CARNEGIE MELLON UNIVERSITY

POWER PREDICTION in LARGE SCALE MULTIPLE TESTING: A FOURIER APPROACH

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

for the degree

DOCTOR OF PHILOSOPHY
In
STATISTICS

by

AVRANIL SARKAR

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

August 28, 2010

Abstract

A problem that is frequently found in large-scale multiple testing is that, in the present stage of experiment (e.g. gene microarray, functional MRI), the signals are so faint that it is impossible to attain a desired level of testing power, and one has to enroll more samples in the follow-up experiment. Suppose we *are going to* enlarge the sample size by a times in the follow-up experiment, where $a > 1$ is not necessary an integer. A problem of great interest is, given data based on the *current* stage of experiment, how to predict the testing power after the sample size is enlarged by a times.

We consider test z-scores and model the test statistics in the current experiment as $X_j \sim N(\mu_j, 1)$, $1 \leq j \leq n$. We propose a Fourier approach to predicting the testing power after n replicates. The approach produces a very accurate prediction for moderately large values of a ($a \leq 7$). Finally, we discuss potential applications of this method on real data with emphasis on gene microarray data.

Acknowledgements

I express my sincere gratitude to my advisor Jiashun Jin. His guidance and experience helped in my academic and intellectual development. I also acknowledge Peter Huggins, Alessandro Rinaldo, Isabella Verdinelli and Larry Wasserman for being in my committee. I would specially like to thank Larry and Isa for their support and for taking their time to go through this document and give valuable suggestions to improve the document. I also thank Steve Fienberg for his valuable advice which enriched my experience in research. I thank the fellow students of the department, especially Gaia Bellone, Daniel Manrique, Zhanwu Liu and Han Liu for making the department environment so enjoyable.

Finally I would like to thank my parents and my little brother for their love and support.

Contents

List of Figures

Chapter 1

Introduction

Until recently, "simultaneous inference" meant considering just a handful of hypothesis tests at the same time. Rapid progress in technology, particularly in genomics and imaging, has vastly upped the ante for simultaneous inference problems giving rise to large scale multiple testing. Now 500 or 5,000 or even 50,000 tests may need to be evaluated simultaneously, raising new problems for the statistician, but also opening new analytic opportunities.

Examples of testing problems in biomedical and genomic research include the following:

- The identification of differentially expressed genes in high-throughput gene expression experiments such as microarray experiments, i.e., genes whose expression measures are associated with possibly censored biological and clinical covariates and outcomes

- The identification of co-expressed genes in high-throughput gene expres- sion experiments, i.e., pairs or sets of genes with correlated expression measures across

biological samples

- Tests of association between gene expression measures and biological an- notation metadata, e.g., Gene Ontology annotation
- Tests of association between phenotypes and codon/amino acid mutations, e.g., association between viral replication capacity and HIV-1 sequence variation

Simultaneous hypothesis testing begins with a collection of null hypotheses,

$$H_{01}, H_{02}, \ldots H_{0n}$$

corresponding test statistics, possibly not independent,

$$X_1, X_2, \ldots, X_n$$

and their corresponding p-values,

$$P_1, P_2, \ldots, P_n$$

with i^{th} p-value, P_i, measuring how strongly x_i, the observed value of X_i, contradicts H_{0i}; "Large-scale" means that n is a large number.

One of the problems frequently faced in large scale multiple testing is that the signals contained in the test statistics are faint. If the signals are faint, then distinguishing the null hypothesis from the alternative is a challenging problem. However, if the more replicates of the sample are collected, then the signal strength increases. As a result, the average power of the testing procedure improves. The main focus of this thesis is to answer the following question: *how many replicates are required to attain a pre-specified level of power*?

We study this problem in the context of the normal means problem

$$X_j = \mu_j + \epsilon_j \quad \epsilon_j \overset{iid}{\sim} N(0,1) \quad j = 1, \ldots, n \tag{1.0.1}$$

where the j^{th} case is a signal if $\mu_j \neq 0$.

1.1 Motivation

As a motivating example, we consider the example of differentially expressed genes in details. In this type of data, the main focus is to identify which genes are responsible for a particular disease. First two groups of people are chosen, one group comprising of the patients and the other one, controls. Then for a large number of genes, the difference in average gene expression of both the groups are recorded. This is equivalent to testing simultaneously a large number of null hypotheses, one for each gene. The null hypothesis, for a particular gene, corresponds to that gene being not differentially expressed. See for example, Brown and Botstein (1999); Lander (1999). Given a set of hypotheses to be tested and a set of test statistics, one for each hypothesis, a particular test statistic is said to contain a signal if the corresponding null hypothesis is false. The test statistic in this case is the normalized difference in average gene expression.

Now, the challenge is how to detect the genuine signals from noisy data. It has been pointed out in Pan et al. (2002) that it may be necessary to design an experiment that uses multiple arrays containing multiple measurement for each gene. One reason is that because of a high noise-to-signal ratio, a single array may not provide enough information that can be reliably extracted Lee et al. (2000). An important and natural question often asked by biologists is how many replicates are required ?

1.2 Main problem in details

The question asked by biologists in the above set-up can be generalized in the following way. In a lot of practical applications, the signal-to-noise ratio is found to be high. This makes reliably testing hypotheses difficult. At least four factors determine the power of simultaneous multiple testing:

- the proportion of no-null hypotheses.
- the distribution of the signals.
- measurement variability
- sample size.

Only the latter is under the experimenter's control. Moreover, if the signals are too weak then an increase in sample size will also lead to stronger signals. Given a set of hypotheses and a test-statistic corresponding to each hypothesis, *how many more copies of test-statistic (samples) for each hypothesis one needs for reliable testing.* ? The larger the number of copies, which we call replication multiplicity, the smaller the signal-to-noise ratio.

However, collecting more samples is a costly procedure. One would like to know beforehand the minimal number of samples required to achieve a certain degree of reliability for testing, from the available sample. This gives rise a lot of interesting estimation problems for the enlarged sample from the current sample. Several quantities are of interest. A few examples are given below.

- *Average power*: When a null hypothesis is rejected, it is a positive. A positive may be a true positive (TP) or a false positive (FP), depending on whether the hypothesis is correctly or incorrectly rejected. The average power of a procedure is the fraction of true positives that it yields. Larger sample size leads to decrease in signal-to-noise to ratio. Decrease in signal-to-noise ratio implies increase in average power for testing. So it is important to consider the prediction of the expected average power which can be obtained from an enlarged sample.

- *False Discovery Rate (FDR)*: The false discovery rate (FDR) of a test is defined as the expected proportion of false positives among the declared significant results [Benjamini and Hochberg (1995), Benjamini and Hochberg (2000), Keselman et al. (2002)]. Because of this directly useful interpretation, FDR is a more convenient scale to work on instead of the P-value scale. For example, if we declare a collection of 100 genes with a maximum FDR of 0.10 to be differentially expressed, then we expect a maximum of 10 genes to be false positives. No such interpretation is available from the $p-$value. Control of the FDR has been widely accepted as a criterion in multiple testing. The FDR level serves as an important guideline for practitioner. Increase in sample size means discovery of more signals and hence, decrease in FDR. Prediction of FDR for the enlarged sample is also of great interest.

- *Required replication multiplicity*: A larger sample usually means a larger power and a better control of the FDR. It is of interest to know the minimum sample size or replication multiplicity that is required to achieve a pre-specified level of average power or FDR , or both.

All the quantities of interest described above can be computed using the two key quantities ϵ, the proportion of non-null hypotheses, and PR, the expected positive rate

for the enlarged sample. Positive rate is the expected proportion of positives. More details about positive rate are in §(3.4). From hereon, we will refer to the positive rate for the enlarged sample as the future positive rate. Future expected positive rate refers to the expected fraction of null hypotheses rejected by the test statistic computed on an enlarged sample, had it been available.

We emphasize the role of the proportion of non-null hypotheses, ϵ and the future positive rate, PR in the light of the following equations.

$$\epsilon = \frac{1}{n} \cdot \#\{j : H_j \text{ is false}\}.$$

The (future) expected positive rate is related to the true positive rate (TPR) and false positive rate (FPR) through

$$PR = (1 - \epsilon) \cdot FPR + \epsilon \cdot TPR,$$

and that the FDR is related to these quantities through

$$FDR = \frac{(1 - \epsilon) \cdot FPR}{PR} = \frac{(1 - \epsilon) \cdot FPR}{(1 - \epsilon) \cdot FPR + \epsilon \cdot TPR.}$$

Therefore, the problems of estimating the proportion and the positive rate are of great interest. (Note that the positive rate associated with current data is relatively easy to estimate, but that associated with the future data is much harder to estimate).

The problem of estimating ϵ, the proportion of non-null hypotheses, has been studied before in great detail. For example, see Genovese and Wasserman (2004), Jin (2008),Jin and Cai (2007), Meinshausen and Rice (2006). The main focus of this thesis is to propose

an efficient estimation procedure for the future expected positive rate, PR and study the properties of the estimator.

1.3 Summary of results

The main technique used in estimating the expected positive rate is as follows. Expected positive rate, as a function of the unknown vector μ, in (1.0.1) can be expressed as a simple average of identical functions over the coordinates of μ. The Fourier transform of PR, depends on the unknown vector μ only through the average characteristic function of the coordinates of μ, which can be estimated using standard deconvolution technique.

Our main results are the following:

1. For a broad class of models $\mathcal{G}_1$, with no restriction on the proportion of non-null hypotheses, ϵ, we give rate of convergence of the mean-squared error of the estimator of PR (Theorem (4.1.1)).

2. An asymptotic lower bound for the minimax risk of estimating PR in the class $\mathcal{G}_1$ is established (Theorem (4.2.1)). This shows that the MSE of our proposed estimator is optimal upto a logarithmic factor.

3. We also consider class of models $\mathcal{G}_2$, where the proportion of non-null hypotheses, $\epsilon \to 0$ as the total number of hypotheses $n \to \infty$. We propose some subtle changes in the estimation procedure in this case, as compared to $\mathcal{G}_1$ in (4.3). The rate of MSE of our estimator for PR in this case is established (Theorem 4.3.1).

This thesis is organized as follows. In §2, we review the recent methods in large scale multiple testing which are important for applications such as false discovery rate

control, etc. We also review some estimation problems related to multiple testing. In §3, we describe the set-up under which we consider testing a large number of hypotheses. In particular, we describe the distribution of the test-statistics and the structure of the signals. We consider two important aspects with regard to the structure of the signals. One is their distribution, and the other is sparsity of the signals. We also present in §3 a Fourier approach for constructing estimators for general functionals.In §4 we define the positive rate (PR) which is the main quantity of interest, in terms of the parameters under the set-up we consider in §3. We also propose an estimator for estimating PR using the Fourier approach and study the asymptotic properties of this estimator under different regimes. In particular, we obtain theoretical rates of convergence for the mean squared error for our proposed estimator. In §5.1 we conduct some simulations to demonstrate the efficiency of the proposed estimator. In §5.2, we apply our method to some real datasets in gene microarray such as, the colon data Alon et al. (1999) and the leukemia data Golub et al. (1999). In §6 we make some concluding remarks and mention some future directions of research in this area.

Chapter 2

Review

In this chapter we review some of the inference problems related to the model,

$$X_j = \mu_j + \epsilon_j \quad \epsilon_j \overset{iid}{\sim} N(0,1) \quad j = 1, \ldots, n \tag{2.0.1}$$

which is more popularly known as the many normal means problem in the statistical literature. The j^{th} case is a signal if $\mu_j \neq 0$, otherwise it is noise. One of the foremost works on normal means was usage of shrinkage estimators by James and Stein in 1961 James and Stein (1961). In this chapter we will discuss some of the testing and estimation procedures developed in recent years for the model in (2.0.1).

2.1 Testing

The testing problems we are going to discuss are of the following two types:

- Simultaneous testing of the set of hypotheses

$$H_{0j} : \mu_j = 0 \; vs. \; H_{1j} : \mu_j \neq 0 \quad 1 \leq j \leq n$$

- Overall testing or testing for checking if there is any signal at all for a given set of hypotheses. Letting $\epsilon = \frac{1}{n}\#\{j : \mu_j \neq 0\}$, the proportion of signals, this is equivalent to testing if $\epsilon = 0$ or not.

2.1.1 Simultaneous testing

Simultaneous testing of a large number of hypotheses is very frequent these days in a variety of problems such as detection of differentially expressed genes, imaging etc. Control of type I error in this kind of situation is not very effective. Controlling type I error amounts to controlling the probability of at least one false positive. Generally this procedure is carried out in practice using the Bonferroni correction. Assuming that there are n hypotheses to be tested and we want control the familywise error rate at level α, the Bonferroni method tests each individual hypothesis at level $\frac{\alpha}{n}$. But it is known that this type of correction is very conservative and its power is too low. This was discussed in Dudoit et al. (2003) who showed, by comparing several testing procedures, such as Bonferreni method, false discovery rate and per-comparison error rate, that the power of Bonferroni method is much lower compared to others.

The most significant development to overcome this kind of problem, was proposed by Benjamini and Hochberg (1995). They argued that the idea of controlling familywise error rate is not necessary because when there is a large number of hypotheses to be tested, then controlling the probability of at least one false rejection will lead to very

low power. Instead, they proposed to control the expected ratio of the number of falsely rejected hypotheses to the total number of rejected hypotheses, which they called the False Discovery Rate (FDR). More formally, let V be the number of falsely rejected null hypotheses and S the number of correctly rejected null hypotheses. Let Q be defined as

$$Q = \begin{cases} \frac{V}{V+S} & \text{if } V+S \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

FDR is defined as $E[Q]$.

A simple procedure for controlling FDR at level q is the following where q is chosen by the user. Let there be n null hypotheses to be tested, $H_{01}, \ldots, H_{0n}$ and let $p_1, \ldots, p_n$ be the corresponding $p-$values. Let the $p-$values be arranged in increasing order of magnitude by $p_{(1)} \leq \ldots \leq p_{(n)}$. Also let the null hypothesis in the increasing order of $p-$values be $H_{(01)}, \ldots, H_{(0n)}$. Let

$$\widehat{k}_{FDR} = \max\{k : p_{(k)} \leq q\frac{k}{n}\}$$

and reject the null hypotheses $H_{(01)}, \ldots, H_{(0\widehat{k}_{FDR})}$. Benjamini and Hochberg (1995) showed that for the above testing procedure, $E[FDR] \leq q$. In fact in the case of n independent hypotheses with n_0 true null hypotheses, they showed that

$$E[FDR] = q\frac{n_0}{n} \leq q$$

The main idea of FDR is that if there is a large number of hypotheses to be tested simultaneously, instead of controlling the probability of incorrectly rejecting at least one null hypothesis, we let some null hypotheses to be incorrectly rejected only controlling for FDR, then we get a much more powerful testing procedure.

2.1.2 Overall testing

Overall testing is carried out when we are interested in testing if there is any signal at all. In particular we consider the following. We assume there are n hypotheses to be tested with X_j being the test statistic for the j^{th} hypothesis. We test,

$$H_{0j} : X_j \sim N(0,1) \; vs. \; X_j \sim N(\mu_j, 1) \quad \mu_j > 0$$

In fact, Donoho and Jin (2004) considers a slightly simpler model for this problem.

$$H_0 : X_j \overset{i.i.d}{\sim} N(0,1) \quad 1 \leq j \leq n$$
$$H_1 : X_j \overset{i.i.d}{\sim} (1-\epsilon)N(0,1) + \epsilon N(\mu, 1) \quad 1 \leq j \leq n$$

where ϵ is the proportion of non-null hypotheses. In Donoho and Jin (2004), ϵ and μ were calibrated in the following way. They chose $\epsilon = n^{-\beta}$ with $\beta \in (\frac{1}{2}, 1)$ and $\mu = \sqrt{2r \log n}$ with $0 < r < 1$. In this setting, with μ and ϵ known, the optimal test is the likelihood ratio test. It was shown in Ingster (1998), that there is a detection boundary described by a function $\rho^*(\beta)$ such that if $r > \rho^*(\beta)$ then the likelihood ratio test can successfully determine whether the null hypothesis is true or not. Letting $p_{(1)} \leq \ldots \leq p_{(n)}$ be the $p-$values arranged in increasing order. Donoho and Jin (2004) proposed the higher criticism statistic,

$$HC^* = \max_{0 \leq j \alpha_0 n} \sqrt{n}[j/n - p_{(j)}] / \sqrt{p_{(j)}(1 - p_{(j)})}$$

The higher criticism test statistic rejects the null hypothesis for large values of HC^*. Donoho and Jin (2004) showed that it can detect whether the null hypothesis is true or

not, in the same region of the detection region as the likelihood ratio test, adapting to the unknown values of β and r.

