The Annals of Statistics
2017, Vol. 45, No. 5, 2151–2189
DOI: 10.1214/16-AOS1522

PHASE TRANSITIONS FOR HIGH DIMENSIONAL CLUSTERING AND RELATED PROBLEMS

BY JIASHUN JIN, ZHENG TRACY KE AND WANJIE WANG

Carnegie Mellon University, University of Chicago and University of Pennsylvania

Consider a two-class clustering problem where we observe $X_i = \ell_i \mu + Z_i$, $Z_i \overset{i.i.d.}{\sim} N(0, I_p)$, $1 \le i \le n$. The feature vector $\mu \in R^p$ is unknown but is presumably sparse. The class labels $\ell_i \in \{-1, 1\}$ are also unknown and the main interest is to estimate them.

We are interested in the statistical limits. In the two-dimensional phase space calibrating the rarity and strengths of useful features, we find the precise demarcation for the *Region of Impossibility* and *Region of Possibility*. In the former, useful features are too rare/weak for successful clustering. In the latter, useful features are strong enough to allow successful clustering. The results are extended to the case of colored noise using Le Cam's idea on comparison of experiments.

We also extend the study on statistical limits for clustering to that for signal recovery and that for global testing. We compare the statistical limits for three problems and expose some interesting insight.

We propose classical PCA and Important Features PCA (IF-PCA) for clustering. For a threshold $t > 0$, IF-PCA clusters by applying classical PCA to all columns of X with an L^2-norm larger than t. We also propose two aggregation methods. For any parameter in the Region of Possibility, some of these methods yield successful clustering.

We discover a phase transition for IF-PCA. For any threshold $t > 0$, let $\xi^{(t)}$ be the first left singular vector of the post-selection data matrix. The phase space partitions into two different regions. In one region, there is a t such that $\cos(\xi^{(t)}, \ell) \to 1$ and IF-PCA yields successful clustering. In the other, $\cos(\xi^{(t)}, \ell) \le c_0 < 1$ for all $t > 0$.

Our results require delicate analysis, especially on post-selection random matrix theory and on lower bound arguments.

1. Introduction. Motivated by the interest on gene microarray study, we consider a clustering problem where we have n subjects from two different classes (e.g., normal and diseased), measured on the same set of p features (i.e., gene expression level). To facilitate the analysis, we assume that two classes are equally likely so the class labels satisfy

$$\ell_i \overset{i.i.d.}{\sim} 2\,\text{Bernoulli}(1/2) - 1, \qquad 1 \le i \le n. \tag{1.1}$$

Received March 2015; revised June 2016.
MSC2010 subject classifications. Primary 62H30, 62H25; secondary 62G05, 62G10.
Key words and phrases. Clustering, comparison of experiments, feature selection, hypothesis testing, L^1-distance, lower bound, low-rank matrix recovery, phase transition.

We also assume that the p-dimensional data vectors X_i's are standardized, so that for a contrast mean vector $\mu \in R^p$,

$$X_i = \ell_i \mu + Z_i, \qquad Z_i \overset{i.i.d.}{\sim} N(0, I_p), 1 \le i \le n. \tag{1.2}$$

Throughout this paper, we call feature j, $1 \le j \le p$, a "useless feature" or "noise" if $\mu(j) = 0$ and a "useful feature" or "signal" otherwise.

The paper focuses on the problem of clustering (i.e., estimating the class labels ℓ_i). Such a problem is of interest, especially in the study of complex disease [31]. In the two-dimensional phase space calibrating the signal rarity and signal strengths, we are interested in the following limits.[1]

- *Statistical limits*. This is the precise boundary that separates the Region of Impossibility and Region of Possibility. In the former, the signals are so rare and (individually) weak that it is impossible for any method to correctly identify most of the class labels. In the latter, the signals are strong enough to allow successful clustering, and it is desirable to develop methods that cluster successfully.
- *Computationally tractable statistical limits*. This is similar to the boundary above, except that for both Possibility and Impossibility, we only consider statistical methods that are computationally tractable.

We use Region of Possibility and Region of Impossibility as generic terms, which may vary from occurrence to occurrence. The paper also contains three closely related objectives as follows, which we discuss in Sections 1.4 and 2, Section 3 and Section 4, respectively:

- Performance of the recent idea of Important Features PCA (IF-PCA).
- Limits for recovering the support of μ (signal recovery).
- Limits for testing whether X_i's are *i.i.d.* samples from $N(0, I_p)$, or generated from Model (1.2) (hypothesis testing).

Our work on sparse clustering is related to Azizyan *et al.* [7] and Chan and Hall [12] (see also [35, 36, 40, 46]): the three papers share the same spirit that we should do a feature selection before we cluster. Our work on support recovery is related to recent interest on sparse PCA (e.g., Amini and Wainwright [3], Johnstone and Lu [29], Vu and Lei [43], Wang *et al.* [44], Arias-Castron and Verzelen [5]), and our work on hypothesis testing is related to recent interest on matrix estimation and matrix testing (e.g., Arias-Castro and Verzelen [5], Cai *et al.* [10]). However, our work is different in many important aspects, especially for our focus on the limits and on the Rare/Weak models. See Section 6 for more discussion.

[1] All limits in this paper are with respect to the ARW model introduced in Section 1.2.

1.1. *Four clustering methods.* Denoting the data matrix by X, we write

$$X' = [X_1, X_2, \ldots, X_n], \qquad X = [x_1, x_2, \ldots, x_p].$$

We introduce two methods: a feature aggregation method and IF-PCA. Each method includes a special case, which can be viewed as a different method.

The first method $\hat{\ell}_N^{(\mathrm{sa})}$ targets on the case where the signals are rare but individually strong ("sa": *Sparse Aggression*; N: tuning parameter; usually, $N \ll p$), so feature selection is desirable. Denote the support of μ by

$$S(\mu) = \{1 \le j \le p : \mu(j) \ne 0\}. \tag{1.3}$$

The procedure first estimates $S(\mu)$ by optimizing ($\|\cdot\|_1$: vector L^1-norm)

$$\hat{S}_N^{(\mathrm{sa})} = \mathrm{argmax}_{\{S \subset \{1,2,\ldots,p\}:|S|=N\}} \Big\{ \Big\| \sum_{j \in S} x_j \Big\|_1 \Big\}, \tag{1.4}$$

and then cluster by aggregating all selected features $\hat{\ell}_N^{(\mathrm{sa})} = \mathrm{sgn}(\sum_{j \in \hat{S}_N^{(\mathrm{sa})}} x_j)$.[2]

An important special case is $N = p$, where $\hat{\ell}_N^{(\mathrm{sa})}$ reduces to the method of *Simple Aggregation* which we denote by $\hat{\ell}_*^{(\mathrm{sa})}$.[3] This procedure targets on the case where the signals are weak but less sparse, so feature selection is hopeless. Note that $\hat{\ell}_N^{(\mathrm{sa})}$ is generally NP-hard but $\hat{\ell}_*^{(\mathrm{sa})}$ is not.

The second method is IF-PCA, denoted by $\hat{\ell}_q^{(\mathrm{if})}$, where $q > 0$ is a tuning parameter. The method targets on the case where the signals are rare but individually strong. To use $\hat{\ell}_q^{(\mathrm{if})}$, we first select features using the χ^2-tests:

$$\hat{S}_q^{(\mathrm{if})} = \{1 \le j \le p : Q(j) \ge \sqrt{2q \log(p)}\}, \qquad Q(j) = (\|x_j\|^2 - n)/\sqrt{2n}. \tag{1.5}$$

We then obtain the first left singular vector $\xi^{(q)}$ of the post-selection data matrix $X^{(q)}$ (containing only columns of X where the indices are in $\hat{S}_q^{(\mathrm{if})}$):

$$\xi^{(q)} = \xi(X^{(q)}), \tag{1.6}$$

and cluster by $\hat{\ell}_q^{(\mathrm{if})} = \mathrm{sgn}(\xi^{(q)})$. IF-PCA includes the classical PCA (denoted by $\hat{\ell}_*^{(\mathrm{if})}$) as a special case, where the feature selection step is skipped, and $\xi^{(q)}$ reduces to the first singular vector of X.[4]

In Table 1, we compare all four methods. Note that for more complicated cases [e.g., the nonzero $\mu(j)$'s may be both positive and negative], we may consider a variant of $\hat{\ell}_N^{(\mathrm{sa})}$ which clusters by $\hat{\ell}_N^{(\mathrm{sa})} = \mathrm{sgn}(X\hat{\mu})$, with $\hat{\mu}$ being

[2]For any vector $x \in R^n$, $\mathrm{sgn}(x) \in R^n$ is the vector where the ith entry is $\mathrm{sgn}(x_i)$, $1 \le i \le n$ [$\mathrm{sgn}(x_i) = -1, 0, 1$ according to $x_i < 0, = 0$, or > 0].

[3]The superscript "sa" now loses its original meaning, but we keep it for consistency.

[4]The superscript "if" now loses its original meaning, but we keep it for consistency.

TABLE 1
Comparison of basic characteristics of four methods

Methods	**Simple aggregation** $\hat{\ell}_*^{(\mathrm{sa})}$	**Sparse aggregation** $\hat{\ell}_N^{(\mathrm{sa})}$ $(N \ll p)$	**Classical PCA** $\hat{\ell}_*^{(\mathrm{if})}$	**IF-PCA** $\hat{\ell}_q^{(\mathrm{if})}$ $(q > 0)$
Signals	Less sparse/weak	Sparse/strong*	Moderately sparse/weak	Very sparse/strong
Feature selection	No	Yes	No	Yes
Comp. complexity	Polynomial	NP-hard	Polynomial	Polynomial
Need tuning	No	Yes	No	Yes†

∗: signals are comparably stronger but still weak. †: a tuning-free version exists.

$\mathrm{argmax}_{\{\mu(j)\in\{-1,0,1\},\|\mu\|_0=N\}} \|X\mu\|_q$, where $q > 0$. If we let $q = 1$ and restrict $\mu(j) \in \{0, 1\}$, it reduces to the current $\hat{\ell}_N^{(\mathrm{sa})}$. Note that when $N = p$ and $q = 2$, approximately, $\hat{\mu}$ is proportional to the first right singular vector of X and $\hat{\ell}_N^{(\mathrm{sa})}$ is approximately the classical PCA. Note also that $\hat{\ell}_q^{(\mathrm{if})}$ can be viewed as the adaption of IF-PCA in Jin and Wang [28] to Model (1.2). The version in [28] is a tuning free algorithm for analyzing microarray data and is much more sophisticated. The current version of IF-PCA is similar to that in Johnstone and Lu [29] but is also different in purpose and in implementation: the former is for estimating ℓ and uses the first *left* singular vector of the post-selection data matrix, and the latter is for estimating μ and uses the first *right* singular vector. The theory two methods entail are also very different. See Sections 1.8 and 6 for more discussion.

1.2. *Rare and weak signal model.* To study all these limits, we invoke the Asymptotic Rare and Weak (ARW) model [11, 17, 18, 24]. In ARW, for two parameters (ε, τ), we model the contrast mean vector μ by

$$\mu(j) \overset{i.i.d.}{\sim} (1-\varepsilon)\nu_0 + \varepsilon\nu_\tau, \qquad 1 \le j \le p, \tag{1.7}$$

where ν_a denotes the point mass at a. In Model (1.7), all signals have the same sign and magnitude. Such an assumption can be largely relaxed; see Sections 1.6 and 6. We use p as the driving asymptotic parameter and tie (n, ε, τ) to p by fixed parameters. In detail, fixing $(\theta, \beta) \in (0, 1)^2$ and $\alpha > 0$, we model

$$n = n_p = p^\theta, \qquad \varepsilon = \varepsilon_p = p^{-\beta}, \qquad \tau = \tau_p = p^{-\alpha}. \tag{1.8}$$

In our model, $n \ll p$ for we focus on the modern "large n, really large p" regime [39]. The study can be conveniently extended to the case of $n \gg p$.

1.3. *Limits for clustering.* Let Π be the set of all possible permutations on $\{-1, 1\}$. For any clustering procedure $\hat{\ell}$ (where $\hat{\ell}_i$ takes values from $\{-1, 1\}$), we

measure the performance by the Hamming distance:

$$\text{Hamm}_p(\hat{\ell}, \alpha, \beta, \theta) = n^{-1} \inf_{\pi \in \Pi} \Big\{ \sum_{i=1}^{n} P(\hat{\ell}_i \neq \pi \ell_i) \Big\}, \tag{1.9}$$

where the probability is evaluated with respective to (μ, ℓ, Z). Fixing $\theta \in (0, 1)$, introduce a curve $\alpha = \eta_\theta^{\text{clu}}(\beta)$ in the β-α plane by

$$\eta_\theta^{\text{clu}}(\beta) = \begin{cases} (1-2\beta)/2, & \beta < (1-\theta)/2, \\ \theta/2, & (1-\theta)/2 < \beta < (1-\theta), \\ (1-\beta)/2, & \beta > (1-\theta). \end{cases}$$

THEOREM 1.1 (Statistical lower bound[5]). *Fix $(\theta, \beta) \in (0, 1)^2$ and $\alpha > 0$ such that $\alpha > \eta_\theta^{\text{clu}}(\beta)$. Consider the clustering problem for Models (1.1)–(1.2) and (1.7)–(1.8). For any procedure $\hat{\ell}$, $\liminf_{p\to\infty} \text{Hamm}_p(\hat{\ell}, \alpha, \beta, \theta) \geq 1/2$.*

THEOREM 1.2 (Statistical upper bound for clustering). *Fix $(\theta, \beta) \in (0, 1)^2$ and $\alpha > 0$ such that $\alpha < \eta_\theta^{\text{clu}}(\beta)$, and consider the clustering problem for Models (1.1)–(1.2) and (1.7)–(1.8). As $p \to \infty$:*

- $\text{Hamm}_p(\hat{\ell}_*^{(\text{sa})}, \alpha, \beta, \theta) \to 0$, *if* $0 < \beta < (1-\theta)/2$.
- $\text{Hamm}_p(\hat{\ell}_N^{(\text{sa})}, \alpha, \beta, \theta) \to 0$, *if* $(1-\theta)/2 < \beta < 1$ *and* $N = \lceil p\varepsilon_p \rceil$.[6]

As a result, the curve $\alpha = \eta_\theta^{\text{clu}}(\beta)$ divides the β-α plane into two regions: Region of Impossibility and Region of Possibility. In the former, the signals are so weak that successful clustering is impossible. In the latter, the signals are strong enough to allow successful clustering.

Consider computationally tractable limits. We call a curve $r = \eta_\theta(\beta)$ in the β-α plane a *Computationally Tractable Upper Bound (CTUB)* if for any fixed (θ, α, β) such that $\alpha < \eta_\theta(\beta)$, there is a computationally tractable clustering method $\hat{\ell}$ such that $\text{Hamm}_p(\hat{\ell}, \alpha, \beta, \theta) \to 0$. A CTUB $r = \eta_\theta(\beta)$ is tight if for any computationally tractable method $\hat{\ell}$ and any fixed (θ, α, β) such that $\alpha > \eta_\theta(\beta)$, $\liminf_{p\to\infty} \text{Hamm}_p(\hat{\ell}, \alpha, \beta, \theta) \geq 1/2$. In this case, we call $r = \eta_\theta(\beta)$ the Computationally Tractable Boundary (CTB). Define

$$\tilde{\eta}_\theta^{\text{clu}}(\beta) = \begin{cases} (1-2\beta)/2, & \beta < (1-\theta)/2, \\ (1+\theta-2\beta)/4, & (1-\theta)/2 < \beta < 1/2, \\ \theta/4, & 1/2 < \beta < 1-\theta/2, \\ (1-\beta)/2, & 1-\theta/2 < \beta < 1. \end{cases}$$

[5]The "lower bound" refers to the information lower bound as in the literature, not the lower bound for the curves in Figure 1 (say). Same for the "upper bound."

[6]$\lceil x \rceil$ denotes the smallest integer that is no smaller than x.

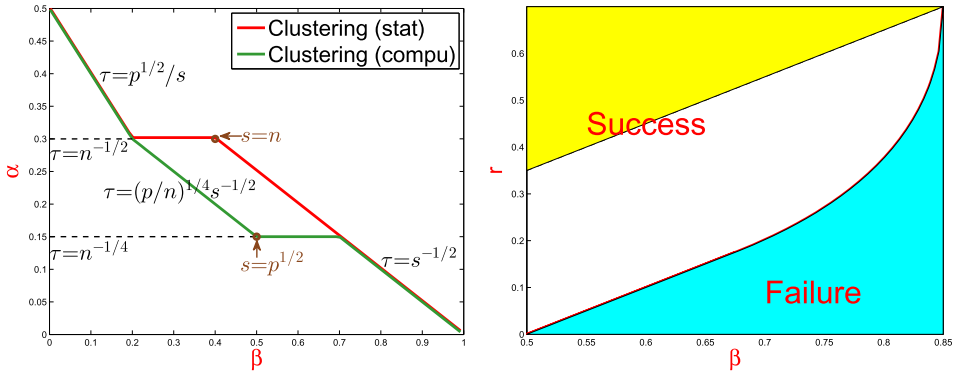

FIG. 1. *Left*: *the statistical limits* (*red*) *and the CTUB* (*green*) *for clustering* ($s = p\varepsilon_p$ *is the expected number of signals*). *Right*: *phase transition of IF-PCA*. *White region*: *successful clustering is possible but successful feature selection is impossible* (*using column-wise* χ^2 *scores*). *Yellow region*: *both successful clustering and feature selection are possible*.

