Submitted to the Annals of Statistics

INFLUENTIAL FEATURES PCA FOR HIGH DIMENSIONAL CLUSTERING

By Jiashun Jin[*] and Wanjie Wang[*]

Carnegie Mellon University

We consider a clustering problem where we observe feature vectors $X_i \in R^p$, $i = 1, 2, \ldots, n$, from K possible classes. The class labels are unknown and the main interest is to estimate them. We are primarily interested in the modern regime of $p \gg n$, where classical clustering methods face challenges.

We propose Influential Features PCA (IF-PCA) as a new clustering procedure. In IF-PCA, we select a small fraction of features with the largest Kolmogorov-Smirnov (KS) scores, obtain the first $(K-1)$ left singular vectors of the post-selection normalized data matrix, and then estimate the labels by applying the classical k-means procedure to these singular vectors. In this procedure, the only tuning parameter is the threshold in the feature selection step. We set the threshold in a data-driven fashion by adapting the recent notion of Higher Criticism. As a result, IF-PCA is a tuning-free clustering method.

We apply IF-PCA to 10 gene microarray data sets. The method has competitive performance in clustering. Especially, in three of the data sets, the error rates of IF-PCA are only 29% or less of the error rates by other methods. We have also rediscovered a phenomenon on empirical null by Efron (2004) on microarray data.

With delicate analysis, especially post-selection eigen-analysis, we derive tight probability bounds on the Kolmogorov-Smirnov statistics and show that IF-PCA yields clustering consistency in a broad context. The clustering problem is connected to the problems of sparse PCA and low-rank matrix recovery, but it is different in important ways. We reveal an interesting phase transition phenomenon associated with these problems and identify the range of interest for each.

1. Introduction. Consider a clustering problem where we have feature vectors $X_i \in R^p$, $i = 1, 2, \ldots, n$, from K possible classes. For simplicity, we assume K is small and is known to us. The class labels y_1, y_2, $\ldots$, y_n take values from $\{1, 2, \ldots, K\}$, but are unfortunately unknown to us, and the main interest is to estimate them.

Our study is largely motivated by clustering using gene microarray data. In a typical setting, we have patients from several different classes (e.g., nor-

[*] Supported in part by NSF grants DMS-1208315 and DMS-1513414.

MSC 2010 subject classifications: Primary 62H30, 62G32; secondary 62E20, 62P10.

Keywords and phrases: Empirical null, Feature selection, Gene microarray, Hamming distance, Phase transition, Post-selection spectral clustering, Sparsity.

mal, diseased), and for each patient, we have measurements (gene expression levels) on the same set of genes. The class labels of the patients are unknown and it is of interest to use the expression data to predict them.

Table 1 lists 10 gene microarray data sets (arranged alphabetically). Data sets 1, 3, 4, 7, 8, and 9 were analyzed and cleaned in Dettling (2004), Data set 5 is from Gordon et al. (2002), Data sets 2, 6, 10 were analyzed and grouped into two classes in Yousefi et al. (2010), among which Data set 10 was cleaned by us in the same way as by Dettling (2004). All the data sets can be found at `www.stat.cmu.edu/~jiashun/Research/software/GenomicsData`. The data sets are analyzed in Section 1.4, after our approach is fully introduced.

In these data sets, the true labels are given but (of course) we do not use them for clustering; the true labels are thought of as the 'ground truth' and are only used for comparing the error rates of different methods.

TABLE 1
Gene microarray data sets investigated in this paper. Note that K is small and $p \gg n$ (p: number of genes; n: number of subjects).

#	Data Name	Abbreviation	Source	K	n	p
1	Brain	Brn	Pomeroy (02)	5	42	5597
2	Breast Cancer	Brst	Wang et al. (05)	2	276	22215
3	Colon Cancer	Cln	Alon et al. (99)	2	62	2000
4	Leukemia	Leuk	Golub et al. (99)	2	72	3571
5	Lung Cancer(1)	Lung1	Gordon et al. (02)	2	181	12533
6	Lung Cancer(2)	Lung2	Bhattacharjee et al. (01)	2	203	12600
7	Lymphoma	Lymp	Alizadeh et al. (00)	3	62	4026
8	Prostate Cancer	Prst	Singh et al. (02)	2	102	6033
9	SRBCT	SRB	Kahn (01)	4	63	2308
10	SuCancer	Su	Su et al (01)	2	174	7909

View each X_i as the sum of a 'signal component' and a 'noise component':

$$X_i = E[X_i] + Z_i, \qquad Z_i \equiv X_i - E[X_i]. \tag{1.1}$$

For any numbers $a_1, a_2, \ldots, a_p$, let $\text{diag}(a_1, a_2, \ldots, a_p)$ be the $p \times p$ diagonal matrix where the i-th diagonal entry is a_i, $1 \leq i \leq p$. We assume

$$Z_i \overset{iid}{\sim} N(0, \Sigma), \qquad \text{where} \qquad \Sigma = \text{diag}\big(\sigma^2(1), \sigma^2(2), \ldots, \sigma^2(p)\big), \tag{1.2}$$

and the vector $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(p))'$ is unknown to us. Assumption (1.2) is only for simplicity: our method to be introduced below is not tied to such an assumption, and works well with most of the data sets in Table 1; see Sections 1.1 and 1.4 for more discussions.

Denote the overall mean vector by $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$. For K different vectors $\mu_1, \mu_2, \ldots, \mu_K \in R^p$, we model $E[X_i]$ by (y_i are class labels)

$$E[X_i] = \bar{\mu} + \mu_k, \qquad \text{if and only if} \qquad y_i = k. \tag{1.3}$$

For $1 \le k \le K$, let δ_k be the fraction of samples in Class k. Note that

$$\delta_1\mu_1 + \delta_2\mu_2 + \ldots + \delta_K\mu_K = 0, \tag{1.4}$$

so $\mu_1, \mu_2, \ldots, \mu_K$ are linearly dependent. However, it is natural to assume

$$\mu_1, \mu_2, \ldots, \mu_{K-1} \text{ are linearly independent.} \tag{1.5}$$

DEFINITION 1.1. *We call feature j a useless feature (for clustering) if $\mu_1(j) = \mu_2(j) = \ldots = \mu_K(j) = 0$, and a useful feature otherwise.*

We call μ_k the *contrast mean vector* of Class k, $1 \le k \le K$. In many applications, the contrast mean vectors are sparse in the sense that only a small fraction of the features are useful. Examples include but are not limited to gene microarray data: it is widely believed that only a small fraction of genes are differentially expressed, so the contrast mean vectors are sparse.

We are primarily interested in the modern regime of $p \gg n$. In such a regime, classical methods (e.g., k-means, hierarchical clustering, Principal Component Analysis (PCA) (Hastie, Tibshirani and Friedman, 2009)) are either computationally challenging or ineffective. Our primary interest is to develop new methods that are appropriate for this regime.

1.1. *Influential Features PCA (IF-PCA).*

Denote the data matrix by:

$$X = [X_1, X_2, \ldots, X_n]'.$$

We propose IF-PCA as a new spectral clustering method. Conceptually, IF-PCA contains an IF part and a PCA part. In the IF part, we select features by exploiting the sparsity of the contrast mean vectors, where we remove many columns of X leaving only those we think are influential for clustering (and so the name of Influential Features). In the PCA part, we apply the classical PCA to the post-selection data matrix.[1]

We normalize each column of X and denote the resultant matrix by W:

$$W(i,j) = [X_i(j) - \bar{X}(j)]/\hat{\sigma}(j), \qquad 1 \le i \le n, \; 1 \le j \le p,$$

where $\bar{X}(j) = \frac{1}{n}\sum_{i=1}^n X_i(j)$ and $\hat{\sigma}(j) = [\frac{1}{n-1}\sum_{i=1}^n (X_i(j) - \bar{X}(j))^2]^{1/2}$ are the empirical mean and standard deviation associated with feature j, respectively. Write

$$W = [W_1, W_2, \ldots, W_n]'.$$

[1] Such a two-stage clustering idea (i.e., feature selection followed by post-selection clustering) is not completely new and can be found in Chan and Hall (2010) for example. Of course, their procedure is very different from ours.

For any $1 \leq j \leq p$, denote the empirical CDF associated with feature j by

$$F_{n,j}(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{W_i(j) \leq t\}.$$

IF-PCA contains two 'IF' steps and two 'PCA' steps as follows.

Input: data matrix X, number of classes K, and parameter t.
Output: predicted $n \times 1$ label vector $\hat{y}_t^{IF} = (\hat{y}_{t,1}^{IF}, \hat{y}_{t,2}^{IF}, \ldots, \hat{y}_{t,n}^{IF})$.

- IF-1. For each $1 \leq j \leq p$, compute a Kolmogorov-Smirnov (KS) statistic by

  $$\psi_{n,j} = \sqrt{n} \cdot \sup_{-\infty < t < \infty} |F_{n,j}(t) - \Phi(t)|, \qquad (\Phi\text{: CDF of } N(0,1)). \tag{1.6}$$

- IF-2. Following the suggestions by Efron (2004), we renormalize by

  $$\psi_{n,j}^{*} = [\psi_{n,j} - \text{mean of all } p \text{ KS-scores}]/\text{SD of all } p \text{ KS-scores}.^{2} \tag{1.7}$$

- PCA-1. Fix a threshold $t > 0$. For short, let $W^{(t)}$ be the matrix formed by restricting the columns of W to the set of retained indices $\hat{S}_p(t)$, where

  $$\hat{S}_p(t) = \{1 \leq j \leq p : \psi_{n,j}^{*} \geq t\}. \tag{1.8}$$

  Let $\hat{U}^{(t)} \in R^{n,K-1}$ be the matrix consisting the first $K-1$ (unit-norm) left singular vectors of $W^{(t)}$.[3] Define a matrix $\hat{U}_{*}^{(t)} \in R^{n,K-1}$ by truncating $\hat{U}^{(t)}$ entry-wise with threshold $T_p = \log(p)/\sqrt{n}$.[4]
- PCA-2. Cluster by applying the classical k-means to $\hat{U}_{*}^{(t)}$ assuming there are $\leq K$ classes. Let $\hat{y}_t^{IF}$ be the predicted label vector.

In the procedure, t is the only tuning parameter. In Section 1.3, we propose a data-driven approach to choosing t, so the method becomes tuning-free. Step 2 is largely for gene microarray data, and is not necessary if Models (1.1)-(1.2) hold.

[2] Alternatively, we can normalize the KS-scores with sample median and Median Absolute Deviation (MAD); see Section 1.5 for more discussion.

[3] For a matrix $M \in R^{n,m}$, the k-th left (right) singular vector is the eigenvector associated with the k-th largest eigenvalue of the matrix MM' (of the matrix $M'M$).

[4] That is, $\hat{U}_{*}^{(t)}(i,k) = \hat{U}(i,k)1\{|\hat{U}(i,k)| \leq T_p\} + T_p \mathrm{sgn}(\hat{U}(i,k))1\{|\hat{U}(i,k)| > T_p\}$, $1 \leq i \leq n$, $1 \leq k \leq K-1$. We usually take $T_p = \log(p)/\sqrt{n}$ as above, but $\log(p)$ can be replaced by any sequence that tends to ∞ as $p \to \infty$. The truncation is mostly for theoretical analysis in Section 2 and is not used in numerical study (real or simulated data).

In Table 2, we use the Lung Cancer(1) data to illustrate how IF-PCA performs with different choices of t. The results show that with t properly set, the number of clustering errors of IF-PCA can be as low as 4. In comparison, classical PCA (column 2 of Table 2; where $t = .000$ so we do not perform feature selection) has 22 clustering errors.

TABLE 2
Clustering errors and # of selected features for different choices of t (Lung Cancer(1) data). Columns highlighted correspond to the sweet spot of the threshold choice.

Threshold t	.000	.608	.828	**.938**	**1.048**	**1.158**	1.268	1.378	1.488
# of selected features	12533	5758	1057	**484**	**261**	**129**	63	21	2
Clustering errors	22	22	24	**4**	**5**	**7**	38	39	33

In Figure 1, we compare IF-PCA with classical PCA by investigating $\hat{U}^{(t)}$ defined in Step 3 for two choices of t: (a) $t = .000$ so $\hat{U}^{(t)}$ is the first singular vector of pre-selection data matrix W, and (b) a data-driven threshold choice by Higher Criticism to be introduced in Section 1.3. For (b), the entries of $\hat{U}^{(t)}$ can be clearly divided into two groups, yielding almost error-free clustering results. Such a clear separation does not exist for (a). These results suggest that IF-PCA may significantly improve classical PCA.

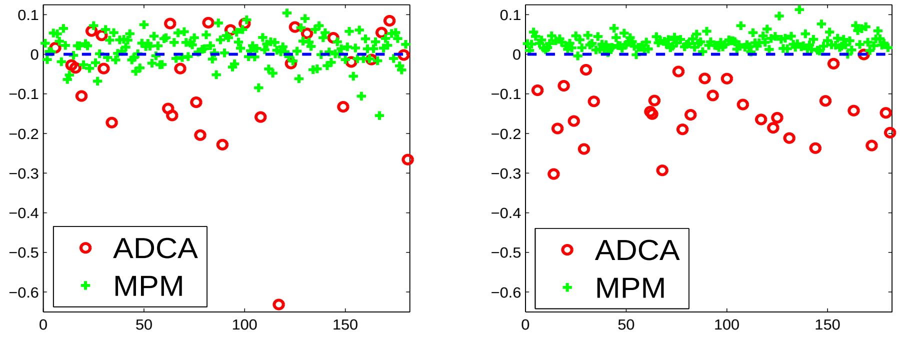

FIG 1. *Comparison of $\hat{U}^{(t)}$ for $t = .000$ (left; no feature selection) and $t = 1.057$ (right; t is set by Higher Criticism in a data-driven fashion); note $\hat{U}^{(t)}$ is an $n \times 1$ vector since $K = 2$. y-axis: entries of $\hat{U}^{(t)}$, x-axis: sample indices. Plots are based on Lung Cancer(1) data, where ADCA and MPM represent two different classes.*

Two important questions arise:

- In (1.7), we use a modified KS statistic for feature selection. What is the rationale behind the use of KS statistics and the modification?
- The clustering errors critically depend on the threshold t. How to set t in a data-driven fashion?

In Section 1.2, we address the first question. In Section 1.3, we propose a data-driven threshold choice by the recent notion of Higher Criticism.