2.2 Estimation

In this section, we will review the following three problems for the model in (2.0.1):

- Estimation of the vector μ under the assumption that the proportion of non-zero means, $\epsilon \to 0$ as the number of observations $n \to \infty$.
- Estimation of the null distribution under a slightly general set up

$$X_j \sim N(\mu_j, \sigma_j^2) \quad X_j \perp X_{j'},\ j \neq j'$$

$$H_{0j} : (\mu_j, \sigma_j) = (\mu_0, \sigma_0) \quad vs.\ H_{1j} : (\mu_j, \sigma_j) \neq (\mu_0, \sigma_0)$$

  where μ_0 and σ_0 are unknown.

2.2.1 Estimation of μ

In Abramovich et al. (2006), an estimator for μ was proposed for the model (2.0.1) based on hard thresholding. It was argued that in the sparse case, i.e. when the proportion of non-zero components of μ, ϵ, is small then estimating μ by hard thresholding would be a sensible strategy. Formally, for any threshold t, hard thresholding at t gives the j^{th} component of the estimator $\widehat{\mu}$ as proposed as

$$\widehat{\mu}_j = \begin{cases} x_j & \text{if } |x_j| > t \\ 0 & \text{otherwise} \end{cases} \tag{2.2.2}$$

A review of Donoho et al. (1992) and Donoho and Johnstone (1994a) on minimax estimation on μ over the space

$$\ell_0(\epsilon) = \{\mu : \frac{1}{n}\sum_{j=1}^{n} |\mu_j| \leq \epsilon\}$$

shows that the ideal threshold t in (2.2.2) is

$$t_\beta = \sqrt{2(1-\beta)\log n} \text{ where } \epsilon = n^{\beta-1} \quad \beta \in (0,1)$$

The smaller the value of β, the smaller the value of ϵ. A small value of ϵ means the data is very sparse. However the parameter β which is related to sparsity, is unknown in practice. Abramovich et al. (2006) showed the following estimator adapts to this unknown sparsity:

- Take the order statistics $|x|_{(1)} \geq \ldots \geq |x|_{(n)}$.
- Compare them to the series of right tail Gaussian quantiles $t_k = z(q/2 \cdot k/n)$ where q is the chosen by the FDR controlling procedure.
- Choose k_{FDR} to be the largest index k for which $|x|_{(k)} \geq t_k$.

Then the estimator $\widehat{\mu}$ of μ obtained by thresholding at $\widehat{t}_{k_{FDR}} = \widehat{t}_F$, is

$$\widehat{\mu}_{F,j} = \begin{cases} x_j & \text{if } |x_j| > \widehat{t}_F \\ 0 & \text{otherwise} \end{cases}$$

The threshold adapts to the unknown sparsity and attains the minimax rate over $\ell_0(\epsilon)$ provided the false discovery rate $q = q_n \to 0$ as $n \to \infty$.

2.2.2 Estimation of null distribution

Cai and Jin (2010) considered a more general version of the model in (2.0.1)

$$X_j \overset{i.i.d.}{\sim} (1-\epsilon)\phi(\frac{x-\mu_0}{\sigma_0}) + \epsilon \int \phi(\frac{x-u}{\sigma}) g(u,\sigma)\, du, \quad 1 \le j \le n \tag{2.2.3}$$

with the constraints that

$$|\xi|^\alpha \widehat{g}(\xi|\sigma) \le A \quad \text{and } \epsilon \le \epsilon_0 n^{-\beta} \tag{2.2.4}$$

with $\alpha > 0$ and $\beta \in [0, \frac{1}{2})$ and $\epsilon_0 \in (0,1)$.

The main problem was to estimate μ_0 and σ_0. This problem is of practical importance as Efron et al. (2001) pointed out in the case of microarray data for breast cancer, that the true null distribution is not $N(0,1)$ i.e. $(\mu_0, \sigma_0) \neq (0,1)$ in (2.2.3). Efron et al. (2001) cited a number of reasons for this phenomenon such as unobserved covariates, correlations across arrays etc.

Cai and Jin (2010) showed at high frequencies, the characteristic function of the X_j's can be used to estimate the parameters μ_0 and σ_0. More formally, they showed that if $r(\xi) = E[e^{i\xi X_j}]$ is the characteristic function of X_j, $1 \le j \le n$, then

$$r(\xi) \approx (1-\epsilon)e^{-\sigma_0^2\xi^2/2} \cdot e^{i\mu_0\xi}$$

Now, we can retrieve σ_0 and μ_0 from the characteristic function r above in the following way:

$$\widehat{\sigma}_0^2(\gamma) = -\left(\frac{\frac{d}{ds}|r(s)|}{s|r(s)|}\right)\Bigg|_{s=\xi_n(\gamma)}, \quad \widehat{\mu}_0(\gamma) = \left(\frac{1}{r^2(s)}\text{Im}(\bar{r}(s)r^{'}(s))\right)\Bigg|_{s=\xi_n(\gamma)} \tag{2.2.5}$$

where $\xi_n(\gamma) = \inf\{\xi : \xi > 0, \quad |r(\xi)| \leq n^{-\gamma}\}$. Finally since, $r(\xi) = E[e^{i\xi X_j}]$ is unknown, they substituted it with the empirical characteristic function $\widehat{r}_n(\xi) = \frac{1}{n}\sum_{j=1}^{n} e^{i\xi X_j}$ and followed the same procedure as above. Cai and Jin (2010) proved that the above procedure gives minimax estimator both for μ_0 and σ_0^2 in the class of models described by (2.2.3) and (2.2.4).

Chapter 3

Methodology

The main focus of this thesis is to predict power of testing procedure in large scale multiple testing, when the sample size is increased. In this section we will introduce the model of the test-statistics, describe what we mean by enlarged sample and also justify our model assumptions with the help of a simple example. Next, we define and explain our main quantity of interest, the positive rate (PR), and also describe the importance of estimating it. Finally, we introduce a general approach using tools from Fourier expansion for estimating functionals of the type of positive rate.

3.1 Gaussian Model

Let there be n independent hypotheses to be tested, the null hypothesis

$$H_{0j} : \mu_j = 0 \ \ vs. \ \ H_{1j} : \mu_j \neq 0 \quad j = 1, \ldots, n. \tag{3.1.1}$$

Let X_j be the test statistic for the j^{th} hypothesis. We assume that X_j's are independent and

$$X_j \sim N(\mu_j, 1), \quad 1 \leq j \leq n \tag{3.1.2}$$

The test-statistics X_j, $1 \leq j \leq n$ come from the sample available to us which we will refer to as the "current sample". The assumption about the model of the test-statistics X_j, $1 \leq j \leq n$ will be justified in §(3.3). Since, our main focus is to study the increase in power for testing in case of (3.1.1) as a result of increase in sample size, we first describe in details what we mean by an enlarged sample. We refer to the enlarged sample by "future sample", since it is not available to us while we are doing inference.

3.2 Future Sample

Next, we consider the future sample. By future sample we mean, in the future more observations will be collected. Let us assume the sample size in future will be a times the current sample size, with $a > 1$. In order to clarify what we mean by future sample, we assume a is an integer although it is not necessary, as we will see later in §(3.3) . Corresponding to each observation X_j in the current sample as in (3.3.2), we assume we have a independent and identically distributed copies of X_j in the future sample for $1 \leq j \leq n$. The future sample can be described as the following:

Current Sample: X_j

Future Sample: $(X^f_{1j}, \ldots, X^f_{aj})$

$X^f_{jk} \overset{i.i.d.}{\sim} N(\mu_j, 1), \quad 1 \leq k \leq a, \qquad 1 \leq j \leq n$

where μ_j is same as in (3.1.2) for $1 \leq j \leq n$. Now we can construct the test-statistic for the future sample by taking a simple average across replications and normalizing by the

standard deviation for each observation i.e.

$$X_j^f = \frac{1}{\sqrt{a}} \sum_{k=1}^{a} X_{kj}^f \quad 1 \leq j \leq n$$

Denoting the test-statistic for the j^{th} case by X_j^f for the future sample, we have

$$X_j^f \sim N(\sqrt{a}\mu_j, 1), \quad 1 \leq j \leq n$$

In the next section we justify the assumptions about the model of the test-statistics X_j, $1 \leq j \leq n$ from the current sample in (3.1.2) as well as the distribution of the test statistics X_j^f, $1 \leq j \leq n$, in future sample.

3.3 Justification for model assumptions

We justify the model described above for the test statistics X_j, $1 \leq j \leq n$ in the context of hypothesis testing in (3.1.1) with a simple example. We consider the problem of detection of differentially expressed genes, described as one of the main motivations in the introduction. Suppose, gene expression measurements are collected on n genes from two groups of people, control and patients. One group has p_1 people and the other one has p_2 people. The problem of interest, is to detect which genes between these two groups we differentially expressed. Hence, in this case the total number of hypotheses to be tested is equal to n. Each gene yielded a two-sample $t-$statistic. For the j^{th} gene, the $t-$statistic is

$$T_j = \frac{\overline{G}_{1j} - \overline{G}_{2j}}{\sqrt{Var(\overline{G}_{1j}) + Var(\overline{G}_{2j})}} \tag{3.3.1}$$

where $\overline{G}_{1j}$ and $\overline{G}_{2j}$ are the mean expression levels for the j^{th} gene in the first group and the second group respectively. μ_j is the expected value of T_j for the j^{th} gene for $j = 1, \ldots, n$. If the degrees of freedom of T_j is moderately large for all j, it can be assumed that

$$T_j \overset{approx}{\sim} N(\mu_j, 1) \quad 1 \leq j \leq n$$

Hence, we assume the test statistic, X_j comes from $N(\mu_j, 1)$ for $1 \leq j \leq n$.

Additionally denoting the proportion of non-null hypotheses in (3.1.1) by ϵ, and assuming the non-zero coordinates of the vector μ come from a density g, the whole set up and the model can be described in the following way. Let $\mathcal{F}$ be the subset of $\{1, \ldots, n\}$ with $|\mathcal{F}| = n\epsilon$ such that if $j \in \mathcal{F}$, then H_{0j} is false for $j = 1, \ldots, n$. This simply means out of the n hypotheses to be tested, $\mathcal{F}$ is the set of coordinates for which the null hypothesis is false. Also assume that if H_{0j} is false, then $\mu_j \sim g$ for some density g for $1 \leq j \leq n$. Here, ϵ, $\mathcal{F}$ and g are unknown. Then,

$$\begin{aligned} X_j &\overset{iid}{\sim} N(0,1) \, \forall \, j \in \mathcal{F}^c \\ X_{j'} &\overset{iid}{\sim} \int \phi(x-u)\, g(u)\, du \, \forall \, j' \in \mathcal{F} \\ X_j &\text{ is independent of } X_{j'} \, \forall j \in \mathcal{F}^c,\, j' \in \mathcal{F} \end{aligned} \tag{3.3.2}$$

where $\phi(\cdot)$ is the density of $N(0, 1)$.

Now we focus on the distribution of the future test-statistics, X_j^f for $1 \leq j \leq n$. Going back to the example of differentially expressed genes, we consider again the two-sample $t - test$ as in 3.3.1. For the j^{th} hypothesis, let T_j^f be the future two-sample $t-$statistic,

for $1 \leq j \leq n$.

$$T_j^f = \frac{\overline{G}_{1j}^f - \overline{G}_{2j}^f}{\sqrt{Var(\overline{G}_{1j}^f) + Var(\overline{G}_{2j}^f)}} \tag{3.3.3}$$

where $\overline{G}_{1j}^f$ and $\overline{G}_{2j}^f$ are the mean expression levels for the j^{th} gene in the first group and the second group respectively for the future sample. Now,

$$E[\overline{G}_{1j} - \overline{G}_{2j}] = E[\overline{G}_{1j}^f - \overline{G}_{2j}^f]$$

However,

$$Var[\overline{G}_{1j}^f] + Var[\overline{G}_{2j}^f] = \Big(Var[\overline{G}_{1j}] + Var[\overline{G}_{2j}]\Big) \times \frac{1}{\sqrt{a}}$$

Hence,

$$T_j^f \overset{approx}{\sim} N(\sqrt{a}\mu_j, 1) \quad 1 \leq j \leq n$$

Letting Y_j be the test statistic for the future data for the j^{th} hypothesis, we have $Y_j \sim N(\sqrt{n}\mu_j, 1)$ for $j = 1, \ldots, n$. Since the coordinates out of $\{1, \ldots, p\}$ for which the null hypotheses are false, remain the same as in the case of the current sample, the marginal density of the future test statistics is

$$\begin{aligned} Y_j &\overset{iid}{\sim} N(0,1) \,\forall\, j \in \mathcal{F}^c \\ Y_{j'} &\overset{iid}{\sim} \int \phi(x - \sqrt{a}u)\, g(u)\, du \,\forall\, j' \in \mathcal{F}^c \\ Y_j &\text{ is independent of } Y_{j'} \,\forall j \in \mathcal{F}^c,\, j' \in \mathcal{F} \end{aligned} \tag{3.3.4}$$

3.4 Testing Procedure and Positive Rate

Having described the hypotheses and the distribution of the test statistics, we turn our attention to the testing procedure that we are going to consider. Note that the model we consider from §3.1 is,

$$X_j \overset{iid}{\sim} N(\mu_j, 1) \quad j = 1, \ldots, n \tag{3.4.1}$$

where only a small proportion ϵ of the vector μ are significantly large and the locations of these components are not known in advance. In such situations, an appropriate testing procedure should be based on hard thresholding. To be more specific, we choose an appropriate threshold t, and then decide that,

$$\begin{aligned} &H_{0j} \text{ is false if } |X_j| > t \\ &H_{0j} \text{ is true, otherwise} \end{aligned} \tag{3.4.2}$$

for $j = 1, \ldots, n$.

The most immediately compelling motivation for this strategy is provided by wavelet analysis, since the wavelet representation of many smooth and piecewise smooth signals is sparse in precisely our sense. For more on this, see Abramovich et al. (2006). The threshold can be chosen in various ways. One of these ways is to choose it based on controlling false discovery rate by Benjamini and Hochberg (1995). Our main focus, though is not on the choice of the threshold t. Our main focus is to estimate a quantity called the positive rate (PR), which we will introduce below, for a range of interesting thresholds t.

In (3.4.2), we discussed the testing strategy for the current (available) sample. Since our quantity of interest is related to the future (enlarged) sample, we discuss very shortly

the testing strategy for the future sample. Note that for the future sample, the future test-statistics,

$$Y_j \overset{iid}{\sim} N(\sqrt{a}\mu_j, 1) \quad j = 1, \ldots, n \tag{3.4.3}$$

If the future test-statistics were available to us, then similar to (3.4.2), the testing strategy should be,

$$\begin{aligned} &H_{0j} \text{ is false if } |Y_j| > t \\ &H_{0j} \text{ is true, otherwise} \end{aligned} \tag{3.4.4}$$

for $j = 1, \ldots, n$. Note that, we use t as a generic threshold here. Before going into further details about positive rate, we revisit some standard terminology for multiple testing.

Given a set of hypotheses and a testing procedure, we have the following:

- When a null hypothesis is rejected, we call it a positive.
- When a null hypothesis is rejected by the test procedure, but in reality it is true, we call it a false positive.
- When a null hypothesis is rejected by the test procedure, and also in reality it is false, we call it a true positive.

