise ratios.

Recall that we have two independent data sets $X_i \overset{\text{ind}}{\sim} N(z_i\theta, I_p)$ and $Y_i \overset{\text{ind}}{\sim} N(\sigma_i\eta, I_q)$ for $i \in [n]$. In this section, we first assume that SNRs of the two data sets are equal. In other words, $\|\theta\| = \|\eta\|$. The general case of potentially unequal SNRs is more complicated and will be studied in Section 3.

First, we show that we can apply dimension reduction to both data sets without losing any information for testing (3). Consider $\{X_i\}_{1 \le i \le n}$. Since the clustering structure only appears in the direction of θ, we can project all X_i's to the one-dimensional subspace spanned by the unit vector $\theta/\|\theta\|$. After projection, we obtain $\theta^T X_i/\|\theta\| \sim N(z_i\|\theta\|, 1)$ for $i \in [n]$. Given θ, $\theta^T X_i/\|\theta\|$ is a sufficient statistic for parameter z_i. Therefore, we conclude that the projected data set $\{\theta^T X_i/\|\theta\|\}_{1 \le i \le n}$ preserves all clustering information. The same argument

also applies to $\{Y_i\}_{1\le i\le n}$. In the rest of this section, we write

$$\widetilde{X}_i = \theta^T X_i/\|\theta\| \quad \text{and} \quad \widetilde{Y}_i = \eta^T Y_i/\|\eta\| \tag{10}$$

for $i \in [n]$ and work with these one-dimensional random variables when constructing tests. On the other hand, we shall establish lower bounds of the testing problem directly in the original multidimensional setting.

2.1. *A connection to sparse signal detection.*

A related sparse mixture detection problem. For simplicity, let us suppose for now that $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i\neq\sigma_i\}} \le \frac{1}{2}$, so that $\ell(z,\sigma) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i\neq\sigma_i\}}$. Under H_0 in (3), both $\widetilde{X}_i \overset{\text{ind}}{\sim} N(z_i\|\theta\|, 1)$ and $\widetilde{Y}_i \overset{\text{ind}}{\sim} N(z_i\|\theta\|, 1)$ for all $i \in [n]$, which motivates us to compute scaled differences $(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}$, $i \in [n]$.

Note that the null and the alternative distributions of $\{(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}\}_{1\le i\le n}$ under (3) are the same as those in a sparse signal detection problem. Indeed, $(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2} \overset{\text{iid}}{\sim} N(0,1)$ for $i \in [n]$ under the null, and at least an ϵ fraction of the statistics follow either $N(\sqrt{2}\|\theta\|, 1)$ or $N(-\sqrt{2}\|\theta\|, 1)$ under the alternative. A well-studied Bayesian version of this sparse signal detection problem is given by the following form: for $i \in [n]$,

$$H_0 : U_i \overset{\text{iid}}{\sim} N(0,1), \tag{11}$$

$$H_1 : U_i \overset{\text{iid}}{\sim} (1-\epsilon)N(0,1) + \frac{\epsilon}{2}N\big(-\sqrt{2}\|\theta\|, 1\big) + \frac{\epsilon}{2}N\big(\sqrt{2}\|\theta\|, 1\big). \tag{12}$$

In what follows, we refer to (11)–(12) (and any such Bayesian version of the problem) as a sparse mixture detection problem. There are two noticeable differences between (12) and the distribution of $\{(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}\}_{1\le i\le n}$ under H_1 in (3):

1. The number of nonnull signals in (12) is a binomial random variable while it is deterministic in (3);
2. The probabilities that a nonnull signal is from $N(\sqrt{2}\|\theta\|, 1)$ and from $N(-\sqrt{2}\|\theta\|, 1)$ are equal in (12) while in (3) there is no restriction on how many nonnull signals follow either of the two distributions.

However, these differences are inconsequential as long as our focus is on the phase diagrams of these testing problems with the calibration we now introduce.

For either the hypothesis testing problem (11)–(12) or (3) with $\|\theta\| = \|\eta\|$, introduce the calibration

$$\epsilon = n^{-\beta} \quad \text{and} \quad \sqrt{2}\|\theta\| = \sqrt{2r\log n}. \tag{13}$$

For (11)–(12),[2] it was proved in [16, 17] for that the likelihood ratio test is consistent when $\beta < \beta^*_{\text{IDJ}}(r)$ and no test is consistent when $\beta > \beta^*_{\text{IDJ}}(r)$, where the threshold function is

$$\beta^*_{\text{IDJ}}(r) = \begin{cases} \dfrac{1}{2} + r, & 0 < r \le \dfrac{1}{4}, \\ 1 - (1-\sqrt{r})^2_+, & r > \dfrac{1}{4}. \end{cases} \tag{14}$$

Note that $\beta < \beta^*_{\text{IDJ}}(r)$ is equivalent to (8). Moreover, Donoho and Jin [11] proposed a higher criticism (HC) test that rejects H_0 when

$$\sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|U_i|^2>t\}} - n\mathbb{P}(\chi^2_1 > t)|}{\sqrt{n\mathbb{P}(\chi^2_1 > t)(1-\mathbb{P}(\chi^2_1 > t))}} > \sqrt{2(1+\delta)\log\log n},$$

[2]The non-Bayesian version of the problem has also been studied in [16, 17].

where χ^2_m denotes a chi-square distribution with m degrees of freedom and $\delta > 0$ is some arbitrary fixed constant. They proved that the HC test adaptively achieves consistency when $\beta < \beta^*_{\mathrm{IDJ}}(r)$. We refer interested readers to [12, 18] for more discussions on HC tests.

Result for testing equivalence of clustering. Turn to (3). We need to slightly modify the HC test to accommodate the possibility of label switching in the clustering context. Define

$$T_n^- = \sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|\widetilde{X}_i - \widetilde{Y}_i|^2/2 > t\}} - n\mathbb{P}(\chi_1^2 > t)|}{\sqrt{n\mathbb{P}(\chi_1^2 > t)(1 - \mathbb{P}(\chi_1^2 > t))}},$$

$$T_n^+ = \sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|\widetilde{X}_i + \widetilde{Y}_i|^2/2 > t\}} - n\mathbb{P}(\chi_1^2 > t)|}{\sqrt{n\mathbb{P}(\chi_1^2 > t)(1 - \mathbb{P}(\chi_1^2 > t))}}.$$

Based on these two statistics, we define

$$\psi = \mathbf{1}_{\{T_n^- \wedge T_n^+ > \sqrt{2(1+\delta)\log\log n}\}}, \tag{15}$$

where $\delta > 0$ is an arbitrary fixed constant. Taking the minimum of T_n^+ and T_n^- makes the test invariant to label switching since if $\ell(z, \sigma) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i \neq -\sigma_i\}}$ then the previous discussion applied to $\{(\widetilde{X}_i + \widetilde{Y}_i)/\sqrt{2}, i \in [n]\}$ rather than $\{(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}, i \in [n]\}$.

PROPOSITION 2.1. *For testing* (3) *with the assumption that* $\|\theta\| = \|\eta\|$ *and the calibration in* (13), *the test* (15) *satisfies* $\lim_{n\to\infty} R_n(\psi, \theta, \eta, \epsilon) = 0$ *as long as* $\beta < \beta^*_{\mathrm{IDJ}}(r)$.

Proposition 2.1 shows that the test (15) consistently distinguishes two clustering structures under the same condition that implies consistency in the sparse mixture detection problem (11)–(12). This being said, it is not clear at this point whether $\beta^*_{\mathrm{IDJ}}(r)$ is the detection boundary for (3) under the equal SNR assumption and the calibration (13), which if were true, would require that no consistent test exists when $\beta > \beta^*_{\mathrm{IDJ}}(r)$.

REMARK 2.1. Another straightforward way to testing (3) is to first estimate z and σ and then reject H_0 if the two estimators are not sufficiently close. Let $\widehat{z}$ and $\widehat{\sigma}$ be minimax rate optimal estimators of z and σ that satisfy the error bounds (1). A natural test is then $\psi_{\mathrm{estimation}} = \mathbf{1}_{\{\ell(\widehat{z},\widehat{\sigma}) > \epsilon/2\}}$. It can be shown that $\lim_{n\to\infty} R_n(\psi_{\mathrm{estimation}}, \theta, \eta) = 0$ when $\beta < r/2$ under the calibration (13). Compared with the condition $\beta < \beta^*_{\mathrm{IDJ}}(r)$ required by the test (15), $\psi_{\mathrm{estimation}}$ needs a stronger SNR to achieve consistency and is hence inferior.

2.2. *The lost information.* The natural follow-up question is whether the condition $\beta < \beta^*_{\mathrm{IDJ}}(r)$ in Proposition 2.1 is necessary for consistently testing (3) with the equal SNR assumption and the calibration (13). In order to address this lower bound question, let us continue to suppose $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i \neq \sigma_i\}} \leq \frac{1}{2}$ so that we ignore label switching temporarily. A key observation is that by reducing the data from $(\widetilde{X}_i, \widetilde{Y}_i)$ to $(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}$, we have thrown away all the information in $(\widetilde{X}_i + \widetilde{Y}_i)/\sqrt{2}$. To see its consequence, we now study the sequence $\{(\widetilde{X}_i + \widetilde{Y}_i)/\sqrt{2}\}_{1\leq i\leq n}$.

