Important Feature PCA for high dimensional clustering

Article · July 2014

Source: arXiv

2 authors:

Jiashun Jin
Carnegie Mellon University

Wanjie Wang
University of Pennsylvania

Important Features PCA for high dimensional clustering

Jiashun Jin and Wanjie Wang

Carnegie Mellon University

Abstract

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 Important 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, where the threshold is chosen by adapting the recent notion of Higher Criticism, 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 to these singular vectors. It can be seen that 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 [16] 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.

Keywords: Empirical null, Feature Selection, gene microarray, Hamming distance, phase transition, post-selection spectral clustering, sparsity.

1 Introduction

Consider a clustering problem where we have n different p-dimensional feature vectors from K possible classes:

$$X_i, \qquad i = 1, 2, \ldots, n.$$

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. In a typical setting, we have patients from several different classes (e.g., normal, 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 gene expression data to predict them.

Table 1 tabulates 10 gene microarray data sets (arranged alphabetically). Data sets 1, 3, 4, 7, 8, and 9 were analyzed and cleaned in [9], Data set 5 is from [20], Data sets 2, 6, 10 were analyzed and grouped into two classes in [45], among which Data set 10 was cleaned in the same way as by [9]. All the data sets can be found at `www.stat.cmu.edu/~jiashun/Research/software/GenomicsData`, each in a separate sub-folder, with detailed information in `ReadMe.txt`. The data sets are analyzed in Section 1.6, after our approach is fully introduced.

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

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

Table 1: 10 gene microarray data sets. Note that K is small and $p \gg n$ (p is the number of genes and n is number of subjects).

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 $p \times p$ matrix where the i-th diagonal entry is a_i, $1 \leq i \leq p$. We assume

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

and the vector $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(p))'$ is unknown to us.

Assumption (1.2) is only for simplicity, and may not fit very well with the gene microarray data. Fortunately, 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 1.3 and 1.6 for more discussions.

Denote the overall mean vector by

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} E[X_i]. \tag{1.3}$$

For K different vectors $\mu_1, \mu_2, \ldots, \mu_K \in R^p$, we model $E[X_i]$ by

$$E[X_i] = \bar{\mu} + \mu_k, \qquad \text{if and only if} \qquad y_i = k. \tag{1.4}$$

For $1 \leq k \leq K$, let δ_k be the fraction of samples in Class k. By (1.3)-(1.4),

$$\sum_{k=1}^{K} \delta_k \mu_k = 0, \tag{1.5}$$

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

$$\mu_1, \mu_2, \ldots, \mu_{K-1} \text{ are linearly independent.} \tag{1.6}$$

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, $1 \leq k \leq K$. 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 this regime, classical methods such as the k-means [35], hierarchical methods [22], and classical Principle Component Analysis (PCA) are either computationally challenging or ineffective. Our primary interest is to develop new methods that are appropriate for the regime of $p \gg n$.

We propose *Important Features PCA (IF-PCA)* as a new spectral clustering method. Conceptually, IF-PCA has two steps. In the first step, we exploit the sparsity of the contrast mean vectors and perform a feature selection, where we remove many columns of the data matrix X leaving only those we think are important for clustering (and so the name of Important Features). In the second step, we apply the classical PCA to the post-selection data matrix.

1.1 Normalizations

Before we proceed further, we normalize the data and introduce some notations. Denote the data matrix by (as long as there is no confusion, we may drop the subscripts of the matrices for simplicity):

$$X = X_{n,p} = [X_1, X_2, \ldots, X_n]'.$$

Introduce the empirical mean and variance associated with feature j by

$$\bar{X}(j) = \frac{1}{n}\sum_{i=1}^{n} X_i(j) \text{ and } \hat{\sigma}^2(j) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(j) - \bar{X}(j))^2, \quad 1 \leq j \leq p.$$

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

$$W(i,j) = \frac{X_i(j) - \bar{X}(j)}{\hat{\sigma}(j)}, \qquad 1 \leq i \leq n, 1 \leq j \leq p.$$

Introduce

$$\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}^2(1), \hat{\sigma}^2(2), \ldots, \hat{\sigma}^2(p)), \qquad \widetilde{\Sigma} = E[\hat{\Sigma}].$$

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 complications in notations but such a nuisance is unavoidable. Denote for short

$$\Lambda = \Sigma^{1/2}\widetilde{\Sigma}^{-1/2}. \tag{1.7}$$

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 $L = L_{n,K}$ be the $n \times K$ matrix where the i-th row is e_k' if and only if $i \in$ Class k, and let M be the $K \times p$ matrix

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

Similarly, we can write W as the sum of a 'signal component' and a 'noise component':

$$W = L_{n,K}M_{K,p}\Lambda + (Z\Sigma^{-1/2}\Lambda + R), \tag{1.8}$$

where R stands for the reminder term

$$R = \mathbf{1}_n(\bar{\mu} - \bar{X})'\hat{\Sigma}^{-1/2} + [L_{n,K}M_{K,p}\Sigma^{1/2} + Z](\hat{\Sigma}^{-1/2} - \widetilde{\Sigma}^{-1/2}). \tag{1.9}$$

The 'signal component' $LM\Lambda$ a low-rank matrix containing the most direct information of the class labels and so it is of great interest. For the 'noise component', the major term is $Z\Sigma^{-1/2}$, the entries of which are *iid* samples from $N(0,1)$.

1.2 Classical PCA and the challenge it faces

By (1.6), $rank(LM\Lambda) = K - 1$. We have the following Singular Value Decomposition (SVD):

$$LM\Lambda = UDV', \qquad U = U_{n,K-1},\ D = D_{K-1,K-1},\ V = V_{p,K-1}, \tag{1.10}$$

where D is the $(K-1) \times (K-1)$ diagonal matrix with the diagonals being the singular values arranged in the descending order, and $U'U = V'V = I_{K-1}$. Columns of U and V are called *left and right singular vectors* of $LM\Lambda$, respectively. Below, in Lemma 2.1, we have the following.

$U = U_{n,K-1}$ has K distinct rows, according to which the rows of U partition into K different groups, and each group share the same value. The partition coincides with the partition of the samples into K different clusters. The ℓ^2-norm between two distinct rows of U is no less than $1/\sqrt{n}$.

Example 1. *The case of $K = 2$ is illustrative, in which case U reduces to an $n \times 1$ vector, the i-th row of which equals to $\delta_1/\sqrt{n\delta_1\delta_2}$ if $i \in$ Class 1 and $-\delta_2/\sqrt{n\delta_1\delta_2}$ if $i \in$ Class 2.*

Principle Component Analysis (PCA) is a standard approach to clustering [42]. Since rank$(LM\Lambda) = K-1$, a conventional use of PCA is a two-step procedure as follows: (a) Obtain an $n \times (K-1)$ matrix $\hat{U}$ where the k-th column is the k-th left singular-vector of W, (b) Clustering by applying the classical k-means algorithm [22] to $\hat{U}$, assuming there are no more than K clusters. See Section 1.3 for more detailed descriptions.

To differentiate from IF-PCA to be introduced, we call the above procedure the *Classical PCA*. Classical PCA is connected to SpetralGem by [33] and SCORE by [24] but is different.

The rationale behind Classical PCA lies in the *hope* that the 'noise component' in (1.9) has a negligible effect when we obtain the matrix $\hat{U}$ consisting the $(K-1)$ leading left singular vectors of W, and so

$$\hat{U} \approx U. \tag{1.11}$$

In fact, if $\hat{U}$ and U are 'sufficiently close' to each other, then by the aforementioned results on U, the rows of $\hat{U}$ partition into K different groups: at the center of each group is one of the K distinct rows of U, and each row in this group fall very close to this center; also, the partition coincides with the partition of samples into K different classes. In such a scenario, Classical PCA works just fine.

Unfortunately, in modern application examples, we rarely have (1.11) or the idealized scenario above. In fact, for (1.11) to hold, it is necessary that

$$\|Z\Sigma^{-1/2}\| \ll \|LM\Lambda\|. \tag{1.12}$$

However, in modern applications, it is frequently the case that $p \gg n$ and signals are rare and weak, and it is hard for (1.12) to hold. Rare/weak signals is a recent notion which in the current setting means that most columns of $LM\Lambda$ are 0, and the ℓ^2-norm of each nonzero columns of $LM\Lambda$ is relatively small. See [13] for more discussions.

To fix this problem, a standard approach is feature selection: we select a small fraction of features which we think are important, and then apply Classical PCA to the post-selection data matrix. Recall that Λ is approximately the identity matrix, so for notational simplicity, we think $LM\Lambda$ as LM for a minute. The key idea is that each time when we remove a useless feature, the spectral norm of 'signal part' LM remains the same, while that of 'noise part' $Z\Sigma^{-1/2}$ decreases. In detail, if we denote $\widehat{M}$, and $\widehat{Z\Sigma^{-1/2}}$ by the post-selection counterparts of $M, Z\Sigma^{-1/2}$ correspondingly (formed by restricting the columns of the former to the selected features). For the current approach to work, we only need

$$\|L\widehat{M}\| \gg \|\widehat{Z\Sigma^{-1/2}}\|, \tag{1.13}$$

which is much less stringent than (1.12).

The discussions above contain the central idea of IF-PCA, where we use Kolmogorov-Smirnov (KS) statistics for feature selection. In Section 1.3, we introduce IF-PCA and use a gene microarray data to show how it overcomes the challenge faced by Classical PCA. Sections 1.4-1.5

contain further developments of IF-PCA: Section 1.4 explains the rationale of using KS for feature selection and Section 1.5 proposes a data-driven approach for setting the threshold in the feature selection of IF-PCA.

1.3 Important Features PCA

Write $W' = [W_1, W_2, \ldots, W_n]$, and let $F_{n,j}(t)$ be the empirical CDF associated with feature j:

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

Fixing $T_p = \log(p)$ (see the remark below), IF-PCA contains four steps. The first two are 'IF' steps where we select features. The last two are 'PCA' steps where we apply Classical PCA to post-selection data matrix.

1. For each $1 \leq j \leq p$, we compute a Kolmogorov-Smirnov (KS) [39] statistic by

$$\psi_{n,j} = \sqrt{n} \cdot \sup_{-\infty < t < \infty} |F_{n,j}(t) - \Phi(t)|, \qquad (\Phi\text{: CDF of } N(0,1)). \tag{1.14}$$

2. For analysis of gene microarray data, we follow the suggestions by [16] and renormalize the KS scores by

$$\psi^*_{n,j} = \frac{\psi_{n,j} - \text{mean of all } p \text{ different } KS\text{-scores}}{\text{SD of all } p \text{ different } KS\text{-scores}}. \tag{1.15}$$

   Such a step is largely designed for gene microarray data, and is not necessary if rows of Z are iid $N(0, \Sigma)$ as in model (1.1).

3. For a threshold $t > 0$ to be determined, let

$$\hat{S}_t = \{1 \leq j \leq p : \psi^*_{n,j} \geq t\}. \tag{1.16}$$

   For short, let $W^{(t)}$ by the $n \times |\hat{S}_t|$ matrix formed by restricting the columns of W to the set of indices in $\hat{S}_t$, and let $\hat{U}^{(t)}$ be the $n \times (K-1)$ matrix where the k-th column is the (normalized) k-th left leading singular vector of $W^{(t)}$. Define an $n \times (K-1)$ matrix $\hat{U}^{(t)}_*$ by $\hat{U}^{(t)}_*(i,k) = \hat{U}(i,k)1\{|\hat{U}(i,k)| \leq T_p/\sqrt{n}\} + T_p \mathrm{sgn}(\hat{U}(i,k))1\{|\hat{U}(i,k)| > T_p/\sqrt{n}\}$, $1 \leq i \leq n, 1 \leq k \leq K-1$.

4. Apply the classical k-means algorithm to $\hat{U}^{(t)}_*$, that is, to find an $n \times (K-1)$ matrix $\mathcal{U}$ with no more than K distinct rows to minimize $\|\hat{U}^{(t)}_* - \mathcal{U}\|_F$, where $\|\cdot\|_F$ denotes the matrix Frobenius norm. In the $\mathcal{U}$ we find, each distinct row determines a class; the resultant class labels are our estimated labels, denoted by $\hat{y}^{IF}_{t,1}, \hat{y}^{IF}_{t,2}, \ldots, \hat{y}^{IF}_{t,n}$.

Remark. $\hat{U}^{(t)}_*$ is the matrix of truncating $\hat{U}$ entry-wise at threshold $T_p/\sqrt{n}$. This is largely for convenience in proofs and is rarely necessary in practice. We usually take $T_p = \log(p)$ but $\log(p)$ can be replaced by any sequences that tend to ∞ as $p \to \infty$. In this sense, we only have one tuning parameter t; later, we set t in a data driven fashion.

In Table 2, we investigate IF-PCA with the Lung Cancer(1) data listed in Table 1. We tabulate the clustering errors by IF-PCA for t from .3876 to 1.4876 (and the number of selected features range from 12529 to 2. For this data set, the clustering error of IF-PCA is as small as 4 with t set properly (and 3 if in the last step of IF-PCA, we clustering by using the signs of the first singular vector of $W^{(t)}$, instead of using k-means). In comparison, the clustering error of Classical PCA (where we don't use a feature selection step) is as large as 22; see the first row of Table 2. This suggests IF-PCA significantly improves over Classical PCA.

For further illustration, we use the same data set and compare the first left singular matrix associated with the pre-selection data matrix W and its post-selection counterpart $W^{(t)}$ (for the later, the threshold t is set in a data driven fashion by the Higher Criticism to be introduced in Section 1.5). For this data set, $K = 2$, so in both cases, the left singular matrix reduces to

Threshold	No. of selected features	Errors
.3876	12529	22
.6076	5758	22
.8276	1057	24
.9376	**484**	**4**
1.0476	**261**	**5**
1.1576	**129**	**7**
1.2676	63	38
1.3776	21	39
1.4876	2	33

Table 2: The number of selected features and clustering errors for Lung Cancer (1) data with different thresholds. Rows highlighted correspond to the sweet spot of the threshold choice.

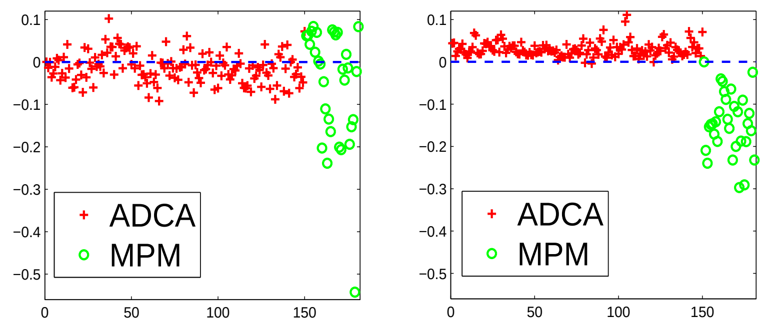

Figure 1: First leading singular vector of data matrix W before feature selection (left) and after feature selection (right; the threshold $t = 1.0573$, set by Higher Criticism in a data driven fashion). y-axis: entries of the left singular vector, x-axis: sample indices. Plots are based on Lung Cancer(1) data, where ADCA and MPM represent two different classes.

a vector. The two vectors are displayed in Figure 1. In the latter (right panel), the entries can be clearly divided into two groups which almost coincide with the truth; such a clear separation does not exist in the former (left panel).

With these being said, two important questions are not answered.

- In (1.15), 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 1.4, we address the first question. In Section 1.5, we propose an approach to threshold choice by the recent notion of Higher Criticism.

1.4 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 for data matrix satisfying Models (1.1)-(1.6), and the focus of Step 2 is to adjust Step 1 in a way so to work well with the 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

$$H_{0,j}: \qquad X_i(j) \overset{iid}{\sim} N(\bar{\mu}(j), \sigma^2(j)), \qquad i = 1, 2, \ldots, n, \tag{1.17}$$

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

$$H_{1j}: \qquad X_i(j) \overset{iid}{\sim} \sum_{k=1}^{K} \delta_k N(\bar{\mu}(j) + \mu_k(j), \sigma^2(j)), \qquad i = 1, 2, \ldots, n, \tag{1.18}$$

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 testing problem above is a well-known difficult problem. For example, Neyman-Pearson's likelihood is known to be unbounded, so that classical Likelihood Ratio Test (LRT) can not be applied directly (e.g., [7]). An alternative is to use moment-based tests, but since $\bar{\mu}(j)$ and $\sigma(j)$ are unknown, the options are tests based on the empirical cumulant or kurtosis. Unfortunately, such tests are vulnerable to outliers and may have undesirable testing powers.

Kolmogorov-Smirnov test is a well-known goodness-of-fit test. Compared to LRT or moment-based tests, KS test is more robust and is ideal for such settings. Our numerical studies, on simulated data and on real data, support the advantage of KS tests. This explains the rationale of Step 1.

We now discuss the rationale of Step 2. We discover an interesting phenomenon, where for illustration, we use the Lung Cancer (1) data set. Ideally, if the normality assumption (1.2) is valid for this data set, then the histogram of the KS scores should fit well the density function of the KS under Model (1.17), obtained by simulations. Unfortunately, this is not the case, and the discrepancy in fitting is substantial. On the other hand, if we renormalize the density function in a way so that it has the same mean and variance as the KS-scores associated with Lung Cancer (1) data set, then it fits very well with the histogram.

Such a phenomenon was discovered earlier by [16], but is on Studentized t-statistics and is in different settings. Following [16], if we call the distribution of KS statistic associated with Model (1.17) the *theoretic null distribution*, and the distribution with mean and variance adjusted to fit with those of the real data the *empirical null distribution*, then the phenomenon we discover is: the histogram of the real KS scores do not fit well with the theoretic null but fit well with the empirical null. See the left panel of Figure 2 for details.

These observations gives rise to the standardization in Step 2. The point can be further illustrated by comparing the empirical survival function of the standardized KS scores in Step 2 with the survival function of the theoretical null; this is the right panel of Figure 2, which suggests that two survival functions are reasonably close to each other.

These also suggest that our proposed approach does not critically depend on the normality assumption: with the adjustment in Step 2, IF-PCA works well for microarray data. This is further validated in Section 1.5.

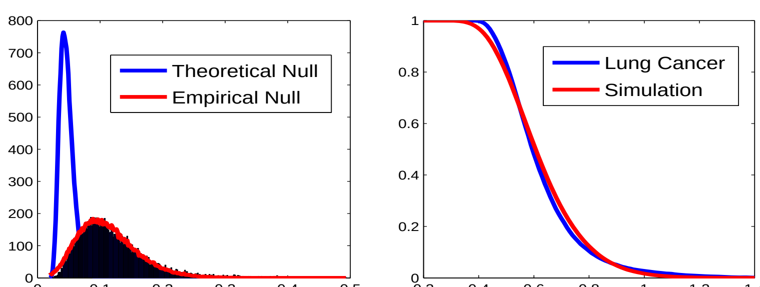

Figure 2: Left: The histogram of KS scores of the Lung Cancer(1) data. The two lines in blue and red denote the theoretical null and empirical null densities, respectively. Right: comparison of the empirical survival function of the standardized KS scores based on Lung Cancer (1) data (red) and the survival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

1.5 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 [11] (see also [3, 21, 23]) as a method for large-scale multiple testing. In [12], HC was also found to be useful for setting threshold for features selection in the context of classification. HC is also useful in many other settings. See [13, 25] 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 before. The approach contains three simple steps as follows.