Comparing (3.4.1) and (3.4.3), it is evident that the signal strength for the future sample (i.e. when the sample size is increased by a times) the signal strength is increased by $\sqrt{a}$ times. An increase in signal strength implies that one should be able to discover more signals in the future sample. Our main goal, as described in the introduction in §1, is to investigate how much one can gain by increasing the sample size. One way to quantify this is, given the testing procedure described in (3.4.4), what should be its power

for the future sample. In other words, given a value of the replication multiplicity a and a threshold t, *how can we estimate power for the future sample given the current data* $\{X_j\}_{j=1}^n$. In the case of a collection of hypothesis, by power of a testing procedure we mean the expected proportion of true positives discovered by it.

From here on, we will refer to expected proportion of true positives as true positive rate (TPR), expected proportion of false positives as false positive rate (FPR) and expected proportion of positives as positive rate (PR). As described above, our main quantity of interest is TPR for the future sample (enlarged by a times). First we derive formulas for TPR, FPR and PR from which the relationship among them will be evident. Then we will explain why we want to estimate PR instead of TPR. From the testing procedure for the future sample as described in (3.4.4), given a threshold value t and replication multiplicity a, and also using (3.4.3),

$$\begin{aligned} TPR(t,a) &= \frac{E_{H_{1j}\text{ true}}[\sum_{j=1}^n I(|Y_j| > t)]}{\{\#j : H_{1j}\text{ true}\}} \\ &= \frac{\sum_{j=1}^n P_{H_{1j}\text{ true}}(|Y_j| > t)}{n\epsilon} \\ &= \frac{n\epsilon[1 - \int \Psi(u;t,n)g(u)\,du]}{n\epsilon} = 1 - \int \Psi(u;t,n)g(u)\,du \end{aligned} \tag{3.4.5}$$

where

$$\Psi(u;t,a) = 1 - [\bar{\Phi}(t - \sqrt{a}\cdot u) + \bar{\Phi}(t + \sqrt{a}\cdot u)], \tag{3.4.6}$$

with $\bar{\Phi} = 1 - \Phi$ being the survival function of $N(0,1)$

Similarly for the false positive rate and the positive rate we have,

$$\begin{aligned} FPR(t,a) &= \frac{E_{H_{0j}\text{ true}}[\sum_{j=1}^{n} I(|Y_j| > t)]}{\{\#j : H_{0j}\text{ true}\}} \\ &= \frac{\sum_{j=1}^{n} P_{H_{0j}\text{ true}}(|Y_j| > t)}{n(1-\epsilon)} \\ &= \frac{n(1-\epsilon)[1-\Psi(0;t,a)]}{n(1-\epsilon)} = 1-\Psi(0;t,a) \end{aligned} \tag{3.4.7}$$

$$\begin{aligned} PR(t,a) &= \frac{1}{n}\sum_{j=1}^{n} P(|Y_j| > t) \\ &= \frac{1}{n}E_{H_{0j}\text{ true}}[\sum_{j=1}^{n} I(|Y_j| > t)] + \frac{1}{n}E_{H_{1j}\text{ true}}[\sum_{j=1}^{n} I(|Y_j| > t)] \\ &= \frac{1}{n}\cdot n(1-\epsilon)[1-\Psi(0;t,a)] + \frac{1}{n}\cdot n\epsilon[1-\int \Psi(u;t,n)g(u)\,du] \\ &= 1-(1-\epsilon)\Psi(0;t,a) - \epsilon\int \Psi(u;t,a)g(u)\,du. \end{aligned} \tag{3.4.8}$$

where Ψ is as described in (3.4.6).

Combining (3.4.5), (3.4.7) and (3.4.8) we have

$$PR = (1-\epsilon)\cdot FPR + \epsilon\cdot TPR \tag{3.4.9}$$

Now, the positive rate PR, being a simple average over all the components of μ (as will be shown in §4) is much easier to estimate compared to TPR. This is because TPR is a function of only the signals i.e. the non-zero components of μ whose coordinates are unknown. However, from the relationship between TPR and PR in (3.4.9), the only unknown quantity involved is ϵ. Note that from (3.4.7) and (3.4.6), given a threshold value t, FPR is known. However the problem of estimating the proportion of non-null

hypothesis, ϵ is well-studied [Genovese and Wasserman (2004),Jin (2008),Jin and Cai (2007),Meinshausen and Rice (2006)]. We will discuss more about estimating ϵ in §4. Now, suppose we have an efficient estimator of the proportion of non-null hypothesis, ϵ. Also suppose we have an estimator $\widehat{PR}$ of the positive rate PR. Then from (3.4.9), we can have an estimator for TPR,

$$\widehat{TPR} = \frac{\widehat{PR} - (1 - \widehat{\epsilon}) \cdot FPR}{\widehat{\epsilon}}$$

Another reason why estimating the positive rate, PR, is important can be given from the perspective of false discovery rate (FDR). Suppose we choose a threshold t and apply the testing procedure in (3.4.2) to the current data. Let us denote the false discovery rate for the current data by FDR^c. Let, for the choice of the same threshold t, the testing procedure be applied to the future sample as in (3.4.4). Let us denote the false discovery rate for the future sample by FDR^f. Now, it can be derived easily, that

$$\frac{FDR^f}{FDR^c} = \frac{PR^c}{PR^f}$$

where PR^c and PR^f refers to the current and future positive rates respectively. Since the future sample is enlarged in size, more signals are expected to be discovered for the same threshold t. Moreover, the number of false signals are expected to be the same since increase in sample size only affects the true signals. Now the current positive rate, for a threshold t is very easy to estimate. It is nothing but the survival function,

$$\widehat{PR}^c = \frac{1}{n} \sum_{j=1}^{n} I(|X_j| > t)$$

Hence an estimator of $\widehat{PR}^f$, which we refer to above as just PR allows us to estimate the decrease in false discovery rate due to enlarging the sample size.

Hence, the central problem for us is then how to estimate $PR(t, a)$ for a given value of the threshold t and replication multiplicity a, from the sample available at the current stage i.e. X_j for $j = 1, \ldots, n$. In the next subsection a Fourier approach is proposed for estimating general functionals which are of the same form as the positive rate, PR.

3.5 A Fourier approach for estimation

From (3.4.8), it follows that the positive rate for a given value of threshold t and replication multiplicity a can be written as

$$1 - PR(t, a) = \frac{1}{n} \sum_{j=1}^{n} P(|Y_j| \leq t) = \frac{1}{n} \sum_{j=1}^{n} E[\Psi(\mu_j; t, a)] \tag{3.5.1}$$

In this section we focus on estimating general functionals of the form of the positive rate i.e. functionals which can be represented as simple average over all components of the vector μ. Suppose we want to estimate functionals of a much broader class of the form

$$T(h) = \frac{1}{n} \sum_{j=1}^{n} E[h(\mu_j)] \tag{3.5.2}$$

for some function h. Also note that the function $\Psi(u; t, a)$ in (3.4.6) has a nice decay as the value of $|u|$ gets large as is evident from Figure (3.5). Hence, the Fourier transform of $\Psi(\cdot; t, a)$ exists. We also assume, for the functional $T(h)$, h also has a Fourier transform. It should be noted, that from (3.4.8) and (3.4.6), it follows that PR is of the form of (3.5.2) with h replaced by Ψ. For simplicity of notations, we also assume h is a symmetric

function.

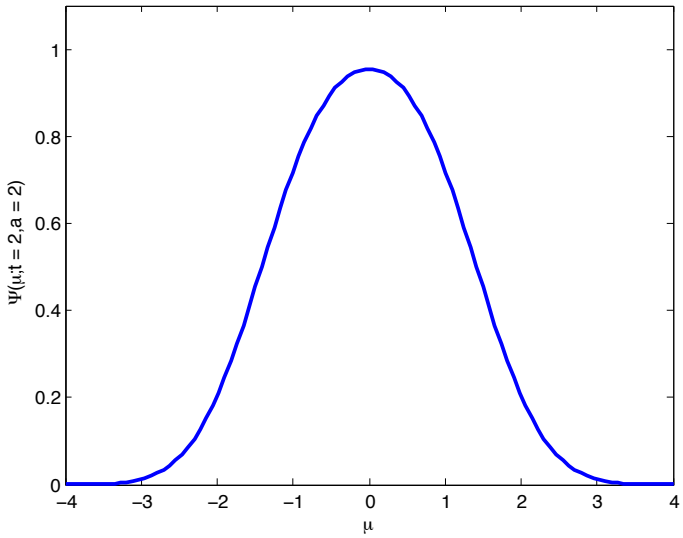

Figure 3.1: Display of $\Psi(u;t,a)$ with the threshold value $t = 2$, and replication multiplicity $a = 2$.

The idea of using Fourier transforms for estimating functionals of the form of T was proposed in Jin (2008) for estimating the proportion of non-null hypothesis, ϵ. Note that the proportion ϵ of non-null hypothesis can be represented as,

$$\epsilon = 1 - \frac{1}{n}\sum_{j=1}^{n} I(\mu_j = 0).$$

Although, ϵ is a discontinuous function at 0, it can be approximated by smooth function which has a Fourier transform. For more details see, Jin (2008). One way of looking at

estimating functionals T of the form (3.5.2) is the following. Note that,

$$\begin{aligned} T(h) &= \frac{1}{n}\sum_{j=1}^{n} E[h(\mu_j)] \\ &= \frac{1}{n}\sum_{j=1}^{n} E\Big[\int \widehat{h}(\xi)\cos(\xi\mu_j)\,d\xi\Big] \\ &= E\Big[\int \widehat{h}(\xi)\frac{1}{n}\sum_{j=1}^{n}\cos(\xi\mu_j)\,d\xi\Big] \\ &= \int \widehat{h}(\xi)[(1-\epsilon)+\epsilon\widehat{g}(\xi)]\,d\xi \end{aligned} \tag{3.5.3}$$

where for any function $r(\cdot)$, $\widehat{r}(\cdot)$ denotes its Fourier transform. We also made use of the fact that $(1-\epsilon)$ proportion of the $\mu_j's$ are 0 and the rest come from an unknown density g which follows from (3.1.1). The distribution of the elements of μ can also be described as

$$\mu_j \overset{iid}{\sim} (1-\epsilon)\delta_0 + \epsilon \cdot g \quad j = 1, \ldots, n$$

Let ϕ_μ be the characteristic function of the distribution of μ_j for any j, $j = 1, \ldots, n$. Also let, ϕ_X be the characteristic function of X_j, for any j, $j = 1, \ldots, n$. Also let $\widehat{\phi}$ be the characteristic function of standard normal distribution. Using deconvolution, it follows that

$$\phi_\mu(\xi) = \frac{\phi_X(\xi)}{\widehat{\phi}(\xi)}$$

Then from (3.5.3), it follows that,

$$T(h) = \int \widehat{h}(\xi)\phi_\mu(\xi)\,d\xi = \int \widehat{h}(\xi)\frac{\phi_X(\xi)}{\widehat{\phi}(\xi)}\,d\xi \tag{3.5.4}$$

Since we observe X, we can estimate ϕ_X by its empirical characteristic function

$$\tilde{\phi}_X(\xi) = \frac{1}{n}\sum_{j=1}^{n}\cos(\xi X_j)$$

Using the fact that $\widehat{\phi}(\xi) = e^{\xi^2/2}$, an estimator of the functional T of the form (3.5.2) is,

$$\widehat{T}(h) = \int \widehat{h}(\xi)\frac{\tilde{\phi}_X(\xi)}{\widehat{\phi}(\xi)}\,d\xi = \int \widehat{h}(\xi)\frac{\frac{1}{n}\sum_{j=1}^{n}\cos(\xi X_j)}{e^{-\xi^2/2}}\,d\xi \tag{3.5.5}$$

Now the standard deviation in using an estimator of the form of $\widehat{T}(h)$ in (3.5.5) is,

$$O\Big(\frac{1}{\sqrt{n}}\Big)\int e^{\xi^2/2}|\widehat{h}(\xi)|\,d\xi$$

Unless we restrict the integral in the Fourier domain to an appropriately chosen compact support, the error is going to blow up. Thus, we can use a slightly modified version of the estimator $\widehat{T}(h)$ in (3.5.5) as

$$\widehat{T}(h;\omega) = \int \widehat{h}(\xi)\omega(\xi)\frac{\frac{1}{n}\sum_{j=1}^{n}\cos(\xi X_j)}{e^{-\xi^2/2}}\,d\xi \tag{3.5.6}$$

where the function ω controls the standard deviation of $\widehat{T}(h;\omega)$.

We present another way of looking at the problem of estimating the functional T. The idea is to construct an appropriate function $f(x)$, and estimate $T(h)$ with

$$\widehat{T}(h) = \frac{1}{n}\sum_{j=1}^{n} f(X_j).$$

In fact, direct calculations show that

$$E[\widehat{T}(h)] = \frac{1}{n}\sum_{j=1}^{n} E[f(X_j)] = \frac{1}{n}\sum_{j=1}^{n} E[(f * \phi)(\mu_j)]$$

where $*$ is the usual convolution. So ideally, the estimator would be unbiased if it were possible to construct an f such that

$$f * \phi \equiv h. \tag{3.5.7}$$

However, for such an f to exist, in the frequency domain f should satisfy

$$\widehat{f} \cdot \widehat{\phi} = \widehat{h}, \qquad \text{or} \qquad \widehat{f}(\xi) = e^{\xi^2/2} \cdot \widehat{h}(\xi). \tag{3.5.8}$$

where $\widehat{r}$ denotes the Fourier transform of r and ϕ is the standard normal density, $\widehat{\phi}(\xi) = e^{\xi^2/2}$. Generally, the function $(\widehat{f}(\xi) = e^{\xi^2/2} \cdot \widehat{h}(\xi))$ is not integrable and hence such an f does not exist. This is the case of PR with $h = \Psi$ and also of ϵ with $h(u) = 1_{\{u=0\}}$.

To overcome this difficulty i.e. to construct an f such that $\widehat{f}$ is integrable and f approximately satistfies (3.5.8), a symmetric continuous function $\omega(\xi)$ is chosen, which will be referred to as a *kernel*, so that the function $\omega(\xi) \cdot e^{\xi^2/2} \cdot \widehat{h}(\xi)$ is integrable. Then $\widehat{f}(\xi)$ in (3.5.8) is replaced by

$$\widehat{f}(\xi;\omega) = \omega(\xi) \cdot e^{\xi^2/2} \cdot \widehat{h}(\xi). \tag{3.5.9}$$

By symmetry and inverse Fourier transformation, the unique f that satisfies (3.5.9) is

$$f(x;\omega) = \int \omega(\xi) \cdot e^{\xi^2/2} \cdot \widehat{h}(\xi) \cos(\xi x)\, d\xi. \tag{3.5.10}$$

Note that a desirable kernel ω should be such that ω has sufficiently thin tail i.e. $\omega(\xi) \approx 0$ for large values of ξ in order to make the existence of f possible, but at the same time for small values of ξ, $\omega(\xi) \approx 1$ so that,

$$f(\cdot;\omega) * \phi \approx h. \tag{3.5.11}$$

Figure (3.2) shows the plot of a desirable kernel ω.

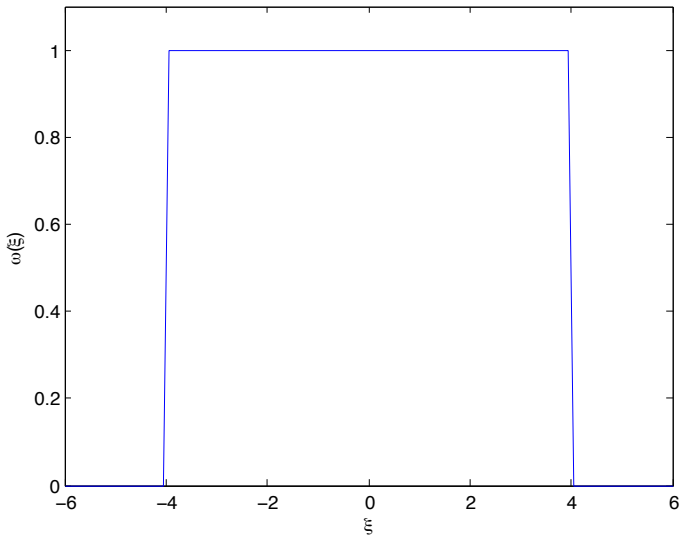

Figure 3.2: Display of a desirable kernel $\omega(\xi)$.

