CARNEGIE MELLON UNIVERSITY

CLUSTERING PROBLEMS FOR HIGH DIMENSIONAL DATA

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY
In
STATISTICS

by
WANJIE WANG

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

June, 2014

Abstract

We consider a clustering problem where we observe feature vectors $X_i \in R^p$, $i = 1, 2, \ldots, n$, from several possible classes. The class labels are unknown and the main interest is to estimate these labels.

We propose a three-step clustering procedure where we first evaluate the significance of each feature by the Kolmogorov-Smirnov statistic, then we select the small fraction of features for which the Kolmogorov-Smirnov scores exceed a preselected threshold $t > 0$, and then use only the selected features for clustering by one version of the Principal Component Analysis (PCA).

In this procedure, one of the main challenges is how to set the threshold t. We propose a new approach to set the threshold, where the core is the so-called *Signal-to-Noise Ratio (SNR)* in post-selection PCA. SNR is reminiscent of the recent innovation of Higher Criticism; for this reason, we call the proposed threshold the Higher Criticism Threshold (HCT), despite that it is significantly different from the HCT proposed earlier by [Donoho 2008] in the context of classification.

Motivated by many examples in Big Data, we study the spectral clustering with HCT for a model where the signals are both rare and weak for two-classes clustering case. Through delicate PCA, we forge a close link between the HCT and the ideal threshold choice, and show that the HCT yields optimal results in the spectral clustering approach. The approach is successfully applied to three gene microarray data sets, where it compares favorably with existing clustering methods.

Our analysis is subtle and requires new development in the Random Matrix Theory (RMT). One challenge we face is that most results in the RMT can not be applied directly to our case: existing results are usually for matrices with *i.i.d.* entries, but the object of interest in the current case is the post-selection data matrix, where (due to feature selection) the columns are non-independent and have hard-to-track distributions. We develop intricate new RMT to overcome this problem.

We also find the theoretical approximation for the tail distribution of Kolmogorov-Smirnov Statistic under null hypothesis and alternative hypothesis. With the theoretical approximation, we can claim the effectiveness of KS statistic.

Besides, we also find the fundamental limits for clustering problem, signal recovery problem, and detection problem under the Asymptotic Rare and Weak model. We find the boundary such that when the model parameters are beyond the boundary, then the inference is unavailable, otherwise there are some methods (usually exhausted search) to achieve the inference.

Acknowledgements

First and foremost, I want to thank my advisor, Prof. Jiashun Jin for all his support in my education. His enthusiasm and curiosity in our project always inspired me to explore the whole story, and make the final thesis. Prof. Jin also helps a lot in pushing me towards academic perfection. His dedication in academia has always been admired by me and guided me for future path. With his support and advices, I have accomplished many that I could not imagine and gained a deep understanding of my career.

Secondly, I want to express my gratitude towards committee members for my thesis, and professors who impacted me much on my understanding of statistics: Prof. Christopher Genovese, Prof. Sungkyu Jung, Prof. Rob Kass, Prof. Alessandro Rinaldo, Prof. Kathryn Roeder, and Prof. Larry Wasserman. I really admire your kind instructions and patience for me when I have little idea about research, and your thoughts helped me to generalize my horizon on my project.

Last but not the least, I am thankful for all the support and friendship I have received from the community in Carnegie Mellon University, especially Department of Statistics. When I entered CMU five years ago, I thought academia is just about research and math skills. My graduate program has shown me the full spectrums of this community. Now, I think academia is more of a stationary process that consists of lovely people. It is these lovely people solidify my faith in research, and help me much in my Ph.D life.

To my beloved ones who are separated from me
because of my Ph.D. program:
Baoquan, Honghua and Xin.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Contents

1.1 Background

Nowadays, we talk more and more about "Big Data", where the data matrix X consists of millions of observations and variables. In other words, both the number of columns and the number of rows are very large for X, hence X is a high dimensional matrix. With the development of technology, this kind of data set appears in many fields, such as genomics, astronomy, network, and so on. Unfortunately, for these large data sets, the classical statistical methods become either computational challenging or ineffective because of the curse of dimensionality ([Donoho 2000]). Hence, new statistical methods

need to be developed to exploit the underlying associations and patterns between variables, and the corresponding inference for Big Data. There are many studies on this topic, see ([Tibshirani 1996, Zhou 2010, Johnstone 2009]).

Although the data matrix X is huge, the real problem usually indicates some sparsity assumptions. For example, for gene analysis where millions of genes are measured for thousands of samples, it is natural to assume that there are only a few genes that take effect on specific inference. Hence, the useful genes are sparse. Based on the sparsity, we assume that the data matrix is a low rank information matrix covered by a noise matrix. The information matrix contains the inference we need, and we hope to recover this information matrix. This setting happens to fit many high dimensional problems.

1.1.1 Two-class Clustering Problem

Consider a high dimensional two-class clustering problem, where we observe X_i, $i = 1, 2, \ldots, n$, from two possible classes, and $X_i \in \mathrm{R}^p$ are feature vectors. The class labels $\{\ell_i\}_{1\le i\le n}$ take values from $\{-1, 1\}$. However, the labels are unknown to us and it is of major interest to estimate them.

One example is the cancer clustering with gene microarray data. Take the Lung Cancer Data for example [Gordon 2002]. The data set consists of 181 tissue samples from two classes MPM and ADCA (31 from MPM and 150 from ADCA), where for each sample, expression levels are measured on the same set of 12533 genes. A problem of interest is how to use the measured features to predict the class labels.

Despite that the class labels were originally given in [Gordon 2002], we assume that they are unknown, and focus our discussion on how to predict them. On the other hand, the true labels are used as ground truth for comparing the performance of different methods.

For this problem, we rewrite the model in matrix form. Fixing $\delta \in (0, 1)$, we assume δ fraction of the labels are 1, and model the feature vectors as

$$X_i = \mu_0 + y_i\mu + Z_i, \qquad i = 1, 2, \ldots, n, \tag{1.1.1}$$

where μ_0, μ, Z_i are vectors in R^p, representing the overall means, contrast means between two classes, and measurement noise correspondingly, and y_i are *adjusted labels*

$$y_i = \begin{cases} (1-\delta), & \ell_i = 1, \\ -\delta, & \ell_i = -1. \end{cases} \tag{1.1.2}$$

In matrix form, we can rewrite by

$$X = \mathbf{1}\mu_0' + y\mu' + Z, \tag{1.1.3}$$

where $X = X_{n,p} = [X_1, X_2, \ldots, X_n]'$, $Z = Z_{n,p} = [Z_1, Z_2, \ldots, Z_n]'$, $y = (y_1, y_2, \ldots, y_n)'$ and $\mathbf{1}$ is the $n \times 1$ vector of 1's.

In this model, $\mathbf{1}\mu_0' + y\mu'$ is the information matrix with rank two, and Z is the noise matrix. The inference of interest is the label vector ℓ, or $\mathrm{sgn}(y)$, which is contained in the information matrix. The data matrix has 181 rows and 12533 columns for Lung Cancer data example, which is a high dimensional two-class clustering problem.

In this problem, not only the data matrix is high dimensional, but also the number of variables (genes) is much larger than the number of observations (tissue samples). In this regime, the classical methods of K-means [Lloyd 1982] or hierarchical methods [Hastie 2009] can be either computationally challenging or ineffective. In this paper, we propose a new method which is a careful refinement of the classical spectral method, specifically designed for the case of $p \gg n$.

1.1.2 Signal Recovery Problem

With Model (1.1.3), there is also another unknown vector–the signal vector μ. In some applications, the recovery of μ is very important, and we call it as Signal Recovery Problem.

Take the electrocardiogram data (ECG data) as an example. Iain Johnstone and Arthur Lu ([Johnstone 2009]) has worked on the ECG data provided by Jeffery Froning and Victor Froelicher. In this data, beat sequences are record from patients, and then decomposed to about 60 cycles according to the peak of the beat. For each cycle, they use 512 interpolations to denote the recordings. The target is to find the cycle-to-cycle variation of the beat sequences, which is μ.

For this example, we model the baseline pattern for each cycle as μ_0, the variation between cycles as μ, and denote the random effect for i-th cycle as y_i. So, the feature vectors can be modeled as

$$X_i = \mu_0 + y_i\mu + Z_i, \qquad i = 1, 2, \ldots, n. \tag{1.1.4}$$

Rewrite it in the matrix form, and we have that

$$X = \mathbf{1}\mu_0' + y\mu' + Z, \tag{1.1.5}$$

where $Z_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$ with unknown parameterσ.

This is an application of high dimensional signal recovery problem. The data matrix has about 60 observations (cycles) and 512 features (interpolations). And the information, signal vector μ, is contained in the low rank matrix $\mathbf{1}\mu_0' + y\mu'$, which should be weak as the variation is small. We can see that the model for this problem has a similar form as clustering problem in previous section, with different interests.

1.1.3 Detection Problem

For some applications, we want to test whether there is information or not. Recall that we assume the data matrix is the summation of an information matrix and a noise

matrix, so it means that we want to test whether the information matrix is zero or not. We call this problem as Detection Problem.

An application of this problem is that high dimensional testing problem, where we assume the data comes from multivariate normal distribution with mean 0 and covariance matrix Σ. The problem is to detect whether Σ is identity matrix or a "spiky" matrix, which means that there are some nonzero off-diagonal entries. Some related works has been done by Emmanuel Candes and Yaniv Plan ([Candès 2009]), and other researchers.

1.1.4 Summary

With the introduction of three problems, we can find that there are some common facts about them. First, for all these problems, the number of variables is much larger than the number of observations, and both are large compared to classical problems. Hence the classical methods become inefficient. Second, the model can be decomposed as a low rank information matrix and a noise matrix, and the target is hidden in the information matrix. Especially, the information matrix has rank 2 for clustering problem and signal recovery problem. The low rank property makes things easier for us. These common facts indicate the possibility that the mathematical analysis for one problem can also be applied to another, and similar results can be achieved.

In my thesis, I will study all these three problems, with the focus on clustering problem. In Section 1.2, I will briefly introduce the clustering method I proposed for the clustering problem. In the following section, I will talk about the results about fundamental limits about the three problems.

1.2 Methodology

How to solve the clustering problem? Given that the information matrix is a low rank matrix, we recall the principal component analysis idea from Karl Pearson ([Pearson 1901]), with which we recall the classical spectral clustering method to apply. However, as we are dealing with high dimensional data, so we have to refine the spectral clustering method to add a feature selection step. To make sure that the feature selection procedure could achieve best results with spectral clustering, we learn experiences from accumulated wisdoms for this step. In this section, I will introduce the spectral clustering with feature selection approach.

1.2.1 Classical Spectral Clustering

It is worthwhile to take a look at the classical spectral method first. To bring out the idea, assume the overall mean vector μ_0 can be efficiently estimated by sample averages,

so we can consider a simplified version of (1.1.3) where μ_0 is removed from the model:

$$X = y\mu' + Z.$$

Consider the $n \times n$ *empirical dual covariance matrix* XX' (in contrast to the $p \times p$ covariance matrix $X'X$ [Lee 2010]). It is seen that

$$XX' = I + II, \qquad I = \|\mu\|^2 yy', \qquad II = y\mu' Z' + Z\mu y' + ZZ'. \tag{1.2.6}$$

In the classical setting where $p \ll n$, we frequently find that the spectral norm of II is much smaller than that of I, and so approximately,

$$XX' \propto yy'.$$

In this case, the leading eigenvector of XX' is (approximately) proportional to y, and the label vector ℓ can be estimated by $\ell_i = \mathrm{sgn}(y_i)$, so the clustering problem is settled.

Unfortunately, in the modern regime of $p \gg n$, we frequently find that the spectral norm of II is non-negligible compared to that of I. In such a case, a brute-forth implementation of the classical spectral method yields unsatisfactory results; see Table 1.3 for more discussion.

To overcome this challenge, a standard response is the feature selection, that is, using only a small fraction of carefully selected features for spectral clustering.

1.2.2 Feature Selection with Accumulated Wisdoms

In the past century, the statistics community has accumulated several noteworthy wisdom on feature selection. Putting them together is the following three step algorithm.

- Measure the significance of each gene by Kolmogorov-Smirnov statistic [Shorack 2009].
- Correct the theoretic null by the empirical null as suggested by Efron [Efron 2004].
- Feature selection by wavelet thresholding [Donoho 2006].

We now describe each step with more details.

Write the data matrix as $X = (x_{ij})_{1\le i\le n, 1\le j\le p}$. Fixing $1 \le j \le p$, the n data points associated with the j-th gene is

$$x_{ij} = \mu_0(j) + y_i\mu(j) + z_{ij}, \qquad i = 1, 2, \ldots, n.$$

We assume z_{ij} has the same distribution that does not depend on i (but may depend on j); the distribution is unknown to us but the mean is 0. Note that the adjusted labels y_i are unknown and the gene is differentially expressed if and only if $\mu(j) \neq 0$. In

such a situation, a well-known summarizing statistic is the Kolmogorov-Smirnov (KS) statistic, which is a well-known *goodness-of-fit* measure.

In detail, let

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \text{and} \qquad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

be the sample mean and sample variance corresponding to the j-th feature. We measure the significance of the j-th feature by the Kolmogorov-Smirnov score

$$KS_j = \sup_{-\infty<t<\infty} |F_{n,j}(t) - \Phi(\bar{x}_j + ts_j)|, \tag{1.2.7}$$

where $F_{n,j}(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_{ij} \le t\}$ is the empirical CDF of the j-th feature, and Φ is the CDF of $N(0,1)$. Compared to likelihood ratio test, KS is much more robust. Compared to moment-based tests (e.g. tests based on cumulants or kurtosis), KS is much more efficient.

In Figure 1.1, we display the KS scores for all genes. The figure suggests a major discrepancy between the scores and the theoretic null distribution: the asymptotic distribution of KS statistics as $n \to \infty$ under the assumption that $\mu(j) = 0$ and that z_{ij} are *i.i.d.* zero mean normals (here, i ranges and j is fixed).

Efron [Efron 2004] explained the causes of the discrepancy in detail, and suggests that in many applications the empirical null fits well with the summarizing scores. The empirical null is the distribution constructed by shifting and translating the theoretic null, where the two parameters (shift and translation) are chosen so that mean and variance of the empirical null fits with that of the data.

In light of this, Efron [Efron 2004] suggests the following simple approach to standardize the KS scores, which turns out to be effective. In detail, we standardize all p KS scores as follows, where SD stands for Standard Deviation:

$$KS_j = \frac{KS_j - \text{mean of all } p \text{ } KS\text{-scores}}{\text{SD of all } p \text{ } KS\text{-scores}};$$

for notational simplicity, we still denote the resultant scores by KS_j, $1 \le j \le p$. The standardized scores are displayed in Figure 1.1, which suggests a good fit between the empirical null and the data.

Roughly say, the larger the KS score KS_j, the more significantly the gene is differentially expressed. A reasonable approach to feature selection is then to keep those with larger KS-scores. Fix a threshold $t > 0$, we retain (remove) a feature j if and only if

$$KS_j \ge t \qquad (KS_j < t).$$

In the literature, this is called *wavelet hard-thresholding*.

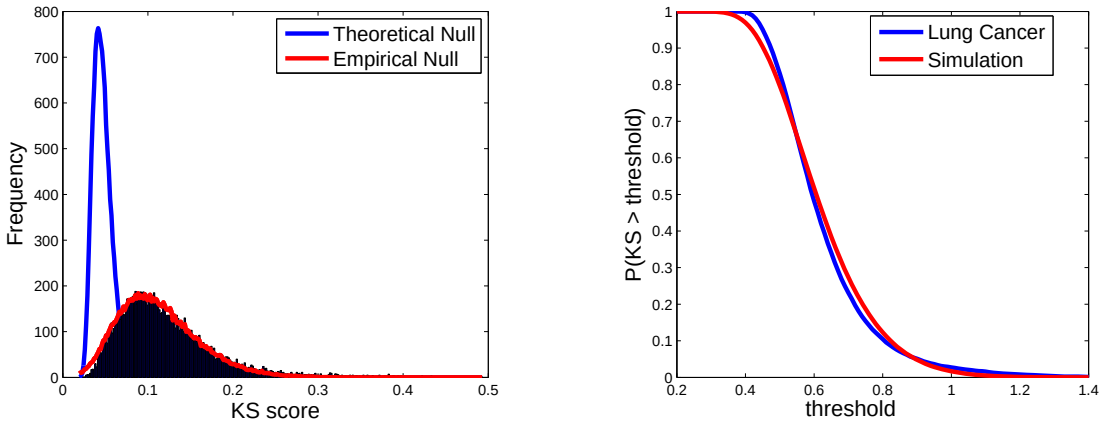

Figure 1.1: Left figure: The histogram is KS statistic from lung cancer data set. The two lines denote the distribution of KS statistic under theoretical null (blue) and empirical null (red). Right figure: The two lines denot the tail distribution of KS statistic from lung cancer data (blue) and simulated data (red).

1.2.3 Post-Selection Spectral Clustering with Applications

Denote the set of all the retained features by

$$\hat{S}(t) = \{1 \le j \le p : KS_j \ge t\},$$

and denote the post-selection data matrix and post-selection dual covariance matrix by $X^{(t)}$ and $H^{(t)}$, respectively. Note that

$$H^{(t)} = X^{(t)}(X^{(t)})',$$

and the sizes of $X^{(t)}$ and $H^{(t)}$ are $n \times |\hat{S}(t)|$ and $n \times n$, respectively. Similar to (1.2.6), we have the following decomposition,

$$H^{(t)} = I + II, \qquad I = \|\mu^{(t)}\|^2 yy', \qquad II = y\mu^{(t)\prime}Z^{(t)\prime} + Z^{(t)}\mu^{(t)}y' + Z^{(t)}Z^{(t)\prime},$$

where $\mu^{(t)}$ is an $|\hat{S}(t)| \times 1$ vector that μ is restricted on S_t. If the feature selection is successful, $\mu^{(t)}$ retains most of the significant features, but the dimension is much smaller than p. We therefore hope that spectral norm of II is negligible, compared to that of I, and so the leading eigenvector of $H^{(t)}$ is approximately proportional to y.

This motivates the post-selection spectral method, where we let $\xi^{(t)}$ be the leading eigenvector of $H^{(t)}$, the method clusters by estimating the labels with

$$\hat{\ell}_i = \mathrm{sgn}(\xi_i^{(t)}), \qquad 1 \le i \le n.$$

Threshold	No. of selected features	Errors
0.3876	12529	57
0.4976	10493	51
0.6076	5758	36
0.7176	2523	16
0.8276	1057	7
0.9376	**484**	**3**
1.0476	**261**	**2**
1.1576	129	10
1.2676	63	63
1.3776	21	55
1.4876	2	45

Table 1.1: The no. of selected features and clustering errors for lung cancer data with different thresholds. The first column shows the threshold we choose for feature selection, ranging from 0.3876 to 1.4876. The second column shows the number of selected features with the corresponding threshold, and the third row is the number of errors.

How does the method perform? With threshold t properly set, the above approach turns out to be quite successful. In Table 1.1. we tabulate the clustering errors of the above approach with t from 0.0486 to 0.1506 for lung cancer dataset. It is seen that with t properly set, this method outperforms the K-means and hierarchical clustering method, by a much smaller clustering errors. Therefore, with threshold set properly, the post-selection PCA can be truly successful.

With that being said, a problem emerges: how to set the threshold t. Setting the threshold too high may result in too few retained features, and setting it too low may result in too many retained features. Seemingly, this is the crucial problem.

1.2.4 Threshold Choice by Higher Criticism

We approach this by the recent innovation of Higher Criticism. Higher Criticism was first introduced Donoho and Jin ([Donoho 2004]) as a tool for multiple testing. Later, it was found to be also useful for feature selection in cancer classification ([Donoho 2008, Jin 2009]). Here, we define a statistic similarly as Higher Criticism statistic to choose the threshold t for spectral clustering problem, which we call HC.

To apply the Higher Criticism, we need the following three simple steps.

- For each $1 \leq j \leq p$, calculate a p-value $\pi_j = 1 - F_0(KS_j)$, where F_0 is the CDF of KS_j defined in (1.2.7) in the case where $x_{ij} \overset{i.i.d.}{\sim} N(u, \sigma^2)$, and the parameters (u, σ) are unknown.

- Sort all p-values in the ascending order $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.
- Define the Higher Criticism functional by

$$HC(p,j) = \frac{\sqrt{p}(j/p - \pi_{(j)})}{\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p\}}},$$

  and let $\hat{j}$ be the index that $HC_{p,j}$ reaches the maximum.

We then keep all the $\hat{j}$-features whose Kolmogorov-Smirnov scores KS_j are among the largest (e.g. if $\hat{j} = 50$, then we keep the 50 features with the largest scores). We call this method the Spectral-HCT. In figure 1.2, the red line shows the threshold we find at last.

1.2.5 Comparison with Other Methods

How does Spectral-HCT preform? Surprisingly well. In Figure 1.2.5, we compared the behavior of leading eigenvector of empirical covariance matrix without thresholding and with HCT. We can see that the one with HCT can be clearly divided into two groups which coincide with the truth, while the one without thresholding is messy.

In Table 1.3, we report the clustering error of Spectral-HCT. The clustering error is comparably to ideal thresholding, and is much smaller than that of K-means or hierarchical clustering methods. Here we use 3 microarray data sets. We have talked about lung cancer data before. The leukemia data set consists of 72 tissue samples from two classes ALL and AML (47 from ALL, 25 from AML), where each sample is measured on 7129 different genes. In colon data, there are 62 colon tissue samples from two classes tumor and normal (40 from tumor, 22 from normal), where 2000 different genes are measured for each sample. Although we know the truth, we *pretend* we do not know and try to do clustering.

We applied K-means, hierarchical clustering, spectral analysis without thresholding and with HCT for the data set. As the threshold to differentiate 2 clusters can be cut off not exactly at 0, so we also tried different thresholds for spectral analysis, and the result is in bracket.

It is seen that HCT outperforms all other methods by a much small error rate. Even more *surprisingly*, the *clustering* error of HCT is found to be sometimes smaller then the *classification* errors by the FAIR classifier proposed by Fan and Fan (2008) [J. 2008]. In [J. 2008], both data sets are investigated for the classification error of FAIR, where labels in the training sets (the training set contains 32 samples in the lung cancer data, and 38 samples in ALL) *are assumed as known*, and the reported classification errors for the test data are 7 for lung cancer data, and 1 for leukemia data.

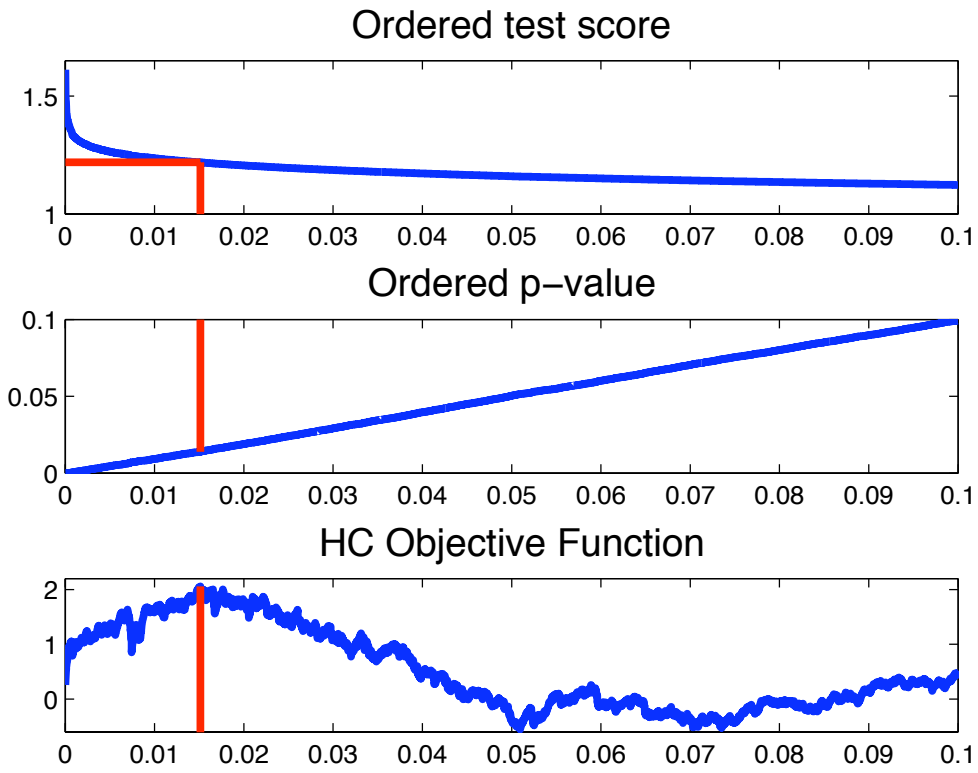

Figure 1.2: In the top figure, we have the ordered KS score. According to the KS score, we have the ordered p-value as the middle figure. With KS function, we have the HC objective function in the bottom figure, and we choose the one with largest HC score, which is the red line shows.

Data	K-means	Hier	Spectral	HCT
Leukemia	20/72	26/72	19(13)/72	**4(3)/72**
Lung	22/181	32/181	57(21)/181	**3(1)/181**
Colon	23/62	24/62	30(21)/62	**24(19)/62**

Table 1.2: The error rate for 3 datasets with different method. In each cell, the denominator is the number of samples, and the numerator is the number of errors by the correponding method. For Sepctral clustering without thresholding and with HCT, the number in bracket is the number of errors using optimal cut off.

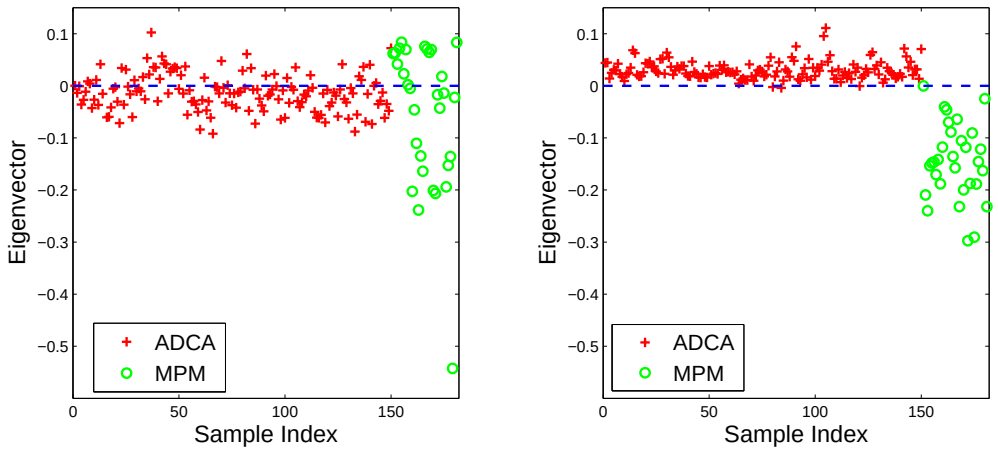

Figure 1.3: Left figure: The leading eigenvector of empirical dual covariance matrix without thresholding. Right figure: The leading eigenvector of empirical dual covariance matrix with thresholding. Crosses denote ADCA group, and circles denote MPM group. The dot line shows the estimated clustering result.

1.2.6 Key Idea of the Approach

We explain why HCT is a right way to set the threshold. Fix a threshold t. Suppose we apply the post-selection PCA proposed in Section 1.4 to the data matrix, and suppose we measure the performance of the procedure by the *Hamming error for clustering*:

$$\mathrm{Hamm}_p(t) = \text{expected number of misclassified samples over all samples}, \qquad (1.2.8)$$

then ideally, one would choose t as the *Ideal Threshold*—the threshold that minimizes the Hamming error. Unfortunately, $\mathrm{Hamm}_p(t)$ depends on the underlying distribution of data matrix in a complicated way, so it is hard to characterize the Ideal Threshold.

Our idea is to sought a functional that offers a nice approximation to the Hamming error functional, and that is easier to analyze. Recall that $\hat{\xi}^{(t)}$ denotes the leading eigenvector of the post-selection empirical dual covariance matrix. It turns out that for t in the range of interest,

$$\hat{\xi}^{(t)} \propto HC(t) \cdot y + z + rem, \qquad (1.2.9)$$

where $z \sim N(0, I_n)$, rem is a vector each coordinate of which is much smaller than $O(1)$, and $HC(t)$ is a non-stochastic term which can be viewed as the signal-to-noise ratio.

Combining (1.2.8)-(1.2.9), it is not surprising that there is an intimate relationship between $\mathrm{Hamm}_p(t)$ and $HC(t)$. In fact, we show that when signals are rare and weak,

$$\mathrm{Hamm}_p(t) \approx \exp\big(-[HC(t)]^2/2\big). \tag{1.2.10}$$

In comparison, $HC(t)$ has a much simpler formula, and so the term on the right hand side provides a more tractable approximation of the term on the left hand side.

Somewhat surprisingly, it turns out that the explicit form of $HC(t)$ is reminiscent of the *Higher Criticism*—a notion that is developed in seemingly unrelated settings (e.g., signal detection [Donoho 2004]), and classification [Donoho 2008, Jin 2009]). Therefore, for literature reasons, we call $HC(t)$ the *Higher Criticism functional*, despite that the actual form of $HC(t)$ is significantly different from those of the previous versions of Higher Criticism.

We call the threshold that maximizes $HC(t)$ the *Ideal HC Threshold*. On one hand, (1.2.10) suggests that the Ideal HC Threshold approximates the Ideal Threshold. On the other hand, HCT is a data-driven threshold that is carefully designed so that it maximizes the empirical counterpart of the HC functional, and so HCT can be viewed as the *stochastic counterpart* of the Ideal HC Threshold. In other words,

$$\text{HCT} \quad \approx \quad \text{Ideal HC Threshold} \quad \approx \quad \text{Ideal Threshold},$$

which explains why HCT is a good threshold choice.

1.2.7 Extension

We also try to extend the method to multi-class case. For multi-class clustering, we take $K-1$ eigenvectors with the $K-1$ eigenvalues with largest magnitude. Say that these eigenvectors form an $n \times (K-1)$ matrix U. Apply K-means method for U under the condition that the number of classes is smaller than K. The labels found by K-means is the label we need.

We use this method for some data sets with number of classes $K > 2$, and the result is very good. Take the Brain data ([Pomeroy 2002]) as an example. In this data, there are 42 samples, and the expression level of 5597 genes are recorded for each sample. There are 5 classes for different tumor types, and we want to differentiate them. For Lymphoma data ([Alizadeh 2000]), there are 62 samples, and the expression levels of 4026 genes are measured for this data set. There are 3 classes in this data set depending on different subtypes of lymphoma. The data set SRBCT ([Khan 2001]) includes 63 samples, and the expression levels of 2308 genes are measured for each sample. There are 4 classes in this data set, divided by different tumor types. For these three data set, we apply K-means, hierarchical clustering, SpectralGem and our method. The comparison of different methods can be found in the following table.

Data	K-means	Hier	SpectralGem	Spectral-HCT
Brain	15/42	23/42	14/42	**10/42**
Lymphoma	20/62	29/62	14/62	**3/62**
SRBCT	38/63	34/63	34/63	**33/63**

Table 1.3: The error rate for 3 datasets with different method. In each cell, the denominator is the number of samples, and the numerator is the number of errors by the corresponding method.

From this table, we can see that our method does improve the clustering performance in some way. For some data sets such as Lymphoma data, our method performs a great improvement, while for SRBCT, it is not so obvious. The improvement depends on the data itself. However, feature selection does help.

1.3 Fundamental Limits

In fact, under proper model, the phase diagram of Spectral-HCT approach can be recovered. It arises our interest in the fundamental limit of two-class clustering problem. What is fundamental limit? It is to study that if there is an area, such that when the model parameters drop in this area, then any method will fail to get inference. Otherwise, the inference can be achieved.

To find the boundary of the region of impossibility and region of possibility for the desired inference, we need a lower bound of the boundary and an upper bound of the boundary. To get the desired inference, the difference of distribution between null hypothesis and alternative hypothesis must be large. In other words, if we could show that the difference between two distributions goes to 0 when p goes to infinity, then the inference is unable to be recovered. It gives us a lower bound. For the upper bound, we have to examine the existed methods to get. If the upper bound meets the lower bound, then the corresponding boundary can be found.

As the model for clustering problem, signal recovery problem, and detection problem is the same, the research of lower bound and upper bound of clustering problem can be generalized to the other problems. Similarly, we get the fundamental limits for all the three problems. The details will be shown in Chapter 6.

1.4 Related Motivated Topics

In our analysis, some related topics are also motivated to develop new theories. In my thesis, I also include these topics.

1.4.1 The Random Matrix Theory

The post-selection spectral clustering requires new development in the Random Matrix Theory (RMT). While most work in RMT addresses data matrxix with $i.i.d$ entries, the data matrix we face in post-selection spectral clustering is much more complicated: the columns of the post-selection data matrix $X^{(t)}$ not only depend on each other in a complicated way, but also have a distribution that is very different from that of pre-selection columns. To overcome this difficulty, we have to develop new RMT.

In Chapter 2, we introduce the results about random matrices with $i.i.d$ entries, and then show the results about the leading eigenvalue and leading eigenvector for the post-selection random matrix under our model. The result will be applied to prove the main result for spectral-HCT approach.

1.4.2 Tail Distribution of Kolmogorov-Smirnov Statistic

In our approach, we use Kolmogorov-Smirnov (KS) statistic to test the significance of genes. To find the corresponding p-value, the current method is to simulate a series of KS statistics with the same n under null hypothesis, and then compare the KS statistic from data with them to calculate an approximate p-value. This method works for real data analysis. However, when n becomes large, it will be time consuming to simulate the KS statistics. What's more, we are also interested in the theoretical distribution of KS statistic.

In my thesis, I will also introduce the boundary-cross approach ([Loader 1992]), from which we find the approximation of empirical KS statistic. I will also show the simulation result about the difference between theoretical distribution and empirical distribution.

1.5 Content

In Chapter 2, I will discuss the research on RMT as the preliminary knowledge for my thesis. In Chapter 3, I will talk about the method in detail. I will briefly show the idea why our method works, and talk about the chosen of threshold. In Chapter 4, I will find the precise proof and boundary for our method, under specific model (ARW model). Some simulations and extensions are also discussed in this chapter. Then, in Chapter 5, I introduce our study on KS statistic, which is very important in the feature selection step. We will talk about the theoretical tail distribution for it. At last, in Chapter 6, the fundamental limits for matrix recovery problems are introduced, as an extension of the clustering problem.

Chapter 2

The Random Matrix Theory

Contents

2.1 Background

For spectral clustering, an important step is to calculate the leading eigenvector of empirical dual covariance matrix, and use this eigenvector to estimate the labels. However, to examine the efficiency of this approach, we need the theoretical results about the leading eigenvector of empirical dual covariance matrix, and hence the results about the leading eigenvector of the random matrix XX'. That's why we are interested in RMT.

Besides, the results about RMT is also important in information theory, signal processing, and small-world networks. The various applications of RMT motivates the study on it. Nowadays, there are many results about the asymptotic and non-asymptotic results about the leading eigenvalue and leading eigenvector of a random matrix with $i.i.d$ sub-Gaussian entries. The study and results are summarized in [Tulino 2004, Vershynin 2010].

Even though there are many results, I can not use them directly in my case. In post-selection random matrix, the distribution of entries are complicated, with dependence on other columns. The complicated distribution requires new development of RMT. That's what I will introduce in this chapter. In Section 2.2, I will briefly review the past results needed in my thesis, and then I will show the main results in Section 2.3. In Section 2.4, the proof of the main theorem will be shown.

2.2 Review of Past Research

Let Z be a standard real Gaussian $n \times p$ matrix with $i.i.d$ real zero-mean Gaussian entries with variance $\sigma = 1$, where $p \geq n$. We call the $n \times n$ matrix ZZ' as Wishart matrix, and denote it by H_0. What we care about is the largest eigenvalue of H_0, denoted by $\lambda_{max}(H_0)$, and the smallest eigenvalue of H_0, denoted by $\lambda_{min}(H_0)$. As H is very near to an identical matrix times p, both $\lambda_{max}(H_0)$ and $\lambda_{min}(H_0)$ would be around p.

With these definitions, some results about $\lambda_{max}(Z)$ and $\lambda_{min}(Z)$ can be found as following ([Vershynin 2010, Page 22]), and hence results about $\lambda_{max}(H_0)$ and $\lambda_{min}(H_0)$ follow.

Corollary 2.2.1 *Let Z be an $n \times p$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1-2\exp(-t^2/2)$ one has*

$$\sqrt{p} - \sqrt{n} - t \leq \lambda_{min}(Z) \leq \lambda_{max}(Z) \leq \sqrt{p} + \sqrt{n} + t.$$

This corollary indicates that the eigenvalues are around $\sqrt{p}$, and the eigenvalues of H_0 are around p.

The result can also be extended to sub-Gaussian random variables. To begin with, we need the following definitions.

Definition 2.2.1 *A real-valued random variable X is said to be sub-Gaussian with parameter $\sigma > 0$, if for any $t \in R$ there is*

$$E[e^{tX}] \leq e^{\sigma^2 t^2/2}.$$

It can be easily derived that when X is sub-Gaussian distributed with parameter σ, there is that $E[X] = 0$ and $Var(X) \leq \sigma^2$.

Definition 2.2.2 *A real-valued random variable X is said to be sub-exponential distributed, if there exists a constant K, such that $(E|X|^p)^{1/p} \leq Kp$, for all $p \geq 1$. When X is sub-exponential distributed, define the sub-exponential norm of X to be*

$$\|X\|_{\psi_1} = \sup_{p \geq 2} p^{-1}(E|X|^p)^{1/p}.$$

These two definitions define the distribution for entries in the random matrix. However we also need the relationship between different entries, as the following definition.

Definition 2.2.3 *(Isotropic random vectors). A random vector X in R^n is called isotropic if $\Sigma(X) = I$. Equivalently, X is isotropic if*

$$E\langle X, x\rangle = \|x\|_2^2 \qquad \text{for all } x \in R^n.$$

With these definitions, there is the following theorem ([Vershynin 2010, Page 26]) for a random matrix with independent sub-Gaussian rows.

Theorem 2.2.1 *(Sub-gaussian rows). Let Z be an $p \times n$ matrix whose rows A_i are independent sub-Gaussian isotropic random vectors in R^n. Then for every $t \geq 0$, with probability at least $1 - 2exp(-ct^2)$ one has*

$$\sqrt{p} - C\sqrt{n} - t \leq \lambda_{min}(Z) \leq \lambda_{max}(A) \leq \sqrt{p} + C\sqrt{n} + t.$$

Here $C = C_K$, $c = c_K > 0$ depend only on the sub-Gaussian norm $K = max_i \|A_i\|_{\psi_1}$ of the rows.

The proof of theorem paves the way for my research.

These results are important, however not enough for my thesis. I will introduce the subtle development I did on RMT.

2.3 Main Result

For the random matrix in spectral-HCT approach, there are two main differences from what we talk above. First, the feature selection step would cause correlation between rows. Recall that we use KS statistic to do thresholding, which is calculated from the feature of n observations, or a column of the data matrix. If we choose this feature, then the KS statistic should be large, hence there is correlation in the post-selection data matrix for each column (however the columns are dependent with each other). Second, the distribution for different columns are different. As we have signals and noise, the corresponding selected columns for them are different. For the column associated with signals, it will be easier to be selected; and the ones associated with noise is harder. However, the number of signals is comparatively small.

Based on the approach, I introduce the random matrix as following. For the random matrix Z, which is an $n \times p$ random matrix with independent column vectors. Assume that there is $p \geq k \geq 1$, such that the column vectors can be decomposed into two sets with k columns and $p - k$ columns. The distribution of the column vectors is the same for one set and can be bounded for another set. The bound will be introduced more in the theorem.

2.3.1 Main Result

Based on the idea of the model, we have the following result.

Theorem 2.3.1 *Fix $p > n > 1$ and $p > k > 1$, with $p > 4n\log(9)$. Let $Z = Z_{n,m} = [z_1, z_2, \cdots, z_m]$ be an $n \times p$ random matrix with independent column vectors z_i, $1 \leq i \leq m$. Suppose that for any non-random unit-norm vector $a \in R^n$, Q satisfies that*

(a) *For all* $1 \leq i \leq p$, $[a'z_i]^2$ *is sub-exponential distributed, and* $\|(a'z_i)^2\|_{\psi_1} \leq K$, *where* K *is a constant;*

(b) *If there is a partition* $\{1, 2, \cdots, p\} = S_0 \cup S_1$ *and a constant* c_0, *such that* $|S_0| = p - k$ *and* $|S_1| = k$, *with that* $E[(a'z_i)^2] = 1 + c_0\sqrt{\frac{\log(p)}{n}}$ *for all* $i \in S_0$, *and that* $\left|E[(a'q_i)^2] - 1\right| \leq c_0\sqrt{\frac{\log(m)}{n}}$ *for all* $i \in S_1$.

Then with probability at least $1 - 9^{-n}$, *all singular values of* Q *fall between*

$$\sqrt{p(1 + c_0\sqrt{\log(p)/n})} \pm [8eK_0\sqrt{\log(9)}\sqrt{n} + 2c_0k\sqrt{\log(p)/np}],$$

where $K_0 = K + 1 + c_0\sqrt{\log(m)/n}$.

2.4 Proof

2.4.1 Proof of Theorem 2.3.1

We write for short $\eta = c_0\sqrt{\frac{\log(p)}{n}}$, and $\xi = 4eK_0\sqrt{\log(9)}$. To show the claim, it is sufficient to show with probability at least $1 - 9^{-n}$,

$$\|(1/p)ZZ' - (1 + \eta)I_n\| \leq 2\xi\sqrt{n/p} + 2\eta k/p. \tag{2.4.1}$$

Once this is shown, then we combine it with the algebraic fact that

$$|(\frac{z}{\sqrt{1+\eta}})^2 - 1| \geq |\frac{z}{\sqrt{1+\eta}} - 1|,$$

and take z as the singular values of $Q/\sqrt{p}$. The result follows that with probability at least $1 - 9^{-n}$, all the singular values of Q fall between

$$\sqrt{m(1 + \eta)} \pm [2\lambda\sqrt{n} + 2\eta k/\sqrt{p}]. \tag{2.4.2}$$

To show (2.4.1), we need some preparations. Let (X, d) be a metric space and let $\alpha > 0$ be a small fixed constant. A subset $\mathcal{N}_\alpha$ of X is called an α-net of X if for every point $x \in X$, there is a point $y \in \mathcal{N}_\epsilon$ such that $d(x, y) \leq \alpha$. The minimal cardinality of an α-net of X, if finite, is denoted $N(X, \alpha)$ and called the covering number of X. The following lemmas are well-known in the literature of the Random Matrix Theory and can be found in [Vershynin 2010, Page 8].

Lemma 2.4.1 *Fix* $n > 1$ *and* $\alpha > 0$. *For the unit sphere* S^{n-1} *in* R^n *equipped with the Euclidean metric,* $N(S^{n-1}, \alpha) \leq (1 + \frac{2}{\alpha})^n$.

Lemma 2.4.2 *Let* A *be a symmetric* $n \times n$ *matrix, and let* $\mathcal{N}_\alpha$ *be an* α*-net of* S^{n-1} *for some* $\alpha \in (0, 1/2)$. *Then* $\|A\| = \sup_{x \in S^{n-1}}\{|a'Aa|\} \leq (1 - 2\alpha)^{-1}\sup_{a \in \mathcal{N}_\alpha}\{|a'Aa|\}$.

In particular, if we take $\alpha_0 = 1/4$, then there is an $1/4$ net $\mathcal{N}_{1/4}$ such that

$$|\mathcal{N}_{1/4}| \leq 9^n. \tag{2.4.3}$$

Moreover, for any $x > 0$ and any $n \times n$ symmetric random matrix A,

$$P(\|A\| \geq x) \leq P(2 \sup_{a \in \mathcal{N}_{1/4}} \{|a'Aa|\} \geq x) \leq \sum_{a \in \mathcal{N}_{1/4}} P(|a'Aa| \geq x/2). \tag{2.4.4}$$

Combining (2.4.3) and (2.4.4), to show (2.4.1), it is sufficient to show that,

$$P(|a'((1/p)ZZ' - (1+\eta)I_n)a| \geq \lambda\sqrt{n/p} + \eta k/p) \leq 9^{-2n}, \qquad a \in N_{1/4}. \tag{2.4.5}$$

Now we try to show (2.4.5). Fix $a \in S^{n-1}$. Denote $w_j = Z_j'a$ and $w = (w_1, w_2, \ldots, w_p)'$. So, we have that $\|w\|^2 = a'(ZZ')a$, and that $E[w_j]^2 = (1+\eta)$ for $j \in S_0$, and that $|E[w_j]^2 - 1| \leq \eta$ for $j \in S_1$. So there is $\big[E[\|w\|_2^2] - (1+\eta)p\big] \leq 0$, and then there is

$$\begin{array}{rl} & P(a'[(1/p)ZZ' - (1+\eta)I_n]a \geq x/\sqrt{p} + \eta k/m) \\ = & P(\|w\|_2^2 - E[\|w\|_2^2] \geq \sqrt{p}x + \eta k + (1+\eta)p - E[\|w\|_2^2]) \\ \leq & P(\|w\|_2^2 - E[\|w\|_2^2] \geq \sqrt{p}x). \end{array} \tag{2.4.6}$$

Similarly, we have that

$$P(a'[(1/p)ZZ' - (1+\eta)I_n]a \leq -x/\sqrt{p} - \eta k/p) \leq P(\|w\|_2^2 - E[\|w\|_2^2] \leq -\sqrt{p}x). \tag{2.4.7}$$

Note that for w_j, we know that w_j^2 is sub-exponential distributed, with sub-exponential norm K. According to the Remark in Appendix, $w_j^2 - E[w_j^2]$ is centered sub-exponential random variables, with sub-exponential norm $\|w_j^2 - E[w_j^2]\|_{\psi_1} \leq K + 1 + \eta = K_0$. Using Theorem 2.4.4, for any $0 < x < 2eK_0\sqrt{p}$,

$$P(\big|\|w\|_2^2 - E[\|w\|_2^2]\big| > \sqrt{p}x) \leq 2e^{-\frac{1}{8e^2K_0^2}x^2}. \tag{2.4.8}$$

Combining (2.4.8) with (2.4.6) and (2.4.7), there is

$$P(|a'((1/p)ZZ' - (1+\eta)I_n)a| \geq \lambda\sqrt{n/p} + \eta k/p) \leq 2 \cdot 9^n e^{-\frac{1}{8e^2K_0^2}x^2},$$

and (2.4.1) follows by taking $x = 4eK_0\sqrt{n\log(9)} = \lambda\sqrt{n}$ and basic algebra. □

2.4.2 Details about sub-Gaussian R.V.

For independent sub-Gaussian random variables, the linear combination of them is also sub-Gaussian distributed.

