# Estimating the Local False Discovery Rate via a Bootstrap Solution to the Reference Class Problem: Application to Genetic Association Data

Farnoosh Abbas Aghababazadeh

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa



[1]The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

# Abstract

Modern scientific technology such as microarrays, imaging devices, genome-wide association studies or social science surveys provide statisticians with hundreds or even thousands of tests to consider simultaneously. Testing many thousands of null hypotheses may increase the number of Type $I$ errors. In large-scale hypothesis testing, researchers can use different statistical techniques such as family-wise error rates, false discovery rates, permutation methods, local false discovery rate, where all available data usually should be analyzed together. In applications, the thousands of tests are related by a scientifically meaningful structure. Ignoring that structure can be misleading as it may increase the number of false positives and false negatives. As an example, in genome-wide association studies each test corresponds to a specific genetic marker. In such a case, the scientific structure for each genetic marker can be its minor allele frequency.

In this research, the local false discovery rate as a relevant statistical approach is considered to analyze the thousands of tests together. We present a model for multiple hypothesis testing when the scientific structure of each test is incorporated as a co-variate. The purpose of this model is to incorporate the co-variate to improve the performance of testing procedures. The method we consider has different estimates depending on the tuning parameter. We would like to estimate the optimal value of that parameter by considering observed statistics. Thus, among those estimators, the one which minimizes the estimated errors due to bias and to variance is chosen by

applying the bootstrap approach. Such an estimation method is called an adaptive reference class method. Under the combined reference class method, the effect of the co-variates is ignored and all null hypotheses should be analyzed together.

In this research, under some assumptions for the co-variates and the prior probabilities, the proposed adaptive reference class method shows smaller error than the combined reference class method in estimating the local false discovery rate, when the number of tests gets large. We describe the adaptive reference class method to the coronary artery disease data, and we use simulation data to evaluate the performance of the estimator associated with the adaptive reference class method.

# Acknowledgements

First and foremost I would like to express my special appreciation and thanks to my supervisors Dr. Mayer Alvo and Dr. David R. Bickel. I appreciate all their contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. My sincere thanks also goes to my thesis examining committee.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa. I take this opportunity to express gratitude to the Department of Mathematics and Statistics of University of Ottawa, Carleton University and Faculty of Graduate and Postdoctoral Studies which provided me the excellent academic environment. I also appreciate the financial support from University of Ottawa Admission Scholarship.

This research would not have been possible without using the `Biobase` and `locfdr` packages of R which facilitated the computational work. This study makes use of the data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of this data set is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

I would like to thank my beloved parents and my brother for their encouragement

and patience. My parents who provided unconditional love and care, I would not have made it this far without them.

Last but not the least, I would like also thank the members of Dr. Bickel's lab who have contributed immensely to my personal and professional time at the university of Ottawa. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful my wonderful and generous friends at the Department of Mathematics and Statistics, University of Ottawa. Thank you for providing support and friendship that I needed.

# Dedication

I would like to dedicate my thesis to my beloved parents and brother, ***Mina***, ***Fari-bourz*** and ***Farzam***, for their endless love, support and encouragement.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Overview

Modern scientific technology, specially in genomics, imaging and social science surveys, normally produce large-scale hypothesis testing problems with hundreds or even thousands of tests to consider simultaneously. As an example, quick progress in microarray technology allows researchers to measure the expression of thousands of genes simultaneously. Such development raises various questions such as, which genes expression levels are different between the patient and normal subjects. Also genome-wide association (GWA) studies represent a new approach to identify genetic markers that are associated with a particular disease or other traits. These two types of studies are explained in Chapter 2. In both microarray and GWA studies, hundreds or even thousands of tests are considered at the same time, where each hypothesis test corresponds to a test for a gene in microarray study or a genetic marker in GWA study. Solving such problems requires statistical techniques that appropriately control the probability of the occurrence of erroneous conclusions. We review some of those approaches in Chapter 3.

Testing multiple hypotheses is much more complicated than single hypothesis testing. The Type $I$ error rate is routinely controlled in a single-hypothesis test, whereas in testing multiple hypotheses controlling a compound error rate must be

considered. In large-scale hypothesis testing, it is ambiguous how to control compound errors. Historically, the family-wise error rate (FWER) [51] was first proposed to control compound errors, which is the probability of making at least one Type $I$ error among all the hypotheses. When the number of tests gets very large, the FWER becomes too strict for identifying non-null features (i.e. genetic markers, genes, voxels). Later, Benjamini and Hochberg in 1995 [6] introduced the false discovery rate (FDR), which is a more appropriate error measure to control the expected proportion of false discoveries among all the rejected hypotheses. Benjamini and Hochberg proposed a testing algorithm to control the FDR at level $q \in (0, 1)$ by considering the ordered observed $p$-values. For multiple comparisons, the $p$-value approaches prevent severely any false positives among thousands of tests, which may cause many non-null features (i.e. genetic markers, genes, voxels) to be hidden.

In 2001, Efron et al. [30] proposed a simple Bayesian framework and introduced the Bayesian false discovery rate (i.e. tail-area false discovery rate) to control compound errors. From the Bayesian view, each hypothesis test has an unknown prior probability. Then, the Bayesian false discovery rate is a posterior probability that the null hypothesis is true given the observed test statistic belongs to the rejection region. A simple empirical Bayes approach was developed to determine such unknown posterior probabilities. We review some frequentist and Bayesian approaches in Chapter 2. We also mention the connection between controlling the FDR and the empirical Bayes false discovery rate.

Efron et al. [30] introduced the local false discovery rate (LFDR), which is a natural extension of the tail-area Bayes false discovery rate. The LFDR is a posterior probability that the null hypothesis is true given the individual data for each test. It is more appropriate from the Bayesian view to consider the LFDR than the tail-area Bayes false discovery rate in order to identify the non-null features (i.e. genetic markers, genes, voxels). In such a case, the posterior probability is still unknown since it depends on some unknown quantities, such as the prior probability and both

the null and non-null densities of the test statistics. Unlike the $p$-value approaches, the LFDR estimate is easily explained as an approximate posterior probability that the null hypothesis is true. Under the different approaches, such quantities are estimated. The LFDR estimation does not suffer from the criticism of the $p$-value approaches. The estimation of the LFDR is not determined according to the choice of subjective or default prior distributions. In Chapter 4, we review both parametric and non-parametric methods for estimating the LFDR. My contribution to a paper on identifying genetic associations [84] was in designing, coding, and running some of the reported simulations under a parametric model for estimating the LFDR. Yang et al. considered two models for analyzing genetic association data: a parametric mixture model (PMM), and a semi-parametric mixture model (SMM). Such models are applied to estimate the LFDR. The application of these models to the coronary artery disease (CAD) data is described, and we use simulation data to evaluate the relative performance of each of the estimators associated with the two models.

The statistical approaches such as the FWER, controlling the FDR, and the LFDR, tend to assume that the thousands of tests should be analyzed together. Such an assumption may influence individual inferences. By considering the hypotheses together for some data sets, discovering the non-null features (i.e. genetic markers, genes, voxels) is more difficult, which is one of the main concerns in large-scale hypothesis testing. Some concerns related to the combination of hypothesis testing problems are reviewed in Chapter 5. As the motivating example in that chapter shows, analyzing all tests together is not an appropriate technique because of the structure of the data. Combining all tests has an effect on finding the number of false positives and false negatives. In applications, thousands of tests are related by a scientifically meaningful structure. As an example, in GWA studies each test corresponds to a specific genetic marker; in microarray study each test corresponds to a specific gene.

The principal goal of this research is to develop an estimate of the LFDR when the scientific structure of each test is incorporated as a co-variate. For example in

GWA studies, the minor allele frequency for each genetic marker may provide useful guidance for identifying that marker is disease-associated or not. Under the proposed model, the defined LFDR is the posterior probability that the null hypothesis is true given both the individual test statistic and the co-variate for each test. The *combined reference class* (CRC) method refers to the situation where the co-variates are not taken into account. In Chapter 6, we propose an adaptive reference class (ARC) method. Under the ARC method, some assumptions hold locally for each test which depends on a tuning parameter. For each test, changing the tuning parameter value yields different estimates for the LFDR. Among all those estimates, the one which minimizes the errors due to bias and to variance is chosen by using the bootstrap approach. Under some assumptions, the LFDR estimator under the ARC method indicates less error compared with the CRC method for large number of null hypotheses. We describe our application of these estimation methods to the coronary artery disease (CAD) data [82], and we use simulation data to evaluate the performances of the estimators associated with those estimation methods.

# Chapter 2

# Nature of Data

In this chapter, we introduce some concepts in genetics. Two important types of studies, *genome-wide association* and *microarray* studies are considered. The format of data in such type of studies is explained. At the end, we introduce some real data sets used throughout the other chapters.

## 2.1 Basic Concepts in Genetics

*Genetics* is the study of heredity, the process in which a parent passes certain genes on their children. The basic unit of genetic information is the *gene*. Genes determine specific traits, such as height, hair color, eye color and skin color. Natural talents, mental abilities, and the possibility of getting certain diseases are other characteristics affected by heredity. We have over 25,000 genes. Each gene contains instructions for a specific section or function of our body ([46], [39]).

Genes are made up of *deoxyribonucleic acid* (DNA). The hereditary material in humans and almost all other organisms is related to DNA. The information in DNA is stored as a code to tell our body how to grow and develop. Every cell of the human body contains our complete DNA genetic code. More than 90% of DNA is shared

by all people, even those not directly related. Less than 10% of our DNA is different and this makes each of us unique. DNA is made up of molecules called *nucleotides* which contain the *nitrogenous* bases. There are different nitrogenous bases called Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The genetic information is determined by the order of these bases. DNA bases pair up with each other, A with T, C with G, to form units called *base pairs*. In the DNA information, each word is a combination of three of these four chemical letters A, G, C and T ([46], [39]).

*Chromosomes* carry hereditary information. Each chromosome consists of two DNA chains running in opposite directions. Each chromosome has a centromere which divides the chromosome into two sections (or arms). The short arm of the chromosome is labeled the "p arm". The long arm of the chromosome is labeled the "q arm". The location of the centromere on each chromosome gives the chromosome its characteristic shape and can be used to identify the location of specific genes. Every species has its own characteristic number of different chromosomes. For every pair of chromosomes, one is inherited from the mother of an individual and one is inherited from the father of an individual. Humans have 23 pairs of chromosomes, 22 autosomes and one pair of sex chromosomes. The 22 autosomal chromosomes are numbered in order of decreasing length from 1 to 22 ([46], [39]).

The most common type of genetic variation among people are *single nucleotide polymorphisms*, frequently called SNPs. SNPs are single base-pair changes in the DNA sequence that happen with high frequency in the human genome. For example, a SNP may replace the nucleotide Guanine (G) with the nucleotide Adenine (A) in a certain stretch of DNA. Such variations are found in the DNA between genes. These variations can help scientists identify genes that are associated with disease ([46], [39]).

The chromosomes in a pair carry the same genes in the same locations, but there are different versions of the same genes. An *allele* is one of two or more types of a gene. An individual inherits two alleles for each gene, one from each parent. As an

example, the gene for eye colour has an allele for blue eye colour and an allele for brown eye colour. Some alleles are dominant over others. An allele that produces the same phenotype whether its paired allele is identical or different is named a *dominant allele*, whereas a *recessive allele* is an allele that produces its characteristic phenotype only when its paired allele is identical. Within a population there are two typically occurring base-pair possibilities for a SNP location. For each SNP, the frequency of the least common allele is named the *minor allele frequency* (MAF). As an example, a SNP with a minor allele (T) frequency of 30% indicates that 30% of a population has the allele $T$, whereas the more common allele is found in 70% of the population ([46], [39]).

At each location (except the sex chromosome), two genes constitute the individual's *genotype* at that location. *Linkage disequilibrium* occurs when genotypes at the two locations are not independent of another. When starting a genetic study, the main focus should be on identifying trait genetic variation influences. Thus the expression of the genotype is termed a *phenotype*. There are two basic classes of phenotypes: categorical (often binary case/control) or quantitative. Some examples of the former are hair color, eye color, and presence or absence of a disease, whereas examples of the latter are weight and height ([46], [39]).

A complete set of DNA of any organism is named a *genome*. Each genome contains all of the information needed to build and protect that organism. Determining the sequence of the human genome and identifying the genes that it contains are given by the Human Genome Project. This project has allowed researchers to learn more about the functions of genes and proteins. These studies have a major effect in the fields of medicine, biotechnology and the life sciences ([46], [39]).

A genetic disease or disorder is the result of changes in an individual's DNA sequence that make up a gene. Some rare genetic diseases such as Huntington's disease and cystic fibrosis are caused by changes in the DNA sequence of a single gene. Many common genetic diseases are caused by a combination of multiple genetic and

environmental factors. Some examples are Alzheimer's disease, asthma, Parkinson's disease, kidney diseases, cancer disease, diabetes, heart disease and mental illnesses ([46], [39], [41]).

## 2.2 Genome-wide Association Study

Genome-wide association (GWA) studies have burst over the last ten years into powerful tools to specify the genetic design [18]. The GWA studies aim to identify SNPs that are associated with a certain disease by scanning SNPs across the complete sets of DNA or genomes of many people. Each study can look at hundreds or thousands of SNPs at the same time. Researchers use data from this type of study to find SNPs that may contribute to a person's risk of a certain disease. The GWA studies are useful in finding the disease-associated SNPs by comparing genetic marker frequency differences between patient and healthy groups. Such studies are useful in determining SNPs that may contribute to a person's risk of developing common diseases with multifactorial reasons such as ashtma, cancer, diabetes, heart disease and mental illnesses.

Three developments have made GWA studies feasible and powerful [5]. First, genotyping has become more accurate and more affordable via the availability of genotyping chips which contain sets of thousands of SNPs across the human genome. Second, SNPs can now be selected on the basis of the linkage disequilibrium patterns observed across the human genome via the International HapMap resource [20]. Third, suitably large and well-specified clinical samples have been collected for many common diseases. Determining the disease-associated SNPs can potentially lead to new treatments and better disease diagnosis and prevention [82].

### 2.2.1 Case-control Study Design

Following Lewis [49], one of the widely used study designs for association studies is a *case-control* study. The most important problem in the case-control studies is making sure that any genetic difference between two groups of individuals, cases diagnosed as affected with disease and controls, known to be unaffected with disease, is related to the disease under study. In other words, cases and controls should be sampled from the same racial groups, the same geographical area or by checking the birth place of grandparents to ensure the distribution of cases is the same as the distribution of controls.

Under the case-control studies, the frequencies of SNP alleles are considered. The presence of a SNP allele may increase the risk of disease if the frequency of a SNP allele or genotype in cases is greater than in controls.

Suppose a single SNP with alleles $A$ and $a$ is tested in a case-control study. Let the number of cases and controls be denoted respectively as $n_{case}$ and $n_{cont}$. Let $n = n_{case} + n_{cont}$ be the total number of tested individuals. Two alleles can combine to give the genotypes; $AA$, $Aa$, and $aa$.

Table 2.1: Full genotype table for a general genetic model

| | $AA$ | $Aa$ | $aa$ | Total |
|---|---|---|---|---|
| Cases | $n_{\text{case}}^{AA}$ | $n_{\text{case}}^{Aa}$ | $n_{\text{case}}^{aa}$ | $n_{\text{case}}$ |
| Controls | $n_{\text{cont}}^{AA}$ | $n_{\text{cont}}^{Aa}$ | $n_{\text{cont}}^{aa}$ | $n_{\text{cont}}$ |
| Total | $n_{AA}$ | $n_{Aa}$ | $n_{aa}$ | $n$ |

Table 2.1 displays a contingency table for a case-control study without providing any ordering across the genotypes. To compare the frequency of genotypes in two groups of cases and controls, the chi-square statistic tests can be applied. Therefore, the observed value for genotype $aa$ in cases ($O = n_{\text{case}}^{aa}$) is compared with its expected value ($E = n_{case}n_{aa}/n$). Assume both cases and controls have the same frequency for each genotype (i.e. $AA$, $Aa$, and $aa$). Under the null hypothesis, the full test statistic

is

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_2,$$

where $O_i$ and $E_i$ represent the observed value and expected value of the $i$th cell. The data may also be analyzed by assuming a pre-specified genetic model.

A possible genetic model is additive. Under this model, carrying allele $a$ increases the disease risk. Thus, genotype $Aa$ has disease risk of $r$ and $2r$ for $aa$ genotype. Under this model, we may consider a logistic regression model for each SNP, where the response variable's outcome is binary, taking on the value 0 (i.e. control) or 1 (i.e. case) and the explanatory variable contains three levels of genotypes that are labelled by $\{0, 1, 2\}$. We explain in the following example how the genetic additive model is used for a specific genetic marker.