THEOREM 1.3 (A CTUB for clustering). *Fix* $(\theta, \beta) \in (0,1)^2$ *and* $\alpha > 0$ *such that* $\alpha < \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$, *and consider the clustering problem for Models* (1.1)–(1.2) *and* (1.7)–(1.8). *As* $p \to \infty$:

- $\mathrm{Hamm}_p(\hat{\ell}_*^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$, *if* $0 < \beta < (1-\theta)/2$.
- $\mathrm{Hamm}_p(\hat{\ell}_*^{(\mathrm{if})}, \alpha, \beta, \theta) \to 0$, *if* $(1-\theta)/2 < \beta < 1/2$.
- $\mathrm{Hamm}_p(\hat{\ell}_q^{(\mathrm{if})}, \alpha, \beta, \theta) \to 0$, *if* $1/2 < \beta < 1$ *and we take* $q \geq 3$.

We now discuss CTB. We discuss the cases (a) $0 < \beta < (1-\theta)/2$, (b) $(1-\theta)/2 < \beta < 1/2$, (c) $1/2 < \theta < 1-\theta/2$, and (d) $1-\theta/2 < \beta < 1$ separately. Note that the CTB is sandwiched by two curves $\alpha = \eta_\theta^{\mathrm{clu}}(\beta)$ and $\alpha = \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$. In (a) and (d), $\tilde{\eta}_\theta^{\mathrm{clu}}(\beta) = \eta_\theta^{\mathrm{clu}}(\beta)$, so our CTUB (i.e., CTUB given in Theorem 1.3) is tight. For (b), we are not sure but we conjecture that our CTUB is tight.[7] For (c), we have good reasons to believe that our CTUB is tight. In fact, our model is intimately connected to the spike model [29]; see Section 1.8. The tightness of our CTUB under the spike model has been well-studied (e.g., [9, 34]). Translating their results[8] to our setting suggests that there is a small constant $\delta > 0$ such that when $1/2 < \beta < 1/2 + \delta$, our CTUB is tight. Note that for (c), the CTUB $\alpha = \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$ is flat. By the monotonicity of CTB (see below), our CTUB is tight for (c). See Figure 1.

[7] We know that CTB crosses two points $(\beta, \alpha) = (1/2, \theta/4)$ and $(\beta, \alpha) = ((1-\theta)/2, \theta/2)$. A natural guess is that the CTB in this part is a line segment connecting the two points.

[8] Consider the hypothesis testing in the spike model. [9] proves that, with the "planted clique" conjecture, for $n < p$ and $s = o(\sqrt{p})$, if $\|\mu\|_0 = s$ and $\|\mu\|^2 \leq s\sqrt{\log(p)/n}$, there is no polynomial-time test that is powerful. In ARW, since $\|\mu\|^2 \approx s\tau^2$, the above translates to (ignoring the logarithmic factor) $\alpha > \theta/4 = \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$.

REMARK (Monotonicity of CTB). We show the CTB is monotone in β (with θ fixed). Fix $\delta > 0$ and consider a new experiment, where for each column of the data matrix, we keep the column with probability $p^{-\delta}$ and replace it with an independent column drawn from $N(0, I_n)$ with probability $1 - p^{-\delta}$. Compare this with the original experiment. The parameters (α, θ) are the same, but β has become $(\beta + \delta)$. The second experiment is harder, for it is the result of the original experiment by sub-sampling the columns. This shows that the CTB is monotone in β. The monotonicity now follows by Le Cam's results on comparison of experiments [33].

1.4. *Phase transition for IF-PCA*. IF-PCA is a flexible clustering method that is easy to use and computationally efficient. In [28], we developed a tuning free version of IF-PCA using Higher Criticism [17, 19, 25] and applied it to 10 microarray data sets with satisfactory results. The success of IF-PCA in real data analysis motivates us to investigate the method in depth. To facilitate delicate analysis, we consider the version of IF-PCA in Section 1.1, and reveal an interesting phase transition.

To this end, we investigate a very challenging case (not covered in Theorems 1.1–1.3) where (α, β) fall exactly on the CTUB in Theorem 1.3:

$$\alpha = \tilde{\eta}_\theta^{\mathrm{clu}}(\beta).^{9} \tag{1.10}$$

Also, note that a key step in IF-PCA is the column-wise χ^2-screening. In our model, a column x_j is either distributed as $N(0, I_n)$ or $N(\tau_p \ell, I_n)$, where $\tau_p = p^{-\alpha}$. For the χ^2-screening to be nontrivial, we further require that

$$1/2 < \beta < 1 - \theta/2. \tag{1.11}$$

For β in this range, the curve $\alpha = \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$ is flat, that is, $\tilde{\eta}_\theta^{\mathrm{clu}}(\beta) \equiv \theta/4$, and so $\tau_p = p^{-\theta/4} = n^{-1/4}$. For β outside this range, (1.10) dictates that either $\tau_p \ll n^{-1/4}$ (so that the signals are too weak that is, that the χ^2-screening bounds to fail) or $\tau_p \gg n^{-1/4}$ (so that the signals are too strong that the χ^2-screening is relatively trivial). See Figure 1.

We now restrict our attention to (1.10)–(1.11), where we recall that $\tau_p = p^{-\theta/4}$. To make the case more interesting, we adjust the calibration of τ_p slightly by an $O(\log^{1/4}(p))$ factor:

$$\tau_p^* = p^{-\theta/4}\big(4r\log(p)\big)^{1/4} \qquad \text{where } 0 < r < 1 \text{ is a fixed parameter.} \tag{1.12}$$

With this calibration, the χ^2-screening could be successful but nontrivial.

[9] The case $\alpha < \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$ is comparably easier to study, and the case $\alpha > \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$ belongs to the Region of Impossibility for computationally tractable methods; see our conjectures.

Introduce the *standard phase function*[10] [17, 18]

$$\rho^*(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases} \tag{1.13}$$

Define the *phase function* for IF-PCA by

$$\rho_\theta^*(\beta) = (1-\theta) \cdot \rho^*\Big(1/2 + \frac{\beta - 1/2}{1-\theta}\Big), \qquad 1/2 < \beta < 1 - \theta/2. \tag{1.14}$$

For any two vectors x and y in R^n, let $\cos(x, y) = |\langle x/\|x\|, y/\|y\|\rangle|$.

THEOREM 1.4 (Phase transition for IF-PCA). *Fix $(\theta, \beta, \alpha, r) \in (0,1)^4$ and $q > 0$ such that (1.10)–(1.11) hold. Consider IF-PCA $\hat{\ell}_q^{(\mathrm{if})}$ for Models (1.1)–(1.2) and (1.7)–(1.8), where τ_p is replaced by the new calibration τ_p^* in (1.12), and let $\xi^{(q)}$ be the leading left singular vector as in (1.6). As $p \to \infty$:*

- *If $r > \rho_\theta^*(\beta)$, then with probability at least $1 - o(p^{-2})$, $\cos(\xi^{(q^*)}, \ell) \to 1$ with $q^* = (\beta - \theta/2 + r)^2/(4r)$ for $r > (\beta - \theta/2)/3$ and $q^* = 4r$ otherwise.*
- *If $r < \rho_\theta^*(\beta)$, then with probability at least $1 - o(n^{-1})$, there is a constant $c_0 \in (0,1)$ such that $\cos(\xi^{(q)}, \ell) \le c_0$ for any fixed $0 < q < 1$.*

Theorem 1.4 is proved in Section 2, using delicate spectral analysis on the post-selection data matrix [and so the term of *post-selection Random Matrix Theory (RMT)*]. Compared to many works on RMT where the data matrix has independent entries [42], the entries of the post-selection data matrix are complicatedly correlated, so the required analysis is more delicate. We conjecture that when $r < \rho_\theta^*(\beta)$, $\cos(\xi^{(q)}, \ell) \to 0$ for any fixed $0 < q < 1$. For now, we can only show this for q in a certain range; see the proof for details.

Figure 1 (right) displays the phase diagram for IF-PCA. For fixed (α, β) in the interior of the white region, successful feature selection is impossible (by column-wise χ^2-screening) but successful clustering is possible. This shows that feature selection and clustering are related but different problems.

REMARK. For the IF-PCA considered here, we use column-wise χ^2-tests for screening which is computationally inexpensive. Alternatively, we may use some regularization methods for screening (e.g., [13, 32, 47]). However, these methods are computationally more expensive, need tuning parameters that are hard to set, and are designed for feature selection, not clustering. For these reasons, it is unclear whether such alternatives may really help.

[10] It was introduced in the literature to study the phase transitions of multiple testing and classification with rare/weak signals.

1.5. *Clustering when the noise is colored.* Consider a new version of ARW where (ℓ, μ) are the same as in Models (1.1), (1.7)–(1.8), but Model (1.2) is replaced by a colored noise model:

$$X = \ell\mu' + AZB, \qquad Z_i(j) \overset{i.i.d.}{\sim} N(0,1), 1 \le i \le n, 1 \le j \le p, \tag{1.15}$$

where A and B are two nonrandom matrices.

DEFINITION 1.1. We use $L_p > 0$ to denote a generic multi-log(p) term which may vary from occurrence to occurrence such that for any fixed $\delta > 0$, $L_p p^{-\delta} \to 0$ and $L_p p^{\delta} \to \infty$, as $p \to \infty$.

THEOREM 1.5 (Statistical lower bound for clustering with colored noise). *Consider the ARW model* (1.1)–(1.2) *and* (1.7)–(1.8). *Theorem* 1.1 *continues to hold if we replace the model* (1.2) *by* (1.15) *where* $\max\{\|A\|, \|A^{-1}\|\} \le L_p$ *and* $\max\{\|B\|, \|B^{-1}\|\} \le L_p$.

Theorem 1.5 is proved in Section 5, using Le Cam's comparison of experiments [33]. The idea is to construct a new experiment that is easy to analyze and that the current one can be viewed as the result of adding noise to it. Since "adding noise always makes the inference harder," analyzing the new experiment provides a lower bound we need for the current experiment. The idea has been used in Hall and Jin [22], but for very different settings.

Consider the case $A = I_n$. In this case, the matrix AZB has independent rows (but the columns may be correlated and heteroscedastic), and all four methods we proposed earlier continue to work, except that in IF-PCA we need $q \ge 3\max\{\mathrm{diag}(B'B)\}$. The following theorem is proved in Section 3.

THEOREM 1.6 (Upper bounds for clustering with colored noise). *Consider the ARW model* (1.1)–(1.2) *and* (1.7)–(1.8). *Theorems* 1.2–1.3 *continue to hold if we replace the model* (1.2) *by* (1.15) *with* $A = I_n$ *and* B *such that* $\max\{\|B\|, \|B^{-1}\|\} \le L_p$ *and that all diagonals of* $B'B$ *is upper bounded by a constant* $c > 0$, *where we set* $q \ge 3c$ *in IF-PCA*.

Practically, it is desirable to have a method that does not depend on the unknown parameter c. One way to attack this is to replace the column-wise χ^2-test by a plug-in χ^2-test where we estimate the variance column-wise by median absolute deviation (say). However, such methods usually involve statistics of higher order moments; see [5] for discussions along this line.

1.6. *Limits for signal recovery and hypothesis testing.* For a more complete picture, we study the limits for signal recovery and hypothesis testing.

The goal of signal recovery is to recover the support of μ. For any feature selector $\hat{S}$, we measure the error by the (normalized) Hamming distance

$\mathrm{Hamm}_p(\hat{S}, \alpha, \beta, \theta) = (p\varepsilon_p)^{-1} \sum_{j=1}^{p} [P(\mu(j) = 0,\ j \in \hat{S}) + P(\mu(j) \neq 0,\ j \notin \hat{S})]$, where $p\varepsilon_p$ is the expected number of signals. Define

$$\eta_\theta^{\mathrm{sig}}(\beta) = \begin{cases} \theta/2, & \beta < (1-\theta), \\ (1+\theta-\beta)/4, & \beta > (1-\theta), \end{cases}$$

and

$$\tilde{\eta}_\theta^{\mathrm{sig}}(\beta) = \begin{cases} \theta/2, & \beta < (1-\theta)/2, \\ (1+\theta-2\beta)/4, & (1-\theta)/2 < \beta < 1/2, \\ \theta/4, & \beta > 1/2. \end{cases}$$

The curve $r = \eta_\theta^{\mathrm{sig}}(\beta)$ can be viewed as the counterpart of $r = \eta_\theta^{\mathrm{clu}}(\beta)$, which divides the two-dimensional phase space into the Region of Impossibility and Region of Possibility. For any fixed (β, α) in the former and any $\hat{S}$, $\mathrm{Hamm}_p(\hat{S}, \alpha, \beta, \theta) \gtrsim 1$. For any fixed (β, α) in the latter, there is an $\hat{S}$ such that $\mathrm{Hamm}_p(\hat{S}, \alpha, \beta, \theta) \to 0$. The curve $r = \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$ can be viewed as the counterpart of $r = \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$ and provides a CTUB for the signal recovery problem. See Section 3 for more discussion.

The goal of (global) hypothesis testing is to test a null hypothesis $H_0^{(p)}$ that the data matrix X has *i.i.d.* entries from $N(0, 1)$ against an alternative hypothesis $H_1^{(p)}$ that X is generated according to Model (1.2). Define

$$\begin{aligned} \eta_\theta^{\mathrm{hyp}}(\beta) &= \max\{\eta_\theta^{\mathrm{hyp},1}(\beta), \eta_\theta^{\mathrm{hyp},2}(\beta)\}, \\ \tilde{\eta}_\theta^{\mathrm{hyp}}(\beta) &= \max\left\{\eta_\theta^{\mathrm{hyp},1}(\beta), \frac{\theta}{4}\right\}, \end{aligned} \tag{1.16}$$

and $\eta_\theta^{\mathrm{hyp},1}(\beta) = (2+\theta-4\beta)/4$, $\eta_\theta^{\mathrm{hyp},2} = \min\{\theta/2, (1+\theta-\beta)/4\}$. Similarly, the curve $r = \eta_\theta^{\mathrm{hyp}}(\beta)$ divides the two-dimensional phase space into the Region of Impossibility and Region of Possibility. Fix (β, α) in the former, the sum of Type I and Type II errors $\gtrsim 1$ for any testing procedures. Fixing (β, α) in the latter, there is a test such that the sum of Type I and Type II errors tends to 0. Also, the curve $r = \tilde{\eta}_\theta^{\mathrm{hyp}}(\beta)$ provides a CTUB for the hypothesis testing problem. See Section 4 for more discussion.

The statistical limits for hypothesis testing here are different from those in Arias-Castro and Verzelen [5]. For the less sparse case ($\beta < \theta/2$), the signal strength needed in our model is weaker, because all signals have the same sign. More interestingly, we find a phase transition phenomenon that is not seen in [5]: when $\theta < 2/3$, there are three segments for the statistical limits; when $\theta > 2/3$, there are only two segments.[11]

[11] The curve $r = \eta_\theta^{\mathrm{hyp}}(\beta)$ is the maximum of the boundary achievable by Simple Aggregation (a line segment) and that by Sparse Aggregation (two line segments). Depending on where two boundaries cross each other, $r = \eta_\theta^{\mathrm{hyp}}(\beta)$ may consist of 2 or 3 line segments.

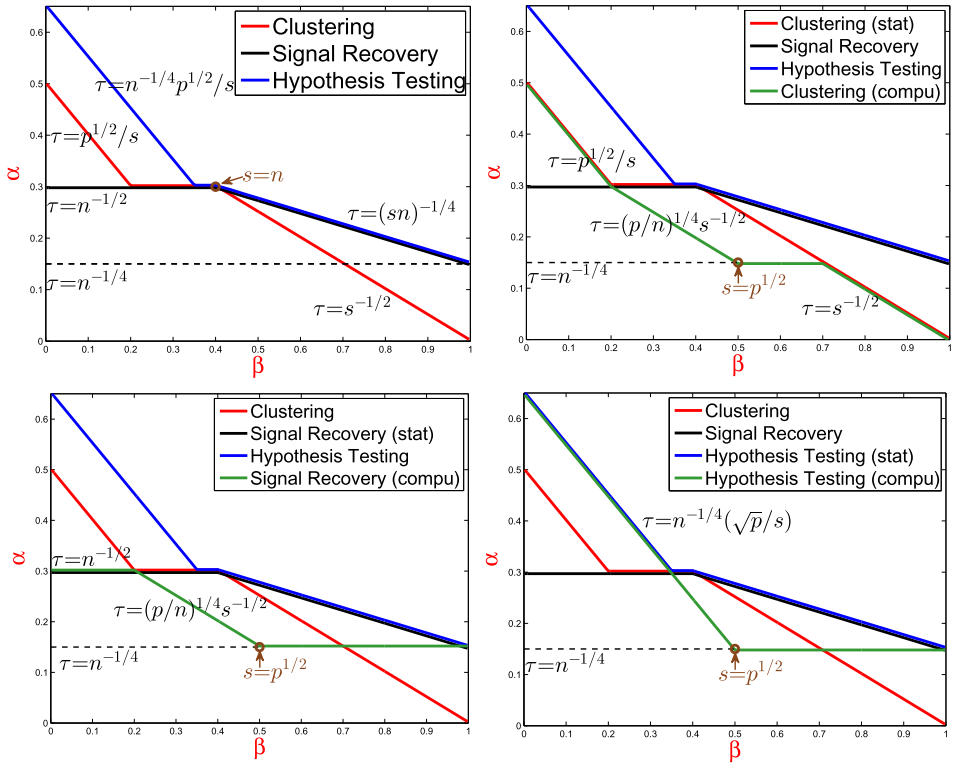

FIG. 2. *Top left*: *statistical limits for clustering* (*red*), *signal recovery* (*black*) *and hypothesis testing* (*blue*); $s = p\varepsilon_p$. *Other three panels*: *CTUB for clustering* (*top right*), *signal recovery* (*bottom left*) *and hypothesis testing* (*bottom right*), *respectively* (*the three statistical limits in the top left panel are also shown for comparison*).