Theorem 2.4.1 *If X_i are independent sub-Gaussian random variables with common parameter σ, $1 \le i \le n$, then for any $n \times 1$ constant vector a, the random variable $a'X$ is sub-Gaussian with parameter $\|a\|\sigma$.*

Proof. Note that

$$E[e^{tX_i}] \le e^{\sigma^2 t^2/2}, \qquad 1 \le i \le n,$$

so we have

$$E[e^{t\sum_{i=1}^n a_i X_i}] = \prod_{i=1}^n E[e^{ta_i X_i}] \le e^{\sum_{i=1}^n a_i^2 \sigma^2 t^2/2} = e^{\|a\|^2 \sigma^2 t^2/2}.$$

So the claim follows. □

Also, there is some equivalent conditions for sub-Gaussian random variables.

Theorem 2.4.2 *For a centered random variable X, the following statements are equivalent:*

1. *$\exists \sigma > 0$, such that $E[e^{tX}] \le e^{\sigma^2 t^2/2}$, any $t \in R$;*
2. *$\exists K_1 > 0$, such that for any $s > 0$, $P(|X| \ge s) \le 2e^{-K_1 s^2}$;*
3. *$\exists K_2 > 0$, such that $E[e^{K_2 X^2}] \le 2$.*

Proof. $1 \to 2$. With Markov inequality, combining with property 1 that $E[e^{tX}] \le e^{\sigma^2 t^2/2}$, we have that

$$P(|X| > s) \le e^{-ts} E[e^{tX}] \le e^{-ts + t^2\sigma^2/2}.$$

Choose that $t = \frac{s}{\sigma^2}$ to minimize the right side, and we get that

$$P(|X| > s) \le e^{-s^2/2\sigma^2}.$$

Take $K_1 = 1/2\sigma^2$ and property 2 follows.

$2 \to 3$. Take $K_2 = K_1/3$. With property 2, we have that

$$P(e^{K_2 X^2} \ge e^{K_2 s^2}) = P(|X| \ge s) \le 2e^{-K_1 s^2}.$$

So, there is

$$E[e^{K_2 X^2}] = \int_0^\infty P(e^{K_2 X^2} \ge e^{K_2 s^2}) d(e^{K_2 s^2}) \le 1 + 2\int_0^\infty 2K_2 e^{K_2 - K_1 s^2} d(s^2).$$

Introduce in $K_2 = K_1/3$, then we have

$$2\int_0^\infty 2K_2 e^{K_2 - K_1 s^2} d(s^2) = \int_0^\infty \frac{2}{3} K_1 e^{-2/3 K_1 s^2} d(s^2) = 1.$$

So, there is $E[e^{K_2 s^2}] \leq 2$ when $K_2 = K_1/3$.

$3 \to 1$. The proof can be found at [Rivasplata 2012, Page 6].

Assume that $E[e^{K_2 X^2}] \leq 2$ for some $K_2 > 0$. Recalling that X is centered, we have

$$E[e^{tX}] = 1 + \int_0^1 (1-y)E[(tX)^2 e^{ytX}]dy \leq 1 + \frac{t^2}{2}E[X^2 e^{|tX|}].$$

As $|tX| \leq \frac{1}{2}(t^2/K_2 + K_2 X^2)$, so we have

$$\frac{t^2}{2}E[X^2 e^{|tX|}] \leq \frac{t^2}{2}e^{t^2/(2K_1)}E[X^2 e^{K_2 X^2/2}] \leq \frac{t^2}{2}e^{t^2/(2K_2)}E[e^{K_2 X^2}].$$

So, we have

$$E[e^{tX}] \leq 1 + \frac{t^2}{2}e^{t^2/(2K_2)}E[e^{K_2 X^2}] \leq (1 + t^2/K_2)e^{t^2/(2K_2)} \leq e^{3t^2/(2K_2)}$$

So X is sub-Gaussian with parameter $\sigma = \sqrt{3/K_2}$. □

Besides sub-Gaussian random variables, we also introduce another type of random variables: sub-exponential random variable, which have tails heavier than Gaussian tails. Recall the definition of sub-exponential random variable as following.

Definition 2.4.1 *A real-valued random variable X is said to be sub-exponential distributed, if $\exists K$, such that $(E|X|^p)^{1/p} \leq Kp$, for all $p \geq 1$. When X is sub-exponential distributed, define the sub-exponential norm of X to be*

$$\|X\|_{\psi_1} = \sup_{p \geq 2} p^{-1}(E|X|^p)^{1/p}.$$

Remark. According to triangle inequality, $\|X - E[X]\|_{\psi_1} \leq \|X\|_{\psi_1} + |E[X]|$. So, when a sub-exponential random variable is not centered, we could always make it to be centered sub-exponential distributed.

Theorem 2.4.3 *For a random variable X, if there exists $c > 0$ and $s > 0$, such that*

$$P(|X| > t) \leq c\exp(-st), \qquad t > 0.$$

Then X is sub-exponential distributed.

Proof. We need to show that, there exists $K > 0$, such that for any p, there is

$$p^{-1}(E|X|^p)^{1/p} \leq K. \tag{2.4.9}$$

So, we try to calculate $E|X|^p$. As $|X|^p > 0$ with probability 1, with elementary statistics, we have

$$E[|X|^p] = \int_0^\infty P(|X|^p > t^p)d(t^p) = \int_0^\infty P(|X| > t)pt^{p-1}d(t).$$

Combining with that $P(|X| > t) \leq c\exp(-st)$, we have that

$$E[|X|^p] \leq \int_0^\infty ce^{-st}pt^{p-1}d(t) = \frac{cp}{s^p}\Gamma(p).$$

With Stirling's approximation, when p is large, there is $\Gamma(p) = p! \leq \sqrt{2\pi p}(p/e)^p e^{1/(12p)}$. Introduce it into $E[|X|^p]$, and we have

$$E[|X|^p] \leq \frac{cp\sqrt{2\pi p}}{s^p}(p/e)^p e^{1/(12p)}.$$

So, when $p \to \infty$, we have that

$$\limsup_{p\to\infty} p^{-1}(E[|X|^p])^{1/p} \leq 1/(se).$$

As $E[|X|^p]$ is bounded when p is finite, so there exists K, such that (2.4.9) holds. □

Lemma 2.4.3 *Let X be a centered sub-exponential random varaible. Then, for t such that $|t| \leq 1/(2e\|X\|_{\psi_1})$, there is*

$$E[e^{tX}] = \exp(2e^2t^2\|X\|_{\psi_1}^2).$$

Proof. Note that

$$E[e^{tX}] = 1 + tE[X] + \sum_{p=2}^{\infty} t^p \sum E|X|^p/p! \leq 1 + \sum_{p=2}^{\infty}(tp\|X\|_{\psi_1})^p/p!.$$

When $|t| < 1/(2e\|X\|_{\psi_1})$, the second term on the right can be controlled as

$$\sum_{p=2}^{\infty}(tp\|X\|_{\psi_1})^p/p! \leq \sum_{p=2}^{\infty}(e\|X\|_{\psi_1}|t|)^p \leq 2(e\|X\|_{\psi_1}|t|)^2.$$

So, we have

$$E[e^{tX}] \leq 1 + 2(e\|X\|_{\psi_1}|t|)^2 \leq \exp(2e^2t^2\|X\|_{\psi_1}^2).$$

□

With this lemma, we show the large deviation result for sums of independent sub-exponential random variables, which can be viewed as an extension of Berstein inequality.

Theorem 2.4.4 *Let $X_1, \cdots, X_p$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $p \times 1$ constant vector a, and any $t \geq 0$, we have*

$$P(|\sum_{i=1}^{p} a_iX_i| \geq t) \leq 2\exp(-\frac{1}{4e}\min\{\frac{t^2}{2eK^2\|a\|^2}, \frac{t}{K\|a\|_\infty}\}).$$

Proof. The proof can be found in [Vershynin 2010, Page 14], and here is a copy of that proof. Without loss of generality, we assume that $K = 1$ by replacing X_i with X_i/K and t with t/K. With Markov inequality, we have that

$$P(|\sum_{i=1}^{p} a_i X_i| \geq t) \leq e^{-\lambda t} E[e^{\lambda(\sum_{i=1}^{p} a_i X_i)}] = e^{-\lambda t} \prod_{i=1}^{p} E[e^{\lambda a_i X_i}].$$

Combining with Lemma 2.4.3, if $|\lambda| \leq 1/(\|a\|_\infty 2e)$ then $|\lambda a_i| \leq 1/(2e)$, so there is

$$P(\sum_{i=1}^{p} a_i X_i \geq t) \leq \exp(-\lambda t + 2e^2 t^2 \lambda^2 \|a\|^2).$$

Choose $\lambda = \min\{\frac{t}{4e^2\|a\|^2}, 1/(2e\|a\|_\infty)\}$, we obtain that

$$P(\sum_{i=1}^{p} a_i X_i \geq t) \leq \exp(-\min\{\frac{t^2}{8e^2\|a\|^2}, \frac{t}{4e\|a\|_\infty}\}).$$

Repeating this argument for $-X_i$ instead of X_i, we obtain the same bound for $P(-\sum_{i=1}^{p} a_i X_i \geq t)$. A combination of these two bounds completes the proof. □

Chapter 3

Methodology

Contents

3.1 Background

In Chapter 1, I introduce the clustering problem, and state the difficulty for the clustering problem with high dimensional data. The difficulty is not only for two-class clustering problem, but also more complicated models with number of classes $K > 2$. I have introduced how Spectral-HCT method can be extended to multi-class clustering case. However, why can we extend the method in this way? Why it works? I will show the idea in this chapter.

3.1.1 Model

When there are more than two classes, how do we model the data? Here I will talk about this case. Assume that there are K classes in total.

For a feature vector X_i, take the label of X_i as $e_{l(i)} = (0, 0, \cdots, 1, \cdots, 0)'$, where $e_{l(i)}$ is a $K \times 1$ vector with only one nonzero element 1 at the location $l(i)$. The location of the element 1, which is $l(i)$, indicates the class of X_i. Then, the feature vector can be written as

$$X_i = e_{l(i)} M' + z, \qquad z \sim N(0, I_n), \tag{3.1.1}$$

where M is a $p \times K$ matrix, with $M(j, i)$ indicates the mean of j-th variable for class i.

In matrix form, it can be rewritten as

$$X = LM' + Z, \tag{3.1.2}$$

where

$$L = \begin{pmatrix} 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{3.1.3}$$

which is an $n \times K$ matrix.

Assume that the proportion of class i is δ_i, then the overall mean for j-th variable is that

$$d_j = \sum_{i=1}^{K} \delta_i M(j,i). \tag{3.1.4}$$

Take $d_{j1} = M(j,1) - d_j$, $d_{j2} = M(j,2) - d_j$, and so on, then the data matrix can be rewritten as

$$X = \mathbf{1}\mathbf{d}' + LD' + Z, \tag{3.1.5}$$

where $\mathbf{1}$ is an $n \times 1$ vector with all elements as 1, and $\mathbf{d}$ is a $p \times 1$ vector with overall mean for each gene. L is the same as before, and D is a $p \times K$ matrix with elements $D(j,i) = d_{ji}$.

It is obvious that

$$\sum_{i=1}^{K} \delta_i d_{ji} = \sum_{i=1}^{K} \delta_i (\mu_{ji} - d_j) = \sum_{i=1}^{K} \delta_i \mu_{ji} - d_j = 0,$$

so there is

$$d_K = \sum_{i=1}^{K-1} \frac{\delta_i}{\delta_K} d_{ji}.$$

Introduce it into the model (3.1.5), then the model can be written as

$$X = \mathbf{1}\mathbf{d}' + \tilde{L}\tilde{D}' + Z, \tag{3.1.6}$$

where $\tilde{L}$ is an $n \times (K-1)$ matrix. When sample i is in group l, $1 \le l \le K-1$, then $L_{il} = 1$, and other elements are 0. When sample i is in group K, then

$$L_i = (-\frac{\delta_1}{\delta_K}, -\frac{\delta_2}{\delta_K}, \cdots, -\frac{\delta_{K-1}}{\delta_K}). \tag{3.1.7}$$

Matrix $\tilde{D}$ is a $p \times (K-1)$ matrix, which is the first $K-1$ columns of matrix D.

In this way, the empirical dual covariance matrix $H = (X - \mathbf{1}\mathbf{d}')(X - \mathbf{1}\mathbf{d}')'$ is

$$H = \tilde{L}\tilde{D}'\tilde{D}\tilde{L}' + \tilde{L}\tilde{D}'Z' + Z\tilde{D}\tilde{L}' + ZZ'. \tag{3.1.8}$$

$\tilde{D}'\tilde{D}$ is an $(K-1)\times(K-1)$ matrix with rank $K-1$. So $\tilde{L}\tilde{D}'\tilde{D}\tilde{L}'$ also has rank $K-1$.

When the eigenvalues for the information matrix $\tilde{L}\tilde{D}'\tilde{D}\tilde{L}'$ is much larger than the leading eigenvalue of noise part, then the recovered eigenvectors could indicate the class for the samples. That's why we want to do feature selection to reduce the magnitude of leading eigenvalue for the noise matrix.

3.2 Methodology

In this section, I will introduce the clustering algorithm, and discuss the assumptions of the algorithm for real data. At last, I will compare the new approach with other approaches on several real data sets.

3.2.1 Algorithm

There are four steps in the algorithm. I will talk about them step by step.

1. Rank the features by Kolmogorov-Smirnov (KS) statistic.

   For a feature vector X_i, the Kolmorov-Smirnov statistic is defined as following equation,

   $$KS_n = \sqrt{n} \sup_{-\infty<t<\infty} |F_n(t) - \Phi(t; \hat{\mu}_i, \hat{\sigma}_i)|, \tag{3.2.9}$$

   where $F_n(t)$ is the empirical cumulative density function from the feature vector X_i, and $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the corresponding empirical mean and variance for data. Kolmogorov-Smirnov statistic is to test the maximal difference between empirical CDF and normal distribution.

   Then, adjust the KS statistic by normalizing the mean and variance. According to Efron's paper ([Efron 2004]), the parameters for empirical null distribution is different from that for theoretical null distribution. With the lighting of this idea, we have to adjust the KS statistic by their mean and variance, to be at the same scale for empirical null distribution.

   With the adjusted KS statistic and simulated KS statistics, we can find the corresponding p-value for each gene. Denote the p-value by π_i. Sort p-values in an ascending way such that

   $$\pi_{(1)} \leq \pi_{(2)} \leq \cdots \leq \pi_{(p)}.$$

2. Feature selection by thresholding the p-value.

   I use Higher Criticism Thresholding (HCT) for feature selection step. With the sorted p-values, define the Higher Criticism functional by

$$HC(p,j) = \frac{\sqrt{p}(j/p - \pi_{(j)})}{\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p\}}},$$

   and let $\hat{j}$ be the index that $HC_{p,j}$ reaches the maximum. We then keep all the $\hat{j}$-features whose p-values π_i are among the smallest (e.g. if $\hat{j} = 50$, then we keep the 50 features with the smallest p-values). Then, we get the post-selection data matrix $X^{(t)}$.

3. Post-selection PCA.

   With the post-selection data matrix $X^{(t)}$, calculate the empirical dual covariance matrix $H^{(t)} = X^{(t)}(X^{(t)})'$. Assume that K is given, then we take $K-1$ eigenvectors with largest $K-1$ eigenvalues. Denote the $n \times (K-1)$ matrix combined by these eigenvectors as U.

4. K-means clustering.

   Apply K-means clustering algorithm for the matrix U we get in last step. It is to find an $n \times (K-1)$ matrix, with K identical rows, with smallest difference to the matrix U in the sense of Frobenious norm. The labels found by K-means are what we need.

3.2.2 Normality for Micro-array Data Sets

In Spectral-HCT algorithm, I assume that the noise is normal distributed. However, is it true for real data sets? If not, we have to consider some other hypothesis testing methods. To solve this problem, I investigate the normality assumption for some micro-array data sets.

For the micro-array data sets not normally distributed, I find Dettling's pre-processing method ([Dettling 2003, Dudoit 2002]). It includes three steps: thresholding the features by some lower bound and upper bound; excluding the features that the maximum and the minimum are too near; and logarithm base 10 transformation. After the pre-processing, most data sets performs as normally distributed. It can be seen from the following figures about histogram and pp-plot for KS statistics of Lung2001 ([Bhattacharjee 2001]) data set and Breast ([Wang 2005]) data set.

With the figures, we can see that the noise distribution can be assumed to be normal. So the normal assumption won't disturb the result much.

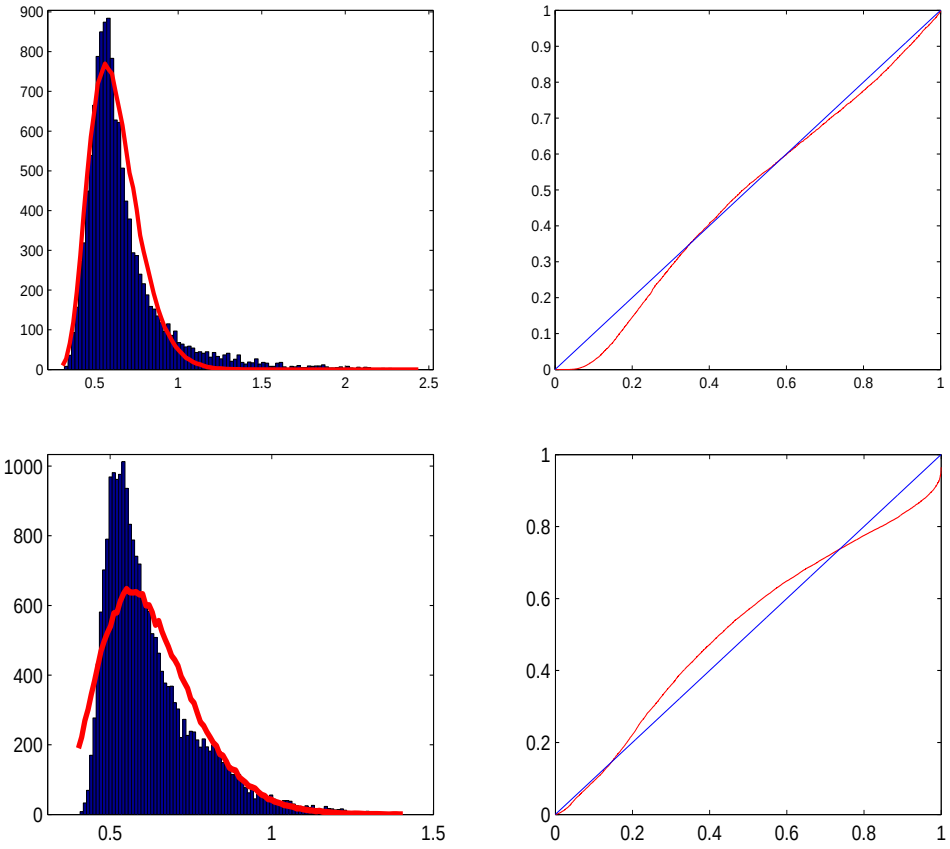

Figure 3.1: Top Left: histogram for Lung2001 and corrected null distribution (red line). Top Right: pp plot for Lung2001 data. Bottom Left: histogram for Breast data and corrected null distribution (red line). Bottom Right: pp plot for Breast data.

3.2.3 Comparison with Other Methods

How does this method work? I compare it with the other methods on different data sets, including Leukemia data, Colon data, Lung cancer data, Brain data, Lymphoma data, SRBCT data I introduced before. What's more, I also compared Spectral-HCT with other clustering methods on Lung2001 data, Breast cancer data, Prostate cancer data ([Singh 2002]), and Su-Cancer data sets ([Su 2001]). A brief introduction of these data sets is in the following table.

Data Name	Source	K	n	p
Brain	Pomeroy (02)	5	42	5597
Breast Cancer	Wang et al. (05)	2	276	22215
Colon Cancer	Alon et al. (99)	2	62	2000
Leukemia	Golub et al. (99)	2	72	3571
Lung Cancer	Gordon et al. (02)	2	181	12533
Lung2001	Bhattacharjee et al. (01)	2	203	12600
Lymphoma	Alizadeh et al. (00)	3	62	4062
Prostate Cancer	Singh et al. (02)	2	136	6033
SRBCT	Kahn (01)	4	63	2308
Su-Cancer	Su et al (01)	2	174	7909

Table 3.1: The brief introduction of data sets.

There are 10 data sets in total, with number of classes from 2 to 5, number of genes ranging from 2000 to 20,000, and year ranging from 1999 to 2005, which is very broad. So, the comparison is very persuading.

For these data sets, I apply Spectral-HCT method, K-means, Hierarchical clustering, and SpectralGem. The result is in Table 3.2. It can be seen that Spectral-HCT works well and improves the results for most data sets. For some data set it does not behave well (such as Colon data), it does not did worse either.

3.2.4 Computation Cost

How about the computation cost? As the HC functional is very easy to compute (computation cost $O(p)$), the only time consuming step is to simulate null distribution of KS statistics. For each data set, we have to simulate 20,000 KS statistics with sample size n. The computation cost for this step is about $O(n)$, yet the constant is large.

How large it is? We record the computation cost for several data sets as Table 3.3. In this table, we can see that it costs less than one minute even for data sets with 12533 genes. It is very fast compared to other methods with feature selection step.

Data Name	K	K-means	Hier	SpectralGem	Spectral-HCT
Brain	5	.433	.500	.436	.425
Breast Cancer	2	.357	.548	.333	.238
Colon Cancer	2	.461	.387	.484	.403
Leukemia	2	.278	.361	.264	.069
Lung Cancer	2	.122	.182	.315	.033
Lung2001	2	.433	.300	.478	.212
Lymphoma	3	.323	.468	.226	.048
Prostate	2	.422	.422	.480	.382
SRBCT	4	.603	.540	.540	.524
Su-Cancer	2	.452	.448	.483	.362

Table 3.2: The clustering error rate of different clustering methods for data sets. In each box, the decimal shows the corresponding error rate.

Data set	Leukemia	Lung	Prostate	Colon
Computation Time	36	49s	48.5s	35s

Table 3.3: Computation cost for Leukemia data, Lung data, Prostate data and Colon data.

Chapter 4

Spectral-HCT Approach

Contents

4.1 Background

In this chapter, I will introduce the elegant theory for Spectral-HCT algorithm in the case of two-class clustering. I will introduce the model I work on, and the main result about upper bound and lower bound for Spectral-HCT algorithm, and then the corresponding phase diagram. Some related topic, such as false discovery rate (FDR) is also discussed. In Section 5.4, I will show my simulation results for Spectral-HCT algorithm under more complicated case. It can be seen that even with more complicated noise and model, Spectral-HCT still improves the result. At last, I will show the proofs.

The paper has contributions in the following perspectives.

- *Higher Criticism for threshold choice in spectral clustering.* We find an intimate relationship between the Hamming error functional of spectral clustering and the recent notion of Higher Criticism. HC was developed earlier in a seemingly unrelated settings (e.g., signal detection [Donoho 2004] and classification [Donoho 2008, Jin 2009]), but the link we forge between HC and spectral clustering is new.

- *Phase transition on PCA.* Our study involves very delicate PCA, where the main goal is to derive an explicit formula for the signal-to-noise ratio $HC(t)$ in (1.2.10). Our study is closely related but is more delicate than that [Paul 2007]. Johnstone and Paul [Paul 2007] considers a setting where we have independent samples from $N(0, \Sigma)$, for which the covariance matrix Σ is "spiky" in the sense that most eigenvalues are 1 except for a few of them are larger than 1. They reveal an interesting phase transition regarding the inner product between the leading eigenvector of the empirical covariance matrix and that of Σ: depending how large the leading eigenvalues of Σ is, the inner product converges to 1 or 0 as p diverges. In comparison, our study in (1.2.10) reveals a similar phase transition on the inner product between the label vector y and the leading eigenvector of $H^{(t)}$. However, to find a good approximation to the Ideal Threshold, the study on the phase transition

is inadequate, and we must derive an explicit formula for the signal-to-noise ratio $HC(t)$, which involves much more delicate analysis.

- *Phase transition in spectral clustering.* Our study reveals interesting phase transition of spectral clustering. In our model $X = y\mu' + z$, we call the $p \times 1$ vector μ the signal vector. If we call the two-dimensional space calibrating the signal sparsity and signal strength, then the phase space partitions into two separate regions, in one of them the signals is sufficiently strong so that successful spectral is possible (say, we use post-selection PCA with the threshold set by HCT), in the other region the signals are merely too rare and weak so spectral clustering must fail. Such a phase transition has been found in signal detection [Donoho 2004], classification [Donoho 2008, Jin 2009], and variable selection [Ji 2010], but not in spectral clustering as far as we know.

4.2 Main results

In this section, we present the main theoretic results, Theorems 4.2.1-4.2.4. In Section 4.2.1, we introduce an asymptotic framework which we call the *Asymptotic Rare and Weak* (ARW) model. Then in Section 4.2.2, we present the main results on post-selection PCA, and introduce the notion of post-selection Signal-to-Noise Ratio (ps-SNR). In Section 4.2.3, we introduce the ideal threshold (the threshold that maximizes the ps-SNR) and the concept of phase diagram. We show that the post-selection PCA yields optimal phase diagram if we set the threshold as the ideal threshold. In Section 4.2.5, we introduce the notion of ideal HCT, the non-stochastic counterpart of HCT, and elaborates the close connection between the ideal HCT and the ideal threshold, and show that spectral-HCT yields the optimal phase diagram.

4.2.1 Rare and weak signal model

Following the discussion in Chapter 1, we consider a model where the data matrix $X = X_{n,p}$ satisfies

$$X = y\mu' + Z, \qquad \text{where } Z = Z_{n,p} \text{ has } i.i.d. \text{ entries from } N(0,1). \tag{4.2.1}$$

Here, μ is the $p \times 1$ signal vector as in Model (1.1.3) and y is the $n \times 1$ adjusted label vector as in (1.1.2).

Under this model, the noise distribution is given, and the most natural summary statistic for each feature is the one that based on the χ^2-statistic. Define

$$T_j = \big(\sum_{i=1}^{n} x_{ij}^2 - n\big)/\sqrt{2n}. \tag{4.2.2}$$

If we let non-central $\chi^2_n(\nu)$ be the χ^2-distribution with $df = n$ and non-central parameter ν, then

$$(\sqrt{2n}T_j + n\big|\mu, y) \sim \chi^2_n(\|y\|^2\mu^2(j)); \tag{4.2.3}$$

note that when $\mu(j) = 0$, $(\sqrt{2n}T_j + n\big|\mu(j) = 0, y) \sim \chi^2_n(0)$.

Fix $\delta \in (0, 1/2)$. We model the label vector ℓ in a way so that $(\ell_i + 1)/2$ are *i.i.d.* samples from Bernoulli(δ). As a result, the adjusted label vector y has independent coordinates that satisfy

$$y_i = \begin{cases} (1-\delta), & \text{with probability } \delta, \\ -\delta, & \text{with probability } (1-\delta); \end{cases} \tag{4.2.4}$$

note that $\|y\|^2/n \approx \delta(1-\delta)$. For large n, by (4.2.3), we have that approximately,

$$T_j \sim N\big(\delta(1-\delta)\sqrt{n/2}\mu^2(j),\ 1\big).$$

In light of this, we re-scale $\mu(j)$ and suppose

$$\sqrt{n/2}\delta(1-\delta)\mu^2(j) \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon F, \tag{4.2.5}$$

where ν_0 is the point mass at 0 and F is a distribution that has no mass at 0, and $\epsilon \in (0, 1)$ is the parameter for signal sparsity.

We adopt a rare and weak signal model where p is the driving asymptotic parameter, and other quantities (e.g., (n, ϵ, F)) are tied to p through fixed parameters. Fixing $\theta \in (0, 1)$ and $\beta \in (0, 1)$ we model the sample size n and the sparsity parameter ϵ by

$$n = n_p = p^\theta, \qquad \epsilon = \epsilon_p = p^{-\beta}. \tag{4.2.6}$$

As p also tends to ∞, n_p tends to ∞ but $n \ll p$, and ϵ_p tends to 0 so that the signals get increasingly sparse, but the number of signals (which is approximately $p\epsilon_p$) tends to ∞.

For signals this sparse, it turns out that the most interesting range for the signal distribution F is that F concentrates its mass at the order of $O(\sqrt{\log(p)})$, a quantity that goes to ∞ but in a very slow rate. The feature selection problem is relatively easy when signals are much stronger than this, but is impossible when signals are much weaker. For this reason, we fix $r \in (0, 1)$ and let

$$\tau_p = \sqrt{2r\log(p)}. \tag{4.2.7}$$

For some of our results (e.g., Theorems 4.2.2-4.2.4), we assume $F = F_p$ is a point mass at τ_p:

$$F = F_p = \nu_{\tau_p}. \tag{4.2.8}$$

This models the situation where all useful features have the same strength. This is not an unusual assumption for literature on phase diagrams and Higher Criticism. Despite

the seemingly simplicity of the model, the analysis it entails is already very subtle, and the insight gained is valid for much broader cases. For some or our result (e.g., Theorems 4.2.1 and 4.2.3, and Section 4.6), we consider a much broader model where we allow unequal signal strengths.

Definition 4.2.1 *We call Model (4.2.1) and (4.2.4)-(4.2.8) the Asymptotic Rare and Weak model ARW*$(\beta, r, \theta, \delta)$.

4.2.2 Post-selection Signal-to-Noise Ratio (SNR)

In this section, we derive the post-selection Signal-to-Noise Ratio (ps-SNR). In the following context, we will talk about ps-SNR only. So we use SNR for short without confusion. To show that the results hold much more generally than the $ARW(\beta, r, \theta, \delta)$, we digress in this section without the assumptions in Model (4.2.4)-(4.2.8).

To apply the post-selection PCA, we first select feature by thresholding T_j with some threshold $t > 0$, where T_j is as in (4.2.2). To this end, denote the *empirical survival function* associated with T_j by

$$\bar{F}_p(t) = \frac{1}{p}\sum_{j=1}^{p} 1\{T_j \geq t\}, \tag{4.2.9}$$

and denote $\widetilde{F}_p(t) = \widetilde{F}_p(t, \mu, y)$ by

$$\widetilde{F}_p(t, \mu, y) = E[\bar{F}_p(t)\big|\mu, y] = \frac{1}{p}\sum_{j=1}^{p} P(T_j \geq t\big|\mu, y). \tag{4.2.10}$$

At the same time, let

$$\widetilde{W}_p(t) = \widetilde{W}_p(t, \mu, y) = \frac{1}{p}\sum_{j=1}^{p} \mu^2(j) P(T_j \geq t\big|\mu, y). \tag{4.2.11}$$

We define the SNR functional by

$$\widetilde{SNR}(t) = \widetilde{SNR}(t, \mu, y, n, p) = \frac{p\widetilde{W}_p(t)}{\sqrt{p\widetilde{F}_p(t)/n + p\widetilde{W}_p(t)}}.$$

At the same time, let $\hat{S}(t)$ be the set of all retained features:

$$\hat{S}(t) = \hat{S}_p(t, X) = \{1 \leq j \leq p : T_j \geq t_p\}.$$

We define $X^{(t)}$ as the $n \times |\hat{S}|$ sub-matrix of X formed by restricting the columns of X to $\hat{S}(t)$. As before, denote the $n \times n$ post-selection empirical dual covariance matrix by

$$H^{(t)} = X^{(t)}(X^{(t)})',$$

and denote leading eigenvector of $H^{(t)}$ by $\xi^{(t)}$. The following theorem is one of the main result of this paper, the underlying idea of which is discussed in detail in Section 4.4.

Theorem 4.2.1 *(Post-selection PCA). Consider a sequence of clustering problems $Y = y\mu' + z$ that satisfy (4.2.1) and a sequence of thresholds t, where as p ranges, $n = n_p$, $t = t_p$ with $a\sqrt{\log(p)} < t_p < \sqrt{2\log(p)}$ for some constant a, $y = y^{(n_p)}$ and $\mu = \mu^{(p)}$ are non-stochastic quantities that depend on p. As $p \to \infty$, if $n_p \to \infty$ and there exists a constant c, such that*

$$p^{-c}\widetilde{SNR}(t_p, \mu^{(p)}, y^{(n_p)}, n_p, p) \to \infty,$$

then with probability at least $1 - o(1/p)$,

$$\xi^{(t_p)} \propto \widetilde{SNR}(t_p, \mu^{(p)}, y^{(n_p)}, n_p, p)[1 + o(1)] \cdot y^{(n_p)} + z^{(n_p)} + rem^{(n_p)},$$

where $z^{(n_p)} \sim N(0, I_{n_p})$, and $\|rem^{(n_p)}\|_2^2/n_p = o(1)$.

4.2.3 Ideal Threshold and success of post-selection PCA

We now move to $ARW(\beta, r, \theta, \delta)$. As $p \to \infty$, some regularity appears, and

$$\widetilde{F}_p(t, \mu, \ell, n_p) \approx \bar{f}_p(t, \epsilon_p, \tau_p, n_p), \qquad \widetilde{W}_p(t, \mu, \ell, n_p) \approx w_p(t, \epsilon_p, \tau_p, n_p),$$

where $\bar{f}_p(t) = \bar{f}_p(t, \epsilon_p, \tau_p, n_p)$ and $w_p(t) = w_p(t, \epsilon_p, \tau_p, n_p)$ ($\bar{f}_p(t)$ is an unfortunate notation and we should not misinterpret it as a density function) are defined by

$$\bar{f}_p(t, \epsilon_p, \tau_p, n_p) = E_{\epsilon_p, \tau_p}[\widetilde{F}_p(t, \mu, y, n_p)], \qquad w_p(t, \epsilon_p, \tau_p, n_p) = E_{\epsilon_p, \tau_p}[\widetilde{W}_p(t, \mu, y, n_p)],$$

and so

$$\widetilde{SNR}(t, \mu^{(p)}, \ell^{(p)}, n_p) \approx \sqrt{p} \cdot \widetilde{snr}(t, \epsilon_p, \tau_p, n_p),$$

where $\widetilde{snr}$ is a non-stochastic function which we call the *pseudo SNR*.

$$\widetilde{snr}(t, \epsilon_p, \tau_p, n_p) = \frac{w_p(t, \epsilon_p, \tau_p, n_p)}{\sqrt{\bar{f}_p(t, \epsilon_p, \tau_p, n_p)/n_p + w_p(t, \epsilon_p, \tau_p, n_p)}}. \tag{4.2.12}$$

Ideally, one would pick the threshold t to maximizes the pseudo SNR. This is the ideal threshold.

Definition 4.2.2 *The ideal threshold is the threshold that maximizes $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$: $t_p^{ideal} = t_p^{ideal}(\epsilon_p, \tau_p, n_p) = \arg\max_{0<t<\infty}\{\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)\}$.*

We now further characterize $\widetilde{snr}(t)$ and t_p^{ideal}. The following notations are frequently used in this paper.

Definition 4.2.3 *$L_p > 0$ denotes a generic multi-$\log(p)$ term which may vary from occurrence to occurrence, and which satisfies that for any $c > 0$, $L_p p^{-c} \to 0$ and $L_p p^c \to \infty$ as $p \to \infty$.*

Definition 4.2.4 *For any $n \geq 1$ and $\nu \geq 0$, $\bar{G}_\nu(t, n) = 1 - G_\nu(t, n)$ denotes the survival function of $(T - n)/\sqrt{2n}$, where for $T \sim \chi^2_n(\sqrt{2n}\nu)$.*

Note that for large n, $\bar{G}_{\tau_p}(t, n)$ is very close to the survival function of $N(\tau_p, 1)$. By direct calculations,

$$\bar{f}_p(t, \epsilon_p, \tau_p, n_p) = (1 - \epsilon_p)\bar{G}_0(t, n_p) + \epsilon_p \bar{G}_{\tau_p}(t, n_p), \tag{4.2.13}$$

and

$$w_p(t, \epsilon_p, \tau_p, n_p) = \frac{\sqrt{2}}{\delta(1-\delta)} \epsilon_p \tau_p n_p^{-1/2} \bar{G}_{\tau_p}(t, n_p). \tag{4.2.14}$$

For any fixed $q > 0$, let

$$t_p(q) = \sqrt{2q \log(p)}. \tag{4.2.15}$$

Plugging (4.2.13)-(4.2.14) into (4.2.12), it follows from basic algebra and the closeness between $\bar{G}_{\tau_p}$ and the survival function of $N(\tau_p, 1)$ that

$$\widetilde{snr}(t_p(q), \epsilon_p, \tau_p, n_p) \equiv \widetilde{snr}(t, \epsilon_p, \tau_p, n_p)\big|_{t=t_p(q)} = L_p p^{-\Delta(q,\beta,r,\theta)},$$

where

$$\Delta(q, \beta, r, \theta) = \begin{cases} \beta - \frac{1}{2}\min\{q, \beta - \theta/2\}, & q \leq r; \\ \beta + (\sqrt{q} - \sqrt{r})^2 - \frac{1}{2}\min\{q, \beta - \theta/2 + (\sqrt{q} - \sqrt{r})^2\}, & q > r. \end{cases} \tag{4.2.16}$$

See $\Delta(q, \beta, r, \theta)$ as a function of q, where (β, r, θ) are fixed. A noticeable feature is that, for some (β, r, θ), the function is flat on the top, so there is an interval over which $\Delta(q, \beta, r, \theta)$ does not depend on q, and remains as a constant (that depends on (β, r, θ)). From a practical point of view, the flatness causes some unappealing features that can be further improved. We address this in Section 4.3.

As a result, let

$$\Delta^*(\beta, r, \theta) = \min_{0<q<\infty} \big\{\Delta(q, \beta, r, \theta)\big\}.$$

It follows that

$$\widetilde{snr}(t_p^{ideal}, \epsilon_p, \tau_p, n_p) = \sup_{0<t<\infty} \widetilde{snr}(t, \epsilon_p, \tau_p, n_p) = L_p p^{-\Delta^*(\beta,r,\theta)}, \tag{4.2.17}$$

and

$$\Delta^*(\beta, r, \theta) = \begin{cases} \beta - r, & r < (\beta - \theta/2)/3, \\ \frac{\theta}{2} + \frac{(\beta - \theta/2 + r)^2}{8r}, & (\beta - \theta/2)/3 < r < \beta - \theta/2, \\ \frac{\theta}{4} + \frac{\beta}{2}, & r > \beta - \theta/2. \end{cases} \tag{4.2.18}$$

At the same time, with basic calculation we have that

$$t_p^{ideal} \sim \sqrt{2q^* \log(p)},$$

where $q^* = q^*(\beta, r, \theta)$ satisfies

$$q^*(\beta, r, \theta) = \begin{cases} 4r, & r < (\beta - \theta/2)/3, \\ \frac{(\beta-\theta/2+r)^2}{4r}, & (\beta - \theta/2)/3 < r < \beta - \theta/2, \end{cases} \tag{4.2.19}$$

and

$$q_-(\beta, r, \theta) \le q^*(\beta, r, \theta) \le q_+(\beta, r, \theta), \qquad \text{if } r > \beta - \theta/2, \tag{4.2.20}$$

where

$$q_-(\beta, r, \theta) = \beta - \theta/2, \qquad q_+(\beta, r, \theta) = r. \tag{4.2.21}$$

By Theorem 4.2.1 and (4.2.17), we expect so see that

$$\widetilde{SNR}(t_p^{ideal}, \mu, \ell, n_p, p) \approx \sqrt{p}\widetilde{snr}(t_p^{ideal}, \epsilon_p, \tau_p, n_p, p) = L_p p^{1/2-\Delta^*(\beta,r,\theta)}.$$

Therefore, the success of post-selection PCA hinges on the positivity of the exponent of

$$1/2 - \Delta^*(\beta, r, \theta),$$

Introduce the *clustering phase function* by

$$\rho_\theta^*(\beta) = \begin{cases} 0, & \beta < 1/2, \\ \beta - 1/2, & \frac{1}{2} < \beta < \frac{3-\theta}{4}, \\ (\sqrt{1-\theta} - \sqrt{1+\theta/2-\beta})^2, & \frac{3-\theta}{4} < \beta < 1 - \frac{\theta}{2}. \end{cases} \tag{4.2.22}$$

By direct calculations, it is seen that

$$1/2 - \Delta^*(\beta, r, \theta) > 0 \qquad \text{if and only if} \qquad r > \rho_\theta^*(\beta).$$

Correspondingly, there is a phase-transition phenomenon associated with the post-selection PCA. These are the following two theorems, which are proved in 4.7.1 - 4.7.14.

Theorem 4.2.2 *Possibility. Under the conditions of ARW model, fix $0 < q < 1$ and suppose $r > \rho_\theta^*(\beta)$. As $p \to \infty$, with probability at least $1 - o(1/p)$,*

$$\xi^{(t_p^{ideal})} \propto L_p p^{1/2-\Delta^*(\beta,r,\theta)} \cdot y^{(n_p)} + z^{(n_p)} + rem^{(n_p)},$$

where $z^{(n_p)} \sim N(0, I_{n_p})$ and is independent of $y^{(n_p)}$, and $\|rem^{(n_p)}\|_2 = o(\sqrt{n_p})$. Consequently, if we use post-selection PCA where the threshold is set as t_p^{ideal}, then the clustering error $\to 0$.

For q in this range, the inner product $(\xi^{(t_p(q))}, y)/\|y\|$ tends to 1 as $p \to \infty$ algebraically fast, and the leading eigenvector $\xi^{(t_p(q))}$ is informative for estimating the adjusted label vector y.

4.2.4 Lower bound for post-selection PCA, phase diagram

Theorem 4.2.3 *Impossibility. Under the conditions of ARW model, fix* $(\theta, \beta, r, q) \in (0,1)^4$ *such that* $r < \rho^*_\theta(\beta)$. *Then as* $p \to \infty$, *with probability* $1 + o(1/n)$, *the inner product of* $\xi^{(t_p(q))}$ *and the adjusted label vector* y *satisfies*

$$(\xi^{(q)}, y)/\|y\| \leq L_p p^{1/2-\Delta(q,\beta,r,\theta)},$$

where $L_p p^{1/2-\Delta(q,\beta,r,\theta)} \to 0$ *for any* q.

This says that when $r < \rho_\theta(\beta)$, the signal strength fall below a critical threshold, and the leading eigenvector of $H^{(t_p(q))}$ is almost orthogonal to the adjusted label vector, and is non-informative for clustering.

Together, Theorems 4.2.2-4.2.3 depict an interesting phase transition: when $r < \rho_\theta(\beta)$, the signals are too rare and weak and post-selection PCA bounds to fail. When $r > \rho_\theta(\beta)$, the signals are strong enough that post-selection PCA can be successful, provided that we pick the thresholds in an appropriate range.

4.2.5 Ideal Higher Criticism Threshold (Ideal HCT)

Ideal HCT is the non-stochastic counterpart of HCT introduced earlier, and is also the threshold that HCT tries to estimate. Recall that $\bar{G}_0(t,n)$ is the survival function of T_j when $\mu(j) = 0$ (see (4.2.2)). Introduce the functional that is defined over all survival function $\bar{H} = 1 - H$ associated with a positive random variable:

$$HC(t,H) = HC(t,H;\bar{G}_0,n_p) = \frac{\bar{H}(t) - \bar{G}_0(t,n)}{\sqrt{\bar{H}(t) + \sqrt{n_p}[\max\{\bar{H}(t) - \bar{G}_0(t,n_p), 0\}]}}. \quad (4.2.23)$$

There are two survival functions that are of particular interest: the empirical survival function associated with T_j, $1 \leq j \leq p$, and its non-stochastic counterpart (both are introduced slightly earlier in Section 4.2.3):

$$\bar{F}_p(t) = \frac{1}{p}\sum_{j=1}^{p} 1\{T_j > t\}, \qquad \bar{f}_p(t) = E[\bar{F}_p(t)]; \quad (4.2.24)$$

recall that $\widetilde{F}_p(t) = E[\bar{F}_p(t)|\mu, y]$ and that

$$\bar{f}_p(t) = E_{\epsilon_p,\tau_p}[\widetilde{F}_p(t,\mu,y,n_p)] = E[\bar{F}_p(t)]. \quad (4.2.25)$$

Evaluating $HC(t,\bar{H})$ at $\bar{H} = \bar{F}_p$ and $\bar{H} = \bar{f}_p$ gives rise to the HCT and Ideal HCT.

Definition 4.2.5 *We call the threshold* t *that maximizes* $HC_p(t,\bar{F}_p)$ *the Higher Criticism Threshold, and denote it by* $\hat{t}_p^{HC}$, *and we call the threshold* t *that maximizes* $HC_p(t,\bar{f}_p)$ *the Ideal Higher Criticism Threshold, and denote it by* $t_p^{idealHC}$.

In disguise, $\hat{t}_p^{HC}$ is the HCT we introduce earlier in Section 1.2.4. See details therein.

We explain why three thresholds we introduce are all close to each other (except for a small probability):

$$\hat{t}_p^{HC} \approx t_p^{idealHC} \approx t_p^{ideal}. \tag{4.2.26}$$

On one hand, if we denote

$$V_p(t) = V_p(t, X, n_p) = \bar{F}_p(t) - \bar{G}_0(t, n_p),$$

then we can rewrite $HC(t, \bar{F}_p)$ and $HC(t, \bar{f}_p)$ by

$$HC(t, \bar{F}_p) = \frac{V_p(t)}{\sqrt{\bar{F}_p(t) + \sqrt{n_p}\max\{V_p(t), 0\}}}.$$

Similarly to discussions in Section 4.2.3, as p grows to ∞,

$$\bar{F}_p(t) \approx \bar{f}_p(t), \qquad V_p(t) \approx v_p(t),$$

where $\bar{f}_p(t)$ is as in (4.2.25) and

$$v_p(t, \epsilon_p, \tau_p, n_p) = E[V_p(t)] = \bar{f}_p(t) - \bar{G}_0(t, n_p) > 0.$$

Therefore, we expect to see that

$$HC(t, \bar{F}_p) \approx \frac{v_p(t)}{\sqrt{\bar{f}_p(t) + \sqrt{n_p} v_p(t)}} \equiv HC(t, \bar{f}_p).$$

and so with overwhelming probability,

$$\hat{t}_p^{HC} \approx t_p^{ideal}. \tag{4.2.27}$$

On the other hand, note that in comparison,

$$\widetilde{snr}(t, \epsilon_p, \tau_p, n_p) = \frac{\sqrt{n_p} w_p(t)}{\sqrt{\bar{f}_p(t) + n_p w_p(t)}}, \qquad HC(t, \bar{f}_p) = \frac{v_p(t)}{\sqrt{\bar{f}_p(t) + \sqrt{n_p} v_p(t)}}.$$

By definition and that for t in the range of interest, $\bar{G}_0(t, n_p) \ll \bar{G}_{\tau_p}(t, n_p)$,

$$v_p(t, \epsilon_p, \tau_p, n_p) = \epsilon_p[\bar{G}_{\tau_p}(t, n_p) - \bar{G}_0(t, n_p)] \approx \epsilon_p \bar{G}_{\tau_p}(t, n_p), \tag{4.2.28}$$

where by recalling that

$$\sqrt{n_p} w_p(t, \epsilon_p, \tau_p, n_p) = \frac{\sqrt{2}}{\delta(1-\delta)} \epsilon_p \tau_p \bar{G}_{\tau_p}(t, n),$$

the right hand side of (4.2.28) is proportional to $w_p(t, \epsilon_p, \tau_p, n_p)$, with a factor of $\frac{\sqrt{2}}{\delta(1-\delta)}\tau_p$. Since the factor $\frac{\sqrt{2}}{\delta(1-\delta)}\tau_p$ only has a secondary effect over the the maximizing point of either $HC(t, \bar{f}_p)$ or $\widetilde{snr}(t)$, so we expect to see that

$$t_p^{idealHC} \approx t_p^{ideal}. \tag{4.2.29}$$

Together, (4.2.27) and (4.2.29) explain (4.2.26).