### Coronary Artery Disease Data

The Wellcome Trust Case Control Consortium (WTCCC) [82] brought together over 50 research groups from the UK that are active in researching the genetics of common human diseases, with expertise ranging from clinical; through genotyping; to design and analysis of GWA studies. Sampled individuals were living in England, Scotland, and Wales (Great Britain) and the vast majority had self-identified themselves as white Europeans. Coronary artery disease (CAD) is one of the common diseases of major public health importance in the UK. The control individuals came from two sources: 1,500 individuals from the 1958 British Birth Cohort (58C) and 1,500 individuals selected from blood donors recruited as part of this project (UK Blood Services (UKBS) controls). Therefore, there is a combined set of 3,000 controls and 2,000 cases. The total number of genotyped SNPs on 22 autosomal chromosomes is 500,568.

The WTCCC use some quality control filtering methods to exclude the SNPs based on the exact test of Hardy-Weinberg equilibrium, individuals or SNPs with too

much missing genotype. The total of $N = 455,568$ SNPs and 4,864 individuals (1,926 cases, 2,938 controls ) passed those quality control filters. We also excluded SNPs with MAF (see in Section 2.1) under 1 percent. A total of $N = 394,838$ SNPs were used in identifying disease-associated SNPs.

The CAD data is analyzed in the following chapters. In general, determining the number of disease-associated SNPs is the main problem. Therefore, in each chapter, a specific approach is applied to identify the disease-associated SNPs. The following example indicates the application of both the genetic additive model in Section 2.2.1 and logistic regression model for a single SNP from CAD data.

**Example 2.2.1** *From CAD [82], SNP rs*2980300 *is considered. Table 2.2 represents a contingency table for a case-control study. For SNP rs*2980300*, we fit the logistic regression model under the genetic additive model.*

Table 2.2: CAD data: SNP rs2980300

| | $CC$ | $CT$ | $TT$ | Total |
|---|---|---|---|---|
| Cases | 1381 | 516 | 29 | 1926 |
| Controls | 2088 | 801 | 49 | 2938 |

*From Table 2.2, under the genetic additive model in Section 2.2.1, the response variable $Y$ is defined to have the possible outcomes: person develops CAD by carrying allele $T$. These outcomes may be coded* 1 *and* 0 *respectively. The response variable is binary. The explanatory variable $X$ is assumed to be known constants $\{0, 1, 2\}$ (i.e. the genotypes $TT$, $CT$, and $CC$ are labeled as 2, 1, and 0 respectively).*

*The simple logistic regression model is applied. Let $Y_i$ for $i = 1, ..., 4864$ be independent Bernoulli random variables. Under the logistic regression model, we assume*

$$E(Y_i) = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)},$$

*where the unknown parameters $\beta_0$ and $\beta_1$ have to be estimated. The maximum likelihood approach estimates such of these parameters as*

$$\widehat{\beta}_0 = -0.4124 \ , \ \widehat{\beta}_1 = -0.0328.$$

*Thus, $\widehat{E(Y_i)}$ denoting the fitted value, is given by*

$$\widehat{E(Y_i)} = \frac{exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_i)}{1 + exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_i)}.$$

*Also we can express the fitted values as follows,*

$$\log_e \big(\frac{\widehat{E(Y_i)}}{1 - \widehat{E(Y_i)}}\big) = \widehat{\beta}_0 + \widehat{\beta}_1 X_i,$$

*that is the logarithm of the estimated odds. There exist different approaches to make inference about the logistic regression coefficients $\beta_0$ and $\beta_1$ [54].*

The interpretation of $\widehat{\beta}_1$ is not straightforward compared with the interpretation of the slope in a linear regression model. The reason is that the effect of a unit increase in $X$ varies for the logistic regression model. To see this effect, suppose the fitted value for $X_i$ is

$$\log_{\mathrm{e}}(\mathrm{odds}_1) \equiv \log_{\mathrm{e}} \big(\frac{\widehat{E(Y_i)}}{1 - \widehat{E(Y_i)}}\big) = \widehat{\beta}_0 + \widehat{\beta}_1 X_i,$$

while for $X_i + 1$, the fitted value is computed as follows,

$$\log_{\mathrm{e}}(\mathrm{odds}_2) \equiv \log_{\mathrm{e}} \big(\frac{\widehat{E(Y_i)}}{1 - \widehat{E(Y_i)}}\big) = \widehat{\beta}_0 + \widehat{\beta}_1 (X_i + 1).$$

Hence the difference between the two fitted values is

$$\log_e(\text{odds}_2) - \log_e(\text{odds}_1) = \log_e\left(\frac{\text{odds}_2}{\text{odds}_1}\right) = \widehat{\beta}_1.$$

Therefore, $\widehat{\beta}_1$ represents the logarithm of the odds ratio when $X_i$ increases by one unit [54].

## 2.3 Microarray Data Analysis

The same genes are not active in every cell. However, all of the cells in the human body contain identical genetic material. Scientists try to understand how these cells function normally and how they are affected when some genes do not perform properly by identifying which genes are active and which are inactive in different cells. In the past, scientists have been able to consider these genetic analyses on a few genes at once. Through the use of DNA microarrays, it is now possible to measure the expression levels of thousands of genes in a biological sample simultaneously. DNA microarrays have been used to record the gene expression during important biological processes such as the reaction to environmental changes. After the gene expression is measured, we wish to identify genes with differential expression under two experimental conditions, such as case and control groups, and then interpret the results. A main challenge to the development of statistical methods for microarray gene expression data is the fact that the sample size (e.g. the number of cases and controls groups) is usually small, whereas the number of measurements per gene is very large. As an example, we can mention the HIV data [77] discussed later involving $7,680$ genes measured from 8 individuals. Additionally, the expression levels for individual genes may not be independent. The absence of independence restricts the statistical models that can be used. The data set is a matrix $\mathbf{X} = [x_{ij}]$, where $x_{ij}$ represents the observed expression level for the $i^{\text{th}}$ gene on sample $j$. In the case-control study, the

sample is the combination of case and control groups [4].

### Prostate Data

Following Singh et al. [67], the prostate data with genetic expression levels measured on $N = 6,033$ genes were obtained from $n_1 = 52$ prostate cancer patients and $n_2 = 50$ normal controls. The total number of individuals is $n = 102$. The prostate data consist of a $6033 \times 102$ matrix of measurements $\mathbf{X} = [x_{ij}]$, where $x_{ij}$ represents the observed expression level for the $i^{\text{th}}$ gene on the $j^{\text{th}}$ individual. The prostate data is analyzed in Chapters 2 and 3 to identify genes with different expression levels between case and control groups.

### HIV Data

HIV data [77] discussed in [36] is an example of microarray data analysis with the total number of measured expression levels of $N = 7,680$ genes. Those genes were obtained from $n_1 = 4$ HIV-positive patients and $n_2 = 4$ healthy controls. The HIV data is a $7,680 \times 8$ matrix of measurements $\mathbf{X} = [x_{ij}]$, where $x_{ij}$ represents the observed logged expression level for the $i^{\text{th}}$ gene on the $j^{\text{th}}$ individual. The HIV data is analyzed in Chapter 3 to show how mistakenly choosing the null distribution directly influences the result.

This section introduced most of the data sets that are used in the following chapters. The Brain data described below is another interesting data set to be analyzed later.

### Brain Data

Schwartaman et al. [65] used an advanced MRI technology to measure water diffusion in the human brain by scanning the brain, called DTI. The DTI is used to map and characterize the three-dimensional diffusion of a water molecule randomly moving in

brain tissue to provide information about the direction of diffusion. The measured diffusivity, that is diffusion coefficient, relates diffusive flux to a concentration gradient [75] and has units of ($mm^2/s$). The data for this example is a DTI subset that is aimed finding the dyslexic-normal difference at the $i^{\text{th}}$ brain location related to reading development in children aged 713 [21]. In this study, twelve children $n = 12$ were tested, $n_1 = 6$ dyslexic and $n_2 = 6$ normal. Each child received Diffusion Tensor Imaging (DTI) brain scans in $N = 15,443$ locations which is represented by its own voxel's response. The Brain data in Chapter 4 is an appropriate example to demonstrate how the combination of tests in large-scale hypothesis testing can be misleading as it may increase the numbers of false positives and false negatives.

# Chapter 3

# Large-scale Hypothesis Testing

In this chapter, we introduce the large-scale hypothesis testing problem. There exist some frequentist and Bayesian approaches to this problem. We review both of these and indicate their connections. In the last section, we analyze the prostate and the CAD data sets mentioned in Chapter 2. The large-scale hypothesis testing problems are demonstrated by using those data sets.

## 3.1 Introduction

The basic pattern for single-hypothesis testing is defined as follows. A null hypothesis $H_0$ and non-null hypothesis $H_a$ are formulated and tested based on a test statistic $T$. An observed value $t$ is considered a realization of $T$. For a given rejection region $\mathcal{R}$, the null hypothesis $H_0$ is rejected when $T \in \mathcal{R}$. The null hypothesis $H_0$ is not rejected when $T \notin \mathcal{R}$.

The Type $I$ error, also known as a *false positive*, occurs when the null hypothesis $H_0$ is rejected, when the null hypothesis is really true. The Type $II$ error, also known as a *false negative*, occurs when the null hypothesis is not rejected, when the non-null

hypothesis $H_a$ is really true. For $\alpha \in (0,1)$, and a rejection region $\mathcal{R}_\alpha$,

$$\mathrm{P}_0(T \in \mathcal{R}_\alpha) = \alpha, \tag{3.1.1}$$

where $\mathrm{P}_0$ refers to the probability distribution of $T$ under the null hypothesis $H_0$. To choose the rejection region $\mathcal{R}_\alpha$, the acceptable Type $I$ error probability is fixed at some level $\alpha$. Then we consider all rejection regions that have Type $I$ error less than or equal to $\alpha$ and choose the one with the lowest type $II$ error probability. Thus the rejection region is chosen by controlling the Type $I$ error. By fixing the type $I$ error at level $\alpha$, a rejection region with optimal power is found.

The $p$-value $p(t)$ corresponding to the observed test statistic $t$ is given by

$$p(t) = \inf_{\alpha}\{t \in \mathcal{R}_\alpha\}. \tag{3.1.2}$$

When $p(t) \leq \alpha$, the null hypothesis $H_0$ is rejected. For any value of $u \in (0,1)$, the event $p(T) \leq u$ is equivalent to $T \in \mathcal{R}_u$, thus

$$\mathrm{P}_0(p(T) \leq u) = \mathrm{P}_0(T \in \mathcal{R}_u) = u,$$

which indicates $p(T)$ has a uniform distribution under the null hypothesis $H_0$. The $p$-value $p(t)$ is transformed to the $z$-value given by

$$z = \Phi^{-1}(p(t)), \tag{3.1.3}$$

when $\Phi$ denotes the cumulative distribution function (CDF) for standard normal. Thus any test statistic $T$ may be transformed to the statistic $Z$. The statistic $Z$ has a standard normal distribution under the null hypothesis,

$$\mathrm{P}_0\big(Z \leq z\big) = \mathrm{P}_0\big(\Phi^{-1}(p(T)) \leq z\big) = \mathrm{P}_0\big(p(T) \leq \Phi(z)\big) = \Phi(z). \tag{3.1.4}$$

In the multiple hypothesis testing problem, we have to consider thousands of hypothesis tests at once contrary to classical frequentist testing theory. Multiple hypothesis testing is concerned with testing several statistical hypothesis simultaneously, where the occurrence of erroneous conclusions may be increased. In multiple hypothesis testing, it is important to assess accurately whether discoveries (i.e. rejected hypotheses) are true or false. Any erroneous conclusions would result in either false or positive discoveries. In the multiple hypotheses testing problem, defining statistical significance, that is rejecting the null hypothesis, is a more complex problem.

The multiple-testing problem should be based on a statistical method that appropriately controls the probability of drawing erroneous conclusions, whereas the current statistical methods are limited in their ability to provide clear information about the probability that discoveries are true or false.

In 1995, Benjamini and Hochberg [6] introduced the false discovery rate (FDR) as a useful approach in multiple hypothesis testing. Storey ([69], [70]) improved such frequentist methods. In 2001, Efron et al. [32] proposed the empirical Bayes method which introduced Bayesian ideas without the strong Bayesian and frequentist assumptions. Storey and Tibshirani [72] and Storey [71] explained the connection between the frequentist FDR control procedure and the Bayesian paradigm. In the following sections, we review some frequentist and Bayesian approaches in testing multiple hypothesis to assess the probability that discoveries are true or false.

## 3.2 Frequentist Approach in Large-scale Hypothesis Testing

In the early 1960s, simultaneous hypothesis testing was a popular topic. Miller [51] generated some frequentist techniques that focused on controlling the overall Type $I$ error rate to test multiple hypotheses usually involving a small number of hypotheses.

For a single-hypothesis test, the Type $I$ error rate is normally controlled, while for multiple-hypothesis tests we control a compound error rate. As examples, we can mention family wise error rate (FWER) or controlling the false discovery rate (FDR).

Suppose $N$ hypothesis tests are considered. For each $i = 1, ..., N$, under the $i^{\text{th}}$ null hypothesis $H_{0i}$, the $p$-value $p_i$ is determined according to the test statistic $T_i$. The observed test statistic vector $\underline{t} = (t_1, ..., t_N)^T \in \mathbb{R}^N$ is considered a realization of the test statistic vector $\underline{T} = (T_1, ..., T_N)^T$. Following Efron [32], it is much more convenient to use $z$-values (3.1.3). The $z$-value vector $\underline{z} = (z_1, ..., z_N)^T$ is considered a realization of $\underline{Z} = (Z_1, ..., Z_N)^T$.

Suppose $\mathcal{R}$ is a decision rule that yields a decision for each of the $N$ hypotheses. Table 3.1 reports the various outcomes that occur when testing $N$ hypothesis based on applying a significant threshold $\alpha \in (0, 1)$ to their corresponding $p$-values.

Table 3.1: Outcomes when testing $N$ hypotheses

| Hypothesis | Accept | Reject | Total |
|---|---|---|---|
| Null true | $N_0^\star(\mathcal{R}) - N_0(\mathcal{R})$ | $N_0(\mathcal{R})$ | $N_0^\star(\mathcal{R})$ |
| Non-null true | $N_1^\star(\mathcal{R}) - N_1(\mathcal{R})$ | $N_1(\mathcal{R})$ | $N_1^\star(\mathcal{R})$ |
| Total | $N - N_+(\mathcal{R})$ | $N_+(\mathcal{R})$ | $N$ |

Let $N_0^\star(\mathcal{R})$ and $N_1^\star(\mathcal{R})$ be the total number of true null and non-null hypotheses respectively. Suppose $N_0(\mathcal{R})$ is the false positive results, and $N_1(\mathcal{R})$ is the false negative results. Here neither $N_0(\mathcal{R})$ nor $N_1(\mathcal{R})$ are observable, but $N_+(\mathcal{R})$ represents the total number of $t_i$ falling into $\mathcal{R}$ given by

$$N_+(\mathcal{R}) = N_0(\mathcal{R}) + N_1(\mathcal{R}),$$

which is observable. The family wise error rate was the first approach to control the entire Type $I$ error rate in the multiple hypothesis tests. It is the probability of making at least one Type $I$ error among all hypotheses.

**Definition 3.2.1** *The family wise error rate (FWER) is the probability of rejecting any true null hypothesis*

$$\begin{aligned} FWER(\mathcal{R}) &\equiv P\{Reject\ any\ true\ H_{0i}\} \\ &= P(N_0(\mathcal{R}) > 0). \end{aligned}$$

When testing multiple hypotheses, instead of controlling the probability of a Type $I$ error at level $\alpha$ for each test, the FWER is controlled at level $\alpha$. After controlling the FWER at level $\alpha$, a rejection region $\mathcal{R}$ is found where each test has the same rejection region. In order to control the FWER at some level $\alpha$, each hypothesis test needs to be controlled at lower levels. The following example indicates the Bonferroni procedure in controlling the FWER approach.

**Example 3.2.1** *The $i^{th}$ null hypothesis $H_{0i}$ is rejected if p-value $p_i$ satisfies the following equation,*

$$p_i \leq \frac{\alpha}{N} \ , \ \forall \alpha \in (0,1). \tag{3.2.1}$$

*Let $I_0$ denote the index of the true null hypothesis with length $N_0^\star(\mathcal{R})$. The FWER is controlled at level $\alpha$,*

$$\begin{aligned} FWER(\mathcal{R}) &\equiv P\{\cup_{i\in I_0}(p_i \leq \frac{\alpha}{N})\} \\ &\leq \sum_{i\in I_0} P\{p_i \leq \frac{\alpha}{N}\} \\ &= N_0^\star(\mathcal{R})\frac{\alpha}{N} \\ &\leq \alpha. \end{aligned}$$

Let $\text{FWER}_\alpha(\underline{t})$ indicate the FWER level $\alpha$ test based on the observed statistic $\underline{t}$. From (3.1.2), the general definition of *adjusted p-value* for the $i^{\text{th}}$ null hypothesis

$H_{0i}$ is given by

$$\widetilde{p}_i(\underline{t}) = \inf_{\alpha}\{H_{0i} \text{ rejected by } \text{FWER}_\alpha(\underline{t})\}. \tag{3.2.2}$$

As an example, the adjusted $p$-value in the Bonferroni approach is $\widetilde{p}_i = \min(Np_i, 1)$, where $p_i$ is the $p$-value defined in (3.1.2) for the $i^{\text{th}}$ null hypothesis $H_{0i}$.