The tightness of CTUB for signal recovery and hypothesis testing can be addressed similarly to that for clustering. For signal recovery, the CTUB is tight in the less sparse case $[0 < \beta < (1-\theta)/2]$ for it matches the statistical limits; we have good reasons to believe it is tight in the sparse case $(1/2 < \beta < 1)$, due to results in [9, 34]; we are not sure for the moderate sparse case $[(1-\theta)/2 < \beta < 1/2]$. For hypothesis testing, we have similar arguments except that the cases of "less sparse" and "moderate sparse" refer to that of $0 < \beta < (2-\theta)/4$ and that of $(2-\theta)/4 < \beta < 1/2$, respectively.

Figure 2 compares the limits for all three problems: clustering, signal recovery and hypothesis testing. See details therein.

REMARK. Consider an extension of ARW where (1.7) is replaced by a more complicated signal configuration: $\mu(j) \overset{i.i.d.}{\sim} (1-\varepsilon)\nu_0 + a\varepsilon\nu_{-\tau} + (1-a)\varepsilon\nu_\tau$, where $0 \le a \le 1/2$ is a constant ($a = 0$: original ARW). When $0 < a < 1/2$, our results on statistical limits and CTUB for all three problems continue to hold, provided

with a slight change in the definition of the Hamming distance for signal recovery. The case of $a = 1/2$ is more delicate, but the changes in statistical limits (compared to the case of $a = 0$) can be explained with Figure 2 (top left): (a) the black curve (signal recovery) remains the same, (b) the red curve (clustering) remains the same, except for the segment on the left is replaced by $\tau^4 = p/(ns^2)$, (c) for the blue curve (hypothesis testing), the right most segment remains the same, while the other two segments coincide with those of the red curve. The CTUBs also change correspondingly. See the supplementary material [27], Appendix D, for a more detailed discussion.

1.7. *Practical relevance and a real data example.* The relatively idealized model we use allows very delicate analysis, but also raises practical concerns. In this section, we investigate IF-PCA with a real data example and illustrate that many ideas in previous sections are relevant in much broader settings.

We use the leukemia data set on gene microarrays. This data set was cleaned by Dettling [15], consisting of $p = 3571$ measured genes for $n = 72$ samples from two classes: 47 from ALL (acute lymphoblastic leukemia), and 25 from AML (acute myeloid leukemia). The data set is available at www.stat.cmu.edu/~jiashun/Research/software/GenomicsData/ALL.

To implement IF-PCA, one noteworthy difficulty is the heteroscedasticity across genes in the data set. We apply IF-PCA with small modifications. In detail, arrange the data matrix as $X = [x_1, \ldots, x_p]$ as before. Let $\bar{x}(j) = (1/n)\sum_{i=1}^n x_j(i)$, $m(x_j) = \text{median}(x_j)$ and $d(j) = \text{median}\{|x_j(1) - m(x_j)|, \ldots, |x_j(n) - m(x_j)|\}$ be the Median Absolute Deviation (MAD). We normalize by $x_j^*(i) = 0.6745 \cdot (x_j(i) - \bar{x}(j))/d(j)$, $1 \le i \le n$, $1 \le j \le p$.[12] For $q > 0$ to be determined, we select feature j if and only if $(2n)^{-1}|\|x_j^*\|^2 - n| > \sqrt{2q\log(p)}$. We then obtain the leading left singular vector $(\xi^*)^{(q)}$ of the post-selection data matrix $[x_1^*, \ldots, x_p^*]$ and cluster by applying the standard k-means algorithm to the leading eigenvector. In the last step, we can also cluster by the sign vector of $(\xi^*)^{(q)}$ and the results are similar. The k-means algorithm has a slightly better performance.

Table 2 displays the clustering errors for different numbers of selected features (each corresponds to a choice of q). The table suggests that IF-PCA works nicely, with an error rate as low as $1/72$, if q is set appropriately.

Figure 3 compares $(\xi^*)^{(q)}$ for three choices of q: (a) the q determined by applying the FDR controlling procedure [8] with the FDR parameter of 0.05 and simulated P-values under the null $x_j \sim N(0, I_n)$, (b) the q associated with the ideal number of selected features (see Table 2), and (c) the q corresponding to classical PCA (any q that allows us to skip the feature selection step works). This suggests that IF-PCA works well if q is properly set.[13]

[12] The value 0.6745 is such that $E[(x_j^*(i))^2] = 1$ when $x_j(i) \sim N(0, \sigma^2)$ for any $\sigma > 0$.

[13] A hard problem is how to set q in a data-driven fashion. This is addressed in [26].

TABLE 2
The clustering errors for leukemia data with different numbers of selected features

#{selected features}	Errors	#{selected features}	Errors	#{selected features}	Errors
1	34	1419	3	2847	5
347	8	**1776**	**1**	3204	7
704	6	**2133**	**1**	3561	11
1062	5	**2490**	**1**		

Rows highlighted correspond to the threshold choices that yield lowest clustering errors.

We compare IF-PCA with classical methods of k-means and hierarchical clustering [23], k-means++ (a recent revision of the classical k-means; [6]),[14] SpectralGem (classical PCA applied to X^*; [31]), and sparse k-means (a modification of k-means with sparse feature weights in the objective; [46]). The error rates are in Table 3, suggesting IF-PCA is effective in this case.

1.8. *Comparison to works on the spike model.* In our model (1.1)–(1.2), if we replace the Bernoulli model for ℓ_i in (1.1) by a Gaussian model where $\ell_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, then it becomes the spike model (Johnstone and Lu [29]).

In the spike model, while ℓ_i's are also of interest, the feature vector μ captures most of the attention: most recent works on the spike model (e.g., [4, 32, 44]) have been focused on signal recovery (and especially, sparse PCA). The two problems,

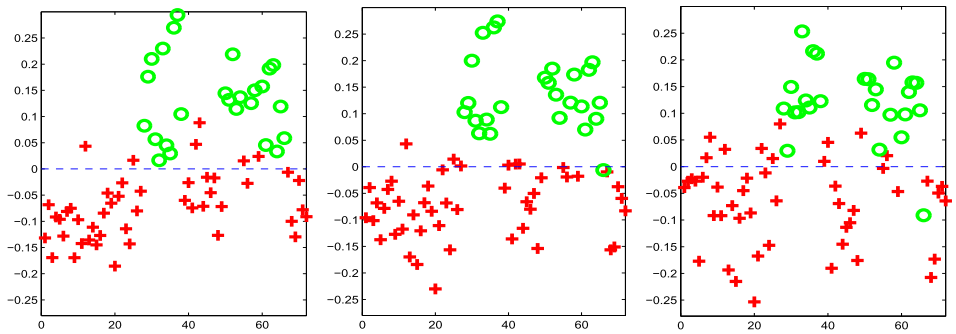

FIG. 3. *Leading left singular vector of the data matrix X with very few features selected by FDR choice* (*left*; 931 *features chosen*), *with ideal number of features selected* (*middle*; 2133 *features chosen*), *and without feature selection* (*right*). *y-axis*: *entries of the left singular vector*, *x-axis*: *sample indices*. *Plots are based on Leukemia data*, *where red and green dots represent samples from the two classes ALL and AML*, *respectively.*

[14]For k-means, we use the built-in Matlab package (parameter 'replicates' equals 30). For k-means++, we run the program 30 times, and compute the average clustering errors.

TABLE 3
Comparison of clustering errors (*leukemia data*)

	Method					
	***k*-means**	***k*-means++**	**Hierarchical**	**SpectralGem**	**Sparse *k*-means**	**IF-PCA**
Error rate	20/72	18.5/72	20/72	21/72	20/72	1/72

Columns 2–7: numerator is the number of clustering errors, and denominator is the number of subjects.

signal recovery and clustering, are different. There are parameter settings where successful clustering is possible but successful signal recovery is impossible, and there are settings where the opposite is true; see Sections 1.4 and 1.6. Therefore, a direct extension of sparse PCA methods to clustering does not always work well.

Our work is also different from existing works on the spike model in terms of motivation and validation. Our model is motivated by cancer (subject) clustering, where the class labels ℓ_i's can be conveniently validated in many applications (e.g., see Section 1.7). In contrast, it is not easy to find real data sets where the feature vector μ is known, so it is comparably harder to validate the methods/theory on signal recovery or sparse PCA. Given the growing awareness of reproducibility and replicability [21], it becomes increasingly more important to develop methods and theory that can be directly validated by real applications. In a sense, our model extends the spike model to a new direction, and it helps strengthen (we hope) the ties between the recent theoretical interests on the spike model with real applications.

1.9. *Content and notation.* Section 2 studies the phase transition of IF-PCA, where we prove Theorem 1.4. Section 3 studies the statistical limits for signal recovery, where we prove Theorems 1.2, 1.3, 1.6, as well as Theorems 3.2–3.3 (to be introduced). Section 4 studies the statistical limits for hypothesis testing, where we prove Theorems 4.2–4.3 (to be introduced). Section 5 studies the lower bounds for all three problems and proves Theorems 1.1 and 1.5, as well as Theorems 3.1 and 4.1 (to be introduced). Other proofs are in the supplementary material [27]. Section 6 is for discussion.

In this paper, $L_p > 0$ denotes a generic multi-log(p) term; see Section 1.5. When ξ is a vector, $\|\xi\|_q$ denotes the vector L^q-norm, $0 \le q \le \infty$ (the subscript is dropped for simplicity if $q = 2$). When ξ is a matrix, $\|\xi\|$ denotes the matrix spectral norm, and $\|\xi\|_F$ denotes the matrix Frobenius norm. For two vectors ξ, η, $\langle \xi, \eta \rangle$ denotes the inner product of them, and $\cos(\xi, \eta) = |\langle \xi/\|\xi\|, \eta/\|\eta\| \rangle|$. For any two probability densities f and g, $\|f - g\|_1$ and $H(f, g)$ are the L^1-distance and the Hellinger distance, respectively. For any real value a, $\lceil a \rceil$ is the smallest integer that is no smaller than a. We say two positive sequences $a_n \sim b_n$, $a_n \lesssim b_n$ and

$a_n \gtrsim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$, $\limsup_{n\to\infty} a_n/b_n \le 1$ and $\liminf_{n\to\infty} a_n/b_n \ge 1$, respectively. For two sets A, B, $A\Delta B = (A \setminus B) \cup (B \setminus A)$.

2. Phase transition for IF-PCA. In this section, we prove Theorem 1.4. Our proofs need very precise characterization of the spectra of the post-selection Gram matrix $X^{(q)}(X^{(q)})'$. Specifically, we need both a tight upper bound on the range of the spectra of $X^{(q)}(X^{(q)})'$ (Lemma 2.1) and a tight lower bound for the largest eigenvalue of $X^{(q)}(X^{(q)})'$ (Lemma 2.2). The main challenges are that, due to feature selection:

- the entries of $X^{(q)}$ are no longer independent,
- the conditional distribution of each survived column is unclear.

For this reason, existing results on RMT do not apply directly and we need to develop new theory on post-selection RMT. Our analysis adapts that in Vershynin [42] and uses the results of covering number in Rogers [37].

REMARK. For the spike model, there are results about the spectra of a different post-selection Gram matrix $(X^{(q)})'X^{(q)}$ (e.g., Theorem 2 of [29]). Since feature selection is column-wise, the leading eigenvectors of $(X^{(q)})'X^{(q)}$ and $X^{(q)}(X^{(q)})'$ have very different behaviors. Moreover, the settings of [29], Theorem 2, implicitly force $X^{(q)}$ to have much more rows than columns (which we call the "skinny" case), but our results do not have such a restriction.

To show the claim, it suffices to show the claim for any fixed realization of (ℓ, μ) in the event

$$D_p = \{\mu : \big||S(\mu)| - p\varepsilon_p\big| \le \sqrt{6p\varepsilon_p \log(p)}\};$$

note that $P(D_p^c) = O(p^{-3})$ and the event only has a negligible effect. Fixing $0 < q < 1$ and a realization of (ℓ, μ) in D_p. Let $\hat{S}_q^{(\mathrm{if})}(\ell, \mu)$ be the set of all survived features. In our model, $X = \ell\mu' + Z$, and $Z = [z_1, z_2, \ldots, z_p]$. Introduce a vector $\mu^{(q)} = \mu^{(q)}(\ell, \mu) \in R^p$ and a matrix $Z^{(q)} = [z_1^{(q)}, \ldots, z_p^{(q)}] \in R^{n,p}$ by

$$\mu^{(q)}(j) = \mu(j) \cdot 1\{j \in \hat{S}_q^{(\mathrm{if})}(\ell, \mu)\},$$
$$z_j^{(q)} = z_j \cdot 1\{j \in \hat{S}_q^{(\mathrm{if})}(\ell, \mu)\}, \qquad 1 \le j \le p,$$

and so the post-selection data matrix $X^{(q)} = X^{(q)}(\ell, \mu)$, viewed as an $n \times p$ matrix with many zero columns, satisfies

$$X^{(q)}(\ell, \mu) = \ell\mu^{(q)}(\ell, \mu) + Z^{(q)}(\ell, \mu).$$

Fixing $(\beta, \theta, r) \in (0, 1)^3$ and $q > 0$, and assuming $z \sim N(0, I_n)$, introduce $m_0^{(q)}(\mu) = (p - |S(\mu)|) \cdot P(\|z\|^2 > n + 2\sqrt{qn\log(p)})$, $m_1^{(q)}(\ell, \mu) = |S(\mu)| \cdot$

$P(\|z+\tau_p^*\ell\|^2 > n+2\sqrt{qn\log(p)})$, and $m^{(q)}(\ell,\mu) = m_0^{(q)}(\ell,\mu) + m_1^{(q)}(\ell,\mu)$. Note that $m_0^{(q)}(\ell,\mu)$ and $m_1^{(q)}(\ell,\mu)$ are the expected numbers of survived useless/useful features, respectively. We also need the following counterpart of $m^{(q)}(\ell,\mu)$:

$$m_*^{(q)}(\ell,\mu) = \big(p-|S(\mu)|\big)\cdot n^{-1}E\big(\|z\|^2 1\{\|z\|^2 > n+2\sqrt{qn\log(p)}\}\big) + |S(\mu)|\cdot n^{-1}E\big(\|z\|^2 1\{\|z+\tau_p^*\ell\|^2 > n+2\sqrt{qn\log(p)}\}\big).$$

The dependence on (ℓ,μ) is tedious, so for notational simplicity, we may drop them without further notices.

The term $m^{(q)}$ is the expected number of selected features, and plays an important role. By tail properties of chi-square distributions (see Section B.1 of the supplementary material), with probability $1-O(p^{-3})$,

$$m_*^{(q)} \sim m^{(q)} \sim L_p\big[p^{1-q} + p\varepsilon_p p^{-[(\sqrt{q}-\sqrt{r})_+]^2}\big], \tag{2.1}$$

where as before L_p is a generic multi-$\log(p)$ term. Recalling $n=p^\theta$, define

$$\tilde{q}(\beta,\theta,r) = \begin{cases} \max\{1-\theta, (\sqrt{1-\beta-\theta}+\sqrt{r})^2\}, & \beta < 1-\theta, \\ 1-\theta, & \beta > 1-\theta. \end{cases}$$

By (2.1) and basic algebra, it is seen that there are two different cases:

- (*"Fat"*). When $q < \tilde{q}(\beta,\theta,r)$, $m^{(q)}/n \to \infty$ and $X^{(q)}$ has much more columns than rows.
- (*"Skinny"*). When $q > \tilde{q}(\beta,\theta,r)$, $m^{(q)}/n \to 0$ and $X^{(q)}$ has much more rows than columns.

LEMMA 2.1 [Upper bound for the range of eigenvalues of $Z^{(q)}(Z^{(q)})'$]. *Suppose conditions of Theorem 1.4 hold. There exists a universal constant $C>0$ such that for any fixed $q>0$, as $p\to\infty$, conditioning on any realization of (ℓ,μ) from the event D_p, with probability at least $1-O(p^{-3})$:*

- (*"Fat" case*). *When $q < \tilde{q}(\beta,\theta,r)$, all eigenvalues of $Z^{(q)}(Z^{(q)})'$ fall between $m_*^{(q)} \pm [C\sqrt{nm^{(q)}\log(p)} + o(m_1^{(q)})]$.*
- (*"Skinny" case*). *When $q > \tilde{q}(\beta,\theta,r)$, all nonzero eigenvalues of $Z^{(q)}(Z^{(q)})'$ fall between $n \pm C\sqrt{nm^{(q)}\log(p)}$.*

REMARK. Noting that $m^{(q)}$ is the expected number of columns of $Z^{(q)}$, our results are very similar to the well-known results on eigenvalues of RMT in the case where we have an $n\times m^{(q)}$ matrix with *i.i.d.* $N(0,1)$ entries. However, we need more sophisticated proofs, as the rows of $Z^{(q)}$ are dependent and the distribution of the columns of $Z^{(q)}$ is unknown and hard to characterize.