- For $1 \leq j \leq p$, calculate a P-value $\pi_j = 1 - F_0(KS_{n,j})$, 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 functional by

$$HC_{p,j} = \sqrt{p}(j/p - \pi_{(j)})/\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p}. \tag{1.19}$$

  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.

For example, if $\hat{j} = 50$, then HC selects only the 50 features with the largest KS scores. In the Lung Cancer(1) data set, $\hat{j} = 251$ and $t_p^{HC} = 1.0573$; this is the threshold we use in Figure 1. In the definition, we require $\pi_{(\hat{j})} \leq \log(p)/p$. For the a few very small j, $HC_{p,j}$ is known to be very heavy tailed [11], and we should not rely on them for threshold choice.

Remark. When we apply HC to microarray data, we follow the discussions in Section 1.4 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.

Once the threshold t is determined, we can plug it into (1.16) in Step 3. As a result, IF-PCA becomes a tuning free procedure.

We now explain the rationale behind the HC threshold choice. For any threshold t, the expected number of misclassified samples by IF-PCA is a non-stochastic function of t, denoted by $err(t) = err(t; L, M, \Sigma, n, p)$. Ideally, if $err(t)$ is known to us, then we would set t as the value that minimizes $err(t)$ (we call such a threshold the *optimal threshold*):

$$t_p^{opt} = \text{argmin}_{t>0}\{err(t; L, M, \Sigma, n, p)\}.$$

Now, as in Section 1.3, let $\hat{U}^{(t)}$ be the left singular vector of $W^{(t)}$. In a companion paper [26], we show that if $K = 2$ and the signals are rare and weak, then for t in a range of interest,

$$\hat{U}^{(t)} \propto \widetilde{snr}(t) \cdot U + z + rem, \tag{1.20}$$

where U is as in the SVD of $LM\Lambda$ (i.e., (1.10)), $\widetilde{snr}(t) = \widetilde{snr}(t; L, M, \Sigma, n, p)$ is a non-stochastic function, $z \sim N(0, I_n)$, and rem is the error vector the entries of which are of much smaller magnitude than that of z or $\widetilde{snr}(t) \cdot U$. In a sense, $\widetilde{snr}(t)$ can be viewed as the post-selection Signal-to-Noise Ratio (SNR), so in order for IF-PCA to have the best performance, we hope to choose t to maximize $\widetilde{snr}(t)$ (we call such a threshold the *Ideal Threshold*):

$$t_p^{ideal} = \text{argmin}_{t>0}\{\widetilde{snr}(t; L, M, \Sigma, n, p)\}.$$

In comparison, $\widetilde{snr}(t)$ has a much simpler form than that of $err(t)$, so t_p^{ideal} is much easier to track than t_p^{opt}. On the other hand, it can be shown that in a broad context [26] that

$$t_p^{opt} \approx t_p^{ideal}. \tag{1.21}$$

In [26], we have derived explicit forms of the leading term of $\widetilde{snr}(t; L, M, \Sigma, n, p)$. Using such forms, we are able to derive the leading term of t_p^{ideal}. The central surprise is that, there is an intimate connection between $\widetilde{snr}(t; L, M, \Sigma, n, p)$ and the HC functional defined in (1.19), and so with high probability,

$$t_p^{HC} \approx t_p^{ideal}. \tag{1.22}$$

Combining (1.20)-(1.22) gives $t_p^{HC} \approx t_p^{ideal} \approx t_p^{opt}$, and so HCT is a good threshold choice.

The relationships (1.20)-(1.22) are justified in [26]. The proofs are rather long (70 manuscript pages in Annals of Statistics format), so instead of including it here, it is better to report it separately.

The ideas above are similar to that in [12] and [19], but their focus is on classification and our focus is on clustering; our version of HC is also very different from theirs.

1.6 Applications to gene microarray data

We compare IF-PCA with four other clustering methods: SpectralGem [33] which is the same as Classical PCA introduced earlier, classical k-means, classical hierarchical method [22], and k-means++ [5]. In classical k-means, we apply the textbook k-means algorithm to the normalized data matrix directly, without feature selection. 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 for 30 times (in the last step of either Classical PCA or IF-PCA, k-means is also used where similarly the parameter of 'replicate' is set as 30). The k-means++ [5] is a recent modification of k-means. It improves the performance of k-means in numerical study but is still NP hard in theory. For hierarchical clustering, we choose the linkage function to be 'complete'; using a different linkage function gives more or less the same results. In IF-PCA, the threshold is determined by HCT, so it does not require any tuning parameter. In HCT, the p-values associated with the KS scores are computed using simulated KS scores under the nul with $2 \times 10^3 \times p$ independent replications; see Section 1.5 for remarks on how to compute the p-values.

Table 3: 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.

#	Data set	K	kmeans	kmeans++	Hier	SpecGem	IF-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 3. Since all methods except the hierarchical clustering have algorithmic randomness (they use build-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 3,

$$r = \frac{\text{error rate of IF-PCA}}{\text{minimum of the error rates of the other 4 methods}}. \tag{1.23}$$

The smaller the value of r, the more substantial the improvement that IF-PCA has over the other 4 methods. 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. In Figure 3, we plot the values of r versus the denominators in (1.23) across all 8 data sets with $r < 1$ (a regression line is also added for illustration). The figure suggests 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, unexpectedly, IF-PCA is inferior to Classical PCA with $r = 1.83$, but still outperforms other methods. In Figure 4 we show the clustering errors by IF-PCA (y-axis) for different number of selected features (x-axis; evenly from 1 to p with an increment of 20), for Lung Cancer(1) (top) and Brain data (bottom). Lung Cancer(1) data represents what we typically have in the 10 data sets: there is clearly a sweet spot for threshold choice (and HCT pins down to such a range). Brain data is an 'outlier', where using all features gives the smallest error. Possible reasons include (a) bad data, (b) useful features are not sparse, and (c) the sample size is very small ($n = 42$) so the useful features are individually very weak. When (b)-(c) happen, it is almost impossible to successfully separate the useful features from useless ones, and it is preferable to use Classical PCA. This is related to the phase transition phenomenon to be introduced Section 1.7; see details therein. Second, for Colon Cancer, all methods behave unsatisfactorily, and IF-PCA slightly underperforms Hierarchical method ($r = 1.04$). Colon Cancer is known to be a difficult data set, even when the task is classification where class labels in the training set are given [12]. 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 turn out to be significantly skewed to the right. Therefore, instead of using the normalization (1.15), 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 1.5 for remarks on how to compute the p-value.

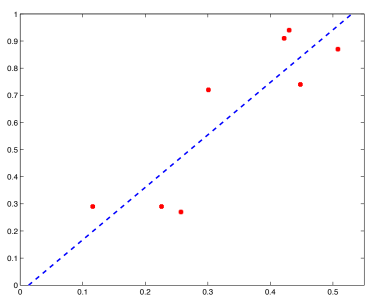

Figure 3: Each of the 8 dots represents a different data set (x-axis: minimum of the error rates for two k-means, SpectralGem, and hierarchical methods. y-axis: the r value).

1.7 Connection to sparse PCA, a phase transition phenomenon

The study is closely related to the recent interest on sparse PCA [29] (see also [4, 30, 34, 37, 46]), but is different in important ways. Take model (1.1) for example, where we have

$$X = LM\Sigma^{1/2} + Z, \qquad \text{rows of } Z \text{ are } iid \text{ from } N(0, \Sigma).$$

In this model, the focus of sparse PCA is to estimate the matrix $M\Sigma^{1/2}$ or to recover the support of $M\Sigma^{1/2}$ (i.e., the set of all useful features [29, 2]), while our focus is to estimate the class label matrix L. We recognize that, the two problems—estimating $M\Sigma^{1/2}$ and clustering—are essentially two different problems, and addressing one successfully does not necessarily addresses 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.

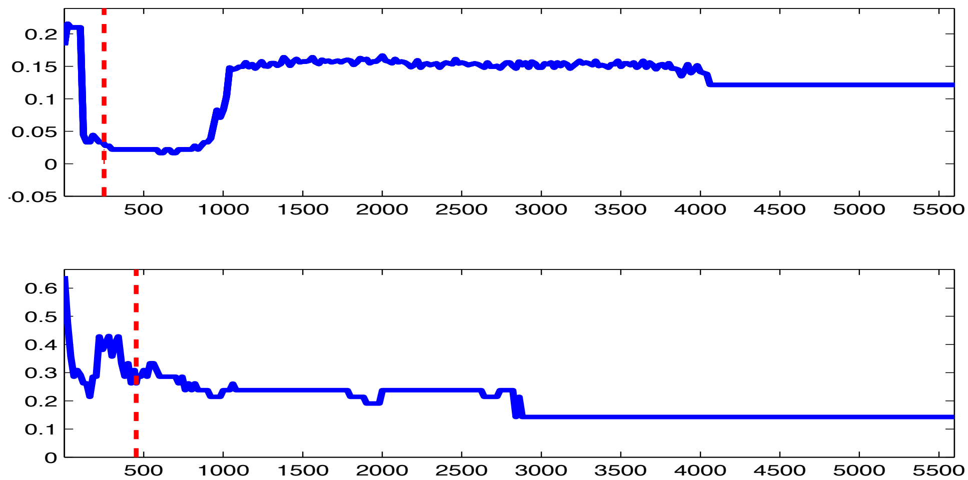

Figure 4: Error rates by IF-PCA (y-axis) with different number of selected features (x-axis) for Lung Cancer(1) data (top) and Su-Cancer data (bottom). Dashed vertical line: number of selected features corresponding to HCT.

- Most of the useful features are very weak and only a few of them are very strong. The later are easy to identify and yield a consistent clustering, still, it may be impossible to estimate $M\Sigma^{1/2}$ well, as most of the useful features are very weak.

Our focus on the study of clustering using microarray data is also different from that on sparse PCA (e.g., [29, 2]). A similar argument can be drawn for the connection between the clustering problem and the recent interest on low-rank matrix recovery.

In a closely related context, our study reveals an interesting phase transition phenomenon that further confirms the observation we made above. For illustration, we take a special case where $K = 2$, all diagonals of Σ are bounded from above and below by a constant, and all nonzero $\mu_k(j)$ 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

$$u_0 \leq |\mu_k(j)| \leq Cu_0, \qquad \text{for any } (k, j) \text{ such that } \mu_k(j) \neq 0. \tag{1.24}$$

Let s be the number of useful features and recall $1 \ll n \ll p$. The phase transition is as follows.

- *Feature selection is trivial but clustering is impossible.* $1 \ll s \ll n^{1/3}$ and $n^{-1/6} \ll u_0 \leq 1/\sqrt{s}$. The useful feature are very sparse but also individually very strong, so it is trivial to recover the support of $M\Sigma^{1/2}$ by thresholding all p different KS scores. However, the useful features are also very sparse and 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. As r increases, consistent clustering by IF-PCA maybe impossible/possible but non-trivial/trivial. 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.

Seemingly, the second case is the most interesting one; we revisit the phase transition in Section 2.3. In a forthcoming manuscript [26], we investigate the phase transition with much more refined studies.

1.8 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 [16] 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 2.4. 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 [43]. Our results follow the work by [40] and [36] 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 as well as subtle tradeoff between bias and variance associated with feature selection. 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 SpetralGem approach by [33], but our focus 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 SNR associated with post-selection eigen-analysis. As mentioned in Section 1.5, 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 [10], that for real data analysis, we prefer to use simple models and methods that allow sophisticated theoretical analysis than complicate and computationally intensive methods (as an increasing trend in some other scientific communities).

1.9 Content and some notations

The paper is organized as follows. In Section 2, we present the main theoretic results, where we show IF-PCA is consistent in clustering under some regularity conditions, and derive tight tail probability bounds for the KS statistic, both for useless features and for useful features. The main results are proved in Section 2.6. Section 3 contains the numeric studies, and Section 4 discusses connection to other work and addresses some future research. Section 5 contains the proofs for all secondary theorems and lemmas.

In this paper, if A is a vector, then $\|A\|$ denotes the ℓ^2-norm, and if A is a matrix, then $\|A\|$ denotes the matrix spectral norm and $\|A\|_F$ denotes the matrix Frobenius norm. For any positive semi-definite matrix A with rank r, if all nonzero eigenvalues values are simple, denoted by $\lambda_1(A) > \lambda_2(A) > \lambda_r(A) > 0$, then the eigen-spacing of A is defined by $eigsp(A) = \min\{\lambda_r(A), \min_{1\leq k\leq r-1}\{\lambda_k(A) - \lambda_{k+1}(A)\}\}$. When any of nonzero eigenvalue of A is not simple, $eigsp(A) = 0$. L_p denotes a multi-log(p) term, as defined in Section 2.

2 Main results

This section is organized as follows. In Section 2.1, we introduce the theoretic framework and some necessary notations. In Section 2.2, we discuss the regularity conditions required for clustering consistency by IF-PCA and then state the main results. Section 2.3 presents two corollaries (which can viewed as the simplified versions of the main theorem) and interpretations of the main results. Section 2.4 discusses the tail probability of KS statistics and Section 2.5 contains eigen-analysis of the post-selection data matrix; these two sections can be thought of as preparations for the proofs of the main results, but the results there are of interest for their own sake. The main theorems and corollaries are proved Section 2.6.

2.1 The Asymptotic Clustering Model and some notations

Recall that the model we consider is equivalent to

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

where Λ is 'nearly' the $p \times p$ identity matrix, M is the $K \times p$ matrix consisting of the normalized contrast mean parameters $m_k(j) = \mu_k(j)/\sigma(j)$ (M is non-stochastic throughout this paper), and R is the remainder term in (1.9).

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 $p \to \infty$, $p \gg n \gg 1$. Denote the set of useful features by

$$S_p = S_p(M) = \{1 \leq j \leq p : m_k(j) \neq 0 \text{ for some } 1 \leq k \leq K\}, \tag{2.2}$$

and let

$$s_p = s_p(M) = |S_p(M)| \tag{2.3}$$

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), (1.4), and (1.6) the Asymptotic Clustering Model if (2.1) and (2.4) hold, and denote it by $ACM(\vartheta, \theta)$.*

We need some notations. Recall that δ_k is the fraction of samples in class k, $1 \leq k \leq K$. Introduce two $p \times 1$ vectors $\kappa = (\kappa(1), \kappa(2), \ldots, \kappa(p))'$ and $\tau = (\tau(1), \tau(2), \ldots, \tau(p))'$ by

$$\kappa(j) = \kappa(j; M, p, n_p) = \big(\sum_{k=1}^{K} \delta_k m_k^2(j)\big)^{1/2}, \tag{2.5}$$

and

$$\tau(j) = \tau(j; M, p, n_p) = \frac{1}{6\sqrt{2\pi}} \cdot \sqrt{n_p} \cdot \big|\sum_{k=1}^{K} \delta_k m_k^3(j)\big|, \tag{2.6}$$

and introduce

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

Note that $\tau(j)$ and $\kappa(j)$ 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. 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. It is proved in Lemma 2.1 that

$$eigsp(LM\Lambda^2 M'L') = n \cdot eigsp(A\Omega A);$$

$eigsp(\cdot)$ denotes the eigen-spacing (e.g., see Section 1.9). Note that $\|A\Omega A\| \leq \|\kappa\|^2$, and that when $\|m_1\|, \cdots \|m_K\|$ have comparable magnitudes, $eigsp(A\Omega A)$ has the same magnitude. In light of this, we introduce the ratio

$$\rho_2(L, M) = \rho_2(L, M; p, n_p) = \|\kappa\|^2 / eigsp(A\Omega A).$$

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 none of these two ratios is unduly large.

Remark. Let B be the $(K-1) \times (K-1)$ matrix such that $B(k,\ell) = \sqrt{\delta_k \delta_\ell} m_k' \Lambda^2 m_\ell$, and ξ be the $(K-1)$-vector $(\sqrt{\delta_1}, \sqrt{\delta_2}, \ldots, \sqrt{\delta_{K-1}})'/\sqrt{\delta_K}$, then the nonzero eigenvalues of $A\Omega A$ are the same as those of $B(I_{K-1} + \xi\xi')$. Following this formula, when $K = 2$, $eigsp(A\Omega A) = [\delta_1/(1-\delta_1)] \cdot (m_1' \Lambda^2 m_1) \sim [\delta_1/(1-\delta_1)]\|m_1\|^2$, $\|\kappa\|^2 = [\delta_1/(1-\delta_1)]\|m\|^2$, and $\rho_2(L,M) \sim 1$.

2.2 Main theorems and clustering consistency by IF-PCA

In this paper, 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:

$$\min_{1 \leq k \leq K} \{\delta_k\} \geq C^{-1}, \qquad \text{and} \qquad \max_{1 \leq j \leq p} \{\sigma(j) + \sigma^{-1}(j)\} \leq C. \tag{2.7}$$

Introduce two quantities

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

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

$$\epsilon(M) \to 0. \tag{2.8}$$

IF-PCA is essentially a two-step procedure, a screening step and a PCA step. Below, we discuss two steps separately, explaining how each of them works and what regularity conditions it requires for success.

We first discuss the screening step. The key is to study the tail behavior of the KS statistic $\psi_{n_p,j}$. For short, let $a_0 > 0$ be the constant

$$a_0 = [(\pi - 2)/(4\pi)]^{1/2}.$$

In Section 2.4, roughly saying, we show that (note that τ is defined in (2.6))

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

For this reason, $\tau(j)$ 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 above also suggests that the problem of feature selection we consider here is very similar to the signal recovery problem with a Stein's normal means model; the latter is more or less well-understood, see [1, 11] for example. As a result, the most interesting range for $\tau(j)$ is $\tau(j) \geq O(\sqrt{\log(p)})$. In fact, if $\tau(j)$ 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

$$\tau_{min} \geq a_0 \cdot \sqrt{2r \log(p)}. \tag{2.9}$$

By the way $\tau(j)$ is defined, the interesting range for $m_k(j)$ is

$$|m_k(j)| \geq O\big((\log(p)/n)^{1/6}\big). \tag{2.10}$$

Similarly, for the threshold t in (1.16) we use for the KS scores, the interesting range of the threshold is $t = O(\sqrt{\log(p)})$. In light of this, we are primarily interested in threshold of the form

$$t_p(q) = a_0 \cdot \sqrt{2q \log(p)}, \qquad \text{where } q > 0 \text{ is a constant.} \tag{2.11}$$

For any threshold $t > 0$, denote the set of retained features by

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

and for any $n \times p$ matrix W, we let

$$W^{\hat{S}_p(t)}$$

be the $n \times p$ matrix formed by replacing all columns of W with the index $j \notin \hat{S}_p(t)$ by the vector of zeros. If we threshold the KS scores at $t_p(q)$, then by (2.9)-(2.11) and similar argument as in signal recovery of Stein's normal means setting, we expect to

- recover most of the useful features, except for a fraction $\leq Cp^{-[(\sqrt{r}-\sqrt{q})_+]^2/K}$,
- retain $\lesssim p^{1-q}$ useless features,
- have a total number of retained features such that

$$|\hat{S}_p(t_p(q))| \leq C[p^{1-\vartheta} + p^{1-q} + \log(p)].$$

These are validated by Theorems 2.3-2.4 below.