We note that whether $z_i = \sigma_i$ not only changes the distribution of $(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}$, but also the distribution of $(\widetilde{X}_i + \widetilde{Y}_i)/\sqrt{2}$. In fact, we have

$$\frac{1}{\sqrt{2}}(\widetilde{X}_i + \widetilde{Y}_i) \sim \begin{cases} N(\pm\sqrt{2}\|\theta\|, 1), & z_i = \sigma_i, \\ N(0, 1), & z_i \neq \sigma_i. \end{cases}$$

Since there is at least an ϵ fraction of clustering labels that do not match, a natural corresponding sparse mixture detection problem is the following:

$$H_0 : V_i \overset{\text{iid}}{\sim} \frac{1}{2}N(-\sqrt{2}\|\theta\|, 1) + \frac{1}{2}N(\sqrt{2}\|\theta\|, 1), \tag{16}$$

$$H_1 : V_i \overset{\text{iid}}{\sim} \frac{1-\epsilon}{2}N(-\sqrt{2}\|\theta\|, 1) + \frac{1-\epsilon}{2}N(\sqrt{2}\|\theta\|, 1) + \epsilon N(0, 1). \tag{17}$$

Compared with (11)–(12), the roles of $N(0, 1)$ and $\frac{1}{2}N(-\sqrt{2}\|\theta\|, 1) + \frac{1}{2}N(\sqrt{2}\|\theta\|, 1)$ have been switched in (16)–(17). To our limited knowledge, the testing problem (16)–(17) has not been studied in the literature before. With the same calibration (13), its fundamental limit is given by the following theorem.

THEOREM 2.1. *Consider testing* (16)–(17) *with calibration* (13). *Define*

$$\bar{\beta}^*(r) = 1 \wedge \frac{r+1}{2}. \tag{18}$$

When $\beta < \bar{\beta}^*(r)$, *the likelihood ratio test is consistent. When* $\beta > \bar{\beta}^*(r)$, *no test is consistent.*

Theorem 2.1 shows that the optimal threshold (in terms of the calibration (13)) for the testing problem (16)–(17) is $\bar{\beta}^*(r)$. It is easy to check that

$$\bar{\beta}^*(r) \le \beta^*_{\text{IDJ}}(r) \quad \text{for all } r > 0.$$

This indicates that the sequence $\{(\widetilde{X}_i + \widetilde{Y}_i)/\sqrt{2}\}_{1\le i\le n}$ does contain information, but not as much as that in $\{(\widetilde{X}_i - \widetilde{Y}_i)/\sqrt{2}\}_{1\le i\le n}$. Similar to (15), we can also design an HC-type test as motivated by (16)–(17). Define

$$\bar{T}_n^+ = \sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{(\widetilde{X}_i+\widetilde{Y}_i)^2/2\le t\}} - \mathbb{P}(\chi^2_{1,2\|\theta\|^2} \le t)|}{\sqrt{n\mathbb{P}(\chi^2_{1,2\|\theta\|^2} \le t)(1 - \mathbb{P}(\chi^2_{1,2\|\theta\|^2} \le t))}},$$

$$\bar{T}_n^- = \sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{(\widetilde{X}_i-\widetilde{Y}_i)^2/2\le t\}} - \mathbb{P}(\chi^2_{1,2\|\theta\|^2} \le t)|}{\sqrt{n\mathbb{P}(\chi^2_{1,2\|\theta\|^2} \le t)(1 - \mathbb{P}(\chi^2_{1,2\|\theta\|^2} \le t))}}.$$

Here and after, $\chi^2_{m,a}$ denotes a noncentral chi-square distribution with m degrees of freedom and noncentrality parameter a. In addition to $\bar{T}_n^+$, we need $\bar{T}_n^-$ to accommodate the possibility of $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{z_i\ne\sigma_i\}} > \frac{1}{2}$. The overall test for our original problem is then

$$\bar{\psi} = \mathbf{1}_{\{\bar{T}_n^- \wedge \bar{T}_n^+ > \sqrt{2(1+\delta)\log\log n}\}}, \tag{19}$$

where $\delta > 0$ is an arbitrary fixed constant.

THEOREM 2.2. *For testing* (3) *with the assumption that* $\|\theta\| = \|\eta\|$ *and the calibration in* (13), *the test* (19) *satisfies* $\lim_{n\to\infty} R_n(\bar{\psi}, \theta, \eta, \epsilon) = 0$ *as long as* $\beta < \bar{\beta}^*(r)$.

2.3. *Combining the two views.* Proposition 2.1 and Theorem 2.2 show that the original testing problem (3) can be tested based on both pairwise differences and pairwise sums. Due to their mutual independence, the two views are complementary. Both are nontrivial and lead to tests for the original problem (3) that achieve consistency under appropriate conditions. We now show that to achieve information-theoretic limit in the original testing problem (3) with all data under equal SNR assumption with the calibration (13), we need to combine the two views. In what follows, we first explain how to combine two views in sparse mixture

detection. This is followed by our main result for testing equivalence of clustering as in (3) with equal SNR assumption. Interestingly, Cai et al. [7] discovered a similar phenomenon that one achieves additional power by using a complementary sequence in a different context, namely two sample multiple testing.

Sparse mixture detection. We now study the combination of the two views (11)–(12) and (16)–(17), which can be formulated as testing for $i \in [n]$:

$$
\begin{aligned}
H_0 : (U_i, V_i) &\overset{\text{iid}}{\sim} \frac{1}{2} N(0,1) \otimes N\big(-\sqrt{2}\|\theta\|, 1\big) \\
&\quad + \frac{1}{2} N(0,1) \otimes N\big(\sqrt{2}\|\theta\|, 1\big) \quad \text{vs.}
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
H_1 : (U_i, V_i) &\overset{\text{iid}}{\sim} \frac{1-\epsilon}{2} N(0,1) \otimes N\big(-\sqrt{2}\|\theta\|, 1\big) \\
&\quad + \frac{1-\epsilon}{2} N(0,1) \otimes N\big(\sqrt{2}\|\theta\|, 1\big) \\
&\quad + \frac{\epsilon}{2} N\big(-\sqrt{2}\|\theta\|, 1\big) \otimes N(0,1) + \frac{\epsilon}{2} N\big(\sqrt{2}\|\theta\|, 1\big) \otimes N(0,1).
\end{aligned} \tag{21}
$$

The critical values in (14) and (18) can now be viewed as detection boundaries for testing (20)–(21) when only $\{U_i\}_{1 \le i \in n}$ and only $\{V_i\}_{1 \le i \in n}$ are used, respectively. The two components U_i and V_i behave very differently under null and alternative. The value of $|U_i|$ tends to be smaller under H_0 and larger under H_1, while the value of $|V_i|$ behaves in the opposite way. This motivates us to combine the two pieces of information by working with $|U_i| - |V_i|$, which tends to be smaller under H_0 and larger under H_1. Since there is on average an ϵ fraction of nonnulls under H_1, we may reject H_0 if $\sum_{i=1}^n \mathbf{1}_{\{|U_i|-|V_i|>t\}}$ is too large for some t. This intuition motivates us to consider the following HC-type test. Define the survival function

$$
S_{\|\theta\|}(t) = \mathbb{P}_{(U^2,V^2)\sim\chi_1^2\otimes\chi^2_{1,2\|\theta\|^2}}\big(|U| - |V| > t\big).
$$

We reject H_0 when

$$
\sup_{t\in\mathbb{R}} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|U_i|-|V_i|>t\}} - nS_{\|\theta\|}(t)|}{\sqrt{nS_{\|\theta\|}(t)(1-S_{\|\theta\|}(t))}} > \sqrt{2(1+\delta)\log\log n}, \tag{22}
$$

where $\delta > 0$ is an arbitrary fixed constant. The test statistic depends on U_i and V_i only through $|U_i|$ and $|V_i|$, the null and alternative distributions of which remain unchanged if we replace the mixing proportions $\{\frac{1}{2}, \frac{1}{2}\}$ in (20)–(21) to $\{\alpha, 1-\alpha\}$ for any $\alpha \in [0,1]$. This observation is key to the fact that the test in (22) continues to work for any cluster fraction and for deterministic cluster labels.

THEOREM 2.3. *Consider testing* (20)–(21) *with calibration* (13). *Define* $\beta^*(r)$ *as in* (9). *When* $\beta < \beta^*(r)$, *the likelihood ratio test and the HC-type test* (22) *are consistent. When* $\beta > \beta^*(r)$, *no test is consistent.*

We plot the three detection boundaries $\bar{\beta}^*(r)$ (red), $\beta^*_{\text{IDJ}}(r)$ (orange) and $\beta^*(r)$ (blue) in Figure 1. Since $\bar{\beta}^*(r) \le \beta^*_{\text{IDJ}}(r) \le \beta^*(r)$ for all $r > 0$, in view of the discussion following (20)–(21) we can conclude that pooling information in $\{U_i\}_{1 \le i \in n}$ and $\{V_i\}_{1 \le i \in n}$ leads to a more powerful test than using either single sequence.

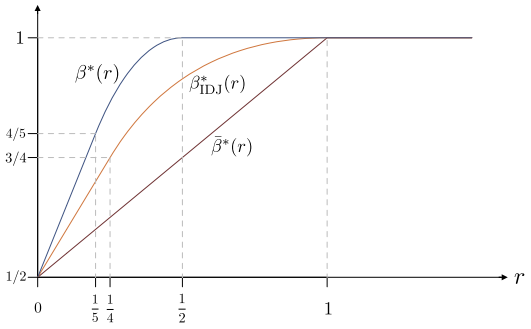

FIG. 1. *Comparison of three detection boundaries.*

Testing equivalence of clustering. We are now in a position to show that $\beta^*(r)$ in (9) is also the detection boundary for testing (3) with all data under the equal SNR assumption and the calibration (13). Motivated by (22) and taking into account possible label switching, we define

$$\check{T}_n^- = \sup_{t\in\mathbb{R}} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|\widetilde{X}_i-\widetilde{Y}_i|-|\widetilde{X}_i+\widetilde{Y}_i|>\sqrt{2}t\}} - nS_{\|\theta\|}(t)|}{\sqrt{nS_{\|\theta\|}(t)(1-S_{\|\theta\|}(t))}},$$

$$\check{T}_n^+ = \sup_{t\in\mathbb{R}} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|\widetilde{X}_i+\widetilde{Y}_i|-|\widetilde{X}_i-\widetilde{Y}_i|>\sqrt{2}t\}} - nS_{\|\theta\|}(t)|}{\sqrt{nS_{\|\theta\|}(t)(1-S_{\|\theta\|}(t))}},$$

and

$$\check{\psi} = \mathbf{1}_{\{\check{T}_n^- \wedge \check{T}_n^+ > \sqrt{2(1+\delta)\log\log n}\}}, \tag{23}$$

where $\delta > 0$ is an arbitrary fixed constant.