In the literature, it is frequently seen that tampering a function significantly in the frequency domain may only result in a change that is uniformly small in the spatial domain. In this case, the ideal $f(\cdot)$ as described in (3.5.7) cannot be constructed. The function $f(\cdot,\omega)$ in (3.5.10) is an approximate version of $f(\cdot)$ where the approximation is done in the Fourier domain. The difference of $f(\cdot)$ and $f(\cdot,\omega)$ is uniformly small although $\widehat{f}(\cdot)$ and $\widehat{f}(\cdot,\omega)$ are significantly different.

Having constructed $f(\cdot, \omega)$ in (3.5.10) , the functional $T(h)$ can be estimated with

$$\widehat{T}(h; \omega) = \frac{1}{n} \sum_{j=1}^{n} f(X_j;\ \omega). \tag{3.5.12}$$

From both perspectives, in (3.5.6) and (3.5.12), we arrive at the conclusion that it is reasonable to estimate a functional $T(\cdot)$ of the form

$$T(h) = \frac{1}{n} \sum_{j=1}^{n} E[h(\mu_j)] \tag{3.5.13}$$

with estimators of the form,

$$\widehat{T}(h; \omega) = \int \widehat{h}(\xi)\omega(\xi) \frac{1}{n} \sum_{j=1}^{n} \cos(\xi X_j) \cdot e^{\xi^2/2}\, d\xi \tag{3.5.14}$$

for an appropriately chosen kernel ω.

In §4, we give an explicit form of the estimator for positive rate using the Fourier approach described so far. We also study its asymptotic properties in detail. In particular, we give the rate of convergence for the mean squared error of this estimator. We also focus on the sparse case i.e. when the proportion of non-null hypothesis, $\epsilon \to 0$ as $n \to \infty$. In the process, we also derive an efficient kernel for estimating positive rate.

Chapter 4

Main Results

In this chapter, first we construct an estimator $\widehat{PR}(\omega; t, a)$ for positive rate $PR(t, a)$ for a generic kernel ω in (4.0.2). In §(4.1), we obtain an upper bound, for a general kernel ω, of the mean squared error of the estimator in a broad class of models $\mathcal{G}_1$. Then we choose an efficient kernel $\tilde{\omega}$ by minimizing the upper bound uniformly over the class $\mathcal{G}_1$. For this kernel $\tilde{\omega}$ we obtain its rate of convergence. However, in many real datasets, for example in the gene microarray data for leukemia Golub et al. (1999) and colon Alon et al. (1999) data, it is believed that the proportions of signals is very small i.e. ϵ is very small. In these cases, the estimator of PR in (4.0.1) can be modified, to get a better rate of convergence. For this reason, in §(4.3) we consider also another class of models $\mathcal{G}_2$ where the proportion on non-null hypothesis, ϵ, is very small. We derive results for convergence for $\mathcal{G}_2$ as in the case of $\mathcal{G}_1$. In §§(4.3.1), we discuss how to estimate the proportion of non-null hypothesis, ϵ, which is needed for estimating the positive rate using the class $\mathcal{G}_2$.

From here,we will denote $PR(t, a)$ simply with PR for notational simplicity. Unless otherwise mentioned, it should be understood that PR denotes the positive rate for a

threshold value t and a replication multiplicity a. For constructing an estimator of the positive rate, we apply the general framework of Fourier approach introduced in Section 3.5 . From (3.5.1), it follows that

$$1 - PR = \frac{1}{n}\sum_{j=1}^{n} E[\Psi(\mu_j)]$$

Now, the positive rate PR is of the form $T(h)$ as in (3.5.2) with $h(\cdot) = \Psi(\cdot; t, a)$. From (3.5.10) and (3.5.12) it follows that for a general kernel ω, an estimator $\widehat{PR}(\omega; t, a)$ of $PR(t, a)$ can be constructed as

$$\widehat{PR}(\omega) = 1 - \frac{1}{n}\sum_{j=1}^{n} f(X_j; \omega) \tag{4.0.1}$$

where

$$f(x; \omega) = \frac{1}{2\pi} \cdot \int \omega(\xi) \cdot e^{\xi^2/2} \cdot \widehat{\Psi}(\xi; t, a) \cos(x\xi)\, d\xi, \tag{4.0.2}$$

where $\widehat{\Psi}$ denotes the Fourier transform of Ψ with

$$\widehat{\Psi}(\xi; t, a) = \frac{2t}{\sqrt{a}} \cdot e^{-\xi^2/(2a)} \cdot \frac{\sin(t\xi/\sqrt{a})}{t\xi/\sqrt{a}}. \tag{4.0.3}$$

Next, we calculate an upper bound for the bias and the variance of $\widehat{PR}(\omega)$. From §3.1, $X_j \overset{iid}{\sim} N(\mu_j, 1)$, $\frac{1}{n}\{\#j : \mu_j \neq 0) = \epsilon$ and if $\mu_j \neq 0$ then $\mu_j \sim g$ for some density g for $j = 1, \ldots, p$. So this class of models can be parametrized by ϵ and g. We consider the following broad class $\mathcal{G}_1$

$$\mathcal{G}_1 = \{(\epsilon, g) : 0 \leq \epsilon \leq 1 \text{ where } g \text{ is any density }\} \tag{4.0.4}$$

However, as mentioned in the beginning of this chapter, in many real datasets, the proportions of signals, ϵ, is very small. We index ϵ by ϵ_n, assuming that as the number of hypotheses n increases, ϵ_n decreases. Hence, we consider a second class of models $\mathcal{G}_2$, where we consider smooth densities g characterized by the tail behavior of $\widehat{g}$, the Fourier transform of g. Hence, we take

$$\mathcal{G}_2 = \{(\epsilon, g) : \epsilon \in (0, \epsilon_n) \quad \text{and } |\xi|^\alpha |\widehat{g}(\xi)| \leq A \text{ for large } |\xi|\} \tag{4.0.5}$$

where

$$\epsilon_n = C \cdot n^{-\beta} \text{ with } \beta \in [0, \frac{1}{2}], \epsilon_0 \in (0, 1) \quad \text{and } \alpha > 0$$

This class $\mathcal{G}_2$ is smaller than $\mathcal{G}_1$ and it has been proposed before in Cai and Jin (2010).

In the following section, our goal is to find a uniform upper bound for the MSE for $\widehat{PR}$ over $\mathcal{G}_1$ for any kernel ω and then find an optimal kernel by minimizing the upper bound with respect to the kernel ω. §(4.3) deals with the same problem for $\mathcal{G}_2$.

4.1 Estimating positive rate for $\mathcal{G}_1$

Lemma (4.1.1) gives a uniform upper bound of the MSE of $\widehat{PR}(\omega)$ over $\mathcal{G}_1$ for any given kernel ω.

Lemma 4.1.1 *Fix $a \geq 1$ and $t > 0$, and let ω be any kernel. Over the class $\mathcal{G}_1$,*

$$\left(E[\widehat{PR}(\omega)] - PR\right)^2 \leq \frac{t}{\pi\sqrt{a}} \int e^{-\xi^2/a} \cdot (\omega(\xi) - 1)^2 \, d\xi. \tag{4.1.1}$$

and

$$\mathrm{Var}(\widehat{PR}(\omega) \leq \frac{t}{\pi\sqrt{a}} \int \frac{1}{n} e^{(1-\frac{1}{a})\xi^2} \cdot \omega^2(\xi)\, d\xi. \tag{4.1.2}$$

Proof of lemma (4.1.1) is in the appendix. Lemma (4.1.1) gives an upper bound to the MSE of $\widehat{PR}(\omega)$,

$$MSE(\widehat{PR}(\omega)) \leq \frac{t}{\pi\sqrt{a}} \int e^{-\xi^2/a}[(\omega(\xi)-1)^2 + \frac{1}{n} e^{\xi^2} \cdot \omega^2(\xi)]\, d\xi, \tag{4.1.3}$$

Now, the optimal kernel ω is derived in Lemma (4.1.2) by minimizing the right hand side of (4.1.3) using standard variation principle.

Lemma 4.1.2 *Fix $a \geq 1$ and $t > 0$. A continuous compactly-supported kernel that minimizes the right hand side of (4.1.3) is given by*

$$\tilde{\omega}(\xi) = (1 + \frac{1}{n} e^{\xi^2})^{-1}, \qquad -\infty < \xi < \infty \tag{4.1.4}$$

Proof: From (4.1.3) it follows that the optimization problem amounts to minimizing

$$F(\omega) = \frac{1}{\pi} \frac{t}{\sqrt{a}} \int_{-\infty}^{\infty} \left[(\omega(\xi)-1)^2 e^{-\xi^2/a} + \frac{1}{n}\omega^2(\xi) e^{\xi^2(1-\frac{1}{a})}\right] d\xi$$

with respect to ω. In order to minimize F, we use calculus of variation principle. Let ω_1 be any smooth and symmetric function. If F has a minimum at $\tilde{\omega}$, then $F(\tilde{\omega} + \epsilon\omega_1)$ should have a derivative equal to 0 with respect to ϵ at $\epsilon = 0$.

$$\frac{\partial F(\tilde{\omega} + \epsilon\omega_1)}{\partial \epsilon}\bigg|_{\epsilon=0} = \frac{2}{\pi} \frac{t}{\sqrt{a}} \int_{-\infty}^{\infty} \left[(\tilde{\omega}(\xi)-1) e^{-\xi^2/a} + \frac{1}{n}\tilde{\omega}(\xi) e^{\xi^2(1-\frac{1}{a})}\right] \omega_1(\xi)\, d\xi$$

Using the fact that, $\frac{\partial F(\tilde{\omega}+\epsilon\omega_1)}{\partial \epsilon}\Big|_{\epsilon=0} = 0$ and ω_1 smooth, we get

$$(\tilde{\omega}(\xi) - 1)e^{-\xi^2/a} + \frac{1}{n}\tilde{\omega}(\xi)e^{\xi^2(1-\frac{1}{a})} = 0 \forall\ \xi$$

Hence it follows that, $\tilde{\omega}(\xi) = \frac{1}{1+\frac{1}{n}e^{\xi^2}}$. □

A plot of the kernel $\tilde{\omega}$ in (4.1.4) is given in Figure (4.1). The kernel $\tilde{\omega}$ is a compactly supported function which is approximately equal to 1 around 0. This matches our intuition about the kernel in Figure (3.2). Having obtained an efficient kernel $\tilde{\omega}$, we can construct

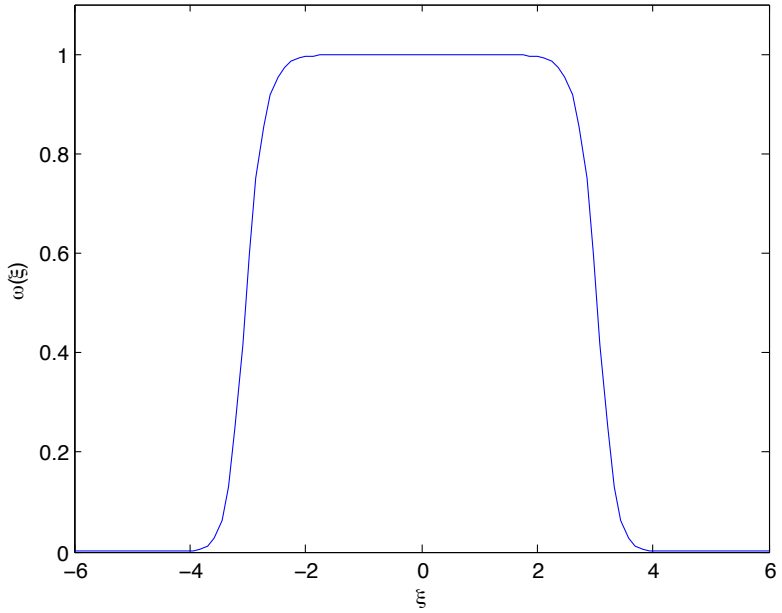

Figure 4.1: Display of kernel $\tilde{\omega}$ in (4.1.4).

the estimator $\widehat{PR}(\tilde{\omega})$ from (4.0.1). The following theorem characterizes the MSE of $\widehat{PR}(\tilde{\omega})$.

Theorem 4.1.1 *Fix $a \geq 1$ and $t > 0$. For sufficiently large n, there is a constant $C =$*

$C(a,t) > 0$ *such that*

$$MSE(\widehat{PR}(\tilde{\omega})) \leq \frac{C \cdot a^2}{\log^2(n)} \cdot n^{-1/a}.$$

The proof of theorem (4.1.1) is in the appendix. Our main goal behind the Fourier approach in (3.5.11) and (4.0.2) was to construct an f using a kernel ω such that

$$f(\cdot;\omega) * \phi(u) = \Psi(u)$$

Since the interesting range of threshold values t are $O(\sqrt{\log n})$, we also give in (4.1.2) the rate of integrated mean squared error for estimating positive rate in the interval $t \in [q_1\sqrt{\log n}, q_2\sqrt{\log n}]$.

Theorem 4.1.2 *Fix $a \geq 1$ and $t > 0$. For sufficiently large n, there is a constant $C = C(a,t,q_1,q_2) > 0$ such that*

$$\int_{q_1\sqrt{\log n}}^{q_2\sqrt{\log n}} MSE(\widehat{PR}(\tilde{\omega};t,a))\,dt \leq \frac{C \cdot a^2}{\log n} \cdot n^{-1/a}.$$

The proof of theorem (4.1.2) is in the appendix. Our main goal behind the Fourier approach in (3.5.11) and (4.0.2) was to construct an f using a kernel ω such that

$$f(\cdot;\omega) * \phi(u) = \Psi(u)$$

With the optimal kernel $\tilde{\omega}$, the difference between

$$f(\cdot;\tilde{\omega}) * \phi(u) \qquad \text{and} \qquad \Psi(u) \tag{4.1.5}$$

is surprisingly small as shown in Figure(4.2), that illustrates the approximation in (4.1.5), where we compare the two functions for $t = 2$ and $a = 2, 4, 6, 8$. For $a \leq 4$, the difference between two functions is very small. As a increases, the rate of convergence of the bias decreases and hence the approximation becomes less accurate.

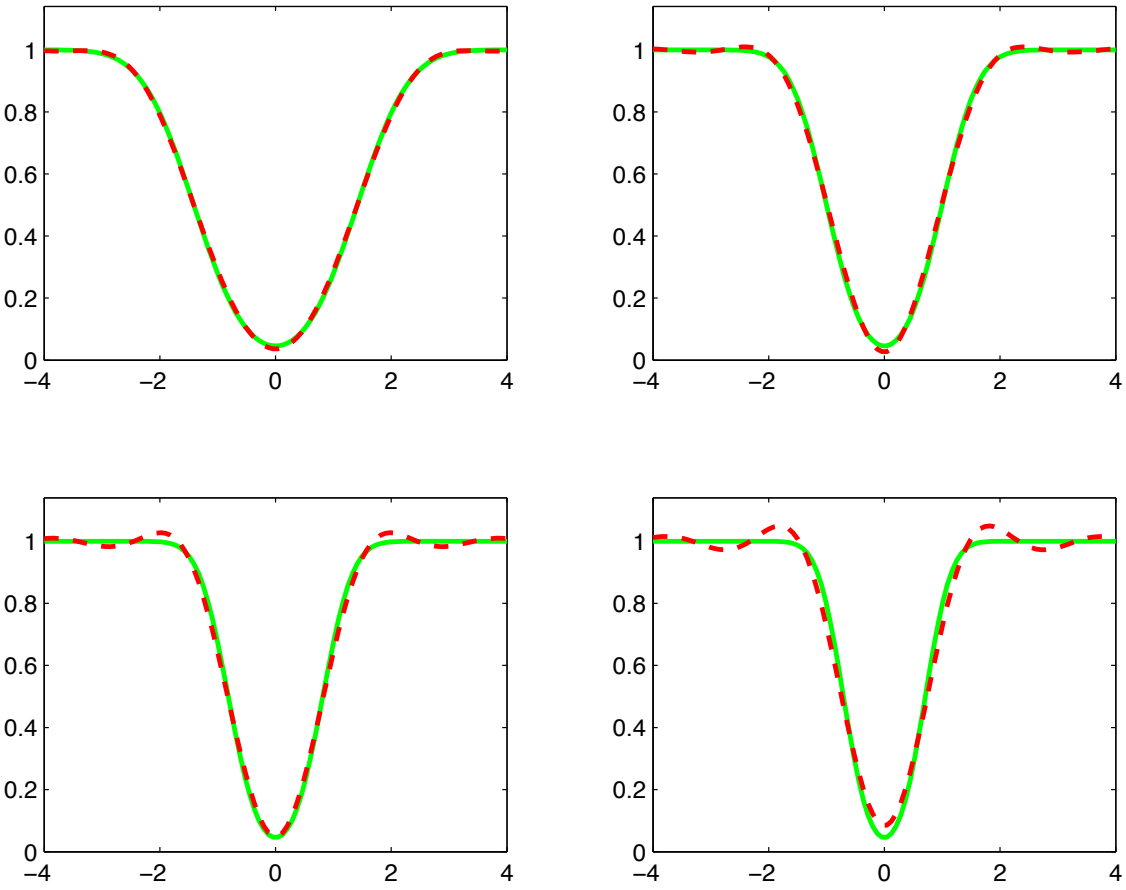

Figure 4.2: Display of $1 - f(\cdot;\tilde{\omega}) * \phi(u)$ (dashed) and $1 - \Psi(u)$ (solid) with $t = 2$, and $a = 2, 4, 6, 8$ from left to right then from top to bottom.