1.2. *KS statistic, normality assumption, and Efron's empirical null.* The goal in Steps 1-2 is to find an easy-to-implement method to rank the features. The focus of Step 1 is on a data matrix satisfying Models (1.1)-(1.5), and the focus of Step 2 is to adjust Step 1 in a way so to work well with microarray data. We consider two steps separately.

Consider the first step. The interest is to test for each fixed j, $1 \leq j \leq p$, whether feature j is useless or useful. Since we have no prior information about the class labels, the problem can be reformulated as that of testing whether all n samples associated with the j-th feature are iid Gaussian

$$H_{0,j}: \qquad X_i(j) \overset{iid}{\sim} N(\bar{\mu}(j), \sigma^2(j)), \qquad i = 1, 2, \ldots, n, \tag{1.9}$$

or they are iid from a K-component heterogenous Gaussian mixture:

$$H_{1,j}: \qquad X_i(j) \overset{iid}{\sim} \sum_{k=1}^{K} \delta_k N(\bar{\mu}(j) + \mu_k(j), \sigma^2(j)), \qquad i = 1, 2, \ldots, n, \tag{1.10}$$

where $\delta_k > 0$ is the prior probability that $X_i(j)$ comes from Class k, $1 \leq k \leq K$. Note that $\bar{\mu}(j)$, $\sigma(j)$, and $\big((\delta_1, \mu_1(j)), \ldots, (\delta_K, \mu_K(j))\big)$ are unknown. The above is a well-known difficult testing problem. For example, in such a setting, the classical Likelihood Ratio Test (LRT) is known to be not well-behaved (e.g., Chen and Li (2009)).

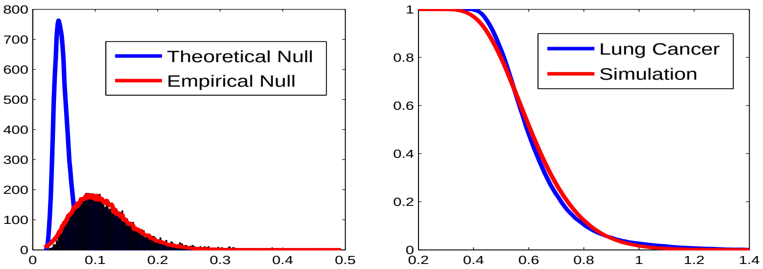

FIG 2. *Left: The histogram of KS-scores of the Lung Cancer(1) data. The two lines in blue and red denote the theoretical null and empirical null densities, respectively. Right: empirical survival function of the adjusted KS-scores based on Lung Cancer(1) data (red) and the survival function of theoretical null (blue).*

Our proposal is to use the Kolmogorov-Smirnov (KS) test, which measures the maximum difference between the empirical CDF for the normalized data and the CDF of $N(0,1)$. The KS test is a well-known goodness-of-fit test (e.g., Shorack and Wellner (1986)). In the idealized Gaussian Model (1.9)-(1.10), the KS test is asymptotically equivalent to the optimal moment-based tests (e.g., see Section 2), but its success is not tied to a specific model for the alternative hypothesis, and is more robust against occasional outliers. Also, Efron's null correction (below) is more successful if we use KS instead of moment-based tests for feature ranking. This is our rationale for Step 1.

We now discuss our rationale for Step 2. We discover an interesting phenomenon which we illustrate with Figure 2 (Lung Cancer(1) data). Ideally, if the normality assumption (1.2) is valid for this data set, then the density function of the KS statistic for Model (1.9) (the blue curve in left panel; obtained by simulations) should fit well with the histogram of the KS-scores based on the Lung Cancer(1) data. Unfortunately, this is not the case, and there is a substantial discrepancy in fitting. On the other hand, if we translate and rescale the blue curve so that it has the same mean and standard deviation as the KS-scores associated with Lung Cancer(1) data, then the new curve (red curve; left panel of Figure 2) fits well with the histogram.[5]

A related phenomenon was discussed in Efron (2004), only considering Studentized t-statistics in a different setting. As in Efron (2004), we call the density functions associated with two curves (blue and red) the *theoretical null* and the *empirical null*, respectively. The phenomenon is then: the theoretical null has a poor fit with the histogram of the KS-scores of the real data, but the empirical null may have a good fit.

In the right panel of Figure 2, we view this from a slightly different perspective, and show that the survival function associated with the adjusted KS-scores (i.e., $\psi^*_{n,j}$) of the real data fits well with the theoretical null.

The above observations explain the rationale for Step 2. Also, they suggest that IF-PCA does not critically depend on the normality assumption and works well for microarray data. This is further validated in Section 1.4.

Remark. Efron (2004) suggests several possible reasons (e.g., dependence between different samples, dependence between the genes) for the discrepancy between the theoretical null and empirical null, but what has really caused such a discrepancy is not fully understood. Whether Efron's empirical null is useful in other application areas or other data types (and if so, to what extent) is also an open problem, and to understand it we need a good grasp on the mechanism by which the data sets of interest are generated.

[5] If we replace sample mean and standard deviation by sample median and MAD, respectively, then it gives rises to the normalization in the second footnote of Section 1.1.

1.3. *Threshold choice by Higher Criticism.*

The performance of IF-PCA critically depends on the threshold t, and it is of interest to set t in a data-driven fashion. We approach this by the recent notion of Higher Criticism.

Higher Criticism (HC) was first introduced in Donoho and Jin (2004) as a method for large-scale multiple testing. In Donoho and Jin (2008), HC was also found to be useful to set a threshold for feature selection in the context of classification. HC is also useful in many other settings. See Donoho and Jin (2015); Jin and Ke (2016) for reviews on HC.

To adapt HC for threshold choice in IF-PCA, we must modify the procedure carefully, since the purpose is very different from those in previous literature. The approach contains three simple steps as follows.

- For $1 \leq j \leq p$, calculate a P-value $\pi_j = 1 - F_0(\psi_{n,j})$, where F_0 is the distribution of $\psi_{n,j}$ under the null (i.e., feature j is useless).
- Sort all P-values in the ascending order $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.
- Define the Higher Criticism score by

$$HC_{p,j} = \sqrt{p}(j/p - \pi_{(j)})/\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p}. \tag{1.11}$$

  Let $\hat{j}$ be the index such that $\hat{j} = \text{argmax}_{\{1 \leq j \leq p/2, \pi_{(j)} > \log(p)/p\}}\{HC_{p,j}\}$. The HC threshold t_p^{HC} for IF-PCA is then the $\hat{j}$-th largest KS-scores.

Combining HCT with IF-PCA gives a tuning-free clustering procedure IF-HCT-PCA, or IF-PCA for short if there is no confusion. See Table 3.

TABLE 3
Pseudocode for IF-HCT-PCA (for microarray data; threshold set by Higher Criticism)

```
    Input: data matrix $X$, number of classes $K$. Output: class label vector $\hat{y}_{HC}^{IF}$.
1.  Rank features: Let $\psi_{n,j}$ be the KS-scores as in (1.6) and $F_0$ be the CDF of $\psi_{n,j}$ under null, $1 \leq j \leq p$.
2.  Normalize KS-scores: $\psi_n^* = (\psi_n - mean(\psi_n))/SD(\psi_n)$.
3.  Threshold choice by HCT: Calculate $P$-values by $\pi_j = 1 - F_0(\psi_{n,j}^*)$, $1 \leq j \leq p$ and sort them by
    $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$. Define $HC_{p,j} = \sqrt{p}(j/p - \pi_{(j)})/\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p}$, and let
    $\hat{j} = \text{argmax}_{\{j: \pi_{(j)} > \log(p)/p, j < p/2\}}\{HC_{p,j}\}$. HC threshold $t_p^{HC}$ is the $\hat{j}$-largest KS-score.
4.  Post-selection PCA: Define post-selection data matrix $W^{(HC)}$ (i.e., sub-matrix of $W$ consists of all
    column $j$ of $W$ with $\psi_{n,j}^* > t_p^{HC}$). Let $U \in R^{n,K-1}$ be the matrix of the first $(K-1)$ left singular
    vectors of $W^{(HC)}$. Cluster by $\hat{y}_{HC}^{IF} = kmeans(U, K)$.
```

For illustration, we again employ the Lung Cancer(1) data. In this data set, $\hat{j} = 251$, $t_p^{HC} = 1.0573$, and HC selects 251 genes with the largest KS-scores. In Figure 3, we plot the error rates of IF-PCA applied to the k features of W with the largest KS-scores, where k ranges from 1 to $p/2$ (for different k, we are using the same ranking for all p genes). The figure shows that there is a 'sweet spot' for k where the error rates are the lowest. HCT corresponds to $\hat{j} = 251$ and 251 is in this sweet spot. This suggests that HCT gives a reasonable threshold choice, at least for some real data sets.

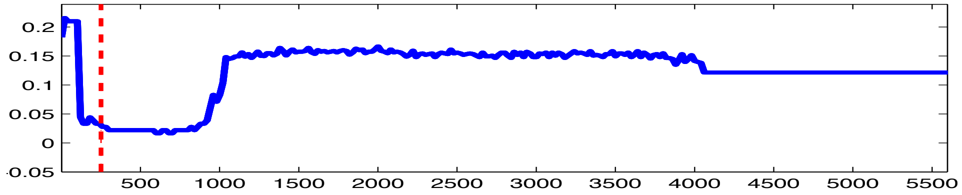

FIG 3. *Error rates by IF-PCA (y-axis) with different number of selected features k (x-axis) (Lung Cancer(1) data). HCT corresponds to* 251 *selected features (dashed vertical line).*

Remark. When we apply HC to microarray data, we follow the discussions in Section 1.2 and take F_0 to be the distribution of $\psi_{n,j}$ under the null but with the mean and variance adjusted to match those of the KS-scores. In the definition, we require $\pi_{(\hat{j})} > \log(p)/p$, as $HC_{p,j}$ may be ill-behaved for very small j (e.g., Donoho and Jin (2004)).

The rationale for HCT can also be explained theoretically. For illustration, consider the case where $K = 2$ so we only have two classes. Fixing a threshold $t > 0$, let $\hat{U}^{(t)}$ be the first left singular vector of $W^{(t)}$ as in Section 1.1. In a companion paper (Jin, Ke and Wang, 2015a), we show that when the signals are rare and weak, then for t in the range of interest,

$$\hat{U}^{(t)} \propto \widetilde{snr}(t) \cdot U + z + rem, \tag{1.12}$$

where U is an $n \times 1$ non-stochastic vector with only two distinct entries (each determines one of two classes), $\widetilde{snr}(t)$ is a non-stochastic function of t, $z \sim N(0, I_n)$, and rem is the remainder term (the entries of which are asymptotically of much smaller magnitude than that of z or $\widetilde{snr}(t) \cdot U$). Therefore, performance of IF-PCA is best when we maximize $\widetilde{snr}(t)$ (though this is unobservable). We call such a threshold the *Ideal Threshold*: $t_p^{ideal} = \text{argmin}_{t>0}\{\widetilde{snr}(t)\}$.

Let $\bar{F}_p(t)$ be the survival function of $\psi_{n,j}$ under the null (not dependent on j), and let $\hat{G}_p(t) = \frac{1}{p}\sum_{j=1}^{p} 1\{\psi_{n,j} \geq t\}$ be the empirical survival function. Introduce $HC_p(t) = \sqrt{p}[\hat{G}_p(t) - \bar{F}_p(t)]/\sqrt{\hat{G}_p(t) + \sqrt{n}[\max\{\hat{G}_p(t) - \bar{F}_p(t), 0\}]}$, and let $\psi_{(1)} > \psi_{(2)} > \ldots > \psi_{(p)}$ be the sorted values of $\psi_{n,j}$. Recall that $\pi_{(k)}$ is the k-th smallest P-value. By definitions, we have $\hat{G}_p(t)|_{t=\psi_{(k)}} = k/p$ and $\bar{F}_p(t)|_{t=\psi_{(k)}} = \pi_{(k)}$. As a result, we have $HC_p(t)\big|_{t=\psi_{(k)}} = [k/p - \pi_{(k)}]/\sqrt{k/p + \sqrt{n}\max\{k/p - \pi_{(k)}, 0\}}$, where the right hand side is the form of HC introduced in (1.11). Note that $HC_p(t)$ is a function which is only discontinuous at $t = \psi_{(k)}$, $1 \leq k \leq p$, and between two adjacent discontinuous points, the function is monotone. Combining this with the definition of t_p^{HC}, $t_p^{HC} = \text{argmax}_t\{HC_p(t)\}$.

Now, as $p \to \infty$, some regularity appears, and $\hat{G}_p(t)$ converges to a non-stochastic counterpart, denoted by $\bar{G}_p(t)$, which can be viewed as the survival function associated with the marginal density of $\psi_{n,j}$. Introduce $IdealHC(t) = \sqrt{p}[\bar{G}_p(t) - \bar{F}_p(t)]/\sqrt{\bar{G}_p(t) + \sqrt{n}[\max\{\bar{G}_p(t) - \bar{F}_p(t), 0\}]}$ as the ideal counterpart of $HC_p(t)$. It is seen that $HC_p(t) \approx IdealHC(t)$ for t in the range of interest, and so $t_p^{HC} \approx t_p^{idealHC}$, where the latter is defined as the non-stochastic threshold t that maximizes $IdealHC(t)$.

In Jin, Ke and Wang (2015a), we show that under a broad class of rare and weak signal models, the leading term of the Taylor expansion of $\widetilde{snr}(t)$ is proportional to that of $IdealHC(t)$ for t in the range of interest, and so $t_p^{idealHC} \approx t_p^{ideal}$. Combining this with the discussions above, we have $t_p^{HC} \approx t_p^{idealHC} \approx t_p^{ideal}$, which explains the rationale for HCT.

The above relationships are justified in Jin, Ke and Wang (2015a). The proofs are rather long (70 manuscript pages in Annals of Statistics format), so we will report them in a separate paper. The ideas above are similar to that in Donoho and Jin (2008) but the focus there is on classification and our focus is on clustering; our version of HC is also very different from theirs.