THEOREM 2.4. *For testing* (3) *with the assumption that* $\|\theta\| = \|\eta\|$ *and the calibration in* (13), *the test* (23) *satisfies* $\lim_{n\to\infty} R_n(\check{\psi}, \theta, \eta, \epsilon) = 0$ *as long as* $\beta < \beta^*(r)$. *Moreover, we have* $\liminf_{n\to\infty} R_n(\theta, \eta, \epsilon) > 0$, *that is no test is consistent, when* $\beta > \beta^*(r)$.

We conclude this section with several remarks on Theorem 2.4. First, the theorem suggests that the two-dimensional sparse mixture testing problem (20)–(21) contains the mathematical essence of the original testing equivalence of clustering problem (3). In particular, they share the same detection boundary. In addition, it shows that using either the pairwise difference or the pairwise sum sequence only results in a suboptimal solution (see Figure 1). It is worth noting that even the Bayesian version of the testing problem (3), namely (20)–(21), is different from the sparse mixture detection problem (11)–(12) that has been well studied in the literature. Furthermore, it suffices to work with the projected data sets $\{(\theta^T X_i/\|\theta\|, \eta^T Y_i/\|\eta\|)\}_{1\le i\le n}$ when constructing tests, as they are sufficient statistics for (z, σ) with any given values of θ and η.

3. The general phase diagram. In this section, we study the general case of testing (3) where $\|\theta\|$ and $\|\eta\|$ are not necessarily equal. This is a more complicated problem than the equal SNR case studied in Section 2. For the general case, we adopt the following calibration:

$$\epsilon = n^{-\beta}, \qquad \frac{2\|\theta\|\|\eta\|}{\sqrt{\|\theta\|^2+\|\eta\|^2}} = \sqrt{2r\log n}, \qquad \frac{|\|\theta\|^2-\|\eta\|^2|}{\sqrt{\|\theta\|^2+\|\eta\|^2}} = \sqrt{2s\log n}. \tag{24}$$

With this calibration, we are interested in all (r, s) pairs in $(0,\infty)\times[0,\infty)$. Although there are other ways to parametrize $\|\theta\|$ and $\|\eta\|$, we find (24) convenient and interpretable. In

(24), r characterizes average signal strength and s quantifies the level of difference in SNRs of the two samples. When $s = 0$, (24) reduces to (13). With this natural reduction, all results in Section 2 can be obtained by setting $s = 0$ in results for the general case which we shall derive in this section. Furthermore, the following identities can be derived from (24):

$$\|\theta\|^2 + \|\eta\|^2 = 2(r+s)\log n, \tag{25}$$

$$\|\theta\|^2 \vee \|\eta\|^2 = (r+s+\sqrt{s}\sqrt{r+s})\log n, \tag{26}$$

$$\|\theta\|^2 \wedge \|\eta\|^2 = (r+s-\sqrt{s}\sqrt{r+s})\log n. \tag{27}$$

3.1. *A related sparse mixture detection problem.* With $\widetilde{X}_i \sim N(z_i\|\theta\|, 1)$ and $\widetilde{Y}_i \sim N(\sigma_i\|\eta\|, 1)$ as defined in (10), it is natural to consider

$$\frac{\|\eta\|\widetilde{X}_i - \|\theta\|\widetilde{Y}_i}{\sqrt{\|\theta\|^2+\|\eta\|^2}} \sim N\Big(\frac{\|\theta\|\|\eta\|(z_i-\sigma_i)}{\sqrt{\|\theta\|^2+\|\eta\|^2}}, 1\Big), \quad i \in [n]. \tag{28}$$

Moreover, to avoid information loss, we also consider the following complementary sequence:

$$\frac{\|\theta\|\widetilde{X}_i + \|\eta\|\widetilde{Y}_i}{\sqrt{\|\theta\|^2+\|\eta\|^2}} \sim N\Big(\frac{\|\theta\|^2 z_i + \|\eta\|^2\sigma_i}{\sqrt{\|\theta\|^2+\|\eta\|^2}}, 1\Big), \quad i \in [n]. \tag{29}$$

The sequences (28) and (29) are mutually independent since they are Gaussian and uncorrelated. Since $(\widetilde{X}_i, \widetilde{Y}_i)$ is sufficient for (z_i, σ_i) and has one-to-one correspondence with $(\frac{\|\eta\|\widetilde{X}_i - \|\theta\|\widetilde{Y}_i}{\sqrt{\|\theta\|^2+\|\eta\|^2}}, \frac{\|\theta\|\widetilde{X}_i + \|\eta\|\widetilde{Y}_i}{\sqrt{\|\theta\|^2+\|\eta\|^2}})$, there is no information loss.

Without loss of generality,[3] let us further assume $\|\theta\| \geq \|\eta\|$. We note that when $z_i = \sigma_i$, the two sequences have means 0 and $\pm\sqrt{\|\theta\|^2+\|\eta\|^2}$, respectively. When $z_i \neq \sigma_i$, they have means $\pm\frac{2\|\theta\|\|\eta\|}{\sqrt{\|\theta\|^2+\|\eta\|^2}}$ and $\pm\frac{|\|\theta\|^2-\|\eta\|^2|}{\sqrt{\|\theta\|^2+\|\eta\|^2}}$, respectively. Therefore, a natural corresponding sparse mixture detection problem is for $i \in [n]$,

$$\begin{aligned} H_0 : (U_i, V_i) &\overset{\text{iid}}{\sim} \frac{1}{2}N(0,1)\otimes N\big(-\sqrt{2(r+s)\log n}, 1\big) \\ &\quad + \frac{1}{2}N(0,1)\otimes N\big(\sqrt{2(r+s)\log n}, 1\big), \end{aligned} \tag{30}$$

$$\begin{aligned} H_1 : (U_i, V_i) &\overset{\text{iid}}{\sim} \frac{1-\epsilon}{2}N(0,1)\otimes N\big(-\sqrt{2(r+s)\log n}, 1\big) \\ &\quad + \frac{1-\epsilon}{2}N(0,1)\otimes N\big(\sqrt{2(r+s)\log n}, 1\big) \\ &\quad + \frac{\epsilon}{2}N(\sqrt{2r\log n}, 1)\otimes N(\sqrt{2s\log n}, 1) \\ &\quad + \frac{\epsilon}{2}N(-\sqrt{2r\log n}, 1)\otimes N(-\sqrt{2s\log n}, 1). \end{aligned} \tag{31}$$

When $s = 0$, the testing problem (30)–(31) reduces to (20)–(21).

Similar to Section 2, as a first step, we derive the detection boundaries of tests that use only $\{U_i\}_{1\leq i\leq n}$ or only $\{V_i\}_{1\leq i\leq n}$.

[3] We only use $\|\theta\| \geq \|\eta\|$ to motivate the testing problem (30)–(31). All the theorems in the paper hold with general θ and η that admit the calibration (24).

THEOREM 3.1. *Consider testing (30)–(31) with $\epsilon = n^{-\beta}$. Define*

$$\bar{\beta}^*(r,s) = \begin{cases} \frac{1}{2} + r - 2\sqrt{s}(\sqrt{r+s} - \sqrt{s}), & 3s > r \text{ and } (\sqrt{r+s} - \sqrt{s})^2 \le \frac{1}{4}, \\ \frac{1+r-s}{2}, & 3s \le r \text{ and } r+s \le 1, \\ r - 2(\sqrt{r+s} - \sqrt{s})(\sqrt{r+s} - 1), & r+s > 1 \text{ and } \frac{1}{4} < (\sqrt{r+s} - \sqrt{s})^2 \le 1, \\ 1, & (\sqrt{r+s} - \sqrt{s})^2 > 1. \end{cases}$$

For any fixed constant $\delta > 0$, we have the following two conclusions:

1. *When $\beta < \beta^*_{\mathrm{IDJ}}(r)$, the test that rejects when*

$$\sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|U_i|^2 > t\}} - n\mathbb{P}(\chi_1^2 > t)|}{\sqrt{n\mathbb{P}(\chi_1^2 > t)(1 - \mathbb{P}(\chi_1^2 > t))}} > \sqrt{2(1+\delta)\log\log n}$$

is consistent. When $\beta > \beta^*_{\mathrm{IDJ}}(r)$, no test that only uses $\{U_i\}_{1\le i\le n}$ is consistent.

2. *When $\beta < \bar{\beta}^*(r,s)$, the test that rejects when*

$$\sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{|V_i|^2 \le t\}} - n\mathbb{P}(\chi^2_{1,2(r+s)\log n} \le t)|}{\sqrt{n\mathbb{P}(\chi^2_{1,2(r+s)\log n} \le t)(1 - \mathbb{P}(\chi^2_{1,2(r+s)\log n} \le t))}} > \sqrt{2(1+\delta)\log\log n}$$

is consistent. When $\beta > \bar{\beta}^*(r,s)$, no test that only uses $\{V_i\}_{1\le i\le n}$ is consistent.