The closeness of three thresholds suggests that using HCT is a good threshold choice, and spectral-HCT method should be successful. This is captured in the following theorem, which is proved in Section 6.4.

Theorem 4.2.4 *Under the conditions of ARW model and fix the parameters such that $r > \rho_\theta^*(\beta)$, say that the ideal threshold $t_p^{ideal} = \sqrt{2q^{ideal}\log p}$ and the HC thresholding under χ^2 test is $\hat{t}_p^{HC} = \sqrt{2q^{HC}\log p}$, then we have, when $p \to \infty$, with probability at least $1 - o(1/p)$,*

$$\begin{cases} \frac{\hat{t}_p^{HC}}{t_p^{ideal}} \to 1, & r < \beta - \theta/2; \\ \underline{\lim}\, \frac{\hat{t}_p^{HC}}{\sqrt{2q_-\log p}} \geq 1, \ \textit{and} \ \overline{\lim}\, \frac{\hat{t}_p^{HC}}{\sqrt{2q_+\log p}} \leq 1, & r > \beta - \theta/2. \end{cases}$$

Also, under the ARW model, in the successful region for ideal threshold, the error rate with HCT clustering also goes to 0 with probability $1 - o(1/p)$.

4.3 Variants of HCT and connection to FDR methodology

In the preceding section, we see that for (β, r, θ) such that $r > \beta - \theta/2$, then the function $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$ is flat in the interval

$$q_-(\beta, r, \theta) \leq q \leq q_+(\beta, r, \theta);$$

over which, every q attains the maximum of the function. In this case, at least asymptotically, many threshold choices would work almost equally well as does the ideal threshold.

While this is true *asymptotically*, it is worthy to reinvestigate from a practical point of view, where we have finite n and p: among many "equally good" threshold choices (in terms of clustering behaviors), it is still desirable to select the threshold that has the smallest (feature) False Discovery Rate (FDR). Call a feature useful or useless according to $\mu(j) \neq 0$ or $\mu(j) = 0$. The (feature) FDR is the expected fraction of selected features that are useless:

$$FDR = E\left[\frac{\#\{\text{number of selected features that are useless}\}}{\#\{\text{total number of selected features}\}}\right].$$

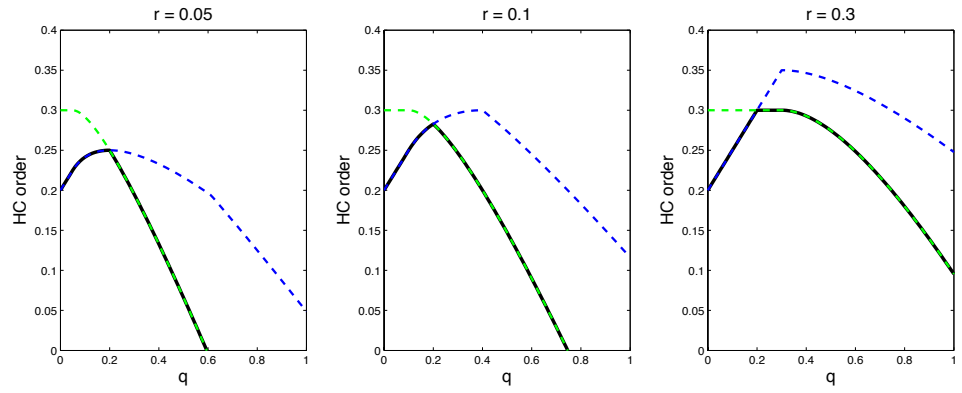

Figure 4.1: The order of HC when q changes. When r is large (in the right figure), there is a flat area between $\beta - \theta/2$ (0.2) and r (0.3).

In light of this, we propose a small variant of HCT. The variant has a similar clustering behavior as the HCT in the preceding sections, but has a FDR that is more appealing from a practical perspective.

To this end, we first investigate where the flatness of the function $\Delta(q, \beta, r, \theta)$ comes from. For any $0 < q < 1$,

$$\widetilde{snr}(t_p(q), \epsilon_p, \tau_p, n_p) = \frac{v_p(t_p(q))}{\sqrt{\bar{f}_p(t_p(q)) + \sqrt{n_p} v_p(t_p(q))}}.$$

By similar calculations, when (β, r, θ) satisfy $r > \beta - \theta/2$, for any fixed q such that $q_-(\beta, r, \theta) < q < q_+(\beta, r, \theta)$,

$$v_p(t_p(q)) \approx \epsilon_p \bar{\Phi}(t_p(q) - \tau_p) = \epsilon_p[1 - L_p p^{-(\sqrt{q}-\sqrt{r})^2}], \qquad \frac{\bar{f}_p(t_p(q))}{\sqrt{n_p} v_p(t_p(q))} = L_p p^{-(q+\theta/2-\beta)};$$

so there is a constant $c(q, \beta, r, \theta) = \min\{(\sqrt{q} - \sqrt{r})^2, q + \theta/2 - \beta\} > 0$ such that

$$\widetilde{snr}(t_p(q), \epsilon_p, \tau_p, n_p) = p^{-(\beta/2+\theta/4)}[1 - L_p p^{-c(q,\beta,r,\theta)}].$$

This says that in the interval of $q_-(\beta, r, \theta) < q < q_+(\beta, r, \theta)$, the effect of q is only on the algebraically small term, and so $\widetilde{snr}(t_p(q))$ is almost flat. See Figure 4.1 for illustration.

It is an easy fix of the flatness if consider the variant of $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$ by

$$\widetilde{\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p) = a(t)\widetilde{snr}(t, \epsilon_p, \tau_p, n_p),$$

where $a(t)$ is monotone function that is slowly increasing for $0 \le t \le \sqrt{2\log(p)}$ such that $a(0) = 1$ and $a(\sqrt{2\log(p)}) = 1 + o(1)$. There are many choices of such function, and a convenient choice is

$$a(t) = 1 - \frac{\log \bar{G}_0(t, n_p)}{\log(p)}. \tag{4.3.30}$$

In comparison, while $\widetilde{\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p) \approx \widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$ for all $0 \le t \le \sqrt{2\log(p)}$, the function $\widetilde{\widetilde{snr}}$ tilde to the right hand slightly, to which the maximizing point is not only much easier to pin down, and is also more robust to noise corruption.

Definition 4.3.1 *We call a threshold* $\arg\max_{t>0}\{\widetilde{\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p)\}$ *the variant of the ideal threshold, and denote it by* $t_p^{idealvariant}$.

Compared to the ideal threshold, the variant is more appealing from a practical perspective, in terms of the FDR. We have the following corollary.

Corollary 4.3.1 *Fix* $(\beta, r, \theta) \in (0,1)^3$. *Under the conditions of ARW*(β, r, θ), *the (feature) FDR associated with the* $t_p^{idealvariant}$ *satisfies*

$$FDR(t_p^{idealvariant}) \sim \begin{cases} \frac{p^{-3r}}{p^{-3r}+p^{-\beta}}, & r < (\beta - \theta/2)/3, \\ \frac{1}{1+p^{-\theta/2}}, & (\beta-\theta/2)/3 < r < \beta - \theta/2, \\ \frac{p^{-r}}{p^{-\beta}+p^{-r}}, & r > \beta - \theta/2. \end{cases}$$

Correspondingly, we can modify HCT slightly so that it gives a good approximation of $t_p^{idealvariant}$. The variant of HCT is defined with three similar steps as before, except the function in step 3 is slightly different.

- Let $\pi_j = \bar{G}_0(T_j, n)$ be the p-values of T_j.
- Let $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$ be the sorted p-values.
- Let $\hat{j}$ be the maximizing index of

$$\frac{\sqrt{p}(j/p - \pi_{(j)})}{\sqrt{\max\{\sqrt{n}(j/p - \pi_{(j)}), 0\} + j/p}}(1 + \frac{\log(1/\pi_{(i)})}{\log p}).$$

The variant of HCT is then defined as $\hat{T}_{\hat{j}}$ that has corresponding p-value $\pi_{\hat{j}}$.

The following theorem is similar to that of Theorem 4.2.4, and is proved in Section 6.4.

Theorem 4.3.1 *Under the conditions of the ARW model and that* $r > \rho_\theta^*(\beta)$, *with probability* $1 - o(1/p^2)$,

$$\frac{\hat{t}_p^{HCvariant}}{t_p^{idealvariant}} \to 1, \qquad p \to \infty;$$

and the error rate of spectral clustering with HCT variant also goes to 0 with probability $1 - o(1/p^2)$.

4.4 Proof of Theorem 4.2.1

One of the major contribution of this paper is Theorem 4.2.1, which consists of delicate analysis on post-selection PCA. In this paper, we illustrate the main ideas underlying Theorem 4.2.1, and provides the proof in the end of the section. The proofs of other theorems are also nontrivial, but are deterred to Section 6.4 for reasons of space and exposition.

We review some notations we have before, and also introduce some new ones. We are interested in Model (4.2.1), where $X = y\mu' + Z$. For any threshold $t > 0$,

$$\hat{S}(t) = \hat{S}(t, X, p) = \{1 \le j \le p : T_j \ge t\} \tag{4.4.31}$$

denotes the set of all the retained features, $X^{(t)}$ and $Z^{(t)}$ denote the $n \times |\hat{S}(t)|$ sub-matrix of X formed by restricting the columns of X to $\hat{S}(t)$, respectively, and $\mu^{(t)}$ denotes the $|\hat{S}(t)| \times 1$ sub-vector of μ formed by restricting the rows of μ to $\hat{S}(t)$. Note that even when μ is non-stochastic, $\mu^{(t)}$ is stochastic for it depends on $\hat{S}(t)$. Note that

$$X^{(t)} = y(\mu^{(t)})' + Z^{(t)}, \tag{4.4.32}$$

We denote

$$H^{(t)} = (X^{(t)})(X^{(t)})', \qquad H_0^{(t)} = (Z^{(t)})(Z^{(t)})',$$

so that $H^{(t)}$ is the $n_p \times n_p$ empirical dual covariance matrix before, and $H_0^{(t)}$ is the empirical dual covariance matrix in the special case of $\mu = 0$.

We are primarily interested in the first leading eigenvector $\xi^{(t)}$ of $H^{(t)} = (X^{(t)})(X^{(t)})'$. To this end, we introduce four stochastic functions $C_{ij}^{(t)}(\lambda) = C_{ij}^{(t)}(\lambda, y, \mu, Z, p, n_p)$, $1 \le i, j \le 2$, defined over over all real values λ that are not an eigenvalue of $H_0^{(t)}$ by

$$C_{11}^{(t)}(\lambda) = C_{11}^{(t)}(\lambda, y, \mu, Z, p, n_p) = y'[\lambda I_{n_p} - H_0^{(t)}]^{-1} y, \tag{4.4.33}$$

$$C_{12}^{(t)}(\lambda, t) = C_{21}^{(t)}(\lambda, y, \mu, Z, p, n_p) = y'[\lambda I_{n_p} - H_0^{(t)}]^{-1} Z^{(t)} \mu^{(t)}, \tag{4.4.34}$$

and

$$C_{22}^{(t)}(\lambda) = C_{22}^{(t)}(\lambda, y, \mu, Z, p, n_p) = (Z^{(t)}\mu^{(t)})'[\lambda I_{n_p} - H_0^{(t)}]^{-1} Z^{(t)} \mu^{(t)}; \tag{4.4.35}$$

where λ is an eigenvalue of $H_0^{(t)}$, we let $C_{ij}^{(t)}(\lambda) = |i - j|$. Moreover, we introduce $h^{(t)}(\lambda) = h(\lambda, y, \mu, Z, p, n_p)$ and $g(\lambda) = g^{(t)}(\lambda, y, \mu, Z, p, n_p)$ by

$$h^{(t)}(\lambda) = C_{11}^{(t)}(\lambda)(\|\mu^{(t)}\|^2 + C_{22}^{(t)}(\lambda)) - (1 - C_{12}^{(t)}(\lambda))^2, \qquad g^{(t)}(\lambda) = \frac{1 - C_{12}^{(t)}(\lambda)}{C_{11}^{(t)}(\lambda)}. \tag{4.4.36}$$

4.4.1 Some useful lemmas

The following lemma characterizes the eigenvalues and eigenvectors of $H^{(t)}$, and is proved in the appendix.

Lemma 4.4.1 *Fix p, $n = n_p$, and $\lambda > 0$. For any λ that $[\lambda I_{n_p} - H_0^{(t)}]$ is non-singular, we have that, with probability 1, it is equivalent with $h^{(t)}(\lambda) = 0$ that λ is an eigenvalue of $H^{(t)}$, and the corresponding eigenvector is*

$$\xi \propto [\lambda I_{n_p} - H_0^{(t)}]^{-1}(g^{(t)}(\lambda)y + Z^{(t)}\mu^{(t)}). \tag{4.4.37}$$

In this paper, we are primarily interested in the largest eigenvalue of $H^{(t)}$. In Lemma 4.4.3, we show that with probability 1, $[\lambda I_n - H_0^{(t)}]$ is non-singular if λ is the largest eigenvalue of $H^{(t)}$ when SNR goes to infinity.

By Lemma 4.4.1, to study the leading eigenvalues/eigenvectors of $H^{(t)}$, the key is to have a good understanding of $h^{(t)}(\lambda)$ and $g^{(t)}(\lambda)$; the analysis it entails is subtle as it involves the post-selection random matrix $Z^{(t)}$. Note that $Z^{(t)}$ is the noise matrix associated with *only* the features that survived the feature selection. As a result, the columns of $Z^{(t)}$ are non-independent and have a distribution that is not easy to analyze. For this reason, existing results on PCA of random matrices (which usually deals with matrices with *i.i.d.* entries) do not directly apply, and we have to develop new techniques. The following lemma plays a key role in our analysis, and is proved in the appendix.

In Chapter 2, Theorem 2.3.1 extends well-known results on random matrices with *i.i.d.* Gaussian entries (e.g., [Vershynin 2010]) to the more difficult case of post-selection random matrices, where the ϵ-net argument in [Vershynin 2010] is very helpful. Theorem 2.3.1 is useful in evaluating many quantities, including the largest eigenvalue of $H_0^{(t)}$, $C_{ij}^{(t)}(\lambda)$, $h^{(t)}(\lambda)$, and $g^{(t)}(\lambda)$. In particular, we have the following lemma, which is proved in Section 6.4.

Note that the leading term is very important here. So we should consider two cases, (a) $Z^{(t)}$ has more rows than columns, which means that $n_p/p\widetilde{F}_p(t) \to \infty$; and (b) $Z^{(t)}$ has more columns than rows, which means that $n_p/p\widetilde{F}_p(t) \to 0$. As we talk about the asymptotic behavior when $p \to \infty$, we mean that there exists a constant c, such that $p^{-c}n_p/p\widetilde{F}_p(t) \to \infty$ when we say $n_p/p\widetilde{F}_p(t) \to \infty$, and it is similar for the case $n_p/p\widetilde{F}_p(t) \to 0$. For case (b), the leading term is $p\widetilde{F}_p(t)$, which is m in Theorem 2.3.1.

There is another term k associated with the singular values, which can be denoted as the features with signals. Even though the term k is not the leading term here, we still need a denotation for it in our case. So we have the following definition.

Definition 4.4.1 *Define $\bar{TP}_p(t) = \frac{1}{p}\sum_{j=1}^{p} 1\{T_j \geq t\}1\{\mu(j) \neq 0\}$, and $\widetilde{TP}_p(t) = \widetilde{TP}_p(t, \mu, y) = E[\bar{TP}_p(t)|\mu, y]$.*

With the lemmas, we have the result for the leading eigenvalue of $H_0^{(t)}$ for the cases as following.

Lemma 4.4.2 *Fixing $0 < q < 1$, let $t_p < \sqrt{2\log(p)}$ and $Z = Z_{n_p,p}$ be the same as that in Theorem 4.2.1. As $p \to \infty$, with probability at least $1 - o(1/p^2)$, there is*

$$\|H_0^{(t_p)}\|_2 \le 2n_p, \qquad n_p/p\widetilde{F}_p(t_p) \to \infty,$$

and, when $n_p/p\widetilde{F}_p(t_p) \to 0$,

$$\|H_0^{(t_p)} - p\widetilde{F}_p(t_p)(1 + c_t\sqrt{\frac{\log(p)}{n}})I_n\|_2 \le 8eK_0\sqrt{np\widetilde{F}_p(t_p)} + \sqrt{\frac{8}{n}}t_p\widetilde{TF}_p(t_p),$$

where $K_0 = 4\sqrt{2\log(p)+1} + 1 + 2\sqrt{\log(p)/n}$, and c_t is a constant that depends on t_p only.

4.4.2 The largest eigenvalue of $H^{(t_p)}$

Fix $0 < q < 1$ and let $t_p < \sqrt{2\log(p)}$ as Theorem 4.2.1. In this section, we characterize the largest eigenvalue of $H^{(t_p)}$, $\lambda_1(H^{(t_p)})$.

Definition 4.4.2 *Fix $n \ge k \ge 1$. For any $n \times n$ symmetric matrix M, $\lambda_k(M)$ denotes the k-th largest eigenvalue of M.*

Take the case $n_p/\widetilde{F}(t_p(q)) \to \infty$ as an example. By (4.4.32), it is seen that

$$H^{(t_p)} = H_0^{(t_p)} + A. \tag{4.4.38}$$

Let $\|\cdot\|_2$ be the spectral norm of matrix, and by triangular inequality we have that

$$\big|\|H^{(t_p)}\|_2 - \|A\|_2\big| \le \|H_0^{(t_p)}\|_2. \tag{4.4.39}$$

On one hand, A is a rank 2 matrix, the eigenvalue of which is easy to derive, which is

$$\lambda_\pm(A) = \frac{\|y\|^2\|\mu^{(t_p)}\|^2}{2}(1 \pm \sqrt{1 + \frac{4y'Z^{(t_p)}\mu^{(t_p)}}{\|y\|^2\|\mu^{(t_p)}\|^2} + \frac{4\|Z^{(t_p)}\mu^{(t_p)}\|^2}{\|y\|^2\|\mu^{(t_p)}\|^4}}) + y'Z^{(t_p)}\mu^{(t_p)}. \tag{4.4.40}$$

In fact, by basic techniques on large-deviation (proved in Section 4.7.6), with probability at least $1 - o(1/p^2)$, there is

$$|y'Z^{(t_p)}\mu^{(t_p)}| \le \sqrt{6}\|y\|\|\mu^{(t_p)}\|^2\log(p). \tag{4.4.41}$$

Combining this with (4.4.40), with probability at least $1 - o(1/p^2)$, we have

$$|\lambda_1(A) - \|y\|^2\|\mu^{(t_p)}\|^2| \le \sqrt{6}\|y\|\|\mu^{(t_p)}\|^2\log(p). \tag{4.4.42}$$

Combining (4.4.39) with (4.4.42) and Lemma 4.4.2, expanding quadratic term and ignoring the lower order terms, it can be derived that with probability at least $1 - o(1/p^2)$,

$$|\lambda_1(H^{(t_p)}) - \|y\|^2 p\widetilde{W}_p(t_p)| \le 2n_p. \tag{4.4.43}$$

Similarly, when $n_p/p\widetilde{F}(t_p) \to 0$, define $Ctr_p(t_p) = Ctr(p, t_p, n_p) = p\widetilde{F}_p(t_p)(1 + c_t\sqrt{\log(p)/n})$ for short, and we have that $H^{(t_p)} = [A + Ctr_p(t_p)I_n] + H_0^{(t_p)} - Ctr_p(t_p)I_n$. With triangle inequality, there is

$$\big|\|H^{(t_p)}\|_2 - \|A + Ctr_p(t_p)I_n\|_2\big| \le \|H_0^{(t_p)} - Ctr_p(t_p)I_n\|_2.$$

Combining with (4.4.42) and Lemma 4.4.2, we have that, with probability at least $1 - o(1/p^2)$,

$$|\lambda_1(H^{(t_p)}) - \|y\|^2 p\widetilde{W}_p(t_p) - Ctr_p(t_p)| \le err_p(t_p), \tag{4.4.44}$$

where $err_p(t_p) = 8eK_0\sqrt{np\widetilde{F}_p(t_p)} + \sqrt{\frac{8}{n}}t\widetilde{TF}_p(t_p)$, with K_0 defined in Lemma 4.4.2 as a constant.

Now, if we restrict λ to the interval

$$\|y\|^2 p\widetilde{W}_p(t_p) + \begin{cases} (-2n_p, 2n_p), & n_p/p\widetilde{F}(t_p) \to \infty, \\ \big(Ctr_p(t_p) - err_p(t_p), Ctr_p(t_p) + err_p(t_p)\big), & n_p/p\widetilde{F}_p(t_p) \to 0 \end{cases}, \tag{4.4.45}$$

when $p\widetilde{W}_p(t_p) \to \infty$, there exists one and only one solution of $h(\lambda)$ with probability 1, which is the leading eigenvalue of $H^{(t_q)}$.

Lemma 4.4.3 *Under the conditions of Theorem 4.2.1, as $p \to \infty$, with probability $1 - o(1/p^2)$, the leading eigenvalue λ has behavior as*

$$\begin{cases} \big|\lambda - \|y\|^2 p\widetilde{W}_p(t_p)\big| \le 2n_p, & n_p/p\widetilde{F}(t_p) \to \infty, \\ \big|\lambda - [Ctr_p(t_p) + \|y\|^2 p\widetilde{W}_p(t_p)]\big| \le err_p(t_p), & n_p/p\widetilde{F}_p(t_p) \to 0. \end{cases}, \tag{4.4.46}$$

where $err_p(t_p) = 8eK_0\sqrt{n_p p\widetilde{F}_p(t_p)} + \sqrt{\frac{8}{n_p}}t\widetilde{TP}_p(t_p)$, with $K_0 = 4\sqrt{2\log(p)} + 2$.

4.4.3 Proof of Theorem 4.2.1

By Lemma 4.4.1, the (first) leading eigenvector satisfies

$$\xi^{(t_p)} \propto [\lambda_1(H^{(t_p)})I_{n_p} - H_0^{(t_p)}]^{-1}(g^{(t_p)}(\lambda)y + Z^{(t_p)}\mu^{(t_p)}).$$

The notations become rather complicate, and it is desirable to use short-hand notations. Fo this end, we write $\lambda_1 = \lambda_1(H^{(t_p)})$, $a = Ctr_p(t_p)1\{p\widetilde{F}_p(t_p) > n_p\}$, $H_0 = H_0^{(t_p)}$, $I_n = I_{n_p}$, and $g(\lambda) = g^{(t_p)}(\lambda)$. Also, we let M be the matrix

$$M = [H_0 - aI_{n_p}]/(\lambda_1 - a), \tag{4.4.47}$$

By our analysis in the previous section, it is seen that the largest eigenvalue of M is $\frac{\lambda_1(H_0)-a}{\lambda_1-a} \leq \frac{\max\{2n_p, err_p(t_p)\}}{\|y\|^2 p\widetilde{W}_p(t_p)}$. When $\sqrt{p\widetilde{F}_p(t_p)/n_p} \ll p\widetilde{W}_p(t_p)$, which indicates that $err_p(t_p) \ll \|y\|^2 p\widetilde{W}_p(t_p)$, the largest eigenvalue of M is much smaller than 1, so we expect to have

$$(\lambda_1 I - H_0)^{-1} = (\lambda_1 - a)^{-1}[I - M]^{-1} \approx (\lambda_1 - a)^{-1}[I + M],$$

In light of this, approximately,

$$\xi^{(t_p)} \propto (\lambda_1 - a)^{-1}[I_n + M][g(\lambda_1)y + Z^{(t_p)}\mu^{(t_p)}] \propto y + I + II,$$

where

$$I = \frac{1}{g(\lambda_1)} Z^{(t_p)}\mu^{(t_p)} + My, \qquad II = \frac{1}{g(\lambda_1)} MZ^{(t_p)}\mu^{(t_p)}.$$

According to our restriction of leading eigenvalue λ, on the interval (4.4.45), when $p \to \infty$, with probability $1 + o(1/p)$, we have that

$$g(\lambda) \sim \frac{\lambda - a}{\|y\|^2}. \tag{4.4.48}$$

Also, we have following lemmas for I and II.

Lemma 4.4.4 *Under the conditions of Theorem 4.2.1, when $\|M\| \leq 1$, let*

$$I = \frac{1}{g(\lambda_1)} Z^{(t_p)}\mu^{(t_p)} + My, \tag{4.4.49}$$

then, with probability $1 - o(1/p)$, each coordinate of I can be decomposed as

$$I_i = \sqrt{\frac{1}{p\widetilde{W}(t_p)} + \frac{\widetilde{F}(t_p)}{\|y\|^2 p\widetilde{W}^2(t_p)}} z + error,$$

where $z \sim N(0,1)$, and $|error|$ is $o(1)$ term compared to the coefficient of z.

Lemma 4.4.5 *Under the conditoin of Theorem 4.2.1, when $\|M\| \leq 1$, let*

$$I = \frac{1}{g(\lambda_1)} Z^{(t_p)}\mu^{(t_p)} + My, \qquad II = (\sum_{i=1}^{\infty} M^i)I.$$

Then with probability $1 - o(1/p)$,

$$\|II\| \leq \sqrt{\frac{n_p \max\{n_p, p\widetilde{F}_p(t_p)\}}{\|y\|^2 p\widetilde{W}_p(t_p)}} \|I\|(1 + o(1)) = o(1) \cdot \|I\|,$$

which is neglible.

Combining the Taylor expansion with Lemma 4.4.4 and Lemma 4.4.5, it turns out that the leading eigenvector of $H^{(t_p)}$ can be decomposed as

$$\xi \propto y + \sqrt{\frac{1}{p\widetilde{W}(t_p)} + \frac{p\widetilde{F}(t_p)}{\|y\|^2 p^2 \widetilde{W}^2(t_p)}}z + error \propto \frac{1}{\sqrt{\frac{1}{p\widetilde{W}(t_p)} + \frac{p\widetilde{F}(t_p)}{\|y\|^2 p^2 \widetilde{W}^2(t_p)}}}y + z + error.$$

It is easy to find that $\frac{1}{\sqrt{\frac{1}{p\widetilde{W}(t_p)}+\frac{p\widetilde{F}(t_p)}{\|y\|^2 p^2 \widetilde{W}^2(t_p)}}} = \frac{p\widetilde{W}(t_p)}{\sqrt{p\widetilde{W}(t_p)+p\widetilde{F}(t_p(q))/n_p}}$, which is SNR. So, Theorem 4.2.1 is proved.

4.5 Simulations

We conducted a simulation study to investigate the numerical performance of spectral clustering with HCT and HCT variant, compared with spectral clustering without feature selection, and hierarchical clustering. The method k-means is not included as it costs too much time when p is large that the computer cannot load it. The thing is the same for hierarchical clustering with feature selection and k-means with feature selection.

In the simulation, we set the parameters (β, θ, δ), signal strength τ and p, and assume $\sigma = 1$ for noise. Then, with the parameters, we do the following steps.

1. Set $n_p = p^\theta$, $\epsilon_p = p^{-\beta}$.
2. Generate $b = (b_1, \cdots, b_p)$ with $b_i \overset{i.i.d.}{\sim} Bernoulli(\epsilon_p)$, and set $\mu = \sqrt{\tau_p/\sqrt{n}}b$.
3. Generate $\ell = (\ell_1, \cdots, \ell_n)$ with $\ell_i \overset{i.i.d.}{\sim} Bernoulli(1-\delta) - (1-\delta)$.
4. Generate $n \times p$ matrix Z, where each column $Z_i \sim N(0, \Omega)$. Set $X = \ell\mu' + Z$, and apply spectral clustering with HCT, HCT variant, without thresholding, and hierarchical clustering to X.
5. Repeat 1 - 4 independently, and record the average Hamming distance.

The simulations contain 4 different experiments, which we will show separately as following.

Experiment 1. In this experiment, we study the effect of μ when it has equal signals and unequal signals. Also, we study the effect of δ. We set the covariance matrix as $\Omega = I_p$. Fix $(p, \theta) = (4 \times 10^4, 0.65)$, and then there is $n = p^\theta \approx 1000$. Let $\beta \in \{0.55, 0.62\}$, and let $\tau \in \{6, 8, 10, 12\}$. The experiment contains 3 sub-experiments 1a-1c.

In Experiment 1a, we set $\delta = 1/2$, which means that the size for two classes is equal. Set the signal μ is a vector with coordinates either 0 or $\sqrt{\tau/\sqrt{n}}$. The average Hamming errors of 50 experiments are reported in Table 4.1.

In Experiment 1b, we still set $\delta = 1/2$, and the setting for μ changes. Let $|n^{1/4}\mu_j| \overset{i.i.d.}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p\nu_F$, where $F \sim 0.8\sqrt{\tau} + 0.2\sqrt{\tau}(1+V/3)$, $V \sim \chi^2_1$. The average Hamming errors of 50 experiments are reported in Table 4.2.

In Experiment 1c, we set $\delta = 1/3$, so the size for two classes are different. Set μ as Experiment 1b. Let $|n^{1/4}\mu_j| \overset{i.i.d.}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p\nu_F$, where $F \sim 0.8\sqrt{\tau} + 0.2\sqrt{\tau}(1+V/3)$, $V \sim \chi^2_1$. The average Hamming errors of 50 experiments are reported in Table 4.3.

The results show that spectral clustering with HCT is much better than spectral clustering without thresholding or hierarchical clustering. Spectral clustering with HCT variant does a bit worse than that with HCT, but also goes to 0 quickly.

	τ	Spectral Clustering	HCT	HCT variant	Hierarchical Clustering
$\beta = 0.55$	6	0.4596	0.4095	0.4148	0.4860
	8	0.3127	0.1573	0.1796	0.4829
	10	0.1787	0.0518	0.0610	0.4872
	12	0.0992	0.0154	0.0187	0.4924
$\beta = 0.62$	6	0.4826	0.4737	0.4739	0.4872
	8	0.4776	0.4148	0.4211	0.4879
	10	0.4768	0.2977	0.2966	0.4862
	12	0.4471	0.1111	0.1162	0.4881

Table 4.1: Comparison of Hamming errors (Experiment 1a).

	τ	Spectral Clustering	HCT	HCT variant	Hierarchical Clustering
$\beta = 0.55$	6	0.3960	0.2339	0.2525	0.4864
	8	0.2184	0.0848	0.0918	0.4867
	10	0.1214	0.0238	0.0244	0.4871
	12	0.0635	0.0049	0.0067	0.4861
$\beta = 0.62$	6	0.4813	0.4656	0.4688	0.4874
	8	0.4717	0.3207	0.3161	0.4857
	10	0.4523	0.1702	0.1654	0.4864
	12	0.3862	0.0669	0.0684	0.4864

Table 4.2: Comparison of Hamming errors (Experiment 1b).

	τ	Spectral Clustering	HCT	HCT variant	Hierarchical Clustering
$\beta = 0.55$	6	0.4405	0.3736	0.3689	0.4886
	8	0.2916	0.1542	0.1682	0.4800
	10	0.1924	0.0741	0.0809	0.4664
	12	0.1268	0.0361	0.0433	0.4419
$\beta = 0.62$	6	0.4823	0.4713	0.4722	0.4863
	8	0.4792	0.4262	0.4207	0.4865
	10	0.4557	0.2346	0.2318	0.4887
	12	0.4475	0.1517	0.1501	0.4903

Table 4.3: Comparison of Hamming errors (Experiment 1c).

Experiment 2. In this experiment, we examine how HCT and the variant of HCT changes with respect to τ. Set $(\beta, \theta) = (0.62, 0.65)$, and take $\tau \in \{7, 11, 16\}$. Let $n^{1/4}\mu_j \overset{i.i.d.}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_{\sqrt{\tau_p}}$, where $\epsilon = p^{-\beta}$. As it is large deviation result, take $p = (4 \times 10^4, 8 \times 10^4, 1.2 \times 10^5, 1.6 \times 10^5)$, and take correpsonding $n_p = p^\theta$. With the parameters, calculate HCT and threshold with variant of HCT for the data, and then divide over $\sqrt{2 \log p}$. Repeat the process 50 times, and take the average of the threshold.

For $\delta = 1/2$ and $\delta = 1/3$, we have the results. The result is as Table 4.4 and 4.5.

	p	4×10^4	8×10^4	1.2×10^5	1.6×10^5
$\tau = 7$	HCT	0.3938	0.3585	0.3694	0.3519
	HCT variant	0.4364	0.3941	0.4252	0.3964
$\tau = 11$	HCT	0.3922	0.3567	0.3793	0.3679
	HCT variant	0.4570	0.4061	0.4201	0.4184
$\tau = 14$	HCT	0.3935	0.4025	0.4075	0.4202
	HCT variant	0.4672	0.4633	0.4626	0.4725

Table 4.4: Comparison of thresholding (Experiment 2, $\delta = 1/2$).

Experiment 3. In this experiment, we study the effect of covariance matrix for Hamming errors. Fix $p = 4 \times 10^4$, and $n_p = p^\theta \approx 1000$. Fix $\beta = 0.6$, and let $\epsilon = p^{-\beta}$. Let $\tau \in \{7, 8, 9, 10, 11, 12, 13, 14\}$. Generate μ as either 0 or $\sqrt{\tau/\sqrt{n_p}}$. Let $\delta = 1/2$. To set the noise to be correlated, we generate indepedent $p \times 1$ normal variables Z_i, and then take $Z_i = AZ_i$, so that $\Sigma = AA'$. For each τ, we set Σ to be tridiagonal covariance matrix Σ and penta-diagonal matrix.

The experiment contains 2 sub-experiments.

In Experiment 3a, we take Σ to be tridiagonal matrix. To make it, we set the main

	p	4×10^4	8×10^4	1.2×10^5	1.6×10^5
$\tau = 7$	HCT	0.3662	0.3884	0.3716	0.3678
	HCT variant	0.4067	0.4080	0.4045	0.3962
$\tau = 11$	HCT	0.3665	0.3807	0.3696	0.3722
	HCT variant	0.4170	0.4180	0.4246	0.4207
$\tau = 14$	HCT	0.4002	0.3821	0.3675	0.3819
	HCT variant	0.4730	0.4239	0.4179	0.4353

Table 4.5: Comparison of thresholding (Experiment 2, $\delta = 1/3$).

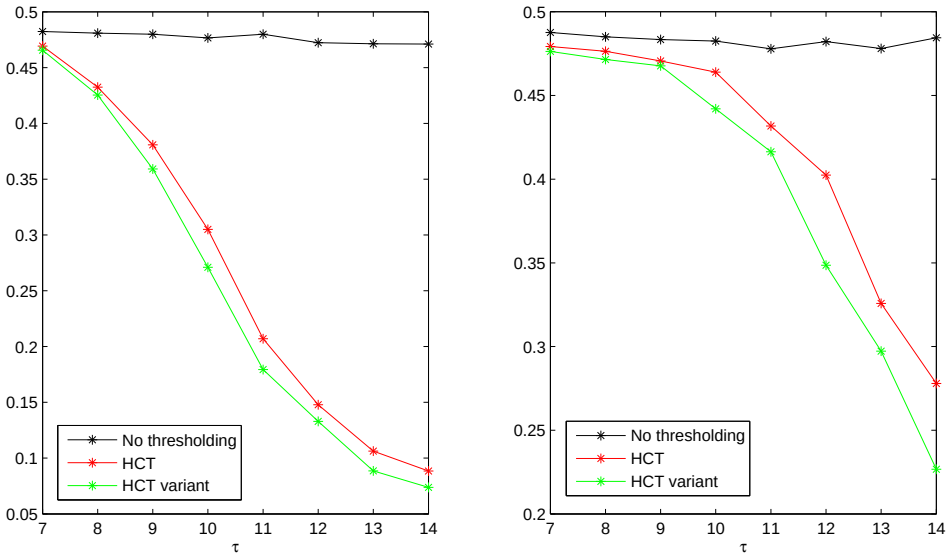

Figure 4.2: Left figure: The average Hamming error for tridiagonal covariance matrix. Right figure: The average Hamming error for penta-diagonal covariance matrix.

diagonal of A as $1/\sqrt{1.04}$, and the first diagonal below main diagonal as $0.2/\sqrt{1.04}$. The denominator is to make sure the noise has variance 1. The average Hamming error over 50 repetitions for spectral clustering without thresholding, with HCT, and the variant of HCT is showed in Figure 4.2.

In Experiment 3b, we take Σ to be penta-diagonal matrix. To make it, we set A as tridiagonal matrix, with $1/\sqrt{1.08}$ on the main diagonal, and $0.2/\sqrt{1.08}$ on the off-diagonals. The average Hamming error over 50 repetitions for spectral clustering without thresholding, with HCT, and variant HCT is showed in Figure 4.2.

Experiment 4. In Experiment 4, we study the behavior of our method for data with noise other than Gaussian noise. As the distribution of noise changes, we cannot use chi-square test to calculate p-values. In this experiment, we apply Kolmogorov-Smirnov test, as following.

1. According to the noise distribution, generate 2×10^6 sample with sample size n, and calculate the corresponding KS value.
2. Calculate the KS value for each feature with data.
3. Find the p-value with simulated KS values. Then apply HC function.

Fix $(p, \theta) = (2 \times 10^4, 0.65)$, and $n_p = p^\theta \approx 600$. Set $\beta = 0.6$, $\delta = 1/3$, and μ as Experiment 1b. Let $\tau \in \{7, 8, 9, 10, 11, 12, 13, 14\}$. As it is hard to take dependent samples for non-Gaussian distributions in matlab, so we assume $\Sigma = I_p$. For each τ, simulate the data, apply our method with KS test, and get the Hamming distance. It contains 3 sub-experiments.

In Experiment 4a, we take the noise as student t distribution, with degree of freedom 5. So the mean is 0, and the variance is 5/3. The left figure in Figure 4.3 reports the average of Hamming distance among 50 repetitions.

In Experiment 4b, we take the noise as $z_{ij} \overset{i.i.d.}{\sim} Exp(1) - 1$. The middle figure in Figure 4.3 reports the average of Hamming distance among 50 repetitions.

In Experiment 4c, we take the noise as $z_{ij} \overset{i.i.d.}{\sim} Unif(-2, 2)$. The right figure in Figure 4.3 reports the average of Hamming distance among 50 repetitions.

The result shows that spectral clustering with HCT/HCT variant is better than the other two methods.

4.6 Discussions and extension

4.6.1 Extension

In the $ARW(\beta, r, \theta, \delta)$ model, we assume that the signal is either a constant or 0. Now, we generalize Model (4.2.1) with the distribution that

$$\mu(j) = \begin{cases} 0, & \text{with Prob 1 - } \epsilon_p, \\ \sim F, & \text{with Prob } \epsilon_p; \end{cases}$$

where F is some distribution on $\mathcal{R}\backslash\{0\}$.

For a better expression, we re-parametrize μ by defining a vector r, with $r(j) = \delta(1-\delta)\mu^2(j)\sqrt{\frac{n}{2\log p}}$. So the distribution F for $\mu(j)$ could be transformed into a distribution for $r(j)$, which is denoted as F_r. The distribution of $r(j)$ is

$$r(j) = \begin{cases} 0, & \text{with Prob 1 - } \epsilon_p, \\ \sim F_r, & \text{with Prob } \epsilon_p. \end{cases} \tag{4.6.50}$$

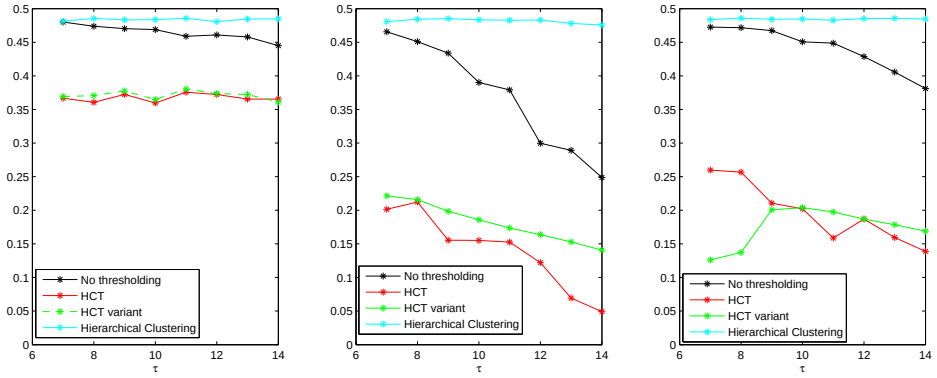

Figure 4.3: Left figure: t noise with df = 5 (Experiment 4a). Middle figure: exponential noise (Experiment 4b). Right figure: Uniform noise (Experiment 4c).

Assume that F_r is some fixed distribution that does not depend on p. The $ARW(\beta, F, \theta, \delta)$ model with (4.6.50) is called *generalized* $ARW(\beta, F, \theta, \delta)$ *model*.

Under the generalized setting, Theorem 4.2.1 still holds. So we need an approximation of $\widetilde{F}_p(t_p)$ and $\widetilde{W}_p(t_p)$ in this case. Note that the expectations are different now, which should be

$$\bar{f}_p(t, \epsilon_p, F_r, n_p) = (1-\epsilon_p)\bar{G}_0(t, n_p) + \epsilon_p \int \bar{G}_{\sqrt{\log(p)}\gamma}(t, n_p) dF_r(\gamma), \tag{4.6.51}$$

$$w_p(t, \epsilon_p, F_r, n_p) = \epsilon_p n_p^{-1/2} \frac{\sqrt{2\log(p)}}{\delta(1-\delta)} \int \gamma \bar{G}_{\sqrt{\log(p)}\gamma}(t, n_p) dF_r(\gamma), \tag{4.6.52}$$

and that $\widetilde{snr}(t, \epsilon_p, F_r, n_p) = \frac{w_p(t,\epsilon_p,F_r,n_p)}{\bar{f}_p(t,\epsilon_p,F_r,n_p)/n_p + w_p(t,\epsilon_p,F_r,n_p)}$. With similar induction in Section 6.2, the threshold that maximizes $\widetilde{snr}(t, \epsilon_p, F_r, n_p)$ approximates the one that maximizes $\widetilde{SNR}(t, \mu^{(p)}, \ell^{(p)}, n_p)$.

Now, if we define HC function as before, the intimate relationship between $\widetilde{snr}(t, \epsilon_p, F_r, n_p)$ and HC function still holds. So, the threshold that maximizes HC function approximates the threshold that maximizes $\widetilde{snr}(t, \epsilon_p, F_r, n_p)$. HCT still achieves the optimal thresholding. According to the behavior of the ideal threshold, we have the following theorem about the phase diagram.

Theorem 4.6.1 *Under the conditions of generalized* $ARW(\beta, F, \theta, \delta)$ *model, set the threshold as* $t_q(p) = q\sqrt{\log p}$*, then we have the following statements:*

- *When the support of* F_r *is bounded above by* $(\rho_\theta^*(\beta))^2/2$*, where* $\rho_\theta^*(\beta)$ *is the clustering phase function defined in (4.2.22), the error rate of spectral clustering with any threshold goes to 1/2;*

- *When the support of F_r is bounded above by some large constant, but with constant positive probability there is $r > (\rho_\theta^*(\beta))^2/2$, then spectral-HCT will succeed with error rate converging to 0 and HCT achieves the ideal threshold;*
- *When the support of F_r could not be bounded above, t^{HC} would give a threshold that spectral clustering approach is successful.*

The theorem states that even in the unequal signal situation, HCT would achieve the optimal threshold when the distribution of signal satisfies some condition. The proof is very similar with the main proof, with an extension to unequal signals, so we do not show it in details here.

4.7 Proofs

4.7.1 Proof of Lemma 4.4.1: Eigenvector

We write for short $H_0 = H_0^{(t)}$, $H = H^{(t)}$, $g(\lambda) = g^{(t)}(\lambda)$, $h(\lambda) = h^{(t)}(\lambda)$, $\mu = \mu^{(t)}$, and $Z = Z^{(t)}$. We now consider the two claims separately.

Consider the first claim that λ is an eigenvalue of $H^{(t)}$ is equivalent with that $h(\lambda) = 0$. With our short-hand notations,

$$H = \|\mu\|^2 yy' + Z\mu y' + y(Z\mu)' + H_0.$$

Since η is an eigenvector of H and λ is the corresponding eigenvalue,

$$\lambda\eta = H\eta = \|\mu\|^2(y,\eta)y + (y,\eta)Z\mu + (\eta, Z\mu)y + H_0\eta, \tag{4.7.53}$$

where $(\cdot,\cdot)$ denotes the inner product. Denote for short

$$a = \|\mu\|^2(y,\eta) + (Z\mu,\eta), \qquad b = (y,\eta). \tag{4.7.54}$$

It follows from (4.7.53) that

$$(\lambda - H_0)\eta = ay + bZ\mu.$$

Now, since that $\lambda I - H_0$ is non-singular, we have

$$\eta = (\lambda I - H_0)^{-1}(ay + bZ\mu). \tag{4.7.55}$$

Introduce the expression of η into the inner product. Using the definitions of $C_{ij}(\lambda)$, it follows from basic algebra that

$$(Z\mu,\eta) = aC_{12}(\lambda) + bC_{22}(\lambda), \tag{4.7.56}$$

and

$$(y, \eta) = aC_{11}(\lambda) + bC_{12}(\lambda). \tag{4.7.57}$$

Combine (4.7.54)-(4.7.57) and re-organize,

$$\begin{cases} [\|\mu\|^2 C_{11}(\lambda) + C_{12}(\lambda) - 1]a + (\|\mu\|^2 C_{12}(\lambda) + C_{22}(\lambda))b = 0, \\ C_{11}(\lambda)a + [C_{12}(\lambda) - 1]b = 0. \end{cases}$$

Since a and b can not equal to 0 simultaneously, the determinant of the following 2 by 2 matrix must be 0:

$$\begin{pmatrix} \|\mu\|^2 C_{11}(\lambda) + C_{12}(\lambda) - 1 & \|\mu\|^2 C_{12}(\lambda) + C_{22}(\lambda) \\ C_{11}(\lambda) & C_{12}(\lambda) - 1 \end{pmatrix}. \tag{4.7.58}$$

By basic algebra, this is equivalent to that of

$$0 = (1 - C_{12}(\lambda))^2 - C_{11}(\lambda)(\|\mu\|^2 + C_{22}(\lambda)), \tag{4.7.59}$$

and the first claim follows by the definition of $h(\lambda)$.