We then discuss the PCA step. Recall $W = [LM + Z\Sigma^{-1/2}]\Lambda + R$; R is the reminder term in normalization and is expected to have a negligible effect. Write

$$W^{\hat{S}_p(t_p(q))} = LM\Lambda + [W^{\hat{S}_p(t_p(q))} - LM\Lambda].$$

Let $LM\Lambda = UDV'$ be the SVD as before, where the diagonals of D are arranged descendingly. Denote the $n \times (K-1)$ matrix of the first $(K-1)$ left singular vectors of $W^{\hat{S}_p(t_p(q))}$ by

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

Both U and $\hat{U}^{(t_p(q))}$ are uniquely determined up to a ± 1 factor in each column. Our goal is to find some regularity conditions under which $\hat{U}^{(t_p(q))} \approx U$. By basic algebra, for this to be hold, it is sufficient that

$$eigsp(LM\Lambda^2 M'L') \gg \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|^2. \tag{2.12}$$

Recall that

$$eigsp(LM\Lambda^2 M'L') = \frac{n}{\rho_2(L,M)}\|\kappa\|^2. \tag{2.13}$$

At the same time, write

$$W^{\hat{S}_p(t_p(q))} - LM\Lambda = (I) + (II),$$

where $(I) = L(M^{\hat{S}_p(t_p(q))} - M)\Lambda$ is 'bias' term associated with the useful features that we have missed in feature selection, and $(II) = [Z\Sigma^{-1/2}\Lambda + R]^{\hat{S}_p(t_p(q))}$ is the noise term. It is shown in Lemmas 2.2 and 2.3 that

$$\|(I)\| \leq L_p \cdot \|\kappa\|\sqrt{n_p} \cdot \big[p^{-(1-\vartheta)/2}\sqrt{\rho_1(L,M)} + p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)}\big],$$

and

$$\|(II)\| \leq L_p\big[\sqrt{n_p} + p^{(1-\vartheta\wedge q)/2} + \|\kappa\| \cdot p^{(\vartheta-q)_+/2}\sqrt{\rho_1(L,M)}\big]$$

Comparing these bounds with (2.13), we denote

$$\begin{aligned} err_p = \rho_2(L,M)\bigg(&\frac{1+\sqrt{p^{(1-\vartheta\wedge q)}/n_p}}{\|\kappa\|} + p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)} \\ &+ \sqrt{p^{-(1-\vartheta)} + p^{(\vartheta-q)_+}/n_p} \cdot \sqrt{\rho_1(L,M)}\bigg), \end{aligned}$$

and (2.12) is satisfied if there is a constant $C > 0$ such

$$err_p \leq p^{-C}. \tag{2.14}$$

We are now ready for the main claims in this paper. For any $K > 1$, let

$$\mathcal{H}_{K-1} = \{\text{All } (K-1)\times(K-1) \text{ diagonal matrices with diagonals taking } \pm 1\}.$$

The main results are the following, which is proved in Section 2.6.

Theorem 2.1. *Fix $(\vartheta, \theta) \in (0,1)^2$, and consider $ACM(\vartheta, \theta)$. Suppose the regularity conditions (2.7), (2.8), (2.9) and (2.14) hold, and the threshold in IF-PCA is set as $t = t_p(q)$. Then there is a matrix H in $\mathcal{H}_{K-1}$ such that as $p \to \infty$, with probability at least $1 - o(p^{-2})$, $\|\hat{U}^{(t_p(q))}H - U\|_F \leq L_p err_p$.*

Finally, the class labels are estimated by truncating the matrix $\hat{U}^{(t_p(q))}$ entry-wise and applying k-means. Let $\hat{y}^{IF}_{t_p(q)} = (\hat{y}^{IF}_{t_p(q),1}, \hat{y}^{IF}_{t_p(q),2}, \hat{y}^{IF}_{t_p(q),n_p})'$ be the estimated labels. We measure the clustering error by the Hamming distance

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

where π is any permutation in $\{1, 2, \ldots, K\}$. We have the following theorem.

Theorem 2.2. *Fix $(\vartheta, \theta) \in (0,1)^2$, and consider $ACM(\vartheta, \theta)$. Suppose the regularity conditions (2.7), (2.8), (2.9) and (2.14) hold, and let $t_p = t_p(q)$ and $T_p = \log(p)$ in IF-PCA. As $p \to \infty$,*

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

By Lemma 2.1, all entries of U is bounded by $C/\sqrt{n_p}$ from above. By the choice of T_p and definitions, the truncated matrix $\hat{U}^{(t_p(q))}_*$ satisfies $\|\hat{U}^{(t_p(q))}_* - U\|_F \leq \|\hat{U}^{(t_p(q))} - U\|_F$. Using this and Theorem 2.1, the proof of Theorem 2.2 is basically an exercise of classical theory on k-means algorithm. For example, the theorem can be proved by adapting that of Theorem 2.2 in [24] to the current setting. For this reason, we skip the proof.

2.3 Corollaries, interpretations, and phase transition revisit.

We consider an idealized case that the two ratios $\rho_1(L,M)$ and $\rho_2(L,M)$ are both bounded a multi-$\log(p)$ term. The following corollary can be viewed as a simplified version of Theorem 2.1, the proof of which is elementary calculation and omitted.

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 $1 - o(p^{-2})$,*

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

Recall that the interesting range for those nonzero $m_k(j)$'s is stated in (2.10), i.e., $|m_k(j)| \asymp L_p n_p^{-1/6}$. It follows that $\|\kappa\| \asymp L_p p^{(1-\vartheta)/2} n_p^{-1/6}$; hence, $\|\kappa\|^{-1}p^{(1-\vartheta)/2} \to \infty$. In this range, we have the following corollary, which is proved in Section 2.6.

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

- *If $(1-\vartheta) < \theta/3$, for any $r > 0$, whatever q is chosen, the upper bound of $\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))}H - U\|_F$ in Corollary 2.1 goes to infinity.*
- *If $\theta/3 < (1-\vartheta) < 1 - 2\theta/3$, for any $r > \vartheta - 2\theta/3$, there exists $q \in (0,r)$ such that $\min_{H \in \mathcal{H}_{K-1}} \|\hat{U}^{(t_p(q))}H - U\|_F \to 0$ 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 $q = 1-\theta$,*

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

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

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

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

To appreciate Corollary 2.2, we relate it to the phase transition phenomenon discussed in Section 1.7, where we have identified three phases: "Feature selection is trivial but clustering is impossible", "Clustering and feature selection are possible but non-trivial", and "Clustering is trivial but feature selection is impossible". In Corollary 2.2, $s_p = p^{1-\vartheta}$, $n_p = p^{\theta}$, the three phases mentioned in Section 1.7 are translated to (a) $(1-\vartheta) < \theta/3$ and $n^{-1/6} \ll u_0 \ll p^{-(1-\vartheta)/2}$, (b) $\theta/3 < (1-\vartheta) < 1-2\theta/3$, $u_0 \asymp (\log(p)/n)^{1/6}$, and (c) $(1-\vartheta) > 1-2\theta/3$ and $p^{(\vartheta-\theta)/2} \ll u_0 \ll n^{-1/6}$, where u_0 is as in (1.24). 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. Case (a) addresses the case of very sparse signals, and we need stronger signals than that in Corollary 2.2 for IF-PCA to be successful (that is, those nonzero $|m_k(j)| \gg n^{-1/6}$). Case (c) addresses the case where signals are relatively dense, and PCA is successful without feature selection (i.e., taking $q = 0$).

2.4 Tail probability of KS statistic

In this section, we present tight bounds on the tail probability of $\psi_{n_p,j}$. We discuss both the case where feature j is a useless feature and when it is a useful feature. In the former, the data points associated with feature j, denoted by $\{X_i(j)\}_{i=1}^n$, are *iid* from $N(\bar{\mu}(j), \sigma^2(j))$. In the latter, the data points split into K classes, and those in Class k are *iid* from $N(\bar{\mu}(j) + \mu_k(j), \sigma^2(j))$.

Recall that $a_0 = \sqrt{(\pi-2)/(4\pi)}$. The following lemma addresses the case where j is a useless feature.

Theorem 2.3. *Fix $\theta \in (0,1)$ and let $n_p = p^{\theta}$. Fix $1 \le j \le p$. 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 $t_p/\sqrt{n_p} \to 0$,*

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

We conjecture that $P(\psi_{n_p,j} \ge t_p) \sim 2 \cdot \frac{1}{\sqrt{2}a_0}\exp(-t_p^2/(2a_0^2))$, but this requires a more complicated proof. Theorem 2.3 says that the right tail of KS statistic behaves similar to that of $N(0, a_0^2)$.

We now consider the case where j is a useful feature. We assume some mild regularity conditions: for some $\delta > 0$,

$$\max_{j \in S_p(M)} \Big\{ \frac{\sqrt{n_p}}{\tau(j)} \sum_{k=1}^K \delta_k m_k^4(j) \Big\} \le C p^{-\delta}, \quad \min_{j \in S_p(M), 1 \le k \le K} |m_k(j)| \ge C^{-1} \sqrt{\frac{\log(p)}{n_p}}. \tag{2.15}$$

Note that (2.15) is mild given (2.8) and (2.10). The following theorem is proved in Section 5.

Theorem 2.4. *Fix $\theta \in (0,1)$. Let $n_p = p^{\theta}$, and $\tau(j)$ be as in (2.6), where j is a useful feature. Suppose (2.9) and (2.15) hold, and the threshold t_p is such that $t_p \to \infty$, that $t_p/\sqrt{n_p} \to 0$, and that $\tau(j) \ge (1+C)t_p$ for some constant $C > 0$. Then as $p \to \infty$,*

$$P(\psi_{n_p,j} \le t_p) \le C\Big(K\exp\big(-\frac{1}{2Ka_0^2}(\tau(j) - t_p)^2\big) + p^{-3}\Big).$$

Note that the left tail probability is bounded (up to a constant) by that of $N(\tau(j), Ka_0^2)$.

Remark. In cases where $\tau(j) = o(\sqrt{\log(p)/n})$ (for example, when the distribution is symmetric, $\tau(j) = 0$), the above result still holds except that we have to replace $\tau(j)$ by

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

where $\phi^{(3)}(y)$ is the third derivative of the standard normal density $\phi(y)$.

In the literature, tight bounds of this kind are only available for the case where X_i are iid samples from a normal with known mean and variance (or $N(0,1)$ without loss of generality; in this case, the bound is derived by [32]; also see [39]). The setting considered here is more complicated, and how to derive tight bounds is an interesting but a 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. In the literature, there are two existing approaches, one is due to [15] which approaches the problem by approximating the stochastic process by a Brownian bridge, the other is due to [36] (also [40, 44]) on the local Poisson approximation of the boundary crossing probability. It is argued in [36] that the second approach is more accurate. Our proofs follow the idea in [40, 36].

2.5 Eigen-analysis of the post-selection data matrix

For any $n \times p$ matrix W, recall that $W^{\hat{S}_p(t_p(q))}$ denotes the sub-matrix of W formed by restricting the columns to $\hat{S}_p(t_p(q))$. In this section, for notational simplicity, we misuse the notations slightly, and think of $W^{\hat{S}_p(t_p(q))}$ as an $n \times p$ matrix where the j-th column is the same of W if $j \in \hat{S}_p(t_p(q))$ and is the vector of 0s otherwise. With such notations,

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

We now analyze the three terms on the right hand side separately.

Consider the first term. Recall that L is the $n \times K$ matrix the i-th row of which is e_k' if and only if $i \in$ Class k, $1 \le i \le n, 1 \le k \le K$, and M is the $K \times p$ matrix the k-th row of which is $m_k' = (\Sigma^{-1/2}\mu_k)'$, $1 \le k \le K$. Also, recall that $A = \mathrm{diag}(\sqrt{\delta_1}\|m_1\|, \ldots, \sqrt{\delta_K}\|m_K\|)$ and Ω is the $K \times K$ matrix with $\Omega(k, \ell) = m_k'\Lambda^2 m_\ell/(\|m_k\| \cdot \|m_\ell\|)$, $1 \le k, \ell \le K$. 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

$$A\Omega A = Q_{K,K-1} \cdot \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{K-1}) \cdot Q_{K,K-1}', \tag{2.16}$$

where the k-th column of $Q_{K,K-1}$ is the k-th eigenvector of $A\Omega A$, and let

$$LM\Lambda = U_{n,K-1} D_{K-1,K-1} (V_{p,K-1})' \tag{2.17}$$

be a SVD of $LM\Lambda$. Introduce

$$G_{K,K} = \mathrm{diag}(\sqrt{\delta_1}, \sqrt{\delta_2}, \ldots, \sqrt{\delta_K}). \tag{2.18}$$

The following lemma is proved in Section 5.

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 $H_{K-1,K-1} \in \mathcal{H}_{K-1}$ such that*

$$U_{n,K-1} = n^{-1/2} L_{n,K}[G_{K,K}^{-1} Q_{K,K-1} H_{K-1,K-1}].$$

For the matrix $G_{K,K}^{-1} Q_{K,K-1} H_{K-1,K-1}$, 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, $1 \le k < \ell \le K$.

By Lemma 2.1 and definitions, it follows that

- For any $1 \le i \le n$ and $1 \le k \le K - 1$, the i-th row of $U_{n,K-1}$ equals to the k-th row of $n^{-1/2} D_{K,K} Q_{K,K-1} H_{K-1,K-1}$ if and only if sample i comes from Class k.
- The rows of $U_{n,K-1}$ partition into K different groups such that rows in the same group are the same but rows in different groups are different. This partition coincides with the partition of the n samples into K different classes.

- $U_{n,K-1}$ has K distinct rows and the ℓ^2-norm between each pair of the distinct rows is no less than $2/\sqrt{n}$.

Consider the second term. This is the 'bias' term causes 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 $1 - o(p^{-2})$,*

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

Consider the last term. 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 $1 - o(p^{-2})$,*

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

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

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

2.6 Proofs of the main results

We now show Theorem 2.1 and Corollary 2.1 (proof of Theorem 2.2 and Corollary 2.2 are omitted). Consider Theorem 2.1. Let

$$G = LM\Lambda^2M'L', \qquad \hat{G} = W^{\hat{S}_p(t_p(q))}(W^{\hat{S}_p(t_p(q))})'.$$

By definition, for some $H \in \mathcal{H}_{K-1}$, $\hat{U}^{(t_p(q))}H$ and U contain the first $(K-1)$ leading eigenvectors of G and $\hat{G}$, respectively. Let $\eta_k(A)$ be the k-th leading eigenvector of an arbitrary symmetric matrix A. It follows that

$$\|\hat{U}^{(t_p(q))}H - U\|_F^2 \leq \sum_{k=1}^{K}\big(\min\{\|\eta_k(\hat{G}) + \eta_k(G)\|, \|\eta_k(\hat{G}) - \eta_k(G)\|\}\big)^2. \tag{2.19}$$

The following lemma is proved in [24].

Lemma 2.4. *For any $n \times n$ symmetric matrix A with rank $K-1$, and an $n \times n$ symmetric matrix B with $\|B\| \leq eigsp(A)/2$,*

$$\min\{\|\eta_k(A+B) + \eta_k(A)\|, \|\eta_k(A+B) - \eta_k(A)\|\} \leq 2\sqrt{2}\frac{\|B\|}{eigsp(A)}, \qquad 1 \leq k \leq K-1.$$

Note that $\|\hat{G} - G\| \leq 2\|LM\Lambda\| \cdot \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\| + \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|^2$. From Lemmas 2.2-2.3 and (2.14), $\|LM\Lambda\| \gg \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|$. Therefore,

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

On the other hand, by Lemma 2.1,

$$eigsp(G) = n \cdot eigsp(A\Omega A') = n\|\kappa\|^2/\rho_2(L,M).$$

Applying Lemma 2.4, for $1 \leq k \leq K-1$,

$$\min\{\|\eta_k(\hat{G}) + \eta_k(G)\|, \|\eta_k(\hat{G}) - \eta_k(G)\|\} \leq \frac{4\sqrt{2}}{\|\kappa\|\sqrt{n}} \cdot \rho_2(L,M) \cdot \|W^{\hat{S}_p(t_p(q))} - LM\Lambda\|, \tag{2.20}$$

where by Lemmas 2.2-2.3, the right hand side is equal to $L_p err_p$. The claim then follows by combining (2.19)-(2.20).

Consider Corollary 2.2. For each $j \in S_p(M)$, $\kappa(j) \geq \epsilon(M)$; hence, $\|\kappa\| \geq L_p p^{(1-\vartheta)/2} n_p^{-1/6} = L_p p^{(1-\vartheta)/2-\theta/6}$. The error bound in Corollary 2.1 reduces to

$$L_p p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)} + L_p \begin{cases} p^{-\theta/3+(\vartheta-q)_+/2}, & \theta < 1-\vartheta, \\ p^{\theta/6-(1-\vartheta)/2+(1-\theta-q)_+/2}, & \theta \geq 1-\vartheta. \end{cases} \tag{2.21}$$

Note that (2.21) 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, we 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 $L_p p^{-\theta/3} = L_p n_p^{-1/3}$.

Next, we consider the case $1-\vartheta \leq \theta < 3(1-\vartheta)$. If $r > 1-\theta$, we can take any $q \in (1-\theta, r)$ and the error bound is $o(1)$, noting that $\theta/6 < (1-\vartheta)/2$. If $r \leq 1-\theta$, noting that $(1-\theta-r)/2 < (1-\vartheta)/2-\theta/6$, there exists $q < r$ 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 $L_p p^{\theta/6-(1-\vartheta)/2} = L_p n_p^{1/6} s_p^{-1/2}$.

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), where in the first one, we set the threshold by HCT as in Section 1.5, and in the second one, the threshold is not data-driven. In both versions, we don't use Step 2 of the IF-PCA procedure described in Section 1.3 (which is specifically for microarray data). 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 may use three different types of linkage [22], "complete", "simple", and "average", but in our study, different choices of linkage only have negligible differences, so for reasons of space, we only report the results using the "complete" linkage.

In each experiment, fixing parameters $(K, p, \theta, \vartheta, r, rep)$ and a vector $\delta = (\delta_1, \cdots, \delta_K)$, we let $n_p = p^\theta$ and $\epsilon_p = p^{1-\vartheta}$; n_p is the sample size, ϵ_p is roughly the fraction of useful features, and rep is the number of repetitions in simulations (see below). We generate the class labels $y_1, y_2, \ldots, y_{n_p}$ *iid* from the Multinomial distribution $MN(K, \delta)$, and let L be the $n_p \times K$ matrix such that the i-th row of L equals to e_k' if and only if $y_i = k$, $1 \leq k \leq K$. Fixing $\gamma = (\gamma_1, \gamma_2, \gamma_3)$, densities g_σ and g_μ defined over $(0, \infty)$ and $g_{\bar{\mu}}$ defined over $(-\infty, \infty)$, for each matrix L as above, we generate the data matrix X as follows.