For the "fat" case, it turns out that Lemma 2.1 is insufficient: we need both an improved upper bound on the range [with the $\sqrt{\log(p)}$ factor eliminated] and a lower bound on the leading eigenvalue.

LEMMA 2.2 [Improved bound ("fat" case)]. *Suppose the conditions of Theorem 1.4 hold and $r < \rho_\theta^*(\beta)$. There exist constants $c_1 > c_2 > 0$ such that as $p \to \infty$, for any fixed $q > 0$, conditioning on any realization of (ℓ, μ) from the event D_p, with probability $1 - O(n^{-2})$:*

- *All singular values of $Z^{(q)}(Z^{(q)})'$ fall between $m_*^{(q)} \pm c_1\sqrt{nm^{(q)}}$;*
- $\lambda_{\max}(Z^{(q)}(Z^{(q)})') \geq m_*^{(q)} + c_2\sqrt{nm^{(q)}}$.

We now prove Theorem 1.4. We show the cases of $r > \rho_\theta^*(\beta)$ (Region of Possibility) and $r < \rho_\theta^*(\beta)$ (Region of Impossibility) separately.

2.1. *Region of possibility.* Consider the case $r > \rho_\theta^*(\beta)$. Recall that

$$q = q^*(\beta, \theta, r) = \begin{cases} 4r, & r < (\beta - \theta/2)/3, \\ \dfrac{(\beta - \theta/2 + r)^2}{4r}, & (\beta - \theta/2)/3 \leq r < 1. \end{cases}$$

Let ξ^* be the first left singular vector of $X^{(q)}$ at $q = q^*(\beta, \theta, r)$. The goal is to show

$$\cos(\ell, \xi^*) \to 1.$$

Write

$$X^{(q)}(X^{(q)})' = \|\mu^{(q)}\|^2 \ell\ell' + Z^{(q)}(Z^{(q)})' + A, \tag{2.2}$$

where $A = \ell(\mu^{(q)})'(Z^{(q)})' + Z^{(q)}\mu^{(q)}\ell'$ for short. On the right-hand side of (2.2), the first matrix has a rank 1, with $n\|\mu^{(q)}\|^2$ being the only nonzero eigenvalue and ℓ being the associated eigenvector. In our model, the expectation of $\|\mu^{(q)}\|^2$ is equal to $(\tau_p^*)^2 m_1^{(q)}$, where by tail properties of chi-square distributions (see Section B.1 of the supplementary material), $m_1^{(q)} = L_p p^{1-\beta} p^{-[(\sqrt{q}-\sqrt{r})_+]^2}$ with overwhelming probabilities. It follows that with a probability at least $1 - O(p^{-3})$,

$$n\|\mu^{(q)}\|^2 \gtrsim n(\tau_p^*)^2 \cdot m_1^{(q)} \sim L_p p^{\Delta(q,\beta,\theta,r)},$$

where $\Delta(q, \beta, \theta, r) = 1 + \theta/2 - \beta - [(\sqrt{q} - \sqrt{r})_+]^2$. Compare this with (2.2). By perturbation theory in matrices [15] [10, 14], to show the claim, it suffices to show

[15] We use [10], Proposition 1, a variant of the sine-theta theorem [14]. By that proposition, if $\hat{\xi}$ and ξ are the respective leading eigenvectors of two symmetric matrices $\hat{G}$ and G, where G has a rank 1, then $\|\hat{\xi}\hat{\xi}' - \xi\xi'\| \leq 2\|G\|^{-1}\|\hat{G} - G\|$. We also note that for two unit-norm vectors $\hat{\xi}$ and ξ, $\cos(\hat{\xi}, \xi) \to 1$ if and only if $\|\hat{\xi}\hat{\xi}' - \xi\xi'\| \to 0$ by linear algebra.

that there is a scalar a^* (either random or nonrandom) and a constant $\delta^* > 0$ so that[16]

$$\|Z^{(q)}(Z^{(q)})' + A - a^* I_n\| \le L_p p^{\Delta(q,\beta,\theta,r)-\delta^*}.$$

To this end, note that by triangle inequality,

$$\|Z^{(q)}(Z^{(q)})' + A - a^* I_n\| \le \|Z^{(q)}(Z^{(q)})' - a^* I_n\| + \|A\|.$$

The following lemma is proved in the supplementary material [27].

LEMMA 2.3. *Suppose conditions of Theorem 1.4 hold. For any fixed $q > 0$, as $p \to \infty$, conditioning on any realization of (ℓ, μ) from the event D_p, with probability $1 - O(p^{-3})$, $\|\ell(\mu^{(q)})'(Z^{(q)})' + Z^{(q)}\mu^{(q)}\ell'\| \le Cn\tau_p^*\sqrt{m_1^{(q)}}$.*

The key to the proof is to control $\|Z^{(q)}\mu^{(q)}\|_\infty$ using the Bernstein inequality [38] and to study the distribution of $Z^{(q)}$. See [27] for details.

Now, when $q > \tilde{q}(\beta, \theta, r)$, we are in the "skinny" case, combining Lemmas 2.1 and 2.3, we have that with probability at least $1 - O(p^{-3})$,

$$\begin{aligned}\|Z^{(q)}(Z^{(q)})' + A\| &\lesssim n + C\Big(n\tau_p^*\sqrt{m_1^{(q)}} + \sqrt{nm^{(q)}\log(p)}\Big)\\ &\le L_p p^{\frac{\theta}{2}+\frac{1}{2}\max\{\theta,\Delta(q,\beta,\theta,r)\}}.\end{aligned}$$

In the last inequality, we have used (2.1) which indicates that $m_0^{(q)} = L_p p^{1-q}$ and $m_1^{(q)} = L_p p^{1-\beta-[(\sqrt{q}-\sqrt{r})_+]^2}$. By the condition of $r > \rho_\theta^*(\beta)$, it can be shown that $\Delta(q, \beta, \theta, r) > \theta$, and the claim follows by letting $a^* = 0$ and $\delta^* = \frac{\Delta-\theta}{2}$. When $q < \tilde{q}(\beta, \theta, r)$, we are in the "fat" case. Combining Lemmas 2.1 and 2.3, with probability at least $1 - O(p^{-3})$,

$$\begin{aligned}\|Z^{(q)}(Z^{(q)})' + A - m_*^{(q)} I_n\| &\le C\Big(n\tau_p^*\sqrt{m_1^{(q)}} + \sqrt{nm^{(q)}\log(p)} + n^{-1}m_1^{(q)}\Big)\\ &\le L_p p^{\frac{\theta}{2}+\frac{1}{2}\max\{\theta,\Delta(q,\beta,\theta,r),1-q\}} + p^{\Delta(q,\beta,\theta,r)-\frac{3\theta}{2}}.\end{aligned}$$

By the condition of $r > \rho_\theta^*(\beta)$, it can be shown that $\Delta(q, \beta, \theta, r) > \max\{\theta, \frac{\theta+1-q}{2}\}$, and the claim follows by letting $a^* = m_*^{(q)}$ and $\delta^* = \min\{\frac{\Delta-\theta}{2}, \Delta - \frac{1-q+\theta}{2}, \frac{3\theta}{2}\}$.

[16] We have used the fact that adding/subtracting a multiple of the identity matrix does not affect the eigenvectors.

2.2. *Region of impossibility.* Consider the case $r < \rho_\theta^*(\beta)$. Fix $0 < q < 1$. Recall that $\xi^{(q)}$ is first left singular vector of $X^{(q)}$. The goal is to show that

$$\cos(\ell, \xi^{(q)}) \le c_0 < 1 \qquad \text{for any } 0 < q < 1, \tag{2.3}$$

where c_0 is a universal constant independent of q. Denote for short $H = X^{(q)}(X^{(q)})'$, $H_0 = Z^{(q)}(Z^{(q)})'$, $\xi = \xi^{(q)}$, and $\tilde{\ell} = \ell/\|\ell\|$. Let the eigenvalues of H be $\lambda_1(H) \ge \lambda_2(H) \ge \cdots \ge \lambda_n(H)$. Write

$$\tilde{\ell} = a\xi + \sqrt{1-a^2}\eta \qquad \text{for a unit-norm vector } \eta \text{ such that } \eta \perp \xi.$$

Note that $\xi' H\eta = \lambda_1 \xi'\eta = 0$ and $\eta' H\eta \ge \lambda_n$, we have $\tilde{\ell}' H\tilde{\ell} = a^2 \xi' H\xi + 2a\sqrt{1-a^2}\xi' H\eta + (1-a^2)\eta' H\eta \ge a^2\lambda_1 + (1-a^2)\lambda_n$. Rearranging it gives $a^2 \le 1 - [\lambda_1(H) - \tilde{\ell}' H\tilde{\ell}]/[\lambda_1(H) - \lambda_n(H)]$. Note that $\cos(\ell, \xi) = |a|$. So to show (2.3), it suffices to show there

$$\frac{\lambda_1(H) - \tilde{\ell}' H\tilde{\ell}}{\lambda_1(H) - \lambda_n(H)} \ge 1 - c_0^2 \qquad \text{for some constant } c_0 \in (0,1). \tag{2.4}$$

The following lemma is proved in the supplementary material [27].

LEMMA 2.4. *Suppose $r < \rho_\theta^*(\beta)$ and the conditions of Theorem 1.4 hold. As $p \to \infty$, for any fixed $q > 0$, conditioning on any realization of (ℓ, μ) from the event D_p, for any $v \in \mathcal{S}^{n-1}$, with probability $1 - O(p^{-3})$, $|v' H_0 v - m_*^{(q)}| \le C\sqrt{m^{(q)}\log(p)}$, and*

$$\|H - H_0\| = \begin{cases} o(n), & q > \tilde{q}(\beta, r, \theta) \text{ (“skinny” case)}, \\ o(\sqrt{nm^{(q)}}), & q < \tilde{q}(\beta, r, \theta) \text{ (“fat” case)}. \end{cases} \tag{2.5}$$

We now show (2.4). Similarly, let $\lambda_1(H_0) \ge \lambda_2(H_0) \ge \cdots \ge \lambda_n(H_0)$ be the eigenvalues of H_0. We prove for the cases of $q > \tilde{q}(\beta, r, \theta)$ and $q < \tilde{q}(\beta, r, \theta)$ separately. Consider the first case. This is the “skinny” case where $m^{(q)} \ll n$. By Lemma 2.1 and the first claim of Lemma 2.4, with probability $1 - O(p^{-3})$, $\lambda_1(H_0) \sim n$, $\lambda_n(H_0) \ge 0$ and $\tilde{\ell}' H_0 \tilde{\ell} = o(n)$. By the second claim of Lemma 2.4, $\|H - H_0\| = o(n)$. Combining the above with Weyl's inequality [45] [i.e., $\max_{1\le i\le n} |\lambda_i(H) - \lambda_i(H_0)| \le \|H - H_0\|$], we have

$$\lambda_1(H) - \lambda_n(H) \le n + o(n), \qquad \lambda_1(H) - \tilde{\ell}' H\tilde{\ell} \ge n - o(n).$$

Inserting these into (2.4) gives the claim.

Consider the second case. This is the “fat” case and $m^{(q)} \gg n$. By Lemma 2.2 and the first claim of Lemma 2.4, there is $\lambda_1(H_0) - \lambda_n(H_0) \le c_2\sqrt{nm^{(q)}}$ and $\lambda_1(H_0) - \tilde{\ell}' H_0 \tilde{\ell} \gtrsim c_1\sqrt{nm^{(q)}}$. Similarly, combining these with the second claim of Lemma 2.4 and Weyl's inequality, we find that

$$\lambda_1(H) - \lambda_n(H) \lesssim c_2\sqrt{nm^{(q)}}, \qquad \lambda_1(H) - \tilde{\ell}' H\tilde{\ell} \gtrsim c_1\sqrt{nm^{(q)}}.$$

Inserting these into (2.4) gives the claim.

3. Limits for signal recovery. In this section, we discuss limits for signal (support) recovery. The results are intertwined with those for clustering (namely, Theorems 1.1–1.3 and Theorem 1.5), so we prove all of them together in the later part of the section.

Compare two problems: signal recovery and clustering. One useful insight is that in the less sparse case, clustering is comparably easier than signal recovery, so we should estimate ℓ first and then use it to estimate $S(\mu)$; in the more sparse case, we should do the opposite.

For the less sparse case, we have introduced two clustering methods, $\hat{\ell}_*^{(\mathrm{sa})}$ and $\hat{\ell}_*^{(\mathrm{if})}$, in Section 1.1. They give rise to two signal recovery methods, $\hat{S}_*^{(\mathrm{sa})}$ and $\hat{S}_*^{(\mathrm{if})}$. In detail, let $y_*^{(\mathrm{sa})} = n^{-1/2}X'\hat{\ell}_*^{(\mathrm{sa})}$ and $y_*^{(\mathrm{if})} = n^{-1/2}X'\hat{\ell}_*^{(\mathrm{if})}$, and let $t_p^* = \sqrt{2\log(p)}$ be the *universal threshold* [20]. Respectively, $\hat{S}_*^{(\mathrm{sa})}$ and $\hat{S}_*^{(\mathrm{if})}$ are defined by

$$\hat{S}_*^{(\mathrm{sa})} = \{1 \le j \le p : |y_{*,j}^{(\mathrm{sa})}| \ge t_p^*\}, \qquad \hat{S}_*^{(\mathrm{if})} = \{1 \le j \le p : |y_{*,j}^{(\mathrm{if})}| \ge t_p^*\}.$$

For the more sparse case, we introduce two methods $\hat{S}_N^{(\mathrm{sa})}$ and $\hat{S}_q^{(\mathrm{if})}$; they are in fact the ones that give rise to the clustering methods $\hat{\ell}_N^{(\mathrm{sa})}$ and $\hat{\ell}_q^{(\mathrm{if})}$ we introduced in Section 1.1. In detail, recalling that $Q(j) = (2n)^{-1/2}(\|x_j\|^2 - n)$ is the column-wise χ^2-statistics,

$$\hat{S}_N^{(\mathrm{sa})} = \mathrm{argmax}_{\{S: S\subset\{1,2,\ldots,p\}, |S|=N\}}\Big\{N^{-1/2}\Big\|\sum_{j\in S} x_j\Big\|_1\Big\}, \tag{3.1}$$

and

$$\hat{S}_q^{(\mathrm{if})} = \{1 \le j \le p : Q(j) \ge \sqrt{2q\log(p)}\}.$$

For any signal (support) recovery procedure $\hat{S}$, we measure the performance by the normalized size of the difference of $\hat{S}$ and the true support

$$\mathrm{Hamm}_p(\hat{S}, \alpha, \beta, \theta) = (p\varepsilon_p)^{-1}E\big(|\hat{S}\Delta S(\mu)|\big), \tag{3.2}$$

where $A\Delta B = (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference of two sets and the expectation is with respective to the randomness of (μ, ℓ, Z). If we think $\hat{S}$ as an estimate of μ, say, $\hat{\mu}$, and $E(|\hat{S}\Delta S(\mu)|)$ is actually the Hamming distance between the two vectors $(\mathrm{sgn}(|\hat{\mu}(1)|), \ldots, \mathrm{sgn}(|\hat{\mu}(p)|))'$ and $(\mathrm{sgn}(\mu(1)), \ldots, \mathrm{sgn}(\mu(p)))'$. For this reason, we call that in (3.2) the (normalized) Hamming distance. In Section 1.6, we have introduced the curves $\alpha = \eta_\theta^{\mathrm{sig}}(\beta)$ and $\alpha = \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$. The following theorem is proved in Section 5.

THEOREM 3.1 (Statistical lower bound for signal recovery). *Fix $(\alpha, \beta, \theta) \in (0, 1)^3$ and suppose $\alpha > \eta_\theta^{\mathrm{sig}}(\beta)$. Consider the signal recovery problem for Models (1.1)–(1.2) and (1.7)–(1.8). For any $\hat{S}$ that is an estimate for the support of S, $\mathrm{Hamm}_p(\hat{S}, \alpha, \beta, \theta) \gtrsim 1$ as $p \to \infty$.*

We also have the following theorems, which are proved below.

THEOREM 3.2 (Statistical upper bound for signal recovery). *Fix* $(\alpha, \beta, \theta) \in (0,1)^3$ *and suppose* $\alpha < \eta_\theta^{\mathrm{sig}}(\beta)$. *Consider the signal recovery problem for Models* (1.1)–(1.2) *and* (1.7)–(1.8). *As* $p \to \infty$:

- $\mathrm{Hamm}_p(\hat{S}_*^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$, *if* $0 < \beta < (1-\theta)/2$.
- $\mathrm{Hamm}_p(\hat{S}_N^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$, *if* $(1-\theta)/2 < \beta < 1$ *and* $N = \lceil p\varepsilon_p \rceil$.

THEOREM 3.3 (CTUB for signal recovery). *Fix* $(\alpha, \beta, \theta) \in (0,1)^3$ *and suppose* $\alpha < \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$. *Consider the signal recovery problem for Models* (1.1)–(1.2) *and* (1.7)–(1.8). *As* $p \to \infty$:

- $\mathrm{Hamm}_p(\hat{S}_*^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$, *if* $0 < \beta < (1-\theta)/2$.
- $\mathrm{Hamm}_p(\hat{S}_*^{(\mathrm{if})}, \alpha, \beta, \theta) \to 0$, *if* $(1-\theta)/2 < \beta < 1/2$.
- $\mathrm{Hamm}_p(\hat{S}_q^{(\mathrm{if})}, \alpha, \beta, \theta) \to 0$, *if* $(1-\theta)/2 < \beta < 1/2$ *and* $q \geq 3$.