When the number of tests is large, the FWER is too strict. Two examples are considered in Section 3.4 to show that by increasing the number of tests, the power to discover interesting results while controlling the FWER is extremely reduced.

In 1995, Benjamini and Hochberg [6] proposed the false discovery rate (FDR) to control the errors in testing multiple hypotheses. The FDR is designed to control the expected proportion of false discoveries among all the rejected hypotheses. It may be an appropriate approach to control the error rate in multiple-hypothesis tests compared with the traditional FWER approach. In general, when the number of hypothesis tests becomes large, controlling the FDR yields more positive discoveries compared with controlling the FWER ([6], [72]).

From Table 3.1, the false discovery proportion is an unobservable quantity given by,

$$\text{Fdp}(\mathcal{R}) = \begin{cases} \frac{N_0(\mathcal{R})}{N_+(\mathcal{R})} & \text{if } N_+(\mathcal{R}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.2.3}$$

Following Benjamini and Hochberg [6], the FDR of rejection region $\mathcal{R}$ is defined by,

$$\text{FDR}(\mathcal{R}) \equiv E\big(\frac{N_0(\mathcal{R})}{N_+(\mathcal{R})}|N_+(\mathcal{R}) > 0\big)\text{P}(N_+(\mathcal{R}) > 0). \tag{3.2.4}$$

That is the expected proportion of false rejections among all rejected hypotheses times the probability of making at least one rejection. In case all null hypotheses are true, the FDR is the same as the probability of making at least one error. In other words, when $N_0(\mathcal{R}) = N_+(\mathcal{R})$, if $N_0(\mathcal{R}) = 0$ then $\text{Fdp}(R) = 0$, and if $N_0(\mathcal{R}) > 0$ then $\text{Fdp}(R) = 1$, which means controlling the FDR controls the traditional FWER

approach. If many null hypotheses are rejected, then instead of controlling the FWER approach as the probability of making at least one error, it may be more appropriate to control the proportion of errors.

Benjamini and Hochberg [6] showed that when the test statistics are independent, the following algorithm controls the FDR at level $q \in (0, 1)$. Such an algorithm allows us to fix the error rate in advance and then estimate the rejection region, which is usually done in traditional multiple hypothesis testing.

### 3.2.1 Benjamini and Hochberg's Testing Algorithm

Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(N)}$ be the ordered observed $p$-values and choose $q \in (0, 1)$. Define

$$i_q = \max\{i : p_{(i)} \leq \frac{i}{N}q\}, \tag{3.2.5}$$

and reject the null hypothesis $H_{0(i)}$ corresponding to $p_{(i)}$ for $i \leq i_q$. Hence, the rejection rule is

$$\mathcal{R}_q = \{\text{Reject all } \ H_{0(i)} \ \text{ when } \ i \leq i_q\}. \tag{3.2.6}$$

If the $p$-values based on the null hypotheses are independent, then the Benjamini and Hochberg's testing algorithm controls the expected false discovery proportion at level $q$ [6],

$$\mathrm{E}\{\mathrm{Fdp}(\mathcal{R}_q)\} \leq q. \tag{3.2.7}$$

Benjamini and Yekutieli [9] showed that when the $p$-values based on the null hypotheses are both independent and continuous, then the inequality becomes an equality.

The FDR control is a less strict, more liberal rejector, and much more powerful compared with FWER ([6], [9], [69], [70]). Equation (3.2.7) indicates that controlling the FDR at level $q$ depends on the independence of the $p$-values which is usually an unrealistic assumption. In 2001, Benjamini and Yekutieli [9] proved under the positive regression dependency on each of the test statistics, the FDR approach is

still powerful compared with the FWER approach.

When $N_+(\mathcal{R}) = 0$, then the ratio $N_0(\mathcal{R})/N_+(\mathcal{R})$ is defined to be 0. To avoid this problem, Storey proposed to consider just $N_+(\mathcal{R}) > 0$ which was called the *positive false discovery rate* [70]. When controlling the FDR at level $q$, if the probability of discoveries is positive, then the FDR can always be made optionally small because of the extra term $\mathrm{P}(N_+(\mathcal{R}) > 0)$, and it is not the case for the positive false discovery rate. Whereas if $N_+(\mathcal{R}) = N_0(\mathcal{R})$, then all rejections are false discoveries and the positive false discovery rate is equal to 1 for any rejection region. In such situations, it is impossible to apply the other frequentist approach such as the FWER to control the multiple hypothesis testing. For Benjamini and Hochberg's algorithm, the sequence of $p$-values is considered. Such $p$-values are computed on the basis of an assumed null hypothesis distribution. In multiple hypothesis testing, the assumed null hypothesis distribution may be chosen mistakenly [23] (see in Section 4.3.1).

Given the above concerns with controlling the FDR, the question remains: is Benjamini and Hochberg's algorithm sufficient enough to make reliable decisions? The question is answered in the next section, by introducing a Bayesian approach. Such a method is connected to the frequentist FDR algorithm of Benjamini and Hochberg [6].

## 3.3 Bayesian Approach in Large-scale Hypothesis Testing

Efron et al. [32] proposed a simple Bayesian framework for multiple hypothesis testing called the *two-groups* model. Starting with a minimum of a priori assumptions, the posterior probability that the null hypothesis is true is considered. Let $A_i$ be an

indicator variable for the event that a non-null hypothesis $H_{ai}$ is true,

$$A_i = \begin{cases} 0 & \text{if } H_{0i} \text{ is true} \\ 1 & \text{if } H_{ai} \text{ is true} \end{cases} \tag{3.3.1}$$

Let $\{A_i\}$ be a sequence of independent and identically distributed indicator random variables with Bernoulli distribution, $A_i \overset{iid}{\sim} \text{Bernoulli}(1-\pi_0)$, where $\pi_0$ is the prior probability that the $i^{\text{th}}$ null hypothesis is true. Suppose under $H_{0i}$, the statistic $Z_i$ has density $f_0(z_i)$ and under $H_{ai}$, the statistic $Z_i$ has density $f_1(z_i)$,

$$\begin{aligned} &\text{under } H_{0i} : Z_i \sim f_0(z_i), \\ &\text{under } H_{ai} : Z_i \sim f_1(z_i). \end{aligned} \tag{3.3.2}$$

Both models in (3.3.1) and (3.3.2) are called the *two-groups* model. The statistic $Z_i$ has a mixture density $f(z_i;\pi_0)$ given by

$$f(z_i;\pi_0) = \pi_0 f_0(z_i) + (1-\pi_0) f_1(z_i). \tag{3.3.3}$$

Let $F_0(\mathcal{R})$ and $F_1(\mathcal{R})$ represent the CDF under null and non-null hypotheses respectively. For $z_i \in \mathcal{R}$,

$$F_0(\mathcal{R}) = \int_{\mathcal{R}} f_0(z_i)\, \mathrm{d}z_i \quad , \quad F_1(\mathcal{R}) = \int_{\mathcal{R}} f_1(z_i)\, \mathrm{d}z_i, \tag{3.3.4}$$

where $\mathcal{R}$ is any subset of the real line, and the mixture CDF is denoted $F(\mathcal{R})$ given by,

$$F(\mathcal{R}) = \pi_0 F_0(\mathcal{R}) + (1-\pi_0) F_1(\mathcal{R}). \tag{3.3.5}$$

**Definition 3.3.1** *The posterior probability that the $i^{th}$ null hypothesis is true given*

$z_i \in \mathcal{R}$ *is*

$$\psi(\mathcal{R}) \equiv P(A_i = 0 | z_i \in \mathcal{R}) = \frac{\pi_0 F_0(\mathcal{R})}{F(\mathcal{R})}$$

*which is called the Bayes false discovery rate (or tail-area false discovery rate) [32].*

If $\mathcal{R}$ is the rejection region, then the Bayes false discovery rate gives the probability of a false discovery. The Bayes false discovery rate contains three unknown quantities: the null prior probability $\pi_0$, the null density $f_0(z_i)$, and the non-null density $f_1(z_i)$. Efron et al. [32] proposed a simple empirical Bayes approach to estimate the Bayes false discovery rate which is called the *empirical Bayes false discovery rate*.

### 3.3.1 Empirical Bayes False Discovery Rate

Following Efron [32], the null density $f_0(z_i)$ is assumed to be standard normal. Such an assumption is called the *theoretical null* hypothesis. Suppose the null prior probability $\pi_0$ is known and close to 1, i.e. $\pi_0 \approx 1$. In such a case, in order to estimate the Bayes false discovery rate, it is sufficient to estimate the mixture CDF $F(\mathcal{R})$. The empirical approach is considered to estimate $F(\mathcal{R})$. Let $\bar{F}(\mathcal{R})$ denote the proportion of the observed $z$-values belonging to the region $\mathcal{R}$,

$$\overline{F}(\mathcal{R}) = \frac{\#\{z_i \in \mathcal{R}\}}{N}. \tag{3.3.6}$$

Substitute $\overline{F}(\mathcal{R})$ into the Bayes false discovery rate equation to get the empirical Bayes false discovery rate is denoted by $\overline{\psi}(\mathcal{R})$,

$$\overline{\psi}(\mathcal{R}) = \frac{\pi_0 F_0(\mathcal{R})}{\overline{F}(\mathcal{R})}. \tag{3.3.7}$$

When $N$ gets large, $\overline{F}(\mathcal{R})$ gets closer to $F(\mathcal{R})$, and the empirical Bayes false discovery rate $\overline{\psi}(\mathcal{R})$ may be a good approximation for $\psi(\mathcal{R})$. Following Storey [69] and Efron and Tibshirani [32], the connection between the frequentist controlling the FDR and

empirical Bayes false discovery rate is represented below.

Let the expectation of random variables $N_0(\mathcal{R})$ and $N_1(\mathcal{R})$ be denoted by $e_0(\mathcal{R})$ and $e_1(\mathcal{R})$ respectively,

$$e_0(\mathcal{R}) \equiv N\pi_0 F_0(\mathcal{R}) \ , \ e_1(\mathcal{R}) \equiv N(1-\pi_0)F_1(\mathcal{R}).$$

The expectation of an observable number of discoveries is denoted by $e_+(\mathcal{R})$,

$$e_+(\mathcal{R}) = e_0(\mathcal{R}) + e_1(\mathcal{R}).$$

For any rejection region $\mathcal{R}$, the Bayes false discovery rate $\psi(\mathcal{R})$, the empirical Bayes false discovery rate $\overline{\psi}(\mathcal{R})$, and the false discovery proportion $\mathrm{Fdp}(\mathcal{R})$ can be rewritten,

$$\psi(\mathcal{R}) = \frac{e_0(\mathcal{R})}{e_+(\mathcal{R})} \ , \ \overline{\psi}(\mathcal{R}) = \frac{e_0(\mathcal{R})}{N_+(\mathcal{R})} \ , \ \mathrm{Fdp}(\mathcal{R}) = \frac{N_0(\mathcal{R})}{N_+(\mathcal{R})}. \tag{3.3.8}$$

Following Efron [28], the accuracy of the empirical Bayes false discovery rate as an estimation of the Bayes false discovery rate is considered.

**Lemma 3.3.1** *Let $\gamma(\mathcal{R})$ denote the squared coefficient of variation of $N_+(\mathcal{R})$. Thus, $\frac{\overline{\psi}(\mathcal{R})}{\psi(\mathcal{R})}$ has approximate mean and variance respectively*

$$\Big(1+\gamma(\mathcal{R}), \gamma(\mathcal{R})\Big). \tag{3.3.9}$$

**Proof:** Apply the Delta method to approximate the mean and variance of the empirical Bayes false discovery rate

$$\begin{aligned}\overline{\psi}(\mathcal{R}) &= \frac{\mathrm{e}_0(\mathcal{R})}{N_+(\mathcal{R})} \\ &= \frac{\mathrm{e}_0(\mathcal{R})}{\mathrm{e}_+(\mathcal{R})} \frac{1}{1+\frac{N_+(\mathcal{R})-\mathrm{e}_+(\mathcal{R})}{\mathrm{e}_+(\mathcal{R})}}\end{aligned}$$

$$= \psi(\mathcal{R})\Big[1 - \frac{N_+(\mathcal{R}) - \mathrm{e}_+(\mathcal{R})}{\mathrm{e}_+(\mathcal{R})} + \Big(\frac{N_+(\mathcal{R}) - \mathrm{e}_+(\mathcal{R})}{\mathrm{e}_+(\mathcal{R})}\Big)^2\Big],$$

where $\frac{N_+(\mathcal{R})-\mathrm{e}_+(\mathcal{R})}{\mathrm{e}_+(\mathcal{R})}$ has mean zero and variance $\gamma(\mathcal{R})$ given

$$\gamma(\mathcal{R}) = \frac{\mathrm{Var}(N_+(\mathcal{R}))}{e_+^2(\mathcal{R})}.$$

∎

Lemma 3.3.1 demonstrates that the accuracy of $\overline{\psi}(\mathcal{R})$ as an estimate of $\psi(\mathcal{R})$ depends on the variance of $\mathrm{N}_+(\mathcal{R})$. When the variance of $\mathrm{N}_+(\mathcal{R})$ becomes small enough, the empirical Bayes false discovery rate may be an unbiased estimate of the Bayes false discovery rate with small variance.

Storey [69] and Efron and Tibshirani [32] showed the connection between the empirical Bayes false discovery rate and the frequentist method controlling the FDR proposed by Benjamini and Hochberg [6].

### Empirical Bayes Interpretation

Denote the $i^{\text{th}}$ ordered $z$-value by $z_{(i)}$. Then the ordered sequence of observed statistics is

$$z_{(1)} \leq z_{(2)} \leq ... \leq z_{(N)}. \tag{3.3.10}$$

For the $i^{\text{th}}$ ordered $z$-value, suppose the rejection region $\mathcal{R}$ contain $z$-values such that $z_j \leq z_{(i)}$, which means $\mathcal{R} = \{z_j : z_j \leq z_{(i)}, j = 1, ..., N\}$. Let $p_{(i)}$ indicate the left-tailed $p$-value which is $p_{(i)} = F_0(z_{(i)})$, and $1 - p_{(i)}$ corresponds to the right-tailed $p$-value. For given rejection region $\mathcal{R}$, the empirical cumulative distribution function of the $z_{(i)}$ satisfies

$$\overline{F}(\mathcal{R}) = \frac{\#\{z_i \in \mathcal{R}\}}{N} = \frac{i}{N}. \tag{3.3.11}$$

On the other hand, the Benjamini and Hochberg algorithm shows that,

$$p_{(i)} \leq \frac{i}{N}q \quad \Rightarrow \quad \frac{Np_{(i)}}{i} \leq q \quad \Rightarrow \quad \frac{F_0(\mathcal{R})}{\overline{F}(\mathcal{R})} \leq q,$$

which implies

$$\overline{\psi}(\mathcal{R}) \leq \pi_0 q, \tag{3.3.12}$$

where $\overline{\psi}(\mathcal{R})$ represents the empirical Bayes false discovery rate in (3.3.7).

Consequently, the Benjamini and Hochberg's testing algorithm depends on an estimated version of the Bayes false discovery rate in (3.2.4), where the cumulative distribution function $F(\mathcal{R})$ is replaced by the empirical cumulative distribution function $\bar{F}(\mathcal{R})$. Efron [28] showed that, when the $Z_i$'s are independent, the empirical Bayes false discovery rate controls the expected proportion of false discoveries. In the empirical approach to estimate the Bayes false discovery rate, there is no need to have the independence assumption for the test statistics.