- Generate the $p \times 1$ vector σ by $\sigma(j) \stackrel{iid}{\sim} g_\sigma$, and construct the covariance matrix $\Sigma = \mathrm{diag}(\sigma^2(1), \sigma^2(2), \cdots, \sigma^2(p))$. Generate an $n_p \times p$ matrix Z with rows *iid* from $N(0, \Sigma)$.
- Generate the overall mean vector $\bar{\mu}$ by $\bar{\mu}(j) \stackrel{iid}{\sim} g_{\bar{\mu}}$, $1 \leq j \leq p$.
- Generate the $K \times p$ matrix M as follows: generate $b_1, b_2, \ldots, b_p$ *iid* from Bernoulli(ϵ_p); for each j such that $b_j = 1$, independently, generate $\{\beta_k(j)\}_{k=1}^{K-1}$ *iid* from Bernoulli$(3, \gamma)$, and $\{h_k(j)\}_{k=1}^{K-1}$ *iid* from g_μ; for $1 \leq k \leq K-1$, set μ_k by

$$\mu_k(j) = \left[72\pi \cdot (2r\log(p)) \cdot n_p^{-1} \cdot h_k(j)\right]^{1/6} \cdot b_j \cdot (\beta_k(j) - 1),$$

and let $\mu_K = -\frac{1}{\delta_K}\sum_{k=1}^{K-1}\delta_k\mu_k$. Let the k-th row of M be $(\Sigma^{-1/2}\mu_k)'$.

- Let $X = \mathbf{1}\bar{\mu}' + LM\Sigma^{1/2} + Z$.

Note that r can be viewed as the signal strength parameter, and that $\beta_k(j) - 1$ takes values from $\{-1, 0, 1\}$, so $\mu_k(j)$ can either be negative, 0, or positive. Once we have X, we apply all 6 algorithms, repeat the above 4 steps independently for *rep* times, and report the average clustering errors. The simulation study consists 4 experiments, which we now describe.

Experiment 1. In this experiment, we study the effect of signal strengths over the clustering results. We set $(K, p, \theta, \vartheta, rep) = (2, 4\times 10^4, .6, .7, 100)$, and the vector $(\delta_1, \delta_2) = (1/3, 2/3)$. Denote the uniform distribution with parameters a and b by $U(a, b)$. 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 take the signal strength parameter $r \in \{.20, .35, .50, .65\}$, and investigate two choices of γ, $\gamma = (.5, 0, .5)$ and $\gamma = (.2, 0, .8)$ (in the former, the signs of signals are equally likely to be ± 1). The threshold for IF-PCA(2) is chosen to be $t_p(q) = \sqrt{2\cdot .06\cdot \log(p)}$.

The results are summarized in Table 4 ($eigsp(LM\Lambda)$ are also included for references). The results suggest that IF-PCA outperform the methods of SpecGem, two versions of k-means, and hierarchical clustering, increasingly so when r increase (so the problem of clustering become easier for the signals become stronger). The results also suggest that two versions of IF-PCA have similar performances, with those of IF-PCA(1) being slightly better. On one hand, this suggests that threshold choice by HCT has advantages: it is not only a data driven choice but also yields satisfactory clustering results. On the other hand, this suggests that IF-PCA is relatively insensitive to different choices of the threshold, as long as they are in a certain range.

Somewhat surprisingly, in this experiment, the signs of the signals do not have major effects over the clustering results (as one might have expected): in Table 4, the results in the top and bottom panels are largely similar, despite that the signal signs are equally likely to be ± 1 in the former, but largely unequally likely in the latter.

r	*eigsp*	IF-PCA(1)	IF-PCA(2)	SpecGem	kmeans	kmeans+	Hier
.20	65.3	.397(.11)	.453(.07)	.401(.08)	.455(.03)	.437(.04)	.439(.05)
.35	69.5	.221(.15)	.331(.15)	.353(.12)	.435(.05)	.433(.04)	.428(.04)
.50	72.2	.103(.12)	.202(.17)	.320(.11)	.424(.08)	.432(.04)	.436(.04)
.65	73.0	.052(.08)	.090(.12)	.305(.12)	.414(.08)	.430(.04)	.426(.05)
r	*eigsp*	IF-PCA(1)	IF-PCA(2)	SpecGem	kmeans	kmeans+	Hier
.20	64.5	.405(.10)	.451(.07)	.406(.08)	.458(.03)	.435(.05)	.431(.04)
.35	67.9	.248(.14)	.342(.15)	.375(.10)	.450(.05)	.434(.04)	.432(.05)
.50	70.6	.128(.13)	.206(.18)	.336(.11)	.446(.06)	.435(.05)	.435(.05)
.65	73.4	.056(.07)	.100(.13)	.286(.12)	.414(.08)	.423(.04)	.430(.05)

Table 4: Comparison of average clustering error rates (Experiment 1). Number in the brackets are standard deviations of the error rates. Top: $\gamma = (.5, 0, .5)$. Bottom: $\gamma = (.2, 0, .8)$. Note that *eigsp* is the short hard notation for $eigsp(LM\Lambda)$; same in all tables in this section.

Experiment 2. The experiment contains two sub-experiments, Experiment 2a and 2b. In Experiment 2a, we investigate the effects of sparsity levels over the clustering results. We set $(K, p, \theta, r, rep) = (2, 4\times 10^4, .6, .3, 100)$ (so $n_p = 577$), $\gamma = (.5, 0, .5)$, $(\delta_1, \delta_2) = (1/3, 2/3)$, and $\vartheta \in \{.68, .72, .76, .80\}$; recall that ϑ is the parameter for sparsity levels. For a random variable $X \sim N(u, b^2)$, denote the conditional density of $(X|a_1 \leq X \leq a_2)$ by $\widetilde{TN}(u, b, a_1, a_2)$, where $\widetilde{TN}$ stands for 'Truncated Normal'. We take $g_{\bar{\mu}}$ as $N(0, 1)$, g_μ as $\widetilde{TN}(1, .1, .8, 1.2)$, and g_σ as $\widetilde{TN}(1, .1, .9, 1.1)$. The threshed for IF-PCA(2) is $t_q(p) = \sqrt{2\cdot .047\cdot \log(p)}$. The results are summarized in Table 5, where in all 4 different sparsity levels, two versions of IF-PCA have similar performances, and each of them significantly outperforms the other 4 methods. The results also suggest that the sparsity levels have major effects over $eigsp(LM\Lambda)$ and so the clustering results: as ϑ increase, $eigsp(LM\Lambda)$ decrease sharply, and the clustering problem become visibly more difficult.

ϑ	$eigsp$	IF-PCA(1)	IF-PCA(2)	SpecGem	kmeans	kmeans+	Hier
.68	84.54	.064(.08)	.072(.10)	.200(.13)	.321(.16)	.440(.04)	.439(.04)
.72	67.14	.222(.16)	.219(.16)	.401(.09)	.454(.04)	.443(.04)	.442(.04)
.76	54.34	.362(.14)	.363(.14)	.465(.03)	.475(.02)	.449(.03)	.448(.03)
.80	40.55	.425(.10)	.398(.12)	.476(.02)	.476(.02)	.410(.06)	.422(.06)

Table 5: Comparison of average clustering error rates (Experiment 2a). Numbers in the brackets are the standard deviations of the error rates.

In Experiment 2b, we use the same setting as in Experiment 2a, except for that $g_{\bar{\mu}}$ is $\widetilde{TN}(1, .316, .3, 1.7)$ and g_σ is the point mass at 1. The results are summarized in Table 6, from which a claim similar to that of Experiment 2a can be drawn.

ϑ	$eigsp$	IF-PCA(1)	IF-PCA(2)	SpecGem	kmeans	kmeans+	Hier
.68	82.51	.074(.10)	.085(.12)	.211(.10)	.354(.13)	.434(.04)	.437(.04)
.72	67.36	.208(.10)	.184(.16)	.398(.09)	.456(.04)	.441(.04)	.436(.04)
.76	54.91	.337(.15)	.323(.15)	.454(.05)	.467(.03)	.437(.04)	.438(.04)
.80	42.43	.443(.06)	.430(.09)	.475(.02)	.476(.02)	.436(.04)	.439(.04)

Table 6: Comparison of average clustering error rates (Experiment 2b). Numbers in the brackets are the standard deviations of the error rates.

In Experiment 3, we study IF-PCA(2) only, and investigate how different choices of thresholds affect its performance. We use the same setting as in Experiment 2b, and investigate $t_q(p) = \sqrt{2q\log(p)}$ for 4 different choices of q: $\{.029, .038, .047, .056\}$ (correspondingly, $t_q(p) = .78, .90, 1.00, 1.09$). The results are summarized in Table 7, which suggest that IF-PCA(1) and IF-PCA(2) have comparable performances, and that IF-PCA(2) is relatively insensitive to different choices of the threshold, as long as they fall in a certain range. On the other hand, the best threshold indeed 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).

	Threshold	$\vartheta = .68$	$\vartheta = .72$	$\vartheta = .76$	$\vartheta = .80$
IF-PCA(1)	HCT (stochastic)	.064(.08)	.222(.16)	.362(.14)	.425(.10)
IF-PCA(2)	.78	.047(.06)	.208(.14)	.379(.12)	.446(.07)
	.90	.050(.07)	.198(.14)	.371(.13)	.419(.11)
	1.00	.072(.10)	.219(.16)	.363(.14)	.398(.12)
	1.09	.130(.15)	.285(.17)	.386(.14)	.390(.14)

Table 7: Comparison of average clustering error rates (Experiment 3). Numbers in the brackets are the standard deviations of the error rates.

Experiment 4. In this experiment, we investigate the effects of correlated noise over the clustering results. We generate the data matrix in the same way as before, except for that Z is replaced by ZA, where A is a $p \times p$ matrix satisfying $A(i,j) = 1\{i = j\} + a \cdot 1\{j = i+1\}$, $1 \leq i, j \leq p$. We set $(K, p, \theta, \vartheta, r, rep) = (4, 2 \times 10^4, .5, .6, .7, 100)$ (so $n_p = 141$), $(\delta_1, \delta_2, \delta_3, \delta_4) = (1/4, 1/4, 1/4, 1/4)$, $\gamma = (.3, .05, .65)$, and $a \in \{.1, .2, .3, .4\}$. 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 $\widetilde{TSE}(.1, .9, -\infty, \infty)$, and g_σ as $\widetilde{TSE}(.1, .9, .9, 1.2)$. The threshold for IF-PCA(2) is chosen to be $t_p(q) = \sqrt{2 \cdot .08 \cdot \log(p)}$. The results are summarized in Table 8, which suggest that IF-PCA continues to work in the presence of short-range correlations among the noise: for all 4 choices of correlation parameter a, IF-PCA significantly outperforms the other 4 methods, increasingly so when a decrease (so that correlations among the noise become weaker).

Experiment 5. In this experiment, we study the effects of non-Gaussianity over the clustering results. Let $t_\nu(0)$ be the central t-distribution with $df = \nu$. We generate the data matrix X

a	IF-PCA(1)	IF-PCA(2)	SpecGem	kmeans	kmeans+	Hier
.1	.077(.11)	.107(.11)	.458(.06)	.470(.09)	.657(.03)	.644(.03)
.2	.188(.12)	.270(.12)	.463(.05)	.470(.09)	.653(.03)	.643(.04)
.3	.312(.10)	.402(.08)	.486(.07)	.488(.09)	.657(.03)	.647(.04)
.4	.390(.08)	.427(.05)	.483(.06)	.501(.09)	.659(.03)	.653(.03)

Table 8: Comparison of average clustering error rates (Experiment 4). Numbers in the brackets are the standard deviations of the error rates.

similarly as before, except for (a) the noise matrix Z has iid entries from $t_\nu(0)$, and (b) $LM\Sigma^{1/2}$ is replaced by $[\nu/(\nu-2)]^{1/2}LM$, for some ν to be determined. We set $(K, p, \theta, \vartheta, r, rep) = (4, 2\times 10^4, .5, .55, 1, 100)$, $(\delta_1, \delta_2, \delta_3, \delta_4) = (1/4, 1/4, 1/3, 1/6)$, $\gamma = (.45, .1, .45)$, and $\nu \in \{5, 10, 15, 20\}$. We take $g_{\bar{\mu}}$ to be $N(0,1)$ and g_μ to be the point mass at 1. The threshold for IF-PCA(2) is set as $t_p(q) = \sqrt{2\cdot .03\cdot \log(p)}$. The results are summarized in Table 9. The results suggest that IF-PCA continues to work in the presence of nonGaussianity: IF-PCA outperforms the other 4 clustering methods, and increasingly so when ν increase (so the noise distributions become closer to that of $N(0,1)$).

ν	$eigsp$	IF-PCA(1)	IF-PCA(2)	SpecGem	kmeans	kmeans+	Hier
5	20.4	.158(.10)	.316(.10)	.351(.09)	.402(.08)	.639(.04)	.586(.06)
10	22.4	.052(.07)	.093(.08)	.307(.09)	.353(.11)	.626(.04)	.554(.06)
15	23.5	.041(.06)	.046(.06)	.303(.10)	.335(.12)	.624(.04)	.550(.05)
20	24.4	.037(.05)	.036(.05)	.292(.10)	.333(.11)	.624(.05)	.549(.06)

Table 9: Comparison of average clustering error rates (Experiment 5). Numbers in the brackets are the standard deviations of the 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 has three elements: using the KS statistic for screening, post-selection PCA, and threshold choice by HC.

Screening is a well-known approach in high dimensional analysis. For example, in variable selection, we use marginal screening for dimension reduction [18], and in cancer classification, we use t-screening to adapt Fisher's LDA to modern settings [17, 12, 8]. However, the setting here is very different for the data contains latent variables and the KS screening is more appropriate.

The KS statistic can be viewed as an ominous test or a measure of goodness-of-fit. The methods and theory we developed on the KS statistic can be useful in many other settings, where it is of interest of 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., [13]).

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 spare PCA and low-rank matrix recovery. Intellectually, the PCA approach is connected to SCORE for community detection on social network [24], but is very different.

Threshold choice by HC is a recent innovation, and was first proposed in [12] (see also [19]) 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 [12, 19]. In particular, this paper requires sophisticated post-selection Random Matrix Theory (RMT), which we do not need in [12, 19]. Our study on RMT is connected to [28, 38] but is very different.

In our settings, we assume different features are independent to each other, and it is of interest to extend IF-PCA to settings where features are correlated. In such settings, exploiting the graphic structure in the correlations could help improve the screening efficiency. One possibility is to combine the KS statistic with the recent innovation of Graphlet Screening [27, 31], a recent innovation in variable selection. Graphic Screening is a graph-guided multivariate screening procedure and has advantages over the better known method of marginal screening and the lasso.

5 Proof

We now prove Theorems 2.3-2.4 and Lemmas 2.1-2.3. Note that Theorems 2.1-2.2 and Corollaries 2.1-2.2 are proved in Section 2.6.

5.1 Proof of Theorem 2.3

We use the techniques developed by [36]. For short, write $u = \bar{\mu}(j)$, $\sigma = \sigma(j)$, $n = n_p$, $\bar{X} = \bar{X}(j)$, $X_i = X_i(j)$, $\psi_n = \psi_{n_p,j}$ and $s = \hat{\sigma}(j)$. Under these notations,

$$\psi_n = \sqrt{n} \sup_{-\infty < v < \infty} \{|\Phi((v - \bar{X})/s) - F_n(v)]|\}.$$

Define

$$\psi_n^{\pm} = \sqrt{n} \sup_{-\infty < v < \infty} \{\mp[\Phi((v - \bar{X})/s) - F_n(v)]\}. \tag{5.1}$$

Writing $P(\psi_n \geq t_p) \leq P(\psi_n^- \geq t_p) + P(\psi_n^+ \geq t_p)$, and noting that by symmetry and time reversal, the two terms on the right hand side equal to each other, it follows that

$$P(\psi_n^- \geq t_p) \leq P(\psi_n \geq t_p) \leq 2P(\psi_n^- \geq t_p). \tag{5.2}$$

At the same time, note that ψ_n^- is an ancillary statistic to the parameters (u, σ), so it is independent of the sufficient statistics $(\bar{X}, s^2)$. Therefore,

$$P(\psi_n^- \geq t_p) = P\big(\psi_n^- \geq t_p | \bar{X} = 0, s^2 = 1\big). \tag{5.3}$$

Combining (5.2)-(5.3) and comparing the result with the theorem, all we need to show is

$$P(\psi_n^- \geq t_p | \bar{X} = 0, s^2 = 1) \sim \sqrt{\frac{2\pi}{\pi - 2}} \exp(-\frac{2\pi}{\pi - 2} t_p^2). \tag{5.4}$$

We now show (5.4). Denote for short $q_v = \Phi(v) - t_p/\sqrt{n}$. It follows from (5.1) and basic algebra that

$$P\big(\psi_n^- \geq t_p | \bar{X} = 0, s^2 = 1\big) \equiv P\big(\inf_{-\infty < v < \infty} \{F_n(v) - q_v\} \leq 0 \big| \bar{X} = 0, s^2 = 1\big). \tag{5.5}$$

Introduce the *first boundary crossing time* by

$$\tau = \inf\{v : F_n(v) < q_v\}.$$

Let v_j be the solution of

$$q_v = j/n, \qquad j = 0, 1, \ldots, n-1.$$

Since $F_n(v)$ is a monotone staircase function taking values from $\{0, 1/n, 2/n, \ldots, 1\}$ and q_v is strictly increasing in v, it is seen that $\{\tau < \infty\} = \{v_0, v_1, \ldots, v_{n-1}\}$ and that $F_n(v_j) = j/n$ given $\tau = v_j$. As a result,

$$\begin{aligned} P\big(\inf_{-\infty < v < \infty} \{F_n(v) - q_v\} \leq 0 \big| \bar{X} = 0, s^2 = 1\big) &= \sum_{j=0}^{n-1} P(\tau = v_j, F_n(v_j) = \frac{j}{n} | \bar{X} = 0, s^2 = 1) \\ &= \sum_{j=0}^{n-1} P\big(\tau = v_j \big| F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big) \cdot P\big(F_n(v_j) = \frac{j}{n} \big| \bar{X} = 0, s^2 = 1\big). \end{aligned} \tag{5.6}$$

Introduce

$$g_0(v) = \Phi(v)\Phi(-v) - \phi^2(v)(1 + v^2/2), \qquad g_1(v) = \Phi(-v) + v\phi(v)(1 + v^2/2).$$

The following lemma is proved in the appendix, using results from ([36]) and ([6]).