1.4. *Applications to gene microarray data.* We compare IF-HCT-PCA with four other clustering methods (applied to the normalized data matrix W directly, without feature selection): (1) SpectralGem (Lee, Luca and Roeder, 2010) which is the same as classical PCA introduced earlier, (2) classical k-means, (3) hierarchical clustering (Hastie, Tibshirani and Friedman, 2009), and (4) k-means++ (Arthur and Vassilvitskii, 2007). In theory, k-means is NP hard, but heuristic algorithms are available; we use the built-in k-means package in Matlab with the parameter 'replicates' equal to 30, so that the algorithm randomly samples initial cluster centroid positions 30 times (in the last step of either classical PCA or IF-HCT-PCA, k-means is also used, where the number of 'replicates' is also 30). The k-means++ (Arthur and Vassilvitskii, 2007) is a recent modification of k-means. It improves the performance of k-means in some numerical studies, though the problem remains NP hard in theory. For hierarchical clustering, we use 'complete' as the linkage function; other choices give more or less the same results. In IF-HCT-PCA, the P-values associated with the KS-scores are computed using simulated KS-scores under the null with $2 \times 10^3 \times p$ independent replications; see Section 1.3 for remarks on F_0. In Table 3, we repeat the main steps of IF-HCT-PCA for clarification, by presenting the pseudocode.

Table 4

Comparison of clustering error rates by different methods for the 10 *gene microarray data sets introduced in Table 1. Column* 5*: numbers in the brackets are the standard deviations (SD); SD for all other methods are negligible so are not reported. Last column: see* (1.13).

#	Data set	K	kmeans	kmeans++	Hier	SpecGem	IF-HCT-PCA	r
1	Brain	5	.286	.427(.09)	.524	.143	.262	1.83
2	Breast Cancer	2	.442	.430(.05)	.500	.438	.406	.94
3	Colon Cancer	2	.443	.460(.07)	.387	.484	.403	1.04
4	**Leukemia**	**2**	**.278**	**.257(.09)**	**.278**	**.292**	**.069**	**.27**
5	**Lung Cancer(1)**	**2**	**.116**	**.196(.09)**	**.177**	**.122**	**.033**	**.29**
6	Lung Cancer(2)	2	.436	.439(.00)	.301	.434	.217	.72
7	**Lymphoma**	**3**	**.387**	**.317(.13)**	**.468**	**.226**	**.065**	**.29**
8	Prostate Cancer	2	.422	.432(.01)	.480	.422	.382	.91
9	SRBCT	4	.556	.524(.06)	.540	.508	.444	.87
10	SuCancer	2	.477	.459(.05)	.448	.489	.333	.74

We applied all 5 methods to each of the 10 gene microarray data sets in Table 1. The results are reported in Table 4. Since all methods except hierarchical clustering have algorithmic randomness (they depend on built-in k-means package in Matlab which uses a random start), we report the mean error rate based on 30 independent replications. The standard deviation of all methods is very small ($< .0001$) except for k-means++, so we only report the standard deviation of k-means++. In the last column of Table 4,

$$r = \frac{\text{error rate of IF-HCT-PCA}}{\text{minimum of the error rates of the other 4 methods}}. \tag{1.13}$$

We find that $r < 1$ for all data sets except for two. In particular, $r \leq .29$ for three of the data sets, marking a substantial improvement, and $r \leq .87$ for three other data sets, marking a moderate improvement. The r-values also suggest an interesting point: for 'easier' data sets, IF-PCA tends to have more improvements over the other 4 methods.

We make several remarks. First, for the Brain data set, unexpectedly, IF-PCA underperforms classical PCA, but still outperforms other methods. Among our data sets, the Brain data seem to be an 'outlier'. Possible reasons include (a) useful features are not sparse, and (b) the sample size is very small ($n = 42$) so the useful features are individually very weak. When (a)-(b) happen, it is almost impossible to successfully separate the useful features from useless ones, and it is preferable to use classical PCA. Such a scenario may be found in Jin, Ke and Wang (2015b); see for example Figure 1 (left) and related context therein.

Second, for Colon Cancer, all methods behave unsatisfactorily, and IF-PCA slightly underperforms hierarchical clustering ($r = 1.04$). The data set is known to be a difficult one even for classification (where class labels of training samples are known (Donoho and Jin, 2008)). For such a difficult data set, it is hard for IF-PCA to significantly outperform other methods.

Last, for the SuCancer data set, the KS-scores are significantly skewed to the right. Therefore, instead of using the normalization (1.7), we normalize $\psi_{n,j}$ such that the mean and standard deviation for the lower 50% of KS-scores match those for the lower 50% of the simulated KS-scores under the null; compare this with Section 1.3 for remarks on P-value calculations.

1.5. *Three variants of IF-HCT-PCA.* First, in IF-HCT-PCA, we normalize the KS-scores with the sample mean and sample standard deviation as in (1.7). Alternatively, we may normalize the KS-scores by $\psi^*_{n,j} = [\psi_{n,j} -$ median of all KS-scores]/[MAD of all KS-scores] (MAD: Median Absolute Deviation), while other steps of IF-HCT-PCA are kept intact. Denote the resultant variant by IF-HCT-PCA-med (med: median). Second, recall that IF-HCT-PCA has two stages: in the first one, we select features with a threshold determined by HC; in the second one, we apply PCA to the post-selection data matrix. Alternatively, in the second stage, we may apply classical k-means or hierarchical clustering to the post-selection data instead (the first stage is intact). Denote these two alternatives by IF-HCT-kmeans and IF-HCT-hier, respectively.

Table 5
Clustering error rates of IF-HCT-PCA, IF-HCT-PCA-med, IF-HCT-kmeans, and IF-HCT-hier.

	Brn	Brst	Cln	Leuk	Lung1	Lung2	Lymp	Prst	SRB	Su
IF-HCT-PCA	.262	.406	.403	.069	.033	.217	.065	.382	.444	.333
IF-HCT-PCA-med	.333	.424	.436	.014	.017	.217	.097	.382	.206	.333
IF-HCT-kmeans	.191	.380	.403	.028	.033	.217	.032	.382	.401	.328
IF-HCT-hier	.476	.351	.371	.250	.177	.227	.355	.412	.603	.500

Table 5 compares IF-HCT-PCA with the three variants (in IF-HCT-kmeans, the ‘replicate’ parameter in k-means is taken to be 30 as before), where the first three methods have similar performances, while the last one performs comparably less satisfactorily. Not surprisingly, these methods generally outperform their classical counterparts (i.e., classical PCA, classical k-means, and hierarchical clustering; see Table 4).

We remark that, for post-selection clustering, it is frequently preferable to use PCA than k-means. First, k-means could be much slower than PCA, especially when the number of selected features in the IF step is large. Second, the k-means algorithm we use in Matlab is only a heuristic approximation of the theoretical k-means (which is NP-hard), so it is not always easy to justify the performance of k-means algorithm theoretically.

1.6. *Connection to sparse PCA.* The study is closely related to the recent interest on sparse PCA (Arias-Castro, Lerman and Zhang (2013); Amini and Wainwright (2008); Johnstone (2001); Jung and Marron (2009); Lei and Vu (2015); Ma (2013); Zou, Hastie and Tibshirani (2006)), but is different in important ways. Consider the normalized data matrix $W = [W_1, W_2, \ldots, W_n]'$ for example. In our model, recall that $\mu_1, \mu_2, \ldots, \mu_K$ are the K sparse contrast mean vectors and the noise covariance matrix Σ is diagonal, we have

$$W \approx M\Sigma^{-1/2} + Z, \qquad \text{where } Z \in R^{n,p} \text{ has } iid\ N(0,1) \text{ entries},$$

and $M \in R^{n,p}$ is the matrix where the i-th row is μ_k' if and only if $i \in$ Class k. This is a setting that is frequently considered in the sparse PCA literature.

However, we must note that the main focus of sparse PCA is to recover the supports of $\mu_1, \mu_2, \ldots, \mu_K$, while the main focus here is subject clustering. We recognize that, the two problems—support recovery and subject clustering—are essentially two different problems, and addressing one successfully does not necessarily address the other successfully. For illustration, consider two scenarios.

- If useful features are very sparse but each is sufficiently strong, it is easy to identify the support of the useful features, but due to the extreme sparsity, it may be still impossible to have consistent clustering.
- If most of the useful features are very weak with only a few of them very strong, the latter will be easy to identify and may yield consistent clustering, still, it may be impossible to satisfactorily recover the supports of $\mu_1, \mu_2, \ldots, \mu_K$, as most of the useful features are very weak.

In a forthcoming manuscript Jin, Ke and Wang (2015b), we investigate the connections and differences between two problems more closely, and elaborate the above points with details.

With that being said, from a practical viewpoint, one may still wonder how sparse PCA may help in subject clustering. A straight-forward clustering approach that exploits the sparse PCA ideas is the following:

- Estimate the first $(K-1)$ right singular vectors of the matrix $M\Sigma^{-1/2}$ using the sparse PCA algorithm as in (Zou, Hastie and Tibshirani, 2006, Equation (3.7)) (say). Denote the estimates by $\hat{\nu}_1^{sp}, \hat{\nu}_2^{sp}, \ldots, \hat{\nu}_{K-1}^{sp}$.
- Cluster by applying classical k-means to the $n \times K - 1$ matrix $[W\hat{\nu}_1^{sp}, W\hat{\nu}_2^{sp}, \ldots, W\hat{\nu}_{K-1}^{sp}]$, assuming there are $\leq K$ classes.

For short, we call this approach Clu-sPCA. One problem here is that, Clu-sPCA is *not* tuning-free, as most existing sparse PCA algorithms have one or more tuning parameters. How to set the tuning parameters in subject clustering is a challenging problem: for example, since the class labels are unknown, using conventional cross validations (as we may use in classification where class labels of the training set are known) might not help.

TABLE 6

Clustering error rates for IF-HCT-PCA and Clu-sPCA. The tuning parameter of Clu-sPCA is chosen ideally to minimize the errors (IF-HCT-PCA is tuning-free). Only SDs that are larger than 0.01 are reported (in brackets).

	Brn	Brst	Cln	Leuk	Lung1	Lung2	Lymp	Prst	SRB	Su
IF-HCT-PCA	.262	.406	.403	**.069**	**.033**	**.217**	**.065**	.382	.444	**.333**
Clu-sPCA	.263	.438	.435	**.292**	**.110**	**.433**	**.190(.01)**	.422	.428	**.437**

In Table 6, we compare IF-HCT-PCA and Clu-sPCA using the 10 data sets in Table 1. Note that in Clu-sPCA, the tuning parameter in the sparse PCA step (Zou, Hastie and Tibshirani, 2006, Equation (3.7)) is *ideally chosen* to minimize the clustering errors, using the true class labels. The results are based on 30 independent repetitions. Compared to Clu-sPCA, IF-HCT-PCA outperforms for half of the data sets (bold face), and has similar performances for the remaining half.

The above results support our philosophy: the problem of subject clustering and the problem of support recovery are related but different, and success in one does not automatically lead to the success in the other.

1.7. *Summary and contributions.* Our contribution is three-fold: feature selection by the KS statistic, post-selection PCA for high dimensional clustering, and threshold choice by the recent idea of Higher Criticism.

In the first fold, we rediscover a phenomenon found earlier by Efron (2004) for microarray study, but the focus there is on t-statistic or F-statistic, and the focus here is on the KS statistic. We establish tight probability bounds on the KS statistic when the data is Gaussian or Gaussian mixtures where the means and variances are unknown; see Section 2.5. While tight tail probability bounds have been available for decades in the case where the data are *iid* from $N(0,1)$, the current case is much more challenging. Our results follow the work by Siegmund (1982) and Loader et al. (1992) on the local Poisson approximation of boundary crossing probability, and are useful for pinning down the thresholds in KS screening.

In the second fold, we propose to use IF-PCA for clustering and have successfully applied it to gene microarray data. The method compares favorably with other methods, which suggests that both the IF step and the post-selection PCA step are effective. We also establish a theoretical framework where we investigate the clustering consistency carefully; see Section 2. The analysis it entails is sophisticated and involves delicate post-selection eigen-analysis (i.e., eigen-analysis on the post-selection data matrix). We also gain useful insight that the success of feature selection depends on the feature-wise weighted third moment of the samples, while the success of PCA depends more on the feature-wise weighted second moment. Our study is closely related to the SpectralGem approach by Lee, Luca and Roeder (2010), but our focus is on KS screening, post-selection PCA, and clustering with microarray data is different.

In the third fold, we propose to set the threshold by Higher Criticism. We find an intimate relationship between the HC functional and the signal-to-noise ratio associated with post-selection eigen-analysis. As mentioned in Section 1.3, the full analysis on the HC threshold choice is difficult and long, so for reasons of space, we do not include it in this paper.

Our findings support the philosophy by Donoho (2015), that for real data analysis, we prefer to use simple models and methods that allow sophisticated theoretical analysis than complicated and computationally intensive methods (as an increasing trend in some other scientific communities).

1.8. *Content and notations.* Section 2 contains the main theoretical results, where we show IF-PCA is consistent in clustering under some regularity conditions. Section 3 contains the numerical studies, and Section 4 discusses connection to other work and addresses some future research. Secondary theorems and lemmas are proved in the supplementary material of the paper. In this paper, L_p denotes a generic multi-log(p) term (see Section 2.3). For a vector ξ, $\|\xi\|$ denotes the ℓ^2-norm. For a real matrix A, $\|A\|$ denotes the matrix spectral norm, $\|A\|_F$ denotes the matrix Frobenius norm, and $s_{\min}(A)$ denotes the smallest nonzero singular value.

2. Main results.

Section 2.1 introduces our asymptotic model, Section 2.2 discusses the main regularity conditions and related notations. Section 2.3 presents the main theorem, and Section 2.4 presents two corollaries, together with a phase transition phenomenon. Section 2.5 discusses the tail probability of the KS statistic, which is the key for the IF step. Section 2.6 studies post-selection eigen-analysis which is the key for the PCA step. The main theorems and corollaries are proved in Section 2.7.

To be utterly clear, the IF-PCA procedure we study in this section is the one presented in Table 7, where the threshold $t > 0$ is given.