In the next section, we derive a lower bound for minimax risk over $\mathcal{G}_1$ for estimating PR.

4.2 Lower bound for the minimax rate

For deriving an asymptotic lower bound for the minimax risk, as $n \to \infty$, for estimating PR over $\mathcal{G}_1$, we model the $X_j's$ in as

$$X_j \overset{i.i.d.}{\sim} f$$

where f is the Gaussian mixture model

$$f(x) = (1-\epsilon)\phi(x) + \epsilon \int \phi(x-u)g(u)\,du \tag{4.2.1}$$

Similarly we take

$$\mathcal{G}_1 = \{f : f(x) = (1-\epsilon)\phi(x) + \epsilon \int \phi(x-u)g(u)\,du, \quad \epsilon \in (0,1) \ \&\ g \text{ any density}\}$$

All the results proved so far for estimating PR are also true for this case. The minimax risk for estimating PR in $\mathcal{G}_1$ is defined as

$$\mathcal{R}(\mathcal{G}_1) = \inf_{\widehat{T}} \sup_{\mathcal{G}_1} E\Big(\widehat{PR} - PR\Big)^2$$

As we shall see in the next theorem, our proposed estimator matches the lower bound except for a $\log n$ term.

The following theorem characterizes the lower bound for the minimax risk.

Theorem 4.2.1 *Fix $a > 1$ and $t > 0$. For sufficiently large n, there exists a constant*

$C = C(a,t) > 0$ *such that*

$$\lim_{n\to\infty} n^{1/a} \cdot \frac{\log^{2+\alpha(1-1/a)+\frac{1}{a}}(n)}{a^2} \cdot \mathcal{R}(\mathcal{G}_1) \geq C$$

In this case, our proposed estimator for PR achieves the optimal rate except for a log *term.*

We sketch the main idea of the proof. In order to find the lower bound we construct two densities f_1 and $f_2 \in \mathcal{G}_1$ such that f_1 and f_2 are indistinguishable in the sense that their Hellinger distance is $o(1/n)$ but the positive rate (PR) associated with f_1 and f_2 are as far as possible.

Now we discuss the construction of f_1 and f_2. Let

$$h(\xi) = \begin{cases} -\pi|\xi| & 0 \leq |\xi| \leq 1 \\ |\xi|^{-2} & |\xi| > 1 \end{cases}$$

Take $\widehat{g}_j(\xi) = e^{-\xi^2/2} + \vartheta_0 \widehat{w}_j(\xi)$ for $j = 1, 2$ with

$$\widehat{w}_1(\xi) = s_1(|\xi|)h(\xi) + s_2(|\xi|)|\xi|^{-2}$$

$$\widehat{w}_1(\xi) = s_1(|\xi|)h(\xi)$$

where $s_1(\xi)$ is a smooth function around $\xi = 1$. Let $\tau_n = \log(n\epsilon^2) - \log\log n$. $s_2(\xi)$ is a smooth function such that

$$s_2(\xi) = \begin{cases} 0 & 0 \leq \xi \leq \sqrt{\tau_n} \\ 1 & \xi \geq \sqrt{\tau_n + c} \end{cases}$$

where the constant c is chosen in such a way that $\frac{t}{\sqrt{a}} \cdot \sqrt{\tau_n + c} = 2k\pi + \pi/2$. For large enough u, both $w_1(u)$ and $w_2(u)$ are equal to $\frac{1}{|u|^2}(1 + O(1))$. Hence, by choosing ϑ_0 small enough g_j is a density and $f_j(x) = (1-\epsilon)\phi(x) + \epsilon \int \phi(x-u)g_j(u) \in \mathcal{G}_1$ for $j = 1, 2$. Clearly,

$$f_2(x) \geq \begin{cases} C \cdot \epsilon(1+|x|)^{-2} & \text{for large enough } x \,(\text{say}, |x| > c_1) \\ Ce^{-\frac{1}{2}x^2} & |x| \leq c_1 \end{cases} ,$$

Lemma 4.2.1 *Let f_i^n be the joint distribution of $X_1, \ldots, X_n$ for $i = 1, 2$. The Hellinger affinity between f_1^n and $f_2^n (= \rho(f_1^n, f_2(n)))$ $\to 1$ as $n \to \infty$.*

The next lemma gives a lower bound on $|PR(f_1) - PR(f_2)|$.

Lemma 4.2.2 *There exists a constant $C > 0$ such that $r \cdot |PR(f_1 - PR(f_2))|^2 \to C$ as $n \to \infty$.*

Now using the above three lemmas, the proof of Theorem 4.2.1 follows since, for any estimator $\widehat{PR}$ of PR, we have

$$\max[E(\widehat{PR} - PR(f_1))^2, E(\widehat{PR} - PR(f_2))^2] \geq C\rho^4(f_1^n, f_2^n)(|PR(f_1 - PR(f_2))|^2) \geq C \cdot r^{-1}$$

Hence ,

$$\lim_{n\to\infty} r \max[E(\widehat{PR} - PR(f_1))^2, E(\widehat{PR} - PR(f_2))^2] \geq \lim_{n\to\infty} C \cdot \rho^4(f_1^n, f_2^n) = C.$$

4.3 Estimating positive rate for $\mathcal{G}_2$

We now study the special case where ϵ is small and g is a smooth density, as in the class $\mathcal{G}_2$ in (4.0.5), a case that arises in many practical situations. In this case, it is possible to

reduce considerably the rate of mean squared error of $\widehat{PR}(\omega)$ in (4.0.1). In (4.0.1) for any kernel ω , it is possible to reduce the bias of $\widehat{PR}(\omega)$ without increasing its variance, by modifying the estimation procedure as follows. First, for any kernel ω. Recall that

$$\begin{aligned} PR(t;a) &= (1-\epsilon)2\bar{\Phi}(t) + \epsilon \cdot \int (1 - \Psi(u;t,a))\, g(u). \\ &= 1 - \frac{(1-\epsilon)}{2\pi} \int \widehat{\Psi}(\xi;t)\, d\xi - \frac{\epsilon}{2\pi} \cdot \int \big[\int \widehat{\Psi}(\xi;t,a) \cdot \cos(\xi u)\, d\xi \big] g(u)\, du \end{aligned} \tag{4.3.1}$$

At the same time, direct calculations show that

$$E[\widehat{PR}(\omega)] = 1 - \frac{(1-\epsilon)}{2\pi} \int \omega(\xi) \cdot \widehat{\Psi}(\xi;t)\, d\xi - \frac{\epsilon}{2\pi} \cdot \int \big[\int \omega(\xi) \cdot \widehat{\Psi}(\xi;t,a) \cdot \cos(\xi u)\, d\xi \big] g(u)\, du. \tag{4.3.2}$$

From (4.3.1) with (4.3.2) it follows that

$$\begin{aligned} &Bias[\widehat{PR}(\omega)] = (1-\epsilon) b_0(\omega) + \epsilon b_1(\omega) \qquad \text{where} \\ &b_0(\omega) = \tfrac{1}{2\pi} \textstyle\int (1 - \omega(\xi)) \cdot \widehat{\Psi}(\xi;t)\, d\xi \text{ and} \\ &b_1(\omega) = \tfrac{1}{2\pi} \cdot \textstyle\int \big[\int (1 - \omega(\xi)) \cdot \widehat{\Psi}(\xi;t,a) \cdot \cos(\xi u)\, d\xi \big] g(u)\, du. \end{aligned} \tag{4.3.3}$$

Now for any kernel ω, $b_0(\omega)$ is known. So, if we use an estimator $\widehat{\epsilon}$ for ϵ, and then estimate the positive rate by

$$\widehat{PR}(\omega, \widehat{\epsilon}) = \widehat{PR}(\omega) + (1 - \widehat{\epsilon}) b_0$$

then,

$$Bias[\widehat{PR}(\omega, \widehat{\epsilon})] = Bias(\widehat{\epsilon}) \cdot b_0(\omega) + \epsilon b_1(\omega) \tag{4.3.4}$$

In case of $\mathcal{G}_2$ where ϵ is very small, $Bias(\widehat{\epsilon})$ is much smaller than $(1-\epsilon)$ and so comparing (4.3.3) with (4.3.4), it is easy to see that the bias for $\widehat{PR}(\omega, \widehat{\epsilon})$ is smaller than $\widehat{PR}(\omega)$. Before discussing any further about estimation of the positive rate for $\mathcal{G}_2$, we discuss the estimation of the proportion ϵ.

4.3.1 Estimation of proportion ϵ

We restrict our attention here only to using a Fourier approach. The problem of estimating ϵ, the proportion of non-null hypotheses, has been extensively studied using Fourier approach in Jin (2008), Jin and Cai (2007) and Cai and Jin (2010). As mentioned in subsection (3.5), the Fourier approach can be applied for estimating the proportion since

$$\epsilon = 1 - \frac{1}{n}\sum_{j=1}^{n} E[I(\mu_j = 0)]$$

which is of the same form as in (3.5.13) with $h(u) = I(u = 0)$. However, the function h in this case is discontinuous at 0. For estimating ϵ, the function $I(u = 0)$ can be approximated by a continuous function and then we can use the Fourier approach as in (3.5.14). For more details see Jin (2008). The estimator proposed in Jin (2008), based on this approach, is

$$\tilde{\epsilon} = 1 - \frac{1}{n}\sum_{j=1}^{n} \int \tilde{r}(\xi) e^{\xi^2/2} \cos(\xi X_j)\, d\xi \tag{4.3.5}$$

where $\gamma_0 \in (0, \frac{1}{2})$ is an appropriately chosen constant, and the kernel $\tilde{r}$ is any symmetric density on $[-\sqrt{2\gamma_0 \log n}, \sqrt{2\gamma_0 \log n}]$. Here $\tilde{r}$ plays the same role as the general kernel ω in (3.5.14). It follows from Cai and Jin (2010) that $\tilde{\epsilon}$ cannot achieve the optimal rate in the class $\mathcal{G}_2$ although simulations show it performs very well numerically. Here we propose another estimator $\widehat{\epsilon}$, a slightly modified version of $\tilde{\epsilon}$ which performs numerically at least

as good as $\tilde{\epsilon}$ (see Figure (5.1)) and also achieves theoretically the optimal rate . The estimator we propose here is,

$$\widehat{\epsilon} = 1 - \frac{1}{n}\sum_{j=1}^{n}\int r(\xi)e^{\xi^2/2}\cos(\xi X_j)\,d\xi \tag{4.3.6}$$

where $\gamma_0 \in (0, \frac{1}{2})$ and the kernel r is a smooth density on $[-\sqrt{2\gamma_0 \log n}, \sqrt{2\gamma_0 \log n}]$ with no mass on $[-\delta, \delta]$ for some small δ. The choice of an appropriate γ_0 and δ as well as r will be discussed in §(5.1.1) . From Cai and Jin (2010), we have

$$\mathrm{Var}(\widehat{\epsilon}) \leq |\mathrm{Bias}(\widehat{\epsilon})|^2 \leq C \cdot \epsilon_n^2 \frac{1}{(\log n)^{\alpha}} \tag{4.3.7}$$

The rate of MSE of $\widehat{\epsilon}$ will be used in the next section for estimating the MSE for PR.

Rate of MSE for positive rate PR

It follows from (4.3.4), that the final estimator proposed for estimating PR in $\mathcal{G}_2$ is, given a generic kernel ω,

$$\widehat{PR}(\omega, \widehat{\epsilon}) = \widehat{PR}(\omega) + \frac{(1-\widehat{\epsilon})}{2\pi}\int (1 - \omega(\xi)) \cdot \widehat{\Psi}(\xi)\,d\xi \tag{4.3.8}$$

Following the same procedure as in lemma (4.1.1), an efficient kernel ω^* is obtained by minimizing the MSE of $\widehat{PR}(\omega, \widehat{\epsilon})$ as a functional of ω. As in lemma (4.1.2), that the solution of this optimization problem is the kernel,

$$\omega^*(\xi, \epsilon_n, \alpha) = \frac{1}{1 + C \cdot \frac{(\log n)^{\alpha}}{n\epsilon_n^2} e^{\xi^2}} \qquad -\infty < \xi < \infty \tag{4.3.9}$$

Proof of (4.3.9) is obtained by minor modifications of lemma (4.1.1), and hence it is omitted. It is important to note that the kernel ω^* compared to the kernel $\tilde{\omega}$ in (4.1.4) in the class $\mathcal{G}_1$, where the support is $O(\sqrt{\log n})$, the support of ω^* is $O(\sqrt{\epsilon_n^2 \log n})$ i.e.the kernel has a decreasing support as a result of sparsity. As sparsity increases i.e. the proportion of non-null hypotheses, ϵ, decreases, the part of the positive rate, PR, involving the non-null hypotheses decreases. As a result the bias decreases. The larger the support of the kernel, the smaller the bias and the larger the variance. As a result of the trade-off between bias and variance, the support of the kernel decreases since the bias decreases with sparsity. Using the kernel ω^* in (4.3.9), the rate of MSE of $\widehat{PR}(\omega^*, \widehat{\epsilon})$ in the class $\mathcal{G}_2$ is given in the following theorem.

Theorem 4.3.1 *Fix $a \geq 1$ and $t > 0$. For sufficiently large n, there is a constant $C > 0$ such that*

$$MSE[\widehat{PR}(\omega^*, \widehat{\epsilon})] \leq \frac{C \cdot a^2}{\log^{2+\alpha(1-1/a)}(n)} \cdot \epsilon_n^{2(1-1/a)} n^{-1/a}.$$

The proof of the above theorem can be obtained by minor modifications of the Theorem 4.1.1 and hence it is omitted. As we can see above, the kernel giving the optimal rate of convergence for the class $\mathcal{G}_2$ depends on ϵ_n as well as α which are unknown in practice. For practical purposes, when it is known beforehand that the proportion of non-null hypotheses, ϵ, is very small, we propose to use the kernel

$$\omega_{plug}(\xi) = \frac{1}{1 + \frac{e^{\xi^2}}{n\widehat{\epsilon}^2}} \qquad -\infty < \xi < \infty \tag{4.3.10}$$

where $\widehat{\epsilon}$ is given in (4.3.6) and estimate positive rate by $\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$ as in (4.3.8). Since in practice, we do not know the value of ϵ_n in $\mathcal{G}_2$ we plugged in the value of the estimator of proportion, $\widehat{\epsilon}$ in place of ϵ_n.