Lemma 5.1. *With t_p in Theorem 2.3, for each $0 \le j \le n-1$,*

$$P\big(\tau = v_j \big| F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big) \sim \frac{g_1(v_j)}{g_0(v_j)} \frac{t_p}{\sqrt{n}},$$

and that

$$P\big(F_n(v_j) = \frac{j}{n} \big| \bar{X} = 0, s^2 = 1\big) \sim \frac{1}{\sqrt{2\pi n g_0(v_j)}} \exp(-\frac{t_p^2}{2g_0(v_j)}).$$

Combining (5.6) with Lemma 5.1, we have

$$P(\psi_n^- \ge t_p) \sim t_p \Big[\frac{1}{n}\sum_{j=0}^{n-1} \frac{g_1(v_j)}{\sqrt{2\pi g_0(v_j)^3}} \exp(-\frac{t_p^2}{2g_0(v_j)})\Big]. \tag{5.7}$$

Moreover, recall that v_j is the solution of $\Phi(v) = j/n + t_p/\sqrt{n}$, so $(v_{j+1} - v_j)\phi(v_j) \sim 1/n$. Inserting this into (5.7) gives

$$P(\psi_n^- \ge t_p) \sim t_p \sum_{j=0}^{n-1} \frac{g_1(v_j)}{\sqrt{2\pi g_0(v_j)^3}} \exp(-\frac{t_p^2}{2g_0(v_j)})\phi(v_j)(v_{j+1} - v_j).$$

Note that the right hand side can be approximated by a Reimman integral and

$$\sim t_p \int_{v_0}^{v_{n-1}} \frac{g_1(v)}{\sqrt{2\pi} g_0^{3/2}(v)} \exp(-\frac{t_p^2}{2g_0(v)})\phi(v)dv. \tag{5.8}$$

The following lemma is proved in the appendix.

Lemma 5.2. *With t_p in Theorem 2.3,*

$$t_p \int_{v_0}^{v_{n-1}} \frac{g_1(v)}{\sqrt{2\pi} g_0^{3/2}(v)} \exp(-\frac{t_p^2}{2g_0(v)})\phi(v)dv \sim \sqrt{\frac{2\pi}{\pi - 2}} \exp(-\frac{2\pi}{\pi - 2} t_p^2). \tag{5.9}$$

Inserting (5.9) into (5.7)-(5.8) gives (5.4). □

5.1.1 Proof of Lemma 5.1

The following lemma is proved in [6], [44], or [36]. Given n samples $X_1, X_2, \ldots, X_n$ from an exponential family

$$f(x; \theta) = h_0(x) \exp(\theta' x - \eta(\theta)), \qquad x \in R^d, \ \theta \in R^d.$$

Let $\hat{\theta}$ be the Maximum Likelihood Estimator (MLE) for θ. Note that a sufficient statistic for θ is $\frac{1}{n}\sum_{i=1}^n X_i$, and then $\hat{\theta}$ is a function of $\frac{1}{n}\sum_{i=1}^n X_i$. Denote the density function of $\frac{1}{n}\sum_{i=1}^n X_i$ by $f_0^{(n)}$.

Lemma 5.3. $f_0^{(n)}(x) = (1 + o(1)) \cdot (2\pi n)^{-d/2} \cdot |det(\eta''(\hat{\theta}))|^{-1/2} \exp(-n\ell(\hat{\theta}, x))$*, where* $\ell(\hat{\theta}, x) = (\hat{\theta} - \theta)' x - (\eta(\hat{\theta}) - \eta(\theta))$.

For preparations, we need some calculations related to density associated with n samples from a normal distribution. Let $f(x)$ be the density of $N(u, \sigma^2)$. If we let

$$y = x^2, \qquad \alpha = -\frac{1}{2\sigma^2}, \qquad \beta = \frac{u}{\sigma^2}, \tag{5.10}$$

then we have

$$f(x) = \exp\big(\alpha y + \beta x + \frac{\beta^2}{4\alpha} - \frac{1}{2}\log(-\frac{\pi}{\alpha})\big),$$

Let $\hat{\theta}^{(1)}$ be the MLE, which is a function of $(\frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n X_i)$. Applying Lemma 5.3 with $d = 2$, $\hat{\theta} = \hat{\theta}^{(1)}$, $\eta = \eta^{(1)} = -\frac{\beta^2}{4\alpha} + \frac{1}{2}\log(-\frac{\pi}{\alpha})$, and $x = x_0^{(1)} = (1, 0)$ (corresponding to $\frac{1}{n}\sum_{i=1}^n X_i^2 = 1$ and $\frac{1}{n}\sum_{i=1}^n X_i = 0$), we have

$$\ell(\hat{\theta}^{(1)}, x_0^{(1)}) = -1/2 - \alpha - \beta^2/4\alpha + \log(-2\alpha)/2,$$

and so

$$P\big(\bar{X} = 0, s^2 = 1\big) = (1 + o(1)) \cdot \frac{1}{2\pi n\sqrt{2}} \exp([\frac{1}{2} + \alpha + \frac{\beta^2}{4\alpha} - \frac{\log(-2\alpha)}{2}]n). \tag{5.11}$$

Alternatively, writing $v = v_j$ for short, we can embed the above normal density into a *three-parameter exponential* family

$$f_v(x; \alpha, \beta, \delta) = \exp(\alpha x^2 + \beta x + \delta 1\{x > v\} - \eta^{(2)}(\alpha, \beta, \delta)), \tag{5.12}$$

where

$$\eta^{(2)}(\alpha, \beta, \delta) = -\frac{\beta^2}{4\alpha} + \frac{1}{2}\log\big(-\frac{\pi}{\alpha}\big) + \log(\Phi(\sqrt{-2\alpha}v - \frac{\beta}{\sqrt{-2\alpha}}) + e^{\delta}\Phi(-\sqrt{-2\alpha}v + \frac{\beta}{\sqrt{-2\alpha}})\big). \tag{5.13}$$

Note that when $\delta = 0$, $f_v(x; \alpha, \beta, \delta) \equiv f(x)$. We let

$$h_v = h_v(\alpha, \beta) = -\sqrt{-2\alpha}v + \frac{\beta}{\sqrt{-2\alpha}}. \tag{5.14}$$

Denote the MLE for $\theta \equiv (\alpha, \beta, \delta)$ by $\hat{\theta}^{(2)}$. We note that $\hat{\theta}^{(2)}$ is a function of sufficient statistics $(\frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n 1\{X_i > v\})$. Let $x_0^{(2)} = (1, 0, 1 - q_v)$, corresponding to $\frac{1}{n}\sum_{i=1}^n X_i^2 = 1$, $\frac{1}{n}\sum_{i=1}^n X_i = 0$, and $\frac{1}{n}\sum_{i=1}^n 1\{X_i > v\} = 1 - q_v$, and denote

$$(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)) = \hat{\theta}^{(2)}\Big|_{x = x_0^{(2)}}.$$

Then $(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v))$ is the solution of the following equation system:

$$\begin{cases} \beta\Phi(h_v)\Phi(-h_v) - \sqrt{(1 - \beta v)}[q_v - \Phi(-h_v)]\phi(h_v) = 0, \\ \alpha = (\beta v - 1)/2, \\ e^{\delta} = \frac{1 - q_v}{q_v}\Phi(-h_v)/\Phi(h_v). \end{cases} \tag{5.15}$$

Applying Lemma 5.3 with $d = 3$, $\hat{\theta} = \hat{\theta}^{(2)}$, $\eta = \eta^{(2)}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v))$, and $x = x_0^{(2)}$ gives

$$\ell(\hat{\theta}^{(2)}, x_0^{(2)}) = (\alpha_n^*(v) - \alpha) + \delta_n^*(v)(1 - q_v) - \eta^{(2)}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)) + \eta(\alpha, \beta), \tag{5.16}$$

and so

$$P\big(F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big) = \frac{(1 + o(1))}{(2\pi n)^{3/2}} \cdot (\det(H_{\eta^{(2)}}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)))^{-1/2} e^{-n\ell(\hat{\theta}^{(2)}, x_0^{(2)})}, \tag{5.17}$$

where $H_{\eta^{(2)}}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v))$ is the 3×3 Hessian matrix of $\eta^{(2)}(\alpha, \beta, \delta)$, evaluated at the point $(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v))$.

Introduce

$$\mu(v, q_v) = \frac{\exp(\alpha_n^*(v)v^2 + \beta_n^*(v)v - \eta^{(1)}(\alpha_n^*(v), \beta_n^*(v)))}{\Phi(-h_v(\alpha_n^*(v), \beta_n^*(v))) + e^{\delta_n^*(v)}\Phi(h_v(\alpha_n^*(v), \beta_n^*(v)))}, \tag{5.18}$$

and

$$\tilde{\ell}(v) = \alpha_n^*(v) + 1/2 + \delta_n^*(v)(1 - q_v) - \eta^{(2)}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)) + \log(\sqrt{2\pi}). \tag{5.19}$$

The following lemma is proved in the appendix.

Lemma 5.4. *When $t_p/\sqrt{n} \to 0$, we have the following approximations for the functions of MLE estimators,*

$$\tilde{\ell}(v) = \frac{1}{2g_0(v)} t_p^2/n + O(t_p^3/n^{3/2}),$$

$$\phi(v) - \mu(v, q_v) = \phi(v)\frac{g_1(v)}{g_0(v)} t_p/\sqrt{n} + O(t_p^2/n),$$

and

$$\det(H_{\eta^{(2)}}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)) = 2g_0(v) + O(t_p/\sqrt{n}).$$

We now proceed to prove Lemma 5.1. Consider the first claim. Now we show the approximation for $P\big(\tau = v_j \big| F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big)$. By Loader [36, (20), (21) and Lemma B.2],

$$P\big(\tau = v_j \big| F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big) = (\frac{\partial q_v}{\partial v})^{-1}(\frac{\partial q_v}{\partial v} - \mu(v, q_v))\Big|_{v=v_j}(1 + o(1)), \tag{5.20}$$

where $\mu(v, q_v)$ is defined in (5.18). Using Lemma 5.4 and recalling that $q_v = \Phi(v) - t_p/\sqrt{n}$, it follows from definitions and direct calculations that

$$P\big(\tau = v_j \big| F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big) = \frac{g_1(v_j)}{g_0(v_j)} t_p/\sqrt{n}(1 + o(1)),$$

and the claim follows.

Consider the second claim. Write

$$P\big(F_n(v_j) = \frac{j}{n} \big| \bar{X} = 0, s^2 = 1\big) = \frac{P\big(F_n(v_j) = \frac{j}{n}, \bar{X} = 0, s^2 = 1\big)}{P\big(\bar{X} = 0, s^2 = 1\big)}, \tag{5.21}$$

where both the denominator and numerator are thought of as the density function at that point; this is a slight misuse of notations. Inserting (5.11) and (5.17) into (5.21) and note that $g_0(v) > 0$ for all v, then

$$P(F_n(v_j) = \frac{j}{n} | \bar{X} = 0, s^2 = 1) = (1 + o(1))\frac{1}{\sqrt{\pi n}}[\det(H_{\eta^{(2)}}(\alpha_n^*(v_j), \beta_n^*(v_j), \delta_n^*(v_j)))]^{-1/2} e^{-n\tilde{\ell}(v)}. \tag{5.22}$$

Using Lemma 5.4, the right hand side reduces to

$$(1 + o(1)) \cdot \frac{\exp(-t_p^2/2g_0(v_j))}{\sqrt{2\pi n g_0(v_j)}},$$

and the claim follows. □

5.1.2 Proof of Lemma 5.2

Recall that $g_0(v) = \Phi(v)\Phi(-v) - \phi^2(v)(1 + v^2/2)$ and $g_1(v) = \Phi(-v) + v\phi(v)(1 + v^2/2)$. Denote for short

$$h(v) = \frac{g_1(v)\phi(v)}{\sqrt{2\pi} g_0^{3/2}(v)}.$$

What we need to show is

$$t_p \int_{v_0}^{v_{n-1}} h(v) \exp(-\frac{t_p^2}{2g_0(v)}) dv \sim \sqrt{\frac{2\pi}{\pi - 2}} \exp(-\frac{2\pi}{\pi - 2} t_p^2).$$

The following results follow from elementary calculus.

(a). Since $t_p \to \infty$ and $t_p/\sqrt{n} \to 0$, it is seen that $v_{n-1} = (\sqrt{n}/t_p)^{1/2}(1 + o(1))$ and $v_0 = -(\log(\sqrt{n}/t_p))^{1/2}(1 + o(1))$.

(b). Since that for all $v > 0$, $g_0'(v) = \phi(v)[(v+v^3)\phi(v) - (2\Phi(v)-1)] < 0$, and that $\lim_{v\to\infty} g_0(v)/\Phi(-v) \sim 1$, $g_0(v)$ is a symmetric and positive function over $(-\infty, \infty)$. Moreover,

$$\frac{1}{g_0(0)} = \frac{4\pi}{\pi-2}, \qquad \frac{d}{dv}(\frac{1}{g_0(v)})\Big|_{v=0} = 0, \qquad \frac{d^2}{dv^2}(\frac{1}{g_0(v)})\Big|_{v=0} = \frac{8\pi}{(\pi-2)^2},$$

(c). $h(v)$ is a positive function with $h(0) = \sqrt{4\pi}/(\pi-2)^{3/2}$, $h'(0) = 0$, and $|h''(v)| \le C$ for some constant C when $|v| \le 1/2$.

Denote $b_n = t_p^{-5/6}$. Note that $b_n \to 0$, $t_p b_n^{3/2} \to 0$, but $t_p b_n \to \infty$. We write

$$t_p \int_{v_0}^{v_{n-1}} h(v) \exp(-\frac{t_p^2}{2g_0(v)}) dv = I + II + III, \tag{5.23}$$

where

$$I = \int_{|v|\le b_n} t_p h(v) \exp(-\frac{t_p^2}{2g_0(v)}) dv, \qquad II = \int_{b_n \le |v| \le 1} t_p h(v) \exp(-\frac{t_p^2}{2g_0(v)}) dv,$$

and

$$III = \int_{|v|>1, v_0 \le v \le v_{n-1}} t_p h(v) \exp(-\frac{t_p^2}{2g_0(v)}) dv,$$

where in I and II, we have used $|v_0| > 1$ and $|v_{n-1}| > 1$.

Consider I. By elementary calculus, It follows that

$$I = \int_{-b_n}^{b_n} t_p h(v) e^{-\frac{t_p^2}{2g_0(v)}} dv = t_p e^{-\frac{t_p^2}{2g_0(0)}} \int_{-b_n}^{b_n} (h(0) + O(b_n^2)) e^{-\frac{2\pi}{(\pi-2)^2} t_p^2 v^2 + O(t_p^2 b_n^3)} dv.$$

Recall that $b_n = o(1)$ and $t_p^2 b_n^3 = o(1)$, it follows from elementary calculus that

$$I \sim \sqrt{\frac{2\pi}{\pi-2}} \exp(-\frac{2\pi}{\pi-2} t_p^2). \tag{5.24}$$

Consider II. It is seen that $h(v) \le C$ for $b_n \le |v| \le 1$. Recall that $g_0(v)$ is symmetric and monotone on $[0, \infty]$,

$$II \le C t_p \exp(-\frac{t_p^2}{2g_0(b_n)}).$$

Moreover, by the first and second derivative of $1/g_0(v)$ in (b), there is a constant $c_0 > 0$ such that

$$\frac{1}{g_0(b_n)} \ge \frac{1}{g_0(0)} + c_0 b_n^2.$$

Inserting this into

$$II \le \exp(-\frac{2\pi}{\pi-2} t_p^2) \cdot [C t_p \exp(-\frac{c_0}{2} t_p^2 b_n^2] = o(1) \exp(-\frac{2\pi}{\pi-2} t_p^2). \tag{5.25}$$

Consider III. By symmetry and elementary calculus, there is a constant C_1, such that

$$h(v) \le C_1 |v| / \sqrt{g_0(v)}, \qquad |v| \ge 1$$

so

$$h(v) \exp(-\frac{t_p^2}{2g_0(v)}) \le \frac{C_1|v|}{\sqrt{g_0(v)}} \exp(-\frac{t_p^2}{2g_0(v)}), \qquad |v| \ge 1.$$

Note that, first, there is a constant $C > 0$ such that

$$\frac{1}{g_0(v)} \ge \frac{1}{g_0(1)} + C(|v| - 1),$$

and second, for sufficiently large t_p, the function $\sqrt{x}e^{-(t_p^2/2)x}$ is monotonely decreasing in $[1/g_0(1), \infty)$, so

$$\frac{|v|}{\sqrt{g_0(v)}}\exp(-\frac{t_p^2}{2g_0(v)}) \leq |v|\sqrt{\frac{1}{g_0(1)} + C(|v|-1)}\exp(-\frac{t_p^2}{2}(\frac{1}{g_0(1)} + C(|v|-1))).$$

Combining these,

$$III \leq C_1 t_p \int_{|v|\geq 1} |v|\sqrt{\frac{1}{g_0(1)} + C(|v|-1)}\exp(-\frac{t_p^2}{2}(\frac{1}{g_0(1)} + C(|v|-1)))dv \leq Ct_p\exp(-\frac{t_p^2}{2g_0(1)}).$$

Since $1/g_0(1) > 4\pi/(\pi-2)$, it follows that

$$III = o(1)\exp(-\frac{2\pi}{\pi-2}t_p^2). \tag{5.26}$$

Inserting (5.24), (5.25), and (5.26) into (5.23) gives the claim. $\square$

5.1.3 Proof of Lemma 5.4

We need some preparations. Throughout this subsection, $\epsilon_n = t_p/\sqrt{n}$ for short. First, we study $(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v))$, which satisfies the equations

$$\begin{cases} \beta\Phi(h_v)\Phi(-h_v) - \sqrt{(1-\beta v)}[q_v - \Phi(-h_v)]\phi(h_v) = 0, \\ \alpha = (\beta v - 1)/2, \\ e^{\delta} = \frac{1-q_v}{q_v}\Phi(-h_v)/\Phi(h_v). \end{cases} \tag{5.27}$$

We solve for $\beta_n^*(v)$ first. Recall that

$$q_v = \Phi(v) - \epsilon_n, \qquad h_v(\alpha, \beta) = -\sqrt{-2\alpha}v + \beta/\sqrt{-2\alpha}.$$

Inserting this and the second equation in (5.27) into the first equation of (5.27),

$$\beta\Phi(h_v(\beta))\Phi(-h_v(\beta)) - \sqrt{1-\beta v}[\Phi(v) - \epsilon_n - \Phi(-h_v(\beta))]\phi(h_v(\beta)) = 0, \tag{5.28}$$

and

$$h_v(\beta) = \frac{\beta(v^2+1) - v}{\sqrt{1-\beta v}}.$$

It is seen that $\beta_n^*(v) = o(1)$, we expand this in the neighborhood of $\beta = 0$. Denote $h_v = h_v(\beta)|_{\beta=\beta_n^*(v)}$ for short.

$$h_v = -v + \beta_n^*(v)(1+v^2/2) + O((\beta_n^*(v))^2), \quad \Phi(h_v) = \Phi(-v) + \phi(v)\beta_n^*(v)(v^2/2+1) + O((\beta_n^*(v))^2). \tag{5.29}$$

Reorganizing this gives

$$\beta_n^*(v)\Phi(-v)\Phi(v) - \phi(v)(\beta_n^*(v)(v^2/2+1)\phi(v) - \epsilon_n) + O((\beta_n^*(v))^2) = 0,$$

and so

$$\beta_n^*(v) \sim -\frac{\phi(v)}{\Phi(v)\Phi(-v) - (v^2/2+1)\phi^2(v)}\epsilon_n. \tag{5.30}$$

Inserting this back into (5.27) gives

$$\begin{cases} \alpha_n^*(v) = (\beta_n^*(v)v - 1)/2, \\ e^{\delta_n^*(v)} = \frac{1-q_v}{q_v}\frac{\Phi(-h_v(\beta_n^*(v)))}{\Phi(h_v(\beta_n^*(v)))}. \end{cases} \tag{5.31}$$

We now show the results. Consider the first claim. By definitions,

$$\tilde{\ell}(v) = \alpha_n^*(v) + 1/2 + \delta_n^*(v)(1-q_v) - \eta^{(2)}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)) + \log(\sqrt{2\pi}).$$

Combining this with (5.31),

$$\tilde{\ell}(v) = \frac{\beta_n^*(v)v}{2} + \frac{1}{2}\log(1-\beta_n^*(v)v) - \frac{\beta_n^*(v)^2}{2(1-\beta_n^*(v)v)} + (1-q_v)\log\frac{1-q_v}{\Phi(h_v)} + q_v\log\frac{q_v}{\Phi(-h_v)}.$$

Recall that $\beta_n^*(v) = O(\epsilon_n)$. Using Taylor expansion and (5.28),

$$\begin{aligned} 2\tilde{\ell}(v) &= \frac{((v^2/2+1)\phi(v)\beta_n^*(v)-\epsilon_n)^2}{\Phi(-v)\Phi(v)} - (v^2/2+1)\beta_n^*(v)^2 + O(\epsilon_n^3) \\ &= \frac{1}{\Phi(v)\Phi(-v)-\phi^2(v)(1+v^2/2)}\epsilon_n^2 + O(\epsilon_n^3), \end{aligned}$$

and the claim follows by recalling $\epsilon_n = t_p/\sqrt{n}$.