2.1. *The Asymptotic Clustering Model.* The model we consider is (1.1), (1.2), (1.3) and (1.5), where the data matrix is $X = [X_1, X_2, \ldots, X_n]'$, with $X_i \sim N(\bar{\mu} + \mu_k, \Sigma)$ if and only if $i \in$ Class k, $1 \leq k \leq K$, and $\Sigma =$

TABLE 7

Pseudocode for IF-PCA (for a given threshold $t > 0$)

```
Input: data matrix $X$, number of classes $K$, threshold $t > 0$. Output: class label vector $\hat{y}_t^{IF}$.
1. Rank features: Let $\psi_{n,j}$, $1 \leq j \leq p$, be the KS-scores as in (1.6).
2. Post-selection PCA: Define post-selection data matrix $W^{(t)}$ (i.e, sub-matrix of $W$ consists of all
   column $j$ with $\psi_{n,j} > t$). Let $U \in R^{n,K-1}$ be the matrix of the first $(K-1)$ left singular vectors
   of $W^{(t)}$. Cluster by $\hat{y}_t^{IF} = kmeans(U, K)$.
```

$\text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$; K is the number of classes, $\bar{\mu}$ is the overall mean vector, $\mu_1, \mu_2, \ldots, \mu_K$ are contrast mean vectors which satisfy (1.5).

We use p as the driving asymptotic parameter, and let other parameters be tied to p through fixed parameters. Fixing $\theta \in (0, 1)$, we let

$$n = n_p = p^{\theta}, \tag{2.1}$$

so that as $p \to \infty$, $p \gg n \gg 1$.[6] Let $M \in R^{K,p}$ be the matrix

$$M = [m_1, m_2, \ldots, m_K]', \qquad \text{where} \quad m_k = \Sigma^{-1/2}\mu_k. \tag{2.2}$$

Denote the set of useful features by

$$S_p = S_p(M) = \{1 \leq j \leq p : m_k(j) \neq 0 \text{ for some } 1 \leq k \leq K\}, \tag{2.3}$$

and let $s_p = s_p(M) = |S_p(M)|$ be the number of useful features. Fixing $\vartheta \in (0, 1)$, we let

$$s_p = p^{1-\vartheta}. \tag{2.4}$$

Throughout this paper, the number of classes K is fixed, as p changes.

DEFINITION 2.1. *We call model (1.1), (1.2), (1.3), and (1.5) the Asymptotic Clustering Model if (2.1) and (2.4) hold, and denote it by $ACM(\vartheta, \theta)$.*

It is more convenient to work with the normalized data matrix $W = [W_1, W_2, \ldots, W_n]'$, where, as before, $W_i(j) = [X_i(j) - \bar{X}(j)]/\hat{\sigma}(j)$, and $\bar{X}(j)$ and $\hat{\sigma}(j)$ are the empirical mean and standard deviation associated with the feature j, $1 \leq j \leq p$, $1 \leq i \leq n$. Introduce $\hat{\Sigma} = \text{diag}(\hat{\sigma}^2(1), \hat{\sigma}^2(2), \ldots, \hat{\sigma}^2(p))$ and $\widetilde{\Sigma} = E[\hat{\Sigma}]$. Note that $\hat{\sigma}^2(j)$ is an unbiased estimator for $\sigma^2(j)$ when feature j is useless but is not necessarily so when feature j is useful. As a result, $\hat{\Sigma}$ is 'closer' to $\widetilde{\Sigma}$ than to Σ; this causes (unavoidable) complications in notations. Denote for short

$$\Lambda = \Sigma^{1/2}\widetilde{\Sigma}^{-1/2}. \tag{2.5}$$

[6] For simplicity, we drop the subscript of n_p as long as there is no confusion.

This is a $p \times p$ diagonal matrix where most of the diagonals are 1, and all other diagonals are close to 1 (under mild conditions). Let $\mathbf{1}_n$ be the $n \times 1$ vector of ones and $e_k \in R^K$ be the k-th standard basis vector of R^K, $1 \leq k \leq K$. Let $L \in R^{n,K}$ be the matrix where the i-th row is e_k' if and only if Sample $i \in$ Class k. Recall the definition of M in (2.2). With these notations, we can write

$$W = [LM + Z\Sigma^{-1/2}]\Lambda + R, \qquad Z\Sigma^{-1/2} \text{ has } iid\ N(0,1) \text{ entries}, \tag{2.6}$$

where R stands for the remainder term

$$R = \mathbf{1}_n(\bar{\mu} - \bar{X})'\hat{\Sigma}^{-1/2} + [LM\Sigma^{1/2} + Z](\hat{\Sigma}^{-1/2} - \widetilde{\Sigma}^{-1/2}). \tag{2.7}$$

Recall that $rank(LM) = K - 1$ and Λ is nearly the identity matrix.

2.2. *Regularity conditions and related notations.* We use $C > 0$ as a generic constant, which may change from occurrence to occurrence, but does not depend on p. Recall that δ_k is the fraction of samples in Class k, and $\sigma^2(j)$ is the j-th diagonal of Σ. The following regularity conditions are mild:

$$\min_{1 \leq k \leq K}\{\delta_k\} \geq C, \qquad \text{and} \qquad \max_{1 \leq j \leq p}\{\sigma(j) + \sigma^{-1}(j)\} \leq C. \tag{2.8}$$

Introduce the following two $p \times 1$ vectors $\kappa = (\kappa(1), \kappa(2), \ldots, \kappa(p))'$ and $\tau = (\tau(1), \tau(2), \ldots, \tau(p))'$ by

$$\kappa(j) = \kappa(j; M, p, n) = \big(\sum_{k=1}^{K} \delta_k m_k^2(j)\big)^{1/2}, \tag{2.9}$$

$$\tau(j) = \tau(j; M, p, n) = (6\sqrt{2\pi})^{-1} \cdot \sqrt{n} \cdot \big|\sum_{k=1}^{K} \delta_k m_k^3(j)\big|. \tag{2.10}$$

Note that $\kappa(j)$ and $\tau(j)$ are related to the weighted second and third moments of the j-th column of M, respectively; τ and κ play a key role in the success of feature selection and post-selection PCA, respectively. In the case that $\tau(j)$'s are all small, the success of our method relies on higher moments of the columns of M; see Section 2.5 for more discussions. Introduce

$$\epsilon(M) = \max_{1 \leq k \leq K, j \in S_p(M)}\{|m_k(j)|\}, \qquad \tau_{min} = \min_{j \in S_p(M)}\{\tau(j)\}.$$

We are primarily interested in the range where the feature strengths are rare and weak, so we assume as $p \to \infty$,

$$\epsilon(M) \to 0.^{7} \tag{2.11}$$

In Section 2.5, we shall see that $\tau(j)$ can be viewed as the Signal-to-Noise Ratio (SNR) associated with the j-th feature and τ_{min} is the minimum SNR of all useful features. The most interesting range for $\tau(j)$ is $\tau(j) \geq O(\sqrt{\log(p)})$. In fact, if $\tau(j)$s are of a much smaller order, then the useful features and the useless features are merely inseparable. In light of this, we fix a constant $r > 0$ and assume

$$\tau_{min} \geq a_0 \cdot \sqrt{2r\log(p)}, \qquad \text{where } a_0 = \sqrt{(\pi-2)/(4\pi)}.^{8} \tag{2.12}$$

By the way $\tau(j)$ is defined, the interesting range for non-zero $m_k(j)$ is $|m_k(j)| \geq O\big((\log(p)/n)^{1/6}\big)$. We also need some technical conditions which can be largely relaxed with more complicated analysis:[9]

(2.13)

$$\max_{j \in S_p(M)} \left\{ \frac{\sqrt{n}}{\tau(j)} \sum_{k=1}^{K} \delta_k m_k^4(j) \right\} \leq Cp^{-\delta}, \quad \min_{\{(j,k): m_k(j) \neq 0]\}} \{|m_k(j)|\} \geq C(\frac{\log(p)}{n})^{1/2},$$

for some $\delta > 0$. As the most interesting range of $|m_k(j)|$ is $O((\log(p)/n)^{1/6})$, these conditions are mild.

Similarly, for the threshold t in (1.8) we use for the KS-scores, the interesting range is $t = O(\sqrt{\log(p)})$. In light of this, we are primarily interested in threshold of the form

$$t_p(q) = a_0 \cdot \sqrt{2q\log(p)}, \qquad \text{where } q > 0 \text{ is a constant.} \tag{2.14}$$

We now define a quantity err_p, which is the clustering error rate of IF-PCA in our main results. Define

$$\rho_1(L, M) = \rho_1(L, M; p, n) = \frac{s_p \|\kappa\|_\infty^2}{\|\kappa\|^2}.$$

[7] This condition is used in the post-selection eigen-analysis. Recall that $W^{(t)}$ is the short-hand notation for the post-selection normalized data matrix associated with threshold t. As $W^{(t)}$ is the sum of a low-rank matrix and a noise matrix, $(W^{(t)})'(W^{(t)})$ equals to the sum of four terms, two of them are "cross terms". In eigen-analysis of $(W^{(t)})'(W^{(t)})$, we need condition (2.11) to control the cross terms.

[8] Throughout this paper, a_0 denotes the constant $\sqrt{(\pi-2)/(4\pi)}$. The constant comes from the analysis of the tail behavior of the KS statistic; see Theorems 2.3-2.4.

[9] Condition (2.13) is only needed for Theorem 2.4 on the tail behavior of the KS statistic associated with a useful feature. The conditions ensure singular cases will not happen so the weighted third moment (captured by $\tau(j)$) is the leading term in the Taylor expansion. For more discussions, see the remark in Section 2.5.

Introduce two $K \times K$ matrices A and Ω (where A is diagonal) by

$$A(k,k) = \sqrt{\delta_k}\|m_k\|, \qquad \Omega(k,\ell) = m_k' \Lambda^2 m_\ell / (\|m_k\| \cdot \|m_\ell\|), \qquad 1 \leq k, \ell \leq K;$$

recall that Λ is 'nearly' the identity matrix. Note that $\|A\Omega A\| \leq \|\kappa\|^2$, and that when $\|m_1\|, \cdots, \|m_K\|$ have comparable magnitudes, all the eigenvalues of $A\Omega A$ have the same magnitude. In light of this, let $s_{\min}(A\Omega A)$ be the minimum singular value of A and introduce the ratio

$$\rho_2(L, M) = \rho_2(L, M; p, n) = \|\kappa\|^2 / s_{\min}(A\Omega A).$$

Define

$$err_p = \rho_2(L,M)\Bigg[\frac{1 + \sqrt{\frac{p^{1-\vartheta \wedge q}}{n}}}{\|\kappa\|} + p^{-\frac{(\sqrt{r}-\sqrt{q})_+^2}{2K}} + \sqrt{p^{\vartheta - 1} + \frac{p^{(\vartheta - q)_+}}{n}}\sqrt{\rho_1(L,M)}\Bigg].$$

This quantity err_p combines the 'bias' term associated with the useful features that we have missed in feature selection and the 'variance' term associated with retained features; see Lemmas 2.2 and 2.3 for details. Throughout this paper, we assume that there is a constant $C > 0$ such that

$$err_p \leq p^{-C}. \tag{2.15}$$

Remark. Note that $\rho_1(L,M) \geq 1$ and $\rho_2(L,M) \geq 1$. A relatively small $\rho_1(L,M)$ means that $\tau(j)$ are more or less in the same magnitude, and a relatively small $\rho_2(L,M)$ means that the $(K-1)$ nonzero eigenvalues of $LM\Lambda^2 M' L'$ have comparable magnitudes. Our hope is that neither of these two ratios is unduly large.

2.3. *Main theorem: clustering consistency by IF-PCA.* Recall $\psi_{n,j}$ is the KS statistic. For any threshold $t > 0$, denote the set of retained features by

$$\hat{S}_p(t) = \{1 \leq j \leq p : \psi_{n,j} \geq t\}.$$

For any $n \times p$ matrix W, let $W^{\hat{S}_p(t)}$ be the matrix formed by replacing all columns of W with the index $j \notin \hat{S}_p(t)$ by the vector of zeros (note the slight difference compared with $W^{(t)}$ in Section 1.1). Denote the $n \times (K-1)$ matrix of the first $(K-1)$ left singular vectors of $W^{\hat{S}_p(t_p(q))}$ by

$$\hat{U}^{(t_p(q))} = \hat{U}(W^{\hat{S}_p(t_p(q))}) = [\hat{\eta}_1, \hat{\eta}_2, \cdots, \hat{\eta}_{K-1}], \quad \text{where } \hat{\eta}_k = \hat{\eta}_k(W^{\hat{S}_p(t_p(q))}).$$

Recall that $W = [LM + Z\Sigma^{-1/2}]\Lambda + R$ and let $LM\Lambda = UDV'$ be the Singular Value Decomposition (SVD) of $LM\Lambda$ such that $D \in R^{K-1,K-1}$ is a diagonal matrix with the diagonals being singular values arranged descendingly, $U \in R^{n,K-1}$ satisfies $U'U = I_{K-1}$, and $V \in R^{p,K-1}$ satisfies $V'V = I_{K-1}$. Then U is the non-stochastic counterpart of $\hat{U}^{(t_p(q))}$. We hope that the linear space spanned by columns of $\hat{U}^{(t_p(q))}$ is "close" to that spanned by columns of U.

Definition 2.2. *$L_p > 0$ denotes a multi-$\log(p)$ term that may vary from occurrence to occurrence but satisfies $L_p p^{-\delta} \to 0$ and $L_p p^{\delta} \to \infty$, $\forall \delta > 0$.*

For any $K \geq 1$, let

$$\mathcal{H}_K = \{\text{All } K \times K \text{ orthogonal matrices}\}. \tag{2.16}$$

The following theorem is proved in Section 2.7, which shows that the singular vectors IF-PCA obtains span a low-dimensional subspace that is "very close" to its counterpart in the ideal case where there is no noise.

Theorem 2.1. *Fix $(\vartheta, \theta) \in (0,1)^2$, and consider $ACM(\vartheta, \theta)$. Suppose the regularity conditions (2.8), (2.11), (2.12), (2.13) and (2.15) hold, and the threshold in IF-PCA is set as $t = t_p(q)$ as in (2.14). Then there is a matrix H in $\mathcal{H}_{K-1}$ such that as $p \to \infty$, with probability at least $1 - o(p^{-2})$, $\|\hat{U}^{(t_p(q))} - UH\|_F \leq L_p err_p$.*

Recall that in IF-PCA, once $\hat{U}^{(t_p(q))}$ is obtained, we estimate the class labels by truncating $\hat{U}^{(t_p(q))}$ entry-wise (see the PCA-1 step and the footnote in Section 1.1) and then cluster by applying the classical k-means. Also, the estimated class labels are denoted by $\hat{y}^{IF}_{t_p(q)} = (\hat{y}^{IF}_{t_p(q),1}, \hat{y}^{IF}_{t_p(q),2}, \hat{y}^{IF}_{t_p(q),n})'$. We measure the clustering errors by the Hamming distance

$$\text{Hamm}^*_p(\hat{y}^{IF}_{t_p(q)}, y) = \min_{\pi} \Big\{ \sum_{i=1}^{n} P(\hat{y}^{IF}_{t_p(q),i} \neq \pi(y_i)) \Big\},$$

where π is any permutation in $\{1, 2, \ldots, K\}$. The following theorem is our main result, which gives an upper bound for the Hamming errors of IF-PCA.