Consider the second claim. By (4.7.58)-(4.7.59), it follows from direct calculations and the definition of $g(\lambda)$ that

$$\begin{pmatrix} a \\ b \end{pmatrix} \propto \begin{pmatrix} [C_{12}(\lambda) - 1]/C_{11}(\lambda) \\ 1 \end{pmatrix} \equiv \begin{pmatrix} g(\lambda) \\ 1 \end{pmatrix}. \tag{4.7.60}$$

Plugging this into (4.7.55), we have that

$$\eta \propto [g(\lambda)y + Z\mu],$$

which gives the second claim. □

4.7.2 Proof of Lemma 4.4.2

There are two cases here, that (a) $n_p/p\widetilde{F}_p(t_p(q)) \to \infty$, and (b) $n_p/p\widetilde{F}_p(t) \to 0$. The proof is different, so we will discuss it case by case.

In the case that $n_p/p\widetilde{F}_p(t) \to \infty$, we need a lemma to bound the eigenvalues of $Z^{(t)}(Z^{(t)})'$, which is the post-selection random matrix. An extension of the random matrix theory could show that,

Lemma 4.7.1 *For an $n_p \times p$ random matrix $Z_{n_p,p}$, where all the entries are i.i.d. standard normal distributed, let Q^S be a sub-matrix of Z that restricted on the set of columns $S = (j_1, j, \cdots, j_m)$. For each sub-matrix Q^S, let $H_0^S = Q^S(Q^S)'$. Given $m < n_p$, for all possible sets S with cardinality $|S| = m$, with probability p^{-m},*

$$\begin{aligned} (\sqrt{n_p} - \sqrt{m} - 2\sqrt{m\log(p)})^2 &\leq \text{ all eigenvalues of } \{H_0^S\}_{|S|=m} \\ &\leq (\sqrt{n_p} + \sqrt{m} + 2\sqrt{m\log(p)})^2. \end{aligned}$$

Now, we introduce in that $\hat{S} = \hat{S}(t) = \{1 \le j \le p : T_j > t\}$, then we have that $|\hat{S}| = p\bar{F}_p(t)$, and $Z^{\hat{S}} = Z^{(t)}$. According to this lemma, when $p\bar{F}_p(t)\log(p) \ll n_p$, for the matrix $Z^{(t)}$, with probability at least $1 - o(p^{-2})$ there is

$$\|H_0^{(t)}\| < (\sqrt{n_p} + \sqrt{p\bar{F}_p(t)} + 2\sqrt{p\bar{F}_p(t)\log(p)})^2 < 2n_p.$$

So we want to prove that the random variable $p\bar{F}_p(t) \ll n_p$ with large probability.

Now we know that $E[p\bar{F}_p(t)] = p\widetilde{F}_p(t) \ll n_p$. For any $b > 0$ and any sequence of independent random variables W_i such that $|W_i| \le b$, $E[W_i] = 0$, and $\mathrm{Var}(W_i) \le \sigma_i^2$ for $1 \le i \le p$. Write for short $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$. Bennett's Lemma [Jin 2012b, Page 38] says that,

$$P(|\sum_{j=1}^{p} |W_j - E[W_j]| \ge s) \le 2\exp(-\frac{c_0}{2\sigma^2}s^2), \qquad \text{if } sb \le \sigma^2,$$

where $c_0 = \psi(1) \approx 0.733$. Applying this with $W_j = 1\{T_j \ge t\} - E[1\{T_j \ge t\}]$, $b = 1$, and $s = 2\sqrt{2\log(p)}\sigma$ and noting that $\sum_{1\le j\le p}\mathrm{Var}(W_j) \le p\widetilde{F}_p(t) = \sigma^2$, with probability at least $1 + o(1/p^2)$,

$$\left|p\bar{F}_p(t) - E[p\bar{F}_p(t)]\right| \le 2\sqrt{2p\widetilde{F}_p(t)\log(p)}. \tag{4.7.61}$$

So, the result follows that $p\bar{F}_p(t) \le p\widetilde{F}_p(t) + 2\sqrt{2p\widetilde{F}_p(t)\log(p)} \le 2p\widetilde{F}_p(t)$. Combine this estimation with the fact that $p\widetilde{F}_p(t) \ll n_p$, and we have that $p\bar{F}_p(t) \ll n_p$. By Lemma 4.7.1, with probability $1 - o(1/p^2)$, the result $\|H_0^{(t_q)}\| \le 2n_p$ follows.

In the case that $n_p/p\widetilde{F}_p(t) \to 0$, Theorem 2.3.1 helps to bound the eigenvalue of column independent random matrix. However, we need to show that the post-selection random matrix $Z^{(t)}$ satisfies the conditions in Theorem 2.3.1, which is shown in Lemma 4.7.2.

Lemma 4.7.2 *Let z be $n_p \times 1$ random vector where $z \sim N(0, I_{n_p})$, and y is $n_p \times 1$ non-stochastic vector with $\|y\|^2 \le \sqrt{2n\log(p)}$. Given threshold $0 < t < \sqrt{2\log(p)}$, define events $A = \{z : \|z\|^2 - n \ge \sqrt{2n}t\}$, and $B = \{z : \|z + y\|^2 - n \ge \sqrt{2n}t\}$.*

Then, for any non-stochastic $n_p \times 1$ vector a with $\|a\| = 1$, $\left[(z'a)^2|A\right]$ is sub-exponential distributed with norm $K \le 4$, and $\left[(z'a)^2|B\right]$ is sub-exponential distributed with norm $K \le 4\sqrt{1 + 2\log(p)}$. Furthermore, as $n_p \to \infty$, we have

$$\begin{cases} E[|z'a|^2|A] = 1 + \sqrt{2/n}\frac{\phi(t)}{\bar{\Phi}(t)}(1 + O(1/\sqrt{n})), \\ 1 \le E[|z'a|^2|B] \le 1 + \sqrt{2/n}\frac{\phi(t)}{\bar{\Phi}(t)}(1 + O(1/\sqrt{n})). \end{cases}$$

Define $\hat{S}$ as the set of selected features, then we have that $Z^{(t)} = Z^{\hat{S}}$. Then $\hat{S} = S_1 \cup S_2$, where $S_1 = \{1 \le j \le p : T_j > t, \mu_j \neq 0\}$ and $S_2 = \{1 \le j \le p : T_j > t, \mu_j = 0\}$. According to Lemma 4.7.2, we have the control on the second moment for the random variables in S_1 and S_2. We also have that $|S_1| = T\bar{P}_p(t)$. Let $m = p\bar{F}_p(t)$, $\eta = \sqrt{2/n_p}t$. According to Theorem 2.3.1, when $p\bar{F}_p(t) \gg n_p$, there is, with probability $1 - o(9^{-n})$,

$$\|Z^{(t)}(Z^{(t)})' - p\bar{F}_p(t)(1 + \sqrt{2/n_p}t)I_{n_p}\| \le 8eK_0\sqrt{n_p m \log 9} + 2\eta T\bar{F}_p(t), \qquad (4.7.62)$$

where $K_0 = 4\sqrt{1 + 2\log(p)} + 1 + t\sqrt{2/n}$.

Combine (4.7.61) with (4.7.62), and note that $c\sqrt{\log p/n_p} = O(\sqrt{\log p/n_p})$, with similar induction we have that, with probability $1 - o(1/p^2)$,

$$\|H_0^{(t)} - p\widetilde{F}_p(t)(1 + O(\sqrt{\frac{\log p}{n_p}}))I_n\| \le 8eK_0\sqrt{n_p p\widetilde{F}_p(t)} + 2\sqrt{2}t\widetilde{TF}_p(t)/\sqrt{n_p}.$$

Combine the two cases, and we have the result.

4.7.3 Proof of Lemma 4.7.1

According to random matrix theory ([Vershynin 2010, Corollary 35]), we have that

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

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

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

Recall that the sub-matrix Q^S is to restrict $Z_{n,p}$ to $S = \{j_1, \cdots, j_m\}$ columns. So, the columns of Q^S is $i.i.d$ normal distributed. Apply with $A = Q^S$, and $N = n$, $n = m$, $u = 2\sqrt{m\log(p)}$, then with probability at least $1 - p^{-2m}$, the eigenvalues of Q_m are between

$$\sqrt{n} - \sqrt{m} \pm 2\sqrt{m\log(p)}.$$

It is easy to find that the number of all the sub-matrices with m columns is less than p^m. All of these sub-matrices have $i.i.d$ standard Gaussian variables, so the lemma about eigenvalue also stands. The probability that the eigenvalue is not in the range $[\sqrt{n} - \sqrt{m} - 2\sqrt{m\log(p)}, \sqrt{n} + \sqrt{m} + 2\sqrt{m\log(p)}]$ is smaller than

$$\sum_{\text{all } Q^S, |S|=m} p^{-2m} \le p^{-m}.$$

It means that, with probability $1 - p^m$, we have

$$\begin{aligned}(\sqrt{n} - \sqrt{m} - 2\sqrt{m\log(p)})^2 &\le \text{ all eigenvalues of } \{H_0^S\}_{|S|=m} \\ &\le (\sqrt{n} + \sqrt{m} + 2\sqrt{m\log(p)})^2.\end{aligned}$$

□

4.7.4 Proof of Lemma 4.7.2

In this proof, we use $\phi(\cdot)$ and $\Phi(\cdot)$ to denote probability density function (pdf) and cumulative density function (cdf) for standard normal distribution. Let $g_a(t,n)$ and $G_a(t,n)$ to denote the pdf and cdf for $(Y-n)/\sqrt{2n}$ where Y is non-central χ^2 distributed random variable with parameter $\sqrt{2n_p}a$. Also, for a random variable X, we use $f_X(s)$ to denote the pdf of X at s.

There are two events here. (a) $A:\{\|z\|^2 \geq n+\sqrt{2nt}\}$, and (b) $B:\{\|z+y\|^2 \geq n+\sqrt{2nt}\}$. We will discuss about them separately.

In case (a), we conditional on $\|z\|^2 \geq n+\sqrt{2nt}$, which means that we cut a ball from the center. To show $(z'a)^2|A$ is sub-exponential distributed, we should start from the pdf of $z'a|A$. To make things easy, take an orthogonal matrix Q, such that the first row of U is a. As $\|a\|=1$, U exists. Let $w=z'U$, then $w_1=z'a$, and $w\sim N(0,I_n)$. As $\|w\|^2=z'UU'z=\|z\|^2$, the event $A=\{z:\|z\|^2\geq n+\sqrt{2nt}\}=\{w:\|w\|^2\geq n+\sqrt{2nt}\}$. So, the pdf for $z'a|A$ is equivalent with $w_1|A$.

The pdf for $w_1|A$ can be written according to the definition of conditional distribution, as

$$f_{w_1|A}(s)=\phi(s)R(s), \tag{4.7.63}$$

where

$$R(s)=P(A|w_1=s)/P(A)=\frac{P(\sum_{i=2}^n w_i^2>n+\sqrt{2nt}-s^2)}{P(A)}.$$

According to Theorem 2.4.3, to show that $(w_1^2|A)$ is sub-exponential distributed, it is sufficient to show that the tail probability for $(w_1^2|A)$ can be controlled by $ce^{-\lambda x}$ for some $\lambda>0$ and $c>0$, which is equivalent with

$$P(|w_1|>s|A)\leq 2e^{-\lambda s^2}, \qquad s>0. \tag{4.7.64}$$

To show (4.7.64), we decompose (4.7.63) into range $|s|<\sqrt{8\log(p)}$ and $|s|>\sqrt{8\log(p)}$.

In the range $|s|<\sqrt{8\log(p)}$, with basic algebra, there is

$$R(s)=1-\frac{1-s^2}{\sqrt{2n}}\frac{g_0(t,n-1)}{\bar{G}_0(t,n)}(1+O(1/\sqrt{n}))\leq 1-\frac{\min\{1-s^2,0\}}{\sqrt{2n}}t\leq e^{s^2/4}.$$

Combining with (4.7.63), we have that

$$f_{w_1|A}(s)\leq\sqrt{1/(2\pi)}e^{-s^2/4}, \qquad n\to\infty. \tag{4.7.65}$$

In the range $|s|>\sqrt{8\log(p)}$, there is $R(s)\leq 1/P(A)\leq t\sqrt{2\pi}e^{t^2/2}$ with basic algebra. As $|t|^2<2\log(p)<s^2/2-\log(p)$, so we have

$$R(s)\leq t\sqrt{2\pi}e^{s^2/4}p^{-1}\leq e^{s^2/4}, \qquad p\to\infty.$$

Combining with (4.7.63), we have that

$$f_{w_1|A}(s) \leq \sqrt{1/(2\pi)}e^{-s^2/4}, \qquad |s| > \sqrt{8\log(p)}. \tag{4.7.66}$$

Combining (4.7.65) and (4.7.66), there is $f_{w_1|A}(s) \leq \sqrt{2/(3\pi)}e^{-s^2/3}$. With basic calculation, we have the tail probability as

$$P(|w_1|^2 \geq s|A) \leq \sqrt{2}e^{-s/4}, \qquad s > 0.$$

According to Theorem 2.4.3, we have that $[w_1^2|A]$ is sub-exponential distributed, with the norm $\|w_1^2\|_{\psi_1} \leq 2$.

Next, we calculate the second moment of $w_1|A$. According to Lemma 4.7.8, we have that $g_0(t, n_p) = \phi(t)(1 + O(1/\sqrt{n}))$. Let $T \sim N(0, 1)$, then we have that

$$E[T|T > t] = \frac{\phi(t)}{\bar{\Phi}(t)}(1 + o(1)). \tag{4.7.67}$$

Combining with $f_Y(t) = \phi(t)(1 + O(1/\sqrt{n}))$, we have

$$E[\|w\|^2|A] = n + \sqrt{2n}\frac{\phi(t)}{\bar{\Phi}(t)}(1 + O(1/\sqrt{n})).$$

As $w|A$ is symmetric for every direction, the conditonal expectation of w_1^2 is $1/n$ fraction of $E[\|w\|^2|A]$. It means that, for any a, we have

$$E[\|z'a\|^2|A] = 1 + \sqrt{2/n}\frac{\phi(t)}{\bar{\Phi}(t)}(1 + O(1/\sqrt{n})). \tag{4.7.68}$$

In case (b), $\|z + y\|^2 \geq n + \sqrt{2nt}$, a ball with center y is cut off.

We have to estimate the pdf for $z'a|B$ first. For any a with $\|a\| = 1$, we construct an orthogonal matrix U with the first row as a. Let $w = z'U$, and $v = (z + y)'U$, then $\|v\|^2 = \|w + y\|^2$. Also, $w \sim N(0, I_n)$, and $v \sim N(y'U, I_n)$. Now, we want to find distribution and second moment of $w_1|B$, where $B = \{z : \|z + y\|^2 > n + \sqrt{2nt}\} = \{w : \|w + y'U\|^2 > n + \sqrt{2nt}\}$.

With basic statistics, the conditional pdf for w_1 is

$$f_{w_1|B}(s) = \phi(s)R(s), \tag{4.7.69}$$

where

$$R(s) = \frac{P(\sum_{i=2}^n v_i^2 > n + \sqrt{2nt} - (s + y'a)^2)}{P(B)}.$$

Still, we decompose it into $|s| > \sqrt{8\log(p)}$ and $|s| < \sqrt{8\log(p)}$.

In the range $|s| < \sqrt{8\log(p)}$, with basic algebra, there is

$$R(s) = 1 - \frac{1 - s^2 - 2y'as}{\sqrt{2n}}\frac{g_{\|y\|^2/\sqrt{2n}}(t, n-1)}{\bar{G}_{\|y\|^2/\sqrt{2n}}(t, n)}(1 + O(1/\sqrt{n})) \leq e^{s^2\sqrt{\frac{\log(p)}{n}}} + 2\sqrt{2}s\log(p).$$

Combining with (4.7.69), we have that

$$f_{w_1|B}(s) \leq \frac{1}{\sqrt{2\pi}} e^{-s^2/4} + 2\log(p)\sqrt{1/\pi} e^{-s^2/4}, \qquad n \to \infty. \tag{4.7.70}$$

In the range $|s| > \sqrt{8\log(p)}$, there is $R(s) \leq 1/P(B) \leq t\sqrt{2\pi}e^{t^2/2}$ with basic algebra. As $|t|^2 < 2\log(p) < s^2/2 - \log(p)$, so we have

$$R(s) \leq t\sqrt{2\pi}e^{s^2/4}p^{-1} \leq 1/\sqrt{2\pi}e^{s^2/4}, \qquad p \to \infty.$$

Combining with (4.7.69), we have that

$$f_{w_1|B}(s) \leq \sqrt{1/(2\pi)}e^{-s^2/4}, \qquad |s| > \sqrt{8\log(p)}. \tag{4.7.71}$$

Combining (4.7.70) and (4.7.71), there is $f_{w_1|B}(s) \leq (1 + 2\sqrt{2}\log(p))/\sqrt{(2\pi)}e^{-s^2/4}$. With basic calculation, we have the tail probability as

$$P(|w_1|^2 \geq s|B) \leq (\sqrt{2} + 4\log(p))e^{-s/4}, \qquad s > 0.$$

According to Theorem 2.4.3, we have that $[w_1^2|B]$ is sub-exponential distributed. With basic calculation, the sub-exponential norm is $\|w_1^2\|_{\psi_1} \leq 4\sqrt{2\log(p)+1}$.

Now, we go on to calculate the second moment of $w_1|B$. Without loss of generality, we suppose $y'a > 0$. What's more, we suppose that $y'a > \log(n)$. In the case that $y'a \leq \log(n)$, a simplification of our proof would work. Note that $R(s) < 1$ when $-y'a - \sqrt{(y'a)^2+1} \leq s \leq -y'a + \sqrt{(y'a)^2+1}$. On the left it is at the order of $y'a$, and on the right it is at constant order.

Now, we decompose the region into two parts, $|s| \leq \log(n)$ and $|s| > \log(n)$.

In the case that $|s| < \log(n)$, we have that $s^2 \leq sy'a \ll \sqrt{n}$. According to Taylor expansion,

$$R(s) = 1 - \frac{1 - s^2 - 2sy'a}{\sqrt{2n}} \frac{g_{\|y\|^2}(t)}{\bar{\Phi}(t - \|y\|^2/\sqrt{2n})}(1 + O(1/n^{1/4})). \tag{4.7.72}$$

Combine (4.7.69) and (4.7.72) on the region $|s| < \log(n)$, with that $g_{\|y\|^2}(t) = \phi(t - \|y\|^2/\sqrt{2n})(1 + O(1/n^{1/4}))$,

$$\int_{-\log n}^{\log n} s^2\phi(s)R(s)ds = 1 + 2\frac{1}{\sqrt{2n}} \frac{\phi(t - \|y\|^2/\sqrt{2n})}{\bar{\Phi}(t - \|y\|^2/\sqrt{2n})}(1 + O(1/n^{1/4})) + o(n^{-c}), \text{for any } c. \tag{4.7.73}$$

In the case that $|s| > \log n$, note that the supereme of $R(s)$ is $\frac{1}{\bar{\Phi}(t-\|y\|^2/\sqrt{2n})}$, and the minimum is larger than $1/2$ by calculation. The integration of $s^2\phi(s)$ over $|s| > \log n$ goes to 0 faster than n^{-c} for any c. So both $\frac{1}{\bar{\Phi}(t-\|y\|^2/\sqrt{2n})}$ and $1/2$ times the integration also go to 0 faster than any n^{-c}.

Combine the two cases, the approximation of $E[(z'a)^2|B] = 1 + 3\frac{1}{\sqrt{2n}}\frac{\phi(t-\|y\|^2/\sqrt{2n})}{\bar{\Phi}(t-\|y\|^2/\sqrt{2n})}(1+O(1/n^{1/4}))$. Using Mill's ratio, there is,

$$1 \le E[(z'a)^2] \le 1 + \frac{2|t-\|y\|^2/\sqrt{2n}|}{\sqrt{2n}} \le 1 + \sqrt{2/n}\frac{\phi(t)}{\bar{\Phi}(t)}.$$

So, the claim follows. □

4.7.5 Proof of (4.4.40)

Lemma 4.7.3 *Let $A = \|\mu^{(t)}\|^2 yy' + Z^{(t)}\mu^{(t)}y' + y(Z^{(t)}\mu^{(t)})'$, then the eigenvalues of A are*

$$\lambda_\pm(A) = \frac{\|y\|^2\|\mu^{(t)}\|^2}{2}(1 \pm \sqrt{1 + \frac{4y'Z\mu}{\|y\|^2\|\mu^{(t)}\|^2} + \frac{4\|Z^{(t)}\mu^{(t)}\|^2}{\|y\|^2\|\mu^{(t)}\|^4}}) + y'Z^{(t)}\mu^{(t)}.$$

Proof. To simplify the notations, we use Z, μ instead of $Z^{(t)}$ and $\mu^{(t)}$.

Note that A is the matrix expanded by $Z\mu$ and y, so there should be two eigenvectors as linear combination of y and $Z\mu$, and all the other eigenvectors are with eigenvalue 0.

Assume the eigenvector is $\xi = ay + bZ\mu$, with eigenvalue λ. So there is $A\xi = \lambda\xi = \lambda ay + \lambda bZ\mu$. Introduce $\xi = ay + bZ\mu$ and A, and we have

$$\begin{aligned} A\xi &= \Big[a\|y\|^2\|\mu\|^2 + ay'Z\mu + b\|\mu\|^2 y'Z\mu + b\|Z\mu\|^2\Big]y \\ &\quad + \Big[a\|y\|^2 + by'Z\mu\Big]Z\mu. \end{aligned}$$

Compare $A\xi$ with $\lambda\xi$, and we get that a system of equations as

$$\begin{pmatrix} \|y\|^2\|\mu\|^2 + y'Z\mu - \lambda & \|\mu\|^2 y'Z\mu + \|Z\mu\|^2 \\ \|y\|^2 & y'Z\mu - \lambda \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

As it is impossible that a and b equal to 0 simultaneously, the determinant of coefficient matrix must be 0, which is equivalent with that

$$(y'Z\mu - \lambda)^2 + (y'Z\mu - \lambda)\|y\|^2\|\mu\|^2 - \|y\|^2\|\mu\|^2 y'Z\mu - \|y\|^2\|Z\mu\|^2 = 0.$$

Elementary calculations show that

$$\lambda_\pm(A) = \frac{\|y\|^2\|\mu\|^2}{2}(1 \pm \sqrt{1 + \frac{4y'Z\mu}{\|y\|^2\|\mu\|^2} + \frac{4\|Z\mu\|^2}{\|y\|^2\|\mu\|^4}}) + y'Z\mu.$$

□

4.7.6 Behavior of $y'Z\mu$

Lemma 4.7.4 *With threshold $t = t_p$, where $c\sqrt{\log(p)} < t_p < \sqrt{2\log(p)}$ for some constant c, $Z^{(t)}$, $\mu^{(t)}$ and y are the post-selection random matrix and vector as we defined. When $\|\mu^{(t)}\| \to \infty$, with probability $1 - o(1/p^2)$, there is*

$$|y'Z^{(t)}\mu^{(t)}| \le \sqrt{6}\|y\|\|\mu^{(t)}\|^2\log(p).$$

Proof. To show the claim, it is suffcient to show that for any given μ, y and t_p, with probability at least $1 - o(1/p^2)$, there is

$$|y'Z^{(t)}\mu^{(t)}| \le \sqrt{6}\|y\|\|\mu^{(t)}\|^2\log(p). \tag{4.7.74}$$

To show (4.7.74), it is sufficient to show that, with probability at least $1 - o(1/p^2)$, there is

$$|y'Z^{(t)}\mu^{(t)} - E[y'Z^{(t)}\mu^{(t)}]| \le 6\|\mu^{(t)}\|\|y\|\sqrt{\log(p)}, \tag{4.7.75}$$

and that

$$|E[y'Z^{(t)}\mu^{(t)}]| \le \frac{\|\mu^{(t)}\|^2\|y\|^2}{\sqrt{n}}\sqrt{6}\log(p) \tag{4.7.76}$$

Combine with that $\|\mu^{(t)}\| \to \infty$, and the claim follows.

To prove (4.7.75) and (4.7.76), we start with the random variable $[Z'y/\|y\||T_j > t]$. Let $\tilde{y} = y/\|y\|$, and construct an $n \times n$ orthogonal matrix U with the first row as $\tilde{y}$. Let $w = Z_j'U$, so $w_1 = Z_j'\tilde{y}$, and $w_i \overset{i.i.d}{\sim} N(0,1)$. With basic algebra, the conditional distribution for w_1 is that

$$f(w_1 = s|\|Z_j + \mu_j y\| > n + \sqrt{2n}t) = \phi(s)R(s), \tag{4.7.77}$$

where

$$R(s) = \frac{\bar{G}_0(\sqrt{\frac{n}{n-1}}t - \frac{1}{\sqrt{2(n-1)}} - \frac{(s+\mu_j\|y\|)^2}{\sqrt{2(n-1)}}, n-1)}{\bar{G}_{\mu_j^2\|y\|^2/\sqrt{2n}}(t,n)}.$$

To solve it, we decompose the region into $|s| > \sqrt{6\log(p)}$ and $|s| < \sqrt{6\log(p)}$.

When $|s| \ge \sqrt{6\log(p)}$, then $R(s) \le p$ and the integration of $\phi(s)$ is smaller than p^{-3}. So the integration of w_1 is smaller than $p^q p^{-3} = o(1/p^2)$.

When $|s| \le \sqrt{6\log(p)}$, with boundary on $R(s)$, we have that

$$\begin{aligned} \tfrac{1}{\sqrt{2\pi}} - \phi(\sqrt{6\log(p)}) &\le \textstyle\int_0^{\sqrt{6\log(p)}} s\phi(s)R(s)ds \\ &\le (\tfrac{1}{\sqrt{2\pi}} - \phi(\sqrt{6\log(p)}))(1 + \tfrac{\mu_j\|y\|}{\sqrt{2n}}2\sqrt{3\log(p)}). \end{aligned} \tag{4.7.78}$$

and

$$\begin{aligned} -[\tfrac{1}{\sqrt{2\pi}} - \phi(\sqrt{6\log(p)})] &\le \textstyle\int_{-\sqrt{6\log(p)}}^0 s\phi(s)R(s)ds \\ &\le -(\tfrac{1}{\sqrt{2\pi}} - \phi(\sqrt{6\log(p)}))(1 - \tfrac{\mu_j\|y\|}{\sqrt{2n}}2\sqrt{3\log(p)}). \end{aligned} \tag{4.7.79}$$

Combine (4.7.78)-(4.7.79) with the analysis, we have that,

$$|E[Z_j'\tilde{y}|T_j > t]| \leq \frac{|\mu_j|\|y\|}{\sqrt{2n}} 2\sqrt{3}\log(p). \tag{4.7.80}$$

Consequently, we have that

$$|E[y'Z^{(t)}\mu^{(t)}]| \leq \frac{\|\mu^{(t)}\|^2\|y\|^2}{\sqrt{2n}} 2\sqrt{3}\log(p).$$

So, (4.7.76) is proved.

Now we want to prove that $[y'Z^{(t)}\mu^{(t)}]$ is near to $E[y'Z^{(t)}\mu^{(t)}]$ with large probability. Let ξ be $p \times 1$ vector, where $\xi_j = [Z_j'\tilde{y}|T_j > t] - E[Z_j'\tilde{y}|T_j > t]$. We show that ξ_j is sub-Gaussian distributed, and then use the tail probability of sub-Gaussian random variables to control.

With basic algebra, we find that

$$E[e^{\xi_j^2/4}] \leq 2.$$

So, ξ is sub-Gaussian distributed with $E[e^{t\xi_j}] \leq e^{9t^2/2}$ according to Theorem 2.4.2. With Theorem 2.4.1, we could get that $\sum_{T_j>t}\xi_j\mu_j$ is also sub-Guassian distributed, with parameter $3\|\mu^{(t)}\|$.

According to the property of sub-Gaussian random variables in Theorem 2.4.2, there is

$$P(|\tilde{y}'Z^{(t)}\mu^{(t)} - E[\tilde{y}'Z^{(t)}\mu^{(t)}]| > \lambda) \leq 2\exp(-\frac{\lambda^2}{18\|\mu^{(t)}\|}).$$

Take $\lambda = 6\|\mu^{(t)}\|\sqrt{\log(p)}$, then with probability at least $1 - o(1/p^2)$, we have

$$|\tilde{y}'Z^{(t)}\mu^{(t)} - E[\tilde{y}'Z^{(t)}\mu^{(t)}]| \leq 6\|\mu^{(t)}\|\|y\|\sqrt{\log(p)}.$$

So, the claim follows. □

4.7.7 Proof of $g^{(t)}(\lambda)$

Recall that for any t,

$$g^{(t)}(\lambda) = \frac{1 - C_{12}^{(t)}(\lambda)}{C_{11}^{(t)}(\lambda)},$$

where λ is any λ in (4.4.46). Also, recall that for short,

$$a_p(t) = a_p(t, \widetilde{F}_p) = Ctr_p(t_p)1\{p\widetilde{F}_p(t) > n_p\}.$$

Lemma 4.7.5 *Let $t = t_p$, where $c\sqrt{\log(p)} \le t_p \le \sqrt{2\log(p)}$ with some constant c. As $p \to \infty$, when $\widetilde{SNR}(t_p, \mu, y, n, p) \to \infty$, with probability at least $1 + o(1/p^2)$,*

$$g^{(t)}(\lambda)\big|_{t=t_p} = (1 + o(1))[(\lambda - a_p(t))]/\|y\|^2, \tag{4.7.81}$$

where $o(1) \to 0$ uniformly for all $\lambda \in \|y\|^2 p\widetilde{W}_p(t) + a_p(t) \pm (-err_p(t), err_p(t))$ as (4.4.46), with

$$err_p(t) = \begin{cases} 2n_p, & n/p\widetilde{F}_p(t) \to \infty, \\ 8eK_0\sqrt{np\widetilde{F}_p(t)} + \sqrt{8/n}t\widetilde{TP}_p(t), & n/p\widetilde{F}_p(t) \to 0. \end{cases} \tag{4.7.82}$$

By simple calculation, to show the claim, it is sufficient to show that with probability at least $1 - o(p^{-2})$,

$$\|C_{11}^{(t)}(\lambda) - (\lambda - a_p(t))^{-1}\|y\|^2\| \le (\lambda - a_p(t))^{-2} err_p(t)\|y\|^2, \tag{4.7.83}$$

and

$$\|C_{12}^{(t)}(\lambda)\| \le \frac{1}{\lambda - a_p(t)}\|y\|p\widetilde{W}_p(t)\log p \sim \frac{\log p}{\|y\|}. \tag{4.7.84}$$

In fact, recall that by $\|y\|^2 p\widetilde{W}_p(t) \to \infty$, for all λ in (4.4.46),

$$\lambda - a_p(t) \sim \|y\|^2 p\widetilde{W}_p(t).$$

So once (4.7.83)-(4.7.84) are proved, then by basic algebra, with probability at least $1 - o(1/p^2)$, uniformly for all λ in (4.4.46),

$$\big|(\lambda - a_p(t))^{-1} g^{(t)}(\lambda)\|y\|^2 - 1\big| \le \max\{\frac{\log p}{\|y\|}, \frac{err_p(t)}{\|y\|^2 p\widetilde{W}_p(t)}\}.$$

As $\widetilde{SNR}(t, \mu, y, n, p) \to \infty$, which means that $p\widetilde{W}_p(t) \to \infty$ and that $p\widetilde{W}_p(t)/\sqrt{p\widetilde{F}_p(t)/n} \to \infty$. So, when $n/p\widetilde{F}_p(t) \to 0$, $err_p(t)/\|y\|^2 p\widetilde{W}_p(t) = C/(p\widetilde{W}_p(t)) \to 0$; when $n/p\widetilde{F}_p(t) \to \infty$, $err_p(t)/\|y\|^2 p\widetilde{W}_p(t) = C_1\sqrt{p\widetilde{F}_p(t)/n}/p\widetilde{W}_p(t) + C_2/\|y\|^2 \to 0$. Then the claim follows.

We now show (4.7.83) and (4.7.84). Since the proofs are similar, we only show the first one. As in (4.4.47), we let $M = [H_0^{(t)} - a_p(t)I_n]/(\lambda - a_p(t))$. By definitions and basic algebra,

$$C_{11}^{(t)}(\lambda) = y'[\lambda I_n - H_0^{(t)}]^{-1} y = (\lambda - a_p(t))^{-1} y'[I_n - M]^{-1} y.$$

By the definition of M and the result of Lemma 4.4.2 about $\|H_0 - aI_p\|$, with probability at least $1 - o(1/p^2)$, there is

$$\|M\| \le \begin{cases} 2n_p/(\lambda - a_p(t)), & n_p/p\widetilde{F}_p(t) \to 0, \\ \big(8eK_0\sqrt{2np\widetilde{F}_p(t)\log(p)} + \sqrt{\frac{8}{n}}t\widetilde{TF}_p(t)\big)/(\lambda - a_p(t)), & n_p/p\widetilde{F}_p(t) \to \infty. \end{cases} \tag{4.7.85}$$

Note that by $\lambda - a_p(t) \sim \|y\|^2 p\widetilde{W}_p(t)$,

$$\|M\| \leq err_p(t)/\|y\|^2 p\widetilde{W}_p(t).$$

At the same time, by basic algebra, for any $n \times n$ matrix (non-stochastic) A such that $\|A\| < 1$,

$$\|I_n - [I_n - A]^{-1}\| \leq \|A\|/(1 - \|A\|).$$

Applying this with $A = M$ gives that with probability at least $1 - o(1/p^2)$,

$$\|I_n - [I_n - M]^{-1}\| \leq \|M\|/(1 - \|M\|) \lesssim \frac{err_p(t)}{\|y\|^2 p\widetilde{W}_p(t)}.$$

Combining with the condition that $p\widetilde{W}_p(t) \to \infty$ and $\frac{\sqrt{p\widetilde{F}_p(t)/n_p}}{p\widetilde{W}_p(t)} \to 0$ and the claim follows. □

4.7.8 Proof of Lemma 4.4.4

In the proof, we use t as shorthand for $t_p(q)$. The goal is to find the marginal distribution of the coordinates of $v = \frac{1}{g^{(t)}(\lambda)} Z^{(t)}\mu^{(t)} + My$. We discuss for the cases (a) $p\widetilde{F}_p(t)/n_p \to 0$, and (b) $p\widetilde{F}_p(t)/n_p \to \infty$, separately. The arguments are similar, so we focus on the first case, and keep it very brief on the second case.

Consider (a). Recall that

$$Z = [z_1, z_2, \ldots, z_p], \qquad \text{and} \qquad Z' = [Z_1, Z_2, \ldots, Z_n].$$

Fix $1 \leq i \leq n$. Let η be the $p \times 1$ (random) vector such that $\eta_j = [\|y\|^2\mu_j + y'Ze_j]1\{T_j > t\}$. By definitions and Lemma 4.7.5, the i-th coordinate of v is

$$v_i = \frac{1}{\lambda}\sum_{j=1}^{p} Z_{ij} \cdot [\|y\|^2\mu_j + y'Ze_j]1\{T_j > t\} \equiv \frac{1}{\lambda}\eta' Z_i.$$

To deal with the weak dependence between η and Z, we use decoupling technique. Let $\tilde{Z}$ be the $n \times p$ matrix formed by replacing the i-th row of Z by a $p \times 1$ vector that is distributed as $N(0, I_p)$ and that is independent of Z. Let $\tilde{T}$ and $\tilde{\eta}$ be the counterpart of T and η, respectively, where Z is replaced by $\tilde{Z}$ but (y, μ) are not changed. Define

$$\tilde{v}_i = \frac{1}{\lambda}\sum_{j=1}^{p} Z_{ij} \cdot [\|y\|^2\mu_j + y'\tilde{Z}e_j]1\{\tilde{T}_j > t\} \equiv \frac{1}{\lambda}\tilde{\eta}' Z_i.$$

Now, first, by definition, $\tilde{\eta}$ is independent of Z_i, and so

$$\tilde{\eta}' Z_i/\|\tilde{\eta}\| \sim N(0, 1).$$

Second, by elementary large deviation results, with probability at least $1 - o(1/p^2)$,

$$\|\tilde{\eta}\|^2 = (1 + o(1))(\|y\|^2 p\widetilde{W}_p(t) + p\widetilde{F}_p(t))\|y\|^2.$$

Compare these with the claim, all we need to show is, with probability $1 - o(1/p^2)$,

$$\lambda|v_i - \tilde{v}_i| \le o(1)(\|y\|\sqrt{\|y\|^2 p\widetilde{W}_p(t) + p\widetilde{F}_p(t)}). \tag{4.7.86}$$

Now, by definitions,

$$\tilde{\eta}_j - \eta_j = (\|y\|^2\mu_j + y'\tilde{Z}e_j)(1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}) + y_i(\tilde{Z}_{ij} - Z_{ij})1\{T_j \ge t\},$$

and so

$$\lambda(\tilde{v}_i - v_i) = \|y\|^2 \cdot Ia + Ib + y_i \cdot Ic,$$

where

$$Ia = \sum_{j=1}^{p} Z_{ij}\mu_j(1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}),$$

$$Ib = \sum_{j=1}^{p} Z_{ij}y'\tilde{Z}e_j(1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}),$$

and

$$Ic = \sum_{j=1}^{p} Z_{ij}(\tilde{Z}_{ij} - Z_{ij})1\{T_j \ge t\},$$

Consider Ia first. Note that μ is sparse, and we decompose the signal μ into the set that $S_1 = \{j : |\mu_j| > c(\log(p)^2/n)^{1/4}\}$ and that $S_2 = \{j : |\mu_j| \le c(\log(p)^2/n)^{1/4}\}$. Then we have that

$$|Ia| \le \sum_{j\in S_1} |Z_{ij}\mu_j||1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}| + \sum_{j\in S_2} |Z_{ij}\mu_j||1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}|.$$

With probability at least $1 - p^{-2}$, $\max_{1\le i\le n, 1\le j\le p}\{|Z_{ij}|\} \le 3\sqrt{\log(p)}$, then we have that

$$|Ia| \le \sum_{j\in S_2}^{p} |Z_{ij}\mu_j||1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}| \le 3\log(p)/n^{1/4} \sum_{\mu_j\ne 0} |1\{\tilde{T}_j \ge t\} - 1\{T_j \ge t\}|. \tag{4.7.87}$$

By definitions and elementary statistics, there is an event with probability at least $1 - o(1/p^2)$ over which

$$|\tilde{T}_j - T_j| \le (Z_{ij}^2 + \tilde{Z}_{ij}^2)/\sqrt{n} \le 4\log(p)/\sqrt{n}.$$

It follows that over the event,

$$|1\{\tilde{T}_j \geq t\} - 1\{T_j \geq t\}| \leq 1\{|\tilde{T}_j - t| \leq 4\log(p)/\sqrt{n}\}, \tag{4.7.88}$$

Inserting this into (4.7.87), with probability at least $1 - o(1/p^2)$,

$$|Ia| \leq C[\log^2(p)/\sqrt{n}] \cdot \sum_{j:\mu_j \neq 0} 1\{|\tilde{T}_j - t| \leq 4\log(p)/\sqrt{n}\}. \tag{4.7.89}$$

Note that the right hand side is distributed as Binomial(m, δ_n), with $m = |\{1 \leq j \leq n : \mu_j \neq 0\}| \sim p\epsilon_p$, and $\delta_n = P(|\tilde{T} - t| \leq 4\log(p)/\sqrt{n}) \sim f_{\tilde{T}}(t)8\log(p)/\sqrt{n}$, where $f_{\tilde{T}}(t)$ is the pdf of $\tilde{T}$ at t. Combining this with (4.7.89), it follows from elementary statistics that with probability at least $1 - o(1/p^2)$,

$$|Ia| \leq C[(\log(p))^2/\sqrt{n}] \cdot (m\delta_n + \sqrt{m\delta_n \log p}).$$

Now, in our notation, $p\widetilde{W}_p(t) \geq (c/\log(p))m\delta_n$. Therefore, with probability at least $1 - o(1/p^2)$,

$$|Ia| \leq 32(\log(p))^2/\sqrt{n}p\widetilde{W}_p(t). \tag{4.7.90}$$

Consider Ib. Let ω be a $p \times 1$ vector with $\omega = \tilde{Z}'y$. It is seen that

$$Ib = \sum_{j=1}^{p} Z_{ij}\omega_j(1\{\tilde{T}_j \geq t\} - 1\{T_j \geq t\}). \tag{4.7.91}$$

Since $\omega/\|y\| \sim N(0, I_p)$, so with probability at least $1 - o(1/p)$,

$$\max_{1\leq j\leq p} |\omega_j| \leq 2\sqrt{\log(p)}. \tag{4.7.92}$$

Note that with probability at least $1 - p^{-2}$, $\max_{1\leq i\leq n, 1\leq j\leq p}\{|Z_{ij}|\} \leq 3\sqrt{\log(p)}$. Combining these, with probability $1 - o(1/p)$,

$$|Ib| \leq \sum_{j=1}^{p} |Z_{ij}\omega_j||1\{\tilde{T}_j \geq t\} - 1\{T_j \geq t\}| \leq \sum_{j=1}^{p} 6\log(p)|1\{\tilde{T}_j \geq t\} - 1\{T_j \geq t\}|.$$

Combine with (4.7.88), with probability $1 - o(1/p)$,

$$|Ib| \leq 12(\log(p))^2 p\widetilde{F}_p(t)/\sqrt{n}. \tag{4.7.93}$$

Consider Ic. Similarly, let ξ be the $p \times 1$ vector such that $\xi_j = Z_{ij}1\{T_j \geq t\}$. It is seen

$$Ic = \xi'\tilde{Z}_i - \|\xi\|^2. \tag{4.7.94}$$

Since ξ is independent of $\tilde{Z}_i$.

$$\xi'\tilde{Z}_i/\|\xi\| \sim N(0, 1),$$

so with probability at least $1-o(1/p^2)$,

$$|\xi' \tilde{Z}_i| \le 2\sqrt{\log(p)}\|\xi\| \tag{4.7.95}$$

Combining (4.7.94) and (4.7.95) gives that with probability $1-o(1/p^2)$,

$$|Ic| \le \|\xi\|(\|\xi\| + 2\sqrt{\log(p)}).$$

We now bound $\|\xi\|$. Introduce event $A = \{|Z_{ij} \le 3\sqrt{\log(p)}\}$. It is seen that $P(A^c) \le 1/p^2$. Letting $Z^*_{ij} = Z_{ij} \cdot 1\{|Z_{ij}| \le 3\sqrt{\log(p)}\}$, then over the event A,

$$\|\xi\|^2 = \sum_{j=1}^{p} (Z^*_{ij})^2 1\{|T_j| \ge t\}.$$

Now, for any $b > 0$ and any sequence of independent random variables W_i such that $|W_i| \le b$, $E[W_i] = 0$, and $\mathrm{Var}(W_i) \le \sigma_i^2$ for $1 \le i \le p$. Write for short $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$. Bennett's Lemma [Jin 2012b, Page 38] says that,

$$P(|\sum_{j=1}^{p} |W_j - E[W_j]| \ge s) \le 2\exp(-\frac{c_0}{2\sigma^2}s^2), \qquad \text{if } sb \le \sigma^2,$$

where $c_0 = \psi(1) \approx 0.733$. Applying this with $W_j = (Z^*_{ij})^2 1\{T_j \ge t\} - E[(Z^*_{ij})^2 1\{T_j \ge t\}]$, $b = 9\sqrt{\log(p)} + 1$, and $s = 2\sqrt{2\log(p)}\sigma$ and noting that $\mathrm{Var}(W_j) = L_p p^{-q}$, with probability at least $1 + o(1/p^2)$,

$$\sum_{j=1}^{p} (Z^*_{ij})^2 1\{|T_j| \ge t\} \le E[\sum_{j=1}^{p} (Z^*_{ij})^2 1\{T_j \ge t\}] + 2\sqrt{6p\widetilde{F}_p(t)\log(p)}.$$

Combining this with the expectation, with probability at least $1 + o(1/p^2)$,

$$\|\xi\|^2 \le p\widetilde{F}_p(t) + 2\sqrt{6p\widetilde{F}_p(t)\log(p)},$$

and so

$$|Ic| \le p\widetilde{F}_p(t) + 6\sqrt{2p\widetilde{F}_p(t)\log(p)}. \tag{4.7.96}$$

Finally, combining (4.7.90), (4.7.93) and (4.7.96) gives

$$|\lambda(\tilde{v}_i - v_i)| \le cp\widetilde{F}_p(t) + c\sqrt{n}(\log p)^2 p\widetilde{W}_p(t).$$

Note that in our case, there is $p\widetilde{F}_p(t)/n_p \to 0$, and $p\widetilde{F}_p(t)/n/\sqrt{p\widetilde{W}_p(t)} \to 0$. With basic calculation, the right hand side is smaller order term of $(\|y\|\sqrt{\|y\|^2 p\widetilde{W}_p(t) + p\widetilde{F}_p(t)})$. Combining with (4.7.86), it gives the claim. □

4.7.9 Proof of Lemma 4.4.5

Note that

$$II = (\sum_{i=1}^{\infty} M^i)I.$$

By the definition of M and the result of Lemma 4.4.2 about $\|H_0 - a_p(t)I_p\|$, with probability at least $1 - o(1/p^2)$, there is

$$\|M\| \leq err_p(t)/(\lambda_1 - a_p(t)),$$

where

$$S(t) = S(t, p, n_p, \mu) = \begin{cases} 2n_p, & n_p/p\widetilde{F}_p(t) \to 0, \\ 4eK_0\sqrt{2n_p p\widetilde{F}_p(t)\log(p)} + \sqrt{\frac{8}{n}}t\widetilde{TP}_p(t), & n_p/p\widetilde{F}_p(t) \to \infty. \end{cases}$$

Combining with Lemma 4.4.3 about λ_1, with probability $1 - o(1/p^2)$, there is

$$\|M\| \leq err_p(t)/\|y\|^2 p\widetilde{W}_p(t).$$

We have shown that when $\widetilde{SNR}(t, \mu, y, n_p, p) \to \infty$, $err_p(t)/\|y\|^2 p\widetilde{W}_p(t) \to 0$. Then $\|M\|$ goes to 0 with probability $1 - o(1/p^2)$.