Consider the second claim. Recall that in $\phi(v) - \mu(v, q_v)$,

$$\mu(v, q_v) = \frac{\exp(\alpha_n^*(v)v^2 + \beta_n^*(v)v - \eta^{(1)}(\alpha_n^*(v), \beta_n^*(v)))}{\Phi(-h_v) + e^{\delta_n^*(v)}\Phi(h_v)}, \tag{5.32}$$

where $\eta^{(1)}(\alpha_n^*(v), \beta_n^*(v))) = -\frac{(\beta_n^*(v))^2}{4\alpha_n^*(v)} + \frac{1}{2}\log(-\pi/\alpha_n^*(v))$. Inserting the second equation of (5.31) into (5.32) and noting that the numerator is the density function for normal distribution with parameter $(\alpha = \alpha_n^*(v), \beta = \beta_n^*(v))$ at point v, it gives that

$$\mu(v, q_v)|_{\alpha_n^*(v),\beta_n^*(v),\delta_n^*(v)} = \frac{q_v}{\Phi(-h_v)}\phi(h_v).$$

Combining with (5.29) and (5.28), we have

$$\mu(v, q_v) \quad = \quad \phi(v) - \phi(v)\frac{\Phi(-v) + v\phi(v)(1+v^2/2)}{\Phi(v)\Phi(-v) - \phi^2(v)(1+v^2/2)}\epsilon_n + O(\epsilon_n^2),$$

it follows that

$$\phi(v) - \mu(v, q_v) = \phi(v) - \mu(v, q_v)\phi(v)\frac{\Phi(-v) + v\phi(v)(1+v^2/2)}{\Phi(v)\Phi(-v) - \phi^2(v)(1+v^2/2)}\epsilon_n + O(\epsilon_n^2),$$

and the claim follows by recalling $\epsilon_n = t_p/\sqrt{n}$.

Consider the last claim. Recall that $\det(H_{\eta^{(2)}}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)))$ is the determinant of the 3×3 Hessian matrix of $\eta^{(2)}(\alpha, \beta, \delta)$ evaluated at the point $(\alpha, \beta, \delta) = (\alpha_n^*(v),\ \beta_n^*(v), \delta_n^*(v))$. By definition and direct calculations, the top left entry of the matrix $H_{\eta^{(2)}}(\alpha, \beta, \delta)$ is

$$\begin{aligned} \frac{\partial^2\eta^{(2)}(\alpha,\beta,\delta)}{\partial\alpha^2} &= \frac{\alpha-\beta^2}{2\alpha^3}\frac{(-\beta/2\alpha+v)^2(e^\delta-1)^2\phi^2(h_v(\alpha,\beta))}{2\alpha(\Phi(-h_v(\alpha,\beta))+e^\delta\Phi(h_v(\alpha,\beta)))^2} \\ &\quad + \frac{\phi(h_v(\alpha,\beta))(e^\delta-1)((v-\frac{\beta}{2\alpha})^2(v+\frac{\beta}{2\alpha})+\frac{3\beta}{4\alpha^2}-\frac{v}{2\alpha})}{\sqrt{-2\alpha}(\Phi(-h_v(\alpha,\beta))+e^\delta\Phi(h_v(\alpha,\beta)))}, \end{aligned}$$

which is $2 + O(\epsilon_n)$ when evaluated at the point $(\alpha, \beta, \delta) = (\alpha_n^*(v),\ \beta_n^*(v), \delta_n^*(v))$. By similar calculations,

$$H_{\eta^{(2)}}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v)) = \begin{pmatrix} 2 & 0 & \phi(v)v \\ 0 & 1 & \phi(v) \\ v\phi(v) & \phi(v) & \Phi(v)\Phi(-v) \end{pmatrix} + \epsilon_n \cdot Rem,$$

where Rem is a 3×3 matrix each entry of which is $O(1)$. As a result,

$$\det(H_{\eta^{(2)}}(\alpha_n^*(v), \beta_n^*(v), \delta_n^*(v))) = 2(\Phi(v)\Phi(-v) - (1+v^2/2)\phi^2(v)) + O(\epsilon_n),$$

and the claim follows by recalling $\epsilon_n = t_p/\sqrt{n}$. $\square$

5.2 Proof of Theorem 2.4

For notational simplicity, we fix j and suppress the dependence of j in all notations. In this section, $m_k = m_k(j)$, $\tau = \tau(j)$, $\bar{X} = \bar{X}(j)$, $X_i = X_i(j)$, and $\hat{\sigma} = \hat{\sigma}(j)$, all of them represent a number instead of a vector. Similarly, X denotes the j-th column of the original data matrix, so it is now an $n \times 1$ vector instead of an $n \times p$ matrix. This is a slight misuse of the notation. When the j-th feature is useful, all samples of the j-th feature partition into K different groups, and sample i belongs to group k if and only if $y_i = k$. Let $\bar{X}^{(k)}$ and $\hat{\sigma}_X^{(k)}$ be the sample mean and sample standard deviation of group k. Decompose

$$F_n(t) - \Phi(\frac{t - \bar{X}}{\hat{\sigma}_X}) = (I) + (II),$$

where

$$(I) = F_n(t) - \sum_{k=1}^{K} \delta_k \Phi(\frac{t - \bar{X}^{(k)}}{\hat{\sigma}_X^{(k)}}), \qquad (II) = \sum_{k=1}^{K} \delta_k \Phi(\frac{t - \bar{X}^{(k)}}{\hat{\sigma}_X^{(k)}}) - \Phi(\frac{t - \bar{X}}{\hat{\sigma}_X}).$$

Consider (I). Introduce

$$M_n^{(k)}(t) \equiv \sqrt{n\delta_k}\Big[\frac{1}{n\delta_k} \sum_{i: y_i = k} 1\{X_i < t\} - \Phi(\frac{t - \bar{X}^{(k)}}{\hat{\sigma}^{(k)}})\Big].$$

We can rewrite

$$(I) = \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sqrt{\delta_k} \cdot M_n^{(k)}(t),$$

where we note that $\frac{1}{n\delta_k} \sum_{i:y_i=k} 1\{X_i < t\}$ is the empirical CDF for the observations in group k. As a result, $\sup_{-\infty<t<\infty} |M_n^{(k)}(t)|$ is the KS statistic for group k. It follows from Theorem 2.3 that $P(\sup_{-\infty<t<\infty} |M_n^{(k)}(t)| > \tilde{\eta}_n) \leq 2\sqrt{2\pi/(\pi-2)} \exp(-\frac{2\pi}{\pi-2}\tilde{\eta}_n^2) \cdot [1 + o(1)]$, for any sequence $\tilde{\eta}_n \to \infty$ and $\tilde{\eta}_n/\sqrt{n} \to 0$. We then have

$$\begin{aligned}
P(\sqrt{n} \sup_{-\infty<t<\infty} |(I)| > \tilde{\eta}_n) &\leq P(\sum_{k=1}^{K} \sqrt{\delta_k} \sup_{-\infty<t<\infty} |M_n^{(k)}(t)| > \tilde{\eta}_n) \\
&\leq \sum_{k=1}^{K} P(\sup_{-\infty<t<\infty} |M_n^{(k)}(t)| > \frac{\tilde{\eta}_n}{\sum_{k=1}^{K} \sqrt{\delta_k}}) \\
&\leq \sum_{k=1}^{K} P(\sup_{-\infty<t<\infty} |M_n^{(k)}(t)| > \frac{\tilde{\eta}_n}{\sqrt{K}}) \\
&\leq 2K\sqrt{\frac{2\pi}{\pi-2}} \exp(-\frac{2\pi}{(\pi-2)K}\tilde{\eta}_n^2) \cdot [1 + o(1)], \qquad (5.33)
\end{aligned}$$

where the third inequality follows from $\sum_{k=1}^{K} \sqrt{\delta_k} \leq \sqrt{K}(\sum_{k=1}^{K} \delta_k)^{1/2} = \sqrt{K}$, by Cauchy-Schwartz inequality.

Consider (II). The following lemma is proved below:

Lemma 5.5. *Under conditions of Theorem 2.4, with probability at least* $1 - O(p^{-3})$*,*

$$\Big| \sup_{-\infty<t<\infty} |(II)| - \frac{1}{\sqrt{n}}\tau \Big| \leq C \sum_{k=1}^{K} \delta_k m_k^4.$$

By Lemma 5.5 and (2.15), with probability at least $1 - O(p^{-3})$,

$$\sqrt{n} \sup_{-\infty<t<\infty} |(II)| = \tau[1 + O(p^{-\delta})]. \qquad (5.34)$$

Combining (5.33)-(5.34), when $\tau \geq (1+C)t_p$,

$$
\begin{aligned}
P(\psi_n \leq t_p) &\leq P\big(\sqrt{n} \sup_{-\infty<t<\infty} |(I)| \geq \tau[1+O(p^{-\delta})] - t_p\big) + O(p^{-3}) \\
&\leq 2K\sqrt{\frac{2\pi}{\pi-2}} \exp\Big\{ -\frac{2\pi}{(\pi-2)K}\big(\tau[1+O(p^{-\delta})] - t_p\big)^2 \Big\} + O(p^{-3}) \\
&\leq 2K\sqrt{\frac{2\pi}{\pi-2}} \exp\Big\{ -\frac{2\pi}{(\pi-2)K}\big(\tau - t_p\big)^2 \Big\}[1+o(1)] + O(p^{-3}),
\end{aligned}
$$

where the last inequality follows from that $\tau \leq L_p$ (recall that L_p is a generic multi-log(p) term). This gives the claim. $\square$

5.2.1 Proof of Lemma 5.5

Let $L_n^{(k)}(t) = \frac{t-\bar{X}^{(k)}}{\hat{\sigma}_X^{(k)}} - \frac{t-\bar{X}}{\hat{\sigma}_X}$. We apply Taylor expansion to (II) and obtain

$$
\begin{aligned}
(II) = \phi(\frac{t-\bar{X}}{\hat{\sigma}_X}) \sum_{k=1}^{K} \delta_k L_n^{(k)}(t) + \frac{1}{2}\phi^{(1)}(\frac{t-\bar{X}}{\hat{\sigma}_X}) \sum_{k=1}^{K} \delta_k [L_n^{(k)}(t)]^2 \\
+ \frac{1}{6}\phi^{(2)}(\frac{t-\bar{X}}{\hat{\sigma}_X}) \sum_{k=1}^{K} \delta_k [L_n^{(k)}(t)]^3 + \frac{1}{24}\sum_{k=1}^{K} \phi^{(3)}(\xi_k)\cdot \delta_k [L_n^{(k)}(t)]^4,
\end{aligned}
\tag{5.35}
$$

where $\phi^{(m)}$ denotes the m-th derivative of the standard normal density function ϕ and ξ_k falls between $\frac{t-\bar{X}}{\hat{\sigma}_X}$ and $\frac{t-\bar{X}}{\hat{\sigma}_X} + L_n^{(k)}(t)$.

To simplify (5.35), we rewrite $L_n^{(k)}(t)$ as

$$
L_n^{(k)}(t) = \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X} + \frac{t-\bar{X}}{\hat{\sigma}_X}\Big(\frac{\hat{\sigma}_X}{\hat{\sigma}_X^{(k)}} - 1\Big) + \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\Big(\frac{\hat{\sigma}_X}{\hat{\sigma}_X^{(k)}} - 1\Big).
$$

Furthermore,

$$
\begin{aligned}
\frac{\hat{\sigma}_X}{\hat{\sigma}_X^{(k)}} - 1 &= \frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{2\hat{\sigma}_X^2(2\hat{\sigma}_X + \hat{\sigma}_X^{(k)})}{(\hat{\sigma}_X^{(k)} + \hat{\sigma}_X)^2\hat{\sigma}_X^{(k)}}\Big(\frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2}\Big)^2 \\
&= \frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{3}{2}\Big(\frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2}\Big)^2[1+o(1)],
\end{aligned}
$$

where in the last inequality, we have used the fact that $|\hat{\sigma}_X - \sigma| = o(1)$ and $|\hat{\sigma}_X^{(k)} - \sigma| = o(1)$, which is easily seen from (5.42)-(5.43) below. Together, we have

$$
L_n^{(k)}(t) = \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X} + \frac{t-\bar{X}}{\hat{\sigma}_X}\frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \epsilon_n^{(k)}(t),
\tag{5.36}
$$

where

$$
\sum_{k=1}^{K} \delta_k |\epsilon_n^{(k)}(t)| \leq \sum_{k=1}^{K} \delta_k \frac{|t-\bar{X}| + |\bar{X}-\bar{X}^{(k)}|}{\hat{\sigma}_X} \cdot \frac{3}{2}\Big(\frac{\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2}\Big)^2 \cdot [1+o(1)].
\tag{5.37}
$$

Plugging (5.36) into (5.35) gives

$$
(II) = (II_1) + (II_2) + (II_3) + err,
\tag{5.38}
$$

where

$$(II_1) = \phi(\frac{t-\bar{X}}{\hat{\sigma}_X})\sum_{k=1}^K \delta_k \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X},$$

$$(II_2) = \phi(\frac{t-\bar{X}}{\hat{\sigma}_X})\frac{t-\bar{X}}{\hat{\sigma}_X}\sum_{k=1}^K \delta_k \frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{1}{2}\phi^{(1)}(\frac{t-\bar{X}}{\hat{\sigma}_X})\sum_{k=1}^K \delta_k \frac{(\bar{X}-\bar{X}^{(k)})^2}{\hat{\sigma}_X^2},$$

$$(II_3) = \phi(\frac{t-\bar{X}}{\hat{\sigma}_X})\sum_{k=1}^K \delta_k \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{1}{6}\phi^{(2)}(\frac{t-\bar{X}}{\hat{\sigma}_X})\sum_{k=1}^K \delta_k \Big(\frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\Big)^3$$
$$+ \frac{1}{2}\phi^{(1)}(\frac{t-\bar{X}}{\hat{\sigma}_X})\sum_{k=1}^K 2\delta_k \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\cdot\frac{t-\bar{X}}{\hat{\sigma}_X}\frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2}.$$

Denote $y = (t-\bar{X})/\hat{\sigma}_X$. We show that $(II_1) = (II_2) = 0$, and that

$$(II_3) = [\phi(y) + y\phi^{(1)}(y)]\sum_{k=1}^K \delta_k \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{1}{6}\phi^{(2)}(y)\sum_{k=1}^K \delta_k \Big(\frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\Big)^3. \quad (5.39)$$

The last claim follows by basic algebra, so we only show the first two claims. Consider the first claim. By definition and elementary calculation,

$$\bar{X} = \sum_{k=1}^K \delta_k \bar{X}^{(k)}, \qquad \hat{\sigma}_X^2 = \sum_{k=1}^K \delta_k (\hat{\sigma}_X^{(k)})^2 + \sum_{k=1}^K \delta_k (\bar{X}^{(k)} - \bar{X})^2.$$

In particular, this implies that

$$\sum_{k=1}^K \delta_k\big((\hat{\sigma}_X^{(k)})^2 - \hat{\sigma}_X^2\big) = \sum_{k=1}^K \delta_k (\bar{X}^{(k)} - \bar{X})^2. \quad (5.40)$$

It follows that $(II_1) = \hat{\sigma}_X^{-1}\phi(y)[\bar{X} - \sum_{k=1}^K \delta_k \bar{X}^{(k)}] = \hat{\sigma}_X^{-1}\phi(y)\cdot 0 = 0$, and the first claim follows. At the same time, using (5.40),

$$(II_2) = y\phi(y)\sum_{k=1}^K \delta_k \frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} + \frac{1}{2}\phi^{(1)}(y)\sum_{k=1}^K \delta_k \frac{(\bar{X}-\bar{X}^{(k)})^2}{\hat{\sigma}_X^2}$$
$$= [y\phi(y) + \phi^{(1)}(y)]\sum_{k=1}^K \delta_k \frac{(\bar{X}-\bar{X}^{(k)})^2}{2\hat{\sigma}_X^2},$$

where the right hand side is 0 as $y\phi(y) + \phi^{(1)}(y) = 0$ for any y. This shows the second claim.

Combining the above with (5.38) gives

$$(II) = (II_3) + err, \quad (5.41)$$

where (II_3) is given by (5.39).