Theorem 2.2. *Fix $(\vartheta, \theta) \in (0,1)^2$, and consider $ACM(\vartheta, \theta)$. Suppose the regularity conditions (2.8), (2.11), (2.12), (2.13) and (2.15) hold, and let $t_p = t_p(q)$ as in (2.14) and $T_p = \log(p)/\sqrt{n}$ in IF-PCA. As $p \to \infty$,*

$$n^{-1}\text{Hamm}^*_p(\hat{y}^{IF}_{t_p(q)}, y) \leq L_p err_p.$$

The theorem can be proved by Theorem 2.1 and an adaption of (Jin, 2015, Theorem 2.2). In fact, by Lemma 2.1 below, the absolute values of all entries of U are bounded by $C/\sqrt{n}$ from above. By the choice of T_p and definitions, the truncated matrix $\hat{U}_*^{(t_p(q))}$ satisfies $\|\hat{U}_*^{(t_p(q))} - UH\|_F \leq \|\hat{U}^{(t_p(q))} - UH\|_F$. Using this and Theorem 2.1, the proof of Theorem 2.2 is basically an exercise of classical theory on k-means algorithm. For this reason, we skip the proof.

2.4. *Two corollaries and a phase transition phenomenon.* Corollary 2.1 can be viewed as a simplified version of Theorem 2.1, so we omit the proof; recall that L_p denotes a generic multi-log(p) term.

Corollary 2.1. *Suppose conditions of Theorem 2.1 hold, and suppose* $\max\{\rho_1(L,M), \rho_2(L,M)\} \leq L_p$ *as* $p \to \infty$. *Then there is a matrix* H *in* $\mathcal{H}_{K-1}$ *such that as* $p \to \infty$, *with probability at least* $1 - o(p^{-2})$,

$$\begin{aligned}\|\hat{U}^{(t_p(q))} - UH\|_F &\leq L_p p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)} \\ &+ L_p(\|\kappa\|^{-1} p^{(1-\vartheta)/2} + 1) \begin{cases} p^{-\theta/2 + [(\vartheta - q)_+]/2}, & \text{if } (1-\vartheta) > \theta, \\ p^{-(1-\vartheta)/2 + [(1-\theta-q)_+]/2}, & \text{if } (1-\vartheta) \leq \theta. \end{cases}\end{aligned}$$

By assumption (2.12), the interesting range for a nonzero $m_k(j)$ is $|m_k(j)| \asymp L_p n^{-1/6}$. It follows that $\|\kappa\| \asymp L_p p^{(1-\vartheta)/2} n^{-1/6}$ and $\|\kappa\|^{-1} p^{(1-\vartheta)/2} \to \infty$. In this range, we have the following corollary, which is proved in Section 2.7.

Corollary 2.2. *Suppose conditions of Corollary 2.1 hold, and* $\|\kappa\| = L_p p^{(1-\vartheta)/2} n^{-1/6}$. *Then as* $p \to \infty$, *the following holds:*

(a) *If* $(1-\vartheta) < \theta/3$, *for any* $r > 0$, *whatever* q *is chosen, the upper bound of* $\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))} - UH\|_F$ *in Corollary 2.1 goes to infinity.*

(b) *If* $\theta/3 < (1-\vartheta) < 1 - 2\theta/3$, *for any* $r > \vartheta - 2\theta/3$, *there exists* $q \in (0, r)$ *such that* $\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))} - UH\|_F \to 0$ *with probability at least* $1 - o(p^{-2})$. *In particular, if* $(1-\vartheta) \leq \theta$ *and* $r > (\sqrt{K(1-\vartheta) - K\theta/3} + \sqrt{1-\theta})^2$, *by taking* $q = 1 - \theta$,

$$\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))} - UH\|_F \leq L_p n^{1/6} s_p^{-1/2};$$

if $(1-\vartheta) > \theta$ *and* $r > (\sqrt{2K\theta/3} + \sqrt{\vartheta})^2$, *by taking* $q = \vartheta$,

$$\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))} - UH\|_F \leq L_p n^{-1/3}.$$

(c) *If* $(1-\vartheta) > 1 - 2\theta/3$, *for any* $r > 0$, *by taking* $q = 0$, $\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))} - UH\|_F \to 0$ *with probability at least* $1 - o(p^{-2})$.

To interpret Corollary 2.2, we take a special case where $K = 2$, all diagonals of Σ are bounded from above and below by a constant, and all nonzero features $\mu_k(j)$ have comparable magnitudes; that is, there is a positive number u_0 that may depend on (n, p) and a constant $C > 0$ such that

$$u_0 \leq |\mu_k(j)| \leq Cu_0, \qquad \text{for any } (k, j) \text{ such that } \mu_k(j) \neq 0. \tag{2.17}$$

In our parametrization, $s_p = p^{1-\vartheta}$, $n = p^{\theta}$, and $u_0 \asymp \tau_{min}^{1/3}/n^{1/6} \asymp (\log(p)/n)^{1/6}$ since $K = 2$. Cases (a)-(c) in Corollary 2.2 translate to (a) $1 \ll s_p \ll n^{1/3}$, (b) $n^{1/3} \ll s_p \ll p/n^{2/3}$, and (c) $s_p \gg p/n^{2/3}$, respectively.

The primary interest in this paper is Case (b). In this case, Corollary 2.2 says that both feature selection and post-selection PCA can be successful, provided that $u_0 = c_0(\log(p)/n)^{1/6}$ for an appropriately large constant c_0. Case (a) addresses the case of very sparse signals, and Corollary 2.2 says that we need stronger signals than that of $u_0 \asymp (\log(p)/n)^{1/6}$ for IF-PCA to be successful. Case (c) addresses the case where signals are relatively dense, and PCA is successful without feature selection (i.e., taking $q = 0$).

We have been focused on the case $u_0 = L_p n^{-1/6}$ as our primary interest is on clustering by IF-PCA. For a more complete picture, we model u_0 by $u_0 = L_p p^{-\alpha}$; we let the exponent α vary and investigate what is the critical order for u_0 for some different problems and different methods. In this case, it is seen that $u_0 \sim n^{-1/6}$ is the critical order for the success of feature selection (see Section 2.5), $u_0 \sim \sqrt{p/(ns)}$ is the critical order for the success of Classical PCA and $u_0 \sim 1/\sqrt{s}$ is the critical order for IF-PCA in an idealized situation where the Screen step finds exactly all the useful features. These suggest an interesting phase transition phenomenon for IF-PCA.

- *Feature selection is trivial but clustering is impossible.* $1 \ll s \ll n^{1/3}$ and $n^{-1/6} \ll u_0 \leq 1/\sqrt{s}$. Individually, useful features are sufficiently strong, so it is trivial to recover the support of $M\Sigma^{1/2}$ (say, by thresholding the KS-scores one by one); note that $M\Sigma^{1/2} = [\mu_1, \mu_2, \ldots, \mu_K]'$. However, useful features are so sparse that it is impossible for any methods to have consistent clustering.
- *Clustering and feature selection are possible but non-trivial.* $n^{1/3} \ll s \ll p/n^{2/3}$ and $u_0 = (r\log(p)/n)^{1/6}$, where r is a constant. In this range, feature selection is indispensable and there is a region where IF-PCA may yield a consistent clustering but Classical PCA may not. A similar conclusion can be drawn if the purpose is to recover the support of $M\Sigma^{1/2}$ by thresholding the KS-scores.
- *Clustering is trivial but feature selection is impossible.* $s \gg p/n^{2/3}$ and $\sqrt{p/(ns)} \leq u_0 \ll n^{-1/6}$. In this range, the sparsity level is low and

> Classical PCA is able to yield consistent clustering, but the useful features are individually too weak that it is impossible to fully recover the support of $M\Sigma^{1/2}$ by using all p different KS-scores.

In Jin, Ke and Wang (2015b), we investigate the phase transition with much more refined studies (in a slightly different setting).

2.5. *Tail probability of KS statistic.* IF-PCA consists of a screening step (IF-step) and a PCA step. In the IF-step, the key is to study the tail behavior of the KS statistic $\psi_{n,j}$, defined in (1.6). Fix $1 \leq j \leq p$. Recall that in our model, $X_i \sim N(\bar{\mu} + \mu_k, \Sigma)$ if $i \in$ Class k, $1 \leq i \leq n$, and that j is a useless feature if and only if $\mu_1(j) = \mu_2(j) = \ldots = \mu_K(j) = 0$.

Recall that $a_0 = \sqrt{(\pi - 2)/(4\pi)}$. Theorem 2.3 addresses the tail behavior of $\psi_{n,j}$ when feature j is useless.

Theorem 2.3. *Fix $\theta \in (0, 1)$ and let $n = n_p = p^\theta$. Fix $1 \leq j \leq p$. If the j-th feature is a useless feature, then as $p \to \infty$, for any sequence t_p such that $t_p \to \infty$ and $t_p/\sqrt{n} \to 0$,*

$$1 \lesssim \frac{P(\psi_{n,j} \geq t_p)}{(\sqrt{2}a_0)^{-1}\exp\big(-t_p^2/(2a_0^2)\big)} \lesssim 2.$$

We conjecture that $P(\psi_{n,j} \geq t_p) \sim 2 \cdot \frac{1}{\sqrt{2}a_0}\exp(-t_p^2/(2a_0^2))$, with possibly a more sophisticated proof than that in the paper.

Recall that τ is defined in (2.10). Theorem 2.4 addresses the tail behavior of $\psi_{n,j}$ when feature j is useful.

Theorem 2.4. *Fix $\theta \in (0, 1)$. Let $n = n_p = p^\theta$, and $\tau(j)$ be as in (2.10), where j is a useful feature. Suppose (2.12) and (2.13) hold, and the threshold t_p is such that $t_p \to \infty$, that $t_p/\sqrt{n} \to 0$, and that $\tau(j) \geq (1+C)t_p$ for some constant $C > 0$. Then as $p \to \infty$,*

$$P(\psi_{n,j} \leq t_p) \leq C\bigg(K\exp\big(-\frac{1}{2Ka_0^2}(\tau(j) - t_p)^2\big) + O(p^{-3})\bigg).$$

Theorems 2.3-2.4 are proved in the supplementary material Jin and Wang (2015). Combining two theorems, roughly saying, we have that

- if j is a useless feature, then the right tail of $\psi_{n,j}$ behaves like that of $N(0, a_0^2)$,
- if j is a useful feature, then the left tail of $\psi_{n,j}$ is bounded by that of $N(\tau(j), Ka_0^2)$.

These suggest that the feature selection using the KS statistic in the current setting is very similar to feature selection with a Stein's normal means model; the latter is more or less well-understood (e.g., Abramovich et al. (2006)).

As a result, the most interesting range for $\tau(j)$ is $\tau(j) \geq O(\sqrt{\log(p)})$. If we threshold the KS-scores at $t_p(q) = \sqrt{2q\log(p)}$, by similar argument as in feature selection with a Stein's normal means setting, we expect that

- All useful features are retained, except for a fraction $\leq Cp^{-[(\sqrt{r}-\sqrt{q})_+]^2/K}$,
- No more than $(1+o(1))\cdot p^{1-q}$ useless features are (mistakenly) retained,
- $\#\{\text{retained features}\} = |\hat{S}_p(t_p(q))| \leq C[p^{1-\vartheta} + p^{1-q} + \log(p)]$.

These facts pave the way for the PCA step; see Sections below.

Remark. Theorem 2.4 hinges on $\tau(j)$, which is a quantity proportional to the "third moment" $\sum_{k=1}^K \delta_k m_k^3(j)$ and can be viewed as the "effective signal strength" of the KS statistic. In the symmetric case (say, $K = 2$ and $\delta_1 = \delta_2 = 1/2$), the third moment (which equals to 0) is no longer the right quantity for calibrating the effective signal strength of the KS statistic, and we must use the fourth moment. In such cases, for $1 \leq j \leq p$, let

$$\omega(j) = \sqrt{n} \sup_{-\infty<y<\infty} \left[\frac{1}{8} y(1-3y^2)\phi(y)\cdot\big(\sum_{k=1}^K \delta_k m_k^2(j)\big)^2 + \frac{1}{24}\phi^{(3)}(y)\cdot\sum_{k=1}^K \delta_k m_k^4(j)\right],$$

where $\phi^{(3)}(y)$ is the third derivative of the standard normal density $\phi(y)$. Theorem 2.4 continues to hold provided that (a) $\tau(j)$ is replaced by $\omega(j)$, (b) the condition (2.12) of $\tau_{min} \geq a_0\sqrt{2r\log(p)}$ is replaced by that of $\omega_{min} \geq a_0\sqrt{2r\log(p)}$, where $\omega_{min} = \min_{j \in S_p(M)}\{\omega(j)\}$, and (c) the first part of condition (2.13), $\max_{j\in S_p(M)} \big\{\frac{\sqrt{n}}{\tau(j)} \sum_{k=1}^K \delta_k m_k^4(j)\big\} \leq Cp^{-\delta}$, is replaced by that of $\max_{j\in S_p(M)} \big\{\frac{\sqrt{n}}{\omega(j)} \sum_{k=1}^K \delta_k |m_k(j)|^5\big\} \leq Cp^{-\delta}$. This is consistent with that in Arias-Castro and Verzelen (2014), which studies the clustering problem in a similar setting (especially on the symmetric case) with great details.