The first conclusion of Theorem 3.1 is obvious, since the marginal distributions of $\{U_i\}_{1\le i\le n}$ under (30) and (31) are exactly the same as those under (11) and (12), respectively. In contrast, the second conclusion shows an intricate behavior of the two-dimensional threshold function $\bar{\beta}^*(r,s)$. We note that $\bar{\beta}^*(r,s)$ can be viewed as an extension of $\bar{\beta}^*(r)$ defined in (18) in the sense that setting $s = 0$ in $\bar{\beta}^*(r,s)$ gives (18). The definition of $\bar{\beta}^*(r,s)$ involves four disjoint regions in the (r,s) domain $(0,\infty)\times[0,\infty)$. When $s = 0$, the second and the third cases become degenerate. Moreover, we also have the relation $\bar{\beta}^*(r,s) \le \bar{\beta}^*(r)$ for all $r, s > 0$, which suggests that the testing problem becomes harder as the gap between $\|\theta\|$ and $\|\eta\|$ gets larger. Last but not least, at each fixed $r > 0$, as $s \to \infty$ we have $\bar{\beta}^*(r,s) \to \frac{1}{2}$.

3.2. *Which event should we count?* Now let us try to solve the testing problem (30)–(31) by considering both $\{U_i\}_{1\le i\le n}$ and $\{V_i\}_{1\le i\le n}$. In order to derive the optimal detection boundary for (30)–(31) and also for the original problem (3), we need to first find the optimal test statistic. By Theorem 3.1, the detection boundary of either single sequence can be achieved by an appropriate HC-type test. For $\{U_i\}_{1\le i\le n}$, the test counts the number of large $|U_i|$'s by $\sum_{i=1}^n \mathbf{1}_{\{|U_i|^2>t\}}$, and for $\{V_i\}_{1\le i\le n}$ the corresponding test counts the number of small $|V_i|$'s by $\sum_{i=1}^n \mathbf{1}_{\{|V_i|^2\le t\}}$. These tests suggest that for testing (30)–(31) we should count the event that either $|U_i|$ is large or $|V_i|$ is small. When the SNRs are equal, we have used $\sum_{i=1}^n \mathbf{1}_{\{|U_i|-|V_i|>t\}}$ in Section 2.3 for this purpose. However, such an event may no longer be appropriate when $\|\theta\| \ne \|\eta\|$.

In order to find out the appropriate event to count, we present the following heuristic argument from a more general perspective. Let us consider the following abstract sparse mixture testing problem:

$$H_0 : W_1, \dots, W_n \overset{\text{iid}}{\sim} P \quad \text{vs.} \tag{32}$$

$$H_1 : W_1, \dots, W_n \overset{\text{iid}}{\sim} (1-\epsilon)P + \epsilon Q, \tag{33}$$

where $\epsilon = n^{-\beta}$ for some constant $\beta \in (0, 1)$. Then the general HC-type test statistic can be written as

$$\sup_{A \in \mathcal{A}} \frac{|\sum_{i=1}^n \mathbf{1}_{\{W_i \in A\}} - nP(A)|}{\sqrt{nP(A)(1 - P(A))}}, \tag{34}$$

where $\mathcal{A}$ is some collection of events. As we shall show, the reason to take the supremum over the collection $\mathcal{A}$ is mostly for the sake of adaptation. When one has knowledge of P and Q, let us consider first the statistic

$$T_n(A) = \frac{\sum_{i=1}^n \mathbf{1}_{\{W_i \in A\}} - nP(A)}{\sqrt{nP(A)(1 - P(A))}}.$$

With the calibration $\epsilon = n^{-\beta}$, by our detailed calculation in Supplementary Material (Section A.5), a sufficient condition for a test of the form $\mathbf{1}_{\{|T_n(A)| > c_n\}}$ to be consistent with some slowly diverging sequence c_n is that

$$\beta < \frac{1}{2} + \frac{\log Q(A)}{\log n} + \frac{1}{2} \min\left(1, \frac{\log \frac{1}{P(A)}}{\log n}\right). \tag{35}$$

To maximize the detection region, we shall consider some event A that makes the right-hand side of (35) as large as possible. Since the right-hand side of (35) is increasing in $Q(A)$ and decreasing in $P(A)$, the maximum is achieved by $A = \{\frac{dQ}{dP}(W) > t\}$ for some appropriate choice of t according to the Neyman–Pearson lemma. This fact naturally motivates the choice of

$$\mathcal{A} = \left\{\left\{\frac{dQ}{dP}(W) > t\right\} : t > 0\right\}$$

in (34), which results in the following HC-type statistic

$$\sup_{t>0} \frac{|\sum_{i=1}^n \mathbf{1}_{\{(dQ/dP)(W_i) > t\}} - nP((dQ/dP)(W) > t)|}{\sqrt{nP((dQ/dP)(W) > t)P((dQ/dP)(W) \le t)}}. \tag{36}$$

3.3. *Likelihood ratio approximation.* The heuristic argument in Section 3.2 suggests that we use the statistic $\sum_{i=1}^n \mathbf{1}_{\{(dQ/dP)(W_i) > t\}}$. We specify P and Q to the setting of (30)–(31) to obtain that

$$\frac{dQ}{dP}(W_i) = \frac{q(U_i, V_i)}{p(U_i, V_i)},$$

where

$$\begin{aligned} p(u, v) = {} & \frac{1}{2}\phi(u)\phi\big(v - \sqrt{2(r + s)\log n}\big) \\ & + \frac{1}{2}\phi(u)\phi\big(v + \sqrt{2(r + s)\log n}\big), \end{aligned} \tag{37}$$

$$\begin{aligned} q(u, v) = {} & \frac{1}{2}\phi(u - \sqrt{2r\log n})\phi(v - \sqrt{2s\log n}) \\ & + \frac{1}{2}\phi(u + \sqrt{2r\log n})\phi(v + \sqrt{2s\log n}). \end{aligned} \tag{38}$$

Here, $\phi(\cdot)$ is the probability density function of $N(0, 1)$. Note that p and q scale with n though the dependence is not explicit in the notation. The following key lemma simplifies the calculation of the likelihood ratio statistic.

LEMMA 3.1. *For $p(u,v)$ and $q(u,v)$ defined above, we have*

$$\sup_{r\geq 0,s>0}\sup_{u,v\in\mathbb{R}}\left|\log\frac{q(u,v)}{p(u,v)}-\sqrt{2\log n}\big(|\sqrt{r}u+\sqrt{s}v|-\sqrt{r+s}|v|\big)\right|\leq\log 2.$$

By Lemma 3.1, $\sqrt{2\log n}(|\sqrt{r}u+\sqrt{s}v|-\sqrt{r+s}|v|)$ is the leading term of $\log\frac{q(u,v)}{p(u,v)}$ as $n\to\infty$. Therefore, from an asymptotic viewpoint, we could simply focus on the sequence

$$\big\{|\sqrt{r}U_i+\sqrt{s}V_i|-\sqrt{r+s}|V_i|\big\}_{1\leq i\leq n},$$

which combines the information in $\{U_i\}_{1\leq i\leq n}$ and $\{V_i\}_{1\leq i\leq n}$. When $s=0$, it reduces to $\{\sqrt{r}(|U_i|-|V_i|)\}_{1\leq i\leq n}$, which further justifies the optimality of the test (22) when $\|\theta\|=\|\eta\|$. As $s\to\infty$, we have $\sqrt{r+s}-\sqrt{s}=\frac{r}{\sqrt{r+s}+\sqrt{s}}\to 0$, and it can shown that the sequence becomes $\{\sqrt{r}U_i\,\mathsf{sign}(V_i)\}_{1\leq i\leq n}$. For any real number a, $\mathsf{sign}(a)=a/|a|$ when $a\neq 0$ and $\mathsf{sign}(0)=0$. In other word, for extremely large values of s only the sign information of the sequence with weaker SNR matters.

3.4. *The three-dimensional phase diagram.* We now move on to determine detection boundaries for (30)–(31) and for (3) in general.

Sparse mixture detection. Consider the sparse mixture detection problem (30)–(31) first. Inspired by Lemma 3.1, we consider the following HC-type test that rejects (30) when

$$\sup_{t\in\mathbb{R}}\frac{|\sum_{i=1}^n\mathbf{1}_{\{|\sqrt{r}U_i+\sqrt{s}V_i|-\sqrt{r+s}|V_i|>t\}}-nS_{(r,s)}(t)|}{\sqrt{nS_{(r,s)}(t)(1-S_{(r,s)}(t))}}>\sqrt{2(1+\delta)\log\log n},\tag{39}$$

where $\delta>0$ is some arbitrary fixed constant. Here, $S_{(r,s)}(t)$ is the survival function of $|\sqrt{r}U_i+\sqrt{s}V_i|-\sqrt{r+s}|V_i|$ under the null distribution, defined by

$$S_{(r,s)}(t)=\mathbb{P}_{H_0}\big(|\sqrt{r}U+\sqrt{s}V|-\sqrt{r+s}|V|>t\big),\tag{40}$$

where H_0 is defined in (30). By Lemma 3.1 and our heuristic argument in Section 3.2, the test statistic in (39) is asymptotically equivalent to (36). Indeed, the test with rejection region (39) achieves the optimal detection boundary of the testing problem (30)–(31), which is summarized as the following theorem.

THEOREM 3.2. *Consider testing (30)–(31) with $\epsilon=n^{-\beta}$. Define*

$$\beta^*(r,s)=\begin{cases}\frac{1}{2}+2(r+s-\sqrt{s}\sqrt{r+s}), & 3s>r \text{ and } r+s-\sqrt{s}\sqrt{r+s}\leq\frac{1}{8},\\ \frac{1}{2}(1+3r-s), & 3s\leq r \text{ and } 5r+s\leq 1,\\ 2\sqrt{r}\sqrt{1-r-s}, & 5r+s>1,\ \frac{1}{8}<r+s-\sqrt{s}\sqrt{r+s}\leq\frac{1}{2},\\ & \text{and } 2(1-r-s)(r+s-\sqrt{s}\sqrt{r+s})>r,\\ \Big[2\sqrt{2(r+s-\sqrt{s}\sqrt{r+s})} & 5r+s>1,\ \frac{1}{8}<r+s-\sqrt{s}\sqrt{r+s}\leq\frac{1}{2},\\ -2(r+s-\sqrt{s}\sqrt{r+s})\Big], & \text{and } 2(1-r-s)(r+s-\sqrt{s}\sqrt{r+s})\leq r,\\ 1, & r+s-\sqrt{s}\sqrt{r+s}>\frac{1}{2}.\end{cases}$$

When $\beta<\beta^*(r,s)$, the test with rejection region (39) is consistent. When $\beta>\beta^*(r,s)$, no test is consistent.