### Poisson-independence Assumption

Suppose the test statistic $Z_i$ follows the two-groups model defined in (3.3.1) and (3.3.2) independently. From Table 3.1, the total number of false discoveries $N_0(\mathcal{R})$ and the total number of true discoveries $N_1(\mathcal{R})$ are distributed as

$$\begin{aligned} \mathrm{N}_0(\mathcal{R}) &\sim \text{Binomial}\Big(N, \pi_0 F_0(\mathcal{R})\Big), \\ \mathrm{N}_1(\mathcal{R}) &\sim \text{Binomial}\Big(N, (1-\pi_0) F_1(\mathcal{R})\Big). \end{aligned} \tag{3.3.13}$$

The squared coefficient of variation for the total number of discoveries $N_+(\mathcal{R})$ is

$$\begin{aligned}\gamma(\mathcal{R}) &= \frac{\mathrm{Var}\Big(N_+(\mathcal{R})\Big)}{e_+^2(\mathcal{R})} \\ &= \frac{1-F(\mathcal{R})}{NF(\mathcal{R})} \\ &= \frac{1-F(\mathcal{R})}{e_+(\mathcal{R})},\end{aligned} \tag{3.3.14}$$

where the rejection region $\mathcal{R}$ is chosen such that $F(\mathcal{R})$ is small, so $\gamma(\mathcal{R}) \approx \frac{1}{e_+(\mathcal{R})}$. The accuracy of estimator $\overline{\psi}(\mathcal{R})$ depends on the expected number of discoveries, $e_+(\mathcal{R})$, that is the expected number of $z$-values falling in $\mathcal{R}$. Therefore, by increasing $e_+(\mathcal{R})$, the squared coefficient $\gamma(\mathcal{R})$ gets closer to zero. From Lemma 3.3.1, the empirical Bayes false discovery rate $\overline{\psi}(\mathcal{R})$ is a reasonably accurate estimate for the Bayes false discovery rate $\psi(\mathcal{R})$. If we add $N \sim \text{Poisson}(\lambda)$ to the independence requirement, then

$$\begin{aligned}\mathrm{N}_0(\mathcal{R}) &\sim \text{Poisson}\Big(\lambda\pi_0 F_0(\mathcal{R})\Big), \\ \mathrm{N}_1(\mathcal{R}) &\sim \text{Poisson}\Big(\lambda(1-\pi_0)F_1(\mathcal{R})\Big).\end{aligned} \tag{3.3.15}$$

**Lemma 3.3.2** *Under the Poisson-independence assumptions in (3.3.13),*

$$E\Big(Fdp(\mathcal{R})\Big) = \psi(\mathcal{R})\Big[1 - exp(-e_+(\mathcal{R}))\Big]. \tag{3.3.16}$$

**Proof:** Under the Poisson-independence assumptions, $N_0(\mathcal{R})$ and $N_1(\mathcal{R})$ are independent Poisson random variables, so

$$N_+(\mathcal{R}) \sim \text{Poisson}\Big(\lambda F(\mathcal{R})\Big)$$

and

$$N_0(\mathcal{R})|N_+(\mathcal{R}) \sim \text{Binomial}\Big(N_+(\mathcal{R}), \psi(\mathcal{R})\Big).$$

Therefore, the probability of no discoveries occurring ($N_+(\mathcal{R}) = 0$), and at least one discovery occurring ($N_+(\mathcal{R}) > 0$) are computed respectively as

$$P\Big(N_+(\mathcal{R}) = 0\Big) = \exp\{-\mathrm{e}_+(\mathcal{R})\} \text{ and } P\Big(N_+(\mathcal{R}) > 0\Big) = 1 - \exp\{-\mathrm{e}_+(\mathcal{R})\}.$$

The expectation of the false discovery proportion is

$$\mathrm{E}\Big[\frac{N_0(\mathcal{R})}{N_+(\mathcal{R})}|N_+(\mathcal{R})\Big] = \psi(\mathcal{R}),$$

when $P\big(N_+(\mathcal{R}) > 0\big) = 1 - \exp\{-\mathrm{e}_+(\mathcal{R})\}$. ∎

Based on Lemma 3.3.2, the Bayes false discovery rate is expected to be close to the expectation of the false discovery proportion for large $e_+(\mathcal{R})$. In applications, the proportion of non-null features (i.e. SNPs, genes, voxels) is often close to zero which implies $e_+(\mathcal{R})$ is quite small. Then $\overline{\psi}(\mathcal{R})$ is a badly biased and highly variable estimator for $\psi(\mathcal{R})$. In such a case, instead of considering the empirical Bayes false discovery rate in (3.3.7), Efron [28] defined a new estimator for the Bayes false discovery rate is denoted by $\widetilde{\psi}(\mathcal{R})$,

$$\widetilde{\psi}(\mathcal{R}) = \frac{e_0(\mathcal{R})}{N_+(\mathcal{R}) + 1}. \tag{3.3.17}$$

**Lemma 3.3.3** *Under the Poisson-independence assumptions in (3.3.13),*

$$E\Big(\widetilde{\psi}(\mathcal{R})\Big) = E\Big(Fdp(\mathcal{R})\Big) = \psi(\mathcal{R})\Big[1 - exp(-e_+\{\mathcal{R}\})\Big]. \tag{3.3.18}$$

**Proof:** From Poisson-independence assumptions in (3.3.13) and (3.3.15), the number of discoveries is distributed $N_+(\mathcal{R}) \sim \text{Poisson}\Big(\lambda F(\mathcal{R})\Big)$, then

$$\begin{aligned}\mathrm{E}\Big(\widetilde{\psi}(\mathcal{R})\Big) &= \mathrm{E}\Big(\frac{e_0(\mathcal{R})}{N_+(\mathcal{R})+1}\Big)\\ &= e_0(\mathcal{R})\mathrm{E}\Big(\frac{1}{N_+(\mathcal{R})+1}\Big).\end{aligned}$$

Let $U = N_+(\mathcal{R})$, then

$$\begin{aligned}\mathrm{E}\Big(\frac{1}{U+1}\Big) &= \sum_{u=0}^{\infty}\frac{1}{u+1}\frac{e^{-\lambda F(\mathcal{R})}\Big(\lambda F(\mathcal{R})\Big)^u}{u!}\\ &= e^{-\lambda F(\mathcal{R})}\sum_{u=0}^{\infty}\frac{1}{u+1}\frac{\Big(\lambda F(\mathcal{R})\Big)^u}{u!}\\ &= \frac{e^{-\lambda F(\mathcal{R})}}{\lambda F(\mathcal{R})}\sum_{u=0}^{\infty}\frac{\Big(\lambda F(\mathcal{R})\Big)^{u+1}}{(u+1)!}\\ &= \frac{1-e^{-\lambda F(\mathcal{R})}}{\lambda F(\mathcal{R})}.\end{aligned}$$

Hence, the expectation of the modified estimator of the Bayes false discovery rate in (3.3.17) is

$$\begin{aligned}\mathrm{E}\Big(\widetilde{\phi}(\mathcal{R})\Big) &= e_0(\mathcal{R})\frac{1-e^{-\lambda F(\mathcal{R})}}{\lambda F(\mathcal{R})}\\ &= \lambda\pi_0 F_0(\mathcal{R})\frac{1-e^{-\lambda F(\mathcal{R})}}{\lambda F(\mathcal{R})}\\ &= \psi(\mathcal{R})\Big(1-e^{-\lambda F(\mathcal{R})}\Big)\\ &= \psi(\mathcal{R})\Big[1-\exp\{-\mathrm{e}_+(\mathcal{R})\}\Big].\end{aligned}$$

∎

Lemma 3.3.3 exhibits the modified empirical Bayes false discovery rate in (3.3.17)

which is unbiased for $\mathrm{E}\Big(\mathrm{Fdp}(\mathcal{R})\Big)$ but still is badly biased for the Bayes false discovery rate. When the expected number of discoveries $e_+(\mathcal{R})$ is not large enough, the usual situation in applications, the modified empirical Bayes false discovery rate in (3.3.17) is not an unbiased estimator for the Bayes false discovery rate in definition 3.3.1.

The Bayes false discovery rate $\psi(\mathcal{R})$ estimates depend on the prior probability $\pi_0$. Storey and Tibshirani [72] suggested a method which is called the *zero assumption* in order to estimate $\pi_0$. That approach considers some set of observed test statistics to estimate the prior probability $\pi_0$. Such a method provides a less conservative estimator compared with the empirical Bayes false discovery rate in (3.3.7).

### Zero assumption

Rather than setting $\pi_0$ equal to its upper bound 1, Storey and Tibshirani [72] used the collection of observed $z$-values to estimate $\pi_0$. Suppose $\mathcal{R}_0$ is a certain subset of $\mathcal{R}$ such that,

$$f_1(z_j) = 0 \ \ , \ \ \forall z_j \in \mathcal{R}_0, \tag{3.3.19}$$

that is all the non-null features (e.g. SNPs, genes, voxels) must give $z$-values outside of $\mathcal{R}_0$, is called the *zero assumption*. The expectation of the number of discoveries $N_+(\mathcal{R}_0)$ is

$$\mathrm{E}\Big(N_+(\mathcal{R}_0)\Big) = N\pi_0 F_0(\mathcal{R}_0).$$

Set

$$\widehat{\pi}_0 = \frac{N_+(\mathcal{R}_0)}{NF_0(\mathcal{R}_0)}, \tag{3.3.20}$$

where $\widehat{\pi}_0$ is an estimator of the prior probability $\pi_0$. To obtain another estimator $\widehat{\psi}(\mathcal{R})$ of the Bayes false discovery rate, substitute $\widehat{\pi}_0$ at (3.3.7),

$$\widehat{\psi}(\mathcal{R}) = \frac{\widehat{\pi}_0 F_0(\mathcal{R})}{\overline{F}(\mathcal{R})}. \tag{3.3.21}$$

Following Efron [28], the region $\mathcal{R}_0$ is selected where the null density $f_0(z_i)$ is standard normal and the subset of $\mathcal{R}_0$ is the central $\alpha_0$ proportion of the null density $f_0(z_i)$ given by

$$\mathcal{R}_0 = \Big[\Phi^{-1}(0.5 - \frac{\alpha_0}{2}), \Phi^{-1}(0.5 + \frac{\alpha_0}{2})\Big].$$

Using estimator $\widehat{\psi}(\mathcal{R})$ compared with the empirical Bayes false discovery rate $\overline{\psi}(\mathcal{R})$ indicates that the number of discoveries increase since $\widehat{\phi}(\mathcal{R})$ provides a larger rejection region. Storey et al. [71] proved that $\widehat{\psi}(\mathcal{R})$ is similar to the empirical Bayes false discovery rate and controls the false discovery proportion. However, if the tests are correlated with each other or if an appropriate null hypothesis distribution is not chosen [23], then the resulting estimates of the prior probability $\pi_0$ may be very biased [23].

Following the two-groups model, the estimate of the posterior probability that a rejected null hypothesis is actually true can be considered as the choice of $q$ in Section 3.2.1. For example, if $q = 0.1$, it means 10% of rejections are false discoveries, and 90% of rejections are true discoveries. The accuracy of the empirical Bayes false discovery rate as an estimator of the Bayes false discovery rate depends on $N$. From (3.3.12) and Lemma 3.3.2, it is concluded that the accuracy of the empirical Bayes false discovery rate is important to make reliable decisions by controlling the FDR approach.

## 3.4 Application

In this section, we consider two examples of large-scale hypothesis testing. We display the effect of the classical approaches and the reasons for moving toward improving the frequentist approaches and of applying the Bayesian approaches.

### 3.4.1 Microarray Data Example

We consider the prostate data in Section 2.3 to identify genes with different expression levels between the case and control groups.

Let $x_{ij}$ be a realization of $X_{ij}$ for $i = 1, ..., N$ and $j = 1, ..., n_1, n_1 + 1, ..., n$. For gene $i$, suppose $X_{i1}, ..., X_{in_1}$ is a random sample from $N(\mu_{i1}, \sigma_{i1}^2)$ and $X_{i(n_1+1)}, ..., X_{in}$ is an independent random sample from $N(\mu_{i2}, \sigma_{i2}^2)$. For gene $i$, the null hypothesis $H_{0i}$ is $X_{ij}$ has the same distribution for the normal and cancer patients,

$$H_{0i} : \mu_{i1} = \mu_{i2} \text{ vs. } H_{ai} : \mu_{i1} > \mu_{i2},$$

with the assumption of equal variances $\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma_i^2$. This example demonstrates the inference on comparing the means of two independent samples which are generated from normal distribution with equal variances. The two-sample $t$-test statistic is computed,

$$T_i = \frac{\bar{X}_{i1} - \bar{X}_{i2}}{\sqrt{S_i^2(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where $\bar{X}_{i1}$ and $\bar{X}_{i2}$ are the sample mean of $X_{ij}$ for case group and control group respectively. The $S_i^2$ is pooled sample variance,

$$S_i^2 = \frac{(n_1 - 1)S_{i1}^2 + (n_2 - 1)S_{i2}^2}{n_1 + n_2 - 2},$$

and

$$S_{i1}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(X_{ij} - \bar{X}_{i1})^2 \quad , \quad S_{i2}^2 = \frac{1}{n_2 - 1}\sum_{j=n_1+1}^{n}(X_{ij} - \bar{X}_{i2})^2.$$

Suppose $t_1, ..., t_N$ are realizations of $T_1, ..., T_N$ such that,

$$\text{under } H_{0i} : T_i \sim t_\nu,$$

with degrees of freedom $\nu = n_1 + n_2 - 2$. The null hypothesis $H_{0i}$ is rejected, if $t_i$ is

large,

$$p\text{-}value: \quad p_i = \mathrm{P}_0(T_i > t_i).$$

Following Efron [23], it may be more convenient to use $z$-values instead of $t$-values. Thus, we transform $t_i$ to $z_i$ by (3.1.3). The observed statistic $z_i$ is considered a realization of $Z_i$ where it has a standard normal distribution under the null hypothesis $H_{0i}$.

We would like to identify disease-associated genes. We consider a histogram of $z$-values. Under $H_{0i}$, the histogram should match the standard normal density curve. Figure 3.1 displays some non-null genes, since the histogram is too low near the center and too high in the tails.

By applying the traditional approach in multiple hypothesis testing, we changed the rejection level for each test from $\alpha = 0.05$ to $\alpha/N = 0.05/6033$ which denotes the Bonferroni bound. By applying the traditional approach, the total number of 397 genes satisfy

$$p_i \leq 0.05\ , \quad i = 1, ..., 6033,$$

that is the number of disease-associated genes, while this number is reduced to 4 by applying the Bonferroni bound as follows

$$p_i \leq \frac{0.05}{6033}\ , \quad i = 1, ..., 6033.$$

Figure 3.1 shows more interesting non-null genes. By considering Benjamini and Hochberg's testing algorithm, the number of disease-associated genes increases to 41, which seems more reasonable compared with the Bonferroni bound. Under the Bayesian false discovery rate, 46 genes are associated with disease.

### 3.4.2 Genome-wide Association Data Example

We consider the CAD data in Section 2.2 to identify the number of disease-associated SNPs. Under the additive genetic model in Section 2.2, we would like to investigate the association between the disease state and the $i^{\text{th}}$ SNP.

The logistic regression model estimates the $i^{\text{th}}$ SNP effect $\beta_i$, that is the logarithm of the odds ratio for the $i^{\text{th}}$ SNP, where $i = 1, ..., N$ with $N$ being the total number of measured SNPs. For the $i^{\text{th}}$ SNP, we perform the following hypothesis test,

$$H_{0i} : \beta_i = 0 \text{ vs. } H_{ai} : \beta_i > 0.$$

The reason of considering one-sided hypothesis test instead of two-sided one is explained in Section 4.3.

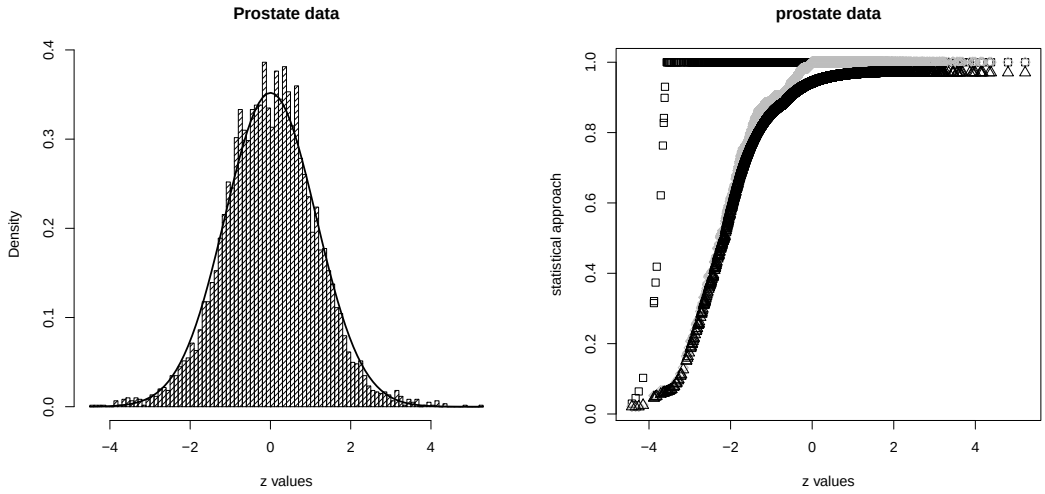



Figure 3.1: Left panel: histogram of $z$-values testing $6,033$ genes to discover association with prostate cancer. Right panel: statistical approaches versus $z$-values, Bonferroni (black square), Benjamini and Hochberg's testing algorithm (gray), and Bayesian false discovery rate (black triangle).