At the same time, by basic algebra, for any $n \times n$ matrix (non-stochastic) A such that $\|A\| < 1$,

$$\|\sum_{i=1}^{\infty} A^i\| \leq \|A\|/(1 - \|A\|).$$

Applying this with $A = M$ gives that with probability at least $1 - o(1/p^2)$,

$$\|\sum_{i=1}^{\infty} M^i\| \leq \|M\|/(1 - \|M\|) \lesssim \frac{err_p(t)}{\|y\|^2 p\widetilde{W}_p(t)},$$

and that

$$\|II\| \leq \frac{\sqrt{err_p(t)}}{\sqrt{\|y\|^2 p\widetilde{W}_p(t)}}\|I\|,$$

and $err_p(t) \to 0$. So the claim follows. □

4.7.10 Proof about $\widetilde{snr}$ for ARW model

Lemma 4.7.6 *Under the conditions of ARW$(\beta, r, \theta, \delta)$ model, given threshold $t = \sqrt{2q\log p}$, where $0 < q < 1$ is a constant, then with probability $1 - o(1/p^2)$, there*

is

$$\left|\widetilde{F}_p(t,\mu,y,n_p) - \bar{f}_p(t,\epsilon_p,\tau_p,n_p)\right| \leq (2\sqrt{\frac{\log(p)}{p\epsilon_p}} + c(\delta)\frac{\log^2(p)}{\sqrt{n}})\bar{f}_p(t,\epsilon_p,\tau_p,n_p), \quad (4.7.97)$$

$$\left|\widetilde{W}_p(t,\mu,y,n_p) - w_p(t,\epsilon_p,\tau_p,n_p)\right| \leq (2\sqrt{\frac{\log(p)}{p\epsilon_p}} + c(\delta)\frac{\log^2(p)}{\sqrt{n}})w_p(t,\epsilon_p,\tau_p,n_p), \quad (4.7.98)$$

where $c(\delta)$ is a non-stochastic term that depends on δ only.

The proof for $\widetilde{F}_p(t,\mu,y,n_p)$ and $\widetilde{W}_p(t,\mu,y,n_p)$ is quite similar. So we will show the proof for $\widetilde{F}_p(t,\mu,y,n_p)$ only.

Let $\tilde{f}_p(t,\epsilon_p,\tau_p,n_p) = E_{\epsilon_p,\tau_p,y}[\widetilde{F}_p(t,\mu,y,n_p)]$. With basic algebra, if we have that, with probability $1-o(1/p^2)$,

$$|\widetilde{F}_p(t,\mu,y,n_p) - \tilde{f}_p(t,\epsilon_p,\tau_p,n_p)| \leq 2\sqrt{\frac{\log(p)}{p\epsilon_p}}\tilde{f}_p(t,\epsilon_p,\tau_p,n_p), \quad (4.7.99)$$

and that

$$|\tilde{f}_p(t,\epsilon_p,\tau_p,n_p) - \bar{f}_p(t,\epsilon_p,\tau_p,n_p)| \leq c(\delta)\frac{\log^2(p)}{\sqrt{n}}\bar{f}_p(t,\epsilon_p,\tau_p,n_p), \quad (4.7.100)$$

the claim follows.

The first approximation can be found by large deviation result. Let $k = \sum_{j=1}^{p} 1\{\mu_j \neq 0\}$, then $k \sim Binomial(p,\epsilon_p)$, and

$$\widetilde{F}_p(t,\mu,y,n_p) = \frac{k}{p}P(T_j > t|\mu_j = u,y) + (1-\frac{k}{p})P(T_j > t|\mu_j = 0,y). \quad (4.7.101)$$

Introduce that $E[k] = p\epsilon_p$, we have

$$\tilde{f}_p(t,\epsilon_p,\tau_p,n_p) = \epsilon_p P(T_j > t|\mu_j = u,y) + (1-\epsilon_p)P(T_j > t|\mu_j = 0,y). \quad (4.7.102)$$

Combining with Berstein's inequality for binomial random variables ([Lugosi 2004, Page 12]), applying with binomial random variable k with parameter p and ϵ_p, we have that

$$P(|\frac{k-E[k]}{p}| > \lambda) \leq 2\exp(-\frac{p\lambda^2}{2[\epsilon_p(1-\epsilon_p)+2\lambda/3]}).$$

Take $\lambda = 2\sqrt{\epsilon_p\log(p)/p}$, when $\frac{1}{p} < \epsilon_p < 1/2$, which means that $0 < \beta < 1$, $2\lambda/3 \ll \epsilon_p(1-\epsilon_p)$. Applying with $E[k] = p\epsilon_p$, we have that, with probability at least $1-o(1/p^2)$, there is

$$|k - p\epsilon_p| < 2\sqrt{p\epsilon_p\log(p)}.$$

Combining with the definition of $\widetilde{F}_p(t,\mu,y,n_p)$ and $\tilde{f}_p(t,\epsilon_p,\tau_p,n_p)$ in (4.7.101) - (4.7.102), with probability at least $1-o(1/p^2)$,

$$|\widetilde{F}_p(t,\mu,y,n_p)-\tilde{f}_p(t,\epsilon_p,\tau_p,n_p)|\leq 2\sqrt{\frac{\log(p)}{p\epsilon_p}}\tilde{f}_p(t,\epsilon_p,\tau_p,n_p). \tag{4.7.103}$$

The second approximation comes from the large deviation result about $\|y\|^2$. Recall that

$$\bar{f}_p(t,\epsilon_p,\tau_p,n_p)=E_{\epsilon_p,\tau_p}[\widetilde{F}_p(t,\mu,y,n_p)]=\epsilon_p\bar{G}_0(t,n_p)+(1-\epsilon_p)\bar{G}_{\tau_p}(t,n_p). \tag{4.7.104}$$

Compare with $\tilde{f}_p(t,\epsilon_p,\tau_p,n_p)$, the difference here is between the term $\bar{G}_{\tau_p}(t,n_p)$ and $P(T_j>t|\mu_j=u,y)$. Given y and μ_j, there is $T_j\sim\chi^2_n(\mu_j^2\|y\|^2)$. So we care about $\|y\|^2$.

Let $l=\sum_{i=1}^n 1\{y_i=1-\delta\}$, then $l\sim Binomial(n,\delta)$, and $\|y\|^2=n\delta^2+l(1-2\delta)$. With Bernstein's inequality for binomial random variables ([Lugosi 2004, Page 12]), applying with random variable l with parameter n and δ, with $0<\delta<1/2$. We have that

$$P(\left|\frac{l-E[l]}{n}\right|>\lambda)\leq 2\exp(-\frac{n\lambda^2}{2[\delta(1-\delta)+2\lambda/3]}).$$

Take $\lambda=2\sqrt{\delta(1-\delta)\log(p)/n}$, then with probability at least $1-o(1/p^2)$, there is

$$|l-E[l]|<2\sqrt{\delta(1-\delta)n\log(p)}.$$

Combining with basic algebra, with probability at least $1-o(1/p^2)$, there is

$$\left|\|y\|^2-E[\|y\|^2]\right|<2(1-2\delta)\sqrt{\delta(1-\delta)n\log(p)}. \tag{4.7.105}$$

With basic algebra, Combining with (4.7.105), with probability at least $1-o(1/p^2)$, there is

$$|P(T_j>t|\mu_j=u,y)-P(T_j>t|\mu_j=u,\|y\|^2=n\delta(1-\delta))|\leq\frac{16(1-2\delta)}{\sqrt{\delta(1-\delta)}}\frac{\log^2(p)}{\sqrt{n}}\bar{\Phi}_{t-\tau_p}(t,n_p). \tag{4.7.106}$$

Note that $\bar{G}_{\tau_p}(t,n_p)=\bar{\Phi}_{t-\tau_p}(t,n_p)(1+O(1/n^{1/4}))$. Combining with the definition of $\tilde{f}_p(t,\epsilon_p,\tau_p,n_p)$ and $\bar{f}_p(t,\epsilon_p,\tau_p,n_p)$, and take $c(\delta)=\frac{16(1-2\delta)}{\sqrt{\delta(1-\delta)}}$, then with probability $1-o(1/p^2)$, there is

$$|\tilde{f}_p(t,\epsilon_p,\tau_p,n_p)-\bar{f}_p(t,\epsilon_p,\tau_p,n_p)|\leq c(\delta)\frac{\log^2(p)}{\sqrt{n}}\bar{f}_p(t,\epsilon_p,\tau_p,n_p). \tag{4.7.107}$$

So, the claim follows. □

4.7.11 Proof about $\Delta(q, \beta, r, \theta)$

Define

$$(\sqrt{q} - \sqrt{r})_+ = \max\{\sqrt{q} - \sqrt{r}, 0\}.$$

When $t_p(q) = \sqrt{2q\log(p)}$, where $0 < q < 1$, with basic algebra, there is $\bar{G}_0(t_p(q), n_p) = \bar{\Phi}(t)(1 + O(1/\sqrt{n}))$. For non-central chisquare distribution with parameter $\lambda = \sqrt{2n}\tau_p$, there is $\bar{G}_{\tau_p}(t_p(q), n_p) = \bar{\Phi}(t-\tau_p)(1+o(\log^2(p)/n^{1/4}))$. So we have that

$$\bar{f}_p(t, \epsilon_p, \tau_p, n_p) = (1-\epsilon_p)\bar{\Phi}(t)(1+O(1/\sqrt{n}))+\epsilon_p\bar{\Phi}(t-\tau_p)(1+o(\log^2(p)/n^{1/4})), \quad (4.7.108)$$

and that

$$w_p(t, \epsilon_p, \tau_p, n_p) = \epsilon_p\tau_p^2\bar{\Phi}(t-\tau_p)/\sqrt{n}(1 + o(\log^2(p)/n^{1/4})). \quad (4.7.109)$$

Using Mill's ratio, we have that

$$\bar{f}_p(t, \epsilon_p, \tau_p, n_p) = L_p(1-\epsilon_p)p^{-q} + L_p p^{-\beta}p^{-(\sqrt{q}-\sqrt{r})_+^2}, \quad (4.7.110)$$

and

$$w_p(t, \epsilon_p, \tau_p, n_p) = L_p p^{-\beta}p^{-(\sqrt{q}-\sqrt{r})_+^2}p^{-\theta/2}. \quad (4.7.111)$$

Introduce them into $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$, and we get $\Delta(q, \beta, r, \theta)$. □

4.7.12 Proof of $q^*(\beta, r, \theta)$

According to different relationship between r, β, and θ, the function about $\Delta(q, \beta, r, \theta)$ is different. So we have to discuss it in two cases: (a) $r < \beta - \theta/2$, and (b) $r > \beta - \theta/2$.

In case (a), $r < \beta - \theta/2$, so when $q < r$, there is $q < \beta - \theta/2$. As q increases, there are 3 stages for $\Delta(q, \beta, r, \theta)$, which is

$$\Delta(q, \beta, r, \theta) = \begin{cases} -\beta + q/2, & q < r, \\ -\beta + r - \frac{1}{2}(\sqrt{q} - 2\sqrt{r})^2, & r < q < \frac{(\beta-\theta/2+r)^2}{4r}, \\ -\frac{1}{2}[\beta + (\sqrt{q} - \sqrt{r})^2] - \theta/4, & q > \frac{(\beta-\theta/2+r)^2}{4r}. \end{cases}$$

In the first stage, $\Delta(q, \beta, r, \theta)$ keeps increasing as q increases. In the third stage, $\Delta(q, \beta, r, \theta)$ keeps decreasing as q increases. So the maximum will be achieved at the second stage or the endpoints of the second stage.

Study the function at the second stage, that $-\beta + r - \frac{1}{2}(\sqrt{q} - 2\sqrt{r})^2$ for $q > 0$. As q increases, the function keeps increasing till $q = 4r$, and then decreases. As the restriction is that $r < q < \frac{(\beta-\theta/2+r)^2}{4r}$, we have to study whether $q = 4r$ is in this

range or not. If $q = 4r$ is not in this range, then the maximum will be achieved at the endpoint $q = \frac{(\beta-\theta/2+r)^2}{4r}$. With basic algebra, there are two cases as following,

$$q^* = \begin{cases} 4r, & r < (\beta - \theta/2)/3, \\ \frac{(\beta-\theta/2+r)^2}{4r}, & (\beta - \theta/2)/3 < r < \beta - \theta/2. \end{cases} \tag{4.7.112}$$

In case (b), $r > \beta - \theta/2$. For this case, there are also 3 stages for $\Delta(q, \beta, r, \theta)$, which is a little different at the second stage, as following,

$$\Delta(q, \beta, r, \theta) = \begin{cases} -\beta + q/2, & q < r, \\ -\beta/2 - \theta/4, & \beta - \theta/2 < q < r, \\ -\frac{1}{2}[\beta + (\sqrt{q} - \sqrt{r})^2] - \theta/4, & q > r. \end{cases} \tag{4.7.113}$$

It is easy to find that $\Delta(q, \beta, r, \theta)$ still increases as q increases in the first stage, and decreases as q increases in the third stage. However, for the second stage, $\Delta(q, \beta, r, \theta)$ is flat, which does not change as q changes.

So, for $\beta - \theta/2 < q < r$, the maximum is achieved.

Combine the two cases and we get the result. □

4.7.13 Proof of Theorem 4.2.2

According to Lemma 4.7.6, $\sqrt{p\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p)$ is very near to $\widetilde{SNR}(t, \mu, y, n_p, p)$ under the $ARW(\beta, r, \theta, \delta)$ model. With Mill's ratio, Section 4.7.11 shows that $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p) \approx L_p p^{-\Delta(q,\beta,r,\theta)}$. Section 4.7.12 shows that, when $q > \rho^*_\theta(\beta)$, the signal $\sqrt{p\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p)$ goes to infinity, which means that $\widetilde{SNR}(t, \mu, y, n_p, p)$ goes to infinity with probability $1 - o(1/p^2)$ too. Combining with Theorem 4.2.1, the result follows. □

4.7.14 Proof of Theorem 4.2.3

For simplification, we use $\tilde{y}$ to denote the standardized y, which is $y/\|y\|$. Also, we use H, H_0, ξ, Z, μ t to denote $H^{(t_p(q))}$, $H_0^{(t_p(q))}$, $\xi^{(t_p(q))}$, $Z^{(t_p(q))}$, $\mu^{(t_p(q))}$, and $t_p(q)$. Recall that $H = H_0 + A$, where A is a rank 2 matrix. For any given matrix B, we use $\lambda_1(B)$ to denote the leading eigenvalue of B.

Let ξ be the leading eigenvector of H, and we want to prove that $(\xi, \tilde{y}) \to 0$ with large probability. By basic algebra, to show the claim, it is sufficient to show that with probability at least $1 - o(1/n)$,

$$\lambda_1(H_0) \geq p\widetilde{F}_p(t) + \frac{1}{2}\sqrt{n_p p \widetilde{F}_p(t)}, \tag{4.7.114}$$

and that with probability at least $1 - o(1/p^2)$,

$$\lambda_1(A) \leq 2n_p p\widetilde{W}_p(t), \tag{4.7.115}$$

and that

$$\tilde{y}' H_0 \tilde{y} \leq p\widetilde{F}_p(t) + 2\sqrt{6p\widetilde{F}_p(t)\log(p)}. \tag{4.7.116}$$

Combining (4.7.114) and (4.7.116), with basic algebra we can find that, there is a vector w orthogonal to $\tilde{y}$, and that $w' H_0 w \geq p\tilde{F}_p(t) + \sqrt{n_p p\widetilde{F}_p(t)}/2 - 2\sqrt{6p\widetilde{F}_p(t)\log(p)}$. Combine this result with (4.7.115), with basic algebra, there is that, with probability $1 + o(1/n)$,

$$(\xi, \tilde{y})^2 \leq 3\frac{2n_p p\widetilde{W}_p(t) + 2\sqrt{6p\widetilde{F}_p(t)\log(p)}}{\sqrt{n_p p\widetilde{F}_p(t)}/2 - 2\sqrt{6p\widetilde{F}_p(t)\log(p)}}.$$

Combining with the condition that $r < \rho^*_\theta(\beta)$, which indicates that $n_p p\widetilde{W}_p(t)/\sqrt{n_p p\widetilde{F}_p(t)} \to 0$, the claim follows. So, what left to prove are the equations (4.7.114) - (4.7.116).

Let $M = H_0 - p\widetilde{F}_p(t) I_n$, then there is $\lambda_1(H_0) = p\widetilde{F}_p(t) + \lambda_1(M)$. Then, to calculate the lower bound of $\lambda_1(H_0)$ is equivalent with the problem to calculate the lower bound of $\lambda_1(M)$.

Let $\|M\|_F$ to denote the Frobenious norm of M if we see it as an $n_p^2 \times 1$ vector. With basic calculation, there is $E[\|M\|_F^2] = n_p^2 p\widetilde{F}_p(t)(1 + O(1/\sqrt{n_p}))$, and $Var(\|M\|_F^2) \leq n_p^3 (p\widetilde{F}_p(t))^2$. With Chebyshev's inequality, there is

$$P(|\|M\|_F^2 - E[\|M\|_F^2]| \geq \eta) \leq \frac{n_p^3 (p\widetilde{F}_p(t))^2}{\eta^2}.$$

Take $\eta = \frac{n_p^2 p\widetilde{F}(t)}{\log p}$, then with probability at least $1 - o(1/n_p)$, there is

$$\|M\|_F^2 \geq n_p^2 p\widetilde{F}_p(t) - \frac{n_p^2 p\widetilde{F}_p(t)}{\log(p)}. \tag{4.7.117}$$

For any $n_p \times n_p$ matrix (non-stochastic) B such that $rank(B) = k$, there is $\|B\| \geq \|B\|_F/\sqrt{k}$. Note that $rank(M) = n_p$ with probability 1. So, with probability at least $1 - o(1/n_p)$,

$$\|M\| \geq \sqrt{n_p p\widetilde{F}_p(t)}(1 - \frac{1}{2\log(p)}). \tag{4.7.118}$$

Combining with the definition of M, and the relationship between $\lambda_1(M)$ and $\lambda_1(H_0)$, then there is, with probability at least $1 - o(1/n_p)$,

$$\lambda_1(H_0) \geq p\widetilde{F}_p(t) + \sqrt{n_p p\widetilde{F}_p(t)}/2. \tag{4.7.119}$$

For the rank 2 matrix A, we have that

$$\|A\| \leq \|\|\mu\|^2 yy'\| + \|Z\mu y'\| + \|y(Z\mu)'\| \leq 3\|y\|^2\|\mu\|^2/2. \tag{4.7.120}$$

Introducing $\|y\|^2$ and $\|\mu\|^2$, with probability at least $1-o(1/p^2)$, there is

$$\|A\| \le 2n_p p\widetilde{W}_p(t). \tag{4.7.121}$$

Now we bound $|\tilde{y}'H_0\tilde{y}|$, which is $\|Z'\tilde{y}\|^2$. Let w be a $p\times 1$ vector where $w_j = \sum_{i=1}^n \tilde{y}_i Z_{ij}$, then $w\sim N(0,I_p)$. Introduce event $E=\{|w_j|\le 2\sqrt{\log(p)}\}$. It is seen that $P(E^c)\le 1/p$. Let $w_j^* = w_j\cdot 1\{|w_j|\le 2\sqrt{\log(p)}\}$, then over the event E,

$$\|Z'\tilde{y}\|^2 = \sum_{j=1}^p (w_j^*)^2 1\{|T_j|\ge t\}.$$

Now, for any $b>0$ and any sequence of independent random variables W_i such that $|W_i|\le b$, $E[W_i]=0$, and $\mathrm{Var}(W_i)\le\sigma_i^2$ for $1\le i\le p$. Write for short $\sigma^2=\sigma_1^2+\sigma_2^2+\cdots+\sigma_p^2$. Bennett's Lemma [Jin 2012b, Page 38] says that,

$$P(|\sum_{j=1}^p |W_j - E[W_j]|\ge s) \le 2\exp(-\frac{c_0}{2\sigma^2}s^2), \qquad \text{if } sb\le\sigma^2,$$

where $c_0=\psi(1)\approx 0.733$. Applying this with $W_j=(w_j^*)^2 1\{|T_j|\ge t\} - E[(w_j^*)^2 1\{|T_j|\ge t\}]$, $b=4\log(p)+1$, and $s=2\sqrt{2\log(p)}\sigma$ and noting that $\mathrm{Var}(W_j)=L_p p^{-q}$, with probability at least $1-o(1/p^2)$,

$$\sum_{j=1}^p (w_j^*)^2 1\{|T_j|\ge t\} \le E[\sum_{j=1}^p (w_j^*)^2 1\{|T_j|\ge t\}] + 2\sqrt{6p\widetilde{F}_p(t)\log(p)}.$$

Combining this with the expectation, with probability at least $1-o(1/p)$,

$$|\tilde{y}'H_0\tilde{y}| \le p\widetilde{F}_p(t) + 2\sqrt{6p\widetilde{F}_p(t)\log(p)}.$$

So, the claim follows. □

4.7.15 Relationship between ideal HC and HC

Lemma 4.7.7 *For all $|t|<\sqrt{2\log p}$, with probability $1-o(1/p)$, there is that*

$$|HC_p(t,\bar{F}_p) - HC_p(t,\bar{f}_p(t))| \le \frac{L_p}{\sqrt{p}} + L_p\frac{\sqrt{p\bar{f}_p(t)}}{p\epsilon G_\tau(t,n)} HC_p(t,\bar{f}_p(t)). \tag{4.7.122}$$

What's more, for $|t|<2\log p$, with probability $1-o(1/p^2)$, we have that

$$|HC_p(t,\bar{F}_p) - HC_p(t,\bar{f}_p(t))| \le \frac{L_p}{\sqrt{p}}. \tag{4.7.123}$$

In the proof, we use n, $\bar{f}$ to denote n_p and $\bar{f}_p(t)$ as shorthand notations. Let $g_0(t,n) = \frac{d\,G_0(t,n)}{dt}$, and $g_\tau(t,n) = \frac{d\,G_{\tau_p}(t,n)}{dt}$. Also, write n for n_p for short, and

To prove the claim, we decompose $(-\sqrt{2\log p}, \sqrt{2\log p})$ into p small intervals equally, and try to prove the claim on each interval. Let

$$t_i = -\sqrt{2\log p} + \frac{2\sqrt{2\log p}}{p} i, \qquad 0 \le i \le p,$$

and take the i-th interval as $(t_i, t_{i+1}]$ for $0 \le i \le p-1$. Then each interval has length $2\sqrt{2\log p}/p$, and the union of intervals covers the whole interval $(-\sqrt{2\log p}, \sqrt{2\log p})$. Now we try to find the relationship between $HC_p(t, \bar{F}_p)$ and $HC_p(t, \bar{f})$ on the interval $(t_i, t_{i+1}]$. Note that

$$HC_p(t, \bar{F}_p) - HC_p(t, \bar{f}) = I + II + III, \tag{4.7.124}$$

where

$$I = HC_p(t, \bar{F}_p) - HC_p(t_{i+1}, \bar{F}_p),$$

$$II = HC_p(t_{i+1}, \bar{F}_p) - HC_p(t_{i+1}, \bar{f}),$$

and that

$$III = HC_p(t_{i+1}, \bar{f}) - HC_p(t, \bar{f}).$$

Consider I first. Take $HC_p(t,h)$ as a bivariate function about t and h, where

$$HC_p(t,h) = \frac{h - \bar{G}_0(t,n)}{\sqrt{(1+\sqrt{n})h - \sqrt{n}G_0(t,n)}}. \tag{4.7.125}$$

Take the derivative of $HC(t,h)$ with respect to h, then there is

$$\frac{\partial\, HC_p(t,g)}{\partial g}\Big|_{g=\bar{F}_p} = \frac{1}{2}\Big(\frac{1}{\bar{F}_p - \bar{G}_0(t,n)} + \frac{\bar{G}_0(t,n)}{(\bar{F}_p - \bar{G}_0(t,n))[(1+\sqrt{n})\bar{F}_p - \sqrt{n}\bar{G}_0(t,n)]}\Big) HC_p(t, \bar{F}_p). \tag{4.7.126}$$

The difference between $p\bar{F}_p(t_i)$ and $p\bar{F}_p(t_{i+1})$ is distributed as binomial random variable with parameter p and $\bar{f}(t_{i+1}) - \bar{f}(t_i)$. Applying Berstein's inequality for binomial random variables, we have that, with probability at least $1 - o(1/p^2)$, there is

$$|\bar{F}_p(t_i) - \bar{F}_p(t_{i+1})| \le 2\sqrt{(\bar{f}(t_{i+1}) - \bar{f}(t_i))/p}\sqrt{\log(p)}.$$

Combining them, and the difference caused by $\bar{F}_p(t)$ is bounded above as

$$L_p \frac{\sqrt{\bar{f}}}{p\epsilon_p \bar{G}_\tau(t,n)} HC_p(t_{i+1}, \bar{F}_p(t_{i+1})). \tag{4.7.127}$$

Note that t also changes in this interval. Take the derivative of $HC_p(t, \bar{F}_p)$ with respect to t, and we have that

$$\frac{\partial\, HC_p(t, \bar{F}_p)}{\partial t} = \frac{g_0(t,n)}{2(\bar{F}_p - \bar{G}_0(t,n))}[3 - \frac{\bar{F}_p}{(1+\sqrt{n})\bar{F}_p - \sqrt{n}\bar{G}_0(t,n)}]HC_p(t, \bar{F}_p). \quad (4.7.128)$$

Combining with $|t - t_{i+1}| \leq 2\sqrt{2\log p}/p$ for any $t \in (t_i, t_{i+1})$, then the difference from t is bounded above by

$$HC(t, \bar{F}_p) \cdot L_p \bar{G_0}(t,n)/(p\epsilon \bar{G}_\tau(t,n)). \quad (4.7.129)$$

Combining the results from h and t, we have that, with probability at least $1 - o(1/p^2)$,

$$|I| \leq L_p \frac{\sqrt{\bar{f}}}{p\epsilon_p \bar{G}_\tau(t,n)} HC_p(t_{i+1}, \bar{F}_p(t_{i+1})). \quad (4.7.130)$$

Now we find an upper bound to II. For II, the only difference is that $HC_p(t,h)$ take $h = \bar{F}_p$ and $h = \bar{f}$. Note that $p\bar{F}_p \sim Binomial(p, \bar{f})$. With Bernstein's inequality, we have that when $0 < q < 1$ and $0 < \beta < 1$, with probability at least $1 - o(1/p^2)$,

$$|\bar{F}_p - \bar{f}| \leq 2\sqrt{\bar{f}/p}\sqrt{\log(p)}. \quad (4.7.131)$$

Combining (4.7.131) and (4.7.126), with basic algebra, we have that, with probability at least $1 - o(1/p^2)$, there is

$$|II| = |HC_p(t_{i+1}, \bar{F}_p) - HC_p(t_{i+1}, \bar{f})| \leq L_p \frac{\sqrt{p\bar{f}}}{p\epsilon_p \bar{G}_\tau(t,n)} HC_p(t_{i+1}, \bar{f}). \quad (4.7.132)$$

Consider III. For III, we should consider $\bar{f}$ as a function of t, and so $HC_p(t, \bar{f}(t))$ is a function of t only. With basic calculation, ignoring small terms, we have that

$$\frac{d\, HC_p(t, \bar{f})}{dt} = \frac{\epsilon_p}{2(\bar{f} - \bar{G}_0(t,n))}[g_0(t,n) - g_\tau(t,n) + \frac{g_0(t,n)\bar{G}_\tau(t,n) - g_\tau(t,n)\bar{G}_0(t,n)}{\bar{f} + \sqrt{n}\epsilon_p \bar{G}_\tau(t,n)}]HC_p(t, \bar{f}). \quad (4.7.133)$$

Combining with that $|t - t_{i+1}| \leq 2\sqrt{2\log(p)}/p$ for $t \in (t_i, t_{i+1})$, then we have that,

$$|III| = |HC_p(t, \bar{f}) - HC_p(t_{i+1}, \bar{f})| \leq \frac{L_p}{p} HC_p(t, \bar{f}) \quad (4.7.134)$$

Combining (4.7.130), (4.7.132) and (4.7.141), we have that, with probability $1 - o(1/p^3)$,

$$|HC_p(t, \bar{F}_p) - HC_p(t, \bar{f})| \leq L_p \frac{\sqrt{p\bar{f}}}{p\epsilon G_\tau(t,n)} HC_p(t, \bar{f}), \qquad t \in (t_i, t_{i+1}),\ 0 \leq i \leq p-1. \quad (4.7.135)$$

Combine the p intervals, and we have that, with probability $1 - o(1/p)$,

$$|HC_p(t, \bar{F}_p) - HC_p(t, \bar{f})| \leq L_p \frac{\sqrt{p\bar{f}}}{p\epsilon G_\tau(t, n)} HC_p(t, \bar{f}), \quad t \in (-\sqrt{2\log(p)}, \sqrt{2\log(p)}). \tag{4.7.136}$$

Introduce in $HC_p(t, \bar{f})$, and we have that, with probability $1 - o(1/p)$,

$$|HC_p(t, \bar{F}_p) - HC_p(t, \bar{f})| \leq \frac{L_p}{\sqrt{p}}. \tag{4.7.137}$$

□

4.7.16 Relationship between $\widetilde{snr}(t)$ and ideal HC

Recall that the definition for ideal HC and $\widetilde{snr}$ is

$$HC_p(t, \bar{f}) = \frac{\bar{f} - \bar{G}_0(t, n)}{\sqrt{\bar{f} + \sqrt{n}(\bar{f} - \bar{G}_0(t, n))}},$$

and

$$\widetilde{snr} = \frac{w_p}{\sqrt{\bar{f}/n + w_p}}.$$

With basic calculation, we can get that

$$\tau_p^2 HC_p(t, \bar{f}) = \frac{\tau_p^2 \epsilon_p \bar{G}_{\tau_p}(t, n)}{\sqrt{\bar{f} + \sqrt{n}\epsilon_p \bar{G}_{\tau_p}(t, n)}} (1 + O(\bar{G}_0(t, n)/\bar{G}_{\tau_p}(t, n))),$$

and

$$\widetilde{snr} = \frac{\tau_p^2 \epsilon_p \bar{G}_{\tau_p}(t, n)}{\sqrt{\bar{f} + \tau_p^2 \sqrt{n}\epsilon_p \bar{G}_{\tau_p}(t, n)}}.$$

So the claim follows. □

4.7.17 Proof of Theorem 4.2.4

To show the claim, it is sufficient to show that, for any $u > \sqrt{\frac{L_p\sqrt{p\bar{f}}}{p\epsilon_p G_{\tau_p}(t^{ideal}, n)}}$, there is

$$HC(t^{ideal} \pm u, \bar{F}_p(t)) < HC(t^{ideal}, \bar{F}_p(t)), \qquad r < \beta - \theta/2,$$

and that when $r > \beta - \theta/2$,

$$HC(\sqrt{2q_+\log(p)} + u, \bar{F}_p(t)) < HC(\sqrt{2q_+\log(p)}, \bar{F}_p(t)),$$
$$HC(\sqrt{2q_-\log(p)} - u, \bar{F}_p(t)) < HC(\sqrt{2q_-\log(p)}, \bar{F}_p(t)).$$

When $r > \rho_\theta^*(\beta)$, with basic calculation, there is $\frac{L_p\sqrt{p\bar{f}}}{p\epsilon_p G_{\tau_p}(t^{ideal},n)} \to 0$, so the claim follows.

To show the claim, first we show that $t^{idealHC}$ is near to t^{ideal}. There are two cases here, case (a): $r < \beta - \theta/2$, and case (b): $r > \beta - \theta/2$.

Introduce a function $h(t)$ as the estimation of $HC(t, \bar{f})$, where

$$h(t) = h(\epsilon_p, \tau_p, \tau, n) = \frac{\epsilon_p \bar{G}_{\tau_p}(t, n)}{\sqrt{\bar{f} + \sqrt{n}\epsilon_p \bar{G}_{\tau_p}(t, n)}},$$

then we have that

$$HC(t, \bar{f}) = h(t)(1 + O(\bar{G}_0(t, n)/\bar{G}_{\tau_p}(t, n))). \tag{4.7.138}$$

So we could go on to analyze $h(t)$.

In case (a), when $r < \beta - \theta/2$. When $u > c\sqrt{\log(p)}$ for some constant c, it is obvious that $t^{ideal} + u$ will change the order of $HC(t, \bar{f})$, and much smaller than $HC(t^{ideal}, \bar{f})$.

Now, we come to the case that $u = o(\sqrt{\log(p)})$. For the difference between $HC(t^{ideal} + u, \bar{f})$ and $HC(t^{ideal}, \bar{f})$, note that we have

$$HC(t^{ideal} + u, \bar{f}) - HC(t^{ideal}, \bar{f}) = I + II + III, \tag{4.7.139}$$

where

$$I = HC(t^{ideal} + u, \bar{f}) - h(t^{ideal} + u),$$
$$II = h(t^{ideal}) - HC(t^{ideal}, \bar{f}),$$

and that

$$III = h(t^{ideal} + u) - h(t^{ideal}).$$

Combining with (4.7.138), we have that

$$|I + II| \leq (h(t^{ideal}) + h(t^{ideal} + u))O(\bar{G}_0(t^{ideal}, n)/\bar{G}_{\tau_p}(t^{ideal}, n)). \tag{4.7.140}$$

With basic calculation, we have that

$$III \leq (e^{-u^2/4} - 1)h(t^{ideal}). \tag{4.7.141}$$

Combining (4.7.139), (4.7.140) and (4.7.141), when $|u| > L_p\sqrt{\bar{G}_0(t^{ideal}, n)/\bar{G}_{\tau_p}(t^{ideal}, n)}$, there is

$$HC(t^{ideal}+u, \bar{f}) - HC(t^{ideal}) \leq h(t^{ideal})(e^{-u^2/4} - 1 + O(\bar{G}_0(t^{ideal}, n)/\bar{G}_{\tau_p}(t^{ideal}, n))) < 0.$$

Case (b), $r > \beta - \theta/2$. In this case we want to show that $t^{idealHC}$ is in the range of $[\sqrt{2q_- \log(p)} - u, \sqrt{2q_+ \log(p)} + u]$. Here, we show just one side that when u is large, there is

$$HC(t^{ideal} + u, \bar{f}) - HC(t^{ideal}, \bar{f}) \leq 0.$$

The analysis for the other side is the same.

Decompose $HC(\sqrt{2q_-\log p}-u,\bar{f})-HC(\sqrt{2q_-\log p},\bar{f})$ in the same as (4.7.139). For I and II, we have the same result. For III, with basic calculation, we have that, when $|u|=o(\sqrt{\log(p)})$,

$$h(\sqrt{2q_-\log(p)}-u)-h(\sqrt{2q_-\log(p)})\leq \exp[u(u-2\sqrt{2q_-\log(p)})/4]h(\sqrt{2q_-\log(p)}). \tag{4.7.142}$$

Combining I, II and III, when $u>p^{-\beta+\theta/2}$, there is $HC(\sqrt{2q_-\log(p)}-u,\bar{f})<HC(\sqrt{2q_-\log p},\bar{f})$.

Combining the two cases, and we find that $t^{idealHC}\in(t^{ideal}-a,t^{ideal}+a)$ when $r<\beta-\theta/2$, and $t^{idealHC}\in(\sqrt{2q_-\log(p)}-a,\sqrt{2q_+\log(p)}+a)$ when $r>\beta-\theta/2$, where

$$a=a(r,\beta,\theta,p)=\begin{cases} L_p\sqrt{\bar{G}_0(t^{ideal},n)/\bar{G}_{\tau_p}(t^{ideal},n)}, & r<\beta-\theta/2,\\ p^{-\beta+\theta/2}, & r>\beta-\theta/2.\end{cases}$$

Now, we will show that $t^{idealHC}$ is near to t^{ideal}. For u, we have that

$$HC(t^{idealHC}+u,\bar{F}_p)-HC(t^{idealHC},\bar{F}_p)=I+II+III, \tag{4.7.143}$$

where

$$I=HC(t^{idealHC}+u,\bar{F}_p)-HC(t^{idealHC}+u,\bar{f}),$$
$$II=HC(t^{idealHC},\bar{f})-HC(t^{idealHC},\bar{F}_p),$$

and that

$$III=HC(t^{idealHC}+u,\bar{f})-HC(t^{idealHC},\bar{f}).$$

Combining I and II with (4.7.136), there is

$$|I+II|\leq L_p\frac{\sqrt{p\bar{f}}}{p\epsilon_p G_{\tau_p}(t^{ideal},n)}(HC(t^{idealHC},\bar{f})+HC(t^{idealHC}+u,\bar{f})).$$

For III, with similar analysis before, we have that

$$III\leq[\exp(au-u^2/4)-1]HC(t^{idealHC},\bar{f}).$$

Combine, with basic algebra, when $u>L_p\sqrt{\frac{L_p\sqrt{p\bar{f}}}{p\epsilon_p\bar{G}_{\tau_p}(t^{ideal},n)}}$, there is

$$I+II+III\leq\frac{L_p\sqrt{p\bar{f}}}{p\epsilon_p G_{\tau_p}(t^{ideal},n)}(HC(t^{idealHC},\bar{f})+HC(t^{idealHC}+u,\bar{f}))-HC(t^{idealHC},\bar{f})u^2/4<0.$$

So, the claim follows. □

4.7.18 HCT variant

Here we show that ideal threshold for $\widetilde{\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p)$ and associated FDR.

In the terms of w_p, $\bar{f}_p$, $\widetilde{\widetilde{snr}}$ can be written as

$$\widetilde{\widetilde{snr}} = \frac{\sqrt{n_p} w_p(t_p(q))}{\sqrt{\bar{f}_p(t_p(q)) + n_p w_p(t_p(q))}}(1 - \frac{\log(\bar{G}_0(t_p(q)))}{\log(p)}). \tag{4.7.144}$$

Introduce in $t_p(q) = \sqrt{2q \log p}$, and $\bar{G}_0(t_p(q)) = L_p p^{-q}$, we have that

$$\widetilde{\widetilde{snr}} = \widetilde{snr}(1 + q + \log(L_p)/\log(p)). \tag{4.7.145}$$

So, the ideal threshold turns to be $t^{idealvariant} \sim \sqrt{2q^{ideal} \log p}$, where

$$q^{ideal} = \begin{cases} 4r, & r < (\beta - \theta/2)/3, \\ \frac{(\beta - \theta/2 + r)^2}{4r}, & (\beta - \theta/2)/3 < r < \beta - \theta/2, \\ r, & r > \beta - \theta/2. \end{cases} \tag{4.7.146}$$

We have that

$$FDR = \frac{p(1-\epsilon_p)\bar{G}_0(t^{idealvariant})}{p(1-\epsilon_p)\bar{G}_0(t^{idealvariant}) + p\epsilon_p \bar{G}_\tau(t^{idealvariant})}. \tag{4.7.147}$$

Combine (4.7.146) and (4.7.147), then we get the FDR related with $t^{idealvariant}$ as

$$FDR(t_p^{idealvariant}) = \begin{cases} \frac{L_p p^{-3r}}{L_p p^{-3r} + p^{-\beta}}, & r < (\beta - \theta/2)/3, \\ \frac{L_p}{L_p + p^{-\theta/2}}, & (\beta - \theta/2)/3 < r < \beta - \theta/2, \\ \frac{L_p p^{-r}}{p^{-\beta} + L_p p^{-r}}, & r > \beta - \theta/2. \end{cases} \tag{4.7.148}$$

4.7.19 Proof of Theorem 4.3.1

Note that the variant of $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$ is that

$$\widetilde{\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p) = \widetilde{snr}(t, \epsilon_p, \tau_p, n_p)(1 - \frac{\log(\bar{G}_0(t, n_p))}{\log(p)}),$$

which does not include $\bar{F}_p$ or $\bar{f}_p$ in the new term. So, the analysis about the difference between $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$ and HC function can be easily extended to the variant of HC functional, with a multiplied term. The details here are left to readers.

However, when t changes, the difference between $\widetilde{\widetilde{snr}}(t, \epsilon_p, \tau_p, n_p)$ and $\widetilde{\widetilde{snr}}(t + \Delta t, \epsilon_p, \tau_p, n_p)$ is larger than that for $\widetilde{snr}(t, \epsilon_p, \tau_p, n_p)$. For the case that $r > \beta - \theta/2$, with basic calculations we have that

$$HC(t^{idealvariant} - u, \bar{f}) - HC(t^{idealvariant}, \bar{f}) = HC(t^{idealvariant}, (\bar{f})\frac{u(u-2t)}{2\log(p)})(1 + o(1)).$$

Combine this with the derivations for HC functional, we get that the result. □

4.7.20 χ^2 distribution

Lemma 4.7.8 *For a random variable $Y = (X-n)/\sqrt{2n}$, where $X \sim \chi^2_n$, recall that the probability density function (pdf) and cumulative density function (cdf) of Y is $g_0(\cdot, n)$ and $G_0(\cdot, n)$. Recall that $\phi(\cdot)$ and $\Phi(\cdot)$ is pdf and cdf for standard normal distribution. For any $c > 0$, and any $|t| < c\sqrt{\log(n)}$, there is*

$$f_Y(t) = \phi(t)(1 + O(1/\sqrt{n})), \tag{4.7.149}$$

and correspondingly,

$$F_Y(t) = \Phi(t)(1 + O(1/\sqrt{n})), \qquad 1 - F_Y(t) = \bar{\Phi}(t)(1 + O(1/\sqrt{n})). \tag{4.7.150}$$

Proof. We start with (4.7.149) to show that pdf for Y is very close to $\phi(t)$, and then show that the cdf is also close to $\Phi(t)$.

To show (4.7.149), we calculate the pdf for Y directly, and compare it with $\phi(t)$. By the transformation, it is easy to find that $g_0(t, n) = \sqrt{2n} f_X(n + \sqrt{2n}t)$. As $X \sim \chi^2_n$, the pdf for X at $x = (n + \sqrt{2n}t)$ is that

$$f_X(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)}, \qquad x \geq 0.$$

Combining with Stirling's approximation for Gamma function, the pdf can be estimated as

$$f_X(x) = 2\sqrt{n/\pi} x^{n/2-1} e^{-x/2} 2^{-n/2} (e/(n/2))^{n/2} (1 + O(1/n)).$$

With basic calculation, it can be written as

$$f_X(x) = \frac{1}{2\sqrt{n\pi}} \exp(\frac{n-x}{2} + (n/2-1)\log(x/n))(1 + O(1/n)).$$

Recall the transformation that $x = n + \sqrt{2n}t$, and $f_Y(t) = \sqrt{2n} f_X(x(t))$, then we have

$$g_0(t, n) = \frac{1}{\sqrt{2\pi}} \exp(-\sqrt{n/2}t + (n/2-1)\log(1 + \sqrt{2/n}t))(1 + O(1/n)).$$

With Taylor's expansion, we have that $\log(1 + \sqrt{2/n}t) = \sqrt{2/n}t - t^2/n + O(t^3/\sqrt{n^3})$. Combining with $f_Y(t)$, with basic calculation, there is

$$f_Y(t) = \sqrt{1/(2\pi)} e^{-(t+\sqrt{2/n})^2} (1 + O(t^3/\sqrt{n} + 1/n)).$$

For $|t| < c_1 \log(n)$, any $c_1 > 0$, there is $f_Y(t) = \phi(t)(1 + O(1/\sqrt{n}))$.

Choose $c_0 > 0$ such that $c_0^2 \log^2(n) > c^2 \log(n) + 1$. As $\log(n)$ is large, there exists such c_0. To show that the cdf for Y is also close to the cdf of standard Gaussian, it is sufficient to show that,

$$|F_Y(-c_0 \log(n)) - \Phi(-c_0 \log(n))| = O(1/\sqrt{n})\Phi(-c\sqrt{\log(n)}), \tag{4.7.151}$$

and

$$[F_Y(t) - F_Y(-c_0\log(n))] = [\Phi(t) - \Phi(c_0\log(n))](1 + O(1/\sqrt{n})). \tag{4.7.152}$$

When (4.7.151) and (4.7.152) are proved, the summation of the two equations shows that

$$F_Y(t) = \Phi(t)(1 + O(1/\sqrt{n})), |t| < c\sqrt{\log(n)}.$$

In the same way, $1 - F_Y(t) = \bar{\Phi}(t)(1 + O(1/\sqrt{n}))$ can also be proved.

We have proved that (4.7.149) holds for $|t| < c_1\log(n)$ for any c_1, and it also holds when $c_1 = c_0$. Integrate the pdf over the interval $(-c_0\log(n), t)$ and we can get (4.7.152).

To show (4.7.151), it is sufficient to show that

$$F_Y(-c_0\log(n)) \le \frac{1}{\sqrt{2\pi}} e^{-c_0^2\log^2(n)}, \tag{4.7.153}$$

and that

$$\Phi(-c_0\log(n)) \le \frac{1}{\sqrt{2\pi}} e^{-c_0^2\log^2(n)}.s \tag{4.7.154}$$

With Mill's ratio, it is easy to show (4.7.154) holds. Then the only thing left to show is that (4.7.153) holds. The cdf for Y at t is that

$$P(Y \le t) = P(X \le n + \sqrt{2nt}) = (x/2)^{n/2} e^{-x/2} \sum_{k=0}^{\infty} \frac{(x/2)^k}{\Gamma(n/2 + k + 1)},$$

where $x = n + \sqrt{2nt}$. Combining with Stirling's approximation, with basic calculation, we have that

$$P(Y \le t) = e^{1-x/2} \sum_{k=0}^{\infty} \sqrt{\frac{2\pi}{n/2+k}} \Big(\frac{xe/2}{n/2+k}\Big)^{n/2+k} (1 + O(1/n)). \tag{4.7.155}$$

We start with $(\frac{ex/2}{n/2+k})^{n/2+k}$. Introduce in that $x = n + \sqrt{2nt}$, and we have

$$\Big(\frac{ex/2}{n/2+k}\Big)^{\frac{n}{2}+k} = \exp\Big(\Big(\frac{n}{2}+k\Big)\log\Big(\frac{x}{n+2k}\Big) + 1\Big) = \exp\Big(\Big(\frac{n}{2}+k\Big)\log\Big(1 - \frac{2k - \sqrt{2nt}}{n+2k}\Big) + 1\Big). \tag{4.7.156}$$

With basic algebra, we have $\log(1-x) \le -x - x^2/2$ when $x < 1$. Combining with (4.7.156), and we get that

$$\Big(\frac{ex/2}{n/2+k}\Big)^{\frac{n}{2}+k} \le \exp\Big[n/2 + \sqrt{n/2}t - \frac{(2k - \sqrt{2nt})^2}{4(n+2k)}\Big].$$

Combining with (4.7.155), with basic calculation we have that

$$P(Y \le t) \le 2e\sqrt{\pi/n} \sum_{k=0}^{\infty} \exp\Big[-\frac{(2k - \sqrt{2nt})^2}{4(n+2k)}\Big](1 + O(1/n)). \tag{4.7.157}$$

Now, we go on to estimate the term $\frac{(2k-\sqrt{2nt})^2}{4(n+2k)}$. It is obvious that

$$\frac{(2k-\sqrt{2nt})^2}{4(n+2k)} = \frac{n+2k}{4} - (n+\sqrt{2nt})/2 + \frac{(n+\sqrt{2nt})^2}{4(n+2k)}.$$

Combining with (4.7.157) and that $\frac{1}{1+x} \geq 1-x$, with basic calculation we have that

$$P(Y \leq t) \leq 2e\sqrt{\pi/n}e^{-t^2/2}\sum_{k=0}^{\infty}\exp\big[\sqrt{2/n}t + t^2/n\big]^k(1+O(1/n)). \tag{4.7.158}$$

Take $t = -c_0\log(n)$, then we have that

$$P(Y \leq t) \leq \frac{2e\sqrt{\pi/n}e^{-c_0^2\log^2(n)/2}}{1-\exp[c_0\log(n)\sqrt{2/n}+c_0^2\log^2(n)/n]}(1+O(1/n)) = \frac{\sqrt{2\pi}e}{c_0\log(n)}e^{-c_0^2\log^2(n)}.$$

As $\log(n)\to\infty$, then (4.7.153) can be proved.