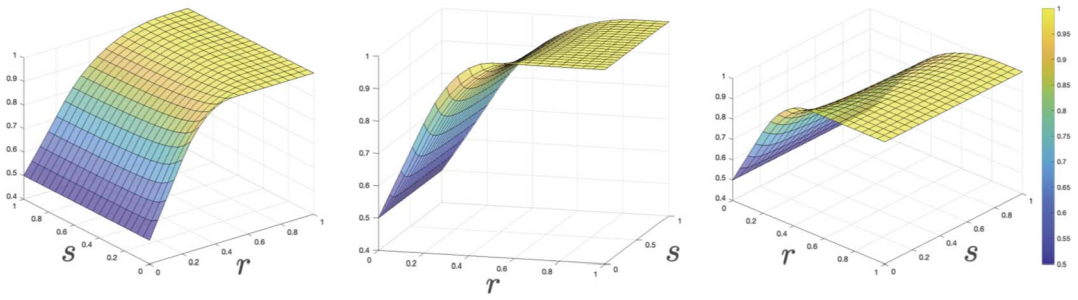

FIG. 2. *3D plot of the detection boundary $\beta^*(r, s)$ in Theorems 3.2 and 3.3.*

Testing equivalence of clustering. Turn to the original testing problem (3). Note that the two sequences (28) and (29) play the same roles as $\{U_i\}_{1\le i\le n}$ and $\{V_i\}_{1\le i\le n}$ do in sparse mixture detection. In view of the parameterization in (24)–(27), we define

$$C^-(X_i, Y_i, \theta, \eta) = |\theta^T X_i - \eta^T Y_i| - |\theta^T X_i + \eta^T Y_i|, \tag{41}$$

$$C^+(X_i, Y_i, \theta, \eta) = |\theta^T X_i + \eta^T Y_i| - |\theta^T X_i - \eta^T Y_i|. \tag{42}$$

For testing (3), we need both $\{C^-(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ and $\{C^+(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ to accommodate the possibility of label switching. Then the HC-type statistics for testing (3) can be defined as

$$\dot{T}_n^- = \sup_{t\in\mathbb{R}} \frac{|\sum_{i=1}^n \mathbf{1}_{\{C^-(X_i,Y_i,\theta,\eta)>t\sqrt{2\log n}\}} - nS_{(r,s)}(t)|}{\sqrt{nS_{(r,s)}(t)(1-S_{(r,s)}(t))}}, \tag{43}$$

$$\dot{T}_n^+ = \sup_{t\in\mathbb{R}} \frac{|\sum_{i=1}^n \mathbf{1}_{\{C^+(X_i,Y_i,\theta,\eta)>t\sqrt{2\log n}\}} - nS_{(r,s)}(t)|}{\sqrt{nS_{(r,s)}(t)(1-S_{(r,s)}(t))}}. \tag{44}$$

They lead to the test

$$\dot{\psi} = \mathbf{1}_{\{\dot{T}_n^- \wedge \dot{T}_n^+ > \sqrt{2(1+\delta)\log\log n}\}}, \tag{45}$$

for some arbitrary fixed constant $\delta > 0$.

THEOREM 3.3. *For testing (3) with calibration (24), the test (45) satisfies $\lim_{n\to\infty} R_n(\dot{\psi}, \theta, \eta, \epsilon) = 0$ as long as $\beta < \beta^*(r, s)$. Moreover, when $\beta > \beta^*(r, s)$, no test is consistent.*

With Theorem 3.3, we fully characterize the detection boundary of the testing problem (3) by the function $\beta^*(r, s)$. To help understanding the behavior of $\beta^*(r, s)$, Figure 2 demonstrates its 3D plot from various angles. In addition, we plot the five regions that divide the domain of $\beta^*(r, s)$, that is, $(0, \infty) \times [0, \infty)$, on the left panel of Figure 3. Furthermore, we fix s and study the behavior of the function $\beta_s^*(r) = \beta^*(r, s)$ as a function of r at some fixed s value. We start with $s = 0$. In this case, the problem reduces to the equal SNR situation, and we are able to recover $\beta_s^*(r) = \beta^*(r)$, where $\beta^*(r)$ is defined in (9). For any fixed $s \in (0, \frac{1}{16})$,

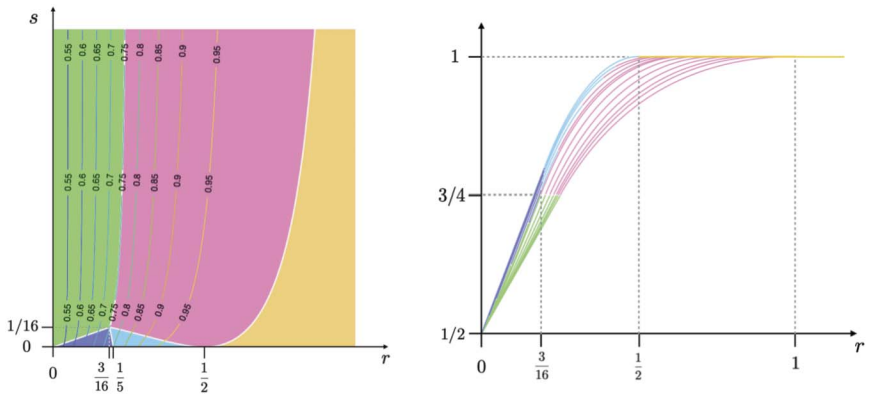

FIG. 3. *The five regions of the (r,s)-plane with the contour of $\beta^*(r,s)$ (left panel). The detection boundaries $\beta_s^*(r)=\beta^*(r,s)$ with s fixed (right panel). The curve moves to the right as the fixed value of s increases. The five colors of the two plots correspond to the five regions of $\beta^*(r,s)$ in the order of green, blue, cyan, magenta and yellow.*

the definition of $\beta_s^*(r)$ involves all the five areas in the left panel of Figure 3, and we have

$$\beta_s^*(r)=\begin{cases}\dfrac{1}{2}+2(r+s-\sqrt{s}\sqrt{r+s}),\\ \qquad 0<r<3s,\\ \dfrac{1}{2}(1+3r-s),\\ \qquad 3s\le r<\dfrac{1-s}{5},\\ 2\sqrt{r}\sqrt{1-r-s},\\ \qquad \dfrac{1-s}{5}\le r<\mathsf{root}(s),\\ 2\sqrt{2(r+s-\sqrt{s}\sqrt{r+s})}-2(r+s-\sqrt{s}\sqrt{r+s}),\\ \qquad \mathsf{root}(s)\le r<\left(\sqrt{\dfrac{1}{2}+\dfrac{s}{4}}+\sqrt{\dfrac{s}{4}}\right)^2-s,\\ 1,\quad r>\left(\sqrt{\dfrac{1}{2}+\dfrac{s}{4}}+\sqrt{\dfrac{s}{4}}\right)^2-s.\end{cases}$$

Here, $r=\mathsf{root}(s)$ is a root of the equation $2(1-r-s)(r+s-\sqrt{s}\sqrt{r+s})=r$. We note that when $s\in(0,\frac{1}{16})$, the equation has a unique real root between $\frac{3}{16}$ and $\frac{1}{2}$. Next, we consider any fixed $s\ge\frac{1}{16}$. In this case, two regions become degenerate, and we have

$$\beta_s^*(r)=\begin{cases}\dfrac{1}{2}+2(r+s-\sqrt{s}\sqrt{r+s}),\\ \qquad 0<r<\left(\sqrt{\dfrac{1}{8}+\dfrac{s}{4}}+\sqrt{\dfrac{s}{4}}\right)^2-s,\\ \left[2\sqrt{2(r+s-\sqrt{s}\sqrt{r+s})}-2(r+s-\sqrt{s}\sqrt{r+s})\right],\\ \qquad \left(\sqrt{\dfrac{1}{8}+\dfrac{s}{4}}+\sqrt{\dfrac{s}{4}}\right)^2-s\le r<\left(\sqrt{\dfrac{1}{2}+\dfrac{s}{4}}+\sqrt{\dfrac{s}{4}}\right)^2-s,\\ 1,\quad r\ge\left(\sqrt{\dfrac{1}{2}+\dfrac{s}{4}}+\sqrt{\dfrac{s}{4}}\right)^2-s.\end{cases}$$

Last but not least, we would like to point out that when $s = \infty$, we obtain the Ingster–Donoho–Jin threshold $\beta_s^*(r) = \beta_{\mathrm{IDJ}}^*(r)$. This agrees with the intuition that the sequence $\{V_i\}_{1\le i\le n}$ is asymptotically noninformative for the testing problem (30)–(31) as $s \to \infty$. The functions $\{\beta_s^*(r)\}$ with various fixed values of s are shown on the right panel of Figure 3, and all the curves are sandwiched between $\beta^*(r)$ and $\beta_{\mathrm{IDJ}}^*(r)$ (also see Figure 1). It is clear that for a fixed s, a larger r makes the testing problem easier. On the other hand, increasing s always makes the problem harder in the sense that $\beta_{s_1}^*(r) \ge \beta_{s_2}^*(r)$ for all $r > 0$ when $s_1 < s_2$.