3.1. *Proofs of Theorems* 1.2–1.3, 1.6 *and* 3.2–3.3. We need two lemmas. The first one is on classical PCA, and it is needed for studying $\hat{\ell}_*^{(\mathrm{if})}$ and $\hat{S}_*^{(\mathrm{if})}$. The second one is a large-deviation inequality for folded normal random variables and it is needed for studying the optimization problem in (3.1).

LEMMA 3.1. *Fix* $(\alpha, \beta, \theta) \in (0,1)^3$ *such that* $(1-\theta)/2 < \beta < 1/2$ *and* $\alpha < \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$. *In Models* (1.1)–(1.2) *and* (1.7)–(1.8), *let* λ *be the first eigenvalue of* XX' *and* ξ *be the corresponding eigenvector. There is a generic constant* $\delta = \delta(\alpha, \beta, \theta) > 0$ *such that with probability* $1 - O(p^{-3})$,

$$\min\{\|\sqrt{n}\xi + \ell\|_\infty, \|\sqrt{n}\xi - \ell\|_\infty\} < p^{-\delta}.$$

The claim continues to hold if we replace the model (1.2) *by* (1.15) *for* $A = I_n$ *and* B *such that* $\max\{\|B\|, \|B^{-1}\|\} \leq L_p$.

LEMMA 3.2 (Large-deviation on Folded Normals). *As* $n \to \infty$, *for any* $h > 0$ *and* $0 \leq x \leq \sqrt{n}/\log(n)$, *and* n *independent samples* z_i *from* $N(0,1)$,

$$P\left(\left|\sum_{i=1}^n \big(|z_i + h| - E[|z_i + h|]\big)\right| \geq \sqrt{n}x\right) \leq 2\exp\big(-(1+o(1))x^2/2\big),$$

where $o(1) \to 0$, *uniformly for all* $h > 0$ *and* $0 < x \leq \sqrt{n}/\log(n)$.

We now show all theorems about upper bound. Since Theorems 1.2–1.3 are special cases of Theorem 1.6 with $B = I_p$, it suffices to show Theorems 1.6 and 3.2–3.3. As there are four methods involved, it is more convenient to prove in a way by grouping the items associated with each method together. Fixing $(\alpha, \beta, \theta) \in$

$(0,1)^3$ and viewing all statements in Theorems 1.6 and Theorems 3.2–3.3, what we need to show can be re-organized as follows (for the statements regarding $\hat{\ell}$, we need to prove that they hold for a general B where $\max\{\|B\|, \|B^{-1}\|\} \le L_p$):

- *(a). Simple Aggregation*. Consider the case $0 < \beta < (1-\theta)/2$. In this range, $\eta_\theta^{\mathrm{sig}}(\beta) < \eta_\theta^{\mathrm{clu}}(\beta)$. All we need to show is that if $\alpha < \eta_\theta^{\mathrm{clu}}(\beta)$, then $\hat{\ell}_*^{(\mathrm{sa})} = \ell$ with probability at least $1 - O(p^{-3})$, and that if additionally $\alpha < \eta_\theta^{\mathrm{sig}}(\beta)$, then $\mathrm{Hamm}_p(\hat{S}_*^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$.
- *(b). Sparse Aggregation*. Consider the case $(1-\theta)/2 < \beta < 1$. In this case, $\eta_\theta^{\mathrm{clu}}(\beta) \le \eta_\theta^{\mathrm{sig}}(\beta)$. Letting $N = \lceil p\varepsilon_p \rceil$, all we need to show is that $\mathrm{Hamm}_p(\hat{S}_N^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$ if $\alpha < \eta_\theta^{\mathrm{sig}}(\beta)$ and $\mathrm{Hamm}_p(\hat{\ell}_N^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$ if additionally $\alpha < \eta_\theta^{\mathrm{clu}}(\beta)$.
- *(c). Classical PCA*. Consider the case $(1-\theta)/2 < \beta < 1/2$ where only computationally tractable bounds are concerned and $\tilde{\eta}_\theta^{\mathrm{clu}}(\beta) = \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$. All we need to show is that if $\alpha < \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$, then $\hat{\ell}_*^{(\mathrm{if})} = \pm\ell$ with probability at least $1 - O(p^{-3})$ and that $\mathrm{Hamm}_p(\hat{S}_*^{(\mathrm{if})}, \alpha, \beta, \theta) \to 0$.
- *(d). IF-PCA*. Consider the case $1/2 < \beta < 1$ where only computationally tractable bounds are concerned and $\tilde{\eta}_\theta^{\mathrm{clu}}(\beta) \le \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$. All we need to show is that if $\alpha < \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$, then $\hat{S}_q^{(\mathrm{if})} = S(\mu)$ with probability at least $1 - O(p^{-3})$; and if additionally $\alpha < \tilde{\eta}_\theta^{\mathrm{clu}}(\beta)$, then $\mathrm{Hamm}_p(\hat{\ell}_q^{(\mathrm{if})}, \alpha, \beta, \theta) \to 0$.

Consider (a). Note that $\hat{\ell}_*^{(\mathrm{sa})} = \mathrm{sgn}(\sum_{j=1}^p x_j)$ and $\sum_{j=1}^p x_j \sim N(\|\mu\|_0\tau\ell, pI_n)$. By (3.3), $\|\mu\|_0\tau = p^{1-\beta-\alpha}(1+o(1))$. Hence, $\alpha < \eta_\theta^{\mathrm{clu}}(\beta)$ implies $\|\mu\|_0\tau \gg \sqrt{p}$, and it follows that $\hat{\ell}_*^{(\mathrm{sa})} = \ell$ with overwhelming probability. Once $\hat{\ell}_*^{(\mathrm{sa})} = \ell$, $y_*^{(\mathrm{sa})} = n^{-1/2}X'\ell \sim N(\sqrt{n}\mu, I_p)$. Noting that $\alpha < \eta_\theta^{\mathrm{sig}}(\beta)$ implies $\sqrt{n}\tau \gg 1$, we have $\mathrm{Hamm}_p(\hat{S}_*^{(\mathrm{sa})}) \to 0$ with overwhelming probability. Consider (c). The first claim is a direct result of Lemma 3.1, and the second claim can be proved similarly as in (a). Consider (d). Recall that the column-wise test statistic $Q(j)$ is approximately distributed as $N(0,1)$ for useless features and $N(\sqrt{n/2}\tau^2, 1)$ for useful features. So $\tau \gg n^{-1/4}$ will assure successful signal recovery, which translates to $\alpha < \tilde{\eta}_\theta^{\mathrm{sig}}(\beta)$. Once $\hat{S}_q^{(\mathrm{if})} = S(\mu)$, we restrict our attention to $X^{S(\mu)}$, the sub-matrix of X restricted to the columns in $S(\mu)$, and the claim of Lemma 3.1 continues to hold by adapting the proof there (see the supplementary material [27] for details). So $\mathrm{Hamm}_p(\hat{\ell}_q^{(\mathrm{if})}) \to 0$ with overwhelming probability. It remains to prove (b).

We now show (b). Define $\hat{\mu}_N^{(\mathrm{sa})}$ such that $\hat{\mu}_N^{(\mathrm{sa})}(j) = \tau_p \cdot 1\{j \in \hat{S}_N^{(\mathrm{sa})}\}$. Write $\hat{S}_N^{(\mathrm{sa})} = \hat{S}$, $\hat{\mu}_N^{(\mathrm{sa})} = \hat{\mu}$, $\hat{\ell}_N^{(\mathrm{sa})} = \hat{\ell}$ and $s_p = p\varepsilon_p$. With probability $1 - O(p^{-3})$,

$$\big|\|\mu\|_0 - s_p\big| \le C\sqrt{s_p \log(p)}. \tag{3.3}$$

Since any event of probability $O(p^{-3})$ has a negligible effect to the Hamming distances, we always condition on a fixed realization (ℓ, μ) that satisfy (3.3); so

the probabilities below are with respective to the randomness of Z. To show (b), all we need to show are:

- (b1). $\mathrm{Hamm}_p(\hat{\ell}, \alpha, \beta, \theta) \to 0$, if $\alpha < \eta_\theta^{\mathrm{clu}}(\beta)$. In this item, the matrix B may be any matrix that satisfies $\max\{\|B\|, \|B^{-1}\|\} \le L_p$.
- (b2). $\mathrm{Hamm}_p(\hat{S}, \alpha, \beta, \theta) \to 0$, if $\alpha < \eta_\theta^{\mathrm{sig}}(\beta)$. In this item, $B = I_p$.

Consider (b1) first. It suffices to show

$$n^{-1}\langle \hat{\ell}, \ell\rangle \to 1. \tag{3.4}$$

For any realized μ, we construct $\tilde{\mu}$ as follows:

- If $\|\mu\|_0 > N$, replace $\|\mu\|_0 - N$ nonzero entries by 0.
- If $\|\mu\|_0 < N$, replace $N - \|\mu\|_0$ zero entries by τ_p.

Let $\tilde{S}$ be the support of $\tilde{\mu}$. Write $X\tilde{\mu} = \|B\tilde{\mu}\| \cdot [Z(B\tilde{\mu}/\|B\tilde{\mu}\|) + \langle \mu, \tilde{\mu}/\|B\tilde{\mu}\|\rangle \ell]$, where $Z(B\tilde{\mu}/\|B\tilde{\mu}\|) \sim N(0, I_n)$ and $\langle \mu, \tilde{\mu}/\|B\tilde{\mu}\|\rangle \gtrsim \|B\|^{-1}\tau_p\sqrt{s_p} = L_p \times p^{(1-\beta-2\alpha)/2}$, with $(1-\beta-2\alpha) > 0$ in our range of interest. According to Mills' ratio [38], with probability $1 - O(p^{-3})$, the absolute value of standard normal variable is bounded by $\sqrt{6\log(p)}$, which is less than $L_p p^{(1-\beta-2\alpha)/2}$ when $p \to \infty$. It follows that with probability at least $1 - O(p^{-3})$, $\mathrm{sgn}(X\tilde{\mu}) = \ell$. Furthermore, $\ell' X\tilde{\mu} = \|X\tilde{\mu}\|_1 = \tau_p \|\sum_{j\in\tilde{S}} x_j\|_1$. Since that $\hat{\ell}' X\hat{\mu} = \|X\hat{\mu}\|_1 = \tau_p\|\sum_{j\in\hat{S}} x_j\|_1$ and that $\hat{S}$ solves the optimization problem (3.1),

$$\hat{\ell}' X\hat{\mu} \ge \ell' X\tilde{\mu}. \tag{3.5}$$

Write $\hat{\ell}' X\hat{\mu} = \langle \hat{\ell}, \ell\rangle\langle \mu, \hat{\mu}\rangle + \hat{\ell}' ZB\hat{\mu}$. We aim to obtain an upper bound for $|\hat{\ell}' ZB\hat{\mu}|$ (an upper bound for $|\ell' ZB\tilde{\mu}|$ can be obtained similarly). Denote by $(ZB)^{\hat{S}}$ the sub-matrix of ZB containing columns in $\hat{S}$. Then $|\hat{\ell}' ZB\hat{\mu}| \le \sqrt{n}\|(ZB)^{\hat{S}}\|\|\mu\| \le \sqrt{ns_p}\tau_p\|(ZB)^{\hat{S}}\|$, where $\|(ZB)^{\hat{S}}\| \le \|B\|\|Z^{\hat{S}}\| \le L_p \times \max_{|S|=N}\|Z^S\|$. By classical RMT [42], $\max_{|S|=N}\|Z^S\| \le L_p \max\{\sqrt{n}, \sqrt{s_p}\}$ with probability at least $1 - O(p^{-3})$. Inserting them into (3.5) gives

$$\langle \hat{\ell}, \ell\rangle\langle \mu, \hat{\mu}\rangle \ge n\langle \mu, \tilde{\mu}\rangle - L_p\sqrt{ns_p}\tau_p(\sqrt{n} + \sqrt{s_p}). \tag{3.6}$$

First, $\langle \mu, \hat{\mu}\rangle \le \max\{\|\mu\|_0, N\}\tau_p^2 \sim s_p\tau_p^2$. Second, by (3.3) and the definition of $\tilde{\mu}$, $\langle \mu, \tilde{\mu}\rangle = s_p\tau_p^2(1+o(1))$. Inserting these into (3.6) gives $n^{-1}\langle \hat{\ell}, \ell\rangle \ge 1 - L_p(\sqrt{n} + \sqrt{s_p})/(\tau_p\sqrt{ns_p})$. When $\alpha < \eta_\theta^{\mathrm{clu}}(\beta)$, the second term on the right-hand side is $\le p^{-\delta}$ for some $\delta = \delta(\alpha, \beta, \theta) > 0$, and (3.4) follows.

We now consider (b2). Let $\tilde{\mu}$ and $\tilde{S}$ be the same as above. Due to (3.3), $|\tilde{S} \cap S(\mu)| \ge |S(\mu)|(1 + o(1))$. It suffices to show that

$$\big|\hat{S} \cap S(\mu)\big| \ge \big|\tilde{S} \cap S(\mu)\big| - o(s_p). \tag{3.7}$$

Since $|\tilde{S}| = |\hat{S}| = N$ and that $\hat{S}$ solves the optimization (3.1),

$$G(\hat{S}) \equiv N^{-1/2} \sum_{i=1}^{n} \Big| \sum_{j \in \hat{S}} X_i(j) \Big| \geq N^{-1/2} \sum_{i=1}^{n} \Big| \sum_{j \in \tilde{S}} X_i(j) \Big| \equiv G(\tilde{S}). \tag{3.8}$$

For any $S \subset \{1, \ldots, p\}$ such that $|S| = N$, we define $w_i(S) = N^{-1/2} \sum_{j \in S} Z_i(j)$ and $h(S) = N^{-1/2} |S \cap S(\mu)| \tau_p$. It follows that $G(S) \overset{(d)}{=} \sum_{i=1}^{n} |w_i(S) + h(S)|$, where $w_i(S) \overset{i.i.d.}{\sim} N(0,1)$, $1 \leq i \leq n$. For any $h > 0$, we define the function $u(h) = E_{X \sim N(0,1)}(|X + h|)$. Let E_p be the event that $\{\max_{S \subset \{1,\ldots,p\}, |S| = N} |G(S) - u(h(S))| \leq \sqrt{6N \log(p)/n}\}$. By Lemma 3.2 and the fact that there are no more than p^N such S, $P(E_p^c) = O(p^{-3})$; so those realizations Z in E_p^c has a negligible effect. Combining it with (3.8) gives

$$u\big(h(\hat{S})\big) \geq u\big(h(\tilde{S})\big) - L_p \sqrt{s_p \log(p)/n}. \tag{3.9}$$

The following lemma is proved in the supplementary material [27].

LEMMA 3.3. *There exists a constant $C > 0$ such that for any $0 < h_1 < h_2$, $u(h_2) - u(h_1) \geq C \min\{(h_2 - h_1), (h_2 - h_1)^2\}$.*

Since $h(S) \leq h(\tilde{S})$ for any S with $|S| = N$, by Lemma 3.3,

$$u\big(h(\tilde{S})\big) \geq u\big(h(\hat{S})\big) - C \min\big\{h(\tilde{S}) - h(\hat{S}), \big[h(\tilde{S}) - h(\hat{S})\big]^2\big\}. \tag{3.10}$$

We combine (3.9)–(3.10). It yields that

$$0 \leq \frac{h(\tilde{S}) - h(\hat{S})}{\sqrt{s_p}\tau_p} \leq L_p \tau_p^{-1} \max\big\{\big(\log(p)/n\big)^{1/2}, \big(\log(p)/ns_p\big)^{1/4}\big\}. \tag{3.11}$$

The assumption $\alpha < \eta_\theta^{\mathrm{sig}}(\beta)$ implies $\tau_p \leq p^{-\delta} \min\{n^{-1/2}, (ns_p)^{-1/4}\}$. So the right-hand side of (3.11) is $o(1)$. Then (3.7) follows.

4. Limits for hypothesis testing. The goal for (global) hypothesis testing is to test a null hypothesis

$$H_0^{(p)} : X_i \overset{i.i.d.}{\sim} N(0, I_p), \qquad 1 \leq i \leq n, \tag{4.1}$$

against a specific alternative in the complement of the null,

$$H_1^{(p)} : X_i\text{'s are generated from Models (1.1)–(1.2) and (1.7)–(1.8).} \tag{4.2}$$

We consider three different tests.