In the literature, tight bounds of this kind are only available for the case where X_i are iid samples from a known distribution (especially, parameters—if any—are known). In this case, the bound is derived by Kolmogorov (1933); also see Shorack and Wellner (1986). The setting considered here is more complicated, and how to derive tight bounds is an interesting but rather challenging problem. The main difficulty lies in that, any estimates of the unknown parameters $(\bar{\mu}(j), \mu_1(j), \ldots, \mu_k(j), \sigma(j))$ have stochastic fluctuations at the same order of that of the stochastic fluctuation of the empirical CDF, but two types of fluctuations are correlated in a complicated way, so it is hard to derive the right constant a_0 in the exponent. There are two existing approaches, one is due to Durbin (1985) which approaches the problem by approximating the stochastic process by a Brownian bridge, the other is due to Loader et al. (1992) (see also Siegmund (1982); Woodroofe (1978)) on the local Poisson approximation of the boundary crossing probability. It is argued in Loader et al. (1992) that the second approach is more accurate. Our proofs follow the idea in Siegmund (1982); Loader et al. (1992).

2.6. *Post-selection eigen-analysis.* For the PCA step, as in Section 2.3, we let $W^{\hat{S}_p(t_p(q))}$ be the $n \times p$ matrix where the j-th column is the same as that of W if $j \in \hat{S}_p(t_p(q))$ and is the zero vector otherwise. With such notations,

$$W^{\hat{S}_p(t_p(q))} = LM\Lambda + L(M - M^{\hat{S}_p(t_p(q))})\Lambda + (Z\Sigma^{-1/2}\Lambda + R)^{\hat{S}_p(t_p(q))}. \tag{2.18}$$

We analyze the there terms on the right hand side separately.

Consider the first term $LM\Lambda$. Recall that $L \in R^{n,K}$ with the i-th row being e_k' if and only if $i \in$ Class k, $1 \leq i \leq n, 1 \leq k \leq K$, and $M \in R^{K,p}$ with the k-th row being $m_k' = (\Sigma^{-1/2}\mu_k)'$, $1 \leq k \leq K$. Also, recall that $A = \mathrm{diag}(\sqrt{\delta_1}\|m_1\|, \ldots, \sqrt{\delta_K}\|m_K\|)$ and $\Omega \in R^{K,K}$ with $\Omega(k,\ell) = m_k'\Lambda^2 m_\ell/(\|m_k\| \cdot \|m_\ell\|)$, $1 \leq k, \ell \leq K$. Note that $rank(A\Omega A) = rank(LM) = K-1$. Assume all nonzero eigenvalues of $A\Omega A$ are simple, and denote them by $\lambda_1 > \lambda_2 > \ldots > \lambda_{K-1}$. Write

$$A\Omega A = Q \cdot \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{K-1}) \cdot Q', \qquad Q \in R^{K,K-1}, \tag{2.19}$$

where the k-th column of Q is the k-th eigenvector of $A\Omega A$, and let

$$LM\Lambda = UDV' \tag{2.20}$$

be an SVD of $LM\Lambda$. Introduce

$$G = \mathrm{diag}(\sqrt{\delta_1}, \sqrt{\delta_2}, \ldots, \sqrt{\delta_K}) \in R^{K,K}. \tag{2.21}$$

The following lemma is proved in the supplementary material (Jin and Wang, 2015, Section **??**).

Lemma 2.1. *The matrix $LM\Lambda$ has $(K-1)$ nonzero singular values which are $\sqrt{n\lambda_1}, \ldots, \sqrt{n\lambda_{K-1}}$. Also, there is a matrix $H \in \mathcal{H}_{K-1}$ (see (2.16)) such that*

$$U = n^{-1/2}L[G^{-1}QH] \in R^{n,K-1}.$$

For the matrix $G^{-1}QH$, the ℓ^2-norm of the k-th row is $(\delta_k^{-1} - 1)^{1/2}$, and the ℓ^2-distance between the k-th row and the ℓ-th row is $(\delta_k^{-1} + \delta_\ell^{-1})^{1/2}$, which is no less than 2, $1 \leq k < \ell \leq K$.

By Lemma 2.1 and definitions, it follows that

- For any $1 \leq i \leq n$ and $1 \leq k \leq K-1$, the i-th row of U equals to the k-th row of $n^{-1/2}G^{-1}QH$ if and only if Sample i comes from Class k.
- The matrix U has K distinct rows, according to which the rows of U partition into K different groups. This partition coincides with the partition of the n samples into K different classes. Also, the ℓ^2-norm between each pair of the K distinct rows is no less than $2/\sqrt{n}$.

Consider the second term on the right hand side of (2.18). This is the 'bias' term caused by useful features which we may fail to select.

LEMMA 2.2. *Suppose the conditions of Theorem 2.1 hold. As $p \to \infty$, with probability at least $1 - o(p^{-2})$,*

$$\|L(M - M^{\hat{S}_p(t_p(q))})\Lambda\| \leq C\|\kappa\|\sqrt{n}\cdot\Big[p^{-(1-\vartheta)/2}\sqrt{\rho_1(L,M)}\cdot\sqrt{\log(p)} + p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)}\Big].$$

Consider the last term on the right hand side of (2.18). This is the 'variance' term consisting of two parts, the part from original measurement noise matrix Z and the remainder term due to normalization.

LEMMA 2.3. *Suppose the conditions of Theorem 2.1 hold. As $p \to \infty$, with probability at least $1 - o(p^{-2})$,*

$$\|(Z\Sigma^{-1/2}\Lambda + R)^{\hat{S}_p(t_p(q))}\| \leq C\Big[\sqrt{n} + \Big(p^{(1-\vartheta\wedge q)/2} + \|\kappa\|p^{(\vartheta-q)_+/2}\sqrt{\rho_1(L,M)}\Big)\cdot(\sqrt{\log(p)})^3\Big].$$

Combining Lemmas 2.2-2.3 and using the definition of err_p,

$$\|W^{\hat{S}_p(t_p(q))} - LM\Lambda\| \leq L_p err_p \cdot \frac{\sqrt{n}\|\kappa\|}{\rho_2(L,M)}. \tag{2.22}$$

2.7. *Proofs of the main results.* We now show Theorem 2.1 and Corollary 2.1. Proof of Theorem 2.2 is very similar to that of Theorem 2.2 in Jin (2015) and proof of Corollary 2.2 is elementary, so we omit them.

Consider Theorem 2.1. Let

$$T = LM\Lambda^2M'L', \qquad \hat{T} = W^{\hat{S}_p(t_p(q))}(W^{\hat{S}_p(t_p(q))})'.$$

Recall that U and $\hat{U}^{(t_p(q))}$ contain the $(K-1)$ leading eigenvectors of T and $\hat{T}$, respectively. Using the sine-theta theorem (Davis and Kahan, 1970) (see also Proposition 1 in Cai, Ma and Wu (2013)),

$$\|\hat{U}^{(t_p(q))}(\hat{U}^{(t_p(q))})' - UU'\| \leq 2s_{\min}^{-1}(T)\|\hat{T} - T\|; \tag{2.23}$$

in (2.23), we have used the fact that T has a rank of $K-1$ so that the gap between the $(K-1)$-th and K-th largest eigenvalues is equal to the minimum nonzero singular value $s_{\min}(T)$. The following lemma is proved in the supplementary material (Jin and Wang, 2015, Section **??**).

Lemma 2.4. *For any integers $1 \leq m \leq p$ and two $p \times m$ matrices V_1, V_2 satisfying $V_1'V_1 = V_2'V_2 = I$, there exists an orthogonal matrix $H \in R^{m,m}$ such that $\|V_1 - V_2 H\|_F \leq \|V_1V_1' - V_2V_2'\|_F$.*

Combine (2.23) with Lemma 2.4 and note that $\hat{U}^{(t_p(q))}(\hat{U}^{(t_p(q))})' - UU'$ has a rank of $2K$ or smaller. It follows that there is an $H \in \mathcal{H}_{K-1}$ such that

$$\|\hat{U}^{(t_p(q))} - UH\|_F \leq 2\sqrt{2K} s_{\min}^{-1}(T)\|\hat{T} - T\|. \tag{2.24}$$

First, $\|\hat{T} - T\| \leq 2\|LM\Lambda\| \cdot \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\| + \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|^2$. From Lemmas 2.2-2.3 and (2.15), $\|LM\Lambda\| \gg \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|$. Therefore,

$$\|\hat{T} - T\| \lesssim 2\|LM\Lambda\|\|W^{\hat{S}_p(t_p(q))} - LM\Lambda\| \leq 2\sqrt{n}\|\kappa\| \cdot \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|.$$

Second, by Lemma 2.1,

$$s_{\min}(T) = n \cdot s_{\min}(A\Omega A') = n\|\kappa\|^2/\rho_2(L, M).$$

Plugging in these results into (2.24), we find that

$$\|\hat{U}^{(t_p(q))} - UH\|_F \leq 4\sqrt{2K}\frac{\rho_2(L, M)}{\sqrt{n}\|\kappa\|}\|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|, \tag{2.25}$$

where by Lemmas 2.2-2.3, the right hand side equals to $L_p err_p$. The claim then follows by combining (2.25) and (2.22).

Consider Corollary 2.2. For each $j \in S_p(M)$, it can be deduced that $\kappa(j) \geq \epsilon(M)$, using especially (2.11). Therefore, $\|\kappa\| \geq L_p p^{\frac{(1-\vartheta)}{2}} n^{-1/6} = L_p p^{\frac{1-\vartheta}{2} - \frac{\theta}{6}}$. The error bound in Corollary 2.1 reduces to

$$L_p p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)} + L_p \begin{cases} p^{-\theta/3 + (\vartheta - q)_+/2}, & \theta < 1 - \vartheta, \\ p^{\theta/6 - (1-\vartheta)/2 + (1-\theta-q)_+/2}, & \theta \geq 1 - \vartheta. \end{cases} \tag{2.26}$$

Note that (2.26) is lower bounded by $L_p p^{\theta/6 - (1-\vartheta)/2}$ for any $q \geq 0$; and it is upper bounded by $L_p p^{-\theta/3 + \vartheta/2}$ when taking $q = 0$. The first and third claims then follow immediately. Below, we show the second claim.

First, consider the case $\theta < 1 - \vartheta$. If $r > \vartheta$, we can take any $q \in (\vartheta, r)$ and the error bound is $o(1)$. If $r \leq \vartheta$, noting that $(\vartheta - r)/2 < \theta/3$, there exists $q < r$ such that $(\vartheta - q)/2 < \theta/3$, and the corresponding error bound is $o(1)$.

In particular, if $r > (\sqrt{2K\theta/3} + \sqrt{\vartheta})^2$, we have $(\sqrt{r} - \sqrt{\vartheta})^2/(2K) > \theta/3$; then for $q \geq \vartheta$, the error bound is $L_p p^{-\theta/3} + L_p p^{-(\sqrt{r}-\sqrt{q})^2/(2K)}$; for $q < \vartheta$, the error bound is $L_p p^{-\theta/3+(\vartheta-q)/2}$; so the optimal $q^* = \vartheta$ and the corresponding error bound is $L_p p^{-\theta/3} = L_p n^{-1/3}$.

Next, consider the case $1 - \vartheta \leq \theta < 3(1 - \vartheta)$. If $r > 1 - \theta$, for any $q \in (1-\theta, r)$, the error bound is $o(1)$; note that $\theta/6 < (1-\vartheta)/2$. If $r \leq 1-\theta$, noting that $(1 - \theta - r)/2 < (1 - \vartheta)/2 - \theta/6$, there is a $q < r$ such that $(1 - \theta - q)/2 < (1 - \vartheta)/2 - \theta/6$, and the corresponding error bound is $o(1)$.

In particular, if $r > (\sqrt{K(1-\vartheta) - K\theta/3} + \sqrt{1-\theta})^2$, we have that $(\sqrt{r} - \sqrt{1-\theta})^2/(2K) > (1 - \vartheta)/2 - \theta/6$; then for $q \geq 1 - \theta$, the error bound is $L_p p^{\theta/6-(1-\vartheta)/2} + L_p p^{-(\sqrt{r}-\sqrt{q})^2/(2K)}$; for $q < 1 - \theta$, the error bound is $L_p p^{\theta/6-(1-\vartheta)/2+(1-\theta-q)/2}$; so the optimal $q^* = 1 - \theta$ and the corresponding error bound is $L_p p^{\theta/6-(1-\vartheta)/2} = L_p n^{1/6} s_p^{-1/2}$.

3. Simulations. We conducted a small-scale simulation study to investigate the numerical performance of IF-PCA. We consider two variants of IF-PCA, denoted by IF-PCA(1) and IF-PCA(2). In IF-PCA(1), the threshold is chosen using HCT (so the choice is data-driven), and in IF-PCA(2), the threshold t is given. In both variants, we skip the normalization step on KS scores (that step is designed for microarray data only). The pseudocodes of IF-PCA(2) and IF-PCA(1) are given in Table 7 (Section 2) and Table 8, respectively. We compared IF-PCA(1) and IF-PCA(2) with 4 other different methods: classical k-means (kmeans), k-means++ (kmeans+), classical hierarchical clustering (Hier), and SpectralGem (SpecGem; same as classical PCA). In hierarchical clustering, we only consider the linkage type of "complete"; other choices of linkage have very similar results.

TABLE 8

Pseudocode for IF-PCA(1) (for simulations; threshold set by Higher Criticism)

```
    Input: data matrix $X$, number of classes $K$. Output: class label vector $\hat{y}^{IF}_{HC}$.
1.  Rank features: Let $\psi_{n,j}$ be the KS-scores as in (1.6), and $F_0$ be the CDF of $\psi_{n,j}$ under null, $1 \leq j \leq p$.
2.  Threshold choice by HCT: Calculate $P$-values by $\pi_j = 1 - F_0(\psi_{n,j})$, $1 \leq j \leq p$ and sort them by
    $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$. Define $HC_{p,j} = \sqrt{p}(j/p - \pi_{(j)})/\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p}$, and let
    $\hat{j} = \mathrm{argmax}_{\{j:\pi_{(j)} > \log(p)/p, j<p/2\}}\{HC_{p,j}\}$. HC threshold $t_p^{HC}$ is the $\hat{j}$-largest KS-score.
3.  Post-selection PCA: Define post-selection data matrix $W^{(HC)}$ (i.e., sub-matrix of $W$ consists of all
    column $j$ of $W$ with $\psi_{n,j} > t_p^{HC}$). Let $U \in R^{n,K-1}$ be the matrix of the first $(K-1)$ left singular
    vectors of $W^{(HC)}$. Cluster by $\hat{y}^{IF}_{HC} = kmeans(U, K)$.
```

In each experiment, we fix parameters $(K, p, \theta, \vartheta, r, rep)$, two probability mass vectors $\delta = (\delta_1, \cdots, \delta_K)'$ and $\gamma = (\gamma_1, \gamma_2, \gamma_3)'$, and three probability densities g_σ, g_μ defined over $(0, \infty)$ and $g_{\tilde{\mu}}$ defined over $(-\infty, \infty)$. With these parameters, we let $n = n_p = p^\theta$ and $\epsilon_p = p^{1-\vartheta}$; n is the sample size, ϵ_p is roughly the fraction of useful features, and rep is the number of repetitions.[10] We generate the $n \times p$ data matrix X as follows.