4. Testing for exact equality. The most stringent version of the testing problem (3) is whether or not the two clustering structures are exactly equal. This can be formulated as the following hypothesis testing problem:

$$H_0 : \ell(z, \sigma) = 0 \quad \text{versus} \quad H_1 : \ell(z, \sigma) > 0. \tag{46}$$

Since the distance function $\ell(z, \sigma)$ only takes value in the set $\{0, n^{-1}, 2n^{-1}, \ldots\}$, the alternative hypothesis of (46) is equivalent to $\ell(z, \sigma) \ge n^{-1}$. Therefore, the testing problem (46) is a special case of (3) with $\beta = 1$. However, Theorem 3.3 only covers $\beta < 1$. Since the lower bound proof of Theorem 3.3 is based on the connection between (3) and (30)–(31), which requires $n\epsilon \to \infty$, the case of $\beta = 1$ is thus excluded.

In this section, we rigorously study testing problem (46). Given a testing procedure ψ, we define its worst-case testing error by

$$R_n^{\mathrm{exact}}(\psi, \theta, \eta) = \sup_{(z,\sigma)\in\mathcal{F}_n^0} P_{(\theta,\eta,z,\sigma)}^{(n)}\psi + \sup_{(z,\sigma)\in\mathcal{F}_n^1(n^{-1})} P_{(\theta,\eta,z,\sigma)}^{(n)}(1-\psi).$$

The minimax testing error is then defined by $R_n^{\mathrm{exact}}(\theta, \eta) = \inf_\psi R_n^{\mathrm{exact}}(\psi, \theta, \eta)$. Our first result gives a necessary and sufficient condition for the existence of a consistent test.

THEOREM 4.1. *Consider testing* (46) *with calibration* (24). *When* $r + s - \sqrt{s}\sqrt{r+s} < \frac{1}{2}$, *we have* $\liminf_{n\to\infty} R_n^{\mathrm{exact}}(\theta, \eta) > 0$. *When* $r + s - \sqrt{s}\sqrt{r+s} > \frac{1}{2}$, *the HC-type test* $\dot\psi$ *defined in* (45) *satisfies* $\lim_{n\to\infty} R_n^{\mathrm{exact}}(\dot\psi, \theta, \eta) = 0$.

Theorem 4.1 shows that whether $r + s - \sqrt{s}\sqrt{r+s}$ is above or below $\frac{1}{2}$ determines the existence of a consistent test for (46). This is compatible with the last regime of the threshold function $\beta^*(r, s)$. See the yellow area in the left panel of Figure 3. Given the relation (27), it is required that both $\|\theta\|^2$ and $\|\eta\|^2$ are greater than $\frac{1}{2}\log n$ for separating the null and the alternative hypotheses. Moreover, the same optimal HC-type test in Theorem 3.3 continues to work for testing exact equality. On a related note, HC tests also work for the boundary case of $\beta = 1$ in the classical settings (11)–(12) and that in [11].

In addition to the HC-type test, we introduce a Bonferroni-type test that is also optimal for (46). To this end, define

$$t^*(r, s) = \begin{cases} \sqrt{r(1-r-s)}, \\ \quad 2(r+s)(r+s+\sqrt{s}\sqrt{r+s}) \le r, \\ \sqrt{2(r+s-\sqrt{s}\sqrt{r+s})} - (r+s-\sqrt{s}\sqrt{r+s}), \\ \quad 2(r+s)(r+s+\sqrt{s}\sqrt{r+s}) > r. \end{cases}$$

The following lemma shows that it characterizes the largest element of the sequence $\{|\sqrt{r}U_i + \sqrt{s}V_i| - \sqrt{r+s}|V_i|\}_{1\le i\le n}$ under the null distribution.

LEMMA 4.1. *Suppose* $\{(U_i, V_i)\}_{1\le i\le n}$ *are generated according to* (30). *Then we have*

$$\frac{\max_{1\le i\le n}(|\sqrt{r}U_i + \sqrt{s}V_i| - \sqrt{r+s}|V_i|)}{\sqrt{2\log n}} \to t^*(r, s) \quad \textit{in probability.}$$

Lemma 4.1 shows that the largest element of the sequence $\{|\sqrt{r}U_i + \sqrt{s}V_i| - \sqrt{r+s}|V_i|\}_{1\le i\le n}$ is asymptotically $t^*(r,s)\sqrt{2\log n}$ under H_0. It is therefore natural to reject H_0 when the random variable $\max_{1\le i\le n}(|\sqrt{r}U_i + \sqrt{s}V_i| - \sqrt{r+s}|V_i|)$ is larger than $t^*(r,s)\sqrt{2\log n}$. In view of the connection between sparse mixture detection and testing clustering equivalence, applying the result to the sequences $\{C^-(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ and $\{C^+(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ in (41)–(42), we obtain the following testing procedure:

$$\psi_{\text{Bonferroni}} = \mathbf{1}_{\{(\max_{1\le i\le n} C^-(X_i,Y_i,\theta,\eta))\wedge(\max_{1\le i\le n} C^+(X_i,Y_i,\theta,\eta))>2t^*(r,s)\log n\}}. \tag{47}$$

THEOREM 4.2. *Consider testing* (46) *with calibration* (24). *When* $r+s-\sqrt{s}\sqrt{r+s} > \frac{1}{2}$, *we have* $\lim_{n\to\infty} R_n^{\text{exact}}(\psi_{\text{Bonferroni}}, \theta, \eta) = 0$.

5. Adaptive tests. In this section, we investigate how to test (3) and (46) when the model parameters $\theta \in \mathbb{R}^p$ and $\eta \in \mathbb{R}^q$ are unknown. We will show that both the HC-type test and the Bonferroni test can be modified into adaptive procedures, as long as some mild growth rate conditions on the ambient dimensions p and q are satisfied.

5.1. *Adaptive Bonferroni test.* We start with testing (46). When designing the adaptive procedures, we adopt a random data splitting scheme. We first draw $d_1, \dots, d_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\frac{1}{2})$, and then define $\mathcal{D}_0 = \{i \in [n] : d_i = 0\}$ and $\mathcal{D}_1 = \{i \in [n] : d_i = 1\}$. Then $\{\mathcal{D}_0, \mathcal{D}_1\}$ forms a random partition of $[n]$. Given some algorithms $\widehat{\theta}(\cdot)$ and $\widehat{\eta}(\cdot)$ that compute estimators of θ and η, we define $\widehat{\theta}^{(m)} = \widehat{\theta}(\{(X_i, Y_i)\}_{i\in\mathcal{D}_m})$ and $\widehat{\eta}^{(m)} = \widehat{\eta}(\{(X_i, Y_i)\}_{i\in\mathcal{D}_m})$ for $m = 0$ and 1. For $m = 0$ and 1, by plugging $\widehat{\theta}^{(m)}$ and $\widehat{\eta}^{(m)}$ into the relation (24), we obtain $\widehat{r}^{(m)}$ and $\widehat{s}^{(m)}$.

Given these estimators of θ and η, we can modify (47) into an adaptive procedure. We replace $\max_{1\le i\le n} C^-(X_i, Y_i, \theta, \eta)$ and $\max_{1\le i\le n} C^+(X_i, Y_i, \theta, \eta)$ by

$$\widehat{C}_m^{\pm} = \max_{i\in\mathcal{D}_m} C^{\pm}\big(X_i, Y_i, \widehat{\theta}^{(1-m)}, \widehat{\eta}^{(1-m)}\big), \quad m = 0, 1.$$

Then we combine these statistics by

$$\widehat{C}^- = \begin{cases} \widehat{C}_0^- \vee \widehat{C}_1^-, & \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|\le 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|\le 1\}} + \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|> 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|> 1\}} = 1, \\ \widehat{C}_0^- \vee \widehat{C}_1^+, & \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|> 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|\le 1\}} + \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|\le 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|> 1\}} = 1, \end{cases} \tag{48}$$

$$\widehat{C}^+ = \begin{cases} \widehat{C}_0^+ \vee \widehat{C}_1^+, & \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|\le 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|\le 1\}} + \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|> 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|> 1\}} = 1, \\ \widehat{C}_0^+ \vee \widehat{C}_1^-, & \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|> 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|\le 1\}} + \mathbf{1}_{\{\|\widehat{\theta}^{(0)}-\widehat{\theta}^{(1)}\|\le 1, \|\widehat{\eta}^{(0)}-\widehat{\eta}^{(1)}\|> 1\}} = 1. \end{cases} \tag{49}$$

The adaptive Bonferroni test is defined by

$$\psi_{\text{ada-Bonferroni}} = \mathbf{1}_{\{\widehat{C}^-\wedge\widehat{C}^+ > 2(1+\frac{1}{\sqrt{\log n}})\widehat{t}\log n\}},$$

where

$$\widehat{t} = \frac{t^*(\widehat{r}^{(0)}, \widehat{s}^{(0)}) + t^*(\widehat{r}^{(1)}, \widehat{s}^{(1)})}{2}.$$

The additional factor $(1+\frac{1}{\sqrt{\log n}})$ is to accommodate the error caused by estimators of θ and η. Before writing down a theorem that gives the desired theoretical guarantee for $\psi_{\text{ada-Bonferroni}}$, let us define the loss functions

$$L(\widehat{\theta}, \theta) = \|\widehat{\theta} - \theta\| \wedge \|\widehat{\theta} + \theta\|, \qquad L(\widehat{\eta}, \eta) = \|\widehat{\eta} - \eta\| \wedge \|\widehat{\eta} + \eta\|.$$

Though θ and η can be of different dimensions, we use the same notation $L(\cdot, \cdot)$ for the two loss functions for simplicity.

THEOREM 5.1. *Consider testing* (46) *with the calibration* (24). *Assume that there is some constant* $\gamma > 0$, *such that*

$$\lim_{n\to\infty} \sup_{\substack{z\in\{-1,1\}^n \\ \sigma\in\{-1,1\}^n}} P^{(n)}_{(\theta,\eta,z,\sigma)}\big(L(\widehat{\theta}^{(0)},\theta) \vee L(\widehat{\theta}^{(1)},\theta) \vee L(\widehat{\eta}^{(0)},\eta) \vee L(\widehat{\eta}^{(1)},\eta) > n^{-\gamma}\big) = 0. \tag{50}$$

When $r + s - \sqrt{s}\sqrt{r+s} > \frac{1}{2}$, *we have* $\lim_{n\to\infty} R_n^{\text{exact}}(\psi_{\text{ada-Bonferroni}}, \theta, \eta) = 0$.