So, the claim follows. □

We also treat with non-central χ^2 distribution. We show that the non-central χ^2 distribution is near to non-central normal distribution.

Lemma 4.7.9 *Define a random variable $Y = (X-n)/\sqrt{2n}$, where $X \sim \chi_n^2(n\delta)$, with $\delta \leq c\sqrt{\log(n)/n}$ for some $c > 0$. Define the cdf of Y is $G_{\sqrt{n/2}\delta}(\cdot, n)$. Recall that $\Phi(\cdot)$ is cdf for standard normal distribution. For any $|t| < c_0\sqrt{\log(n)}$, any $c_0 > 0$, there is*

$$G_{\sqrt{n/2}\delta}(t,n) = \Phi(t-\sqrt{n/2}\delta)(1+O(1/n^{1/4})), \quad 1-G_{\sqrt{n/2}\delta}(t,n) = \bar{\Phi}(t-\sqrt{n/2}\delta)(1+O(1/n^{1/4})). \tag{4.7.159}$$

Proof. According to the definition of non-central χ^2 distribution, let $X = \sum_{i=1}^n(Z_i+\sqrt{\delta})^2$, where $Z_i \overset{i.i.d.}{\sim} N(0,1)$. With basic calcuation, we have that

$$P(Y \leq t) = P(X \leq n+\sqrt{2nt}) = P\bigg(\sum_{i=1}^n Z_i^2 + 2\sqrt{\delta}\sum_{i=1}^n Z_i + n\delta \leq n+\sqrt{2nt}\bigg). \tag{4.7.160}$$

As $\sum_{i=1}^n Z_i \sim N(0,n)$, we have that

$$P(|\sum_{i=1}^n Z_i| \leq c_1\sqrt{n\log(n)}) = \bar{\Phi}(c_1\sqrt{\log(n)}) \leq n^{-c_1^2/2}.$$

Let $c_1 = c_0+1$, then we have that, with probability at least $1-o(n^{-c_0^2/2-1/2})$, there is $|\sum_{i=1}^n Z_i| \leq c_1\sqrt{n\log(n)}$.

Combining with (4.7.160), and it follows that

$$P(Y \leq t) \leq P\bigg(\sum_{i=1}^n Z_i^2 \leq n+\sqrt{2nt}-n\delta+2c_1\sqrt{n\delta\log(n)}\bigg) + o(n^{-c_0^2/2-1/2}).$$

As $\sum_{i=1}^n Z_i^2$ is central χ^2 distributed, combining with Lemma 4.7.8, letting $t_0 = t - \sqrt{n/2}\delta + c_1\sqrt{2\delta\log(n)}$, and we have that

$$P(Y \le t) \le \Phi(t - \sqrt{n/2}\delta + c_1\sqrt{2\delta\log(n)})(1 + O(1/\sqrt{n})) + o(n^{-c_0^2/2-1/2}),$$

As $\delta = O(1/\sqrt{n})$, and $n^{-c_0^2/2-1/2} = O(n^{1/2}\Phi(t - \sqrt{n/2}\delta))$, so we have that

$$P(Y \le t) \le \Phi(t - \sqrt{n/2}\delta)(1 + O(1/n^{1/4})). \tag{4.7.161}$$

On the other hand, similarly we have that

$$\begin{aligned} P(Y \le t) &\ge P\left(\sum_{i=1}^n Z_i^2 \le n + \sqrt{2n}t - n\delta - 2c_1\sqrt{n\delta\log(n)}\right) + o(n^{-c_0^2/2-1/2}) \\ &= \Phi(t - \sqrt{n/2}\delta)(1 + O(1/n^{1/4})). \end{aligned} \tag{4.7.162}$$

Combining (4.7.152) and (4.7.153), and we get the result. □

Lemma 4.7.10 *Define a random variable $Y = (X-n)/\sqrt{2n}$, where $X \sim \chi_n^2(n\delta)$, with $\delta \le c\sqrt{\log(n)/n}$ for some $c > 0$. Define the cdf of Y is $g_{\sqrt{n/2}\delta}(\cdot, n)$. Recall that $\phi(\cdot)$ is pdf for standard normal distribution. For any $|t| < c_0\sqrt{\log(n)}$, any $c_0 > 0$, there is*

$$g_{\sqrt{n/2}\delta}(t, n) = \phi(t - \sqrt{n/2}\delta)(1 + O(1/n^{1/4})). \tag{4.7.163}$$

Proof. To prove the claim, it is sufficient to show that

$$\exp[\sqrt{\frac{n}{2}}t\delta - \frac{n}{4}\delta^2](1 + O(1/n^{1/4})) = \sum_{i=0}^{\infty} \frac{e^{-n\delta/2}(n\delta/2)^i}{i!}\frac{x^i\Gamma(n/2)}{2^i\Gamma(n/2+i)}, \tag{4.7.164}$$

where $x = n + \sqrt{2n}t$. The sufficiency could be shown by that

$$\phi(t - \sqrt{n/2}\delta) = \phi(t)\exp[\sqrt{\frac{n}{2}}t\delta - \frac{n}{4}\delta^2],$$

and that

$$f_Y(t,\delta) = \sqrt{2n}\sum_{i=0}^{\infty} \frac{e^{-n\delta/2}(n\delta/2)^i}{i!} f_{\chi^2_{n+2i}}(x,0) = \sqrt{2n} f_{\chi^2_n}(x,0)\sum_{i=0}^{\infty} \frac{e^{-n\delta/2}(n\delta/2)^i}{i!}\frac{x^i\Gamma(n/2)}{2^i\Gamma(n/2+i)}.$$

As we have that $\phi(t) = \sqrt{2n} f_{\chi^2_n}(x,0)(1 + O(1/\sqrt{n}))$ from Lemma 4.7.8, so what leaves to prove is (4.7.164) only.

With the property of Gamma function, we have that

$$\frac{x^i\Gamma(n/2)}{2^i\Gamma(n/2+i)} = \prod_{j=1}^{i}\frac{x}{n+2j-2} = \exp[\sum_{i=0}^{\infty}\log(\frac{n+\sqrt{2n}t}{n+2j-2})]. \tag{4.7.165}$$

Combining with Taylor expansion for $\log(\frac{n+\sqrt{2nt}}{n+2j-2})$, and we have that

$$\sum_{i=0}^{\infty}\log(\frac{n+\sqrt{2nt}}{n+2j-2}) = \sqrt{\frac{2}{n}}t - \frac{i(i-1)}{n} + O(1/\sqrt{n}), \qquad i < 2c_0\sqrt{n\log(n)}.$$

Introduce it into (4.7.165), then we have

$$\frac{x^i\Gamma(n/2)}{2^i\Gamma(n/2+i)} = \exp\left[i\sqrt{2/nt} - i^2/n + i/n)\right](1+O(1/\sqrt{n})), \qquad i < 2c_0\sqrt{n\log(n)}.$$

Combining with (4.7.164), the right side of (4.7.164) can be written as

$$RHS = e^{-n\delta/2}\sum_{i=0}^{\infty} e^{-i^2/n}\left(\frac{n\delta e^{\sqrt{2/nt}+1/n}}{2}\right)^i/i!(1+O(1/n^{1/4})).$$

With Taylor's expansion for $e^{-i^2/n} = \sum_{k=0}^{\infty}(i^2/n)^k/k!$, we have that

$$RHS = e^{-\frac{n\delta}{2}}\sum_{k=0}^{\infty}\frac{(-n^2\delta^2e^{2\sqrt{2/nt}}/4)^k}{n^kk!}\sum_{i=0}^{\infty}\left(\frac{n\delta e^{\sqrt{2/nt}+1/n}}{2}\right)^i/i!(1+O(1/n^{1/4})).$$

With basic calculation, we can find that

$$RHS = \exp[\frac{n\delta}{2}(e^{\sqrt{2/nt}}-1)]e^{-n\delta^2e^{2\sqrt{2/nt}}/4} = e^{\sqrt{n/2}t\delta - n\delta^2/4}(1+O(1/n^{1/4})) = LHS.$$

So, the claim is proved. □

Chapter 5

The Tail Distribution of Kolmogorov-Smirnov Statistic

Contents

5.1 Introduction

We observe an $n \times p$ matrix W, which can be thought of a rank-1 matrix hidden in noise:

$$W = \ell\mu' + Z, \qquad Z = Z_{n,p}.$$

where ℓ is an $n \times 1$ vector and μ is a $p \times 1$ vector, both of which are unknown to us. Usually, we regard ℓ as a label vector and μ as a sparse feature vector. Such a setting can be found in various application areas.

- *Two-class clustering.* We have gene microarray data from two-classes, but the class labels are unknown and it is of major interest to recover them. In the simplest setting, two classes are equally likely, and $\ell_i = \pm 1$ with equal probabilities. The vector μ is the contrast feature vector, which is unknown but is presumably sparse, and Z is the matrix of measurement noise. The main goal is to recover ℓ. A good reference is our forthcoming manuscript ([Jin 2012a]).

- *Sparse PCA.* In this setting, $\ell_i \overset{iid}{\sim} N(0, \sigma^2)$, μ is the sparse feature vector, and Z is the matrix of noise. The main interest is to recover μ. A good example is [Johnstone 2009].

- *Community detection in network analysis.* In the simplest case we only have two communities. In this example, $p = n$, and $\ell\mu'$ is replaced by a rank-2 $p \times p$ matrix, say, L, where $L(i, j)$ represents the probability that there is an edge between node i and node j. Also, Z is a $p \times p$ matrix the entries of which are (centered) Bernoulli noise. A good reference is [Bickel 2009].

The above studies, especially the clustering problem, motivates us to study the Kolmogorov-Smirnov (KS) statistics. Our study on two gene microarray data sets (Leukemia data by [Golub 1999], and Lung cancer data by [Gordon 2002] shows that using Komogorov-Smirnov test to assess the significances of gens is especially successful. In contrast, using moment-based approaches to assess the significances of genes are comparably much more unsatisfactory.

5.1.1 Two models

In this note, we are interested in the the tail behavior of the KS-statistics in two idealized models. In the first model, we suppose that we have n different (univariate) samples

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2), \tag{5.1.1}$$

where both parameters (μ, σ) are unknown. We are interested in the tail probability of the KS-statistics, with estimations of (μ, σ) plugged in.

In the second model, we have n different (univariate) samples from a two component mixture:

$$X_i \overset{i.i.d.}{\sim} (1-\delta)N(\mu_1, \sigma^2) + \delta(\mu_2, \sigma^2), \tag{5.1.2}$$

where all parameters $(\delta, \mu_1, \mu_2, \sigma)$ are unknown. To misuse the notation a little bit, we let

$$\mu = (1-\delta)\mu_1 + \delta\mu_2, \qquad (1-\delta)d_1 = -\delta d_2 = d_0, \qquad \text{where } d_0 = \mu_2 - \mu_1.$$

Similarly, we are primarily interested in the tail behavior of the KS-statistics with estimations of (μ, σ) plugged in. We assume d_0 is small, so the problem is challenging but is still solvable.

5.1.2 Literature review

In the literature, there are two noteworthy approaches to calculate the tail probability of the KS statistic. The first approach is—probably the earliest of all—is given by Kolmogorov himself [Kolmogorov 1933], which solves the problem in the case where both (μ, σ) are known. This approach uses Kolmogorov's Forward Equation, and relies on explicit formula of the joint density of the j-th and k-th order statistics of the n independent samples from $U(0,1)$, for any $1 \le j, k \le n$. To extend the approach to the case where (μ, σ) are unknown and have to be estimated, we have to calculate the joint density of any such pairs of order statistics conditional on first two moments of the samples. Seemingly, this is a challenging problem, which makes Kolmogorov's approach hard to extend.

Alternatively, there are the modern approaches by Durbin [Durbin 1985] and by Loader [Loader 1992]. These two approaches share a common ground in that they both attempt to approximate the so-called *first passage probability* and use it to approximate the tail probability. However, two approaches are also different in important ways: Durbin [Durbin 1985] uses a Gaussian process approaches, and Loader [Loader 1992] uses a locally Poisson process. In comparison, the latter is found to be more accurate.

In this note, driven by the modern interest of "large p", we are mainly interested in large-deviation results related to the KS statistics. We start with Loader's approximation of first passage probability, and derive simple and explicit large-deviation formula for the tail probability of the KS statistics. Our study covers both the Null model (5.1.1) and the Alternative model (5.1.2), and is readily extendable to exponential families.

5.1.3 Content

In Section 5.2, we derive approximations for the tail probabilities of the KS-statistic associated with Model (5.1.1), and in Section 5.2.3, we discuss that associated with Model (5.1.2). Section 5.4 contains some numerical results, illustrating how accurate

the approximations are in terms of the tail probability as well as the mean and standard deviation.

5.2 Main results

In this section, we first introduce some necessary notations. We then give the main large-deviation formula for the KS statistic associated with Model (5.1.1) and Model (5.1.2).

5.2.1 Notations

In either of the two models, we estimate μ and σ^2 by

$$\hat{\mu}_n = \hat{\mu}_n(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i, \qquad \hat{\sigma}_n^2 = \hat{\sigma}_n^2(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2. \tag{5.2.3}$$

Next, let $\Phi(t;\mu,\sigma)$ be the CDF of $N(\mu,\sigma^2)$:

$$\Phi(t;\mu,\sigma) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \tag{5.2.4}$$

and let $\phi(t;\mu,\sigma)$ be the density of $N(\mu,\sigma^2)$:

$$\phi(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{5.2.5}$$

In the special case of $(\mu,\sigma) = (0,1)$, we write $\Phi(t) = \Phi(t;0,1)$ and $\phi(t) = \phi(t;0,1)$ for short. Also, later in this note,

$$\Phi(t;\hat{\mu}_n,\hat{\sigma}_n) = \Phi(t;\mu,\sigma)\Big|_{\{(\mu,\sigma)=(\hat{\mu}_n,\hat{\sigma}_n)\}}, \qquad \phi(t;\hat{\mu}_n,\hat{\sigma}_n) = \phi(t;\mu,\sigma)\Big|_{\{(\mu,\sigma)=(\hat{\mu}_n,\hat{\sigma}_n)\}}$$

At the same time, denote the empirical CDF by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n 1\{X_i \le t\}. \tag{5.2.6}$$

The KS-statistic with estimations of (μ,σ) plugged-in is defined as

$$KS_n = \sqrt{n}\sup_{-\infty<t<\infty} |F_n(t) - \Phi(t;\hat{\mu}_n,\hat{\sigma}_n)|. \tag{5.2.7}$$

Similarly, we define

$$KS_n^{\pm} = \sqrt{n}\sup_{-\infty<t<\infty} \{\pm[F_n(t) - \Phi(t;\hat{\mu}_n,\hat{\sigma}_n)]\}. \tag{5.2.8}$$

5.2.2 Large deviation approximations for Model (5.1.1)

Consider Model (5.1.1) first, where $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and (μ, σ) are unknown. The main result is

Theorem 5.2.1 *Fix (μ, σ) such that $\sigma > 0$. As $n \to \infty$, if $\eta_n \to \infty$ and $\eta_n/\sqrt{n} \to 0$, then*

$$P(KS_n^{\pm} \geq \eta_n) = (1 + o(1)) \cdot \Big(\sqrt{\frac{2\pi}{\pi - 2}} e^{-\frac{2\pi}{\pi-2}\eta_n^2}\Big),$$

and

$$P(KS_n \geq \eta_n) = (1 + o(1)) \cdot \Big(2\sqrt{\frac{2\pi}{\pi - 2}} e^{-\frac{2\pi}{\pi-2}\eta_n^2}\Big).$$

Theorem 5.2.1 give approximations for $P(KS_n \geq \eta_n)$ (and similar approximations for $KS_n^{\pm}$ by

$$P(KS_n \geq \eta_n) \approx \min\Big\{1,\ 2\sqrt{\frac{2\pi}{\pi - 2}} e^{-\frac{2\pi}{\pi-2}\eta_n^2}\Big\} \tag{5.2.9}$$

In Section 5.4, we further investigate this approximation numerically, focusing on small or moderately large η_n.

We remark that, first, it is not hard to see that the distribution of KS_n and KS_n^- does not depend on (μ, σ), so it is not surprising that all approximations above do not involve (μ, σ). Second, we recall that in the case where (μ, σ) are known, then for any $\eta > 0$,

$$P(KS_n^- \geq \eta) = e^{-2\eta^2},$$

and when $\eta_n \to \infty$,

$$P(KS_n \geq \eta_n) \sim 2e^{-2\eta_n^2}.$$

The inequalities we derive above are quite similar.

5.2.3 The tail probability of the KS-statistics with Model (5.1.2)

Consider Model (5.1.2) where $X_i \overset{i.i.d.}{\sim} (1-\delta)N(\mu_1, \sigma^2) + \delta N(\mu_2, \sigma^2)$. Recall that

$$\mu = (1-\delta)\mu_1 + \delta\mu_2, \qquad (1-\delta)d_1 = -\delta d_2 = (\mu_2 - \mu_1) \equiv d_0.$$

First, we figure out what is the right calibration for d_0. Note that when both (μ, σ) are unknown, the best way for testing $\mu_1 = \mu_2$ is to use the statistics based on the third moment. By direct calculations,

$$\frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n X_i^3}{\sqrt{15}\hat{\sigma}^3} \approx N(\sqrt{n} d_0^3 \frac{\delta(1-\delta)(1-2\delta)}{\sqrt{15}\sigma^3}, 1). \tag{5.2.10}$$

Assume

$$d_0(1 - 2\delta) > 0,$$

so that in the mixture model (5.1.2), the component with the smaller mass always has a larger mean. Note this requires $\delta \neq 1/2$. The case $\delta = 1/2$ is similar, but the detailed calculations are different, so we leave it to later. In light of (5.2.10), we calibrate with

$$\tau_n = \tau_n(\delta, d_0) = \sqrt{n} d_0^3 \frac{\delta(1-\delta)(1-2\delta)}{\sqrt{15}\sigma^3}. \tag{5.2.11}$$

Introduce a constant

$$a_0 = \sqrt{\frac{5}{24\pi}}, \qquad b_0 = (7 - 3\pi)a_0,$$

Asymptotically, as $\tau_n \to \infty$, the statistics KS_n^- and KS_n "centered" at $a_0\tau_n$, and KS_n^+ centered at $\sqrt{\frac{5}{6\pi}e^{-3}}\tau_n$. As KS_n^+ centered at a smaller point, KS_n is decided mostly by KS_n^-.

Theorem 5.2.2 *Fix (μ, σ) such that $\sigma > 0$. As $n \to \infty$, if $\tau_n \to \infty$, we have that*

$$\frac{E[KS]}{a_0\tau_n} \to 1,$$

Theorem 5.2.3 *Fix $a > 0$. As $n \to \infty$, if $\tau_n \to \infty$ and $\tau_n/\sqrt{n} \to 0$, then*

$$P\big(KS_n^- \geq a\tau_n\big) \sim \begin{cases} 1, & a < a_0, \\ \sqrt{\frac{a-a_0}{a-b_0}}\left[\sqrt{\frac{2\pi}{\pi-2}} e^{-\frac{2\pi}{\pi-2}(a-a_0)^2\tau_n^2}\right], & a > a_0, \end{cases}$$

and

$$P\big(KS_n \geq a\tau_n\big) \sim \begin{cases} 1, & a < a_0, \\ \sqrt{\frac{a-a_0}{a-b_0}}\left[\sqrt{\frac{2\pi}{\pi-2}} e^{-\frac{2\pi}{\pi-2}(a-a_0)^2\tau_n^2}\right], & a > a_0, \end{cases}$$

As a result, it is not hard to see that

$$\lim_{n\to\infty}\{E[KS_n^-]/(a_0\tau_n)] = 1.$$

Note that for the claim in the case of $a < a_0$, the requirement of $\tau_n/\sqrt{n} \to 0$ can be removed.

5.3 Proof of Theorem 5.2.1

In this section, we prove Theorem 5.2.1. The proofs for other theorems and lemmas are in Section 6.4.

Since the study is similar, we only discuss that of KS_n^-. Write $W = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)'$ for short. First, since the statistic KS_n^- is ancillary to the parameters (μ, σ), so it is independent of W. In particular,

$$P(KS_n^- \geq \eta) = P\big(KS_n^- \geq \eta | W = (0,n)'\big).$$

We now evaluate $P\big(KS_n^- \geq \eta | W = (0,n)\big)$. Recall $KS_n^- = \sqrt{n}\sup_{-\infty<t<\infty}\{-[F_n(t) - \Phi(t;\hat{\mu}_n,\hat{\sigma}_n)]\}$. We note that given $W = (0,n)$, $(\hat{\mu}_n, \hat{\sigma}_n) = (0,1)$, and so $\Phi(t;\hat{\mu}_n,\hat{\sigma}_n) = \Phi(t;0,1) \equiv \Phi(t)$. Write for short

$$G_n(t) = nF_n(t) = \sum_{i=1}^n 1\{X_i \leq t\}, \qquad q_t = \Phi(t) - \eta/\sqrt{n}, \qquad \epsilon_n = \eta/\sqrt{n}.$$

We have

$$\begin{aligned} P\big(KS_n^- \geq \eta | W = (0,n)'\big) &= P\big(\sqrt{n}\sup_{-\infty<t<\infty}\{-F_n(t) + \Phi(t)\} \geq \eta \big| W = (0,n)'\big) \\ &= P\Big(\inf_{-\infty<t<\infty}\big\{F_n(t) - \Phi(t) + \eta/\sqrt{n}\big\} \leq 0 \big| W = (0,n)'\Big) \qquad (5.3.12) \\ &\equiv P\big(\inf_{-\infty<t<\infty}\{G_n(t) - nq_t\} \leq 0 \big| W = (0,n)'\big). \qquad (5.3.13) \end{aligned}$$

Let t_j be the solution of

$$nq_t = j, \qquad j = 1, \ldots, n-1.$$

Introduce the *first boundary crossing time* by

$$\tau = \inf\{t : G_n(t) < nq_t\}.$$

Since (a) $G_n(t)$ only take integer values, and (b) q_t is strictly increasing in t, two key observation are that $\{\tau < \infty\} = \{t_1, t_2, \ldots, t_{n-1}\}$ with probability 1, and that given $\tau = t_j$, $G_n(t_j) = j$. It follows that

$$\begin{aligned} P\big(\inf_{-\infty<t<\infty}\{G_n(t) - nq_t\} \leq 0 \big| W = (0,n)'\big) &= \sum_{j=1}^n P(\tau = t_j, G_n(t_j) = j | W = (0,n)') \\ &= \sum_{j=1}^n P\big(\tau = t_j \big| G_n(t_j) = j, W = (0,n)'\big) \cdot P\big(G_n = j \big| W = (0,n)'\big). \end{aligned} \qquad (5.3.14)$$

We now analyze $P(\tau = t_j | G_n = j)$ for each $1 \leq j \leq n$. These are the following lemmas, which are proved in the appendix. In the literature, this is referred to

"first-passage" probability, and how to approximate it is a well-known difficult problem. There are two noteworthy approaches, that by Durbin [Durbin 1985] and that by Loader [Loader 1992]. In comparison, the latter is more accurate. In fact, by Loader [Loader 1992, 11],

$$P\big(\tau = t_j | G_n(t_j) = j, W = (0,n)'\big) = 1 - \Big[(\frac{\partial q_t}{dt})^{-1}\mu(t,q_t)\Big]\Big|_{t=t_j} + o(1), \qquad (5.3.15)$$

where $\mu(t,q_t)$ is defined in (5.7.24).

Lemma 5.3.1 *For each* $1 \le j \le n$,

$$P\big(\tau = t_j \big| G_n(t_j) = j\big) = \frac{\Phi(-t_j) + t_j\phi(t_j)(1+t_j^2/2)}{\sqrt{\Phi(t_j)\Phi(-t_j) - \phi^2(t_j)(1+t_j^2/2)}}\epsilon_n(1+o(1)).$$

Lemma 5.3.2 *For each* $1 \le j \le n$,

$$P\big(G_n = j \big| W = (0,n)'\big) = \frac{1}{\sqrt{2\pi n(\Phi(t_j)\Phi(-t_j) - \phi^2(t_j)(1+t_j^2/2))}} e^{-\frac{n\epsilon_n^2}{2(\Phi(t_j)\Phi(-t_j)-\phi^2(t_j)(1+t_j^2/2))}}(1+o(1)).$$

Combining two lemmas, we have

$$P(KS_n^- \ge \eta) = (1+o(1))\sum_{j=1}^{n-1} \frac{\big(\Phi(-t_j) + t_j\phi(t_j)(1+t_j^2/2)\big)\epsilon_n}{\sqrt{2\pi n(\Phi(t_j)\Phi(-t_j) - \phi^2(t_j)(1+t_j^2/2))^3}} e^{-\frac{n\epsilon_n^2}{2(\Phi(t_j)\Phi(-t_j)-\phi^2(t_j)(1+t_j^2/2))}}.$$

where the right hand side converges to a Riemann integral

$$\int_{-\sqrt{-\log(\epsilon_n)}}^{1/\sqrt{\epsilon_n}} \sqrt{n}\epsilon_n h(t) e^{-n\epsilon_n^2 g(t)} dt,$$

where

$$g(t) = \frac{1}{2}\frac{1}{\Phi(t)\Phi(-t) - \phi^2(t)(1+t^2/2)}, \qquad h(t) = \frac{\Phi(-t)\phi(t) + t\phi^2(t)(1+t^2/2)}{\sqrt{2\pi(\Phi(t)\Phi(-t) - \phi^2(t)(1+t^2/2))^3}}.$$

The claim now follows from the following lemma.

Lemma 5.3.3 *As* $\eta_n \to \infty$,

$$\int_{-\sqrt{-\log(\epsilon_n)}}^{1/\sqrt{\epsilon_n}} \sqrt{n}\epsilon_n h(t) e^{-n\epsilon_n^2 g(t)} dt = \sqrt{\frac{\pi}{\pi-2}} e^{-\frac{2\pi}{\pi-2}\eta^2}(1+o(1)).$$

5.4 Simulations

We conducted a small-scale simulation study, investigating the tail behavior of KS_n and KS_n^-, as well as their means and standard deviations. Below, we discuss KS_n^- and KS_n separately.

5.4.1 The KS^- statistics

The study in this section contains two experiments which we now discuss separately.

Experiment 1. In this experiment, for any fixed n, since the distribution of KS_n^- does not depend on (μ, σ), we generate samples $X_1, X_2, \ldots, X_n$ from $N(0,1)$. For any $\eta > 0$, we simulate the value of $P(KS_n^- \geq \eta)$ by repeatedly generating samples for a large number of times, which can be thought of as the true value of $P(KS_n^- \geq \eta)$. At the same time, we compute the approximation of $P(KS_n^- \geq \eta)$ given by Theorem 5.2.1. The experiment contains two sub-experiments, Experiment 1a and 1b.

In Experiment 1a, we let n range in $\{50, 100, 500, 5000\}$ and let η range between 0.2 and 1.3 with an increment of 0.01. For each combination of (n, η), we first simulate the value $P(KS_n^- \geq \eta)$ by 20,000 independent repetitions, and then obtain the approximation given by Theorem 5.2.1 as above. The results are displayed in Figure 5.1. For large n (e.g. $n \geq 500$), the approximation is impressively accurate. Figure 5.2 shows the ratio between empirical tail probability and approximated tail probability. It is very close to 1 when η is large.

In Experiment 1b, we fix $n = 500$, and let η range from 0 to 1.5 with an increment of 0.01. For each η, we simulate the value of $P(KS_n^- \geq \eta)$ using 10^4 independent repetitions. We then use $P(KS_n^- \geq \eta)$ for all these η to compute the mean and standard deviation of KS_n^-, which can be viewed as the true mean and true standard deviation. At the same time, we compute $P(KS_n^- \geq \eta)$ for all η using the approximation formula in (**??**), and then use the results to approximate the mean and standard deviation of KS_n^-. The results are summarized in Table 5.1, which suggest that the approximations for both the mean and for the standard deviation are reasonably accurate.

	Mean	Standard Deviation
True Value	0.5471	0.1528
Simplified Approx.	0.5634	0.1386

Table 5.1: Experiment 1b. Theoretical mean and standard deviation of KS_n^- (based on simulations), and simplified approximated values (based on Theorem 5.2.1).

Experiment 2. We use the same setting as in Experiment 1, except that we generate samples $X_1, X_2, \ldots, X_n$ from Model (5.1.2) instead of Model (5.1.1). Note that similarly, since the distribution of KS_n^- does not depend on (μ, σ), we take $(\mu, \sigma) = (0, 1)$. At

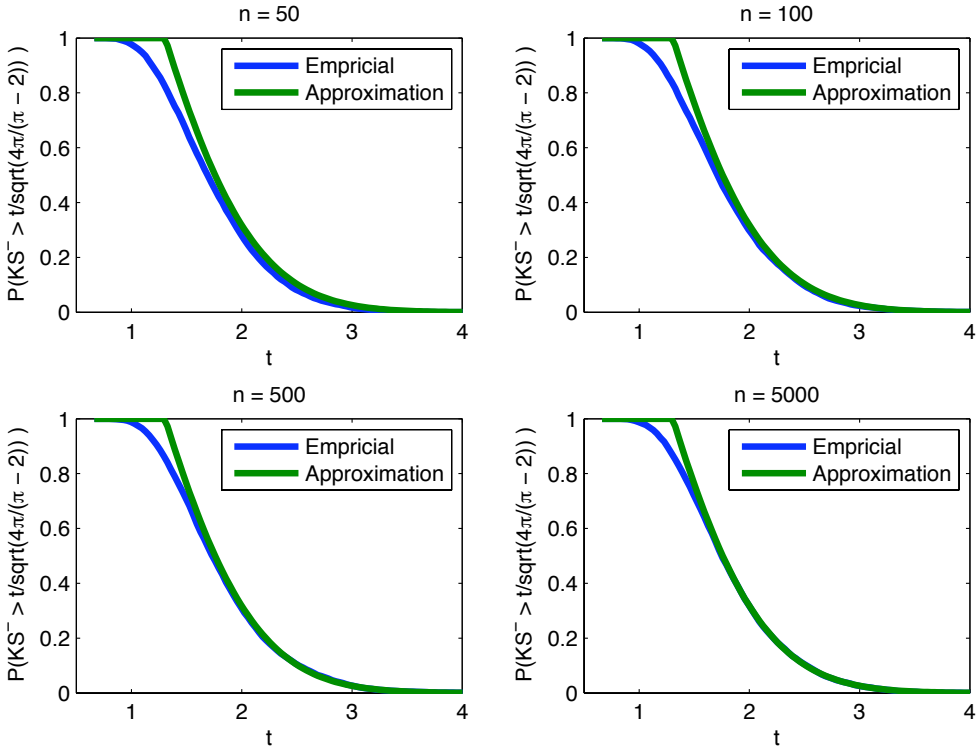

Figure 5.1: Experiment 1a. Comparison of $P(KS_n^- \geq \eta)$ (blue) and the approximation in Theorem 5.2.1 (green). x-axis: η. y-axis: $P(KS_n^- \geq \eta)$.

the same time, we let $\delta = 1/3$ and choose d_0 such that

$$\tau_n = \sqrt{n} d_0^3 \frac{\delta(1-\delta)(1-2\delta)}{\sqrt{15}} = (\frac{5}{6})^3 \sqrt{\log(n)}.$$

The results are displayed in Figure 5.3, which suggest that the approximations are satisfactory, especially when $n \geq 500$. In Figure 5.4, we show the ratio of empirical probability and approximated probability for $n = 500$ and $n = 5000$. The result is not as good as model (5.1.1).

5.4.2 The KS_n statistics

In this section, we investigate the performance of KS_n. The study contains two experiments, Experiment 3 and 4.

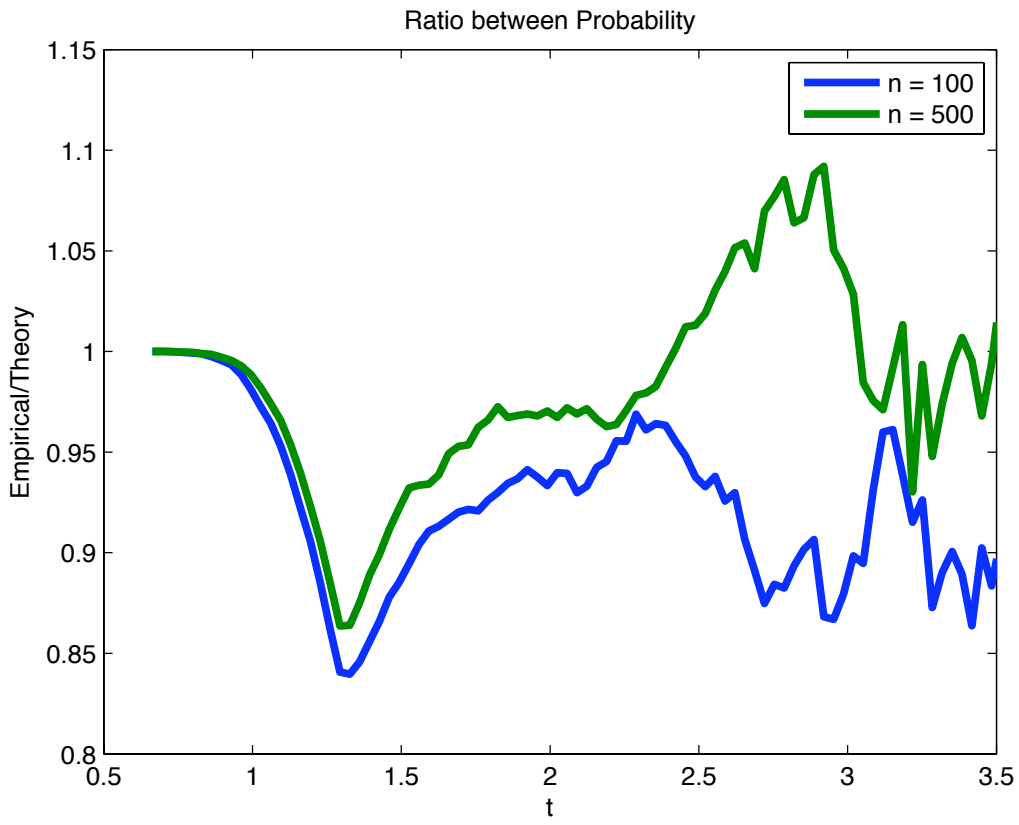

Figure 5.2: Experiment 1a. Ratio of empirical tail probability to approximated tail probability for $n = 100$ (blue) and $n = 500$ (green.) x-axis: η. y-axis: *Empirical/Approximated*.

Experiment 3. The experiment contains two sub-experiments, Experiment 3a and 3b.

In Experiment 3a, we use the same setting as in Experiment 1a, except for that we investigate KS_n instead of KS_n^-. For each combination of (n, η), we report the true value of $P(KS_n \geq \eta)$ using $20,000$ independent simulations, as well as the approximated value given by (5.2.9), respectively. The results are in Figure 5.5, which suggest that the approximation is reasonably good. The ratio plot Figure 5.6 also shows the excellent matching.

In Experiment 3b, we use the same setting as in Experiment 1b, except for that we investigate the mean and standard deviation of KS_n instead of those of KS_n^-. The

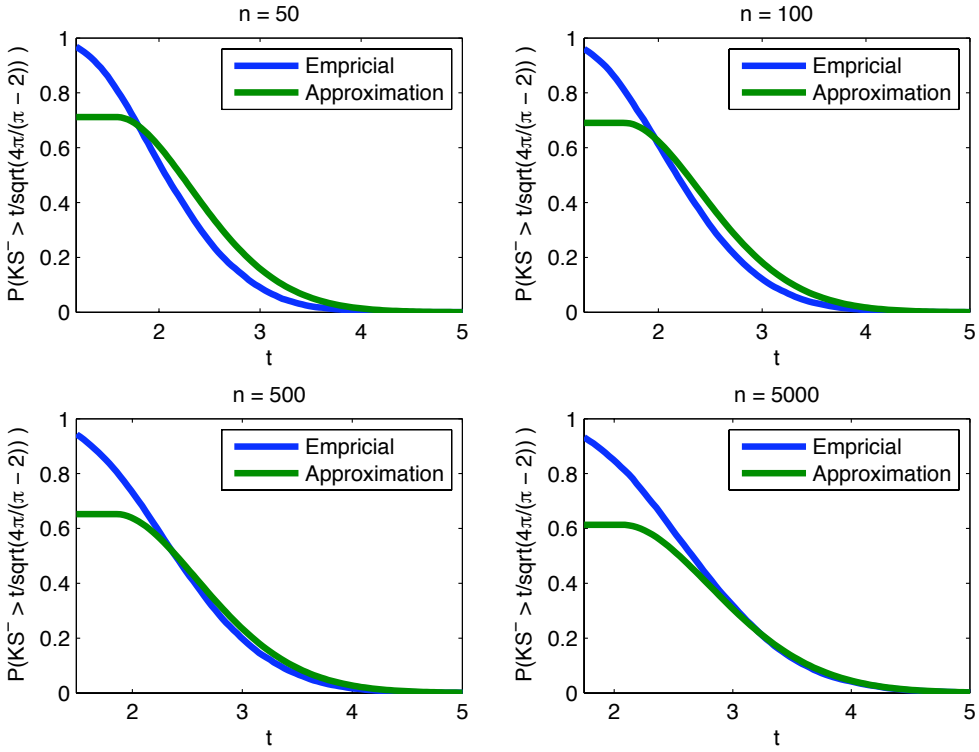

Figure 5.3: Experiment 2. Comparison of $P(KS_n^- \geq \eta)$ (blue) and the associated approximation given by Theorem 5.2.3 (green). x-axis: η. y-axis: $P(KS_n^- \geq \eta)$.

results are reported in Table 5.2. The results suggest that the approximation of mean and standard deviation are reasonably good (note that $n = 500$).

	Mean	Standard Deviation
True Value	0.6299	0.1476
Approximation	0.6695	0.1199

Table 5.2: Experiment 3b. Theoretical mean and standard deviation of KS_n (based on simulations) and approximated values (based on (5.2.9)).

In Experiment 4, we use the same setting as in Experiment 2, except for that we investigate KS_n instead of KS_n^-. The results are displayed in Figure 5.7, which suggest the approximations are reasonably good.

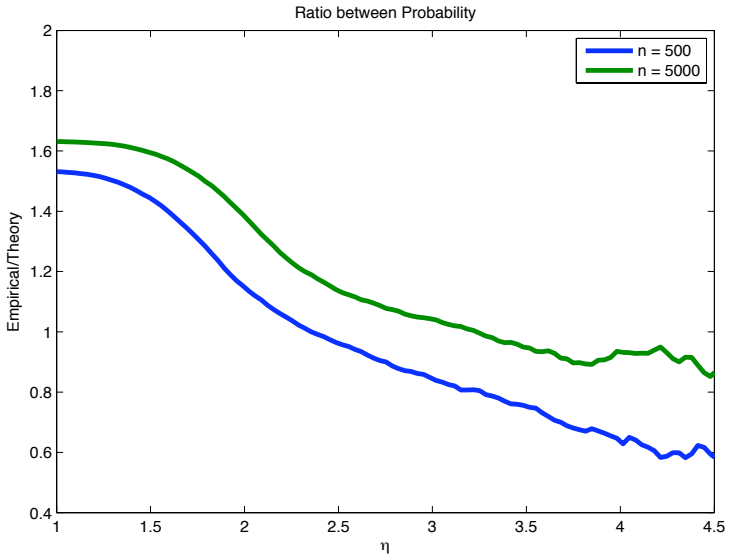

Figure 5.4: Experiment 2. Ratio of empirical tail probability to approximated tail probability for $n = 500$ (blue) and $n = 5000$ (green.) x-axis: η. y-axis: *Empirical/Approximated*.

5.5 Case Study

5.5.1 Cardio Data

Why do we care about KS statistics instead of other statistics? Here we show the comparison of KS statistic and third moment statistic for Cardio cancer data, to have a look at the power and robustness for two statistics. In this data set, there are 44 healthy people, and 19 cardiovascular patients. For each people, the information of 20426 genes are recorded. Through these genes, most of them contain no information about the disease, and only a small fraction of genes have signal. In data analysis, we hope to recover as many as informative genes while we make only a few errors. So, as we have some statistics to do goodness of fitting, we can evaluate the p-value for each gene with the distribution of corresponding statistics under null hypothesis. Here, we assume that under null hypothesis, each gene is Gaussian distributed.

For cardio data set, to find the null distribution for two statistics, we run simulation. As $n = 63$, we take 10k samples, with 63 Gaussian distributed data points for each sample. For each sample, we calculate the KS statistic and third moment, while the mean is standardized to be 0 and variance is standardized to be 1. After we got the distribution, we calculate the tail probability for statistics from data and simula-

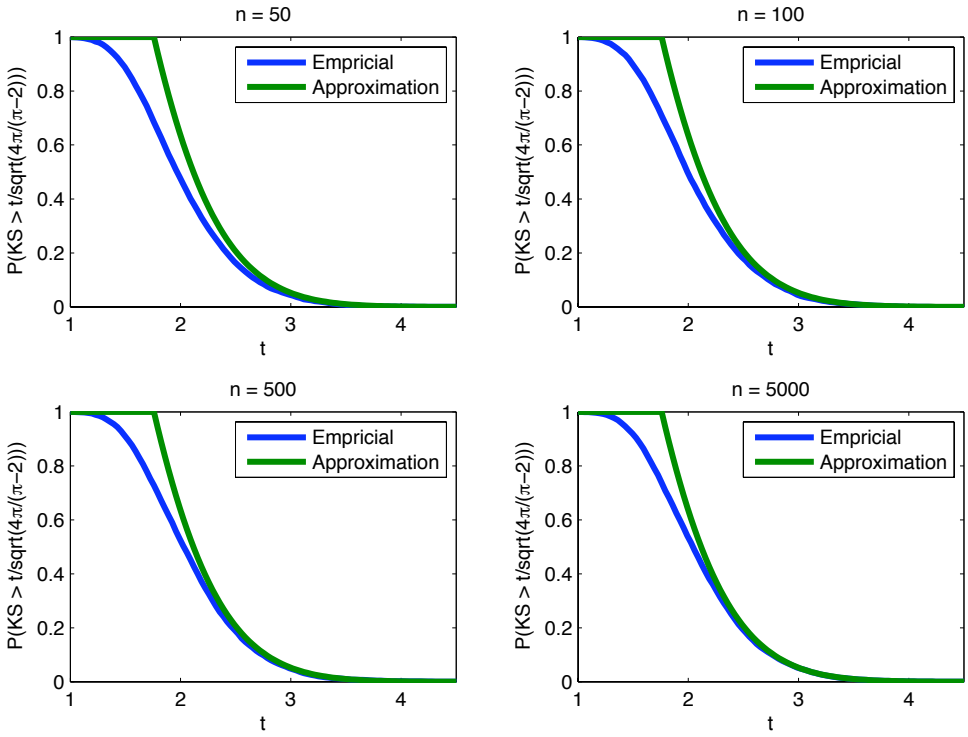

Figure 5.5: Experiment 3a. Comparison of $P(KS_n \geq \eta)$ (blue) and the approximation given in (5.2.9) (green). x-axis: η. y-axis: $P(KS_n^- \geq \eta)$.

tion. Figure 5.9 shows the corresponding fitting result for KS statistic and 3rd moment statistic.

According to Efron's idea, we should correct the parameters of statistics from empirical data.

In Figure 5.9, we can find that the fitting of KS statistic is much better than 3rd moment statistic. On the left figure, the two lines almost overlap, except the blue line is a little right shifted, which is because of signals. On the right figure, the fitting is not so good. The statistic from data is kind of far from the simulated distribution, which means that the fitting is bad.

Interestingly, here we assume the gene distributed as Gaussian distribution when they do not contain information, and the fitting result turns to be good. In fact, for most cases, the null distribution is unknown, and probably not Gaussian distribution. So, KS statistic turns to be much more robust than 3rd moment statistic. Even though

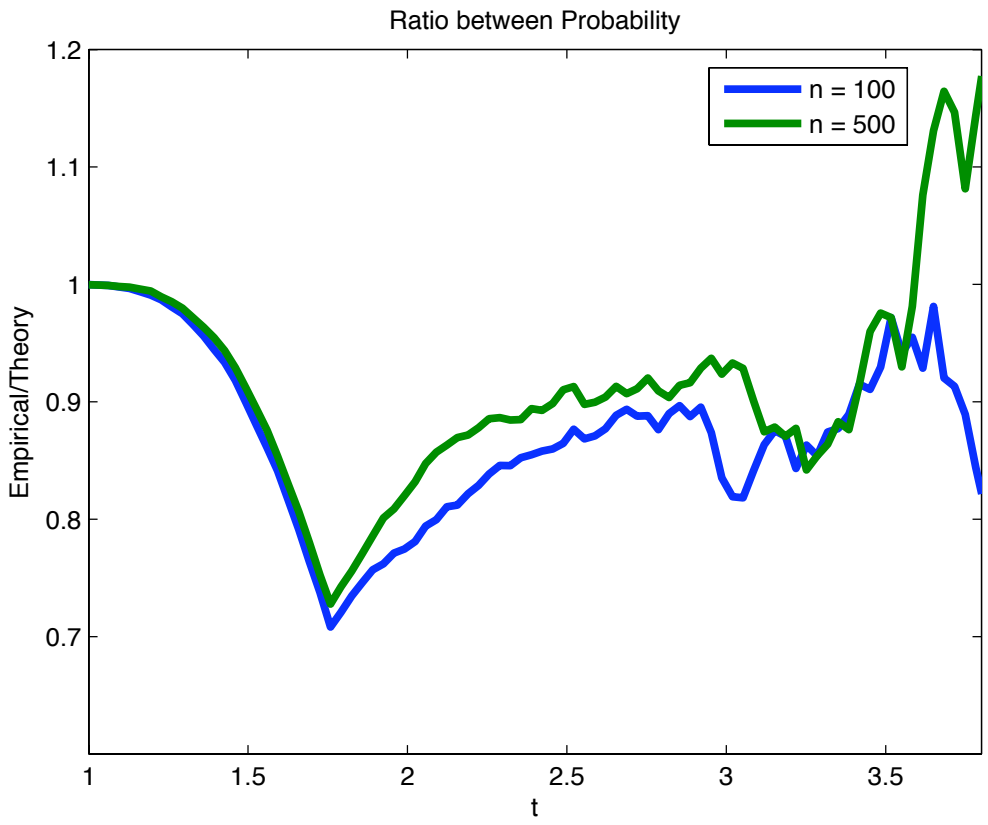

Figure 5.6: Experiment 3a. Ratio of empirical tail probability to approximated tail probability for $n = 100$ (blue) and $n = 500$ (green.) x-axis: η. y-axis: *Empirical/Approximated*.

the hypothesis does not stand for some genes, the performance of KS statistic is also similar to the case that the hypothesis stands, while for 3rd moment statistic, the result changes much.