The lower bound for the minimax risk in the case of $\mathcal{G}_2$ is much harder to obtain compared to $\mathcal{G}_1$. The fact that $\epsilon \to 0$ very fast as $n \to \infty$ makes the problem more difficult. The approach used for the minimax risk in $\mathcal{G}_1$ was that we constructed two densities f_1 and f_2 where the Fourier transform of f_2 is obtained by truncating the Fourier transform of f_1 at a large frequency. This approach does not work in the case of $\mathcal{G}_2$. We believe the MSE of our estimator for $\mathcal{G}_2$ is minimax, at least upto a logarithmic factor, but to prove it some other idea of construction of f_1 and f_2 is needed.

In the next chapter, we will focus about the performance of the estimator $\widehat{\epsilon}$ in (4.3.6) using simulations. We will also simulate data from models in the class $\mathcal{G}_2$ and illustrate the performance of the estimator $\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$ for estimating positive rate. We will compare our method for estimating positive rate with other standard methods and also apply our method for real datasets in gene microarray such as the leukemia data Golub et al. (1999) and the colon data Alon et al. (1999).

Chapter 5

Applications

In this chapter we demonstrate the performance of our estimator for positive rate. We focus only on the case where the proportion of non-null effects, ϵ, is small since it is more relevant for practical purposes. Now our estimator for positive rate in the sparse case, $\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$, as in (4.3.8) with ω_{plug} as in (4.3.10), depends on the estimator, $\widehat{\epsilon}$ of the proportion ϵ in (4.3.6). Hence, in §(5.1) first we show the performance of the estimator $\widehat{\epsilon}$. Then we focus on the performance of $\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$. Then in §(5.2), we apply our estimator for estimating the positive rate for the gene microarray datasets such as the leukemia data Golub et al. (1999) and the colon data Alon et al. (1999).

5.1 Simulation study

In this section, we discuss the choice of the kernel r and δ for estimating the proportion of non-null effects (ϵ), as described in (4.3.6). Then we also test the performance of the resulting estimator by simulation. We also use simulations to test the performance of

$\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$ in (4.3.8) with ω replaced by ω_{plug} in (4.3.10).

5.1.1 Simulation study for estimating proportion

Our proposed estimator from (4.3.6) is

$$\widehat{\epsilon} = 1 - \frac{1}{n}\sum_{j=1}^{n} \int r(\xi) e^{\xi^2/2} \cos(\xi X_j)\, d\xi \tag{5.1.1}$$

where $\gamma_0 \in (0, \frac{1}{2})$ and the kernel r is a smooth density on $[-\sqrt{2\gamma_0 \log n}, \sqrt{2\gamma_0 \log n}]$ with no mass on $[-\delta, \delta]$ for some small δ. In this section, we discuss the choice of the tuning parameters γ_0, δ, and r for estimating ϵ_p. First consider the problem of estimating ϵ_p which involves the choice of the tuning parameter γ_0. Recall from (4.3.6), r is a symmetric density on $[-1, 1]$ with no mass on $[-\delta, \delta]$ for some small δ. We choose $\delta = 0.01$ and

$$r(\xi) = C \cdot e^{\frac{1}{1-\xi^2}}, \qquad \delta < |\xi| < 1$$

and for $\tilde{\epsilon}_p$ in (4.3.5) we again choose

$$\tilde{r}(\xi) = C \cdot e^{\frac{1}{1-\xi^2}}, \qquad |\xi| < 1$$

For more details on choosing the kernel r we refer to Jin (2008). Simulation results show that choosing $\gamma_0 \in [0.2, 0.25]$ gives good numerical result for both $\tilde{\epsilon}$ in (4.3.5) and $\widehat{\epsilon}$. Numerically, their performance depends both on the signal strength and on n. Here, we do simulations for different signal strengths for $n = 5000$ in the following way :
for the signal strength s, we assume under H_1, $\mu \sim U(s, s+1)$ with s fixed in $[1, 4]$. The value for ϵ is taken to be 10%. For each value of s simulate in the following way.

1. Generate μ from $U(s, s+1)$ and then generate an observation X from $(1-\epsilon)N(0,1)+\epsilon N(\mu,1)$.

2. Repeat step 1, n times and estimate ϵ using $\widehat{\epsilon}$ as well as $\tilde{\epsilon}$ with $\gamma_0 = 0.2$.

3. Repeat steps 1 & 2, 100 times and compute the mean squared error of $\widehat{\epsilon}$.

Figure (5.1) illustrates the performance of $\tilde{\epsilon}$ and $\widehat{\epsilon}$. Both seem to perform equally good. However, since $\widehat{\epsilon}$ is theoretically optimal in $\mathcal{G}_2$ while $\tilde{\epsilon}$ is not, we use $\widehat{\epsilon}$ as an estimator of ϵ.

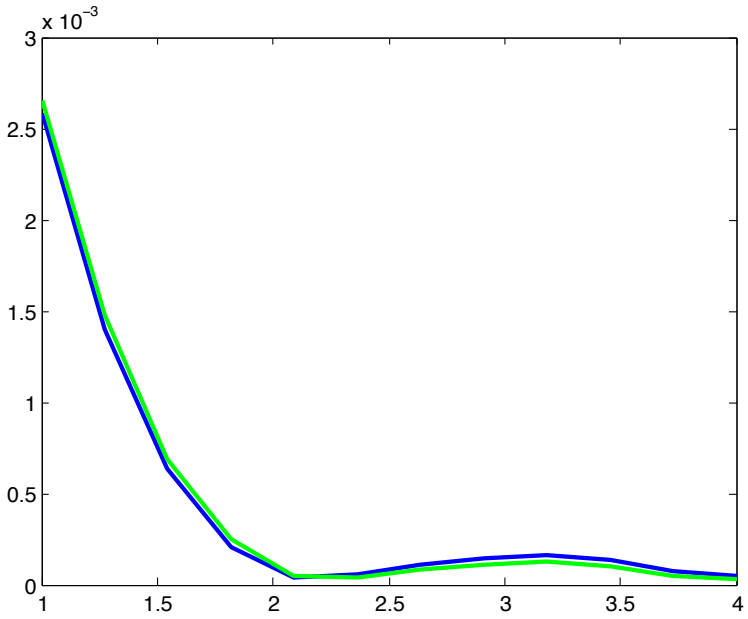

Figure 5.1: The plot displays the mean squared error of $\tilde{\epsilon}$ (green) and $\widehat{\epsilon}$ (blue). The mean squared error is plotted on the $y-$axis versus the signal strength s along the $x-$axis. $\widehat{\epsilon}$ does slightly better for weaker signals whereas $\tilde{\epsilon}$ does slightly better for strong signals, but overall there is not too much difference.

5.1.2 Simulation study for estimating positive rate

We now look at the numerical performance of our estimator $\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$. In this section, our objective is to predict the positive rate,$PR(t, a)$ in (3.4.8), for threshold value t and replication multiplicity a using the available data. Then we compare it with the actual positive rate for a replications.

We set $n = 10,000$. Take the range of the threshold value t to be $(1, 3)$ which is usually the most interesting range for practical purposes. We set the proportion of non-null effects $\epsilon = 5\%$ and $\epsilon = 10\%$. We generate $n\epsilon$ signals from $U(\frac{1}{\sqrt{a}}, \frac{1}{\sqrt{a}} + 1)$ so that $\sqrt{a} \cdot \mu$ is a constant. Our goal is to simulate weak signals inversely proportional to the number of replications a. For the number of replications a,we take $a = 2$ and $a = 4$. Now, for each value of a and each value of ϵ, we do the following steps:

1. Generate $n\epsilon$ values of μ from $U(\frac{1}{\sqrt{a}}, \frac{1}{\sqrt{a}} + 1)$.
2. For each such value of μ generate an observation from $N(\mu, 1)$. Generate $n(1 - \epsilon)$ observations from $N(0, 1)$.
3. Using our estimator $\widehat{PR}(\omega_{plug}, \widehat{\epsilon})$ we predict the positive rate for each value of the threshold t.
4. Repeat steps, 2 and 3 for 100 independent cycles.

As it can be seen in Figure (5.2), for the case $a = 2$, the estimated positive rate almost merges with the true positive with very low variance. For the case, $a = 4$, the mean squared error is slightly larger. The yellow dashed curve in Figure (5.2) represents the positive rate from the available data. In the case of $\epsilon = 10\%$, when the proportion of signals is moderately high, the true positive rate for both replication multiplicities $a = 2$

and $a = 4$, as well as the estimated positive rate is much larger than the current positive rate. This is very encouraging in practice, because our estimator tells that by increasing the number of replications, the power of the testing procedure can be increased.

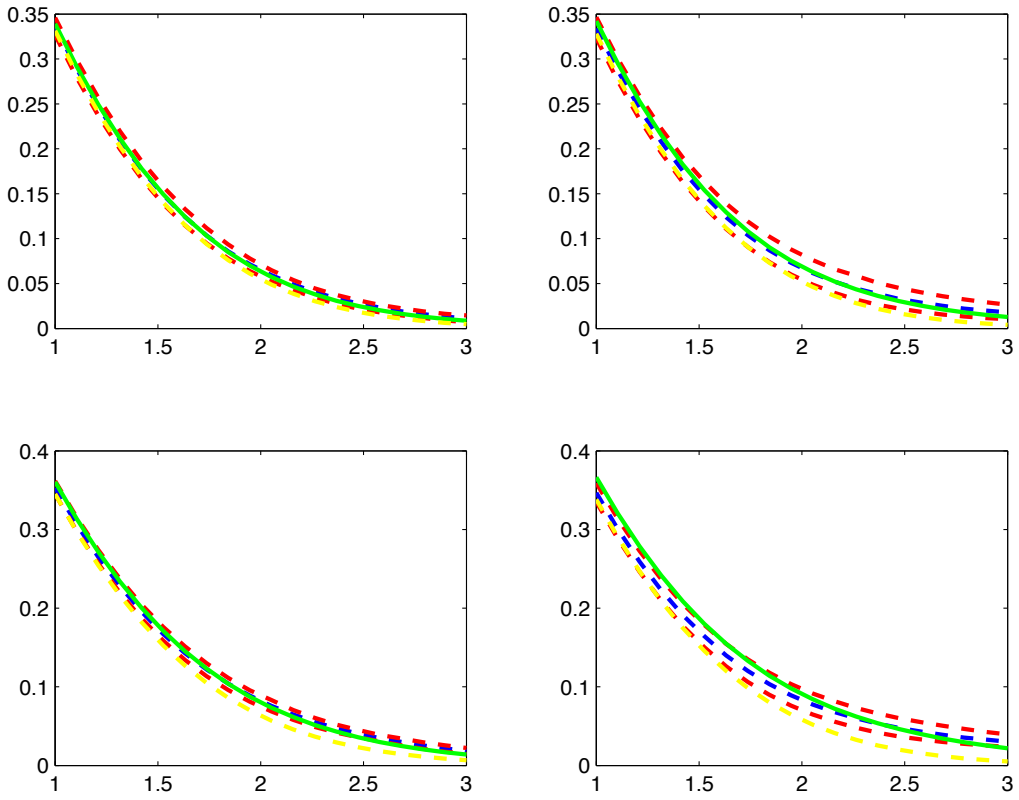

Figure 5.2: Display of the positive rate (PR) for threshold values $t \in [1, 3]$. The top row is for $\epsilon = 5\%$ and the bottom row is for $\epsilon = 10\%$ with $a = 2$ and $a = 4$ replications from left to right. The solid line(green) is the true PR and the blue dashed line is the estimated PR. The yellow dashed line is the PR from the current data. The red dashed line is the 95% confidence interval.

5.2 Application to real datasets

We consider gene microarray data for cancer for colon Alon et al. (1999) and leukemia Golub et al. (1999). Below, we give a brief description of both the datasets.

- **Colon data:** This data is based on cancerous growths (tumors) found in the tissue of colon. This dataset contains 62 samples. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. The total number of genes to be tested is 2000.

- **Leukemia data:** Leukemias are primary disorders of bone marrow. They are malignant neoplasms of hematopoietic stem cells. The total number of genes to be tested is 7129, and number of samples to be tested is 72, which are all acute leukemia patients, either acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML).

In both of the datasets described above, data on gene microarray is collected for two distinct groups and the main problem of interest is — *which genes are differentially expressed between these two groups* ? We want to test the performance of our method assuming only half of the samples is available to us i.e. only 11 controls and 20 patients in case of the colon data and, 24 controls and 12 patients in case of the leukemia data. We want to predict the positive rate (PR) for the whole dataset i.e. for 2 replications ($n = 2$).

Our objective is two-fold: one, we want to see if there is any significant difference between our estimator and the positive rate of the current data , in which case one can get an idea if there were twice the number of samples available, then how many more

true discoveries can be made and second, we want to see how close is our estimator to the positive rate for the whole data having observed only half of the data.

5.2.1 Converting the data to $z-$scores and estimating the null

In order to get the $z-$scores, first we follow the standard procedure of calculating the two sample $t-$statistics for all the genes. Our assumption is for most of the genes the $t-$statistics are independent and identically distributed observation from $N(\mu_0, \sigma_0^2)$ and a small proportion of the genes might come from a mixture of different normally distributed subpopulations. So, we estimate the null distribution using the method in Jin and Cai (2008) and then standardize the $t-$statistics using the estimated null parameters. To give the reader an idea of how these z-scores look, we plotted the histogram for the whole data in Figure 5.3 along with standard normal density. From Figure (5.3) we can say that most of the observations are coming from $N(0,1)$ distribution except for a small proportion. The estimate for the proportion of non-null effects, $\widehat{\epsilon}_p$ turns out to be 7.03% for colon data and 8.56% for the leukemia data.

5.2.2 Power prediction using our estimator

We chose at random two-thirds of the data and computed our estimator. We also used the fact that the proportion of non-null effects is bounded by 5% which is widely believed to be true by biologists. We repeated this procedure 100 times and computed the mean of the estimator as well as the 5% and 95% quantiles over all the 100 permutations. In Figure (5.4), the results can be found. In Figure (5.5), we zoomed in on Figure (5.4) in the region where the predicted PR is significantly higher than the current PR.

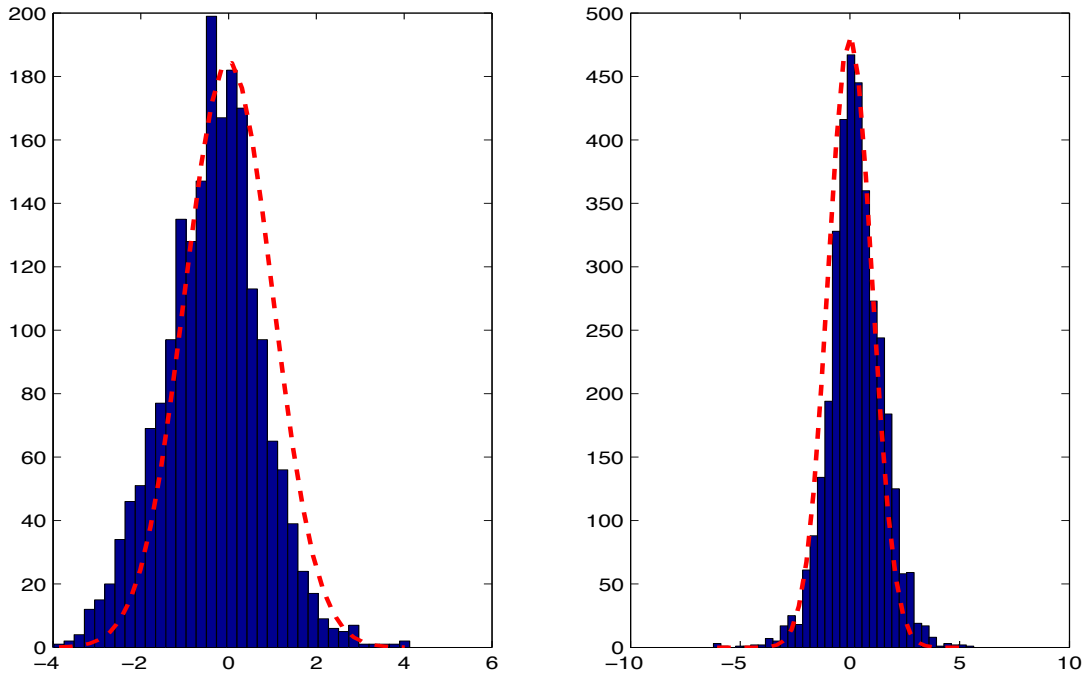

Figure 5.3: Display of histogram for the $z-$scores using the whole dataset. The left plot is for the colon data and the right plot is for the leukemia data. The red dashed line is the $N(0,1)$ density.