The condition (50) may seem abstract at first sight. Later in Section 5.3, we shall give concrete estimators so that it is met when p and q grow sublinearly with n. For full details, see Corollary 5.1.

5.2. *Adaptive HC-type test.* To modify (45) into an adaptive procedure is more involved. This is due to the fact that we not only need to approximate the statistics $\{C^-(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ and $\{C^+(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$, but also need to estimate the survival function $S_{(r,s)}(t)$ defined in (40). Our proposed strategy starts with a random data splitting step. This time we split the data into three parts instead of two. Draw $d_1, \dots, d_n \overset{\text{iid}}{\sim}$ Discrete Uniform($\{0, 1, 2\}$), and then define $\mathcal{D}_m = \{i \in [n] : d_i = m\}$ for $m = 0, 1, 2$.

Given some algorithms $\widehat{\theta}(\cdot)$ and $\widehat{\eta}(\cdot)$ that compute estimators of θ and η, we first define $\widehat{\theta} = \widehat{\theta}(\{(X_i, Y_i)\}_{i\in\mathcal{D}_0})$ and $\widehat{\eta} = \widehat{\eta}(\{(X_i, Y_i)\}_{i\in\mathcal{D}_0})$. We then use $\widehat{\theta}$ and $\widehat{\eta}$ for as projection directions and compute $\widehat{X}_i = \widehat{\theta}^T X_i/\|\widehat{\theta}\|$ and $\widehat{Y}_i = \widehat{\eta}^T Y_i/\|\widehat{\eta}\|$ for all $i \in \mathcal{D}_1 \cup \mathcal{D}_2$. Note that conditioning on $\{d_i\}_{1\le i\le n}$ and $\{(X_i, Y_i)\}_{i\in\mathcal{D}_0}$, $\widehat{X}_i$ and $\widehat{Y}_i$ are distributed according to $N(z_i a, 1)$ and $N(\sigma_i b, 1)$, respectively, where $a = \widehat{\theta}^T\theta/\|\widehat{\theta}\|$ and $b = \widehat{\eta}^T\eta/\|\widehat{\eta}\|$. Given the projected data, we will use those in $\mathcal{D}_1$ to estimate the one-dimensional parameters $|a|$ and $|b|$, and those in $\mathcal{D}_2$ to construct the test statistic. Define

$$\widehat{a} = \Big(\frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1} X_i^2 - 1\Big)_+^{1/2} \quad \text{and} \quad \widehat{b} = \Big(\frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1} Y_i^2 - 1\Big)_+^{1/2}.$$

With $\widehat{a}$ and $\widehat{b}$, we define

$$\widehat{r} = \frac{(2|\widehat{a}||\widehat{b}|)^2}{(2\log n)(\widehat{a}^2 + \widehat{b}^2)} \quad \text{and} \quad \widehat{s} = \frac{|\widehat{a}^2 - \widehat{b}^2|^2}{(2\log n)(\widehat{a}^2 + \widehat{b}^2)}.$$

Then the adaptive HC-type statistics are

$$\widehat{T}_n^{\pm} = \sup_{|t|\le \log n} \frac{|\sum_{i\in\mathcal{D}_2} \mathbf{1}_{\{C^{\pm}(\widehat{X}_i,\widehat{Y}_i,\widehat{a},\widehat{b})>t\sqrt{2\log n}\}} - |\mathcal{D}_2|S_{(\widehat{r},\widehat{s})}(t)|}{\sqrt{|\mathcal{D}_2|S_{(\widehat{r},\widehat{s})}(t)}}.$$

This leads to the adaptive test

$$\widehat{\psi}_{\text{ada-HC}} = \mathbf{1}_{\{\widehat{T}_n^- \wedge \widehat{T}_n^+ > (\log n)^3\}}.$$

Compared with (43) and (44), the adaptive versions $\widehat{T}_n^-$ and $\widehat{T}_n^+$ restrict the supremum to the range $|t| \le \log n$ and does not have an estimator of $1 - S_{(r,s)}(t)$ in the denominator. Moreover, the test uses the threshold $(\log n)^3$ instead of the smaller $\sqrt{2(1+\delta)\log\log n}$. These changes are adopted to accommodate the additional errors caused by estimating the unknown parameters and to avoid extra technicalities.

THEOREM 5.2. *Consider testing* (3) *with calibration* (24). *Assume that there is some constant* $\gamma > 0$, *such that*

$$\lim_{n\to\infty} \sup_{\substack{z\in\{-1,1\}^n \\ \sigma\in\{-1,1\}^n}} P^{(n)}_{(\theta,\eta,z,\sigma)}\big(L(\widehat{\theta},\theta) \vee L(\widehat{\eta},\eta) > n^{-\gamma}\big) = 0. \tag{51}$$

When $\beta < \beta^*(r, s)$, *we have* $\lim_{n\to\infty} R_n(\psi_{\text{ada-HC}}, \theta, \eta) = 0$.

For estimators such that condition (51) can be fulfilled, see Section 5.3.

Randomly splitting the data into three parts is needed for technical details in the proofs. In order for our proof to go through, we need to approximate the statistics $\{C^-(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ and $\{C^+(X_i, Y_i, \theta, \eta)\}_{1\le i\le n}$ and the survival function $S_{(r,s)}(t)$ at different levels of accuracy. In particular, we require that the estimation error of $S_{(r,s)}(t)$ to be at most $\frac{(\log n)^{O(1)}}{\sqrt{n}}$, independent of the dimensions p and q. Therefore, in addition to the two-part data splitting strategy used in building the adaptive Bonferroni test, we need to devote a separate part to estimate projection directions of two one-dimensional subspaces.

REMARK 5.1. One may wonder whether the adaptive HC-type test can also achieve the optimal detection boundary for the problem (46). The answer would be no for the current definition of $\psi_{\text{ada-HC}}$, because under H_1 in (46) with a nontrivial probability, $\mathcal{D}_2$ does not contain the coordinate that has the signal. However, a modification of $\psi_{\text{ada-HC}}$ can resolve this issue. The modification requires rotating the roles of the three parts $\mathcal{D}_0$, $\mathcal{D}_1$ and $\mathcal{D}_2$, and then we can define analogous versions of the HC-type statistics $\widehat{T}_n^-$ and $\widehat{T}_n^+$ on $\mathcal{D}_0$ and $\mathcal{D}_1$. These statistics can be combined in a similar way to (48) and (49). We omit the details.

Computation. We now discuss computation of $\widehat{\psi}_{\text{ada-HC}}$. Note that both $\widehat{T}_n^-$ and $\widehat{T}_n^+$ can be computed efficiently using the p-value interpretation of the HC statistic in [11]. In the ideal situation where θ and η are known, the two sets of p-values are $\{S_{(r,s)}(\frac{C^-(X_i,Y_i,\theta,\eta)}{\sqrt{2\log n}})\}_{1\le i\le n}$ and $\{S_{(r,s)}(\frac{C^+(X_i,Y_i,\theta,\eta)}{\sqrt{2\log n}})\}_{1\le i\le n}$, which are involved in the computation of the test (45). When θ and η are unknown, the following proposition suggests a similar computation strategy.

PROPOSITION 5.1. *Define* $\widehat{p}_i^- = S_{(\widehat{r},\widehat{s})}(\frac{C^-(\widehat{X}_i,\widehat{Y}_i,\widehat{a},\widehat{b})}{\sqrt{2\log n}})$ *and* $\widehat{p}_i^+ = S_{(\widehat{r},\widehat{s})}(\frac{C^+(\widehat{X}_i,\widehat{Y}_i,\widehat{a},\widehat{b})}{\sqrt{2\log n}})$ *for* $i \in \mathcal{D}_2$. *Then, with probability tending to* 1, *we have*

$$\widehat{T}_n^- = \max_{1\le i\le|\mathcal{D}_2|} \sqrt{|\mathcal{D}_2|}\left|\frac{i}{|\mathcal{D}_2|} - \widehat{p}^-_{(i,\mathcal{D}_2)}\right| \Big/ \sqrt{\widehat{p}^-_{(i,\mathcal{D}_2)}}, \tag{52}$$

$$\widehat{T}_n^+ = \max_{1\le i\le|\mathcal{D}_2|} \sqrt{|\mathcal{D}_2|}\left|\frac{i}{|\mathcal{D}_2|} - \widehat{p}^+_{(i,\mathcal{D}_2)}\right| \Big/ \sqrt{\widehat{p}^+_{(i,\mathcal{D}_2)}}, \tag{53}$$

where the subscript $(i, \mathcal{D}_2)$ *indicates the* i*th order statistic within the set* $\mathcal{D}_2$.

The statistics $\widehat{p}_i^-$ and $\widehat{p}_i^+$ can be regarded as estimators of p-values, which is a useful interpretation of $\widehat{\psi}_{\text{ada-HC}}$. Since the formulas (52) and (53) hold with high probability, $\lim_{n\to\infty} R_n(\psi_{\text{ada-HC}}, \theta, \eta) = 0$ will continue to hold when $\beta < \beta^*(r, s)$ if (52) and (53) are used in the computation of $\widehat{\psi}_{\text{ada-HC}}$.

5.3. *Parameter estimation*. We close this section by presenting a simple estimator for θ and η. Since we have that $X_i \sim N(z_i\theta, I_p)$, the empirical second moment $\frac{1}{n}\sum_{i=1}^n X_i X_i^T$ is a consistent estimator of the population counterpart $\theta\theta^T + I_p$ when $p \ll n$. Apply eigenvalue decomposition and we get $\frac{1}{n}\sum_{i=1}^n X_i X_i^T = \sum_{j=1}^p \widehat{\lambda}_j \widehat{u}_j \widehat{u}_j^T$, and then a natural estimator for θ is $\widehat{\theta} = (\widehat{\lambda}_1 - 1)^{1/2}\widehat{u}_1$. This simple estimator enjoys the following property.