- Generate the class labels $y_1, y_2, \ldots, y_n$ iid from $MN(K, \delta)$[11], and let L be the $n \times K$ matrix such that the i-th row of L equals to e_k' if and only if $y_i = k$, $1 \leq k \leq K$.
- Generate the overall mean vector $\bar{\mu}$ by $\bar{\mu}(j) \overset{iid}{\sim} g_{\bar{\mu}}$, $1 \leq j \leq p$.
- Generate the contrast mean vectors $\mu_1, \cdots, \mu_K$ as follows. First, generate $b_1, b_2, \ldots, b_p$ iid from Bernoulli(ϵ_p). Second, for each j such that $b_j = 1$, generate the iid signs $\{\beta_k(j)\}_{k=1}^{K-1}$ such that $\beta_k(j) = -1, 0, 1$ with probability $\gamma_1, \gamma_2, \gamma_3$, respectively, and generate the feature magnitudes $\{h_k(j)\}_{k=1}^{K-1}$ iid from g_μ. Last, for $1 \leq k \leq K-1$, set μ_k by (the factor 72π is chosen to be consistent with (2.10))
$$\mu_k(j) = \big[72\pi \cdot (2r\log(p)) \cdot n^{-1} \cdot h_k(j)\big]^{1/6} \cdot b_j \cdot \beta_k(j),$$
and let $\mu_K = -\frac{1}{\delta_K}\sum_{k=1}^{K-1} \delta_k \mu_k$.
- Generate the noise matrix Z as follows. First, generate a $p \times 1$ vector σ by $\sigma(j) \overset{iid}{\sim} g_\sigma$. Second, generate the n rows of Z iid from $N(0, \Sigma)$, where $\Sigma = \text{diag}(\sigma^2(1), \sigma^2(2), \cdots, \sigma^2(p))$.
- Let $X = \mathbf{1}\bar{\mu}' + L[\mu_1, \cdots, \mu_K] + Z$.

In the simulation settings, r can be viewed as the parameter of (average) signal strength. The density g_σ characterizes noise heteroscedasticity; when g_σ is a point mass at 1, the noise variance of all the features are equal. The density g_μ controls the strengths of useful features; when g_μ is a point mass at 1, all the useful features have the same strength. The signs of useful features are captured in the probability vector γ; when $K = 2$, we always set $\gamma_2 = 0$ so that $\mu_k(j) \neq 0$ for a useful feature j; when $K \geq 3$, for a useful feature j, we allow $\mu_k(j) = 0$ for some k.

For IF-PCA(2), the theoretical threshold choice as in (2.14) is $t = \sqrt{2\tilde{q}\log(p)}$ for some $0 < \tilde{q} < (\pi - 2)/(4\pi) \approx .09$. We often set $\tilde{q} \in \{.03, .04, .05, .06\}$, depending on the signal strength parameter r.

The simulation study contains 5 experiments, which we now describe.

Experiment 1. In this experiment, we study the effect of signal strength over clustering performance, and compare two cases: the classes have unequal or equal number of samples. We set $(K, p, \theta, \vartheta, rep) = (2, 4 \times 10^4, .6, .7, 100)$, and $\gamma = (.5, 0, .5)$ (so that the useful features have equal probability to have

[10] For each parameter setting, we generate the X matrix for rep times, and at each time, we apply all the six algorithms. The clustering errors are averaged over all the repetitions.

[11] We say $X \sim MN(K, \delta)$ if $P(X = k) = \delta_k$, $1 \leq k \leq K$; MN stands for multinomial.

positive and negative signs). Denote by $U(a,b)$ the uniform distribution over $(a-b, a+b)$. We set g_μ as $U(.8, 1.2)$, g_σ as $U(1, 1.2)$, and $g_{\bar{\mu}}$ as $N(0,1)$. We investigate two choices of δ: $(\delta_1, \delta_2) = (1/3, 2/3)$ and $(\delta_1, \delta_2) = (1/2, 1/2)$; we call them "asymmetric" and "symmetric" case, respectively. In the latter case, the two classes roughly have equal number of samples. The threshold in IF-PCA(2) is taken to be $t = \sqrt{2 \cdot .06 \cdot \log(p)}$.

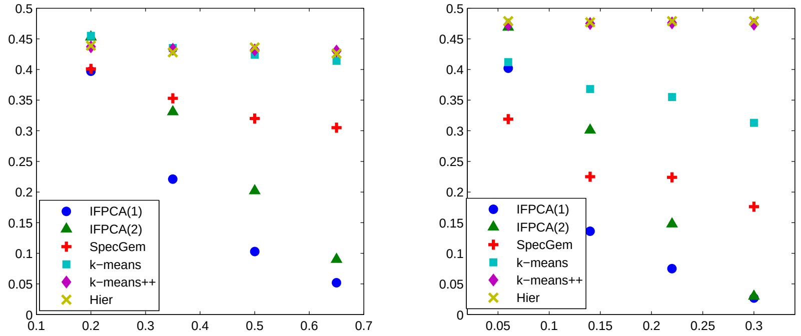

FIG 4. *Comparison of clustering error rates (Experiment 1a). x-axis: signal strength parameter r. y-axis: error rates. Left: $\delta = (1/3, 2/3)$. Right: $\delta = (1/2, 1/2)$.*

In Experiment 1a, we let the signal strength parameter $r \in \{.20, .35, .50, .65\}$ for the asymmetric case, and $r \in \{.06, .14, .22, .30\}$ for the symmetric case. The results are summarized in Figure 4. We find that two versions of IF-PCA outperform the other methods in most settings, increasingly so when the signal strength increases. Moreover, two versions of IF-PCA have similar performance, with those of IF-PCA(1) being slightly better. This suggests that our threshold choice by HCT is not only data-driven but also yields satisfactory clustering results. On the other hand, it also suggests that IF-PCA is relatively insensitive to different choices of the threshold, as long as they are in a certain range.

In Experiment 1b, we make a more careful comparison between the asymmetric and symmetric cases. Note that for the same parameter r, the actual signal strength in the symmetric case is stronger because of normalization. As a result, for $\delta = (1/3, 2/3)$, we still let $r \in \{0.20, 0.35, 0.50, 0.65\}$, but for $\delta = (1/2, 1/2)$, we take $r' = c_0 \times \{0.20, 0.35, 0.50, 0.65\}$, where c_0 is a constant chosen such that for any $r > 0$, r and $c_0 r$ yield the same value of $\kappa(j)$ (see (2.9)) in the asymmetric and symmetric cases, respectively; we note that $\kappa(j)$ can be viewed as the effective signal-to-noise ratio of Kolmogorov-Smirnov statistic. The results are summarized in Table 9. Both versions of IF-PCA have better clustering results when $\delta = (1/3, 2/3)$, suggesting that the clustering task is more difficult in the symmetric case. This is consistent with the theoretical results; see for example Arias-Castro and Verzelen (2014); Jin, Ke and Wang (2015b).

TABLE 9

Comparison of average clustering error rates (Experiment 1). Number in the brackets are standard deviations of the error rates.

	$(\delta_1, \delta_2) = (1/2, 1/2)$		$(\delta_1, \delta_2) = (1/3, 2/3)$	
r	IF-PCA(1)	IF-PCA(2)	IF-PCA(1)	IF-PCA(2)
.20	.467(.04)	.481(.01)	.391(.11)	.443(.08)
.35	.429(.08)	.480(.02)	.253(.15)	.341(.16)
.50	.368(.13)	.466(.05)	.144(.14)	.225(.18)
.65	.347(.13)	.459(.07)	.099(.12)	.098(.11)

Experiment 2. In this experiment, we allow feature sparsity to vary (Experiment 2a), and investigate the effect of unequal feature strength (Experiment 2b). We set $(K, p, \theta, r, rep) = (2, 4 \times 10^4, .6, .3, 100)$ (so $n = 577$), $\gamma = (.5, 0, .5)$ and $(\delta_1, \delta_2) = (1/3, 2/3)$. The threshed for IF-PCA(2) is $t = \sqrt{2 \cdot .05 \cdot \log(p)}$.

In Experiment 2a, we let ϑ range in $\{.68, .72, .76, .80\}$. Since the number of useful features is roughly $p^{1-\vartheta}$, a larger ϑ corresponds to a higher sparsity level. For any μ and $a, b > 0$, let $\widetilde{TN}(u, b^2, a)$ be the conditional distribution of $(X|u-a \leq X \leq u+a)$ for $X \sim N(u, b^2)$, where TN stands for "Truncated Normal". We take $g_{\bar{\mu}}$ as $N(0,1)$, g_μ as $\widetilde{TN}(1, .1^2, .2)$, and g_σ as $\widetilde{TN}(1, .1^2, .1)$. The results are summarized in the left panel of Figure 5, where for all sparsity levels, two versions of IF-PCA have similar performance, and each of them significantly outperforms the other methods.

In Experiment 2b, we use the same setting except that g_μ is $\widetilde{TN}(1, .1, .7)$ and g_σ is the point mass at 1. Note that in Experiment 2a, the support of g_μ is $(.8, 1.2)$, and in the current setting, the support is $(.3, 1.7)$ which is wider. As a result, the strengths of useful features in the current setting have more variability. At the same time, we force the noise variance of all features to be 1, for a fair comparison. The results are summarized in the right panel of Figure 5. They are similar to those in Experiment 2a, suggesting that IF-PCA continues to work well even when the feature strengths are unequal.

Experiment 3. In this experiment, we study how different threshold choices affect the performance of IF-PCA. With the same as those in Experiment 2b, we investigate four threshold choices for IF-PCA(2): $t = \sqrt{2\tilde{q}\log(p)}$ for $\tilde{q} \in \{.03, .04, .05, .06\}$, where we recall that the theoretical choice of threshold (2.14) suggests $0 < \tilde{q} < .09$. The results are summarized in Table 10, which suggest that IF-PCA(1) and IF-PCA(2) have comparable performances, and

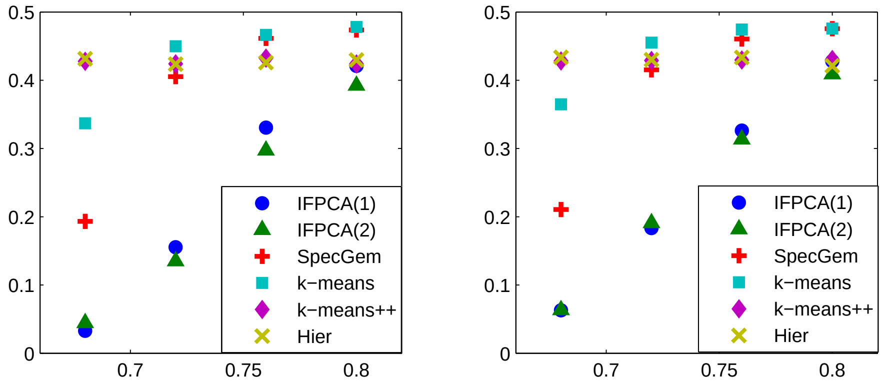

FIG 5. *Comparison of average clustering error rates (Experiment 2). x-axis: sparsity parameter ϑ. y-axis: error rates. Left: g_μ is $\widetilde{TN}(1, .1^2, .2)$ and g_σ is $\widetilde{TN}(1, .1^2, .1)$. Right: g_μ is $\widetilde{TN}(1, .1, .7)$ and g_σ is point mass at 1.*

that IF-PCA(2) is relatively insensitive to different threshold choices, as long as they fall in a certain range. However, the best threshold choice does depend on ϑ. From a practical view point, since ϑ is unknown, it is preferable to set the threshold in a data-driven fashion; this is what we use in IF-PCA(1).

TABLE 10
Comparison of average clustering error rates (Experiment 3). Numbers in the brackets are the standard deviations of the error rates.

	Threshold ($\tilde{q}$)	$\vartheta = .68$	$\vartheta = .72$	$\vartheta = .76$	$\vartheta = .80$
IF-PCA(1)	HCT (stochastic)	.053(.08)	.157(.16)	.337(.14)	.433(.10)
IF-PCA(2)	.03	.038(.05)	.152(.12)	.345(.13)	.449(.06)
	.04	.045(.08)	.122(.12)	.312(.15)	.427(.09)
	.05	.068(.12)	.154(.15)	.303(.16)	.413(.12)
	.06	.118(.15)	.237(.17)	.339(.16)	.423(.10)

Experiment 4. In this experiment, we investigate the effects of correlations among the noise over the clustering results. We generate the data matrix X the same as before, except for that the noise matrix Z is replaced by ZA, for a matrix $A \in R^{p,p}$. Fixing a number $d \in (-1, 1)$, we consider three choices of A, (a)-(c). In (a), $A(i,j) = 1\{i = j\} + d \cdot 1\{j = i+1\}$, $1 \leq i, j \leq p$. In (b)-(c), fixing an integer $N > 1$, for each $j = 1, 2, \ldots, p$, we randomly generate a size N subset of $\{1, 2, \ldots, p\} \setminus \{j\}$, denoted by $I_N(j)$. We then let $A(i,j) = 1\{i = j\} + d \cdot 1\{i \in I_N(j)\}$. For (b), we take $N = 5$ and for (c), we take $N = 20$. We set $d = .1$ in (a)-(c). We set $(K, p, \theta, \vartheta, r, rep) = (4, 2 \times 10^4, .5, .6, .7, 100)$ (so $n = 141$), and $(\delta_1, \delta_2, \delta_3, \delta_4) = (1/4, 1/4, 1/4, 1/4)$, $\gamma = (.3, .05, .65)$. For an

exponential random variable $X \sim Exp(\lambda)$, denote the density of $\big[b+X|a_1 \leq b+X \leq a_2\big]$ by $\widetilde{TSE}(\lambda, b, a_1, a_2)$, where TSE stands for 'Truncated Shifted Exponential'. We take $g_{\bar{\mu}}$ as $N(0,1)$, g_μ as $\widetilde{TSE}(.1, .9, -\infty, \infty)$ (so it has a mean 1), and g_σ as $\widetilde{TSE}(.1, .9, .9, 1.2)$. The threshold for IF-PCA(2) is $t = \sqrt{2 \cdot .03 \cdot \log(p)}$. The results are summarized in the left panel of Figure 6, which suggest that IF-PCA continues to work in the presence of correlations among the noise: IF-PCA significantly outperforms the other 4 methods, especially for the randomly selected correlations.