We now study (II_3). By basics of the normal distribution, with probability at least $1-O(p^{-3})$, $|\bar{X}^{(k)} - \mu_k - \bar{\mu}| \le C\sigma\sqrt{\log(p)/n}$ and $|(\hat{\sigma}_X^{(k)})^2 - \sigma^2| \le C\sigma^2\sqrt{\log(p)/n}$, for all $1 \le k \le K$. It follows that $|\bar{X} - \bar{\mu}| \le C\sigma\sqrt{\log(p)/n}$ and

$$\bar{X}^{(k)} - \bar{X} = \mu_k + \sigma\cdot O\big(\sqrt{\log(p)/n}\big). \quad (5.42)$$

In addition,

$$(\hat{\sigma}_X^{(k)})^2 - \hat{\sigma}_X^2 = \sum_{\ell\neq k}\delta_\ell\big((\hat{\sigma}_X^{(k)})^2 - (\hat{\sigma}_X^{(\ell)})^2\big) - \sum_{\ell=1}^K \delta_\ell(\bar{X} - \bar{X}^{(\ell)})^2 = -\sum_{\ell=1}^K \delta_\ell \mu_\ell^2 + err^{(k)},$$

where $|err^{(k)}| \leq C\sigma^2\sqrt{\log(p)/n}+\sum_{\ell=1}^K \delta_\ell|\mu_\ell||\bar{X}^{(\ell)}-\bar{X}-\mu_\ell| \leq C\sigma^2(1+\sum_{\ell=1}^K \delta_\ell|m_\ell|)\sqrt{\log(p)/n}$. Noting that $\max_\ell |m_\ell| \to 0$, we have $\sum_{\ell=1}^K \delta_\ell|m_\ell| = o(1)$. As a result,

$$\hat{\sigma}_X^2 - (\hat{\sigma}_X^{(k)})^2 = \sum_{\ell=1}^K \delta_\ell \mu_\ell^2 + \sigma^2 \cdot O\big(\sqrt{\log(p)/n}\big). \tag{5.43}$$

From (5.42)-(5.43),

$$\sum_{k=1}^K \delta_k \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X} \frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} = -\sum_{k=1}^K \delta_k\mu_k \frac{\sum_{\ell=1}^K \delta_\ell \mu_\ell^2}{2\hat{\sigma}_X^3} + err_1.$$

Since $\sum_{k=1}^K \delta_k\mu_k = 0$, the first term is equal to 0. In addition, $|err_1| \leq C\sigma\sqrt{\log(p)/n}\frac{\sum_{\ell=1}^K \delta_\ell\mu_\ell^2}{2\hat{\sigma}_X^3} + C\sum_{k=1}^K \delta_k|\mu_k|\frac{\sigma^2\sqrt{\log(p)/n}}{2\hat{\sigma}_X^3}$. Since $\sigma^{-1}\max_\ell|\mu_\ell| = \max_\ell|m_\ell| \to 0$, the second term in $|err_1|$ dominates. Therefore, we have

$$\sum_{k=1}^K \delta_k \frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X} \frac{\hat{\sigma}_X^2-(\hat{\sigma}_X^{(k)})^2}{2\hat{\sigma}_X^2} = O\big(\sqrt{\log(p)/n}\big) \cdot \frac{\sigma^3\sum_{k=1}^K \delta_k|m_k|}{\hat{\sigma}_X^3}. \tag{5.44}$$

Similarly, from (5.42),

$$\sum_{k=1}^K \delta_k \Big(\frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\Big)^3 = \sum_{k=1}^K \delta_k \frac{\mu_k^3}{\hat{\sigma}_X^3} + err_2,$$

where $|err_2| \leq C\sum_{k=1}^K \delta_k|\mu_k|\cdot\sigma^2\log(p)/n + C\sum_{k=1}^K \delta_k\mu_k^2\cdot\sigma\sqrt{\log(p)/n}$. Noting that $\sigma^{-1}\min_\ell|\mu_\ell| = \min_\ell|m_\ell| \geq C\sqrt{\log(p)/n}$, the second term in $|err_2|$ dominates. So we have

$$\sum_{k=1}^K \delta_k \Big(\frac{\bar{X}-\bar{X}^{(k)}}{\hat{\sigma}_X}\Big)^3 = \frac{\sum_{k=1}^K \delta_k\mu_k^3}{\hat{\sigma}_X^3} + O\big(\sqrt{\log(p)/n}\big) \cdot \frac{\sigma^3\sum_{k=1}^K \delta_k m_k^2}{\hat{\sigma}_X^3} \tag{5.45}$$

Plugging (5.44)-(5.45) into (5.39), we obtain

$$(II_3) = \frac{\sigma^3}{\hat{\sigma}_X^3} \cdot \Big[\frac{1}{6}\phi^{(2)}(y)\sum_{k=1}^K \delta_k m_k^3 + O\big(\sqrt{\log(p)/n}\big)\sum_{k=1}^K \delta_k|m_k|\Big]. \tag{5.46}$$

By our assumptions, $\sqrt{\log(p)/n} \leq C|\sum_{k=1}^K \delta_k m_k^3| \leq C\sum_{k=1}^K \delta_k|m_k|^3$. Combining this with Cauchy-Schwartz inequality, $\sqrt{\log(p)/n}\cdot\sum_{k=1}^K \delta_k|m_k| \leq C(\sum_{k=1}^K \delta_k|m_k|)\cdot(\sum_{k=1}^K \delta_k|m_k|^3) \leq C(\sum_{k=1}^K \delta_k m_k^2)^{1/2}\sqrt{(\sum_{k=1}^K \delta_k m_k^2)(\sum_{k=1}^K \delta_k m_k^4)}$, which is $C(\sum_{k=1}^K \delta_k m_k^2)\sqrt{\sum_{k=1}^K \delta_k m_k^4} \leq C\sum_{k=1}^K \delta_k m_k^4$, again by Cauchy-Schwartz inequality. Combining this with (5.46),

$$(II_3) = \frac{\sigma^3}{\hat{\sigma}_X^3} \cdot \Big[\frac{1}{6}\phi^{(2)}(y)\sum_{k=1}^K \delta_k m_k^3 + O\big(\sum_{k=1}^K \delta_k m_k^4\big)\Big]. \tag{5.47}$$

Note that $|\sigma^3/\hat{\sigma}_X^3 - 1| \leq C|\hat{\sigma}-\sigma|/\sigma$. From (5.43), $|\hat{\sigma}-\sigma| \leq C\sigma\big(\sum_{k=1}^K \delta_k m_k^2 + \sqrt{\log(p)/n}\big)$, where $\sum_{k=1}^K \delta_k m_k^2 \cdot \sum_{k=1}^K \delta_k|m_k|^3 \ll (\sum_{k=1}^K \delta_k m_k^2)^2 \leq \sum_{k=1}^K \delta_k m_k^4$ and $\sqrt{\log(p)/n}\sum_{k=1}^K \delta_k|m_k|^3 \leq C\sum_{k=1}^K \delta_k m_k^4$ as $\min_k|m_k| \geq C\sqrt{\log(p)/n}$. Therefore,

$$|\sigma^3/\hat{\sigma}_X^3 - 1| \cdot \sum_{k=1}^K \delta_k|m_k|^3 \leq C\sum_{k=1}^K \delta_k m_k^4. \tag{5.48}$$

Combining (5.47)-(5.48), with probability at least $1-O(p^{-3})$,

$$(II_3) = \frac{1}{6}\phi^{(2)}(y)\sum_{k=1}^K \delta_k m_k^3 + O\big(\sum_{k=1}^K \delta_k m_k^4\big). \tag{5.49}$$

Now, we bound err. It has three parts: (i) the last term in the Taylor expansion (5.35), (ii) those terms related to $\epsilon_n^{(k)}(t)$ in (5.36), and (iii) those terms included in the first three terms of the Taylor expansion but excluded from (II_1)-(II_3). First, consider (i). From (5.36) and (5.42)-(5.43), $|L_n^{(k)}(t)| \leq C\sigma(|m_k| + |y|\sqrt{\log(p)/n} + |y|\sum_{k=1}^K \delta_k m_k^2)$. Noting that $\min_k |m_k| \geq C\sqrt{\log(p)/n}$ and $(1+|y|)^4\phi^{(3)}(y)$ are uniformly bounded, we have

$$\frac{1}{24}\sum_{k=1}^K \phi^{(3)}(\xi_k)\cdot\delta_k[L_n^{(k)}(t)]^4 \leq C\Big[\sum_{k=1}^K \delta_k m_k^4 + \big(\sum_{k=1}^K \delta_k m_k^2\big)^4\Big] \leq C\sum_{k=1}^K \delta_k m_k^4.$$

Second, consider (ii). Those terms related to $\epsilon_n^{(k)}(t)$ will not exceed $\phi(y)\sum_{k=1}^K \delta_k|\epsilon_n^{(k)}(t)|$. Combining (5.37) and (5.42)-(5.43),

$$\sum_{k=1}^K \delta_k|\epsilon_n^{(k)}(t)| \leq C\big[\sum_{k=1}^K \delta_k m_k^2 + \sqrt{\log(p)/n}\big]^2 \leq C\Big[\big(\sum_{k=1}^K \delta_k m_k^2\big)^2 + \log(p)/n\Big] \leq C\sum_{k=1}^K \delta_k m_k^4,$$

where the last inequality comes from $\log(p)/n \leq (\sum_{k=1}^K \delta_k|m_k|)^2 \leq \sum_{k=1}^K \delta_k m_k^2$ and $(\sum_{k=1}^K \delta_k m_k^2)^2 \leq \sum_{k=1}^K \delta_k m_k^4$. Last, consider (iii). We can easily figure out that the dominating terms are the following:

$$\begin{aligned}
\sum_{k=1}^K \delta_k\Big(\frac{\hat\sigma_X^2 - (\hat\sigma_X^{(k)})^2}{2\hat\sigma_X^2}\Big)^2 &= \frac{\sigma^4}{\hat\sigma_X^4}\Big[\frac{1}{4}\big(\sum_{k=1}^K \delta_k m_k^2\big)^2 + O\big(\sqrt{\log(p)/n}\big)\cdot\sum_{k=1}^K \delta_k m_k^2\Big]\\
\sum_{k=1}^K \delta_k\Big(\frac{\bar X - \bar X^{(k)}}{\hat\sigma_X}\Big)^4 &= \frac{\sigma^4}{\hat\sigma_X^4}\Big[\sum_{k=1}^K \delta_k m_k^4 + O\big(\sqrt{\log(p)/n}\big)\cdot\sum_{k=1}^K \delta_k|m_k|^3\Big],\\
\sum_{k=1}^K \delta_k\frac{\hat\sigma_X^2 - (\hat\sigma_X^{(k)})^2}{2\hat\sigma_X^2}\Big(\frac{\bar X - \bar X^{(k)}}{\hat\sigma_X}\Big)^2 &= \frac{\sigma^4}{\hat\sigma_X^4}\Big[\frac{1}{2}\big(\sum_{k=1}^K \delta_k m_k^2\big)^2 + O\big(\sqrt{\log(p)/n}\big)\cdot\sum_{k=1}^K \delta_k m_k^2\Big].
\end{aligned}$$

Since $\min_k|m_k| \geq C\sqrt{\log(p)/n}$, $\sqrt{\log(p)/n}\sum_{k=1}^K \delta_k|m_k|^3 \leq \sum_{k=1}^K \delta_k m_k^4$. By Cauchy-Schwartz inequality, $(\sum_{k=1}^K \delta_k m_k^2)^2 \leq \sum_{k=1}^4 \delta_k m_k^4$. Moreover, $\sqrt{\log(p)/n}\sum_{k=1}^K \delta_k m_k^2 \ll \sqrt{\log(p)/n}\sum_{k=1}^K \delta_k|m_k| \leq C\sum_{k=1}^K \delta_k m_k^4$, as we have seen in deriving (5.49). So these terms are bounded by $C\frac{\sigma^4}{\hat\sigma_X^4}\sum_{k=1}^K \delta_k m_k^4$, where $\hat\sigma = \sigma[1+o(1)]$. Combining the results for (i)-(iii), with probability $1 - O(p^{-3})$,

$$|err| \leq C\sum_{k=1}^K \delta_k m_k^4. \tag{5.50}$$

Now, inserting (5.49) and (5.50) into (5.41) gives that with probability $1 - O(p^{-3})$,

$$(II) = \frac{1}{6}\phi^{(2)}(y)\sum_{k=1}^K \delta_k m_k^3 + O\big(\sum_{k=1}^K \delta_k m_k^4\big), \qquad y = (t - \bar X)/\hat\sigma_X. \tag{5.51}$$

Note that $\sum_{k=1}^K \delta_k m_k^3$ does not depend on t, and that $O(\sum_{k=1}^K \delta_k m_k^4)$ represents a term that $\leq C\sum_{k=1}^K \delta_k m_k^4$ in magnitude, where C does not depend on $(t, \bar X, \hat\sigma_X)$; this is because $y^a\phi^{(b)}(y)$ are always uniformly bounded for all y and any fixed integers $a, b \geq 0$. It follows that with probability $1 - O(p^{-3})$,

$$\sup_{-\infty<t<\infty}|(II)| = \sup_{-\infty<t<\infty}\big\{\frac{1}{6}|\phi^{(2)}(y)|\big\}\cdot\big|\sum_{k=1}^K \delta_k m_k^3\big| + O\big(\sum_{k=1}^K \delta_k m_k^4\big).$$

By elementary calculus, $\sup_{-\infty<t<\infty}\{\frac{1}{6}|\phi^{(2)}(y)|\} = \frac{1}{6\sqrt{2\pi}}$. Recalling $\tau = \frac{1}{6\sqrt{2\pi}}\sqrt{n}|\sum_{k=1}^K \delta_k m_k^3|$ gives the claim. $\square$

5.3 Proof of Lemma 2.1

For simplicity, we drop the subscripts of the matrices as long as there is no confusion. Consider the first two claims. By basic algebra, for any two SVDs of $LM\Lambda$,

$$LM\Lambda = UDV' = \widetilde{U}\widetilde{D}\widetilde{V}',$$

if we require the diagonal entries of D and $\widetilde{D}$ to be arranged in the descending order, then there is a matrix $H \in \mathcal{H}_{K-1}$ such that

$$D = \widetilde{D}, \qquad U = \widetilde{U}H. \tag{5.52}$$

At the same time, we write $A\Omega A = (GM\Lambda)(GM\Lambda)'$, where $G = G_{K,K}$ is as in (2.18). For any SVD of $GM\Lambda$, say, $GM\Lambda = Q^*D^*(V^*)'$, by the way Q is defined, there is a matrix $H \in \mathcal{H}_{K-1}$, $Q = Q^*H$. This says

$$GM\Lambda = QHD^*(V^*)' \tag{5.53}$$

Writing $LM\Lambda = (n^{-1/2}LG^{-1})(\sqrt{n}GM\Lambda)$ and using (5.53),

$$LM\Lambda = (n^{-1/2}LG^{-1})QH\mathrm{diag}(\sqrt{n\lambda_1}, \ldots, \sqrt{n\lambda_{K-1}})(V^*)'. \tag{5.54}$$

Direct calculations show that $(n^{-1/2}LG^{-1}QH)'(n^{-1/2}LG^{-1}QH) = I_{K-1}$. As a result, the right hand side of (5.52) is a SVD of LM, and $\sqrt{n\lambda_1}, \ldots, \sqrt{n\lambda_{K-1}}$ are all the nonzero singular values of $LM\Lambda$. Moreover, by (5.54), there is an $H \in \mathcal{H}_{K-1}$ such that

$$U_{n,K-1} = n^{-1/2}LG^{-1}QH, \qquad Q = Q_{K,K-1}.$$

These prove the first two claims.

Consider the last two claims. Denote ξ by the unit-norm $K \times 1$ vector $(\sqrt{\delta_1}, \ldots, \sqrt{\delta_K})'$ and denote $\widetilde{Q}$ by the $K \times K$ matrix

$$\widetilde{Q} = [Q, \xi].$$

On one hand, by basics on SVD, for any $1 \le k \le K-1$, the k-th column of Q is an eigenvector of the matrix $GM\Lambda^2M'G'$, with λ_k being the associated eigenvalue. On the other hand, recall that $G = \mathrm{diag}(\sqrt{\delta_1}, \ldots, \sqrt{\delta_K})$ and that the k-th row of $M\Lambda$ is $\mu_k'\widetilde{\Sigma}^{-1/2}$,

$$\xi'GM\Lambda = (\sum_{k=1}^{K}\delta_k\mu_k')\widetilde{\Sigma}^{-1/2} = 0,$$

where in the last equality, we have used $\delta_k\mu_k = 0$; see (1.5). It follows that ξ is an eigenvector of $GM\Lambda^2M'G'$, with 0 being the associated eigenvalue. Combining these and noting that $\lambda_1 > \lambda_2 > \ldots \lambda_{K-1} > 0$, it follows from basic algebra that $\widetilde{Q}$ is an orthogonal matrix.

Now, for any $1 \le k \le K$, denote the k-th row of Q by q_k'. Since $\widetilde{Q}$ is orthogonal, $\|q_k\|^2 = (1-\delta_k)$, and the ℓ^2-norm of the k-th row of GQH is

$$\|\delta_k^{-1/2}q_k\| = \delta_k^{-1/2}\|q_k\| = (\delta_k^{-1} - 1)^{1/2}.$$

Moreover, for any $1 \le \ell \le K$ and $k \ne \ell$, again by the orthogonality of $\widetilde{Q}$, $q_k'q_\ell = -\sqrt{\delta_k\delta_\ell}$. The ℓ^2-distance between the k-th row and the ℓ-th row of GQH is then

$$\|\delta_k^{-1/2}q_k - \delta^{-1/2}q_\ell\| = \big[\frac{1}{\delta_k}(1-\delta_k) + \frac{1}{\delta_\ell}(1-\delta_\ell) - \frac{2}{\sqrt{\delta_k\delta_\ell}}(-\sqrt{\delta_k\delta_\ell})\big]^{1/2} = \big[\frac{1}{\delta_k} + \frac{1}{\delta_\ell}\big]^{1/2},$$

where the right hand side $\ge 2/\sqrt{(\delta_k + \delta_\ell)} \ge 2$. Combining these give the claims. □

5.4 Proof of Lemma 2.2

Write $\hat{S} = \hat{S}_p(t_p(q))$ and $t_p = t_p(q)$ for short. Note that

$$\|L(M - M^{\hat{S}})\Lambda\| \leq \|L(M - M^{\hat{S}})\| \cdot \|\Lambda\| \leq \|L(M - M^{\hat{S}})\|_F \cdot \|\Lambda\|.$$

The j-th coordinate of the diagonal matrix Λ is $\sigma(j)/\sqrt{E[\hat{\sigma}^2(j)]}$. By direct calculation (see (5.66) for details), $E[\hat{\sigma}^2(j)] = \sigma^2(j)[1 + \kappa^2(j)] \geq \sigma^2(j)$. So

$$\|\Lambda\| \leq 1. \tag{5.55}$$

Therefore, it suffices to show that with probability $\geq 1 - o(p^{-2})$,

$$\|L(M - M^{\hat{S}})\|_F \leq C\|\kappa\|\sqrt{n} \cdot \big[p^{-(1-\vartheta)/2}\sqrt{\rho_1(L, M)\log(p)} + p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)}\big]. \tag{5.56}$$

Now, we show (5.56). By simple algebra,

$$\|L(M - M^{\hat{S}})\|_F^2 = n \sum_{j \in S_p(M)} \kappa^2(j) \cdot 1\{\psi_{n_p,j} \leq t_p\} \equiv n \sum_{j \in S_p(M)} R_j. \tag{5.57}$$

Here, R_j's are independent and either $R_j = \kappa^2(j)$ or $R_j = 0$. By Theorem 2.4 and the fact that $\tau(j) \geq \tau_{min} \geq a_0 \cdot \sqrt{2r\log(p)}$,

$$P(R_j \neq 0) \leq C[p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}].$$

It follows that

$$\sum_{j \in S_p(M)} E[R_j] \leq C[p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}] \sum_{j \in S_p(M)} \kappa^2(j) = C[p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}] \cdot \|\kappa\|^2. \tag{5.58}$$

To control $\sum_{j \in S_p(M)}(R_j - E[R_j])$, we use Bennet's lemma (see [39, Page 851]): If $R_j \leq b$ and $\sum_{j \in S_p(M)} \mathrm{Var}(R_j) \leq v$, then for any $x \geq 0$

$$P\Big(\sum_{j \in S_p(M)} (R_j - E[R_j]) \geq x\Big) \leq \begin{cases} \exp\big(-\frac{cx^2}{2v}\big), & xb \leq v, \\ \exp\big(-\frac{cx}{2b}\big) & xb > v, \end{cases}$$

where $c = 2\log(2) - 1 \approx .773$. Taking $x = \sqrt{6v\log(p)/c}$ when $6b^2\log(p) \leq cv$, and $x = 6b\log(p)/c$ when $6b^2\log(p) > cv$, we find that with probability $\geq 1 - O(p^{-3})$

$$\sum_{j \in S_p(M)} (R_j - E[R_j]) \leq C\big[\sqrt{v\log(p)} + b\log(p)\big]. \tag{5.59}$$