The Wald test statistic has an asymptotically standard normal distribution under the null hypothesis for large samples [43]. The square of this statistic has approxi-

mately a $\chi^2$ distribution with one degree of freedom,

$$T_i = \frac{\widehat{\beta}_i^2}{\widehat{\mathrm{Var}(\widehat{\beta}_i)}},$$

where $\widehat{\beta}_i$ is the maximum likelihood estimator of $\beta_i$ and $\widehat{\mathrm{Var}(\widehat{\beta}_i)}$ is the estimated variance of $\widehat{\beta}_i$. Let $t_1, ..., t_N$ be considered a realization of $T_1, ..., T_N$ such that

$$\text{under } H_{0i} : T_i \sim \chi_1^2,$$

then the null hypothesis $H_{0i}$ is rejected if $t_i$ is large,

$$p\text{-}value: \quad p_i = \mathrm{P}_0(T_i \geq t_i).$$

For the $i^{\text{th}}$ SNP, the observed $\chi^2$ test statistic $t_i$ is transformed to $z$-value by applying (3.1.3). Under the null hypothesis $H_{0i}$, a statistic $Z_i$ has a standard normal distribution and the usual one-sided hypothesis test is consider to test $H_{0i}$. Figure 3.2 represents the $z$-values for SNPs to identify which one is associated with disease. We consider a histogram of $z$-values. If all SNPs are null, then the histogram should match the standard normal density curve. By applying the traditional approach in multiple-hypothesis testing, we changed the rejection level for each test from $\alpha = 0.05$ to $\alpha/N = 0.05/394838$ which denotes the Bonferroni bound. By applying the traditional approach, 22675 SNPs are associated with disease since their $p$-values are less than 0.05 that is,

$$p_i \leq 0.05 \ , \quad i = 1, ..., 394838,$$

while this number is reduced to 26 by applying the Bonferroni bound since

$$p_i \leq \frac{0.05}{394838} \ , \quad i = 1, ..., 394838.$$

Figure 3.2 shows more interesting non-null SNPs. By considering Benjamini and Hochberg's testing algorithm, the number of disease-associated SNPs increases to 40, which seems more reasonable. Under the Bayesian false discovery rate, 43 SNPs are associated with disease.

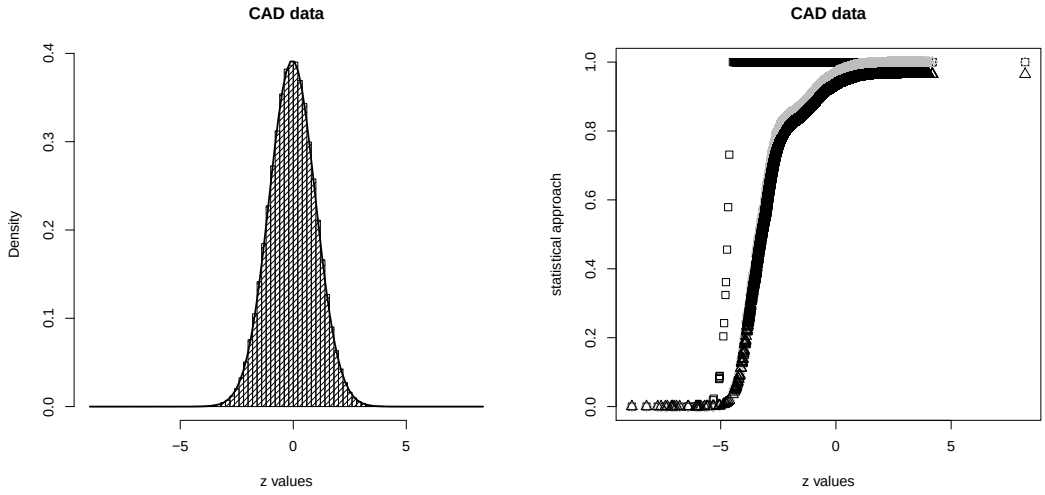



Figure 3.2: Left panel: histogram of $z$-values testing $394,838$ SNPs to discover association with CAD. Right panel: statistical approaches versus $z$-values, Bonferroni (black square), Benjamini and Hochberg's testing algorithm (gray), and Bayesian false discovery rate (black triangle).

# Chapter 4

# Local False Discovery Rate and its Estimate

In this chapter, we introduce the local false discovery rate (LFDR) to analyze large-scale hypothesis testing. We review parametric and non-parametric methods in estimating the LFDR. In applications, the LFDR is considered in identifying interesting results. As an example, in GWA data analysis, we apply the LFDR approach to identify disease-associated SNPs. In [84], Yang et al. considered two models of estimation in identifying the disease-associated SNPs. The LFDR estimate was determined by adapting a simple parametric mixture model (PMM) and a semi-parametric mixture model (SMM). In the simulation study, Yang et al. analyzed the performances of the two models under different values of a prior probability that is approximately equal to the proportion of SNPs that are associated with the specific disease. By considering a specific interval for prior probability, it is concluded that the PMM may be preferred since it has the advantage of supplying an estimate of detectability level of the non-associated SNPs.

## 4.1 Introduction

Efron et al. ([30], [32]) proposed the LFDR approach as an alternative to both Benjamini and Hochberg [6] and the tail area FDR. The advantage of the LFDR is, it provides an interpretation of results for individual features (i.e. SNPs, genes, voxels). The LFDR provides a measure of belief that depends on the exact value of the statistic not belonging to a large set of possible values. The statistical methods based on controlling the FDR may give misleading inference when specific features (i.e. SNPs, genes, voxels) are of interest. On the other hand, the research for multiple comparisons has so far focused on statistical methods based on controlling the FDR. Statistically, it is more difficult to estimate the LFDR than the Bayes false discovery rate in Section 3.3. We review the definition of the LFDR, some models used in estimating the LFDR, as well as characteristics of such estimators.

Suppose $N$ hypothesis tests are considered. For each $i = 1, ..., N$, under the $i^{\text{th}}$ null hypothesis $H_{0i}$, the test statistic $T_i$ is determined. The observed test statistic vector $\underline{t} = (t_1, ..., t_N)^T \in \mathbb{R}^N$ is considered a realization of $\underline{T} = (T_1, ..., T_N)^T$. The $z$-value vector $\underline{z} = (z_1, ..., z_N)^T$ is considered a realization of $\underline{Z} = (Z_1, ..., Z_N)^T$.

The *local false discovery rate* (LFDR) is the posterior probability that the $i^{\text{th}}$ null hypothesis is true given the data

$$\psi(z_i) \equiv \mathrm{P}(A_i = 0 | Z_i = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i; \pi_0)}, \tag{4.1.1}$$

where the prior probability $\pi_0$ is estimated from the data without considering any assigned prior knowledge at the beginning. There is a simple relationship between the Bayes false discovery rate in Section 3.3 and the LFDR [32]. From definition 3.3.1, the Bayes false discovery rate is the posterior probability that the $i^{\text{th}}$ null hypothesis is true given $z_i \in \mathcal{R}$,

$$\psi(\mathcal{R}) \equiv \mathrm{P}(A_i = 0 | z_i \in \mathcal{R}).$$

By averaging the LFDR over $z_i \in \mathcal{R}$, it is possible to compute the Bayes false discovery rate for the specified rejection region $\mathcal{R}$. For ease of notation $E(\psi(z_i))$ indicates the expectation with respect to the random variable $Z_i$.

**Lemma 4.1.1** *For a given rejection region $\mathcal{R}$, the Bayes false discovery rate is the average of the LFDR over $z_i \in \mathcal{R}$*

$$\psi(\mathcal{R}) = E\Big(\psi(z_i)|z_i \in \mathcal{R}\Big),$$

*where E indicates the expectation with respect to the distribution of $Z_i$.*

**Proof:** For the given rejection region $\mathcal{R}$, the Bayes false discovery rate in Section 3.3 is

$$\begin{aligned}
\psi(\mathcal{R}) &= \frac{\pi_0 F_0(\mathcal{R})}{F(\mathcal{R})} \\
&= \frac{\pi_0 \int_{z_i \in \mathcal{R}} f_0(z_i)\, \mathrm{d}z_i}{\int_{z_i \in \mathcal{R}} f(z_i; \pi_0)\, \mathrm{d}z_i} \\
&= \frac{\int_{z_i \in \mathcal{R}} \psi(z_i) f(z_i; \pi_0)\, \mathrm{d}z_i}{\int_{z_i \in \mathcal{R}} f(z_i; \pi_0)\, \mathrm{d}z_i} \\
&= \mathrm{E}\Big(\psi(z_i)|z_i \in \mathcal{R}\Big).
\end{aligned}$$

∎

A decision rule based on the LFDR provides a different set of discoveries compared with the set of discoveries obtained by controlling the FDR. As an example, based on the fixed threshold, the set of discoveries obtained by the LFDR is included in the set of discoveries attained by controlling the FDR. Therefore, the LFDR method is more conservative than controlling the FDR [3].

Typically, the LFDR estimate in (4.1.1) is used to find the rejection region.

The null hypothesis $H_{0i}$ is rejected if $\psi(z_i) < q$, for $q \in (0, 1)$. There is presently no consensus on a standard choice of $q$ for the Benjamini and Hochberg testing algorithm, as is the case for the Type $I$ error in single hypothesis testing. However, Bayesian computations suggest a cutoff threshold $\psi(z_i) \leq 0.2$ in testing the $i^{\text{th}}$ null hypothesis $H_{0i}$ [25]. The posterior odds ratio is given by,

$$\frac{P(A_i = 1|Z_i = z_i)}{P(A_i = 0|Z_i = z_i)} = \frac{1 - \psi(z_i)}{\psi(z_i)} = \frac{(1 - \pi_0)f_1(z_i)}{\pi_0 f_0(z_i)}. \tag{4.1.2}$$

When $\psi(z_i) \leq 0.2$, this posterior ratio is

$$\frac{1 - \psi(z_i)}{\psi(z_i)} \geq 4.$$

Practical applications of the large-scale hypothesis testing usually assume prior probability $\pi_0$ is large, i.e. $\pi_0 \geq 0.90$. Such an assumption identifies a small set of non-null hypotheses that are true. If we assume $\pi_0 \geq 0.9$ and $\psi(z_i) \leq 0.2$, then

$$\frac{f_1(z_i)}{f_0(z_i)} \geq 36,$$

which is in favor of non-null hypothesis $H_{ai}$. The following lemma indicates the expectation of the LFDR in (4.1.1) with respect to the mixture density $f(z_i; \pi_0)$.

**Lemma 4.1.2** *If* $A_i \overset{iid}{\sim}$ *Bernoulli*$\big(1 - \pi_0\big)$, *then* $E\big(\psi(z_i)\big) = \pi_0$.

**Proof:** If $A_i \overset{iid}{\sim}$ Bernoulli$\big(1 - \pi_0\big)$, then $E(A_i) = 1 - \pi_0$ and

$$E(A_i|Z_i = z_i) = \mathrm{P}(A_i = 1|Z_i = z_i) = 1 - \psi(z_i).$$

By taking the expectation with respect to the mixture density $Z_i$,

$$E\big(E(A_i|Z_i = z_i)\big) = E\big(1 - \psi(z_i)\big),$$

$$E(A_i) = 1 - E\big(\psi(z_i)\big),$$

which implies $E\big(\psi(z_i)\big) = \pi_0$. ∎

The LFDR is an unknown quantity that should be estimated. The prior probability $\pi_0$, null density $f_0(z_i)$, and non-null density $f_1(z_i)$ are unknown. The non-null density $f_1(z_i)$ plays a central role in assessing power. In the following sections, different approaches for estimating the LFDR are considered.

## 4.2 Estimate of Local False Discovery Rate

Different techniques were considered in estimating the LFDR to analyze either microarray gene expression and GWA data. The mixture model approach is a simple, flexible and accurate way to estimate the LFDR and the Bayes false discovery rate. This section provides a brief introduction to some parametric, semi-parametric or non-parametric estimation models.

Allison et al. [1], Pounds and Morris [58], Muralidharan [53], Liao et al. [50], and Pan et al. [56] used a parametric discrete mixture model in estimating the LFDR. Allison et al. [1] improved a sequence of procedures containing finite mixture modelling and bootstrap inference to determine the statistical significance of differences in gene expression levels in microarray experiments. The expression levels across genes are assumed to be independent. The density of $p$-values corresponding to the test statistic under the null hypothesis is uniformly distributed on the interval $(0, 1)$ and the non-null density of $p$-values tends to be closer to zero than one. Thus, the entire distribution of $p$-values can be modeled as a finite mixture of unknown Beta distributions. Using the independence assumption, the maximum likelihood approach was applied to estimate unknown parameters for a specified number of mixed Beta dis-

tributions. On the other hand, since the number of components in the finite mixture model is unknown, the bootstrap approach is applied for identifying the number of components. Allison et al. [1] mentioned some issues such as fitting such a model for the non-normal situation under the null hypothesis or even fitting with small sizes in clustering large-scale data.

Pounds and Morris [58] proposed the Beta-uniform mixture distribution for the mixture density, where the $p$-values are independent and identically distributed. Such an assumption is obviously untrue but it provided easy computations. Under the Beta-uniform mixture model, the density of $p$-values under the null hypothesis is uniform distribution on the interval $(0, 1)$, with the one-parameter Beta distribution as the non-null density. Under the independence assumption, the maximum likelihood approach was considered to estimate unknown parameters. It remains to determine how to compute $p$-values in an appropriate way to meet the assumptions. In such a case, non-parametric techniques are used to estimate the mixture density.

Liao et al. [50] proposed a special mixture model by considering the distribution of $p$-values such that the non-null density is stochastically smaller than the null density for all possible $p$-values. The density of $p$-values under the null hypothesis is a uniform on the interval $(0, 1)$, whereas the non-null density of $p$-values is assumed to be a one-parameter Beta distribution. The prior probability $\pi_0$ is the standard conjugate prior Beta(1,1). The Metropolis-Hasting algorithm is applied to estimate the mixture density.

Muralidharan [53] considered a mixture model to solve the Brown-Stein model used by Efron [29]. This model indicates the Bayesian hierarchical model. The statistic $Z_i|\delta_i \sim f_{\delta_i}(z_i)$ and $\delta_i \sim h(\bullet)$ where $f_{\delta_i}(z_i)$ belongs to an exponential family with natural parameter $\delta$. In other words, the mixture density $f(z_i)$ is a convolution of prior $h$ with the density $f_{\delta_i}(z_i)$. Given the prior $h$, the LFDR estimate can be calculated. However, a prior $h$ is not specified in advance. Instead, by applying an empirical Bayes approach, all observed statistics are used to estimate prior $h$ and

then this estimated prior is used to get the LFDR estimates.

Pan et al. [56] proposed a mixture model where both the null density and the mixture density are modeled to have a finite normal mixture density, that can be considered as a non-parametric estimator of a distribution function. Thus, the mixture model is typically fitted by the maximum likelihood method using the expectation-maximization algorithm. To specify the number of components for the proposed mixture model, the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) were suggested.

Aubert et al. [3], Efron ([23], [25]), Efron et al. [32], and Efron and Tibshirani [31] avoid the parametric estimates for mixture density. They considered semi-parametric or non-parametric methods in estimating the LFDR to analyze microarray gene expression data.

Efron et al. [32], Efron and Tibshirani [31] and Efron [23] proposed a non-parametric empirical Bayes approach to estimate the mixture density instead of considering a parametric approach. The test statistic (i.e. discrete Wilcoxon rank-sum statistic or two-sample $t$-test statistic) was applied to do the hypothesis test. They assumed $\pi_0 = 1$ in order to minimize the probability of detecting positive discoveries. Such a choice is conservative. To estimate the mixture density $f(z_i; \pi_0)$, the Poisson regression approach introduced in Section 4.3.1 is proposed.

Aubert et al. [2] proposed a kind of non-parametric technique to estimate LFDR for microarray data. They considered ordered $p$-values that are independently uniformly distributed under the null hypothesis. Without any assumption about the distribution of $p$-values under the non-null hypothesis, smoothing methods like a moving average or lowess fitting, are applied to estimate the LFDR. Aubert et al. [2] indicated that the non-parametric methods usually tend to have large standard errors and to be imprecise. The reason is that the mixture density of the test statistic is estimated by creating bins over the possible observed test statistics in non-parametric techniques.

In contrast with the large number of applications to microarray gene expression data analyses, only a few papers have applied the LFDR estimation to GWA data analysis (see in [17], [37], and [83]).

Greenwood et al. [37] considered the same approach as Efron [32]. They stratified genetic markers into three sets and the goal was to find the optimal selection of genetic markers as a rejection region among those sets. In general, the expected number of false positive results depends on how the set of data is chosen in estimating the LFDR. This means that by increasing the number of hypothesis tests, the number of false positive results will increase.