The first test $\hat{T}_*^{(\mathrm{sa})}$ is connected to the idea of simple aggregation. Recall that $\bar{x}$ is the average of all columns. The idea is to test whether $E[\bar{x}] = 0$ or not using the classical χ^2. This test rejects $H_0^{(p)}$ if and only if

$$(2n)^{-1/2}\big[p\|\bar{x}\|^2 - n\big] \geq 2\sqrt{2\log(p)}.$$

The second test $\hat{T}_N^{(\mathrm{sa})}$ is connected to sparse aggregation. Let $\hat{S}_N^{(\mathrm{sa})}$ be as in (3.1). This test rejects $H_0^{(p)}$ if and only if

$$N^{-1/2}\Big\|\sum_{j\in\hat{S}_N^{(\mathrm{sa})}} x_j\Big\|_1 \geq \sqrt{2/\pi}\,n + \sqrt{2n(N+2)\log(p)}.$$

The third test $\hat{T}^{(\mathrm{hc})}$ is connected to the Higher Criticism in Donoho and Jin [17]. Recalling that $Q(j) = (2n)^{-1/2}(\|x_j\|^2 - n)$ are the column-wise χ^2-tests, the idea is to test whether some of the $Q(j)$'s have nonzero means:

- For $1 \leq j \leq p$, obtain a P-value $\pi_j = P\{(2n)^{-1/2}[\chi^2_n(0) - n] \geq Q(j)\}$.
- Sort the P-values in the ascending order: $\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(p)}$.
- Compute the Higher Criticism statistic $HC_p^* = \max_{\{1\leq i\leq p/2\}} HC_{p,i}$, where $HC_{p,i} \equiv \sqrt{p}[(i/p) - \pi_{(i)}]/[\pi_{(i)}(1-\pi_{(i)})]^{1/2}$.

The test rejects $H_0^{(p)}$ if and only if $HC_p^* \geq 2\sqrt{2\log\log(p)}$.

The test $\hat{T}_N^{(\mathrm{sa})}$ is similar to a test in [5], which is designed for the case that there is (unknown) dependence among features and so the test is more complicated than ours. The other two tests are newly proposed.

For any testing procedure $\hat{T}$ that tests $H_1^{(p)}$ against $H_0^{(p)}$, we measure the performance by the sum of Type I and Type II errors:

$$\mathrm{Err}(\hat{T},\alpha,\beta,\theta) = P_{H_0^{(p)}}\big(\hat{T} \text{ rejects } H_0^{(p)}\big) + P_{H_1^{(p)}}\big(\hat{T} \text{ accepts } H_0^{(p)}\big), \tag{4.3}$$

where the probabilities are with respective to the randomness of (ℓ, μ, Z).

In Section 1.6, we have introduced two curves $\eta_\theta^{\mathrm{hyp}}(\beta)$ and $\tilde{\eta}_\theta^{\mathrm{hyp}}(\beta)$. The following theorem is proved in Section 5.

THEOREM 4.1 (Statistical lower bound for hypothesis testing). *Fix $(\alpha,\beta,\theta) \in (0,1)^3$ with $\alpha > \eta_\theta^{\mathrm{hyp}}(\beta)$. Consider the testing problem (4.1)–(4.2) for Models (1.1)–(1.2) and (1.7)–(1.8). For any test $\hat{T}$, $\mathrm{Err}(\hat{T},\alpha,\beta,\theta) \gtrsim 1$ as $p \to \infty$.*

Consider the upper bound. By the definitions [see (1.16)], when $\alpha < \eta_\theta^{\mathrm{hyp}}(\beta)$, we have either $\alpha < \eta_\theta^{\mathrm{hyp},1}(\beta)$ or $\alpha < \rho_\theta^{\mathrm{hyp},2}(\beta)$, or both.

THEOREM 4.2 (Statistical upper bound for hypothesis testing). *Fix $(\alpha,\beta,\theta) \in (0,1)^3$ such that $\alpha < \eta_\theta^{\mathrm{hyp}}(\beta)$. Consider the testing problem (4.1)–(4.2) for Models (1.1)–(1.2) and (1.7)–(1.8). As $p \to \infty$:*

- $\mathrm{Err}(\hat{T}_*^{(\mathrm{sa})},\alpha,\beta,\theta) \to 0$ *if* $\alpha < \eta_\theta^{\mathrm{hyp},1}(\beta)$.
- $\mathrm{Err}(\hat{T}_N^{(\mathrm{sa})},\alpha,\beta,\theta) \to 0$ *if* $\alpha < \eta_\theta^{\mathrm{hyp},2}(\beta)$ *and we take* $N = \lceil p\varepsilon_p \rceil$.

THEOREM 4.3 (CTUB for hypothesis testing). *Fix* $(\alpha, \beta, \theta) \in (0,1)^3$ *such that* $\alpha < \tilde{\eta}_\theta^{\mathrm{hyp}}(\beta)$. *Consider the testing problem* (4.1)–(4.2) *for Models* (1.1)–(1.2) *and* (1.7)–(1.8). *As* $p \to \infty$:

- $\mathrm{Err}(\hat{T}_*^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$ *if* $0 < \beta < 1/2$.
- $\mathrm{Err}(\hat{T}^{(\mathrm{hc})}, \alpha, \beta, \theta) \to 0$ *if* $1/2 < \beta < 1$.

4.1. *Proofs of Theorems* 4.2–4.3. Similarly, as three tests are involved, it is more convenient to prove the results in a way by grouping items associated with each test separately. Fixing $(\alpha, \beta, \theta) \in (0,1)^3$ and viewing the two theorems, the following is what we need to show:

- (*Simple Aggregation*). When $\alpha < \eta_\theta^{\mathrm{hyp},1}(\beta)$, $\mathrm{Err}(\hat{T}_*^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$.
- (*Sparse Aggregation*). When $\alpha < \eta_\theta^{\mathrm{hyp},2}(\beta)$, $\mathrm{Err}(\hat{T}_N^{(\mathrm{sa})}, \alpha, \beta, \theta) \to 0$ if we take $N = \lceil p\varepsilon_p \rceil$.
- (*HC*). When $1/2 < \beta < 1$ and $\alpha < \theta/4$, $\mathrm{Err}(\hat{T}^{(\mathrm{hc})}, \alpha, \beta, \theta) \to 0$.

In the above, (c) is an easy extension of [17], so we omit its proof. Below, we prove (a) and (b). Consider (a). $\hat{T}_*^{(\mathrm{sa})}$ is defined through $\bar{x}$, where $\bar{x} \sim N(p^{-1}\|\mu\|_0\tau, p^{-1}I_n)$. So the claim follows directly from the tail probability of chi-square distributions. Consider (b). Under $H_0^{(p)}$, for each fixed S with $|S| = N$, we can write $N^{-1/2}\|\sum_{j\in S} x_j\|_1 = \sum_{i=1}^n |w_i|$, where w_i's are *i.i.d.* standard normal variables. Since $E(|w_i|) = \sqrt{2/\pi}$, by Lemma 3.2, $\hat{T}_N^{(\mathrm{sa})} \le \sqrt{(2/\pi)}n + \sqrt{2n(N+2)\log(p)}$ with probability $1 - O(p^{-2})$. On the other hand, it is seen that

$$\hat{T}_N^{(\mathrm{sa})} = \max_{\ell\in\{\pm1\}^n, \mu\in\{0,1\}^p, \|\mu\|_0 = N} \ell' X(\mu/\|\mu\|).$$

Under $H_1^{(p)}$, let $\tilde{\mu}$ be defined in the same way as Section 3.1 and so $\hat{T}_N^{(\mathrm{sa})} \ge \ell' X\tilde{\mu}/\|\tilde{\mu}\| = n\langle\mu, \tilde{\mu}\rangle/\|\tilde{\mu}\| + \ell' Z\tilde{\mu}/\|\tilde{\mu}\|$. Since $|S(\mu)| \sim p\varepsilon_p$ with probability $1 - O(p^{-2})$, $\langle\mu, \tilde{\mu}\rangle/\|\tilde{\mu}\| \ge \|\mu\|(1 + o(1))$. Moreover, $\|\ell' Z\tilde{\mu}\| \le C\sqrt{n}\|\tilde{\mu}\|(\sqrt{n} + \sqrt{N})$ with probability $1 - O(p^{-3})$, by classical RMT [42]. Combining the above gives $\hat{T}_N^{(\mathrm{sa})} \gtrsim n\|\mu\| - C\sqrt{n}(\sqrt{n} + \sqrt{p\varepsilon_p}) \ge n\|\mu\|/2$, where the last inequality is because $\alpha < \eta_\theta^{\mathrm{hyp},2}(\beta)$ implies $\tau_p \gg \max\{n^{-1/2}, s_p^{-1/2}\}$. Therefore, $\hat{T}_N^{(\mathrm{sa})} \gtrsim n\tau_p\sqrt{N}/2 \gg \max\{n, \sqrt{nN\log(p)}\}$, and the claim follows.

5. Proofs of Theorems 1.1, 1.5, 3.1 and 4.1 (lower bounds).

5.1. *Proof of Theorem* 1.1. For each $1 \le i \le n$, consider the testing of two hypotheses, $H_{-1}^{(i)} : \ell_i = -1$ versus $H_1^{(i)} : \ell_i = 1$. Let $f_\pm^{(i)}$ be the joint density of X under $H_{\pm1}^{(i)}$, respectively. Since $\ell_i = \pm1$ with equal probabilities, it follows from the connection between L^1-distance and the sum of Type I and Type II testing

errors [41] that for any clustering procedure $\hat{\ell}$, $P(\hat{\ell}_i \neq \ell_i) \geq 1 - \|f_-^{(i)} - f_+^{(i)}\|_1$. Comparing this with the desired claim, it suffices to show that for all $1 \leq i \leq n$,

$$\|f_-^{(i)} - f_+^{(i)}\|_1 = o(1) \qquad \text{where } o(1) \to 0 \text{ and does not depend on } i. \tag{5.1}$$

We now show (5.1) for every fixed $1 \leq i \leq n$. For short, we drop the superscript "(i)" in $f_\pm^{(i)}$ and $H_{\pm 1}^{(i)}$. Recall that $X = \ell\mu' + Z$. Denote $\tilde{\ell} = \ell - \ell_i e_i$, where e_i is the ith standard basis vector of R^n; note that $\tilde{\ell}_i = 0$. By basic calculus and Fubini's theorem,

$$\begin{aligned}
\|f_- - f_+\|_1 &= E\biggl[\biggl|\int \sinh(X_i'\mu) e^{-\|\mu\|^2/2} e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}\, dF(\mu)\, dF(\tilde{\ell})\biggr|\biggr] \\
&\leq E\biggl[\int \biggl|\int \sinh(X_i'\mu) e^{-\|\mu\|^2/2} e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}\, dF(\mu)\biggr| dF(\tilde{\ell})\biggr] \\
&= \int E\biggl[\biggl|\int \sinh(X_i'\mu) e^{-\|\mu\|^2/2} e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}\, dF(\mu)\biggr|\biggr] dF(\tilde{\ell}),
\end{aligned}$$

where E denotes the expectation under the law of $X = Z$. Seemingly, to show (5.1), it suffices to show that for every realization of $\tilde{\ell}$,

$$E\biggl[\biggl|\int \sinh(X_i'\mu) e^{-\|\mu\|^2/2} e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}\, dF(\mu)\biggr|\biggr] = o(1); \tag{5.2}$$

note that the left-hand side does not depend on i and $\tilde{\ell}$. We now show (5.2) for the cases of $\beta > (1-\theta)$ and $\beta < (1-\theta)$, separately.

Consider the case $\beta < (1-\theta)$ first. Introduce $V = (n-1)^{-1/2} X'\tilde{\ell}$; note that $V \sim N((n-1)^{1/2}\mu, I_p)$. Let $g_-^{(i)}$, $g_+^{(i)}$, and $g_0^{(i)}$ be the joint densities of (X_i, V) for the cases of $X_i = -\mu + z$, $X_i = \mu + z$, and $X_i = z$, where $z \sim N(0, I_p)$ and is independent of μ [in all three cases, $V = (n-1)^{1/2}\mu + \tilde{z}$ where $\tilde{z}$ is independent of (μ, z)]. By the triangle inequality and symmetry, $\|g_-^{(i)} - g_+^{(i)}\|_1 \leq \|g_-^{(i)} - g_0^{(i)}\|_1 + \|g_+^{(i)} - g_0^{(i)}\|_1 = 2\|g_+^{(i)} - g_0^{(i)}\|_1$.

We recognize that the left-hand side of (5.2) is nothing else but $\|g_-^{(i)} - g_+^{(i)}\|_1$. Combining these, to show (5.2), it is sufficient to show

$$\|g_+^{(i)} - g_0^{(i)}\|_1 = o(1). \tag{5.3}$$

Now, denote by $A(f, g)$ the Hellinger affinity for any two densities f and g. Denote $h_p(V(j)) = \varepsilon_p e^{\sqrt{n-1}\tau_p V(j) - (n-1)\tau_p^2/2} / [1 - \varepsilon_p + \varepsilon_p e^{\sqrt{n-1}\tau_p V(j) - (n-1)\tau_p^2/2}]$. By definitions and direct calculations, $A(g_+^{(i)}, g_0^{(i)})$ equals to

$$\begin{aligned}
&\prod_{j=1}^{p} E\bigl\{\bigl[1 + h_p(V(j))\bigl(e^{\tau_p X_i(j) - \tau_p^2/2} - 1\bigr)\bigr]^{1/2}\bigr\} \\
&\qquad = \bigl(E\bigl\{\bigl[1 + h_p(V(1))\bigl(e^{\tau_p X_i(1) - \tau_p^2/2} - 1\bigr)\bigr]^{1/2}\bigr\}\bigr)^p.
\end{aligned}$$

Write for short $u = X_i(1)$ and $w = V(1)$. According to [41], page 221, for any probability densities f and g, $\|f - g\|_1 \le 2\sqrt{2 - 2A(f,g)}$. Combining this with the expression of $A(g_+^{(i)}, g_0^{(i)})$, to show (5.3), it suffices to show

$$E\big[\big(1 + h_p(w)\big[e^{\tau_p u - \tau_p^2/2} - 1\big]\big)^{1/2}\big] = 1 + o(p^{-1}). \tag{5.4}$$

Note that for any $x > -1$, $|\sqrt{1+x} - 1 - x/2| \le Cx^2$,

$$\begin{aligned}\Big|E\big[\big(1 + h_p(w)\big[e^{\tau_p u - \tau_p^2/2} - 1\big]\big)^{1/2}\big] - E\Big[1 + \frac{h_p(w)}{2}\big(e^{\tau_p u - \tau_p^2/2} - 1\big)\Big]\Big| \\ \le C E\big[h_p^2(w)\big(e^{\tau_p u - \tau_p^2/2} - 1\big)^2\big].\end{aligned} \tag{5.5}$$

On one hand, due to the independence between w and u and the fact that $E[e^{\tau_p u - \tau_p^2/2}] = 1$, we have $E[h_p(w)[e^{\tau_p u - \tau_p^2/2} - 1]] = 0$ and $E[h_p^2(w) \times (e^{\tau_p u - \tau_p^2/2} - 1)^2] = E[h_p^2(w)]E[(e^{\tau_p u - \tau_p^2/2} - 1)^2]$. On the other hand, since $h_p(w) \le \varepsilon_p e^{\sqrt{n-1}\tau_p w - (n-1)\tau_p^2/2}$, by direct calculations there is $E[h_p^2(w)] \le \varepsilon_p^2 e^{(n-1)\tau_p^2}$, and $E[(e^{\tau_p u - \tau_p^2/2} - 1)^2] = e^{\tau_p^2} - 1$. Inserting these into (5.5) and invoking $\varepsilon_p = p^{-\beta}$, $\tau_p = p^{-\alpha}$, and $n = p^\theta$,

$$\begin{aligned}&\big|E\big[\big(1 + h_p(w)\big[e^{\tau_p u - \tau_p^2/2} - 1\big]\big)^{1/2}\big] - 1\big| \\ &\quad \le C\varepsilon_p^2\big(e^{\tau_p^2} - 1\big)e^{(n-1)\tau_p^2} \le C p^{-2\beta - 2\alpha} e^{p^{\theta - 2\alpha}}.\end{aligned}$$

By the assumptions of $\alpha > \eta_\theta^{\mathrm{clu}}(\beta)$ and $\beta < (1-\theta)$, we have $2(\beta + \alpha) > 1$ and $\theta < 2\alpha$, and (5.4) follows.

We now consider the case of $\beta > (1-\theta)$. In this case, similarly, by basic algebra and Fubini's theorem, the left-hand side of (5.2) is no greater than

$$\begin{aligned}&E\Big[\int |\sinh(X_i'\mu)| e^{-\|\mu\|^2/2} e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}\, dF(\mu)\Big] \\ &\quad = \int E\big[|\sinh(X_i'\mu)| e^{-\|\mu\|^2/2} e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}\big]\, dF(\mu) \\ &\quad = \int E|\sinh(X_i'\mu)| e^{-\|\mu\|^2/2}\, dF(\mu),\end{aligned} \tag{5.6}$$

where in the last step we have used the independence between X_i and $\{X_k : k \ne i, 1 \le k \le n\}$, and that $E[e^{\tilde{\ell}' X\mu - (n-1)\|\mu\|^2/2}] = 1$. Finally, let A_p be the event of $\{\mu : \|\mu\|_0/(p\varepsilon_p) \le 2\}$, and write

$$\int E|\sinh(X_i'\mu)| e^{-\|\mu\|^2/2}\, dF(\mu) = I + II, \tag{5.7}$$

where $I = \int (E|\sinh(X_i'\mu)| e^{-\|\mu\|^2/2} \cdot 1_{A_p})\, dF(\mu)$, and $II = \int (E|\sinh(X_i'\mu)| \times e^{-\|\mu\|^2/2} \cdot 1_{A_p^c})\, dF(\mu)$.