The fitting of distribution can also be observed from the qq plot, which is as figure 5.10.

In Figure 5.10, the same problem happens, that the 3rd moment statistic fitting is not as robust as KS statistic.

As we zoom in the figure on the range $[0, 0.1]$, which is usually the range for informative genes, we have the following figures. 5.10.

In figure 5.11, we can find some interesting things. First, both statistics can recover

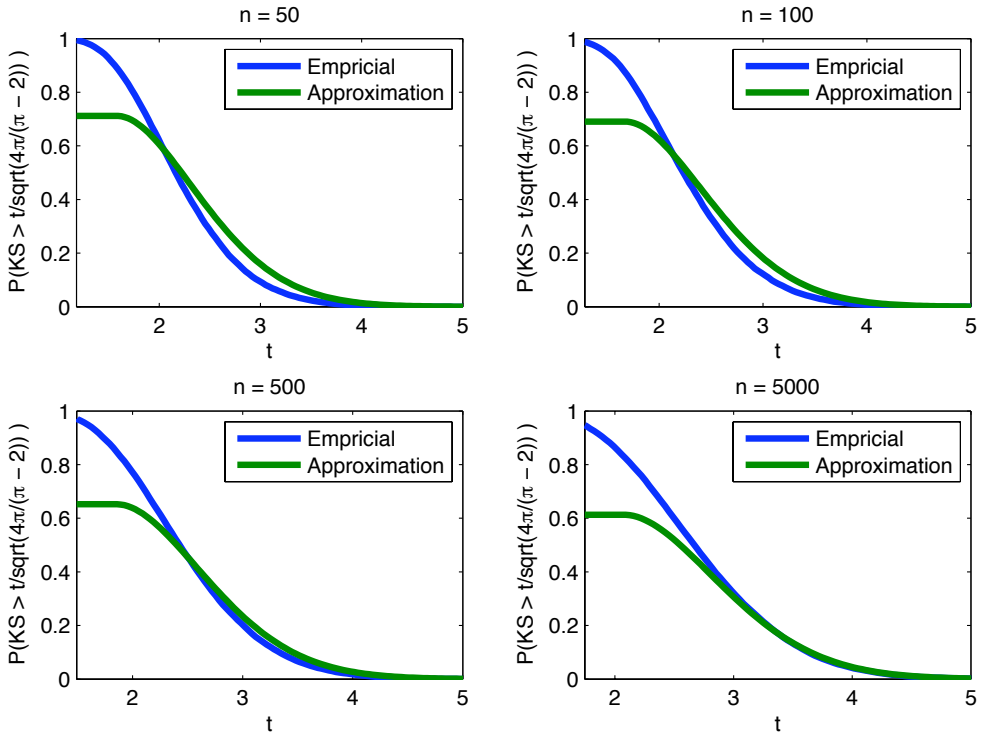

Figure 5.7: Experiment 4. Comparison of $P(KS_n \geq \eta)$ (blue) and the approximation given in Theorem 5.2.3 (green). x-axis: η. y-axis: $P(KS_n \geq \eta)$.

some genes, as the blue line is lower than the red line at the beginning, which means that the informative genes performed. Second, 3rd moment statistic is more powerful. It recovers some genes and then go back to the line quickly, while the KS statistic is always under the red line in left figure. At last, even 3rd moment statistic is powerful, the fitting is bad, even in this small range. It is far away from red line around 0.1. So, even though KS statistic is not as powerful as 3rd moment statistic, it is more robust, which is better for real data. That's why we choose KS statistic instead of moment statistic.

5.5.2 Goodness of fit for multiple data sets

For real data analysis, the approximation of tail probability with this easy formula (5.2.9) also fits data well. Here, we show that for micro-array data sets.

In micro-array data set, we have a set of n patients from two possible classes: normal

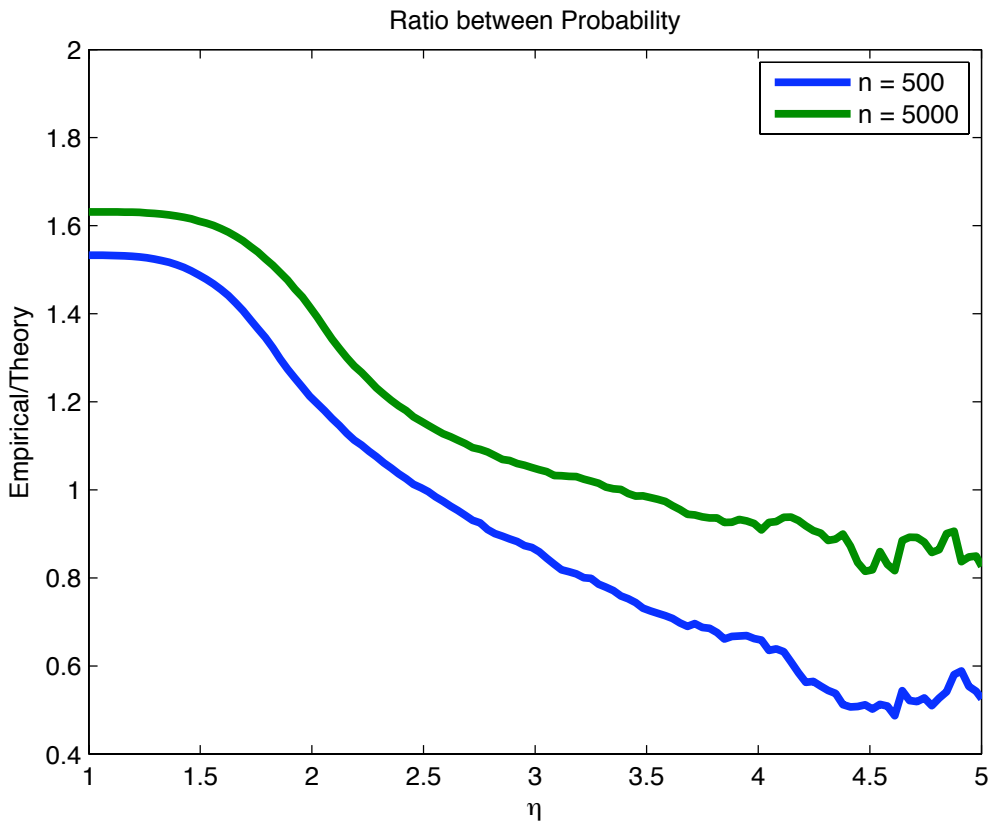

Figure 5.8: Experiment 4. Theoretical mean and standard deviation of KS_n (based on simulations) and approximated values (based on Theorem 5.2.3).

or diseased. The class label for each patient is unknown and it is of interest to find out. For each patient, a gene microarray chip is generated, with measurements on the same set of p genes. Table 5.3 display three such data sets, the Leukemia data analyzed in Golub et al (1999) ([Golub 1999]), the Colon cancer data analyzed in Alon et al (1999) ([Alon 1999]), and the Cardio data analyzed in [**?**]. In these examples, the number of genes p is much larger than the sample size n. Also, only a few fraction of genes contain information of classes, and the other genes have the same distribution for two classes. As most genes are from null distribution, we can assume that the KS statistic from the data is distributed as model (5.1.1), and compared the probability with (5.2.9). Here we assume that the noise is Gaussian distributed, and the result shows that the assumption works.

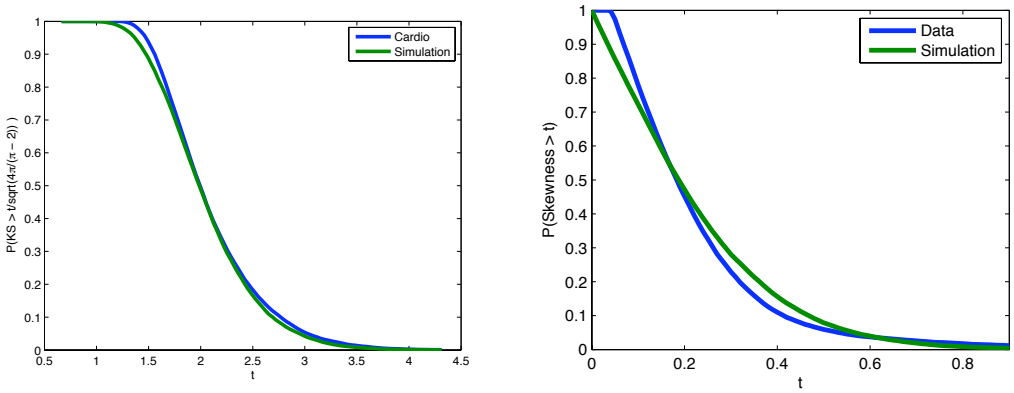

Figure 5.9: Cardio data: On the left, we show the tail probability of KS statistics from data (blue line) and simulation (green line). On right, we show the tail probability of 3rd moment statistics from data (blue line) and simulation (green line).

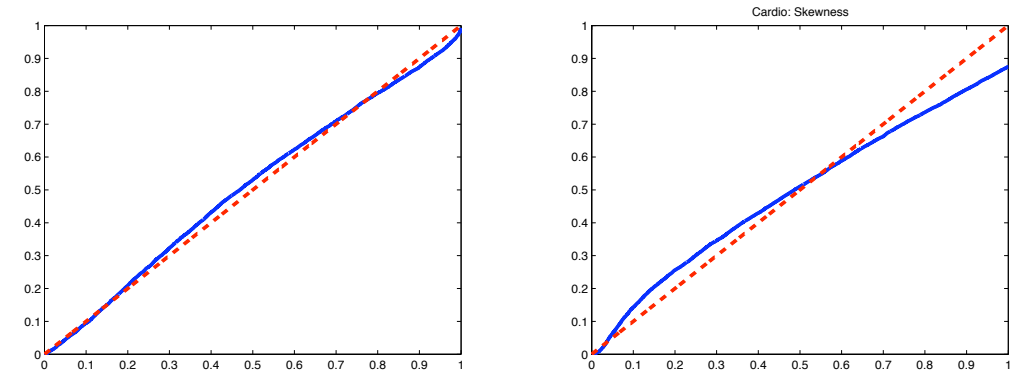

Figure 5.10: Cardio data: On the left, we show the qq plot of KS statistics from data (blue line) and simulation (green line). On right, we show the qq plot of 3rd moment statistics from data (blue line) and simulation (green line).

We remark that it is necessarily to correct the mean and standard deviation of $\{KS_j\}_{1\leq j\leq p}$, following Efron's idea of choosing a null distribution ([Efron 2004]).

Calculate the tail probability for three data sets with corrected mean and standard deviation. The comparisons of real data results and theoretical approximation for the three data sets are shown in Figure 5.12, Figure 5.13, and Figure 5.14. With Efron's correction, the tail distribution of KS statistic for real data sets is very similar with our approximation. It means that our approximation definitely can be used for real data

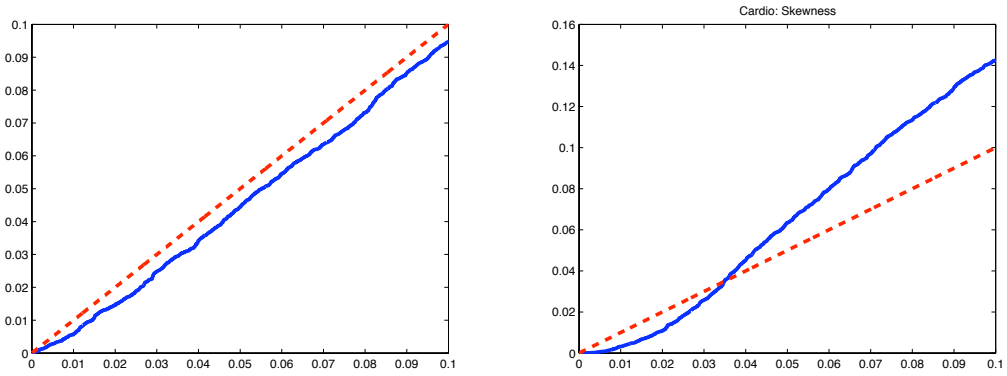

Figure 5.11: Cardio data: On the left, we show the qq plot of KS statistics from data (blue line) and simulation (green line). On right, we show the qq plot of 3rd moment statistics from data (blue line) and simulation (green line).

Data Name	Source	n, # samples	p, # features
Leukemia	[Golub 1999]	72	3571
Colon	[Alon 1999]	62	2000
Cardio	[?]	63	20426

Table 5.3: Three examples of microarray data sets.

analysis when the data fits Gaussian assumption.

5.6 Discussion

The proof depends on (a) an accurate approximation of the first passage probability, and (b) Borovkov and Rogozin approximation. Both approximations hold for general exponential families. Therefore, we expect similar results can be extended to exponential families.

5.7 Proofs

In this section, we gives the proofs of all theorems and lemmas.

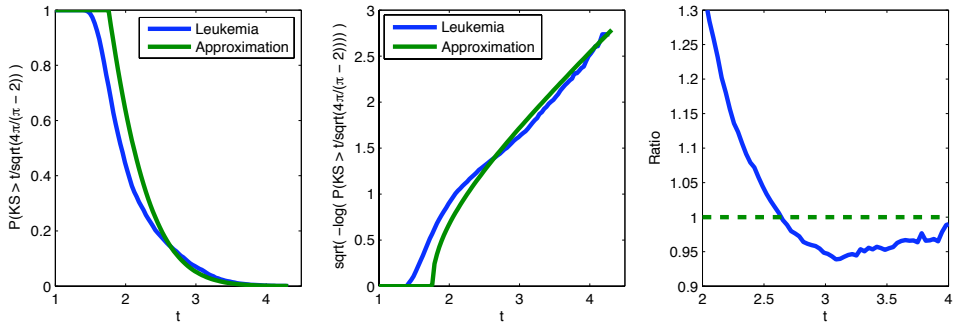

Figure 5.12: Leukemia data. The left figure shows comparison of $P(KS_n^- \geq t)$ (blue) and the approximation (5.2.9). The middle figure shows the comparison of exponential term, $\sqrt{-\log(P(KS > t))}$ for data (blue) and approximation (green). The right figure shows the ratio of the exponential term between data and theory.

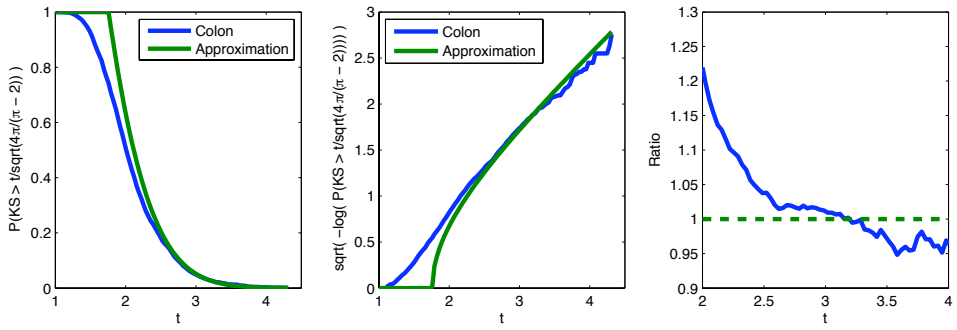

Figure 5.13: Colon data. The left figure shows comparison of $P(KS_n^- \geq t)$ (blue) and the approximation (5.2.9). The middle figure shows the comparison of exponential term, $\sqrt{-\log(P(KS > t))}$ for data (blue) and approximation (green). The right figure shows the ratio of the exponential term between data and theory.

5.7.1 Proof of Lemma 5.3.1

By Loader [Loader 1992, (11) and Lemma B.2],

$$P\big(\tau = t_j | G_n(t_j) = j, W = (0,n)'\big) = 1 - \Big[(\frac{\partial q_t}{dt})^{-1}\mu(t,q_t)\Big]\Big|_{t=t_j} + o(1), \qquad (5.7.16)$$

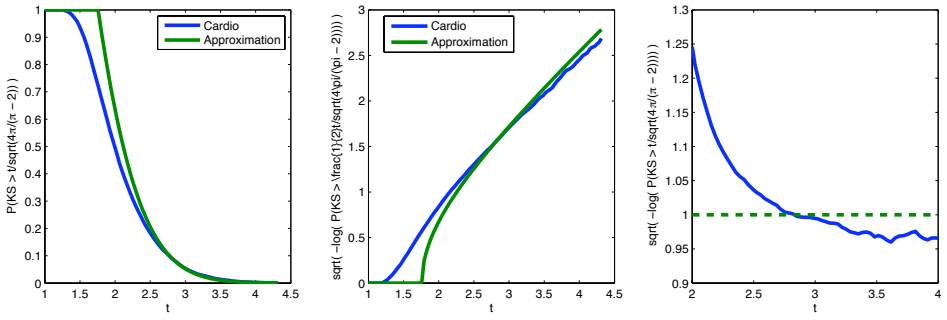

Figure 5.14: Cardio data. The left figure shows comparison of $P(KS_n^- \geq t)$ (blue) and the approximation (5.2.9). The middle figure shows the comparison of exponential term, $\sqrt{-\log(P(KS > t))}$ for data (blue) and approximation (green). The right figure shows the ratio of the exponential term between data and theory.

where $\mu(t, q_t)$ is defined in (5.7.24). We now use Lemma 5.7.2, and it is easy to find that

$$P\big(\tau = t_j \big| G_n(t_j) = j\big) = \frac{\Phi(-t_j) + t_j\phi(t_j)(1 + t_j^2/2)}{\sqrt{\Phi(t_j)\Phi(-t_j) - \phi^2(t_j)(1 + t_j^2/2)}}\epsilon_n(1 + o(1)).$$

So, the lemma is proved. □

5.7.2 Proof of Lemma 5.3.2

Consider $P\big(G_n(t_j) = j|W = (0, n)'\big)$ next. The following theorem is due to Borovkov and Rogozin [Borovkov 1965] and can be found for example in [Woodroofe 1978].

Lemma 5.7.1 *Consider an exponential family $f(x;\theta)$ of the form of*

$$f(x;\theta) = h(x)\exp(< \theta, x > -\psi(\theta)), \qquad x = (x_1, x_2, \cdots, x_d).$$

Then the density for $\sum_{i=1}^n X_i = nx$ is

$$f_0^{(n)}(x) = (1 + o(1)) \cdot (2\pi n)|^{-d/2}(\psi''(\hat{\theta})|)^{-1/2}e^{-n\ell_{\hat{\theta}}(x)},$$

where $\hat{\theta}$ is MLE from sufficient statistics $\sum_{i=1}^n X_i = nx$, and

$$\ell_{\hat{\theta}}(x) =< \hat{\theta} - \theta, x > -(\psi(\hat{\theta}) - \psi(\theta)).$$

We now show the lemma. Write

$$P\big(G_n(t_j)=j|W=(0,n)'\big)=P\big(G_n(t_j)=j,W=(0,n)'\big)/P\big(W=(0,n)'\big). \quad (5.7.17)$$

Consider $P\big(W=(0,n)'\big)$ first. Rewrite the density of $N(\mu,\sigma^2)$ as

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

where

$$y=x^2, \qquad \alpha=-\frac{1}{2\sigma^2}, \qquad \beta=\frac{\mu}{\sigma^2}. \quad (5.7.18)$$

Using Lemma 5.7.1 gives

$$P\big(W=(0,n)'\big)=(1+o(1))\cdot\frac{1}{2\pi n\sqrt{2}}e^{n(1/2+\alpha+\beta^2/4\alpha-\log(-2\alpha)/2)}. \quad (5.7.19)$$

Next, we consider $P\big(G_n(t_j)=j,W=(0,n)'\big)$. For any fixed t, we embed the above normal density into a *three-parameter exponential* family

$$f_t(x;\alpha,\beta,\delta)=\exp(\alpha x^2+\beta x+\delta 1\{x>t\}-\psi_t(\alpha,\beta,\delta)), \quad (5.7.20)$$

where

$$\psi_t(\alpha,\beta,\delta)=\frac{\beta^2}{4\alpha}-\frac{1}{2}\log\big(-\frac{\pi}{\alpha}\big)+\log\big(\Phi(-2\alpha t-\beta)+e^{\delta}\Phi(2\alpha t+\beta)\big). \quad (5.7.21)$$

For clarification, note that t a fixed number and is not a parameter here. Recall that $\Phi(x)$ is the CDF of $N(0,1)$. Also, note that when $\delta=0$, $f_t(x;\alpha,\beta,\delta)\equiv\phi(x;\mu,\sigma)$.

Now, first, we let

$$q_t=q_t(\eta,n)=\phi(t)-\eta/\sqrt{n}, \qquad h_t=h_t(\alpha,\beta)=-\sqrt{-2\alpha}t+\frac{\beta}{\sqrt{-2\alpha}}, \quad (5.7.22)$$

and let $(\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t))$ be the solution of the following equation system:

$$\begin{cases}\beta\Phi(h_t)\Phi(-h_t)-\sqrt{(1-\beta t)}[q_t-\Phi(h_t)]\phi(h_t)=0,\\ \alpha=(\beta t-1)/2,\\ e^{\delta}=\frac{q_t}{1-q_t}\Phi(-h_t)/\Phi(h_t).\end{cases} \quad (5.7.23)$$

Second, we let

$$\mu(t,q_t)=\frac{\exp(\alpha_n^*(t)t^2+\beta_n^*(t)t+\frac{\beta_n^*(t)^2}{4\alpha_n^*(t)}-\frac{1}{2}\log(-\pi/\alpha_n^*(t)))}{\Phi(\sqrt{-2\alpha_n^*(t)}t-\beta_n^*(t)/\sqrt{-2\alpha_n^*(t)})+e^{\delta_n^*(t)}\Phi(-\sqrt{-2\alpha_n^*(t)}t+\beta_n^*(t)/\sqrt{-2\alpha_n^*(t)})}, \quad (5.7.24)$$

$$\ell(t) = \frac{\hat{\beta}t}{2} + \frac{1}{2}\log(1-\hat{\beta}t) - \frac{\hat{\beta}^2}{2(1-\hat{\beta}t)} + (1-q_t)\log\frac{1-q_t}{\Phi(-h_t(\hat{\beta}))} + q_t\log\frac{q_t}{\Phi(h_t(\hat{\beta}))}, \quad (5.7.25)$$

where,

$$h_t(\hat{\beta}) = h_t(\beta)\Big|_{\beta=\hat{\beta}}.$$

Last, we let $\psi_t''(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))$ be the 3×3 Hessian matrix of $\psi_t(\alpha, \beta, \delta)$, evaluated at the point $(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))$, and let $t_0 = t_0(\eta, \hat{\mu}_n, \hat{\sigma}_n, n)$ be the solution of

$$\Phi(t; \hat{\mu}_n, \hat{\sigma}_n) = \eta/\sqrt{n}. \quad (5.7.26)$$

For this exponential family, the sufficient statistics are $(n - G_n(t), W)$. Similarly, applying Lemma 5.7.1 gives

$$P\big(G_n(t_j) = j, W = (0,n)'\big) \sim (2\pi n)^{-3/2}(|\psi_t''(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))|)^{-1/2}\exp(-n\ell_\theta(x)), \quad (5.7.27)$$

where $\psi_t''(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))$ is the Hessian matrix of ψ at the point $(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))$. $\alpha_n^*(t)$, $\beta_n^*(t)$, and $\delta_n^*(t)$ are MLE for α, β, δ. Inserting (5.7.19)-(5.7.27) into (5.7.17) and using the definition of conditional probability, we have gives

$$P\big(G_n(t_j) = j | W = (0,n)'\big) = \frac{\exp(-\frac{n\epsilon_n^2}{2(\Phi(t_j)\Phi(-t_j) - \phi^2(t_j)(1+t_j^2/2))})}{\sqrt{2\pi n(\Phi(t_j)\Phi(-t_j) - \phi^2(t_j)(1+t_j^2/2))}}(1+o(1)). \quad (5.7.28)$$

□

The claim now follows from the following lemma and direct calculations.

Lemma 5.7.2 *When $\epsilon_n = \eta_n/\sqrt{n}$ is very small, we have the following approximations for MLE $(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))$ as*

$$\begin{cases} \alpha_n^*(t) = (\beta_n^*(t)t - 1)/2, \\ e^{\delta_n^*(t)} = \frac{1-q_t}{q_t}\frac{\Phi(\frac{t-\beta_n^*(t)(t^2+1)}{\sqrt{1-\beta_n^*(t)t}})}{\Phi(\frac{\beta_n^*(t)(t^2+1)-t}{\sqrt{1-\beta_n^*(t)t}})}, \\ \beta_n^*(t) = -\frac{\phi(t)}{\Phi(t)\Phi(-t) - (t^2/2+1)\phi^2(t)}\epsilon + O(\epsilon^2). \end{cases}$$

Also, the other terms can be approximated as

$$2\ell(t) = \frac{1}{\Phi(t)\Phi(-t) - \phi^2(t)(1+t^2/2)}\epsilon_n^2 + O(\epsilon_n^3),$$

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

and

$$\det|\psi''(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))| = 2(\Phi(t)\Phi(-t) - (t^2/2+1)\phi^2(t)) + O(\epsilon_n^2).$$

Here, $\alpha_n^*(t)$, $\beta_n^*(t)$ and $\delta_n^*(t)$ are functions of n and t.

Lemma 5.7.2 is proved later in this section. □

5.7.3 Proof of Lemma 5.3.3

We want to analyze

$$\int_{-\sqrt{-\log(\epsilon_n)}}^{1/\sqrt{\epsilon_n}} \sqrt{n}\epsilon_n h(t) e^{-n\epsilon_n^2 g(t)} dt,$$

where $\epsilon_n = \eta_n/\sqrt{n}$ and

$$g(t) = \frac{1}{2}\frac{1}{\Phi(t)\Phi(-t) - \phi^2(t)(1+t^2/2)}, \qquad h(t) = \frac{\Phi(-t)\phi(t) + t\phi^2(t)(1+t^2/2)}{\sqrt{2\pi(\Phi(t)\Phi(-t) - \phi^2(t)(1+t^2/2))^3}}.$$

Lemma 5.7.3 *As a function of t, $\Phi(t)\Phi(-t)-\phi^2(t)(1+t^2/2)$ is symmetric and positive in $t \in (-\infty, \infty)$ and is strictly decreasing in $(0, \infty)$, reaches it maximum at $t = 0$.*

The function $g(t)$ of t is positive in $t \in (-\infty, \infty)$, and the first derivative is 0 at $t = 0$.

Proof. It is sufficient to prove the monotonicity. In fact, first, the symmetry follows trivially. Second, by Mills' ratio, $\Phi(t)\Phi(-t) - \phi^2(t)(1 + t^2/2) \gtrsim \Phi(-t)$ as $t \to \infty$, and the positivity follows by combining this with the monotonicity.

We now show the monotonicity. By direct calculations, for any $t \geq 0$, the derivative of the function at t is

$$\phi(t)[(t + t^3)\phi(t) - (2\Phi(t) - 1)].$$

It can be shown that the term in the bracket is strictly monotonely decreasing in $[0, \infty)$, with a maximum of 0 reached at $t = 0$, and so $\phi(t)[(t + t^3)\phi(t) - (2\Phi(t) - 1)] < 0$ for all $t > 0$. This gives the claim.

For $h(t)$, it is clear that $h(t) > 0$, and with calculation, $h'(0) = 0$. □

We now come back to the integration. Note that for $|t| \geq 1$, there is a constant $C > 0$ such that

$$g(t) \geq g(1) + C|t - 1|.$$

This can be shown by using the monotonicity of g and that

$$\lim_{t\to 1+} \frac{g(t) - g(1)}{t - 1} = g'(1) > 0, \qquad \lim_{t\to\infty} \frac{g(t) - g(1)}{t - 1} = \infty.$$

Now, write for short $b_n = a_n^{-5/12}$ so that $b_n \to 0$, $a_n b_n^3 \to 0$, but $a_n b_n^2 \to \infty$. We write the integral as

$$I + II + III,$$

where

$$I = \int_{|t|\le b_n} \sqrt{a_n} h(t) e^{-a_n g(t)} dt, \quad II = \int_{b_n \le |t| \le 1} \sqrt{a_n} h(t) e^{-a_n g(t)} dt, \quad III = \int_{|t|>1} \sqrt{a_n} h(t) e^{-a_n g(t)} dt.$$

First, note that $h(t)$ can be bounded by some constant as $h(t) \to 0$ and it is continuous. For sufficiently large n, we have

$$\int_{t\ge 1} \sqrt{a_n} h(t) e^{-a_n[g(1)+C|t-1|]} dt \le \sqrt{a_n} e^{-a_n g(1)} \int_{t\ge 1} h(t) e^{-a_n C|t-1|} dt \le \sqrt{a_n} C e^{-a_n g(1)},$$

and

$$\int_{t\le -1} \sqrt{a_n} h(t) e^{-a_n[g(1)+C|t-1|]} dt \le \sqrt{a_n} e^{-a_n g(1)} \int_{t\le -1} e^{-a_n C|t-1|+\log h(t)} dt \le \sqrt{a_n} C e^{-a_n g(1)}.$$

So we have that

$$III \le \sqrt{a_n} C e^{-a_n g(1)}.$$

Second, as $h(t)$ is continuous function on $[-1, -b_n]$ and $[b_n, 1]$, it is bounded by some constant C. Also, using monotonicity of $g(t)$ and basic calculus, for sufficiently large n,

$$II \le 2C\sqrt{a_n} e^{-a_n g(b_n)} \le \sqrt{a_n} C e^{-a_n[g(0)+g''(0)b_n^2/2+O(b_n^3)]},$$

and so

$$II \lesssim C\sqrt{a_n} e^{-a_n g(0) - g''(0) a_n b_n^2/2}.$$

Last,

$$I = \int_{-b_n}^{b_n} \sqrt{a_n} h(t) e^{-a_n g(t)} dt = \sqrt{a_n} e^{-a_n g(0)} \int_{-b_n}^{b_n} (h(0) + O(t^2)) e^{-a_n t^2 g''(0)/2 + O(a_n b_n^3)} dt.$$

By the choice of b_n and that $g(1) > g(0)$, it is seen that

$$I \sim \sqrt{a_n}\sqrt{2\pi} h(0) e^{-a_n g(0)} / \sqrt{a_n g''(0)} = \sqrt{2\pi} h(0) e^{-a_n g(0)} / \sqrt{g''(0)}, \qquad II + III = o(1) \cdot I.$$

Combining these gives that the integral

$$\int_{-\sqrt{-\log(\epsilon_n)}}^{1/\sqrt{\epsilon_n}} \sqrt{n}\epsilon_n h(t) e^{-n\epsilon_n^2 g(t)} dt \sim \sqrt{2\pi} h(0) e^{-a_n g(0)} / \sqrt{g''(0)}.$$

In our case, we have that $a_n = n\epsilon_n^2 = \eta_n^2$. To introduce the result of integration, we need $h(0)$, $g(0)$, and $g''(0)$. With fundamental calculation, we have that

$$g(0) = \frac{2\pi}{\pi - 2}, \qquad g''(0) = \frac{4\pi}{(\pi-2)^2}, \qquad h(0) = \frac{\sqrt{4\pi}}{(\pi-2)^{3/2}}.$$

Introduce these terms in, and we have the one sided probability as

$$P(KS_n^- > \eta_n) \leq \sqrt{\frac{2\pi}{\pi-2}} e^{-\frac{2\pi}{\pi-2}\eta_n^2}, \eta_n \to \infty.$$

So, the two-sided boundary crossing probability is

$$P(KS_n > \eta_n) \leq 2\sqrt{\frac{2\pi}{\pi-2}} e^{-\frac{2\pi}{\pi-2}\eta_n^2}, \eta_n \to \infty.$$

□

5.7.4 Proof of Lemma 5.7.2

To prove this problem, we go over the terms one by one.

First, let's have a look at the MLE. To solve the MLE, we have to solve the equation system (5.7.23) as following:

$$\begin{cases} \beta\Phi(h_t)\Phi(-h_t) - \sqrt{(1-\beta t)}[q_t - \Phi(h_t)]\phi(h_t) = 0, \\ \alpha = (\beta t - 1)/2, \\ e^{\delta} = \frac{1-q_t}{q_t}\Phi(-h_t)/\Phi(h_t). \end{cases}$$

Introduce the second equation $\alpha = (\beta t - 1)/2$ into $h_t(\alpha, \beta)$, and we got

$$h_t(\beta) = h_t(\alpha, \beta) = \frac{\beta(t^2+1) - t}{\sqrt{1-\beta t}}. \tag{5.7.29}$$

So the first equation becomes an equation about β only. The main problem is to calculate β from the first equation. Note that when $\epsilon_n = 0$, the first equation becomes

$$\beta\Phi(h_t)\Phi(-h_t) - \sqrt{1-\beta t}[\Phi(t) - \Phi(h_t)]\phi(h_t) = 0,$$

and the solution is $\beta = 0$ as $h_t = t$ when $\beta = 0$. So we guess that, when there is a small perturbation ϵ_n, the impact on β would be a small perturbation of β from 0. Based on the idea, we express the first equation by Taylor expansion. The corresponding expression is

$$\begin{aligned} &\beta(\Phi(-t)\Phi(t) + \beta(t^2/2+1)\phi(t)(\Phi(t) - \Phi(-t))) - \\ &\quad \sqrt{1-\beta t}(\Phi(t) - \epsilon_n - \Phi(t) + \tfrac{\beta(t^2+1)}{\sqrt{1-\beta t}}\phi(t))(\phi(-t) + \tfrac{\beta(t^2/2+1)}{\sqrt{1-\beta t}}t\phi(t)) = 0. \end{aligned}$$

Simplify it, and drop off the second order terms, we got the simplified result as

$$\beta\Phi(-t)\Phi(t) - \phi(t)(\beta(t^2/2+1)\phi(t) - \epsilon_n) + O(\epsilon_n^2) = 0,$$

so the approximated solution is

$$\beta_n^*(t) = -\frac{\phi(t)}{\Phi(t)\Phi(-t) - (t^2/2+1)\phi^2(t)}\epsilon_n.$$

With the expression of $\beta_n^*(t)$ and the two equations in the equation system, we get the approximation of $\alpha_n^*(t)$ and $\delta_n^*(t)$ as

$$\begin{cases} \alpha_n^*(t) = (\beta_n^*(t)t - 1)/2, \\ e^{\delta_n^*(t)} = \frac{1-q_t}{q_t}\frac{\Phi(-h_t(\beta_n^*(t)))}{\Phi(h_t(\beta_n^*(t)))}. \end{cases}$$

So the result about MLE is proved.

Now, from MLE, we could go on to calculate $\ell(t)$. The equation for $\ell(t)$ is

$$\ell(t) = (\alpha+1/2)+\delta(1-q_t)+\psi(\alpha=1,\beta=0)-\psi(\alpha,\beta)-\log(\Phi(-h_t(\alpha,\beta))+e^{\delta}\Phi(h_t(\alpha,\beta)))$$

Introduce $\alpha_n^*(t) = (\beta_n^*(t)t-1)/2$ and $e^{\delta_n^*(t)} = \frac{1-q_t}{q_t}\frac{\Phi(-h_t(\beta_n^*(t)))}{\Phi(h_t(\beta_n^*(t)))}$, and we get $\ell(t)$ at the point $(\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t))$ as

$$\ell(t)|_{\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t)} = \frac{\hat{\beta}_n t}{2}+\frac{1}{2}\log(1-\hat{\beta}_n t)-\frac{\hat{\beta}_n{}^2}{2(1-\hat{\beta}_n t)}+(1-q_t)\log\frac{1-q_t}{\Phi(h_t(\beta_n^*(t)))}+q_t\log\frac{q_t}{\Phi(-h_t(\beta_n^*(t)))}.$$

The first derivative of $\ell(t)$ about β is

$$\frac{\partial\ell(t)}{\partial\beta} = \frac{\beta(\beta t^3-2t^2+2\beta t-2)}{2(\beta t-1)^2} - \frac{\partial h_t(\beta)}{\partial\beta}\phi(\beta-t)(\frac{1-q_t}{\Phi(\beta-t)}-\frac{q_t}{\Phi(t-\beta)}),$$

and

$$\frac{\partial\ell(t)}{\partial\epsilon_n} = \log\frac{1-q_t}{q_t}\frac{\Phi(-h_t(\beta))}{\Phi(h_t(\beta))}.$$

When $\beta = 0$, we got that $\frac{\partial\ell(t)}{\partial\beta} = 0$, and $\frac{\partial\ell(t)}{\partial\epsilon_n} = 0$, so the first order terms cancel. Calculate the second order term, we have that

$$\frac{\partial^2\ell(t)}{\partial\beta^2}|_{(0,0)} = -(t^2/2+1)\frac{(t^2/2+1)^2\phi^2(t)}{\Phi(-t)\Phi(t)},$$

$$\frac{\partial^2\ell(t)}{\partial\beta\partial\epsilon_n}|_{(0,0)} = \frac{\phi(t)}{\Phi(t)\Phi(-t)}(t^2/2+1),$$

and

$$\frac{\partial^2\ell(t)}{\partial\epsilon_n^2}|_{(0,0)} = \frac{1}{\Phi(t)\Phi(-t)}.$$

So, we got an approximation for $\ell(t)$, which is

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

The result for $\ell(t)$ is proved.

The third term is $\phi(t) - \mu(t, q_t)$. Let's have a look at the approximation of it. The exact for for $\mu(t, q_t)$ is

$$\mu(t, q_t) = \frac{\exp(\alpha t^2 + \beta t + \frac{\beta^2}{4\alpha} - \frac{1}{2}\log(-\pi/\alpha))}{\Phi(-h_t(\alpha, \beta)) + e^{\delta}\Phi(h_t(\alpha, \beta))}.$$

Introduce the MLE $\alpha_n^*(t)$, $\beta_n^*(t)$, and $\delta_n^*(t)$ in the formula, we have

$$\mu(t, q_t)|_{\alpha_n^*(t), \beta_n^*(t), \delta_n^*(t)} = \frac{q_t}{\Phi(-h_t(\beta_n^*(t)))}\phi(h_t(\beta_n^*(t))).$$

Try to simplify it, notice that

$$h_t(\beta) = \frac{\beta(t^2+1) - t}{\sqrt{1 - \beta t}} = -t + \beta(1 + t^2/2) + O(\beta^2),$$

so we have

$$\begin{aligned} \mu(t, q_t) &= \frac{\Phi(t) - \epsilon_n}{\Phi(t) - \beta(1 + t^2/2)\phi(t)}[\phi(t) + t\phi(t)\beta(1 + t^2/2)] \\ &= \phi(t)[1 - \frac{\epsilon_n}{\Phi(t)} + \beta(1 + t^2/2)\frac{\phi(t)}{\Phi(t)} + \beta t(1 + t^2/2)]. \end{aligned}$$

So, the approximation for $\phi(t) - \mu(t, q_t)$ is

$$\begin{aligned} \phi(t) - \mu(t, q_t) &= \frac{\phi(t)}{\Phi(t)}\epsilon_n - \phi(t)(t^2/2 + 1)(\frac{\phi(t)}{\Phi(t)} + t)\beta_n^*(t) \\ &= \phi(t)\frac{\Phi(-t) + t\phi(t)(1 + t^2/2)}{\sqrt{\Phi(t)\Phi(-t) - \phi^2(t)(1 + t^2/2)}}\epsilon_n + O(\epsilon_n^2). \end{aligned}$$

The most complicated term is the Hessian matrix of $\psi(\alpha, \beta, \delta)$. According to delicate calculation, we could find each element is

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

$$\frac{\partial^2\psi(\alpha,\beta,\delta)}{\partial\beta^2} = -\frac{1}{2\alpha} + \frac{(e^\delta-1)^2\phi^2(-h_t(\alpha,\beta))}{2\alpha(\Phi(-h_t(\alpha,\beta))+e^\delta\Phi(h_t(\alpha,\beta)))^2} + \frac{-h_t(\alpha,\beta)\phi(-h_t(\alpha,\beta))(e^\delta-1)}{\sqrt{-2\alpha}(\Phi(-h_t(\alpha,\beta))+e^\delta\Phi(h_t(\alpha,\beta)))},$$
$$\frac{\partial^2\psi(\alpha,\beta,\delta)}{\partial\beta\partial\delta} = \frac{e^\delta\phi(h_t(\alpha,\beta))}{\sqrt{-2\alpha}(\Phi(-h_t(\alpha,\beta))+e^\delta\Phi(h_t(\alpha,\beta)))} - \frac{e^\delta\Phi(h_t(\alpha,\beta)\phi(h_t(\alpha,\beta)e^\delta-1))}{\sqrt{-2\alpha(\Phi(-h_t(\alpha,\beta))+e^\delta\Phi(h_t(\alpha,\beta)))^2}},$$
$$\frac{\partial^2\psi(\alpha,\beta,\delta)}{\partial\delta^2} = \frac{e^\delta\Phi(h_t(\alpha,\beta)\Phi(-h_t(\alpha,\beta)))}{(\Phi(-h_t(\alpha,\beta))+e^\delta\Phi(h_t(\alpha,\beta)))^2}.$$

With these terms we have the Hessian matrix. Introduce the MLE approximation $(\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t))$, we get that the determinant for it is

$$\det|\psi''(\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t))| = 2(\Phi(t)\Phi(-t)-(1+t^2/2)\phi^2(t))+O(\epsilon_n^2).$$

Now, all the terms are calculated, and the result is proved. The approximation here is important, as we can also use it for algorithms. □

5.7.5 Proof of Theorem 5.2.2

For model (5.1.2), note that $\hat{\mu}$ and $\hat{\sigma}$ changes the location and scale only, which does not impact the distribution of KS statistic. So, without loss of generality, we assume that $\sum_{i=1}^n X_i = 0$, $\sum_{i=1}^n X_i^2 = n$.

Lemma 5.7.4 *When the data is distributed under model (5.1.2), with probability* $1+o(1/n)$*, the corresponding KS statistic can be approximated by*

$$KS_n = \sqrt{n}\sup_{-\infty<t<\infty}|F_n(t,Y)-\Phi(t;\hat{\mu}_{n,y},\hat{\sigma}_{n,y})-\phi(t;\hat{\mu}_{n,y},\hat{\sigma}_{n,y})\frac{\sqrt{15}}{6}\tau(1-t^2)/\sqrt{n}+O(\mu^4))|,$$

where $Y=\Phi^{-1}(F(t))$*, which is a sample from standard normal distribution.*

Now we try to find the expectation of KS statistcs. According to this lemma, we have that

$$KS_n \le \sqrt{n}\sup_{-\infty<t<\infty}|F_n(t,Y)-\Phi(t;\hat{\mu}_{n,y},\hat{\sigma}_{n,y})|+\sup_{-\infty<t<\infty}|\phi(t;\hat{\mu}_{n,y},\hat{\sigma}_{n,y})\frac{\sqrt{15}}{6}\tau(1-t^2)|+O(\sqrt{n}\mu^4).$$

The last term $\sqrt{n}\mu^4\to 0$ as $n\to\infty$, so we can ignore it. The first term is the standard KS statistic, where we have found the mean around 0.64, and maximum around 1.4. The second term will achieve maximum at $t=0$, which is $a_0\tau_n$. So, as $\tau_n\to\infty$, the second term would be the most important, and we have that

$$E[KS_n]\le C_1+a_0\tau_n, \qquad E[KS_n]\ge -C_1+a_0\tau_n,$$

where C_1 is the expectation of KS statistics under model (5.1.1), which is a constant around 0.6.

So we have

$$\frac{-C_1}{a_0\tau_n}+1 \le E[KS_n]/a_0\tau_n \le \frac{C_1}{a_0\tau_n}+1,$$

which means that $\frac{E[KS_n]}{\tau_n} \to 1$ when $\tau_n \to \infty$.

Similarly, we can find the center for KS_n^+ and KS_n^-. □

5.7.6 Proof of Theorem 5.2.3

For the second claim, we also use Lemma 5.7.4 to calculate. According to this lemma, the calculation the tail probability for KS statistic under model (1.2) is similar with the KS statistic under model (1.1), with boundary at $q_t = \Phi(t) - \eta/\sqrt{n} + \frac{\sqrt{15}}{6}\tau(1-t^2)/\sqrt{n}$. Recall that $a = \frac{\eta_n}{\tau_n}$, then we have the boundary as

$$q_t = \Phi(t) - (1 - \frac{\sqrt{15}}{6}(1-t^2)\phi(t)/a)\eta_n/\sqrt{n}.$$

Now, there are two possible behavior of a.

- Case 1: when $a < a_0$, then $q_t < \Phi(t)$ over an open interval containing $t = 0$. So, Loader's approximation could not be applied to this case.
- Case 2: when $a > a_0$, we have some relevant results, with the proof very similar to KS statistic under model (1.1).

For case 1, we have that

$$KS_n \ge -\sqrt{n}\sup_{-\infty<t<\infty}|F_n(t,Y)-\Phi(t;\hat\mu_{n,y},\hat\sigma_{n,y})|+\sup_{-\infty<t<\infty}|\phi(t;\hat\mu_{n,y},\hat\sigma_{n,y})\frac{\sqrt{15}}{6}\tau(1-t^2)|+O(\sqrt{n}\mu^4).$$

The maximum of the second term is $a_0\tau_n$, so the probability that $KS_n \le \eta_n$ is smaller than the probability that

$$P(\sqrt{n}\sup_{-\infty<t<\infty}|F_n(t,Y)-\Phi(t;\hat\mu_{n,y},\hat\sigma_{n,y})| \ge \eta_n - a_0\tau_n) = 2\sqrt{\frac{2\pi}{\pi-2}e^{-\frac{2\pi}{\pi-2}(\eta_n-a_0\tau_n)^2}}.$$

When $\eta_n \to \infty$, the probability for it goes to 0. So we have that $P(KS_n \ge \eta_n) \to 1$.