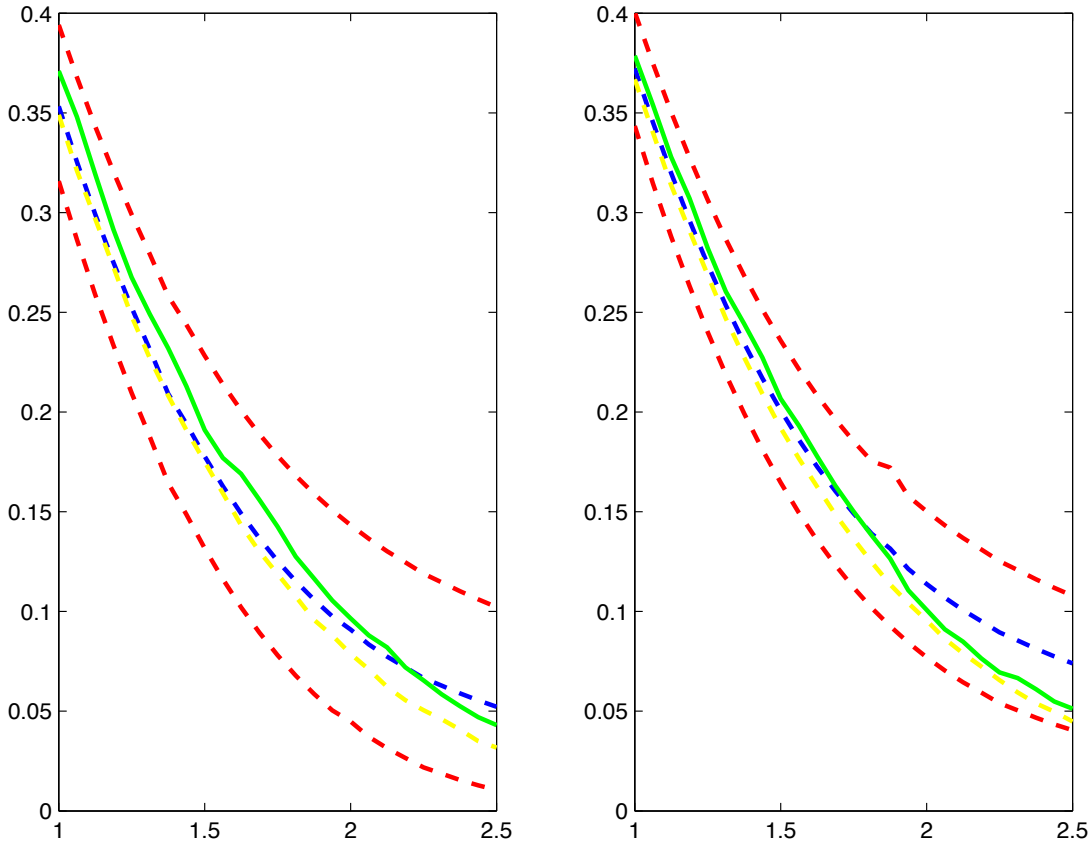

Figure 5.4: Display of the mean of the estimator (blue dashed) along with the 5% and 95% quantiles (red dashed). The green curve is the survival function for the whole data and the dashed yellow curve is the false positive rate

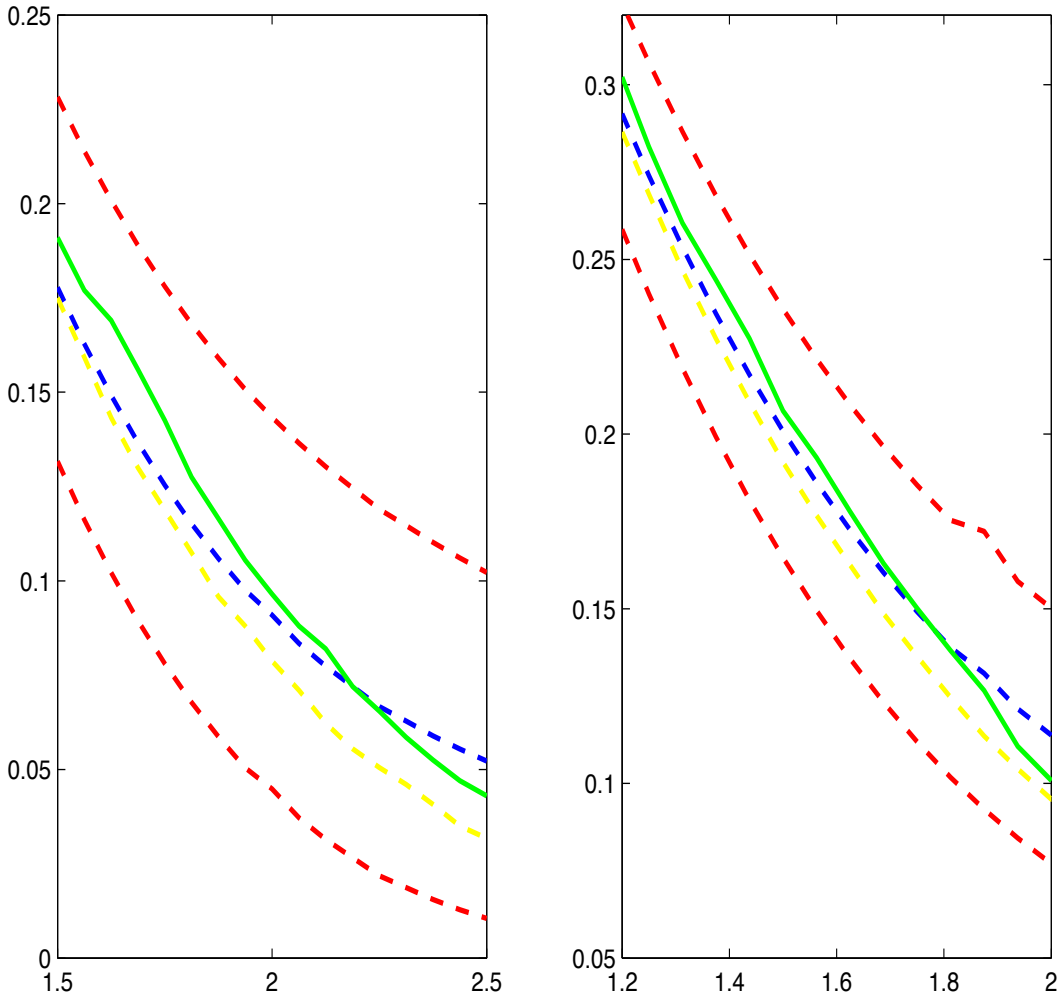

Figure 5.5: Figure (5.4) zoomed in the region where predicted PR exceeds the current PR.

Chapter 6

Conclusion and Future Work

In this thesis, I have given a practical solution to address the problem of estimating the power one can obtain for a given number of replications in multiple testing. I have estimated the positive rate for a given threshold t and replication multiplicity a, which is both convenient to estimate since it is a simple average, as well as serves the purpose of studying the relationship between power and replications as described earlier in the introduction. I have derived good theoretical properties of this estimator. In particular, an asymptotic lower bound for the minimax risk is established for a very general class of models, $\mathcal{G}_1$. Our proposed estimator in this class matches this lower bound except for a logarithmic term. I also obtained the rate of convergence of the MSE in the sparse case, i.e. when the proportion of non-null hypotheses ϵ, is small ($\mathcal{G}_2$). This case is of practical importance, since for many real data such as the leukemia data Golub et al. (1999) and colon data Alon et al. (1999), the proportion of signals is small. I believe the rate of MSE in $\mathcal{G}_2$ is asymptotically minimax, but deriving a sharp lower bound for minimax risk in this case needs a different approach as compared to the case of $\mathcal{G}_1$. This will be discussed

in more details in future work. I also demonstrated very encouraging performance of our estimator through simulations as well as real data.

The next logical steps for this research falls into two categories: studying the minimax risk for estimating PR in the class $\mathcal{G}_2$ and trying to find an estimator for PR which adapts to unknown sparsity in the class$\mathcal{G}_2$. These are discussed in details in the next section.

6.1 Future work

6.1.1 Minimax risk for sparse case

From theorem(4.3.1), it follows that the rate of MSE of $\widehat{PR}(\omega^*, \widehat{\epsilon})$ for estimating PR in $\mathcal{G}_2$ is

$$MSE[\widehat{PR}] \leq \frac{C \cdot a^2}{\log^{2+\alpha(1-1/a)}(n)} \cdot \epsilon_n^{2(1-1/a)} n^{-1/a}.$$

where $X_j \overset{i.i.d.}{\sim} f$, g is a density and

$$\mathcal{G}_2 = \{f : f(x) = (1-\epsilon)\phi(x) + \epsilon \int \phi(x-u)g(u)\,du \quad \epsilon \in (0, \epsilon_n) \ \& \ \ |\xi|^\alpha |\widehat{g}(\xi)| \leq A \text{ for large } |\xi|\}$$

with $\epsilon_n = \epsilon_0 n^{-\beta}$ with $\beta \in [0, \frac{1}{2}]$. Now the minimax rate for estimating PR in $\mathcal{G}_2$ is defined to be

$$\mathcal{R}(\mathcal{G}_2) = \inf_{\widehat{T}} \sup_{\mathcal{G}_2} E[\widehat{T} - PR]^2$$

The lower bound for the minimax risk in the case of $\mathcal{G}_2$ is much harder to obtain compared to $\mathcal{G}_1$. The fact that $\epsilon \to 0$ very fast as $n \to \infty$ makes the problem more difficult. The approach used for the minimax risk in $\mathcal{G}_1$ was that we constructed two densities f_1 and f_2 where the Fourier transform of f_2 is obtained by truncating the Fourier transform of f_1

at a large frequency. This approach does not work in the case of $\mathcal{G}_2$. The main problem is the following. The Fourier transform of PR is of the form $a(\xi)e^{-\xi^2/(2a)}$ which follows from (4.0.3). Moreover, the Fourier transform of standard normal density is $e^{-\xi^2/2}$. Suppose we construct two densities f_1 and f_2 similar to $\mathcal{G}_2$ by truncating such that

$$\widehat{f}_1(\xi) = \begin{cases} \widehat{f}_2(\xi) & \text{if } |\xi| \leq \tau \\ 0 & \text{otherwise} \end{cases} \tag{6.1.1}$$

Now,

$$L^2(f_1, f_2) = \epsilon_n^2 e^{-\tau^2} p(\tau) \quad \text{and} \quad \Big(PR(f_1) - PR(f_2)\Big)^2 = e^{-\tau^2/a} p(\tau)$$

where $p(\cdot)$ is a generic polynomial function. The smoothness of the Fourier transform of PR puts the error in PR for f_1 and f_2 in the same asymptotic scale as the L^2 error of f_1 and f_2. Choosing τ such that $L^2(f_1, f_2) = O(1/n)$ gives us just the lower bound that we get from our proposed estimator. Now, the Hellinger distance between f_1 and f_2 is even smaller because $f_1(x) = (1-\epsilon)\phi(x) + \epsilon \int \phi(x-u)g_1(u)$. The first component of f_1 in the spatial domain which is $\phi(x)$ is very smooth while in the second component there is ϵ which lies in the interval $(0, C \cdot n^{-\beta})$. This increases the Hellinger distance, and hence we need to truncate for a larger interval in which case the lower bound is much smaller.

In deconvolution problems, if the bias dominates the variance or the variance dominates the bias then this kind of frequency matching method works. In this case, the main problem is the smoothness of the Fourier transform of the function to be estimated is of the same order as the smoothness of the characteristic function of standard normal. There can be two possible reasons for the failure of this approach. Either the two point testing argument is not suitable or a different approach for construction of densities for a two point testing argument is necessary.

6.2 Adapting to unknown sparsity

Consider the class $\mathcal{G}_2$ again from above. In (4.3.9), I derived the kernel ω^* which minimizes the MSE of the proposed estimator by a trade-off between the bias and the variance in $\mathcal{G}_2$. Basically this kernel thresholds at frequency $\gamma\ O\Big(\sqrt{\log(n\epsilon_n^2)}\Big)$ in the Fourier domain. However, in practice this threshold is unknown. Hence, I suggested to use the kernel ω_{plug} in practice which plugs in the estimator of ϵ in place of ϵ_n. This estimator performs well in simulations as well as real data as shown in chapter 5.

However, it will be interesting to see if its possible to use a data dependent threshold which can adapt to this unknown sparsity. In the case of estimating the coordinates of the mean vector in the sparse case, Abramovich et al. (2006) showed that hard thresholding using the false discovery rate gives the optimal procedure. As mentioned in chapter (2), the false discovery rate is able to adapt to unknown sparsity. Similar ideas can be used in this case to derive an adaptive estimation strategy.

Appendix A

Proofs

A.1 Proof of Theorem 4.1.1

Denote the bias of $PR^*(t,a)$ by $b^*(\omega;t,a)$. Using (A.2.8) we get,

$$|b^*(\omega;t,a)| \leq \frac{t}{\pi\sqrt{a}} \int |\omega^*(\xi) - 1| \cdot |\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}| \cdot e^{-\xi^2/2a}\, d\xi.$$

Substituting ω^* from Lemma 4.1.2, we get

$$|b^*(\omega;t,a)| \leq \frac{t}{\pi\sqrt{a}} \int \Big[\frac{\frac{1}{n}e^{\xi^2}}{1+\frac{1}{n}e^{\xi^2}} |\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}| \cdot e^{-\xi^2/2a}\Big]\, d\xi.$$

We introduce some more notations for simplicity. Let $a(\xi) = \frac{\frac{1}{n}e^{\xi^2(1-\frac{1}{2a})}}{1+\frac{1}{n}e^{\xi^2}}$, $b(\xi) = |\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}|$ and

$$I_1 = \int \frac{\frac{1}{n}e^{\xi^2(1-\frac{1}{2a})}}{1+\frac{1}{n}e^{\xi^2}} |\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}|\, d\xi. \tag{A.1.1}$$

Then, $I_1 = \int a(\xi)b(\xi)\,d\xi$. The change of variable, $\xi = \sqrt{\log n} + \frac{\eta}{2\sqrt{\log n}}$ in (A.1.1), gives

$$a(\xi) = \frac{\frac{1}{n}e^{\left(\log n+\eta+\frac{\eta^2}{4\log n}\right)(1-\frac{1}{2a})}}{1+\frac{1}{n}e^{\left(\log n+\eta+\frac{\eta^2}{4\log n}\right)}} = n^{-\frac{1}{2a}}\frac{e^{(\eta+\frac{\eta^2}{4\log n})(1-\frac{1}{2a})}}{1+e^{(\eta+\frac{\eta^2}{4\log n})}}$$

$$\text{and } b(\xi) = \Big|\frac{\sin\left(\frac{t\sqrt{\log n}}{\sqrt{a}}(1+\frac{\eta}{2\log n})\right)}{\frac{t\sqrt{\log n}}{\sqrt{a}}(1+\frac{\eta}{2\log n})}\Big| \le \frac{\sqrt{a}}{t\sqrt{\log n}}\frac{1}{|1+\frac{\eta}{2\log n}|}\,, \text{ and hence}$$

$$I_1 = \int a(\xi)b(\xi)\,d\xi \le \frac{\sqrt{a}n^{-\frac{1}{2a}}}{t\log n}\int \frac{1}{|1+\frac{\eta}{2\log n}|}\frac{e^{(\eta+\frac{\eta^2}{4\log n})(1-\frac{1}{2a})}}{1+e^{(\eta+\frac{\eta^2}{4\log n})}}\,d\eta$$

As $n\to\infty$, by Dominated Convergence Theorem,

$$\int \frac{1}{|1+\frac{\eta}{2\log n}|}\frac{e^{(\eta+\frac{\eta^2}{4\log n})(1-\frac{1}{2a})}}{1+e^{(\eta+\frac{\eta^2}{4\log n})}}\,d\eta \sim \int \frac{e^{\eta(1-\frac{1}{2a})}}{1+e^{\eta}}d\eta \sim 2a\cdot C_1 \Rightarrow I_1 \lesssim C_1\Big(\frac{a^{3/2}n^{-\frac{1}{2a}}}{t\log p}\Big) \tag{A.1.2}$$