By the Cauchy–Schwarz inequality, $(E|\sinh(X_i'\mu)|)^2 \le E[(\sinh(X_i'\mu))^2] = (e^{2\|\mu\|^2} - 1)/2$ for any realized μ in A_p. Combining this with basic algebra, it follows that $I \le \int(\sqrt{\sinh(\|\mu\|^2)} \cdot 1_{A_p})\, dF(\mu) \le \sqrt{\sinh(2p\varepsilon_p\tau_p^2)}$, where in the last step, we have used the fact that over the event A_p, $\|\mu\|^2 \le 2p\varepsilon_p\tau_p^2$. By our assumption of $\tau_p = p^{-\alpha}$, $\varepsilon_p = p^{-\beta}$ and $\alpha > \eta_\theta^{\mathrm{clu}}(\beta) = (1-\beta)/2$, $p\varepsilon_p\tau_p^2 = o(1)$. Combining these gives

$$|I| \le o(1). \tag{5.8}$$

At the same time, since $|\sinh(x)| \le \cosh(x)$ for any x,

$$II \le \int \big(E\cosh(X_i'\mu)e^{-\|\mu\|^2/2} \cdot 1_{A_p^c}\big)\, dF(\mu) = P(A_p^c); \tag{5.9}$$

note that $P(A_p^c) = o(1)$, so $II = o(1)$. We insert (5.8)–(5.9) into (5.7), and find that $\int E|\sinh(X_i'\mu)|e^{-\|\mu\|^2/2}\, dF(\mu) = o(1)$. Then (5.2) follows from (5.6).

5.2. *Proof of Theorem 1.5.* Recall that Z has *i.i.d.* entries from $N(0,1)$. By elementary statistics[17] and conditions on A and B, there is a nonstochastic term c_p such that (a) $c_p^{-1} \le L_p$, (b) there is a random matrix $W \in R^{n,p}$ such that $c_pZ + W$ has the same distribution of AZB [W is independent of (ℓ, μ, Z)]. Compare two experiments

$$\text{Experiment 1.} \qquad X = \ell\mu' + c_pZ,$$
$$\text{Experiment 2.} \qquad X = \ell\mu' + c_pZ + W.$$

Fixing $1 \le i \le n$, consider the testing of two hypotheses, $H_{-1}^{(i)} : \ell_i = -1$ versus $H_1^{(i)} : \ell_i = 1$. Let $f_\pm^{(i)}$ be the joint density of X under $H_\pm^{(i)}$, respectively, for Experiment 1, and let $g_\pm^{(i)}$ be the joint density of X under $H_\pm^{(i)}$, respectively, for Experiment 2. By Neyman–Pearson's fundamental lemma on testing [41], for any clustering procedure $\hat{\ell}$, tight lower bounds for $P(\hat{\ell}_i \ne \ell_i)$ (expected Hamming error at location i) associated with the two experiments are $1 - \|f_+^{(i)} - f_-^{(i)}\|_1$ and $1 - \|g_+^{(i)} - g_-^{(i)}\|_1$, respectively, where $\|f - g\|_1$ denotes the L^1-distance between two densities f and g. Le Cam's idea can be solidified as follows.

THEOREM 5.1 (Monotonicity of L^1-distance). $\|g_+^{(i)} - g_-^{(i)}\|_1 \le \|f_+^{(i)} - f_-^{(i)}\|_1$.

[17] Note that ZB has the same distribution as $\tilde{c}_p\tilde{Z} + \tilde{W}$, where $\tilde{Z}$ has *i.i.d.* normal entries, $\tilde{c}_p$ is half of the minimum eigenvalue of BB', and the columns of $\tilde{W}$ follow $N(0, BB' - \tilde{c}_pI_p)$ distribution. Similar analysis for $A(\tilde{c}_p\tilde{Z} + \tilde{W})$ gives the result.

Using this, Theorem 1.5 follows directly from the proof of Theorem 1.1.

It remains to show Theorem 5.1. Without loss of generality, we assume $i = 1$, and drop the superscripts in $g_{\pm}^{(i)}$ and $f_{\pm}^{(i)}$ for simplicity. Let $a \in R^{n-1}$ be the vector such that $a_i \overset{i.i.d.}{\sim} 2\,\text{Bernoulli}(1/2) - 1$. For any realization of a, let $\ell_{\pm} = \ell_{\pm}(a) \in R^n$ be the vectors of $(\pm 1, a')'$, respectively. Let $F(a)$, $F(\mu)$, and $F(w)$ be the CDF of a and μ, respectively, and let $h(z)$ be the (joint) density of the matrix Z. It follows that $g_{\pm}(x) = \int h(x - \ell_{\pm}(a)\mu' - w)\,dF(a)\,dF(\mu)\,dF(w)$, $x \in R^{n,p}$, and $\|g_+ - g_-\|_1$ equals to

$$\int \left| \int \big[h\big(x - \ell_+(a)\mu' - w\big) - h\big(x - \ell_-(a)\mu' - w\big)\big]\,dF(a)\,dF(\mu)\,dF(w) \right| dx.$$

Using Fubini's theorem, this is no greater than $\int G(w)\,dF(w)$, where $G(w) = |\int h(x - \ell_+(a)\mu' - w) - h(x - \ell_-(a)\mu' - w)\,dF(a)\,dF(\mu)|\,dx$. Note that for any fixed $w \in R^{n,p}$, $A(w)$ does not depend on w and equals to $\|f_+ - f_-\|_1$, and the claim follows.

5.3. *Proof of Theorem* 3.1. For each $1 \le j \le p$, consider the testing of two hypotheses, $H_0^{(j)} : \mu(j) = 0$ versus $H_1^{(j)} : \mu(j) = \tau_p$. Let $f_0^{(j)}$ and $f_1^{(j)}$ be the joint density of X under $H_0^{(j)}$ and $H_1^{(j)}$, respectively. Since $P(\mu(j) = \tau_p) = \varepsilon_p$, it follows from the connection between L^1-distance and the sum of Type I and Type II testing errors [41] that for any clustering procedure $\hat{\mu}$,

$$\begin{aligned}
&P\big(\operatorname{sgn}(\hat{\mu}(j)) \neq \operatorname{sgn}(\mu(j))\big) \\
&\quad = (1 - \varepsilon_p) P\big(\hat{\mu}(j) \neq 0 | \mu(j) = 0\big) + \varepsilon_p P\big(\hat{\mu}(j) = 0 | \mu(j) = \tau_p\big) \\
&\quad \ge (1/2)\big[1 - \big\|(1 - \varepsilon_p) f_0^{(j)} - \varepsilon_p f_1^{(j)}\big\|_1\big] \\
&\quad \ge \varepsilon_p\big[1 - (1/2)\big\|f_0^{(j)} - f_1^{(j)}\big\|_1\big],
\end{aligned}$$

where in the last step we have used $\|(1 - \varepsilon_p) f_0^{(j)} - \varepsilon_p f_1^{(j)}\|_1 = \|(1 - 2\varepsilon_p) f_0^{(j)} + \varepsilon_p (f_0^{(j)} - f_1^{(j)})\|_1 \le (1 - 2\varepsilon_p) + \varepsilon_p \|f_0^{(j)} - f_1^{(j)}\|_1$. Comparing this with the desired claim, it suffices to show that for all $1 \le j \le p$,

$$\big\|f_0^{(j)} - f_1^{(j)}\big\|_1 = o(1) \qquad \text{where } o(1) \to 0 \text{ and does not depend on } j. \tag{5.10}$$

We now show (5.10) for every fixed $1 \le j \le p$. We first consider the case $\beta < 1 - \theta$. For short, we drop the superscript "(j)" in $f_0^{(j)}$ and $f_1^{(j)}$. Recall that $X = \ell\mu' + Z = [x_1, x_2, \ldots, x_p]$ and let $\tilde{\mu} = \mu - \mu(j) e_j$, where e_j is the jth standard basis vector of R^p; note that $\tilde{\mu}(j) = 0$. Let E denote the expectation under the law

of $X = Z$. By basic calculus and Fubini's theorem,

$$\begin{aligned}\|f_0 - f_1\|_1 &= E\Big[\Big|\int [1 - e^{\tau_p\langle \ell, x_j\rangle - n\tau_p^2/2}] e^{\ell' X\tilde{\mu} - n\|\tilde{\mu}\|^2/2}\, dF(\tilde{\mu})\, dF(\ell)\Big|\Big] \\ &\le \int E\big[|1 - e^{\tau_p\langle \ell, x_j\rangle - n\tau_p^2/2}| e^{\ell' X\tilde{\mu} - n\|\tilde{\mu}\|^2/2}\big]\, dF(\tilde{\mu})\, dF(\ell) \qquad (5.11)\\ &= \int E\big[|1 - e^{\tau_p\langle \ell, x_j\rangle - n\tau_p^2/2}|\big]\, dF(\ell),\end{aligned}$$

where in the last step, we have used the fact that x_j and $X\tilde{\mu}$ are independent and that $E[e^{\ell' X\tilde{\mu} - n\|\tilde{\mu}\|^2/2}] = 1$. Additionally, note that $E[|1 - e^{\tau_p\langle \ell, x_j\rangle - n\tau_p^2/2}|]$ does not depend on ℓ. Denote $z = n^{-1/2}\langle \ell, x_j\rangle$; note that $z \sim N(0,1)$. Inserting these into (5.11) gives

$$\|f_0 - f_1\|_1 = E_0\big[|1 - e^{\sqrt{n}\tau_p z - n\tau_p^2/2}|\big], \qquad (5.12)$$

where E_0 denotes the expectation under the law of $z \sim N(0,1)$. By the conditions of $\alpha > \eta_\theta^{\mathrm{sig}}(\beta)$ and $\beta < (1-\theta)$, we have $\alpha > \theta/2$, and $n\tau_p^2 = p^{\theta - 2\alpha} = o(1)$. In this simple setting, it is seen that $E_0[|1 - e^{\sqrt{n}\tau_p z - n\tau_p^2/2}|] = o(1)$. Combining (5.10)–(5.12) gives the claim.

We now consider the case $\beta > (1-\theta)$. In this case, $\eta_\theta^{\mathrm{sig}}(\beta) = \eta_\theta^{\mathrm{hyp}}(\beta)$, so intuitively, the claim follows by the argument that "as long as it is impossible to have (global) hypothesis testing, it is impossible to identify the signals." Still, for mathematical rigor, it is desirable to provide a proof using the L^1-distance. Similar to that in the proof on the lower bound for global testing, write $\mu = \tilde{\mu} + \mu(j)e_j$ and let $d_p = (6p\varepsilon_p \log(p))^{1/2}$, A_s be the event $\{\|\tilde{\mu}\|_0 = s\}$ and F_s be the conditional distribution of $\tilde{\mu}$ given the event of A_s, $1 \le s \le p$. Define $a_s = \int e^{\tau_p\langle \ell, x_j\rangle - n\tau_p^2/2} e^{\ell' X\tilde{\mu} - n\|\tilde{\mu}\|^2/2}\, dF_s(\tilde{\mu})\, dF(\ell)$ and $\tilde{a}_s = \int e^{\ell' X\tilde{\mu} - n\|\tilde{\mu}\|^2/2}\, dF_s(\tilde{\mu})\, dF(\ell)$. It suffices to show that for all s such that $|s - p\varepsilon_p| \le d_p$ that

$$E\big[(a_s - \tilde{a}_s)^2\big] = o(1). \qquad (5.13)$$

Let ν be an independent duplicate of μ. By similar arguments and noting that $\mu'\nu = \tilde{\mu}'\tilde{\nu} + \tau_p^2$ and $\tilde{\mu}'\nu = \tilde{\mu}'\tilde{\nu}$, we have $E[a_s^2] = \int [\cosh(\tilde{\mu}'\tilde{\nu} + \tau_p^2)]^n\, dF_s(\tilde{\mu})\, dF_s(\tilde{\nu})$, $E[\tilde{a}_2^2] = \int [\cosh(\tilde{\mu}'\tilde{\nu})]^n\, dF_s(\tilde{\mu})\, dF_s(\tilde{\nu})$, and the cross term $E[\tilde{a}_s a_s] = \int [\cosh(\tilde{\mu}'\tilde{\nu})]^n\, dF_s(\tilde{\mu})\, dF_s(\tilde{\nu})$. Combining these terms and noting that $\cosh(x+y) = \cosh(x)[1 + \tanh(x)\tanh(y)]$, there is

$$E\big[(a_s - \tilde{a}_s)^2\big] = \int \big[\cosh(\tilde{\mu}'\tilde{\nu})\big]^n \big\{\big[1 + \tanh\big(\tau_p^2\big)\tanh\big(\tilde{\mu}'\tilde{\nu}\big)\big]^n - 1\big\}\, dF_s(\tilde{\mu})\, dF_s(\tilde{\nu}).$$

Now, over the event $\{(\tilde{\mu}, \tilde{\nu}) : \|\tilde{\mu}\|_0 = \|\tilde{\nu}\|_0 = s\}$, where $s \sim p\varepsilon_p$, we have $|\tilde{\mu}'\tilde{\nu}| \le s\tau_p^2 \lesssim p\varepsilon_p\tau_p^2 \le p^{(1-\beta-\theta)/2}$; note that by the assumption of $r > \eta_\theta^{\mathrm{hyp}}(\beta)$

and $\beta > (1-\theta)$, the exponent $(1-\beta-\theta)/2 < 0$. As a result, it is seen that $\tanh(\tilde{\mu}'\tilde{\nu})\tanh(\tau_p^2) \lesssim \tilde{\mu}'\tilde{\nu}\tau_p^2 \lesssim p\varepsilon_p\tau_p^4$, where $p\varepsilon_p\tau_p^4 = o(n^{-1})$ by the assumption of $\alpha > \eta_\theta^{\mathrm{sig}}(\beta)$. Inserting this into (5.13) gives

$$E\big[(a_s - \tilde{a}_s)^2\big] = o(1)\cdot \int \big[\cosh\big(\tilde{\mu}'\tilde{\nu}\big)\big]^n \, dF_s(\tilde{\mu})\, dF_s(\tilde{\nu}). \tag{5.14}$$

According to (5.17)–(5.18) in Section 5.4, the second term on the right-hand side of (5.14) is $1 + o(1)$. This gives the claim.

5.4. *Proof of Theorem* 4.1. Recall that $X = \ell\mu' + Z$. Let $f_0(X)$ and $f_1(X)$ be the joint density of Z and X, respectively. It is sufficient to show that as $p \to \infty$, under the conditions of Theorem 4.1,

$$\|f_1 - f_0\|_1 \to 0. \tag{5.15}$$

Recall that $\|\mu\|_0$ and $\|\mu\|$ denote the L^0-norm and the L^2-norm of μ respectively. For $1 \le s \le p$, let A_s be the event $A_s = \{\|\mu\|_0 = s\}$, $F(\ell)$ and $F(\mu)$ be the distributions of ℓ and μ, respectively, and let $F_s(\mu)$ be the conditional distribution of μ given the event of A_s. Introduce a constant $d_p = (6p\varepsilon_p\log(p))^{1/2}$, a set $D_p = \{s : |s - p\varepsilon_p| < d_p\}$, and functions $a_s(X) = \int e^{\ell' X\mu - n\|\mu\|^2/2}\, dF_s(\mu)\, dF(\ell)$, $1 \le s \le p$. Let E be the expectation under the law of $X = Z$. It is seen that $f_1(X)/f_0(X) = \int e^{\ell' X\mu - n\|\mu\|^2/2}\, dF(\mu)\, dF(\ell) = \sum_{s=1}^p P(A_s)a_s(X)$, and so $\|f_1 - f_0\|_1$ equals to

$$E\left|\sum_{s=1}^p P(A_s)\big(a_s(X) - 1\big)\right| \le \sum_{D_p} P(A_s)E\big[|a_s(X) - 1|\big] + rem, \tag{5.16}$$

where $rem = \sum_{D_p^c} P(A_s)E[|a_s(X) - 1|]$. Since $E[|a_s - 1|] \le E[a_s] + 1 = 2$, $rem \le \sum_{D_p^c} 2P(A_s) \le 2P(\|\mu\|_0 \in D_p^c)$. Note that $\|\mu\|_0 \sim \mathrm{Binomial}(p, \varepsilon_p)$, where $p\varepsilon_p = p^{1-\beta}$ with $0 < \beta < 1$, it follows from basic statistics that $rem = o(1)$. At the same time, by the Cauchy–Schwarz inequality, $(E[|a_s(X) - 1|])^2 \le E[(a_s(X) - 1)^2] = E[a_s^2(X)] - 1$. Combining these with (5.16), to show (5.15), it suffices to show that

$$E\big[(a_s^2(X)\big] \le 1 + o(1) \qquad \forall s \in D_p, \tag{5.17}$$

where $o(1) \to 0$ uniformly for all such s as $p \to \infty$.