Bukszár et al. [17] introduced a precise method to estimate the LFDR by considering parametric techniques. They assumed the test statistics under the non-null hypothesis are identically distributed, when the distribution depends on one parameter. Usually the critical step is to estimate the non-null density. Instead of maximizing the log-likelihood function according to the mixture density, they considered the real maximum likelihood function to avoid a loss of precision.

All proposed models in both GWA and microarray data seem to perform reasonably well, though some models, based on a special case of the Beta distribution proposed by Allison et al. [1], are never used in the context of GWA studies. On the other hand, most of the proposed models considered the theoretical null hypothesis that is usually based on assumptions or asymptotic approximations. Efron [23] proposed *permutation* methods to avoid these approximations or assumptions. Permutation methods are essential to the estimation of the null density when considering some microarray data sets. The empirical null hypothesis proposed by Efron [23] is another approach to estimate $f_0(z_i)$ and it is introduced later in Section 4.3.1.

In the next section, two models are considered for analyzing GWA or microarray data: a semi-parametric mixture model (SMM) ([32], [31] and [23]), and a parametric mixture model (PMM) [84].

## 4.3 More Realistic Model

Following Yang et al. [84] and Bickel [55], we describe a more realistic model for analyzing both microarray and GWA data sets. Two estimation models are then considered to estimate the LFDR. Estimating the LFDR helps to identify the disease-associated SNPs in GWA data or the disease-associated genes in microarray data.

### Microarray Data Study

Following Bickel [55], $N$ null hypothesis tests are considered. Each test refers to a gene expression level. Genes are measured from $n$ samples. From Section (2.3), suppose $X_{i1}, ..., X_{in_1}$ is a random sample from $N(\mu_{i1}, \sigma_{i1}^2)$ and $X_{i(n_1+1)}, ..., X_{in}$ is an independent random sample from a $N(\mu_{i2}, \sigma_{i2}^2)$, with the assumption of equal variance $\sigma_{i1}^2 = \sigma_{i2}^2 = \sigma_i^2$. For the $i^{\text{th}}$ hypothesis test, suppose the parameter of interest is the absolute value of the inverse coefficient of variation denoted

$$\theta_i = \frac{|\mu_{i1} - \mu_{i2}|}{\sigma_i}.$$

Let the $i^{\text{th}}$ hypothesis be

$$H_{0i} : \theta_i = 0 \text{ vs. } H_{ai} : \theta_i > 0. \tag{4.3.1}$$

The test statistic $T_i$ is the absolute value of the two-sample $t$-test statistic and is distributed as the absolute value of a random variable from the non-central $t$ distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom and non-centrality parameter

$$\delta_i = (n_1^{-1} + n_2^{-1})^{-1/2}\theta_i.$$

### Genome-wide Association Data Study

Following Yang et al. [84], to investigate the association between the disease state and the $i^{\text{th}}$ SNP, logistic regression estimates the $i^{\text{th}}$ SNP effect $\beta_i$, that is the logarithm of the odds ratio for the $i^{\text{th}}$ SNP, where $i = 1, \ldots, N$ with $N$ being the total number of measured SNPs. For the $i^{\text{th}}$ hypothesis, we perform the Wald test statistic for,

$$H_{0i} : \beta_i = 0 \text{ vs. } H_{ai} : \beta_i > 0, \tag{4.3.2}$$

has an asymptotic standard normal distribution under the null hypothesis [43]. Hence the square $T_i = \widehat{\beta}_i^2 / \text{Var}\left(\widehat{\beta}_i\right)$, called the Wald chi-square test statistic has asymptotically a $\chi^2$ distribution with one degree of freedom under the null hypothesis. The $\widehat{\beta}_i$ is the maximum likelihood estimator of $\beta_i$ and $\widehat{\text{Var}}\left(\widehat{\beta}_i\right)$ is the standard estimate of the variance of $\widehat{\beta}_i$ . Since the Wald $\chi^2$ test statistics can never be negative, the one-sided hypothesis test in Section 3.4 is applied. For all $i = 1, \ldots, N$, the Wald $\chi^2$ test statistic $T_i$ is distributed according to the probability density function $\chi^2$ statistic with one degree of freedom and non-centrality parameter

$$\delta_i = \frac{\beta_i^2}{\text{Var}\left(\widehat{\beta}_i\right)}. \tag{4.3.3}$$

Equation (4.3.3) indicates that the value of $\delta_i$ depends on the parameter of interest $\beta_i$ such that $\delta_i = 0$ if $\beta_i = 0$ and $\delta_i > 0$ if $\beta_i > 0$. In other words, under $H_{0i}$ the test statistic $T_i$ is distributed as $\chi^2$ with one degree of freedom and non-centrality zero, whereas under the non-null hypothesis, it is distributed as $\chi^2$ with one degree of freedom and non-centrality parameter $\delta_i$.

By considering the GWA and microarray studies explained above, we define the

more realistic model [84]. Let the true association indicator be

$$a_i = \begin{cases} 0 & \text{if } \delta_i = 0 \\ 1 & \text{if } \delta_i > 0 \end{cases} \tag{4.3.4}$$

to indicate whether or not the $i^{\text{th}}$ feature is associated with a certain disease. The proportion of the non-associated features is $p_0 = \#\left(a_i = 0\right)/N$, where $\#\left(a_i = 0\right)$ is the number of true non-associated features. In order to estimate $a_i$, we will apply the two-groups model in Section 3.3 [32].

For the $i^{\text{th}}$ hypothesis test, let $A_i$ be an indicator variable of the event that a null hypothesis $H_{0i}$ is true. Suppose $A_1, A_2, ..., A_N$ are independent and identically distributed random variables with Bernoulli distribution, $A_i \overset{iid}{\sim} \text{Bernoulli}(1 - \pi_0)$, where $\pi_0$ is the prior probability that the $i^{\text{th}}$ null hypothesis is true. The following two mixture models will be used to estimate each $a_i$ in (4.3.4) by estimating the LFDR in (4.1.1). For ease of notation, $\widehat{a}_i$ will denote the estimator of $a_i$ for each of the estimating models.

### 4.3.1 Semi-parametric Mixture Model (SMM)

Following Efron et al. [32] and Efron [23], the LFDR in (4.1.1) may be estimated by applying an empirical Bayes approach that does not require any distributional assumption. If the $i^{\text{th}}$ null hypothesis is exactly true, then the statistic $Z_i$ will have a standard normal distribution, $f_0(z_i) \sim N(0, 1)$. The model defined in (3.3.3) with the following methods of estimation is called the *semi-parametric mixture model* (SMM). In estimating the mixture density $f(z_i; \pi_0)$ in (3.3.3), Efron [23] proposed a Poisson regression technique by applying Lindsey's method. He introduced estimation of the mixture density with smooth but flexible parametric models.

### Poisson Regression Approach

Suppose the mixture density $f(z_i;\pi_0)$ (or $f(z_i)$) belongs to the $J$-parameter exponential family,

$$f(z_i) = \exp\Big\{\sum_{j=0}^{J}\gamma_j z_i^j\Big\}, \tag{4.3.5}$$

where the constant $\gamma_0$ can be determined from $\gamma = (\gamma_1, ..., \gamma_J)$ such that $\int_{z_i} f(z_i)\,\mathrm{d}z_i = 1$. If $J = 2$, then the mixture density $f(z_i)$ defined in (4.3.5) is the same as the null density $f_0(z_i)$. The choice of $J$ needs to be greater than 2 in order to detect differences between the mixture density $f(z_i)$ and the null density $f_0(z_i)$. Efron proposed $J = 7$ which yields an exponential family in (4.3.5) between parametric and non-parametric modeling. To obtain the maximum likelihood estimates $\gamma$, Lindsey's method will be applied. Suppose the region $\mathcal{R}$ of $z$-values is partitioned into $K$ bins,

$$\mathcal{R} = \cup_{k=1}^{K}\mathcal{R}_k.$$

Define $y_k$ as the total number of $z$-values in the $k^{\text{th}}$ bin and $x_k$ is the center point of the $k^{\text{th}}$ bin $\mathcal{R}_k$. Lindsey's method assumes the independent $y_k$ are distributed as,

$$y_k \sim \text{Poisson}(\nu_k)$$

with mean $\nu_k = Nlf(x_k;\pi_0)$, where $l$ is the width of each bin. Thus, the mixture density $f(z_i)$ is estimated via a Poisson generalized linear model. The maximum likelihood estimate $\widehat{\gamma}$ maximizes the model

$$\log(\nu_k) = \sum_{j=0}^{J}\gamma_j y_k^j.$$

By Lindsey's algorithm, a smooth function $f_\gamma(y_k)$ may be fitted to the counts $y_k$. Even under dependency, $\hat{f}$ tends to be consistent and close to an unbiased estimate

for the mixture density $f(z_i)$.

Under the theoretical null hypothesis, after estimating the mixture density $f(z_i;\pi_0)$, it remains to estimate the prior probability $\pi_0$ to obtain the LFDR estimate in (4.1.1). Following Efron [31], the prior probability $\pi_0$ is known and close to its upper bound. Since the prior probability estimate $\widehat{\pi}_0$ is not a perfect estimate of $\pi_0$, the LFDR estimate can exceed one. The other option to improve the theoretical null estimate of the LFDR in (4.1.1) is to estimate the null prior probability $\pi_0$ by applying the zero assumption given in Section 3.3.1. In order to increase the accuracy of such an estimate, Efron [28] proposed another approach which is explained later.

Efron [23] showed the theoretical null hypothesis may fail for some practical reasons. In large-scale hypothesis testing, some mathematical assumptions such as independent and identically distributed assumption may not hold. As well when the sample size is small, the asymptotic theory may be in question. Also correlation among features (i.e. gene expression levels in micrroaray data or SNPs in GWA data) can make the theoretical null hypothesis a misleading choice ([23], [28]).

Following Efron [23], in such a case, the null density is assumed normally distributed, but with unknown mean and variance, $f_0(z_i) \sim N(\delta_0, \sigma_0^2)$, is called the *empirical null* hypothesis. Efron proposed either a central matching approach or a maximum likelihood approach for estimating the null density $f_0(z_i)$ and prior probability $\pi_0$. In both approaches, he supposed the non-null density $f_1(z_i)$ is zero around $z_i = 0$ (i.e. zero assumption in Section 3.3.1) to allow estimation of $f_0(z_i)$ and $\pi_0$ from the central histogram counts.

### Central Matching Estimation (CME) Approach

Suppose the LFDR in (4.1.1) is given by

$$\psi(z_i) = \frac{f_{\pi 0}(z_i)}{f(z_i;\pi_0)}, \tag{4.3.6}$$

where $f_{\pi 0}(z_i) = \pi_0 f_0(z_i)$. Under the zero assumption, the mixture density $f(z_i; \pi_0) = f_{\pi 0}(z_i)$ and

$$\log(f_{\pi 0}(z_i)) = \log\Big(\frac{\pi_0}{\sqrt{2\pi\sigma_0^2}}\Big) - \frac{1}{2}\Big(\frac{z_i - \delta_0}{\sigma_0}\Big)^2. \tag{4.3.7}$$

By considering the histogram counts the parameters $\pi_0$, $\delta_0$, and $\sigma_0$ can be estimated. If the zero assumption is true (not usually valid in actual applications), then the central matching approach yields an unbiased estimate for the null density. On the other hand, the zero assumption is more acceptable when $\pi_0$ is large, since the number of non-null hypotheses that are true contributing near $z_i = 0$ is decreased.

### Maximum Likelihood Estimation (MLE) Approach

The prior probability $\pi_0$ and the null density $f_0(z_i)$ are estimated by applying the zero assumption and the truncated normal model properties. Let $\mathcal{I}_0$, $N_0(\mathcal{R}_0)$, and $z_0$ be defined as

$$\begin{aligned} \mathcal{I}_0 &= \{j;\ z_j \in \mathcal{R}_0\} \\ N_0(\mathcal{R}_0) &= \#\{z_j \in \mathcal{R}_0\} \\ z_0 &= \{z_j;\ j \in \mathcal{I}_0\}. \end{aligned}$$

Under the zero assumption, $f_1(z_j) = 0$ for $z_j \in \mathcal{R}_0$. The null density is

$$\mathcal{Q}_{\delta_0,\sigma_0} \sim N(\delta_0, \sigma_0^2), \tag{4.3.8}$$

and the density of $Z_i$ conditional on $z_j \in \mathcal{R}_0$ is the truncated normal distribution

$$\mathrm{P}(z_i | z_j \in \mathcal{R}_0) = \frac{\mathcal{Q}_{\delta_0,\sigma_0}(z_i)}{\mathcal{H}_0(\delta_0, \sigma_0)}, \tag{4.3.9}$$

where $\mathcal{H}_0(\delta_0, \sigma_0)$ is

$$\mathcal{H}_0(\delta_0, \sigma_0) = \int_{\mathcal{R}_0} \mathcal{Q}_{\delta_0,\sigma_0}(z_i)\, \mathrm{d}z_i. \tag{4.3.10}$$

Let $\theta$ be defined as

$$\theta = \pi_0 \mathcal{H}_0(\delta_0, \sigma_0) = \mathrm{P}(Z_j \in \mathcal{R}_0). \tag{4.3.11}$$

Then $Z_j \in \mathcal{R}_0$ has density

$$f_{\delta_0,\sigma_0,\pi_0}(z_i) \propto \theta^{N_0(\mathcal{R}_0)}(1-\theta)^{N-N_0(\mathcal{R}_0)} \prod_{\mathcal{I}_0} \frac{\mathcal{Q}_{\delta_0,\sigma_0}(z_i)}{\mathcal{H}_0(\delta_0, \sigma_0)}, \tag{4.3.12}$$

where $f_{\delta_0,\sigma_0,\pi_0}(z_i)$ is the product of two exponential families that can be maximized separately to estimate parameters $\theta$, $\delta_0$, and $\sigma_0$. By maximizing the Binomial term

$$\hat{\theta} = \frac{N_0(\mathcal{R}_0)}{N},$$

and by maximizing the truncated normal family, the estimators $\delta_0$ and $\sigma_0$ can be determined. The estimate of $\pi_0$ is given by

$$\widehat{\pi}_0 = \frac{\hat{\theta}}{\mathcal{H}_0(\hat{\delta_0}, \hat{\sigma_0})}. \tag{4.3.13}$$

Following Efron [23], the HIV data shows that the theoretical null hypothesis is improper assumption for the null density. It also indicates how choosing the null distribution has an impact on the number of discoveries in the multiple hypothesis testing.

**Example 4.3.1** *The HIV data in Section 2.3 is considered to specify the differential gene expression levels between the case and control groups.*

*The two-sample t-test is applied to get z-values following equations in Section 3.4.1. Figure 4.1 shows the central histogram of z-values is less scattered than a theoretical null hypothesis* $f_0(z_i) \sim N(0, 1)$. *Under the empirical null hypothesis, the*

*null density estimate* $\hat{f}_0 \sim N(0.12, 0.75^2)$ *and the prior probability estimate* $\widehat{\pi}_0 = 0.934$ *can be observed by applying the maximum likelihood approach. However, the prior probability estimate under the theoretical null hypothesis is* $\widehat{\pi}_0 = 1.20$ *which seems more likely that there is something inappropriate about the theoretical null hypothesis. Using the empirical null hypothesis rather than the theoretical null hypothesis increases the number of disease-associated genes from 17 to 160.*

*From Figure 4.2, considering the zero assumption suggest using mainly null features. Thus, around the center,* $z = 0$*, the empirical estimation of LFDR numerator* $\hat{f}_{\pi_0}$ *(i.e. maximum likelihood approach) is approximately similar to the Poisson regression estimate of mixture density* $\hat{f}$*.*

By running the `locfdr` software package in R with the theoretical null hypothesis setting either empirical null setting, we used the non-parametric density estimation procedure (i.e. Poisson regression) to estimate $f(z_i; \pi_0)$ by $\widehat{f}$, to estimate $\pi_0$ by $\widehat{\pi}_0$ or estimate the null density $f_0(z_i)$ in the empirical null setting. The `locfdr` software package defaults to a maximum likelihood fitting for the LFDR estimation when the theoretical null hypothesis is not true. The LFDR estimate is denoted $\widehat{\psi}_{\mathrm{SMM}}(z_i)$ and computed by substituting $\widehat{\pi}_0$ and $\widehat{f}$ into $\pi_0$ and $f(z_i; \pi_0)$ (or $\hat{f}_0$ into $f_0(z_i)$ when the theoretical null hypothesis is not true) in (4.1.1). We choose the estimator $\widehat{a}_i = 1 - \widehat{\psi}_{\mathrm{SMM}}(z_i)$ to estimate the $a_i$ in (4.3.4).