Note that $R_j \leq \|\kappa\|_\infty^2$ and $\mathrm{Var}(R_j) \leq \kappa^4(j) \cdot 2P(R_j \neq 0) \leq C\|\kappa\|_\infty^2 [p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}] \cdot \kappa^2(j)$. We take

$$b = \|\kappa\|_\infty^2, \qquad v = C\|\kappa\|_\infty^2 [p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}] \cdot \|\kappa\|^2.$$

It follows that with probability $\geq 1 - O(p^{-3})$,

$$\sum_{j \in S_p(M)} (R_j - E[R_j]) \leq C\Big[\|\kappa\|_\infty^2 \log(p) + \|\kappa\|_\infty \|\kappa\| \sqrt{(p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3})\log(p)}\Big]. \tag{5.60}$$

Combining (5.58) and (5.60) gives

$$\begin{aligned} \sum_{j \in S_p(M)} R_j &\leq C\Big(\|\kappa\|_\infty \sqrt{\log(p)} + \|\kappa\| \sqrt{p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}}\Big)^2 \\ &= C\|\kappa\|^2 \Big(p^{-(1-\vartheta)/2}\sqrt{\rho_1(L, M)\log(p)} + \sqrt{p^{-[(\sqrt{r}-\sqrt{q})_+]^2/K} + p^{-3}}\Big)^2 \\ &\leq C\|\kappa\|^2 \big[p^{-(1-\vartheta)/2}\sqrt{\rho_1(L, M)\log(p)} + p^{-[(\sqrt{r}-\sqrt{q})_+]^2/(2K)}\big]^2, \end{aligned}$$

where the last inequality is due to that $\rho_1(L, M) \geq 1$ and $1 - \vartheta < 3$. Inserting this into (5.57) gives (5.56), and the claim follows. $\square$

5.5 Proof of Lemma 2.3

Before we show Lemma 2.3, we show that with probability at least $1 - O(p^{-3})$,

$$|\hat{S}(t_p(q))| \leq C\big[p^{1-\vartheta} + p^{1-q} + \log(p)\big]. \tag{5.61}$$

Write for short $\hat{S} = \hat{S}(t_p(q))$ and $t_p = t_p(q)$. Noting that $|\hat{S} \cap S_p(M)| \leq |S_p(M)| = p^{1-\vartheta}$, we only need to bound $|\hat{S} \setminus S_p(M)|$. Write

$$|\hat{S} \setminus S_p(M)| = \sum_{j \notin S_p(M)} 1\{\psi_{n_p,j} \geq t_p\}.$$

Note that $\psi_{n_p,j}$'s are independent. In addition, by Theorem 2.3, for $j \notin S_p(M)$,

$$P(\psi_{n_p,j} \geq t_p) \leq Cp^{-q}.$$

So $E[|\hat{S} \setminus S_p(M)|] \leq Cp^{1-q}$. We now apply Bennett's lemma as in (5.59) with $b = 1$ and $v = p \cdot 2Cp^{-q} \cdot (1 - Cp^{-q})$, it follows that with probability at least $1 - O(p^{-3})$,

$$|\hat{S} \setminus S_p(M)| - E[|\hat{S} \setminus S_p(M)|] \leq C\big[\log(p) + p^{(1-q)/2}\sqrt{\log(p)}\big].$$

Combining the above gives $|\hat{S} \setminus S_p(M)| \leq C[p^{1-q} + \log(p)]$, and (5.61) follows immediately.

We now proceed to show Lemma 2.3. Write $m_p = m_p(q) = p^{1-\vartheta} + p^{1-q} + \log(p)$ and let $\mathcal{B} = \mathcal{B}(t_p)$ be the collection of subsets of $\{1, \cdots, p\}$ with size $\leq Cm_p$. Given (5.61), to show the claim, it suffices to show that with probability at least $1 - O(p^{-3})$,

$$\max_{B \in \mathcal{B}} \|(Z\Sigma^{-1/2}\Lambda + R)^B\| \leq C\Big[\sqrt{n} + \sqrt{m_p \log(p)} + \|\kappa\| \cdot p^{-(1-\vartheta)/2}\sqrt{m_p \rho_1(L, M)\log(p)}\Big]. \tag{5.62}$$

But by triangle inequality, to show (5.62), it suffices to show that with probability at least $1 - O(p^{-3})$,

$$\max_{B \in \mathcal{B}} \|(Z\Sigma^{-1/2}\Lambda)^B\| \leq C\Big[\sqrt{n} + \sqrt{m_p \log(p)}\Big], \tag{5.63}$$

and

$$\max_{B \in \mathcal{B}} \|R^B\| \leq C\Big[\sqrt{n} + \sqrt{m_p \log(p)} + \|\kappa\| \cdot p^{-(1-\vartheta)/2}\sqrt{m_p \rho_1(L, M)\log(p)}\Big]. \tag{5.64}$$

First, we consider (5.63). The random matrix $Z\Sigma^{-1/2}$ has *iid* entries with the standard normal distribution. The following lemma is proved in [41, Corollary 5.35].

Lemma 5.6. *Let A be an $N \times n$ matrix whose entries are independent standard normal random variables. Then for every $x \geq 0$, with probability at least $1 - 2\exp(-x^2/2)$, one has*

$$\sqrt{N} - \sqrt{n} - x \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + \sqrt{n} + x,$$

where $s_{\min}(A)$ and $s_{\max}(A)$ are correspondingly minimum and maximum eigenvalue of random matrix A.

We apply Lemma 5.6 to $A = (Z\Sigma^{-1/2})^B$ (as an $n \times |B|$ matrix by removing zero columns) and $x = \sqrt{2(3 + Cm_p)\log(p)}$. Then for each fixed B, with probability $\geq 1 - 2p^{-(3+Cm_p)}$,

$$\|(Z\Sigma^{-1/2})^B\| \leq \sqrt{n} + \sqrt{|B|} + \sqrt{2(3 + Cm_p)\log(p)}.$$

Noting that $|\mathcal{B}| \leq p^{Cm_p}$ and $m_p \geq 1$, we obtain $\max_{B \in \mathcal{B}} \|(Z\Sigma^{-1/2})^B\| \leq \sqrt{n} + C\sqrt{m_p \log(p)}$. In addition, $\|(Z\Sigma^{-1/2}\Lambda)^B\| = \|(Z\Sigma^{-1/2})^B\| \cdot \|\Lambda\| \leq \|(Z\Sigma^{-1/2})^B\|$, where we have used (5.55). It follows that

$$\max_{B \in \mathcal{B}} \|(Z\Sigma^{-1/2}\Lambda)^B\| \leq \sqrt{n} + C\sqrt{m_p \log(p)},$$

and (5.63) follows.

Next, consider (5.64). Let $G = \Sigma^{1/2}(\hat{\Sigma}^{-1/2} - \tilde{\Sigma}^{-1/2})$, and we can write

$$\begin{aligned} R &= (LM + Z\Sigma^{-1/2})\Sigma^{1/2}(\hat{\Sigma}^{-1/2} - \tilde{\Sigma}^{-1/2}) - \mathbf{1}_n(\bar{X} - \bar{\mu})'\hat{\Sigma}^{-1/2} \\ &= (LM + Z\Sigma^{-1/2})G - \mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2}(\Lambda + G). \end{aligned}$$

Therefore, $\max_{B\in\mathcal{B}} \|R^B\|$ does not exceed

$$\|G\|\big(\max_{B\in\mathcal{B}} \|LM^B\| + \max_{B\in\mathcal{B}} \|(Z\Sigma^{-1/2})^B\|\big) + (1 + \|G\|)\max_{B\in\mathcal{B}} \|[\mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2}]^B\|. \tag{5.65}$$

We now bound $\|G\|$, $\max_{B\in\mathcal{B}} \|[\mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2}]^B\|$, and $\max_{B\in\mathcal{B}} \|LM^B\|$ separately; a bound for $\max_{B\in\mathcal{B}} \|(Z\Sigma^{-1/2})^B\|$ is already given in (5.63).

We first bound $\|G\|$. For sample i that belongs to Class k, $X_i(j) = \bar{\mu} + \mu_k(j) + Z_i(j)$; also, $\bar{X}(j) = \bar{\mu} + \bar{Z}(j)$. Therefore,

$$\begin{aligned} \hat{\sigma}^2(j) &= \frac{1}{n}\sum_{i=1}^n \big[X_i(j) - \bar{X}(j)\big] = \frac{1}{n}\sum_{k=1}^K \sum_{y_i=k} [\mu_k(j) + Z_i(j) - \bar{Z}(j)]^2 \\ &= \sum_{k=1}^K \delta_k \mu_k^2(j) + \frac{1}{n}\sum_{i=1}^n [Z_i(j) - \bar{Z}(j)]^2 + \frac{2}{n}\sum_{k=1}^K \sum_{y_i=k} \mu_k(j)[Z_i(j) - \bar{Z}(j)]. \end{aligned} \tag{5.66}$$

The first term is equal to $\sigma^2(j)\kappa^2(j)$. From basic properties of normal distributions, the second term is $\sigma^2(j) + \sigma^2(j)\cdot O(\sqrt{\log(p)/n})$ with probability $1 - O(p^{-4})$. The third term is equal to

$$2\sum_{k=1}^K \delta_k \mu_k(j)[\bar{Z}^{(k)}(j) - \bar{Z}(j)], \qquad \text{where } \bar{Z}^{(k)}(j) = \frac{1}{n\delta_k}\sum_{y_i=k} Z_i(j).$$

With probability $\geq 1 - O(p^{-4})$, $|\bar{Z}^{(k)}(j) - \bar{Z}(j)| \leq C\sigma(j)\sqrt{\log(p)/(n\delta_k)}$. So the absolute value of this term is bounded by $C\sum_{k=1}^K \sqrt{\delta_k}|\mu_k(j)|\cdot\sigma(j)\sqrt{\log(p)/n}$. By Cauchy-Schwartz inequality, $\sum_{k=1}^K \sqrt{\delta_k} \leq \sqrt{K}$, and this term is further bounded by $C\sqrt{K}\cdot\sigma^2(j)\max_{1\leq k\leq K}|m_k(j)|\cdot\sqrt{\log(p)/n} \ll \sigma^2(j)\sqrt{\log(p)/n}$, recalling (2.8). As a result, with probability $\geq 1 - O(p^{-3})$,

$$\max_{1\leq j\leq p}\Big|\frac{\hat{\sigma}^2(j) - \sigma^2(j)[1 + \kappa^2(j)]}{\sigma^2(j)}\Big| \leq Cn^{-1/2}\sqrt{\log(p)}.$$

In addition, $\tilde{\Sigma}(j,j) = E[\hat{\sigma}^2(j)] = \sigma^2(j)[1+\kappa^2(j)]$, hence, the j-th diagonal of Λ is $1/\sqrt{1+\kappa^2(j)}$. It follows that

$$\|G\| = \max_{1\leq j\leq p}\frac{|\hat{\sigma}(j) - \sigma(j)\sqrt{1+\kappa^2(j)}|}{\hat{\sigma}(j)\sqrt{1+\kappa^2(j)}} \leq C\Big|\frac{\hat{\sigma}^2(j) - \sigma^2(j)[1+\kappa^2(j)]}{\sigma^2(j)}\Big| \leq Cn^{-1/2}\sqrt{\log(p)}. \tag{5.67}$$

Second, we bound $\max_{B\in\mathcal{B}} \|[\mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2}]^B\|$. Let $\zeta(j) = \sqrt{n}[\bar{X}(j) - \bar{\mu}(j)]/\sigma(j)$. Noting that for a rank-1 matrix, its maximum singular value is equal to its Frobenius norm, we obtain

$$\|[\mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2}]^B\| = \|(\mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2})^B\|_F = \|\zeta^B\|.$$

Since $\bar{\mu} = \frac{1}{n}\sum_{i=1}^n E[X_i]$, we write

$$\zeta(j) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i(j) - E[X_i(j)]}{\sigma(j)}.$$

Here, $\zeta(j)$'s are *iid* $N(0,1)$ random variables, so $\|\zeta^B\|^2$ has a $\chi^2_{|B|}(0)$ distribution; hence, $E[\|\zeta^B\|] \leq \sqrt{E[\|\zeta^B\|^2]} = \sqrt{|B|}$. Furthermore, by Gaussian concentration theory (see [14, Proposition 2.5]), for any $x \geq 0$,

$$P\big(\|\zeta^B\| - E[\|\zeta^B\|] > x\big) \leq e^{-x^2/2}.$$

Taking $t = \sqrt{2(3 + Cm_p)\log(p)}$, for each fixed B, with probability at least $1 - 2p^{-(3+Cm_p)}$,

$$\|\zeta^B\| \le \sqrt{|B|} + \sqrt{2(3 + Cm_p(q))\log(p)}.$$

Combining the above results and noting that $|B| \le Cm_p$ for $B \in \mathcal{B}$, with probability $\ge 1 - O(p^{-3})$,

$$\max_{B \in \mathcal{B}} \|[\mathbf{1}_n(\bar{X} - \bar{\mu})'\Sigma^{-1/2}]^B\| \le C\sqrt{m_p \log(p)}. \tag{5.68}$$

Last, we bound $\max_{B\in\mathcal{B}} \|LM^B\|$. Note that $\|LM^B\| \le \|LM^B\|_F$. Therefore,

$$\|LM^B\|^2 \le n \sum_{k=1}^n \delta_k \sum_{j \in B} m_k^2(j) = n \sum_{j \in B} \kappa^2(j) \le n|B|\|\kappa\|_\infty^2.$$

By definition, $\|\kappa\|_\infty^2 \le \rho_1(L, M) \cdot p^{-(1-\vartheta)}\|\kappa\|^2$. For $B \in \mathcal{B}$, $|B| \le Cm_p$. Together, we have

$$\max_{B \in \mathcal{B}} \|LM^B\| \le C\|\kappa\|\sqrt{n} \cdot p^{-(1-\vartheta)/2}\sqrt{m_p \rho_1(L, M)}. \tag{5.69}$$

Finally, with all these bounds, inserting (5.67) and (5.68)-(5.69) into (5.65) gives

$$\max_{B \in \mathcal{B}} \|R^B\| \le C\Big[\sqrt{\log(p)} + \sqrt{m_p \log(p)} + \|\kappa\| \cdot p^{-(1-\vartheta)/2}\sqrt{m_p \rho_1(L, M)\log(p)}\Big], \tag{5.70}$$

and (5.64) follows. □

References

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

[2] A. Amini and M. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, 37:2877–2921, 2008.

[3] E. Arias-Castro, E. Candes, and Y. Plan. Global testing under sparse alternative: Anova, multiple comparisons and the higher criticism. *Ann. Statist.*, 39(5):2533–2556, 2011.

[4] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local pca. *arXiv:1301.2007*, 2013.

[5] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[6] A. Borovkov and B. Rogozin. On the multi-dimensional central limit theorem. *Theory Probab. Appl.*, 10(1):55–62, 1965.

[7] J. Chen and P. Li. Hypothesis test for normal mixture models: The em approach. *Ann. Statist.*, 37(5A):2523–2542, 2009.

[8] A. Delaigle, P. Hall, and J. Jin. Robustness and accuracy of methods for high dimensional data analysis based on student's t-statistic. *J. Roy. Statist. Soc. B*, 73(3):283–301, 2011.

[9] M. Dettling. Bagboosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593, 2004.

[10] D. Donoho. Personal communications. 2014.

[11] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, pages 962–994, 2004.

[12] D. Donoho and J. Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci.*, 105(39):14790–14795, 2008.

[13] D. Donoho and J. Jin. Higher criticism for large-scale inference: especially for rare and weak effects. *Manuscript*, 2013.

[14] R. M. Dudley. *Uniform central limit theorems*, volume 23. Cambridge Univ Press, 1999.

[15] J. Durbin. The first-passage density of a continuous gaussian process to a general boundary. *J. Appl. Probab.*, pages 99–122, 1985.

[16] B. Efron. Large-scale simultaneous hypothesis testing. *J. Amer. Statist. Soc.*, 99(465):96–104, 2004.

[17] B. Efron. Empirical bayes estimates for large-scale prediction problems. *J. Amer. Statist. Soc.*, 104(487):1015–1028, 2009.

[18] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. B*, 70(5):849–911, 2008.

[19] Y. Fan, J. Jin, and Z. Yao. Optimal classification in sparse gaussian graphic model. *Ann. Statist.*, 41(5):2537–2571, 2013.

[20] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967, 2002.

[21] P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.*, 38(3):1686–1732, 2010.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2nd edition, 2009.

[23] L. Jager and J. Wellner. Goodness-of-fit tests via phi-divergence. *Ann. Statist.*, 35(5):2018–2053, 2007.

[24] J. Jin. Fast community detection by score. *arXiv:1211.5803*, 2012.

[25] J. Jin and Z. Ke. Rare and weak effects in large-scale inference: methods and phase diagrams. *Manuscript*, 2014.

[26] J. Jin, Z. Ke, and W. Wang. Optimal spectral clustering by higher criticism thresholding. *Manuscript*, 2014.

[27] J. Jin, C.-H. Zhang, and Q. Zhang. Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.*, page To appear, 2012.

[28] I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, pages 295–327, 2001.

[29] I. Johnstone and A. Lu. On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, 104(486), 2009.

[30] S. Jung and J. Marron. Pca consistency in high dimension, low sample size context. *Ann. Statist.*, 37:4104–4130, 2009.

[31] Z. Ke, J. Jin, and J. Fan. Covariance assisted screening and estimation. *Ann. Statist.*, page To appear, 2012.

[32] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4(1):83–91, 1933.

[33] A. B. Lee, D. Luca, and K. Roeder. A spectral graph approach to discovering genetic ancestry. *Ann. Appl. Statist.*, 4(1):179–202, 2010.

[34] J. Lei and V. Vu. Sparsistency and agnostic inference in sparse pca. *arXiv:1401.6978*, 2014.

[35] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982.

[36] C. R. Loader et al. Boundary crossing probabilities for locally poisson processes. *Ann. Appl. Probab.*, 2(1):199–228, 1992.

[37] Z. Ma. Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, 41(2):772–801, 2013.

[38] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.

[39] G. Shorack and J. Wellner. *Empirical processes with applications to statistics*. John Wiley & Sons, 1986.

[40] D. Siegmund. Large deviations for boundary crossing probabilities. *Ann. Prob.*, 10(3):581–588, 1982.

[41] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*, pages 210–268, 2012.

[42] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[43] J. Wellner. Personal communications. 2012.

[44] M. Woodroofe. Large deviations of likelihood ratio statistics with applications to sequential testing. *Ann. Statist.*, pages 72–84, 1978.

[45] M. R. Yousefi, J. Hua, C. Sima, and E. R. Dougherty. Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, 26(1):68–76, 2010.

[46] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J. Comp. Graph. Stat.*, 15(2):265–286, 2006.