We now show (5.17). Fix an $s \in D_p$. Let $\nu \in R^p$ be an independent copy of μ, and let $F_s(\nu)$ be the distribution of $(\nu|\{\|\nu\|_0 = s\})$. Using basic statistics and the independence of X_i,

$$a_s^2(X) = \int e^{-n\|\mu\|^2/2 - n\|\nu\|^2/2} \prod_{i=1}^n \big[\cosh\big(\mu' X_i\big)\cosh\big(\nu' X_i\big)\big]\, dF_s(\mu)\, dF_s(\nu).$$

First, by the independence of X_i and basic statistics, $E[a_s^2(X)]$ equals to

$$\int [\cosh(\mu'\nu)]^n \, dF_s(\mu)\, dF_s(\nu) = \sum_{k=0}^{n} \int \binom{n}{k} \frac{e^{(2k-n)\mu'\nu}}{2^n} \, dF_s(\mu)\, dF_s(\nu). \tag{5.18}$$

Recalling that any nonzero entry of μ or ν is τ_p, it is seen that over the event $\{\|\mu\|_0 = \|\nu\|_0 = s\}$, $\tau_p^{-2}\langle \mu, \nu\rangle$ is distributed as a hyper-geometric distribution $H(p, s, s)$. Write $\hat{\varepsilon}_p = s/p$. As $s \in D_p$, $\hat{\varepsilon}_p \sim \varepsilon_p$. Following [2], there is a σ-algebra $\mathcal{B}$ and a random variable $b \sim \text{Binomial}(s, \hat{\varepsilon}_p)$ such that $\tau_p^{-2}\langle \mu, \nu\rangle$ has the same distribution as that of $E[b|\mathcal{B}]$. Using Jensen's inequality, $e^{(2k-n)\mu'\nu} \le E[e^{(2k-n)\tau_p^2 b}|\mathcal{B}]$, for $0 \le k \le n$. It follows that

$$\begin{aligned} E\int e^{(2k-n)\mu'\nu} \, dF_s(\mu)\, dF_s(\nu) &\le E\big[e^{(2k-n)\tau_p^2 b}\big] \\ &= \big(1 - \hat{\varepsilon}_p + \hat{\varepsilon}_p e^{(2k-n)\tau_p^2}\big)^s. \end{aligned} \tag{5.19}$$

Inserting (5.19) into (5.18) and rearranging,

$$E\big[a_s^2(X)\big] \le 2^{-n} \sum_{k=0}^{n} \binom{n}{k} \big[1 - \hat{\varepsilon}_p + \hat{\varepsilon}_p e^{(2k-n)\tau_p^2}\big]^s. \tag{5.20}$$

We now analyze the right-hand side of (5.20). Denote S by $\{1, 2, \ldots, n\}$. We split S as the union of three disjoint subsets $S = S_1 \cup S_2 \cup S_3$, where $S_1 = \{k \in S : |2k - n| < \sqrt{n}\log(n)\}$, $S_3 = \{k \in S : |2k - n| > n \wedge \sqrt{2\log(n)np\varepsilon_p}\}$.

Also, let $\tilde{\tau}_p = p^{-\eta_\theta^{\text{hyp}}(\beta)}$. By our assumption of $\alpha > \eta_\theta(\beta)$, there is a constant $\delta = \delta(\theta, \alpha) > 0$ such that $\tau_p^2 = p^{-\delta}\tilde{\tau}_p^2$. We also claim that when $\alpha > \eta_\theta^{\text{hyp}}(\beta)$, $\tau_p^2|2k - n| = o(1)$ for any $k \in S_1 \cup S_2$. In fact, by definitions and direct calculations, we have $\eta_\theta^{\text{hyp}}(\beta) > \theta/2$ when $\beta < \max\{1 - \theta, (2 - \theta)/4\}$ and $\eta_\theta^{\text{hyp}}(\beta) = (1 + \theta - \beta)/4$ otherwise. In the first case, recalling $n = p^\theta$, the claim follows since $\tau_p^2|2k - n| \le \tau_p^2 n = p^{\theta - 2\alpha}$ and $\alpha > \theta/2$. In the second case, noting that $\tau_p^2 = p^{-\delta}\tilde{\tau}_p^2 = p^{-\delta}(np\varepsilon_p)^{-1/2}$, it follows $|2k - n|\tau_p^2 \le \sqrt{2\log(n)np\varepsilon_p} \cdot (p^{-\delta}(np\varepsilon_p)^{-1/2}) = o(1)$ for all $k \in S_1 \cup S_2$, and the claim follows. Now, since for any $x \in (-1, 1)$ and $y \in R$, $1 - \hat{\varepsilon}_p + \hat{\varepsilon}_p e^x \le 1 + 2\hat{\varepsilon}_p|x| \le e^{2\hat{\varepsilon}_p|x|}$, and $1 - \hat{\varepsilon}_p + \hat{\varepsilon}_p e^y \le 1 - \hat{\varepsilon}_p + \hat{\varepsilon}_p e^{|y|} \le e^{|y|}$,

$$\begin{aligned} &\big[(1 - \hat{\varepsilon}_p + \hat{\varepsilon}_p e^{\tau_p^2(2k-n)}\big]^s \\ &\quad \le \begin{cases} \big[1 + 2\hat{\varepsilon}_p\tau_p^2|2k - n|\big]^s \le e^{2p\hat{\varepsilon}_p^2\tau_p^2|2k-n|}, & k \in S_1 \cup S_2, \\ \big(e^{\tau_p^2|2k-n|}\big)^s = e^{s\tau_p^2|2k-n|}, & k \in S_3. \end{cases} \end{aligned} \tag{5.21}$$

If we take $Y \sim \text{Binomial}(n, 1/2)$, then $P(Y = k) = 2^{-n}\binom{n}{k}$. At the same time, by de Moivre–Laplace theorem and Hoeffding inequality [38],

$$P(Y=k)\begin{cases}\sim (\pi n/2)^{-1/2} e^{-(2k-n)^2/(2n)}, & k \in S_1,\\ \le e^{-(2k-n)^2/(2n)}, & k \in S_2 \cup S_3.\end{cases} \tag{5.22}$$

Combining (5.21)–(5.22), we have the following. First, the summation over $k \in S_1$ is smaller than that that [ϕ is the probability density of $N(0,1)$]

$$(2/\sqrt{n}) \sum_{k\in S_1} e^{2p\hat{\varepsilon}_p^2\tau_p^2|2k-n|}\phi\Big(\frac{2k-n}{\sqrt{n}}\Big) \sim \int_{-\log(n)}^{\log(n)} e^{2p\hat{\varepsilon}_p^2\tau_p^2\sqrt{n}x}\phi(x)\,dx. \tag{5.23}$$

By the assumption of $\alpha > \eta_\theta^{\text{hyp}}(\beta)$ and basic algebra, we have $\alpha > (2+\theta-4\beta)/4$. It follows that $p\hat{\varepsilon}_p^2\tau_p^2\sqrt{n} \sim 2p^{-\delta}p\varepsilon_p^2\tilde{\tau}_p^2\sqrt{n} = 2\cdot p^{-\delta}\cdot p^{(2+\theta-4\beta)/2-2\eta_\theta^{\text{hyp}}(\beta)}$, where the exponent is negative. It follows that the right-hand side of (5.23) is $1+o(1)$. Second, let II be the summation over $k \in S_2$, then

$$II \le \sum_{k\in S_2} e^{2p\hat{\varepsilon}_p^2\tau_p^2|2k-n|}e^{-(2k-n)^2/(2n)} \le \sum_{k\in S_2} e^{-(2k-n)^2/(2n)}, \tag{5.24}$$

where the second inequality is because $2p\hat{\varepsilon}_p^2\tau_p^2 \le 2p^{-\delta}/\sqrt{n} \le |2k-n|/(4n)$. The right-hand side does not exceed $ne^{-\log^2(n)} = o(1)$ since $|2k-n| \ge \sqrt{n}\log(n)$. Last, we consider the summation over $k \in S_3$. We only consider the case of $\beta > (1-\theta)$ since only in this case S_3 is nonempty. Note that in this case, $n \wedge \sqrt{2\log(n)np\varepsilon_p} = \sqrt{2\log(n)np\varepsilon_p}$ and that for any $k \in S_3$, $s\tau_p^2 \lesssim p^{-\delta}p\varepsilon_p\tilde{\tau}_p^2 \le |2k-n|/(4n)$,

$$III \le \sum_{k\in S_3} e^{s\tau_p^2|2k-n|-(2k-n)^2/(2n)} \le \sum_{k\in S_3} e^{-(2k-n)^2/(4n)}, \tag{5.25}$$

which $\le ne^{-\log(n)p\varepsilon_p/2} = o(1)$. Combining (5.23)–(5.25) with (5.20) gives the claim.

6. Discussions. We have studied the statistical limits for three interconnected problems: clustering, signal recovery and hypothesis testing. For each problem, in the two-dimensional phase space calibrating the signal sparsity and strength, we identify the exact separating boundary for the Region of Possibility and Region of Impossibility. We have also derived a computationally tractable upper bound (CTUB), part of which is tight, and the other part is conjectured to be tight. Our study on the limits are extended to the case where the parameters fall exactly on the separating boundaries and the case of colored noise.

We propose several different methods, including IF-PCA. IF-PCA is a two-fold dimension reduction algorithm: we first reduce dimensions from (say) 10^4 to a few hundreds by screening, and then further reduce it to just a few by PCA. Each of the

two steps can be useful in other high-dimensional settings. Compared to popular penalization approaches, our approach has advantages for it is highly extendable and computationally inexpensive.

The work is closely related to Jin and Wang [28] but is also very different. The focus of [28] is to investigate the performance of IF-PCA with real data examples and to study the consistency theory. The primary focus here, however, is on the statistical limits for three problems including clustering. The paper is also closely related to the very interesting paper by Arias-Castro and Verzelen [5]. However, two papers are different in important ways:

- The focus of our paper is on clustering, while the focus of their paper is on hypothesis testing (without careful discussion on clustering).
- Both papers addressed signal recovery, but there are important differences: we provided the statistical lower bound but they did not; the CTUB they derived is not as sharp as ours. See Figure 2.
- Both papers studied hypothesis testing, but since the models are different, the separating boundaries (and so the proofs) are also different. See Sections 1.6 and 4 (also Figure 2) for details.
- Both papers studied the case with colored noise, besides the different focuses (clustering vs. hypothesis testing), their setting in the colored case is also different from ours. In their setting, coloration makes a substantial difference to statistical limits.

For these reasons, the methods and theory (especially that on IF-PCA) in our paper are very different from those in [5]. With that being said, we must note that since two papers have overlapping interest, it is not surprising that certain part of this paper overlaps[18] with that in [5] (e.g., some parts of the separating boundaries and some of the ideas and methods).

The paper is related to recent ideas in spectral clustering (e.g., Azizyan *et al.* [7], Chan and Hall [12]; see also [35, 36, 40, 46]). In particular, the high level idea of IF-PCA (i.e., combining feature selection with classical methods) is not new and can be found in [7, 12], but the methods and theory are different. Azizyan *et al.* [7] study the clustering problem in a closely related setting, but they use a different loss function and so the separating boundaries are also different. Chan and Hall [12] use a very different screening idea (motivated by real data analysis) and do not study phase transitions.

Our work is closely related to recent interest in the spike model (e.g., [4, 32, 44]). In particular, *mathematically*, Model (1.2) is similar to the spike model [29],

[18]Compare the critical signal strength required for successful hypothesis testing/signal recovery in our paper with those in [5], we note some discrepancies in terms of some multi-logarithmic factors. This is due to that we choose a simpler calibration than that in [5]: all the parameters (n, ε, τ) are expressed as a (constant) power of p and multi-logarithmic factors are neglected. Such a calibration makes the presentation more succinct.

and theoretical results on one can shed light on those for the other. However, two models are also different from a *scientific perspective*: (a) two models are motivated by different application problems, (b) the primary interest of Model (1.2) is on the class labels ℓ_i, which are sometimes easy to validate in real applications and (c) the primary interest of the spike model is on the feature vector μ, which is relatively hard to validate in real applications. The focus and scope of our study are very different from many recent works on the spike model, and most part of the bounds (especially those for clustering and IF-PCA) we derive are new.

This paper is also related to the recent interest on computationally tractable lower bounds and sparse PCA [9, 10], but it is also very different in terms of our focus on clustering and statistical limits. It is also related to the lower bound for hypothesis testing problem [1] and the sub-matrix detection problem [34], but the model is different. Recovering of ℓ and μ can also be interpreted as recovering a low-rank matrix from the data matrix, which is closely related to the low rank matrix recovery studies [11]. In terms of the phase transitions, the paper is closely related to [17] on signal detection, [18] on classification and [30] on variable selection, but is also very different for the primary focus here is on clustering.

For simplicity, we focus on the ARW model, where we have several assumptions such as $\ell_i = \pm 1$ equally likely, the signals have the same sign and equal strength, etc. Many of these assumptions can be largely relaxed. For example, Theorems 1.1–1.3 continue to hold if we replace the model $\mu(j) \overset{i.i.d.}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p \nu_{\tau_p}$ by that of $\mu(j) \overset{i.i.d.}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p G_p$, where G_p is a distribution supported in the interval $[a_p\tau_p, b_p\tau_p]$ with $0 < \max\{a_p^{-1}, b_p\} \le L_p$ [a multi-log(p) term]. Also, in Section 1.6, we have discussed the case where we replace $\mu(j) \overset{i.i.d.}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p \nu_{\tau_p}$ in Model (1.7) by that of $\mu(j) \overset{i.i.d.}{\sim} (1-\varepsilon_p)\nu_0 + a\varepsilon_p \nu_{-\tau_p} + (1-a)\varepsilon_p \nu_{\tau_p}$ for a constant $0 \le a \le 1/2$. Theorems 1.1–1.3 continue to hold if $a \neq 1/2$. If $a = 1/2$, the left part of the boundaries will change and the aggregation methods need to be modified. We discuss this case in detail in the supplementary material [27], Appendix D. It requires a lot of time and effort to fully investigate how broad the main theorems hold, so we leave it to the future.

The paper motivates an array of interesting problems in post-selection random matrix theory that could be future research topics. For the perspective of spectral clustering, it is of great interest to precisely characterize the limiting behavior of the singular values (bulk and the edge singular values) and leading singular vectors of the post-selection data matrix. These problems are technically very challenging, and we leave them to the future.

Our paper supports the philosophy in Donoho [16], Section 10, that simple and homely methods are just as good as more charismatic methods in Machine Learning for analyzing (real) high dimensional data.

SUPPLEMENTARY MATERIAL

Supplementary Material for "Phase transitions for high dimensional clustering and related problems" (DOI: 10.1214/16-AOS1522SUPP; .pdf). Owing to space constraints, some technical proofs and discussion are relegated a supplementary document [27]. It contains proofs of Lemmas 2.1–2.4 and 3.1–3.3, and discusses an extension of the ARW model.

REFERENCES

[1] ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[2] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—*1983. *Lecture Notes in Math.* **1117** 1–198. Springer, Berlin. MR0883646

[3] AMINI, A. and WAINWRIGHT, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE International Symposium on Information Theory* 2454–2458. IEEE, New York.

[4] AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

[5] ARIAS-CASTRO, E. and VERZELEN, N. (2014). Detection and feature selection in sparse mixture models. arXiv:1405.1478.

[6] ARTHUR, D. and VASSILVITSKII, S. (2007). `k-means++`: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[7] AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *NIPS* 2139–2147.

[8] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300. MR1325392

[9] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066.

[10] CAI, T., MA, Z. and WU, Y. (2013). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* 1–35.

[11] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

[12] CHAN, Y. and HALL, P. (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.* **105**.

[13] D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448. MR2353806

[14] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46.

[15] DETTLING, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583–3593.

[16] DONOHO, D. (2015). 50 years of data science. Manuscript.

[17] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

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

[19] DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference: Especially for rare and weak effects. *Statist. Sci.* **30** 1–25.
[20] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414
[21] DONOHO, D. L., MALEKI, A., RAHMAN, I. U., SHAHRAM, M. and STODDEN, V. (2009). Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* **11** 8–18.
[22] HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357
[23] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer, Berlin.
[24] INGSTER, Y. I., POUET, C. and TSYBAKOV, A. B. (2009). Classification of sparse high-dimensional vectors. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4427–4448. MR2546395
[25] 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
[26] JIN, J., KE, Z. T. and WANG, W. (2014). Optimal spectral clustering by higher criticism thresholding. Manuscript.
[27] JIN, J., KE, Z. T. and WANG, W. (2017). Supplementary material for "Phase transitions for high dimensional clustering and related problems." DOI:10.1214/16-AOS1522SUPP.
[28] JIN, J. and WANG, W. (2016). Influential features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359. MR3576543
[29] 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
[30] KE, Z. T., JIN, J. and FAN, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42** 2202–2242.
[31] LEE, A. B., LUCA, D. and ROEDER, K. (2010). A spectral graph approach to discovering genetic ancestry. *Ann. Appl. Stat.* **4** 179–202.
[32] LEI, J. and VU, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *Ann. Statist.* **43** 299–322.
[33] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901
[34] MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116.
[35] PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.
[36] RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178. MR2268036
[37] ROGERS, C. A. (1963). Covering a sphere with spheres. *Mathematika* **10** 157–164.
[38] SHORACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons, New York.
[39] SPIEGELHALTER, D. J. (2014). Statistics. The future lies in uncertainty. *Science* **345** 264–265.
[40] SUN, W., WANG, J., FANG, Y. et al. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Stat.* **6** 148–167.
[41] VAN DER VAART, A. (2000). *Asymptotic Statistics* **3**. Cambridge Univ. Press, Cambridge.
[42] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170
[43] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947.
[44] WANG, Z., LU, H. and LIU, H. (2014). Nonconvex statistical optimization: Minimax-optimal sparse PCA in polynomial time. arXiv:1408.5352.

[45] WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.* **71** 441–479.

[46] WITTEN, D. M. and TIBSHIRANI, R. (2012). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726.

[47] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

J. JIN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jiashun@stat.cmu.edu

Z. T. KE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: zke@galton.uchicago.edu

W. WANG
DEPARTMENT OF BIOSTATISTICS
AND EPIDEMIOLOGY
PERELMAN SCHOOL OF MEDICINE
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: wanjiew@wharton.upenn.edu