For case 2, where $a > a_0$, introduce it into Lemma 5.7.2. Set the new boundary as

$$q_t = \Phi(t;\hat\mu_{n,y},\hat\sigma_{n,y}) - \frac{\eta_n}{\sqrt{n}} + \phi(t;\hat\mu_{n,y},\hat\sigma_{n,y})\frac{\sqrt{15}}{6}(1-t^2)\tau_n/\sqrt{n},$$

then apply the similar proof as Theorem 5.2.1. As the boundary has changed, we have new estimation of conditional probability $P(G(t_j) = j|W = (0,n)')$ and the first passage probability. Let $\epsilon_n = \eta_n/\sqrt{n}$, we have that

$$2\ell(t) = \frac{(1-\frac{\sqrt{15}}{6}(1-t^2)\phi(t)/a)^2}{\Phi(t)\Phi(-t)-\phi^2(t)(1+t^2/2)}\epsilon_n^2 + O(\epsilon_n^3),$$

$$\phi(t)-\mu(t,q_t)=\phi(t)((1-\frac{\sqrt{15}}{6}(1-t^2)\phi(t)/a))\frac{\Phi(-t)+t\phi(t)(1+t^2/2)}{\sqrt{\Phi(t)\Phi(-t)-\phi^2(t)(1+t^2/2)}}\epsilon_n+O(\epsilon_n^2),$$

and

$$\det|\psi''(\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t))|=2(\Phi(t)\Phi(-t)-(t^2/2+1)\phi^2(t))+O(\epsilon_n^2).$$

So, the probability becomes that

$$P(KS_n^-\geq\eta_n)=\int_{t:q_t>0}^{+\infty}\frac{\phi(t)-\mu(t,q_t)}{\sqrt{\pi\det|\psi''(\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t))|}}e^{-n\ell(t)}(1+o(1)),$$

In this case, note that we still have $\ell(t)$ as symmetric and positive function. Also, $\ell(t)$ achieves the minimum at $t=0$, and $\ell(t)\approx\frac{1}{\Phi(-t)}\to\infty$ when $t\to\infty$. When $0<t\leq\sqrt{3}$, $\ell(t)$ is monotone decreasing function, and when $t>\sqrt{3}$, the minimum of $\ell(t)$ is much larger than $\ell(0)$. So, we can decompose the integration into $I+II+III$, similar as proof of theorem 5.2.1. So, if we set

$$h(t)=\frac{\phi(t)-\mu(t,q_t)}{\sqrt{\pi\det|\psi''(\alpha_n^*(t),\beta_n^*(t),\delta_n^*(t))|}}=\frac{\phi(t)((1-\frac{\sqrt{15}}{6}(1-t^2)\phi(t)/a))\Phi(-t)+t\phi(t)(1+t^2/2)}{\sqrt{2\pi((\Phi(t)\Phi(-t)-(t^2/2+1)\phi^2(t))+O(\epsilon_n^2))^3}},$$

we have that

$$P(KS_n^-\geq\eta_n)\sim\sqrt{2\pi}h(0)e^{-g(0)\eta_n^2}/\sqrt{g''(0)}.$$

As $h(0)=\frac{\sqrt{2}\sqrt{2\pi}}{(\pi-2)\sqrt{\pi-2}}(1-\sqrt{\frac{5}{24\pi}}/a)$, $g(0)=\frac{2\pi}{\pi-2}(1-\sqrt{\frac{5}{24\pi}}/a)^2$, and $g''(0)=\frac{4\pi}{(\pi-2)^2}(1-\sqrt{\frac{5}{24\pi}}/a)(1-(7-3\pi)\frac{5}{24\pi}/a)$, so the one side result is

$$P(KS_n^-\geq\eta_n)=\sqrt{\frac{\pi}{\pi-2}}\sqrt{\frac{\eta_n-\sqrt{5/24\pi}\tau_n}{\eta_n-(7-3\pi)\sqrt{5/24\pi}\tau_n}}e^{-\frac{2\pi}{\pi-2}(\eta_n-\sqrt{5/24\pi}\tau_n)^2}(1+o(1)).$$

Recall that $a_0=\sqrt{5/24\pi}$, $b_0=(7-3\pi)a_0$, we have that

$$P(KS_n^-\geq\eta_n)=\sqrt{\frac{\pi}{\pi-2}}\sqrt{\frac{a-a_0}{a-b_0}}e^{-\frac{2\pi}{\pi-2}(a-a_0)^2\tau_n^2}(1+o(1)).$$

For KS_n^+, it is a little different. as the part $-\phi(t)\frac{\sqrt{15}}{6}\tau(1-t^2)$ would be the most important part, we find that it is smaller than 0 when $t=0$. So, the maximum would achieved when t is large. According to calculation, the maximum will be around $t=\pm\sqrt{3}$, and the corresponding result is $\sqrt{\frac{5}{6\pi}e^{-3}}\tau_n$, which is much smaller than KS_n^-. So, the probability that $KS_n^+>\eta$ will be $o(1)$ order compared to KS_n^-, and then we have the probability for KS_n is the same with KS_n^-. □

Chapter 6

Fundamental Limits for Matrix Recovery Problems

Contents

6.1 Introduction

Nowadays, high dimensional data analysis is a topic of broad interest in both statistic and machine learning. For the high dimensional data, many classical problems have new challenges for statisticians. Take the clustering problem as an example. In classical setting, we solve the clustering problem easily by summing over the features over one sample. When the dimension is low, the noise is weak, and then the sign of summation can differentiate two groups. However, it won't happen in high dimensional data, as the noise will be too strong. This circumstance is called "Curse of Dimensionality". In my thesis, I studied the problem for the clustering problem, signal recovery problem, and detection problem for high dimensional data.

For simplicity, we assume a two-class clustering model. Assume that there are two clusters in the model. In other words, all the samples can be assigned with labels $Y_1, Y_2, \ldots, Y_n$, where $Y_i \in \{-1, 1\}$ without loss of generality. For each sample i, the information of feature is recorded in a $p \times 1$ vector X_i. Here, Y_i are unknown. The problem here is to recover Y. In many applications (Genomics, e.g.), Big Data are of interest, which means that both p and n are large, and $p \gg n$.

Regard to this model, we have the sparsity assumption. Sparsity means that only a few features contain information, and most features act as noise. However, we don't know which are useful features. So there are three main problems regarding to this model for different applications. First, how to recover the labels Y_i? Second, how to recover the features that contain information? Third, how to detect whether there is information in X or not? We call them **Clustering** problem, **Signal Recovery** problem, and **Detection** problem. In our analysis, we focus on clustering problem. The signal recovery problem and detection problem will be referred in Section 6.3. In all, we want to recover the information from the noisy data matrix, that's why it is called matrix recovery problem.

Nowadays, there are many methods to solve the clustering problems. For example, hierarchical clustering, k-means, exhausted search and so on are all generally used classical clustering methods. For high dimensional data, new versions of these classical clustering methods are also be found. With all these methods, a natural question comes out: how good can we do? By intuition, when the signal is too weak and sparse, any method will fail. Otherwise, when the signal is very strong, any method will work. For an appropriate range of signal, only some method work. So, what is the boundary between the case that any method fail and some method work? At what level, all of the methods will fail? On the contrary, which method will work at the boundary? Is there any algorithm for this method? That's the problem we are interested in.

6.1.1 Asymptotic rare and weak model

In the two-class clustering model for high dimensional data, it is assumed that the information is included in a low rank model compared to the high dimensional data matrix for sparsity. In the simplified clustering case, where we take the sample information as an $n \times 1$ vector ℓ, where $\ell_i \in \{-1, 1\}$ is the label for samples. Assume that ℓ is Bernoulli sample from $\{-1, 1\}$, with prameter δ. The feature information is taken as a $p \times 1$ vector μ, where $\mu_j \in \{\eta, 0\}$, and η is some constant. So, we assume that the signal is either a constant or 0, which is a simplified case. In this case, the data matrix X is the rank 1 information matrix $\ell\mu'$ covered by an $n \times p$ noise matrix. When we take the noise as standard normal noise, the model can be written as

$$X = \ell\mu' + Z, \qquad Z_{ij} \overset{i.i.d}{\sim} N(0, 1),$$

where

$$\ell_i = 2Bernoulli(\delta) - 1, \qquad \mu_j = \begin{cases} \eta, & \text{with prob } \epsilon, \\ 0, & \text{with prob } 1-\epsilon. \end{cases}$$

According to the sparsity of the model, we take $\epsilon \to 0$ as $p \to \infty$, which means that the number of useful features is asymptotically sparse. Also, we take $\eta \to 0$ as $p \to \infty$, which means that the signal is asymptotically weak. With these two properties, this model is called "Asymptotic Rare and Weak (ARW)" model. Note that the definition of ARW model here is different from what we defined in Chapter 3.

More specifically, we use p as the driving asymptotic parameter, and say that we have

$$n = n_p = p^{\theta}, \quad \epsilon = \epsilon_p = p^{-\vartheta}, \quad \eta = \eta_p = p^{-\beta}.$$

This is the model for this paper.

Under this model, the 3 problems can be described as

- **Clu**stering Problem: recover ℓ;
- **Sig**nal Recovery Problem: recover μ;
- **Det**ection Problem: test whether $\ell\mu' = 0$ or not.

For problem $(*)$, we hope to find a boundary $f^*(\vartheta, \theta)$, such that when $\beta > f^*(\vartheta, \theta)$, the problem is impossible to solve. When $\beta < f^*(\vartheta, \theta)$, there is some method to solve the problem. Now, even though the problem can be solved, the method might be NP-hard, and we also discuss it in our paper.

Today, we want to focus on clustering problem. For the simplified ARW model, the signal recovery problem and detection problem are also solved, which will be discussed in the extention section.

6.1.2 Fundamental Limit for Clustering Problem under ARW model

Say that $\hat{\ell}$ is the estimation of ℓ from some method. As we have the estimation $\hat{\ell}$, we define the error rate as the Hamming distance,

$$Hamm(\hat{\ell}; \vartheta, r, \theta) = \frac{1}{n} \sum_{i=1}^{n} E[\ell_i \neq \hat{\ell}_i].$$

Obviously, even in the worst case where there is no information about ℓ, the error rate will be no larger than $1/2$. When $Hamm(\hat{\ell}) \to 1/2$ for any method, we say that it is the impossible area. When the method works, $Hamm(\hat{\ell})$ should go to 0 as $p \to \infty$. Now we want to find the boundary between the two cases.

How to find the boundary? There are two statements to confirm, which are

(a) Possibility: When the signal strength is larger than the boundary strength, there is some method with error rate going to 0;

(b) Impossibility: When the signal strength is smaller than the boundary strength, any method would have error rate going to 1/2.

To prove the impossibility, we use L^1 distance. With intuition, when the distance between two hypotheses goes to 0 when $p \to \infty$, then any method will fail. So, the calculation of L^1 distance will give a lower bound of signal strength for impossible area, which means an upper bound of β.

Now that we have the lower bound of signal strength from L^1 distance, we want to find an upper bound. For every method, we have a bound. When the signal strength is at that bound, there is a method that solve the clustering problem. So, it is an upper bound of signal strength. If an upper bound from some method meets the lower bound from L^1 distance, then the boundary is found out. Also, the method works at the boundary is also found out.

With the idea and some proof, we get the main theorem.

Theorem 6.1.1 *In the setting above, the error rate of recovering label ℓ goes to 1/2 for any method when*

$$\beta > f^{clu}(\vartheta, \theta),$$

and either the exhaust search or classical method would have error rate go to 0 when

$$\beta < f^{clu}(\vartheta, \theta),$$

where

$$f^{clu}(\vartheta, \theta) = \begin{cases} \frac{1}{2} - \vartheta, & \vartheta < \frac{1-\theta}{2}, \\ \theta/2, & \frac{1-\theta}{2} < \vartheta < 1 - \theta, \\ \frac{1-\vartheta}{2}, & \vartheta > 1 - \theta. \end{cases}$$

In the theorem, we got the boundary for clustering problem under our model, and the method at the boundary. We will introduce the two methods in Section 6.2.

6.1.3 Related Works

Compare the result with the analysis of combinatorial testing problem([Addario-Berry 2010]). In that paper, when the size of set is given, for all the K sets, the signal strength is required to be in the order of $\mu \sim \sqrt{p/K^2}$. However, in that setting, the label effect is not considered in. It could be extended here.

6.1.4 Content

In Section 6.2, we show the proof for main theorem from two aspects. We also extend the result to signal recovery problem and detection problem under ARW model, with the corresponding result, in Section 6.3.

6.2 Main Proof

There are two statements to prove here. First, when $\beta > f^{clu}(\vartheta, \theta)$, the Hamming distance will go to $1/2$. Second, when $\beta < f^{clu}(\vartheta, \theta)$, the clustering error rate for either exhaust search or classical clustering will go to 0.

We prove them in following sections.

6.2.1 Lower Bound

Assume we know the labels for all but one samples, and we want to find the label for that sample, then it is a classification problem. There is more information in the classification problem. So, if we cannot solve the classification problem, then we cannot solve the clustering problem. According to the analysis, we calculate the L^1 distance for the classification problem.

Say that $\tilde{\ell}$ is the true label for $X_2, \cdots, X_n$, which is an $(n-1) \times 1$ vector. Now we want to calculate the probability density function for X when the label of X_1 is 1 and -1. These two cases are symmetric. So the problem is equivalent with whether the distance goes to 0, which is the distance between the pdf of X when the label of X_1 is 1 and 0. In calculation, there are two cases. When the signal is sparse, we use Cauchy ineuqality to decompose the integral of L^1 distance into the integral over X, μ and X_1, μ. Then, by direct calculation we can find the bound of L^1 distance. When the signal is dense, we have to use Hellinger distance to control L^1 distance, and then apply similar method as classication problem ([Jin 2009]) to find the bound.

As a result, we get the following lemma. The detailed proof can be found in Section 6.4.

Lemma 6.2.1 *Under ARW model, the error rate of recovering label ℓ goes to 1/2 for any method when*

$$\beta > f^{clu}(\vartheta, \theta),$$

where

$$f^{clu}(\vartheta, \theta) = \begin{cases} \frac{1}{2} - \vartheta, & \vartheta < \frac{1-\theta}{2}, \\ \theta/2, & \frac{1-\theta}{2} < \vartheta < 1-\theta, \\ \frac{1-\vartheta}{2}, & \vartheta > 1-\theta. \end{cases}$$

The proof of lower bound indicates us to find possible methods according to whether the signal is dense or sparse, which is the idea in upper bound.

6.2.2 Upper Bound

Now, with the lower bound, we hope that we can find some method that meet the lower bound. For any method, there is some area that the method would work, which is the upper bound for our problem. If there is some method, such that the bound for it meets the lower bound, then we can say that the boundary is exactly the boundary of possible area and impossible area.

Is there some method that just meets the lower bound? Luckily we found it. It differs according to the sparsity of signal. When the signal is sparse, we use exhaust search to find the true label and signals. When the signal is dense, the true set of signal is not so important, and we use classical method to find the true label.

6.2.2.1 Classical method

When we say classical method, we mean that we estimate ℓ by

$$\hat{\ell}_i = \mathrm{sgn}[\sum_{j=1}^{p} X_{ij}].$$

As the signal is dense, the noise generated by $\sum_{j=1}^{p} X_{ij}$ cannot cover the signal, so the sign of it will indicate the label of Y_i. That's the idea of classical method.

According to our assumption, the summation of X_{ij} over j is a normal variable with variance as p and mean as $\eta \times$ #signals. When the mean is much larger than the standard deviation, the classical method will work. Some calculation shows that we need $\eta > \frac{1}{\sqrt{p\epsilon}}$ to make sure that the classical method works. That's the boundary for classical method.

Introduce the assumption that $\eta_p = p^{-\beta}$ and $\epsilon_p = p^{-\vartheta}$, the condition $\eta > \frac{1}{\sqrt{p\epsilon}}$ is equivalent with $\beta < 1/2 - \vartheta$, which meets the lower bound when $\vartheta < \frac{1-\theta}{2}$.

6.2.2.2 Exhaust Search

For the approach of exhaust search, search over all the possible combinations of signals. Without loss of generality, assume that we know the number of signals, as $k = p\epsilon$. Let ν be a $p \times 1$ vector, where $\|\nu\|_0 = k$. When $\nu'\mu = 0$, then $X\nu$ is noise, and there is no information about ℓ. When $\nu = \mu$, then the sign of $X\nu$ indicates the label vector ℓ. Let $\hat{\ell} = \mathrm{sgn}(X\nu)$, then $\hat{\ell}' X\nu$ will be very large when ν is truly discovered.

According to the idea, there are three steps in exhaust search approach.

- Calculate $\|X\nu\|_1 = \sum_{i=1}^{n} |(X\nu)_i|$ over all possible ν with $\|\nu\|_0 = k$;
- Find $\hat{\nu}$ that maximizes $\|X\nu\|_1$;
- $\hat{\ell} = \mathrm{sgn}(X\hat{\nu})$.

When does exhaust search work? Given the size k, then the number of all possible ν is about p^k. When $\nu'\mu = 0$, then $X'\nu$ is standard multivariate normal distribution. According to the properties of folded normal distribution, the mean of it is $n\sqrt{2k\pi}$, and the variance of it is $\frac{\pi-2}{\pi}nk$. When we got ν as the true signal, then the distribution of statistic is $N(nk\eta, nk)$. With some calculation, we need that $\eta > \max\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{k}}\}$. Introduce $k = p\epsilon$, $n = p^\theta$, $\epsilon = p^{-\vartheta}$, and $\eta = p^{-\beta}$, we find that we need

$$\beta < \begin{cases} \frac{\theta}{2}, & \frac{1-\theta}{2} < \vartheta < 1-\theta; \\ \frac{1-\vartheta}{2}, & \vartheta > 1-\theta. \end{cases} \tag{6.2.1}$$

It meets the other part of lower bound we found in Lemma 6.2.1.

According to the analysis and calculation, we have the following lemma about the upper bound of impossible area.

Lemma 6.2.2 *Under ARW model, the error rate of recovering label ℓ goes to 0 when*

$$\beta < f^{clu}(\vartheta, theta).$$

6.3 Some Extensions

6.3.1 Signal Recovery Problem

Theorem 6.3.1 *In the setting above, the features with signal $\mu \neq 0$ could be discovered when*

$$\beta > \rho_\theta^{sig}(\vartheta),$$

and the sum of type I error and type II error would go to 1/2 when

$$\beta < \rho_\theta^{sig}(\vartheta),$$

where

$$\rho_\theta^{sig}(\vartheta) = \begin{cases} \theta/2, & \vartheta < 1-\theta \\ \frac{1+\theta-\vartheta}{4}, & \vartheta > 1-\theta \end{cases}$$

In Amini and Wainwright ([Amini 2008]), the recovery of μ is studied. The setting is similar but a little different. The number of nonzero signals is known as k, and $X = \frac{\sqrt{k}}{\eta}\sqrt{\beta}v\mu^T + \sqrt{\Gamma}G$, where G is noise and $v_i \sim N(0,1)$, and β is a constant relative with Γ. When $G = I$, β could be any positive number. In this setting, we have that

$$X^TX = \beta\frac{k}{\eta^2}v^Tv\mu\mu^T + GG^T + \frac{\sqrt{\beta k}}{\eta}G^Tv\mu^T + \frac{\sqrt{\beta k}}{\eta}\mu v^TG.$$

In this paper, it is stated that with Semi-definite programming, μ could be recovered when $k = O(\log p)$, we need $\frac{n}{k\log p} > \theta_1$ to have rank-one solution, and $\frac{n}{k\log p} > \theta_2$, such

that the solution would converge to the truth. When k is in higher order, they stated the correctness of theorem still holds in simulation, but they did not prove it. When $\frac{n}{k\log p} < \frac{1+\beta}{\beta^2}$, the probability of error of any method is at least $1/2$, which means that μ could not be recovered.

So, the phase diagram of μ recovery is studied in this paper. In our setting, we could approximately treat the number of signals as $p\epsilon_p$. In their setting, the signal strength should be in the order of $1/\sqrt{k} = 1/\sqrt{p\epsilon}$, and μ could not be recovered when $\vartheta < 1-\theta$. When $\mu > 1-\theta$, it could be recovered by diagonal thresholding method. In the range inbetween, it seems that SDP could recover the signal in μ in simulation. However, the optimality of SDP is only proved when $k = O(\log p)$.

Compare it with our result by exhaust search, when $\vartheta > 1-\theta$, both μ and labels ℓ could be recovered when $\eta \sim 1/\sqrt{k}$. The paper provides a feasible method to realize the signals. When the signal strength is weak, the paper does not discuss about it. According to the discussion of exhaust search in following section, if our target is to recover signals only, we need that the difference $diff > \sqrt{2nk^2\log p}$, where the difference is

$$diff = \begin{cases} \frac{1}{\sqrt{2\pi}} nk^{3/2}\eta^2, & k\eta^2 \to 0; \\ n(k\eta - \sqrt{k}\frac{2}{\pi}), & k\eta^2 \to \infty. \end{cases}$$

Calculate the inequality in both cases, and it turns out that the signal strength should be

$$\eta > \max\{\frac{1}{\sqrt{n}}, \frac{1}{(nk)^{1/4}}\},$$

which is an upper bound for the recovery of μ. For classical PCA, the upper bound is that $\eta \sim p^{\vartheta/2-1/4-\theta/4}$.

6.3.2 Detection Problem

Theorem 6.3.2 *In the setting above, the hypothesis that $\mu \neq 0$ could be detected when*

$$\beta > \rho_\theta^{det}(\vartheta),$$

and the sum of type I error and type II error would go to 1/2 when

$$\beta < \rho_\theta^{det}(\vartheta),$$

where

$$\rho_\theta^{Det}(\vartheta) = \begin{cases} (2+\theta-4\vartheta)/4, & \vartheta < \frac{2-\theta}{4} \\ \theta/2, & \frac{2-\theta}{4} < \vartheta < 1-\theta \\ (1+\theta-\vartheta)/4, & \vartheta > 1-\theta \end{cases}$$

As the rank of matrix $y\mu'$ is only 1, it could be viewed as a low-rank matrix recovery problem, with noise terms. In Candes and Plan ([Candes 2011]), the recovery of low

rank matrix M when $y = A(M) + z$ is studied, where A is a linear transformation. Here, in our problem, we could use the setting that $M = \ell\mu'$, $A(\cdot)$ is identity mapping, and z is i.i.d Gaussian noise. So, in their program, they try to find a solution $\hat{M}$, to ensure that

$$\begin{array}{ll} \text{minimize} & \|M\|_1 \\ \text{subject to} & |\sigma_{\max}(y - M)| \leq \lambda, \end{array}$$

where $\|\cdot\|_1$ is the nuclear norm of a matrix, and $|\sigma_{\max}(y - M)|$ is the function norm of a matrix.

In their paper, they proved that in our case, λ should be set as $\lambda = Cp$ to make sure that the true matrix could be one solution. Under this condition, the solution to the problem, $\hat{M}$, satisfies that

$$\|\hat{M} - M\|_F^2 \leq C_0 \min(\sigma^2(M), n\sigma^2),$$

with probability at least $1 - 2e^{-cn}$ for constants C_0. What's more, they have proved that the minimax error over rank 1 matrices is lower bounded by p, which is

$$\sum_{M:rank(M)\leq 1} E\|\hat{M}(y) - M\|_F^2 \geq p.$$

So, the Frobenius norm between recovered $\hat{M}$ and M will be at the order of $\sqrt{p}$.

In our case, as the rank of M is 1, the Frobenius norm is easy to be found as $\|\ell\|\|\mu\| \sim \eta\sqrt{np\epsilon}$. If the Frobenius norm of difference between matrix is very small compared to the truth, the matrix should be recovered. It means that $\eta^2 np\epsilon >> p$, and it turns out to require the signal strength as

$$\eta > p^{\frac{\vartheta-\theta}{2}}.$$

When $\vartheta > \theta$, it means that the signal should go to infinity, which is too strong. When $\vartheta < \theta$, it turns out to be $\frac{1}{\sqrt{n\epsilon}}$. Compared to $\frac{1}{\sqrt{n}}$ when $\frac{1-\theta}{2} < \vartheta < 1 - \theta$, it needs the signal strength to be much stronger to recover the whole matrix.

6.4 Proofs

6.4.1 Proof of Lemma 6.2.2

Proof: Here, we discuss the upper bound in two cases: (a) When signal is dense, we apply classical method; and (b) when signal is sparse, we apply exhaust search. Now, we discuss them one by one.

In case (a), the signal is dense. Then the information part is the main part, and we use the classical method. To apply classical method, we assign the label as $\hat{\ell}_i = \text{sgn}[\sum_{j=1}^p X_{ij}]$. The distribution of statistic $\sum_{j=1}^p X_{ij}$ is

$$N(\ell_i \eta \#signals, p).$$

So, the error rate is $\Phi(-\epsilon \#signals/\sqrt{p})$. When p is large, the number of signals contrates at $p\epsilon$, so we can estimate the error rate as $\Phi(-\epsilon\eta\sqrt{p})$. To make sure that it goes to 0, we need $\epsilon\eta\sqrt{p} \to \infty$, which is equivalent with

$$\eta > \frac{1}{\sqrt{p}\epsilon}.$$

So, the classical method gives an upper bound as

$$\eta \sim p^{\vartheta - \frac{1}{2}}. \tag{6.4.2}$$

In case (b), the signal is sparse. In this case, we have to apply exhaust search method. Here, we assume that we know the number of signals, say it is $k = p\epsilon$. To do the search, search over all the possible combinations of features that have nonzero signals. Let ν be a $p \times 1$ vector, where $\|\nu\|_0 = k$. When $\nu'\mu = 0$, then $X\nu$ is noise, and there is no information about ℓ. When $\nu = \mu$, then the sign of $X\nu$ indicates the label vector ℓ. Let $\hat{\ell} = \mathrm{sgn}(X\nu)$, then $\hat{\ell}' X\nu$ will be very large when ν is truly discovered.

What is the bound given by exhaust search? When $\nu'\mu = 0$, the distribution of statistic is the summation of absolute value of n normal variables with mean 0 and variance k. When we got ν as the true signal, then the distribution of statistic is the summation of absolute value of n normal variables $N(k\eta, k)$. According to the properties of folded normal distribution, in the null case, the mean of statistic is $n\sqrt{2k/\pi}$, and the variance is $\frac{\pi-2}{\pi}nk$. In the alternative case, according to calculation, the statistic has mean as

$$n[\sqrt{2k/\pi}e^{-1/2k\eta^2} + k\mu(1 - 2\Phi(-\sqrt{k}\eta))].$$

So, the difference between the value from true set and the value from wrong set is

$$\begin{cases} \frac{1}{\sqrt{2\pi}}nk^{3/2}\eta^2, & k\eta^2 \to 0; \\ n(k\eta - \sqrt{k}\frac{2}{\pi}), & k\eta^2 \to \infty. \end{cases}$$

Now, there are two things we should consider. First, we have to recover the true signals to do clustering, so we need the true set. As there are about but less than p^k wrong sets, we hope the statistic from true set is larger than the maximum value from all the statistic from p^k sets. Accord to large deviation theory, the maximum value of p^k samples from folded normal distribution is at the order of $n(k\eta - \sqrt{k}\frac{2}{\pi}) + \sqrt{2nk\log(p^k)}$. So, we hope the difference between them is large. Second, even though we get the true signals, we need the data information to find the label. Note that when ν is the true signal vector, $X\nu \sim N(k\eta\ell, k)$. So, the clustering error rate would be $\Phi(-\sqrt{k}\eta)$. To make sure that the clustering rate goes to 0, we need $\sqrt{k}\eta \to 0$. Combine the two requirements, we need that

$$\eta > \max\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{k}}\} \approx \max\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{p\epsilon}}\}.$$

So, the upper bound given by exhaust search is

$$\eta \sim \begin{cases} \frac{1}{\sqrt{n}}, & \vartheta < 1-\theta, \\ \frac{1}{\sqrt{p^\epsilon}}, & \vartheta > 1-\theta. \end{cases} \tag{6.4.3}$$

Combine the two cases, from 6.4.2 and 6.4.3, we get the conclusion. □

6.4.2 Proof of Lemma 6.2.1

Proof: First, we show that when the classification problem cannot be solved, the clustering cannot be solved either. Let $X_{n\times p} = (x_1, x_2, \cdots, x_p)$, where x_i is an $n \times 1$ vector. For label vector ℓ, let $\tilde{\ell} = (\ell_2, \ell_3, \cdots, \ell_n)$, which is an $n-1\times 1$ vector. Without loss of generality, let $\ell_1 = 1$. Take $f(x,\tilde{\ell}) = \prod_{j=1}^p [1-\epsilon+\epsilon e^{\eta\tilde{\ell}' x_j + \eta X_{1j} - \frac{n\eta^2}{2}}]$, and correspondingly $g(x,\tilde{\ell}) = \prod_{j=1}^p [1-\epsilon+\epsilon e^{\eta\tilde{\ell}' x_j - \frac{(n-1)\eta^2}{2}}]$. The information contains in ℓ_1 can be denoted by the L^1 distance between $f(x,\tilde{\ell})$ and $g(x,\tilde{\ell})$, which is

$$E|\int f(x,\tilde{\ell}) - g(x,\tilde{\ell})d\tilde{\ell}| \le \int E|f(x,\ell) - g(x,\ell)|d\tilde{\ell}.$$

On the right side, it happened to be the L^1 distance for classification problem, where $\tilde{\ell}$ is known. According to the symmetry of ℓ, if we could prove that for some ℓ, the classification problem has $o(1)$ L^1 distance, then it holds for other ℓ, and therefore the left side also has $o(1)$ L^1 distance. So, our problem is to find the lower bound for classification problem, which has been solved partly in previous paper ([Jin 2009]).

According to the symmetry of ℓ, without loss of generality, we assume $\tilde{\ell} = (1,1,1,\cdots,1)$. Write $f(x)$ and $g(x)$ instead of $f(x,\tilde{\ell})$ and $g(x,\tilde{\ell})$ for clear notation. Now, let's discuss it in different cases: (a) signal is sparse, and (b) signal is dense.

Case 1: Data is sparse, where $\vartheta > 1-\theta$. When data is sparse, we apply Cauchy Inequality to L^1 distance to find a lower bound. Let

$$X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}.$$

Let $A(X_1,\mu) = e^{X_1'\mu - \|\mu\|^2/2} - 1$, and $B(X,\mu) = \prod_{i=2}^n \big[\cosh(X_i'\mu)e^{-\|\mu\|^2/2}\big]$, and then we have the L^1 distance as

$$E_0|\int A(X_1,\mu)B(X,\mu)dF(\mu)|.$$

According to Cauchy Inequality, the bound for L^1 distance is

$$(E_0|\int A(X_1,\mu)B(X,\mu)dF(\mu)|)^2 \le E_0[\int B(X,\mu)dF(\mu)]E_0[\int A(X_1,\mu)^2 B(X,\mu)dF(\mu)].$$

It is easy to found that $E_0[\int B(X,\mu)dF(\mu)] = 1$, and because of the independence between X_1 and X_i, $2 \le i \le n$, we have

$$E_0[\int A(X_1,\mu)^2 B(X,\mu)dF(\mu)] = \int\int A(X_1,\mu)^2 B(X,\mu)dF(X)dF(\mu) = \int A(X_1,\mu)^2 dF(X_1)dF(\mu),$$

Calculate $\int A(X_1,\mu)^2 dF(X_1)dF(\mu)$ directly, we could get

$$\int A(X_1,\mu)^2 dF(X_1)dF(\mu) = (1-\epsilon+\epsilon e^{\eta^2})^p - 1 \sim e^{p\epsilon\eta^2} - 1.$$

So, when $p\epsilon\eta^2 \to 0$, the L^1 distance is $o(1)$. Equivalently, when $\eta < \frac{1}{\sqrt{p\epsilon}}$, the clustering error rate goes to $1/2$ as $p \to \infty$.

Case 2: Data is dense, where $\vartheta < 1-\theta$. In this case, we have to introduce Hellinger distance to control L^1 distance. Define the Hellinger affinity by $H(f,g) = \int \sqrt{f(x)g(x)}dx$, then we have $\|f-g\|_1 \le 2\sqrt{1-H(f,g)}$. So, when Hellinger affinity goes to 1, the L^1 distance goes to 0, which means that it is impossible to separate the two hypothesis. The Hellinger affinity between $f(x)$ and $g(x)$ is

$$H(f,g) = E\sqrt{\prod_{j=1}^{p}[1-\epsilon+\epsilon e^{\eta\tilde{\ell}'x_j+\eta X_{1j}-\frac{n\eta^2}{2}}][1-\epsilon+\epsilon e^{\eta\tilde{\ell}'x_j-\frac{(n-1)\eta^2}{2}}]}.$$

As different coordinates are independent, we could study each coordinate first, and it is

$$H(f,g) = (E\sqrt{[1-\epsilon+\epsilon e^{\eta\tilde{\ell}'x+\eta X_1-\frac{n\eta^2}{2}}][1-\epsilon+\epsilon e^{\eta\tilde{\ell}'x-\frac{(n-1)\eta^2}{2}}]})^p,$$

where the expectation is over $x \sim N(0, I_{n-1})$ and $X_1 \sim N(0,1)$.

What we need to prove is that

$$m = E\sqrt{[1-\epsilon+\epsilon e^{\eta\tilde{\ell}'x+\eta X_1-\frac{n\eta^2}{2}}][1-\epsilon+\epsilon e^{\eta\tilde{\ell}'x-\frac{(n-1)\eta^2}{2}}]} = 1+o(1/p).$$

Define $h(x) = e^{\eta\tilde{\ell}'x-\frac{(n-1)\eta^2}{2}}$, $a(x) = e^{\eta X_1-\eta^2/2}-1$, and so we have that

$$m = E\sqrt{[1-\epsilon+\epsilon h(x)(a(x)+1)][1-\epsilon+\epsilon h(x)]} = E[1-\epsilon+\epsilon h(x)]\sqrt{1+\frac{\epsilon h(x)a(x)}{1-\epsilon+\epsilon h(x)}}.$$

As $h(x) = e^{\eta\ell'x-\frac{n-1}{2}\eta^2}$, if we regard the expectation is on random variable $x \sim (1-\epsilon)N(0,I_{n-1}) + \epsilon N(\eta, I_{n-1})$, then we could dismiss the term $1-\epsilon+\epsilon h(x)$, and it could be written as

$$m = E\sqrt{1+\frac{\epsilon h(x)a(x)}{1-\epsilon+\epsilon h(x)}}.$$

Since $|\sqrt{1+x}-1-x/2| \leq Cx^2$ for any $x > -1$, there is

$$|E\sqrt{1+\frac{\epsilon h(x)a(x)}{1-\epsilon+\epsilon h(x)}}-1-\frac{1}{2}Ea(x)| \leq CE(\frac{\epsilon h(x)a(x)}{1-\epsilon+\epsilon h(x)})^2.$$

According to calculation,

$$Ea(x) = Ee^{\eta X_1-\eta^2/2}-1 = 0, \qquad Ea^2(x) = E[e^{\eta X_1-\eta^2/2}-1]^2 = e^{\eta^2}-1 \leq \eta^2.$$

For the term $E\frac{\epsilon^2 h^2(x)}{(1-\epsilon+\epsilon h(x))^2}$, let's consider different cases where $x \sim N(0, I_{n-1})$ and $x \sim N(\eta, I_{n-1})$. They are quite similar, so I will discuss the first case in detail and extend it to the second case. Note that when $p \to \infty$, $1-\epsilon+\epsilon h(x) > 1/2$, so $\frac{\epsilon h(x)}{1-\epsilon+\epsilon h(x)} \leq 2\epsilon h(x)$. In case $2\epsilon h(x) > 1$, which means that approximately $\ell' x > \frac{\vartheta \log p}{\eta} + \frac{n-1}{2}\eta$, with probability $\Phi(-\frac{\vartheta \log p}{\sqrt{n-1}\eta} - \frac{\sqrt{n-1}}{2}\eta) \to 0$ anyway. So, $2\epsilon h(x)$ is a good upper bound. So the integration becomes that

$$E\frac{\epsilon^2 h^2(x)}{(1-\epsilon+\epsilon h(x))^2} \leq 4E\epsilon^2 h^2(x) = 4\epsilon^2 e^{(n-1)\eta^2}.$$

Similarly, apply it to the case $x \sim N(\eta, I_{n-1})$, we could get an upper bound $\epsilon^2 e^{3(n-1)\eta^2}$. In all, we have that

$$E\frac{\epsilon^2 h^2(x)}{(1-\epsilon+\epsilon h(x))^2} \leq 4(1-\epsilon)\epsilon^2 e^{(n-1)\eta^2} + 4\epsilon^3 e^{3(n-1)\eta^2},$$

and the whole term

$$m = E\sqrt{1+\frac{\epsilon h(x)a(x)}{1-\epsilon+\epsilon h(x)}} \leq C\eta^2\epsilon^2 e^{(n-1)\eta^2}.$$

When $\eta < \frac{1}{\sqrt{n}}$, $e^{(n-1)\eta^2} \to 1$, and we need that $\eta^2\epsilon^2 \sim o(1/p)$, which means that $\eta \sim o(\frac{1}{\sqrt{p}\epsilon})$. When $\eta > \frac{1}{\sqrt{n}}$, we could not do it in this way. In all, when $\eta < \min\{\frac{1}{\sqrt{n}}, 1/\sqrt{\sqrt{p}\epsilon}\}$, the Hellinger affinity goes to 1, which means that it is impossible to do clustering.

Bibliography

[Addario-Berry 2010] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, Gábor Lugosi*et al.* *On combinatorial testing problems.* The Annals of Statistics, vol. 38, no. 5, pages 3063–3092, 2010. (Cited on page 126.)

[Alizadeh 2000] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu*et al.* *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, vol. 403, no. 6769, pages 503–511, 2000. (Cited on page 12.)

[Alon 1999] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack and Arnold J Levine. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proceedings of the National Academy of Sciences, vol. 96, no. 12, pages 6745–6750, 1999. (Cited on pages 107 and 109.)

[Amini 2008] Arash A Amini and Martin J Wainwright. *High-dimensional analysis of semidefinite relaxations for sparse principal components.* In Information Theory, 2008. ISIT 2008. IEEE International Symposium on, pages 2454–2458. IEEE, 2008. (Cited on page 129.)

[Bhattacharjee 2001] Arindam Bhattacharjee, William G Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette*et al.* *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proceedings of the National Academy of Sciences, vol. 98, no. 24, pages 13790–13795, 2001. (Cited on page 28.)

[Bickel 2009] Peter J Bickel and Aiyou Chen. *A nonparametric view of network models and Newman–Girvan and other modularities.* Proceedings of the National Academy of Sciences, vol. 106, no. 50, pages 21068–21073, 2009. (Cited on page 92.)

[Borovkov 1965] AA Borovkov and BA Rogozin. *On the multi-dimensional central limit theorem.* Theory of Probability & Its Applications, vol. 10, no. 1, pages 55–62, 1965. (Cited on page 111.)

[Candès 2009] Emmanuel J. Candès and Yaniv Plan. *Near-ideal model selection by ℓ_1 minimization.* Ann. Statist., vol. 37, no. 5A, pages 2145–2177, 2009. (Cited on page 4.)

[Candes 2011] E. J. Candes and Y. Plan. *Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements*. IEEE Trans. Inf. Theor., vol. 57, no. 4, pages 2342–2359, April 2011. (Cited on page 130.)

[Dettling 2003] Marcel Dettling and Peter Bühlmann. *Boosting for tumor classification with gene expression data*. Bioinformatics, vol. 19, no. 9, pages 1061–1069, 2003. (Cited on page 28.)

[Donoho 2000] David L Donoho*et al.* *High-dimensional data analysis: The curses and blessings of dimensionality*. AMS Math Challenges Lecture, pages 1–32, 2000. (Cited on page 1.)

[Donoho 2004] D. Donoho and J. Jin. *Higher criticism for detecting sparse heterogeneous mixtures*. Ann. Statist., vol. 32, no. 3, pages 962–994, 2004. (Cited on pages 8, 12, 34 and 35.)

[Donoho 2006] David L Donoho. *Compressed sensing*. Information Theory, IEEE Transactions on, vol. 52, no. 4, pages 1289–1306, 2006. (Cited on page 5.)

[Donoho 2008] David Donoho and Jiashun Jin. *Higher criticism thresholding: Optimal feature selection when useful features are rare and weak*. Proc. Natl. Acad. Sci. U.S.A., vol. 105, no. 39, pages 14790–14795, 2008. (Cited on pages i, 8, 12, 34 and 35.)

[Dudoit 2002] Sandrine Dudoit, Jane Fridlyand and Terence P Speed. *Comparison of discrimination methods for the classification of tumors using gene expression data*. Journal of the American statistical association, vol. 97, no. 457, pages 77–87, 2002. (Cited on page 28.)

[Durbin 1985] James Durbin. *The first-passage density of a continuous Gaussian process to a general boundary*. Journal of Applied Probability, pages 99–122, 1985. (Cited on pages 93 and 98.)

[Efron 2004] Bradley Efron. *Large-scale simultaneous hypothesis testing*. Journal of the American Statistical Association, vol. 99, no. 465, 2004. (Cited on pages 5, 6, 27 and 108.)

[Golub 1999] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri and C. D. Bloomfield. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science, vol. 286, pages 531–537, 1999. (Cited on pages 92, 107 and 109.)

[Gordon 2002] Gavin J Gordon, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, Sridhar Ramaswamy, William G Richards, David J Sugarbaker and Raphael Bueno. *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma*. Cancer research, vol. 62, no. 17, pages 4963–4967, 2002. (Cited on pages 2 and 92.)

[Hastie 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman and R Tibshirani. The elements of statistical learning, volume 2. Springer, 2009. (Cited on page 3.)

[J. 2008] Fan. J. and Y. Fan. *High Dimensional Classification Using Features Annealed Independence Rules*. Ann. Statist., vol. 36, no. 6, pages 2605–2637, 2008. (Cited on page 9.)

[Ji 2010] Pengsheng Ji and Jiashun Jin. *UPS delivers optimal phase diagram in high dimensional variable selection*. Ann. Statist., vol. 40, no. 1, pages 73–103, 2010. (Cited on page 35.)

[Jin 2009] J. Jin. *Impossibility of successful classification when useful features are rare and weak*. Proc. Natl. Acad. Sci. USA, vol. 106, no. 22, pages 8859–8864, 2009. (Cited on pages 8, 12, 34, 35, 127 and 133.)

[Jin 2012a] J. Jin and W Wang. *Optimal spectral clustering by higher criticism thresholding*. Working Manuscript, 2012. (Cited on page 92.)

[Jin 2012b] Jiashun Jin. *Fast community detection by SCORE*. 2012. (Cited on pages 59, 71 and 78.)

[Johnstone 2009] Iain M Johnstone and Arthur Yu Lu. *On consistency and sparsity for principal components analysis in high dimensions*. Journal of the American Statistical Association, vol. 104, no. 486, 2009. (Cited on pages 2, 3 and 92.)

[Khan 2001] Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson*et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*. Nature medicine, vol. 7, no. 6, pages 673–679, 2001. (Cited on page 12.)

[Kolmogorov 1933] Andrey N Kolmogorov. *Sulla determinazione empirica di una legge di distribuzione*. Giornale dell'Istituto Italiano degli Attuari, vol. 4, no. 1, pages 83–91, 1933. (Cited on page 93.)

[Lee 2010] Ann B Lee, Diana Luca and Kathryn Roeder. *A spectral graph approach to discovering genetic ancestry.* The annals of applied statistics, vol. 4, no. 1, page 179, 2010. (Cited on page 5.)

[Lloyd 1982] Stuart Lloyd. *Least squares quantization in PCM.* Information Theory, IEEE Transactions on, vol. 28, no. 2, pages 129–137, 1982. (Cited on page 3.)

[Loader 1992] Clive R Loader*et al. Boundary crossing probabilities for locally Poisson processes.* The Annals of Applied Probability, vol. 2, no. 1, pages 199–228, 1992. (Cited on pages 14, 93, 98 and 110.)

[Lugosi 2004] Gábor Lugosi. *Concentration-of-measure inequalities.* 2004. (Cited on pages 73 and 74.)

[Paul 2007] Debashis Paul. *Asymptotics of sample eigenstructure for a large dimensional spiked covariance model.* Statistica Sinica, vol. 17, no. 4, page 1617, 2007. (Cited on page 34.)

[Pearson 1901] Karl Pearson. *Principal components analysis.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 6, no. 2, page 559, 1901. (Cited on page 4.)

[Pomeroy 2002] Scott L Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M Sturla, Michael Angelo, Margaret E McLaughlin, John YH Kim, Liliana C Goumnerova, Peter M Black, Ching Lau*et al. Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, vol. 415, no. 6870, pages 436–442, 2002. (Cited on page 12.)

[Rivasplata 2012] Omar Rivasplata. *Subgaussian random variables: An expository note.* 2012. (Cited on page 21.)

[Shorack 2009] Galen R Shorack and Jon A Wellner. Empirical processes with applications to statistics, volume 59. SIAM, 2009. (Cited on page 5.)

[Singh 2002] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie*et al. Gene expression correlates of clinical prostate cancer behavior.* Cancer cell, vol. 1, no. 2, pages 203–209, 2002. (Cited on page 30.)

[Su 2001] Andrew I Su, John B Welsh, Lisa M Sapinoso, Suzanne G Kern, Petre Dimitrov, Hilmar Lapp, Peter G Schultz, Steven M Powell, Christopher A Moskaluk, Henry F Frierson*et al. Molecular classification of human carcinomas by use of gene expression signatures.* Cancer research, vol. 61, no. 20, pages 7388–7393, 2001. (Cited on page 30.)

[Tibshirani 1996] Robert Tibshirani. *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996. (Cited on page 2.)

[Tulino 2004] AM Tulino and S Verdú. *Foundations and TrendsTM in Communications and Information Theory: Random Matrix Theory and Wireless Communications*, 2004. (Cited on page 15.)

[Vershynin 2010] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices.* arXiv preprint arXiv:1011.3027, 2010. (Cited on pages 15, 16, 17, 18, 23, 47 and 60.)

[Wang 2005] Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu*et al.* *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* The Lancet, vol. 365, no. 9460, pages 671–679, 2005. (Cited on page 28.)

[Woodroofe 1978] Michael Woodroofe. *Large deviations of likelihood ratio statistics with applications to sequential testing.* The Annals of Statistics, pages 72–84, 1978. (Cited on page 111.)

[Zhou 2010] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candès and Yi Ma. *Stable principal component pursuit.* In Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on, pages 1518–1522. IEEE, 2010. (Cited on page 2.)