Experiment 5. In this experiment, we study how different noise distributions affect the clustering results. We generate the data matrix X the same as before, except for the distribution of the noise matrix Z is different. We consider three different settings for the noise matrix Z: (a) for a vector $a = (a_1, a_2, \ldots, a_K)$, generate row i of Z by $Z_i \overset{iid}{\sim} N(0, a_k I_p)$ if Sample i comes from Class k, $1 \leq k \leq K$, $1 \leq i \leq n$, (b) $Z = \sqrt{2/3}\tilde{Z}$, where all entries of $\tilde{Z}$ are *iid* samples from $t_6(0)$, where $t_6(0)$ denotes the central t-distribution with $df = 6$, (c) $Z = [\tilde{Z} - 6]/\sqrt{12}$, where the entries of $\tilde{Z}$ are *iid* samples from the chi-squared distribution with $df = 6$ (in (b)-(c), the constants of $\sqrt{2/3}$ and $\sqrt{12}$ are chosen so that each entry of Z has zero mean and unit variance). We set $(K, p, \theta, \vartheta, r, rep) = (4, 2 \times 10^4, .5, .55, 1, 100)$, $(\delta_1, \delta_2, \delta_3, \delta_4) = (1/4, 1/4, 1/3, 1/6)$, and $\gamma = (.4, .1, .5)$. We take $g_{\bar{\mu}}$ to be $N(0,1)$. In case (a), we take $(a_1, a_2, a_3, a_4) = (0.8, 1, 1.2, 1.4)$. The threshold for IF-PCA(2) is set as $t = \sqrt{2 \cdot .03 \cdot \log(p)}$. The results are summarized in the right panel of Figure 6, which suggest that IF-PCA continues to outperform the other 4 clustering methods.

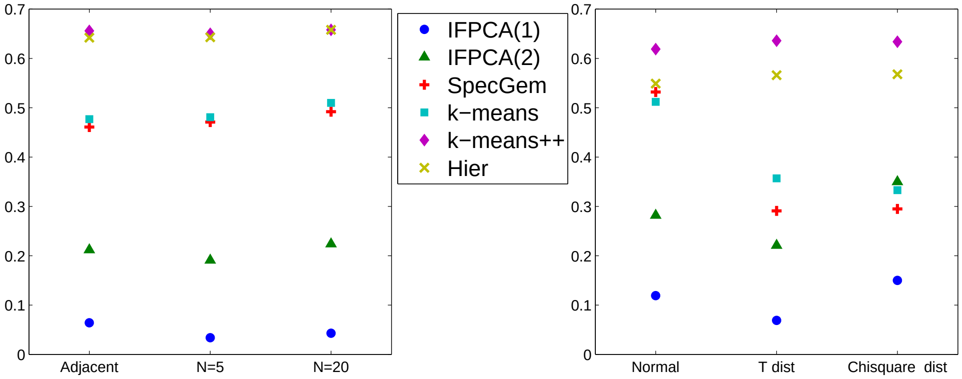

FIG 6. *Comparison of average clustering error rates for Experiment 4 (left panel) and Experiment 5 (right panel). y-axis: error rates*

4. Connections and extensions.

We propose IF-PCA as a new spectral clustering method, and we have successfully applied the method to clustering using gene microarray data. IF-PCA is a two-stage method which consists of a marginal screening step and a post-selection clustering step. The methodology contains three important ingredients: using the KS statistic for marginal screening, post-selection PCA, and threshold choice by HC.

The KS statistic can be viewed as an omnibus test or a goodness-of-fit measure. The methods and theory we developed on the KS statistic can be useful in many other settings, where it is of interest to find a powerful yet robust test. For example, they can be used for nonGaussian detection of the Cosmic Microwave Background (CMB) or can be used for detecting rare and weak signals or small cliques in large graphs (e.g., Donoho and Jin (2015)).

The KS statistic can also be viewed as a marginal screening procedure. Screening is a well-known approach in high dimensional analysis. For example, in variable selection, we use marginal screening for dimension reduction (Fan and Lv, 2008), and in cancer classification, we use screening to adapt Fisher's LDA and QDA to modern settings (Donoho and Jin, 2008; Efron, 2009; Fan et al., 2015). However, the setting here is very different.

Of course, another important reason that we choose to use the KS-based marginal screening in IF-PCA is for simplicity and practical feasibility: with such a screening method, we are able to (a) use Efron's proposal of empirical null to correct the null distribution, and (b) set the threshold by Higher Criticism; (a)-(b) are especially important as we wish to have a tuning-free and yet effective procedure for subject clustering with gene microarray data. In more complicated situations, it is possible that marginal screening is suboptimal, and it is desirable to use a more sophisticated screening method. We mention two possibilities below.

In the first possibility, we might use the recent approaches by Birnbaum et al. (2013); Paul and Johnstone (2012), where the primary interest is signal recovery or feature estimation. The point here is that, while the two problems—subject clustering and feature estimation—are very different, we still hope that a better feature estimation method may improve the results of subject clustering. In these papers, the authors proposed *Augmented sparse PCA (ASPCA)* as a new approach to feature estimation and showed that under certain sparse settings, ASPCA may have advantages over marginal screening methods, and that ASPCA is asymptotically minimax. This suggests an alternative to IF-PCA, where in the IF step, we replace the marginal KS screening by some augmented feature screening approaches. However, the open question is, how to develop such an approach that is tuning-free and practically feasible. We leave this to the future work.

Another possibility is to combine the KS statistic with the recent innovation of Graphlet Screening (Jin, Zhang and Zhang (2014); Ke, Jin and Fan (2014)) in variable selection. This is particularly appropriate if the columns of the noise matrix Z are correlated, where it is desirable to exploit the graphic structures of the correlations to improve the screening efficiency. Graphic Screening is a graph-guided multivariate screening procedure and has advantages over the better known method of marginal screening and the lasso. At the heart of Graphlet Screening is a graph, which in our setting is defined as follow: each feature j, $1 \leq j \leq p$, is a node, and there is an edge between nodes i and j if and only if row i and row j of the normalized data matrix W are strongly correlated (note that for a useful feature, the means of the corresponding row of W are nonzero; in our range of interest, these nonzero means are at the order of $n^{-1/6}$, and so have negligible effects over the correlations). In this sense, adapting Graphlet Screening in the screening step helps to solve highly correlated data. We leave this to the future work.

The post-selection PCA is a flexible idea that can be adapted to address many other problems. Take model (1.1) for example. The method can be adapted to address the problem of testing whether $LM = 0$ or $LM \neq 0$ (that is, whether the data matrix consists of a low-rank structure or not), the problem of estimating M, or the problem of estimating LM. The latter is connected to recent interest on sparse PCA and low-rank matrix recovery. Intellectually, the PCA approach is connected to SCORE for community detection on social networks (Jin, 2015), but is very different.

Threshold choice by HC is a recent innovation, and was first proposed in (Donoho and Jin, 2008) (see also (Fan, Jin and Yao, 2013)) in the context of classification. However, our focus here is on clustering, and the method and theory we need are very different from those in (Donoho and Jin, 2008; Fan, Jin and Yao, 2013). In particular, this paper requires sophisticated post-selection Random Matrix Theory (RMT), which we do not need in (Donoho and Jin, 2008; Fan, Jin and Yao, 2013). Our study on RMT is connected to (Johnstone, 2001; Paul, 2007; Baik and Silverstein, 2006; Guionnet and Zeitouni, 2000; Lee, Zou and Wright, 2010) but is very different.

In a high level, IF-PCA is connected to the approaches by (Azizyan, Singh and Wasserman, 2013; Chan and Hall, 2010) in that all three approaches are two-stage methods that consist of a screening step and a post-selection clustering step. However, the screening step and the post-selection step in all three approaches are significantly different from each other. Also, IF-PCA is connected to the spectral graph partitioning algorithm by (Ng, Jordan and Weiss, 2002), but it is very different, especially in feature selection and threshold choice by HC.

In this paper, we have assumed that the first $(K-1)$ contrast mean vectors $\mu_1, \mu_2, \ldots, \mu_{K-1}$ are linearly independent (consequently, the rank of the matrix M (see (2.6)) is $(K-1)$), and that K is known (recall that K is the number of classes). In the gene microarray examples we discuss in this paper, a class is a patient group (normal, cancer, cancer sub-type) so K is usually known to us as a priori. Moreover, it is believed that different cancer sub-types can be distinguished from each other by one or more genes (though we do not know which) so $\mu_1, \mu_2, \ldots, \mu_{K-1}$ are linearly independent. Therefore, both assumptions are reasonable.

On the other hand, in a broader context, either of these two assumptions could be violated. Fortunately, at least to some extent, the main ideas in this paper can be extended. We consider two cases. In the first one, we assume K is known but $r = \text{rank}(M) < (K-1)$. In this case, the main results in this paper continue to hold, provided that some mild regularity conditions hold. In detail, let $U \in R^{n,r}$ be the matrix consisting the first r left singular vectors of $LM\Lambda$ as before; it can be shown that, as before, U has K distinct rows. The additional regularity condition we need here is that, the ℓ^2-norm between any pair of the K distinct rows has a reasonable lower bound. In the second case, we assume K is unknown and has to be estimated. In the literature, this is a well-known hard problem. To tackle this problem, one might utilize the recent developments on rank detection (Kritchman and Nadler, 2008) (see also (Cai, Ma and Wu, 2013; Birnbaum et al., 2013)), where in a similar setting, the authors constructed a confident lower bound for the number of classes K. A problem of interest is then to investigate how to combine the methods in these papers with IF-PCA to deal with the more challenging case of unknown K; we leave this for future study.

Acknowledgements. The authors would like to thank David Donoho, Shiqiong Huang, Tracy Zheng Ke, Pei Wang, and anonymous referees for valuable pointers and discussion.

SUPPLEMENTARY MATERIAL

Supplementary Material for "Influential Features PCA for high dimensional clustering"
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). Owing to space constraints, the technical proofs are relegated a supplementary document Jin and Wang (2015). It contains three sections, Sections **??**–**??**.

References.

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653.

AMINI, A. and WAINWRIGHT, M. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877-2921.

ARIAS-CASTRO, E., LERMAN, G. and ZHANG, T. (2013). Spectral clustering based on local PCA. *arXiv:1301.2007*.

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

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.

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

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

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084.

CAI, T., MA, Z. and WU, Y. (2013). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815.

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

CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: the EM approach. *Ann. Statist.* **37** 2523–2542.

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7** 1–46.

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

DONOHO, D. (2015). 50 years of data science. *Manuscript.*

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

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

DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Stat. Sci.* **30** 1–25.

DURBIN, J. (1985). The first-passage density of a continuous Gaussian process to a general boundary. *J. Appl. Probab.* 99–122.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing. *J. Amer. Statist. Soc.* **99** 96-104.

EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Soc.* **104** 1015–1028.

FAN, Y., JIN, J. and YAO, Z. (2013). Optimal classification in sparse Gaussian graphic model. *Ann. Statist.* **41** 2537–2571.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. B* **70** 849–911.

FAN, J., KE, Z. T., LIU, H. and XIA, L. (2015). QUADRO: A supervised dimension reduction method via Rayleigh quotient optimization. *Ann. Statist.* **43** 1498-1534.

GORDON, G. J., JENSEN, R. V., HSIAO, L.-L., GULLANS, S. R., BLUMENSTOCK, J. E., RAMASWAMY, S., RICHARDS, W. G., SUGARBAKER, D. J. and BUENO, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research* **62** 4963–4967.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning*, 2nd ed. Springer.

Jin, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89.

Jin, J., Ke, Z. T. and Wang, W. (2015a). Optimal spectral clustering by Higher Criticism Thresholding. *Manuscript.*

Jin, J., Ke, Z. T. and Wang, W. (2015b). Phase transitions for high dimensional clustering and related problems. *arXiv:1502.06952.*

Jin, J. and Ke, Z. T. (2016). Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistica Sinica* **26** 1–34.

Jin, J. and Wang, W. (2015). Supplementary material for "Influential Features PCA for high dimensional clustering". *Manuscript.*

Jin, J., Zhang, C.-H. and Zhang, Q. (2014). Optimality of Graphlet Screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772.

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

Jung, S. and Marron, J. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130.

Ke, Z., Jin, J. and Fan, J. (2014). Covariance assisted screening and estimation. *Ann. Statist.* **42** 2202–2242.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4** 83–91.

Kritchman, S. and Nadler, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemometr. Intell. Lab* **94** 19–32.

Lee, A. B., Luca, D. and Roeder, K. (2010). A spectral graph approach to discovering genetic ancestry. *Ann. Appl. Statist.* **4** 179–202.

Lee, S., Zou, F. and Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist.* **38** 3605.

Lei, J. and Vu, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *Ann. Stat.* **43** 299–322.

Loader, C. R. et al. (1992). Boundary crossing probabilities for locally Poisson processes. *Ann. Appl. Probab.* **2** 199–228.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801.

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **2** 849–856.

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

Paul, D. and Johnstone, I. M. (2012). Augmented sparse principal component analysis for high dimensional data. *arXiv:1202.1242.*

Shorack, G. and Wellner, J. (1986). *Empirical processes with applications to statistics.* John Wiley & Sons.

Siegmund, D. (1982). Large deviations for boundary crossing probabilities. *Ann. Prob.* **10** 581–588.

Woodroofe, M. (1978). Large deviations of likelihood ratio statistics with applications to sequential testing. *Ann. Statist.* 72–84.

Yousefi, M. R., Hua, J., Sima, C. and Dougherty, E. R. (2010). Reporting bias when using real data sets to analyze classification performance. *Bioinformatics* **26** 68–76.

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

J. Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213
USA
E-mail: jiashun@stat.cmu.edu

W. Wang
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213
USA
E-mail: wwang@stat.cmu.edu