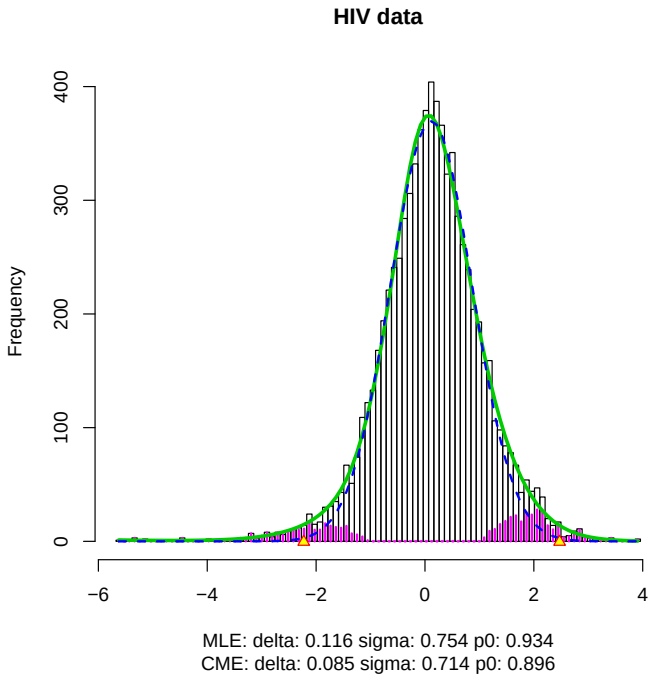



Figure 4.1: HIV data: z-values testing 7,680 genes to discover associated with HIV disease under the theoretical null hypothesis and empirical null hypothesis. Short vertical bars are estimated non-null counts and useful for power calculations.

Efron compared the variability of three estimators, $\widehat{\psi}_{\mathrm{SMM}}(z_i)$, $\widehat{\psi}(\mathcal{R})$ in (3.3.20) and $\overline{\psi}(\mathcal{R})$ in (3.3.7) for fixed prior probability $\pi_0 = 0.95$. The conclusion shows $\widehat{\psi}_{\mathrm{SMM}}(z_i)$ has the least variability and $\overline{\psi}(\mathcal{R})$ has the highest variability [28]. Nevertheless, the maximum likelihood approach gives smaller standard deviation compared with the central matching approach in estimating the null density $f_0(z_i)$ and the prior probability $\pi_0$, while the bias in estimating the null prior probability $\pi_0$ in the central matching approach is less than the maximum likelihood ([23], [28]).

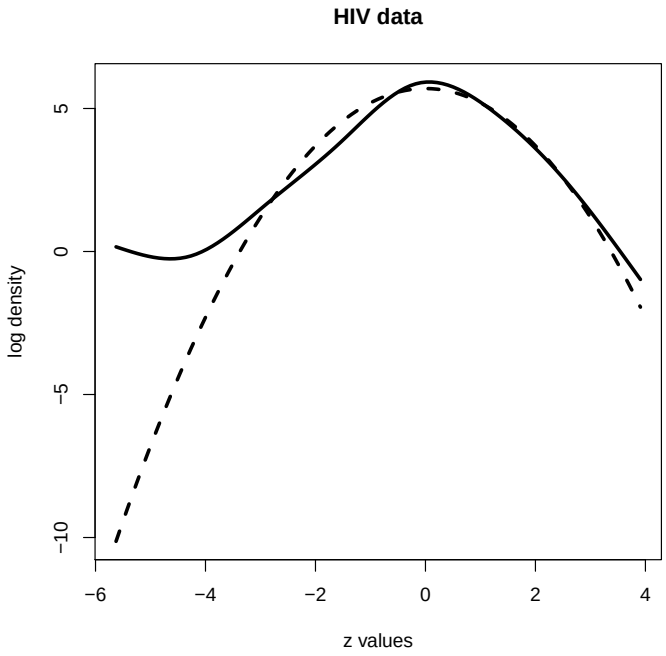



Figure 4.2: HIV data: Empirical estimation of the LFDR numerator $f_{\pi_0}(z_i)$. Heavy curve is log of Poisson regression estimate for mixture density $f(z_i;\pi_0)$, $\log \hat{f}$ and dashed cure is $\log \hat{f}_{\pi_0}$ by the maximum likelihood approach.

### 4.3.2 Parametric Mixture Model (PMM)

Yang et al. [84] proposed the parametric mixture model to improve the precision of the LFDR estimates in identifying the positive discovery results. For the $i^{\text{th}}$ hypothesis test, we perform a specific test, such as Pearson's statistic, logistic regression analysis, or two-sample $t$-test to obtain the observed test statistic $t_i$, where the non-null distribution is approximated by either a non-central chi-square, a non-central $F$, or a non-central $t$ distribution. As explained in Bukszár et al. [17], the critical step of the method consists of estimating the non-centrality parameters of the densities under the non-null hypothesis.

We begin with the Bayesian two-groups model. Suppose $N$ hypothesis tests are performed with test statistics $T_1, ..., T_N$. The probability density function of $T_i$ under

the null hypothesis $H_{0i}$ is $g_0(t_i)$ which is known. The probability density function of $T_i$ under the non-null hypothesis is $g_{\delta_i}(t_i)$ which depends on unknown parameter $\delta_i \in (0, \infty)$. The mixture density for $T_i$ is

$$g(t_i; \pi_0, \delta_i) = \pi_0 g_0(t_i) + (1 - \pi_0) g_{\delta_i}(t_i). \tag{4.3.14}$$

The model defined in (4.3.14) with the following method of estimation is called the *parametric mixture model* (PMM).

Let $T_1, ..., T_N$ be $N$ independent test statistics with mixture density function $g(t_i; \pi_0, \delta_i)$ in (4.3.14). The log-likelihood function is

$$l(\pi_0, \delta_1, ..., \delta_N) = \sum_{i=1}^{N} \log(\pi_0 g_0(t_i) + (1 - \pi_0) g_{\delta_i}(t_i)), \tag{4.3.15}$$

is not very useful since the number of parameters exceeds the number of null hypothesis tests. For this reason, it will be necessary to ease some of the assumptions ([17], [84]). Suppose the $T_i$'s are independent and identically distributed under the non-null hypothesis with $\delta_i = \delta$. In applications, Bukszár et al. [17] also assigned a single non-centrality parameter value to all SNPs that are not associated with disease. Similarly, for microarray data, a three-component mixture model performs as well as more complex models according to [53] with the three components corresponding to null, negative, and positive groups.

Under the i.i.d. assumption for the test statistics, the log-likelihood function now contains only two unknown parameters $\pi_0$ and $\delta$,

$$l(\pi_0, \delta) = \sum_{i=1}^{N} \log(\pi_0 g_0(t_i) + (1 - \pi_0) g_{\delta}(t_i)), \tag{4.3.16}$$

from which we may derive numerically the maximum likelihood estimates $\widehat{\pi}_0$ and $\widehat{\delta}$. Under the $i^{\text{th}}$ hypothesis test, $\widehat{a}_i = 1 - \widehat{\psi}_{\text{PMM}}(t_i)$ estimates $a_i$, where $\widehat{\psi}_{\text{PMM}}(t_i)$ is the

estimator of the LFDR $\psi(t_i)$ in (4.1.1),

$$\widehat{\psi}_{\text{PMM}}(t_i) = \frac{\widehat{\pi}_0 g_0(t_i)}{\widehat{\pi}_0 g_0(t_i) + (1 - \widehat{\pi}_0) g_{\widehat{\delta}}(t_i)},$$

and $g_{\widehat{\delta}}$ is an estimate of $g_\delta(t_i)$.

The likelihood ratio measures the degree to which the data support one hypothesized distribution over another. Since, the non-null hypothesis typically corresponds to an infinite number of probability density functions, a single density function must be selected for the numerator of the likelihood ratio. The full Bayesian solution is to assign priors in order to integrate the likelihoods associated with each hypothesis [81]. Under the PMM, the log-likelihood ratio for the $i^{\text{th}}$ hypothesis test is given by,

$$\Delta_i(t_i) = \log_2\left(\frac{g_\delta(t_i)}{g_0(t_i)}\right), \tag{4.3.17}$$

and is regarded as the ideal *information for discrimination* favoring the hypothesis of non-null ($\delta_i = \delta$) over that of null ($\delta_i = 0$) [48]. The information is called *ideal* because the $\delta$ is unknown. Here, we chose the binary logarithm ($\log_2$) to facilitate interpretation. Following the evidence levels that Bickel [11] considered, the discrimination information indicates strong evidence ($\Delta_i(t_i) > 3$), very strong evidence ($\Delta_i(t_i) > 5$), and overwhelming evidence ($\Delta_i(t_i) > 7$) [65]. Negative discrimination information, $\Delta_i(t_i) < 0$, indicates evidence in favor of the null hypothesis, which cannot be indicated by p-values since they can only quantify the evidence against null hypothesis [81].

## 4.4 Power Diagnostics in Semi-parametric Mixture Model

The FWER technique and the FDR theory both focus on a form of Type $I$ errors control. However, the LFDR method can be useful to assess power by considering the LFDR estimate. The importance of this section is on diagnostic statistics that are simple to calculate. As an example, the prostate data or the HIV data might easily fail to identify the disease-associated genes.

Under the two-groups model in Section 3.3, the non-null density $f_1(z_i)$ plays an important role in assessing power. Under the $i^{\text{th}}$ hypothesis test, from (3.3.3) and (4.1.1) and Lemma 4.1.2, the non-null density can be obtained as

$$f_1(z_i) = \frac{\big(1 - \psi(z_i)\big) f(z_i; \pi_0)}{1 - \pi_0}.$$

Let the LFDR and non-null density estimates be denoted by $\widehat{\psi}(z_i)$ and $\widehat{f}(z_i)$ respectively. The estimated non-null density is given,

$$\widehat{f}_1(z_i) = \frac{\big(1 - \widehat{\psi}(z_i)\big) \widehat{f}(z_i)}{1 - \widehat{\pi}_0},$$

where the estimated prior probability $\widehat{\pi}_0$ is,

$$\widehat{\pi}_0 = \int_{-\infty}^{\infty} \widehat{\psi}(z_j) \widehat{f}(z_j) dz_j.$$

The simple but useful diagnostic statistic for power may be obtained from the comparison of the estimated non-null density $\hat{f}_1(z_i)$ with estimated LFDR $\widehat{\psi}(z_i)$ [24]. The expectation of $\widehat{\psi}(z_i)$ under $\widehat{f}_1(z_i)$, say $\widehat{\psi^{\star}}$, provides a diagnostic statistic is given by

$$\widehat{\psi^{\star}} = \frac{\int_{-\infty}^{\infty} \widehat{\psi}(z_i)\big(1 - \widehat{\psi}(z_i)\big) \widehat{f}(z_i) dz_i}{1 - \int_{-\infty}^{\infty} \widehat{\psi}(z_j) \hat{f}(z_j) dz_j}. \tag{4.4.1}$$

A small value of $\widehat{\psi^\star}$ represents good power. By considering the non-null counts from the Poisson regression approach in estimating the mixture density, the estimate of the non-null expectation is obviously seen.

Suppose $N$ $z$-values are partitioned into $K$ bins of equal width $l$, where $y_k$ shows the total number of $z$-values in $k^{\text{th}}$ bin and $x_k$ is the midpoint of $k^{\text{th}}$ bin. From (4.1.1), the posterior probability that the $i^{\text{th}}$ non-null hypothesis is true given the data is,

$$P(A_i = 1|Z_i = z_i) = 1 - \psi(z_i).$$

An unbiased estimate of the non-null counts in the $k^{\text{th}}$ bin is denoted by $\hat{y}_{1k}$

$$\hat{y}_{1k} = \big(1 - \widehat{\psi}(x_k)\big)y_k.$$

Thus, an obvious estimate of the non-null expectation is

$$\widehat{\psi^\star} = \frac{\sum_{k=1}^{K} \widehat{\psi}(x_k)\hat{y}_{1k}}{\sum_{k=1}^{K} \hat{y}_{1k}}.$$

We considered the HIV data to compute the non-null expectation as a power diagnostic. Based on the empirical null hypothesis, the power diagnostic is $\widehat{\psi^\star} = 0.43$ which demonstrates low power in identifying the disease-associated genes in the list of discoveries. Also, we can have low power for the prostate data, which is indicated in the next Section. Such power diagnostics are computed from the observed data $z$-values, without any prior knowledge, which is one of the advantages of large-scale studies.

## 4.5 Application

### Prostate Data

The prostate data as microarray data described in Section 2.3 is considered to identify genes that are associated with the disease. The two-sample $t$-test is applied to get $z$-values following equation (3.1.3).

We computed the observed statistics $t_i$ for all genes. To perform SMM, we transformed the observed $t_i$ statistics into $z$-values. Figure 4.3 represents the central region of the histogram of the $z$-values; it matches the standard normal distribution very well, and hence the empirical null hypothesis was not needed. Under SMM, a total of 54 genes are associated with disease. Also the power diagnostic is $\widehat{\psi^\star} = 0.43$, which shows the low power in identifying the disease-associated genes.

Instead of considering the one-sided hypothesis tests, we consider the more realistic model in Section 4.3 to identify the number of disease-associated genes under PMM. Instead of working with $z$-values, we analyze the prostate data by considering observed $t_i$ statistics. Under PMM, the proportion of non-associated genes $\pi_0$ and non-centrality $\delta$ are estimated as $\widehat{\pi}_0 = 0.95$ and $\hat{\delta} = 2.33$ respectively. Here the total number of disease-associated genes is 58. Table 4.1 shows the number of genes with very strong or overwhelming evidence in favor of association and of those with evidence in favor of non-association.

Table 4.1: Number of genes with evidence in favor of association/non-association for the prostate data.

| | Negative | Very strong | Overwhelming |
|---|---|---|---|
| PMM | 4,920 | 101 | 35 |

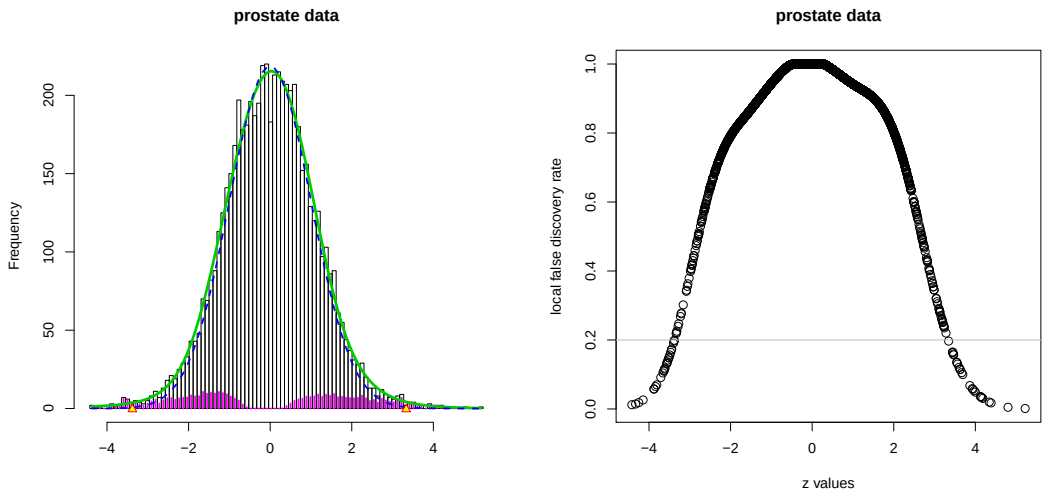



Figure 4.3: Left panel, histogram of $z$-values from the prostate data. The central peak of prostate data histogram closely follows the theoretical null hypothesis $f_0(z_i) \sim N(0,1)$. Short vertical Bars are estimated non-null counts and useful for power calculations. Right panel, the LFDR estimate versus $z$-values for all measured SNPs in the prostate data under SMM. The horizontal line represents the threshold 0.2.

### CAD Data

We now use the described CAD data in Section 2.2 to identify the disease-associated SNPs. We will consider PMM and SMM in estimating the LFDR in (4.1.1). We computed the Wald $\chi^2$ test statistics for all $N$ SNPs (see in Section 3.4.2). In order to implement SMM, we transformed the observed Wald test statistics into $z$-values. From Figure 4.4, the central region of the histogram of the $z$-values matches the standard normal distribution very well. The estimates of $p_0$ under SMM and PMM were denoted by $\widehat{\pi}_0$ (see in (3.3.3) and (4.3.14)). The SMM does not estimate $\delta$ since it estimates instead the non-null density $f_1(z_i)$ using the non-parametric method mentioned in Section 4.3.1.