PROPOSITION 5.2. *Consider independent observations* $X_1, \dots, X_n \sim N(z_i\theta, I_p)$ *with some* $z_i \in \{-1, 1\}$ *for all* $i \in [n]$. *Assume* $p \le n$, *and then there exist universal constants* $C, C' > 0$, *such that* $L(\widehat{\theta}, \theta) = \|\widehat{\theta} - \theta\| \wedge \|\widehat{\theta} + \theta\| \le C\sqrt{p/n}$, *with probability at least* $1 - e^{-C'p}$ *uniformly over all* $z \in \{-1, 1\}^n$ *and all* $\theta \in \mathbb{R}^p$ *that satisfies* $\|\theta\| \ge 1$.

COROLLARY 5.1. *Consider the calibration* (24). *Suppose* $p \vee q < n^{1-\delta}$ *for some constant* $\delta \in (0,1)$, *then there exists some constant* $\gamma > 0$ *depending on* δ, *such that the conditions* (50) *and* (51) *hold.*

Combining Theorem 5.1, Theorem 5.2 and Corollary 5.1, we conclude that the optimal detection boundaries of the testing problems (3) and (46) can be achieved adaptively without the knowledge of (θ, η), as long as the dimensions do not grow too fast in the sense that $p \vee q < n^{1-\delta}$. In a more general setting, one may have $X_i \sim N(z_i\theta, \sigma^2 I_p)$ with both θ and σ^2 unknown. The proposed algorithm still works for estimating θ. To estimate σ^2, one can use $\widehat{\sigma}^2 = \frac{1}{p}\operatorname{Tr}(\frac{1}{n}\sum_{i=1}^n X_i X_i^T - \widehat{\theta}\widehat{\theta}^T)$. The theoretical analysis can be generalized to this case, and we omit the details.

The condition $p \vee q < n^{1-\delta}$ can be weakened if additional sparsity assumptions on θ and η are imposed. This is related to the sparse clustering setting studied in the literature [2, 19, 20], and a sparse PCA algorithm [4, 6, 21, 25, 30] can be applied to estimate θ and η. We would also like to point out a recent work by Roquain and Verzelen [27] that studies the effect of estimating the null distribution for the Benjamini–Hochberg procedure in the setting of multiple testing. In our problem of testing the equivalence of clustering, there may also be very subtle effects of estimating the null distribution without assuming the condition $p \vee q < n^{1-\delta}$. Whether the detection boundary stays the same and how it may change are intriguing problems to be investigated in the future.

6. Some simulation results. This section reports results from a simulation study in the equal SNR case. We assume knowledge of θ and η and so the ambient dimensions p and q do not affect our results.

We set $n = 10^5$, $r = 0.2$ and vary β in $\{0.6, 0.7, 0.8, 0.9\}$. In the null case, we have both cluster label vectors equal to $(1, \ldots, 1, -1, \ldots, -1)^T$, a 10^5-dimensional vector with its first 5×10^4 elements equal to 1 and the remaining ones equal to -1. In each alternative case (as specified by a β value), we set a collection of $\lceil n^{1-\beta} \rceil$ coordinates from 1 to -1. In both null and alternative settings, the cluster label vectors are deterministic.

We consider two tests. The first test is essentially (23). The only modification is that we use an empirical critical value estimated from a simulated null distribution of $\check{T}_n^+ \wedge \check{T}_n^-$ rather than $\sqrt{2(1+\delta)\log\log n}$. We refer to it as the "two-view test." The second test is the single-view test in (15). Here, we also use an empirical critical value estimated from a simulated null distribution of $T_n^+ \wedge T_n^-$. We refer to it as the "single-view test." For either test, the estimated critical value is set to be the 95th percentile of an empirical null distribution over 400 repetitions (corresponding to $\alpha = 0.05$).

Table 1 reports the powers (obtained as average over 200 repetitions in each alternative case) of these two tests for the four different β values. By Proposition 2.1 and Theorem 2.4, phase transitions for the two-view test and the single-view test occur at $\beta = 0.8$ and 0.7, respectively. Results in Table 1 corroborate these theoretical findings. For either test, when

TABLE 1

Comparison of powers and phase transitions: *two-view test versus single-view test. Equal SNR setting with* $r = 0.2$ *and* $n = 10^5$. *Each reported power is the average over* 200 *repetitions. Critical values are obtained svia* 400 *repetitions in the null case with* $\alpha = 0.05$

β	0.6	0.7	0.8	0.9
Two-view test	0.845	0.245	0.085	0.055
Single-view test	0.345	0.090	0.060	0.050

β is no larger than the phase transition point, the test has nontrivial power that is larger than 0.05. When β is larger than the point, power becomes trivial. Across all β values, the two-view test consistently outperforms the single-view test.

Funding The first author was supported in part by NSF CAREER award DMS-1847590 and NSF grant CCF-1934931. The second author was supported in part by NSF CAREER award DMS-1352060.

SUPPLEMENTARY MATERIAL

Supplement to "Testing equivalence of clustering" (DOI: 10.1214/21-AOS2113SUPP; .pdf). The supplement contains proofs of all results in the main text.

REFERENCES

[1] ARIAS-CASTRO, E., HUANG, R. and VERZELEN, N. (2020). Detection of sparse positive dependence. *Electron. J. Stat.* **14** 702–730. MR4057605 https://doi.org/10.1214/19-EJS1675

[2] AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems* 2139–2147.

[3] BINKIEWICZ, N., VOGELSTEIN, J. T. and ROHE, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104** 361–377. MR3698259 https://doi.org/10.1093/biomet/asx008

[4] BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803 https://doi.org/10.1214/12-AOS1014

[5] CAI, T. T., MA, J. and ZHANG, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Ann. Statist.* **47** 1234–1267. MR3911111 https://doi.org/10.1214/18-AOS1711

[6] CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

[7] CAI, T. T., SUN, W. and WANG, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 187–234. MR3928141

[8] CAI, T. T. and WU, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inf. Theory* **60** 2217–2232. MR3181520 https://doi.org/10.1109/TIT.2014.2304295

[9] CURTIS, C., SHAH, S. P., CHIN, S.-F. et al. (2012). The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **486** 346.

[10] DESHPANDE, Y., SEN, S., MONTANARI, A. and MOSSEL, E. (2018). Contextual stochastic block models. In *Advances in Neural Information Processing Systems* 8581–8593.

[11] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

[12] DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751 https://doi.org/10.1214/14-STS506

[13] GAO, C. and MA, Z. (2022). Supplement to "Testing equivalence of clustering." https://doi.org/10.1214/21-AOS2113SUPP

[14] GAO, L. L., BIEN, J. and WITTEN, D. (2020). Are clusterings of multiple data views independent? *Biostatistics* **21** 692–708. MR4164052 https://doi.org/10.1093/biostatistics/kxz001

[15] GAO, L. L., WITTEN, D. and BIEN, J. (2019). Testing for association in multi-view network data. Preprint. Available at arXiv:1909.11640.

[16] INGSTER, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–69. MR1456646

[17] INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446 https://doi.org/10.1007/978-0-387-21580-8

[18] JIN, J. and KE, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statist. Sinica* **26** 1–34. MR3468343

[19] JIN, J., KE, Z. T. and WANG, W. (2017). Phase transitions for high dimensional clustering and related problems. *Ann. Statist.* **45** 2151–2189. MR3718165 https://doi.org/10.1214/16-AOS1522

[20] JIN, J. and WANG, W. (2016). Influential features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359. MR3576543 https://doi.org/10.1214/15-AOS1423

[21] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

[22] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. Preprint. Available at arXiv:1612.02099.

[23] MA, H., YANG, H., LYU, M. R. and KING, I. (2008). Sorec: Social recommendation using probabilistic matrix factorization. In *Proceedings of the* 17*th ACM Conference on Information and Knowledge Management* 931–940. ACM, New York.

[24] MA, H., ZHOU, D., LIU, C., LYU, M. R. and KING, I. (2011). Recommender systems with social regularization. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* 287–296. ACM, New York.

[25] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121 https://doi.org/10.1214/13-AOS1097

[26] MO, Q., WANG, S., SESHAN, V. E., OLSHEN, A. B., SCHULTZ, N., SANDER, C., POWERS, R. S., LADANYI, M. and SHEN, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA* **110** 4245–4250.

[27] ROQUAIN, E. and VERZELEN, N. (2019). On using empirical null distribution in Benjamini–Hochberg procedure. Preprint. Available at arXiv:1912.03109.

[28] SHEN, R., OLSHEN, A. B. and LADANYI, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912.

[29] VERZELEN, N. and ARIAS-CASTRO, E. (2017). Detection and feature selection in sparse mixture models. *Ann. Statist.* **45** 1920–1950. MR3718157 https://doi.org/10.1214/16-AOS1513

[30] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452 https://doi.org/10.1214/13-AOS1151

[31] WANG, H., TERROVITIS, M. and MAMOULIS, N. (2013). Location recommendation in location-based social networks using user check-in data. In *Proceedings of the* 21*st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 374–383. ACM, New York.

[32] ZHAO, S. D., CAI, T. T. and LI, H. (2017). Optimal detection of weak positive latent dependence between two sequences of multiple tests. *J. Multivariate Anal.* **160** 169–184. MR3688697 https://doi.org/10.1016/j.jmva.2017.06.009

[33] ZHAO, S. D. and NGUYEN, Y. T. (2020). Nonparametric false discovery rate control for identifying simultaneous signals. *Electron. J. Stat.* **14** 110–142. MR4047996 https://doi.org/10.1214/19-EJS1663