Combining (A.1.1) and (A.1.2) gives,

$$\big|b^*(\omega;t,a)\big| \lesssim C_1\cdot n^{-(\frac{1}{2a})}\frac{a}{\log n} \le C_1\cdot n^{-(\frac{1}{2a})}\frac{a}{\log n}\,, \tag{A.1.3}$$

uniformly for all ϵ between 0 and 1. Using (A.2.11),

$$\begin{aligned}\mathrm{Var}(PR^*(t,a)) &\le \frac{1}{n}\cdot E\big[\big(\frac{t}{\pi\sqrt{a}}\int \omega^*(\xi)\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}e^{(1-\frac{1}{a})\xi^2/2}\cos(\xi X_1)\,d\xi\big)^2\big]\\ &\le \frac{t^2}{n\pi^2 a}\big(\int |\omega^*(\xi)|\big|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\big|e^{(1-\frac{1}{a})\xi^2/2}\,d\xi\big)^2\end{aligned}$$

Inserting ω^* from Lemma 4.1.2 and using triangle inequality and symmetry around 0 gives,

$$\text{Var}(PR^*(t,a)) \leq 4\frac{t^2}{n\pi^2 a}\Big(\int_0^\infty \frac{1}{1+\frac{1}{n}e^{\xi^2}}\Big[\Big|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\Big|e^{\xi^2(1-\frac{1}{a})\frac{1}{2}}\Big]\,d\xi\Big)^2 \tag{A.1.4}$$

We introduce some more notations.

$$\text{Let}\,, I_{21} = \int_0^\infty \Big|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\Big|\frac{e^{\xi^2(1-\frac{1}{a})\frac{1}{2}}}{1+\frac{1}{n}e^{\xi^2}}\,d\xi \tag{A.1.5}$$

Also let, $a_1(\xi) = \frac{e^{\xi^2(1-\frac{1}{a})\frac{1}{2}}}{1+\frac{1}{n}e^{\xi^2}}$ and, as before, $b(\xi) = \big|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\big|$. Then, $I_{21} = \int_0^\infty b(\xi)a_1(\xi)\,d\xi$. Using the change of variable, $\xi = \sqrt{\log n} + \frac{\eta}{2\sqrt{\log n}}$ in I_{21} we get,

$$a_1(\xi) = \frac{e^{(\log n+\eta+\frac{\eta^2}{4\log n})(1-\frac{1}{a})\frac{1}{2}}}{1+\frac{1}{n}e^{(\log n+\eta+\frac{\eta^2}{4\log n})}} = n^{(1-\frac{1}{a})\frac{1}{2}}\frac{e^{(\eta+\frac{\eta^2}{4\log n})(1-\frac{1}{a})\frac{1}{2}}}{1+e^{\eta+\frac{\eta^2}{4\log n}}} \text{ and as before } b(\xi) \leq \frac{\sqrt{a}}{t\sqrt{\log n}}\frac{1}{|1+\frac{\eta}{2\log n}|}$$

Hence we get,

$$\begin{aligned} I_{21} &\leq \frac{\sqrt{a}}{t}\frac{n^{(1-\frac{1}{a})\frac{1}{2}}}{\log n}\int_{-\infty}^\infty \frac{e^{(\eta+\frac{\eta^2}{4\log n})(1-\frac{1}{a})\frac{1}{2}}}{1+e^{\eta+\frac{\eta^2}{4\log n}}}\frac{\mathbf{1}(\eta > -2\log n)}{|1+\frac{\eta}{2\log n}|}\,d\eta \\ &\sim \frac{\sqrt{a}}{t}\frac{Cn^{(1-\frac{1}{a})\frac{1}{2}}}{\log n}\int_{-\infty}^\infty \frac{e^{\eta(1-\frac{1}{a})\frac{1}{2}}}{1+e^{\eta}}d\eta = C\cdot\frac{\sqrt{a}}{t}\frac{n^{(1-\frac{1}{a})\frac{1}{2}}}{\log n} \end{aligned} \tag{A.1.6}$$

The last approximation in (A.1.6) follows from Dominated Convergence Theorem.

Inserting I_{21} in (A.1.4) gives,

$$\text{Var}(PR^*(t,a)) \leq C\cdot\frac{n^{-\frac{1}{a}}}{\log^2 n}\Big(1+\frac{a^{3/2}}{t\log^{3/2} n}\Big)^2 \sim C\cdot\frac{n^{-\frac{1}{a}}}{\log^2 n} \tag{A.1.7}$$

Together, (A.1.3) and (A.1.7) gives,

$$\mathrm{MSE}(PR^*(t,a)) \lesssim C \cdot \Big(n^{-\frac{1}{a}} \frac{a^2}{\log^2 n} + \frac{n^{-\frac{1}{a}}}{\log^2 n}\Big) \sim C \cdot n^{-\frac{1}{a}} \frac{a^2}{\log^2 p}$$

A.2 Proof of Lemma 4.1.1

Consider the first claim. For short, denote the bias by $b(\omega;t,a) = E[\widehat{PR}(\omega;t,a)] - PR(t;a)$. By (4.0.3) and (4.3.2),

$$\begin{aligned} |b(\omega;t,a)| &= \frac{t}{\pi\sqrt{a}} \Big| \int (\omega(\xi)-1) \frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}} e^{-\xi^2/2a} \cdot \big[(1-\epsilon) + \epsilon \int \cos(\xi u)\, dF(u)\big] d\xi \Big| \\ &\le \frac{t}{\pi\sqrt{a}} \int |\omega(\xi)-1| \cdot \Big| \frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}} \Big| \cdot e^{-\xi^2/2a}\, d\xi. \end{aligned} \tag{A.2.8}$$

Use Hölder inequality,

$$\Big(\int |\omega(\xi)-1| \cdot \Big|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\Big| \cdot e^{-\xi^2/2a}\, d\xi\Big)^2 \le \Big(\int (\omega(\xi)-1)^2 \cdot e^{-\xi^2/a}\, d\xi\Big) \cdot \Big(\int \Big(\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\Big)^2 d\xi\Big), \tag{A.2.9}$$

where by elementary calculus,

$$\int \Big(\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}\Big)^2 d\xi = \frac{\sqrt{a}}{t} \int \frac{\sin^2(\eta)}{\eta^2}\, d\eta = \sqrt{a}\pi/t. \tag{A.2.10}$$

Inserting (A.2.9) and (A.2.10) into (A.2.8) gives the first claim.

Consider the second claim. By definition and symmetry

$$\begin{aligned} \mathrm{Var}[PR(\omega;t,a)] &= \frac{1}{n} \cdot \mathrm{Var}\Big(\frac{t}{\pi\sqrt{a}} \int \omega(\xi) \frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}} e^{(1-\frac{1}{a})\xi^2/2} \cos(\xi X_1)\, d\xi\Big) \\ &\le \frac{1}{n} \cdot E\Big[\Big(\frac{t}{\pi\sqrt{a}} \int \omega(\xi) \frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}} e^{(1-\frac{1}{a})\xi^2/2} \cos(\xi X_1)\, d\xi\Big)^2\Big]. \end{aligned} \tag{A.2.11}$$

Since $|\cos(\xi X_1)| \leq 1$,

$$E[(\int \omega(\xi)\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}} e^{(1-\frac{1}{a})\xi^2/2}\cos(\xi X_1)\,d\xi)^2] \leq (\int |\omega(\xi)|\cdot|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}|\cdot e^{(1-\frac{1}{a})\xi^2/2}\,d\xi)^2. \tag{A.2.12}$$

Now, by similar argument,

$$(\int |\omega(\xi)|\cdot|\frac{\sin(\xi t/\sqrt{a})}{\xi t/\sqrt{a}}|\cdot e^{(1-\frac{1}{a})\xi^2/2}\,d\xi)^2 \leq \frac{t}{\pi\sqrt{a}}\int \omega^2(\xi)e^{(1-\frac{1}{a})\xi^2}\,d\xi. \tag{A.2.13}$$

Inserting (A.2.12) and (A.2.13) into (A.2.11) gives the second claim. □

A.3 Proof of Lemma 4.2.1

The hellinger distance equals $h^2(f_1,f_2) \leq \epsilon^2 \int \frac{(f_1(x)-f_2(x))^2}{f_2(x)}\,dx \leq \epsilon^2(I+II)$ where $I = C\int_{|x|<c_1}(f_1(x)-f_2(x))^2\,dx$ and $II \leq C\frac{1}{\epsilon}\int_{|x|\geq c_1}(1+|x|^2)(f_1(x)-f_2(x))^2\,dx$. Using Parseval's identity and Lemma (**??**) we get,

$$\begin{aligned} I &\leq C\int (f_1(x)-f_2(x))^2\,dx = C\int (\widehat{f}_1(\xi)-\widehat{f}_2(\xi))^2\,d\xi \\ &\leq C\int_{|\xi|>\sqrt{\tau_n}} e^{-\xi^2}|\xi|^{-2}\,d\xi \leq Ce^{-\tau_n}{\tau_n}^{-2-(1/2)} \leq \frac{C}{n\epsilon^2\log n} \end{aligned}$$

$$
\begin{aligned}
II &\leq C\frac{1}{\epsilon}\int (1+|x|^2)(f_1(x)-f_2(x))^2\,dx \\
&= C\frac{1}{\epsilon}\Big[\int (\widehat{f}_1(\xi)-\widehat{f}_2(\xi))^2\,d\xi + \int (\widehat{f}_1^{(1)}(\xi)-\widehat{f}_2^{(1)}(\xi))^2\,d\xi\Big] \\
&\leq C\frac{1}{\epsilon}\int_{|\xi|>\sqrt{\tau_n}} e^{-\xi^2}\xi^{2-4}\,d\xi \\
&\leq C\frac{1}{\epsilon}e^{-\tau_n}\tau_n^{1-2-(1/2)} = \frac{C}{n\epsilon^3\sqrt{\log n}}
\end{aligned}
$$

Hence, $h^2(f_1,f_2) \leq \frac{C}{n}\cdot\frac{1}{\epsilon\sqrt{\log n}}$ which implies the Hellinger affinity between f_1^n and $f_2^n = \rho(f_1^n, f_2^n) \to 1$ as $n\to\infty$.

A.4 Proof of Lemma 4.2.2

Using Parseval's identity, we can write

$$
\begin{aligned}
|PR(f_1)-PR(f_2)| &= \epsilon\Big|\int \Psi(u;t,a)(g_1(u)-g_2(u))\,du\Big| \\
&= \frac{\epsilon}{2\pi}\Big|\int_{\xi>\sqrt{\tau_n}} \widehat{\Psi}(\xi;t,a)s_2(\xi)|\xi|^{-2}\,d\xi\Big|
\end{aligned}
$$

where

$$
\widehat{\Psi}(\xi;t,a) = 2e^{-\xi^2/(2a)}\cdot\frac{\sin(t\xi/\sqrt{a})}{\xi}.
$$

Let, $|PR(f_1)-PR(f_2)| = |III+IV|$ where

$$
III = \frac{\epsilon}{\pi}\int_{\sqrt{\tau_n}<\xi<\sqrt{\tau_n+c}} \widehat{\Psi}(\xi;t,a)s_2(\xi)|\xi|^{-2}\,d\xi
$$

and

$$IV = \frac{\epsilon}{\pi} \int_{\xi \geq \sqrt{\tau_n + c}} \widehat{\Psi}(\xi; t, a) |\xi|^{-2} \, d\xi$$

Observe, that s_2 is an analytic function with bounded derivatives in the interval $\sqrt{\tau_n} < \xi < \sqrt{\tau_n + c}$. Using Taylor's theorem we get, $|s_2(\xi)| \leq C \cdot |\xi - \sqrt{\tau_n}|$ for $\sqrt{\tau_n} < \xi < \sqrt{\tau_n + c}$. Let $r = \frac{n^{1/a}}{\epsilon^{2(1-\frac{1}{a})}} \frac{\log^{2+2(1-1/a)+1/a}(n)}{a^2}$. Now we show that, $r^{1/2}|III| = o(1)$ as $n \to \infty$. Substituting $\xi = \sqrt{\tau_n} + \frac{\eta}{\sqrt{\tau_n}}$, and letting $l(n) = \sqrt{\tau_n}(\sqrt{\tau_n + c} - \sqrt{\tau_n})$, we get

$$\begin{aligned} r^{1/2}|III| &\leq C r^{1/2} \epsilon^{1-1/a} \frac{n^{-1/2a}}{\log^{1+\frac{2}{2}(1-1/a)+\frac{1}{2a}}(n)} \int_0^{l(n)} \frac{|\sin(\sqrt{\tau_n} + \frac{\eta}{\sqrt{\tau_n}})|}{(1+\eta/\tau_n)(1+2)} \cdot \frac{\eta}{\sqrt{\tau_n}} e^{-\eta/a} e^{-\eta^2/2\tau_n} \, d\eta \\ &\leq C \frac{1}{a} \int_0^{l(n)} \frac{|\sin(\sqrt{\tau_n} + \frac{\eta}{\sqrt{\tau_n}})|}{(1+\eta/\tau_n)(1+2)} \cdot \frac{\eta}{\sqrt{\tau_n}} e^{-\eta/a} e^{-\eta^2/2\tau_n} \, d\eta \end{aligned}$$

Noting that $l(n) \to c$ and using dominated convergence theorem we get, $r^{1/2}|III| = O(\frac{1}{\sqrt{\log n}})$

Now we show that $r^{1/2}|IV| \to C > 0$ as $n \to \infty$. Again substituting $\xi = \sqrt{\tau_n + c} + \frac{\eta}{\sqrt{\tau_n + c}}$, we get

$$\begin{aligned} r^{1/2}|IV| &= r^{1/2} \cdot C\pi\epsilon^{1-1/a} \frac{n^{-1/2a}}{\log^{1+\frac{2}{2}(1-1/a)+\frac{1}{2a}}(n)} \left| \int_0^{\infty} \frac{\sin(\sqrt{\tau_n + c} + \frac{\eta}{\sqrt{\tau_n + c}})}{(1+\eta/\tau_n + c)(1+2)} \cdot e^{-\eta/a} e^{-\eta^2/2\tau_n} \, d\eta \right| \\ &\to \frac{C}{a} \int_0^{\infty} e^{-\eta/a} \, d\eta = C \end{aligned}$$

Hence, we get $r \cdot |PR(f_1 - PR(f_2))|^2 \to C > 0$ as $n \to \infty$.

Bibliography

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2006). Special Invited Lecture: Adapting to Unknown Sparsity by Controlling the False Discovery Rate. *The Annals of Statistics*, 34(2):584–653.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). 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*, 96(12):6745.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188.

Brown, P. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *nature genetics*, 21(1 Suppl):33–37.

Cai, T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994.

Donoho, D. and Johnstone, I. (1994a). Minimax risk over lp-balls for lq-error. *Rn*, 1:2.

Donoho, D., Johnstone, I., Hoch, J., and Stern, A. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):41–81.

Donoho, D. and Johnstone, J. (1994b). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al. (1999). Molecular classification of

cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531.

Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence. *Annals of statistics*, 36(1):381.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732.

Ingster, Y. (1998). Minimax Detection of a Signal for l^ n-Balls. *Mathematical Methods of Statistics*, 7(4):401–428.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: held at the Statistical Laboratory, University of California, June 20-July 30, 1960*, page 361. University of California Press.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal-Royal Statistical Society. Series B Statistical Methodology*, 70(3):461.

Jin, J. and Cai, T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.

Keselman, H., Cribbie, R., and Holland, B. (2002). Controlling the rate of Type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology*, 55(1):27–39.

Lander, E. (1999). Array of hope. *nature genetics*, 21(1):3–4.

Lee, M., Kuo, F., Whitmore, G., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):9834.

Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, pages 373–393.

Pan, W., Lin, J., and Le, C. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol*, 3(5):0022–1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.