Table 4.2: "Estimation of parameters $p_0$ and $\delta$ in SMM and PMM for the CAD data" [84]. The difference in numbers is related to using different quality control approaches explained in Section 2.2.

| Models | Estimated $p_0$ | Estimated $\delta$ |
|---|---|---|
| SMM | 0.98 | N/A |
| PMM | 0.94 | 1.17 |

Table 4.3: "Number of SNPs with evidence in favor of association/non-association" [84]. The difference in numbers is related to using different quality control approaches explained in Section 2.2.

| | Negative | Very strong | Overwhelming |
|---|---|---|---|
| PMM | 281,274 | 54 | 24 |

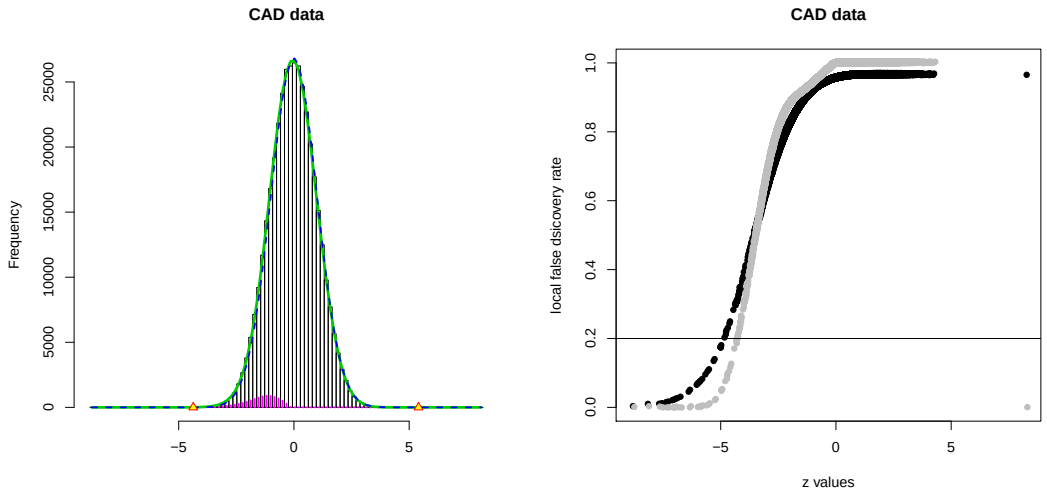



Figure 4.4: Left panel, histogram of $z$-values from the CAD data. The central peak of the CAD data histogram closely follows the theoretical null hypothesis $f_0(z_i) \sim N(0,1)$. Short vertical Bars are estimated non-null counts and useful for power calculations. Right panel, the LFDR estimate versus $z$-values for all measured SNPs in CAD data under PMM (black) and SMM (gray). The horizontal line represents the threshold 0.2. The difference in the number of disease-associated SNPs related to the using different quality control approach explained in Section 2.2.

Figure 4.4 shows the total number of SNPs associated with disease under SMM is 44, while this number under PMM is 31. Under SMM, the large value of estimate $\widehat{\psi^\star} = 0.93$ represents small power in identifying non-null SNPs for the CAD data. Table 4.3 reports the numbers of SNPs with very strong or overwhelming evidence in favor of association and of those with evidence in favor of non-association.

## 4.6 Simulation Study

My contribution to a paper [84] on identifying genetic associations was in designing, coding, and running some of the reported simulations under a parametric model for estimating the LFDR. The performances of the estimators in the two models (see in Sections 4.3.1 and 4.3.2) were compared . The abilities of the two mixture models, PMM and SMM, are compared to estimate $p_0$ and $a_i$. In this Section, $p_0$ denotes the true prior probability that a given SNP is not associated with the disease. Thus, in a particular simulated data set, the proportion of SNPs that are not associated with the disease tends to be very large.

According to (4.3.3), each $\delta_i$ depends on the logarithm of the odds ratio, and SNPs with different values of the odds ratio do not necessarily have the same value of $\delta_i$. In the simulation studies, the observed Wald $\chi^2$ test statistics for the disease-associated SNPs are generated from different distributions with different values of $\delta_{\text{alt}}$. For the non-associated SNPs, the observed Wald $\chi^2$ test statistics are from the same distribution with $\delta_i = 0$. Bukszár et al. [17] similarly simulated test statistics but used different estimators and different measures of performances than those employed herein.

Yang et al. executed 20 simulation studies, each with a different value of $p_1$. The $p_1$ is mentioned as a true prior probability that is approximately equal to the proportion of disease-associated SNPs. It is defined by $p_1 = 1 - p_0$. Among these 20 simulation studies, the values of $p_1$ range from 0.001% to 40%. In each simulation

study, 50 data sets are randomly generated, each corresponding to an artificial case-control study. For each data set, the total number of measured SNPs ($N$) is equal to 300,000. For each of the measured SNPs, a value that is generated from the uniform distribution between 0 and 1 is compared with $p_0$. If the generated value is greater than $p_0$, $a_i$ is assigned as 1, and otherwise assigned as 0. Then, if $a_i$ is equal to 1, the measured SNP is assumed to be associated with the disease. Then, the Wald $\chi^2$ test statistic for this measured SNP is sampled from a non-central $\chi^2$ distribution with one degree of freedom and an assigned non-centrality parameter. The value of this assigned non-centrality parameter $\delta_j$ is between 0.5 and 5. The Wald $\chi^2$ test statistics for the non-associated SNPs ($a_i = 0$) were sampled from a central $\chi^2$ distribution with one degree of freedom.

As the true values of $p_1$ in some simulation studies are extremely small (e.g. 0.001%), some constraints were used on the maximum likelihood estimation of $p_0$ and $\delta$ for the two parametric models. For PMM, the estimated values of $p_0$ were restricted to be between 0.5 and 1, and the estimated values of $\delta$ were restricted to be between 0 and $1.3 \times \max(t_i)$, where $\max(t_i)$ represents the maximum value of the observed Wald $\chi^2$ test statistics for a data set.

The two mixture models, namely PMM and SMM, are compared in terms of estimating $p_0$ and $a_i$. The bias for each estimator of $p_0$ is $\mathrm{E}[\widehat{p}_0 - p_0]$, which is estimated by averaging $\widehat{p}_0 - p_0$ over all 50 simulated data sets. Figure 4 in [84] illustrates the bias for each estimator of $p_0$ under SMM and PMM.

For the $i^{\text{th}}$ SNP at the $b^{\text{th}}$ simulation, the expected loss for the estimator of $a_{ib}$ ($\widehat{a}_{ib}$) is $R\left(a_{ib}, \widehat{a}_{ib}\right) = \mathrm{E}[l\left(a_{ib}, \widehat{a}_{ib}\right)]$, where $l\left(a_{ib}, \widehat{a}_{ib}\right)$ is the loss of $\widehat{a}_{ib}$. In this paper, Yang et al. consider the quadratic loss $l\left(a_{ib}, \widehat{a}_{ib}\right) = \left(\widehat{a}_{ib} - a_{ib}\right)^2$ to evaluate the performance of probability assessments [10]. The expected quadratic loss is called the quadratic risk or the *mean square error* (MSE). To assess the performances of PMM and SMM in estimating $a_{ib}$, Yang et al. also defined the MSE for non-associated SNPs ($\widehat{\mathrm{MSE}}_0$), the MSE for disease-associated SNPs ($\widehat{\mathrm{MSE}}_1$), and the MSE for all SNPs ($\widehat{\mathrm{MSE}}_{\mathrm{all}}$) as

follows:

$$\widehat{\mathrm{MSE}}_0 = \frac{\sum_{b=1}^{B}\sum_{i=1}^{N}\left(\widehat{a}_{ib} - a_{ib}\right)^2\left(1 - a_{ib}\right)}{\sum_{b=1}^{B}\sum_{i=1}^{N}\left(1 - a_{ib}\right)},$$

$$\widehat{\mathrm{MSE}}_1 = \frac{\sum_{b=1}^{B}\sum_{i=1}^{N}\left(\widehat{a}_{ib} - a_{ib}\right)^2 a_{ib}}{\sum_{b=1}^{B}\sum_{i=1}^{N} a_{ib}},$$

and

$$\widehat{\mathrm{MSE}}_{\mathrm{all}} = \frac{\sum_{b=1}^{B}\sum_{i=1}^{N}\left(\widehat{a}_{ib} - a_{ib}\right)^2}{N \times B}.$$

An advantage of the quadratic risk compared to Type $I$ and Type $II$ error rates is that it does not depend on setting hard significance thresholds that separate SNPs into associated and non-associated groups. Methods that perform well according to quadratic risk may be expected to also perform well for LFDR-estimate thresholds chosen subjectively according to the needs of each research group or according to conventional values set by the community. The relationship between quadratic risk and Type $I$ and Type $II$ error rates is explained at the end of this section.

As seen in Figures 5 and 6 of [84], the performances of SMM and PMM depend on the true values of $p_1$ and whether SNPs are associated with disease or not.

- Disease-associated SNPs: PMM has less quadratic risk compared with SMM when $p_1 < 0.1\%$ and when $p_1 > 10\%$. The quadratic risk for two models are equal when $p_1$ lies between 0.1% and 10%.

- Non-associated SNPs: SMM has less quadratic risk compared with PMM when $p_1 < 0.1\%$ and when $p_1 > 10\%$. The quadratic risk for two models are equal when $p_1$ lies between 0.1% and 10%.

- All SNPs: SMM has less quadratic risk when $p_1 < 0.1\%$; PMM has less quadratic risk when $p_1 > 10\%$. The quadratic risk for two models are equal when $p_1$ lies between 0.1% and 10%.

This separation of performance for non-associated SNPs from associated SNPs parallels the traditional separation of Type $I$ error rates from Type $II$ error rates, which depend on setting significance thresholds. The above observations illustrate this. The good performance of PMM relative to SMM for disease-associated SNPs correlates with lower Type $II$ error rates (higher statistical power and efficiency) for PMM, whereas the good performance of SMM relative to PMM for non-associated SNPs correlates with lower Type $I$ error rates (fewer false positives) for SMM. This conservatism of SMM accounts for the fact that it performs better than PMM overall when the number of disease-associated SNPs is extremely small. As discussed in the next section, that number may be larger than had been assumed.

## 4.7 Discussion and Conclusions

The simulation studies in Section 4.6 demonstrated that PMM is robust against model misspecification in the form of multiple non-centrality parameters values when $p_0$ is sufficiently low. Since the true value of $p_1$ is thought to range between $10^{-6}$ and $10^{-4}$ [82], approaches targeting such small values of $p_1$ are a topic of current investigation. Preliminary results indicate that by restricting estimates of $p_1$ to that range, a method of constrained maximum likelihood estimation [15] could, if the true value is within the range, substantially improves estimation of the LFDR. Yang et al. found that the proportion of SNPs with overwhelming evidence of association with CAD is about $10^{-4}$, which is at the upper end of the range currently considered plausible (see in Table 4.3). On the other hand, that upper bound is questionable in light of recent results that indicate that thousands of small-effect SNPs may be

associated with each particular disease [35, 57]. That hypothesis is more consistent with estimates of $p_0$ (see in Table 4.2) and with some of the settings of simulation studies (see in Section 4.6).

The paper [84] provided some insights on the performances of three models under the working hypothesis that the true value of $p_1$ is as small as many currently think. Yang et al. also conducted a thorough analysis on the performances of the three models under higher ranges of $p_1$ since some of those values could be closer to biological reality, as discussed above. According to Figure 6 [84], PMM generally performs better than SMM when $p_1 > 10\%$; SMM outperforms PMM when $p_1 < 0.1\%$; both models are about the same in performance when $p_1$ lies between 0.1% and 10%.

# Chapter 5

# Combining or Separating Tests in Large-scale Hypothesis Testing

This chapter considers the problem of combining tests in large-scale hypothesis testing. We start with a motivating example which illustrates the problem of combined tests in the large-scale hypothesis testing. We review Efron's approach [27] to solve such a problem, and we apply his model on real data to compare the results of combining analysis with separating analysis. This chapter constitutes a main focus of this thesis.

## 5.1 Introduction

Different statistical approaches have been used in large-scale hypothesis testing, such as the FWER in Section 3.2, controlling the FDR in Section 3.3, permutation methods [23], and LFDR in Section 4.1. As mentioned before, testing many thousands, even millions of null hypotheses may worsen the tradeoff between power and Type $I$ error. In such a case, it is more difficult to identify the interesting non-null features (i.e. SNPs, genes, voxels). To solve this issue, controlling the FDR [6], the Bayesian false

discovery rate [32], and the LFDR [32] were introduced. In such approaches, all available tests are usually assumed to be analyzed together. Combining all tests can be misleading as it may yield wrong inference for each test.

On the other hand, hypothesis tests are connected by a scientifically meaningful structure. As an example, in GWA data, each test corresponds to a specific SNP or in microarray data, each test corresponds to a specific gene. Incorporating scientific information as a co-variate may improve the performance of testing procedures.

In multiple hypothesis testing, the hypotheses usually contain class information based on scientific structure related to each hypothesis. Incorporating such scientific structure may increase statistical power in testing multiple hypotheses. Thus the hypotheses can be divided into sub-classes based on the characteristics of the problem. Ignorance of such class structure in data analysis may increase the number of false positives and false negatives.

There have been ways which incorporates the scientific structure into the variety of statistical techniques to handdle multiple hypothesis testing. Some researchers used the idea of incorporating class structure and weights to improve the statistical power. Benjamini and Hochberg [7] used $p$-value weighting method and evaluated different procedures. In 2006, Genovese et al. [34] showed that using $p$-value weighting procedure controls the FWER and FDR while increasing statistical power. Later Wasserman and Roeder introduced an optimal $p$-value weighting procedure for FWER control [79]. Hu et al. [44] proposed a weighting scheme based on a simple Bayesian framework. The proportion of null hypotheses that are true was used within each class. Such an approach controls the FDR for both the independent hypotheses and $p$-values with certain dependence structures. The unknown proportion of true null is estimated within each class. Thus the $p$-value weighting procedure asymptotically controls the FDR for $p$-values under the weak dependence. Efron [29] considered the separate-class model explained in Chapter 5 where the hypotheses were divided into distinct classes. Efron used a simple Bayesian theory to show the advantage of such

separate analysis for FDR methods. The question of separating multiple hypothesis testing problems has not received much recent attention. Few results have been published so far on proper $p$-value weighting approaches in order to control FDR. In many genetic studies, there is usually a natural stratification of the $N$ hypotheses to be tested. Given the FDR framework and the presence of such stratification, Sun et al. [74] considered the estimate of FDR for each stratum, is called *stratified FDR*, and compare it to the FDR estimate for all hypotheses in a single stratum, is called *aggregated FDR*. Sun et al. demonstrated that the aggregated FDR is a weighted average of the stratum FDRs.

We indicate that combining or separating hypothesis tests may affect the number of discoveries. The motivating example below will show that considering all $N$ tests in analyzing large-scale data can distort inference corresponding to each test. We review Efron's approach to indicate how the separate and combined analyses are connected [27].

### Brain Data Example

For the Brain data described in Chapter 1, each test corresponds to a specific brain location. The two-sample $t$-test yields $z$-values at $N = 15,443$ voxels (three-dimensional brain locations). For the $i^{\text{th}}$ voxel, the statistic $z_i$ has been considered to identify the difference between dyslexic and normal children.

Figure 5.1 shows the central region of histogram of the $z$-values matches the standard normal distribution, $f_0(z_i) \sim N(0, 1)$. Thus, the empirical null hypothesis was not needed. If all voxels are null, which means there are no differences between the voxels of dyslexic and normal children, then the histogram should match the standard normal density. Figure 5.1 illustrates the dyslexic-normal differences at some brain locations.

We use SMM from Section 4.3.1 to identify the dyslexic-normal difference at

each brain location. Figure 5.2 shows 184 voxels with $\widehat{\psi}_{SMM} \leq 0.2$. The estimated proportion of no dyslexic-normal difference is $\widehat{\pi}_0 = 0.93$, whereas the low value of power diagnostic estimate $\widehat{\psi^\star} = 0.47$ suggests failure to identify the the dyslexic-normal differences.

Brain data records the distance $x$ from the back toward the front of the brain. The $z$-values for the $N = 15,443$ voxels have been divided into two separated classes. The first class contain $z$-values for front-half (if $x \geq 50$), and the other contain $z$-values for back-half (if $x < 50$). There are $7,661$ voxels in back and $7,782$ in front (see in Figures 5.2 and 5.3). We are interested in identifying the number of discoveries according to these specified classes separately, and then compare these results with the total number of discoveries when all N=15,443 voxels considered together.

Under SMM, by assuming theoretical null hypothesis Figure 5.3 shows front-half data gave 271 voxels with $\widehat{\phi}_{\text{SMM}} \leq 0.2$, but none for the back-half data. The front-half data study indicates the estimated proportion of no dyslexic-normal difference is $\widehat{\pi}_0 = 0.91$ with better power diagnostic value $\widehat{\psi^\star} = 0.37$. Figures 5.2 and 5.3 display the importance of combination and separation of large-scale hypothesis testing. The combined analysis has a direct effect on the discoveries.

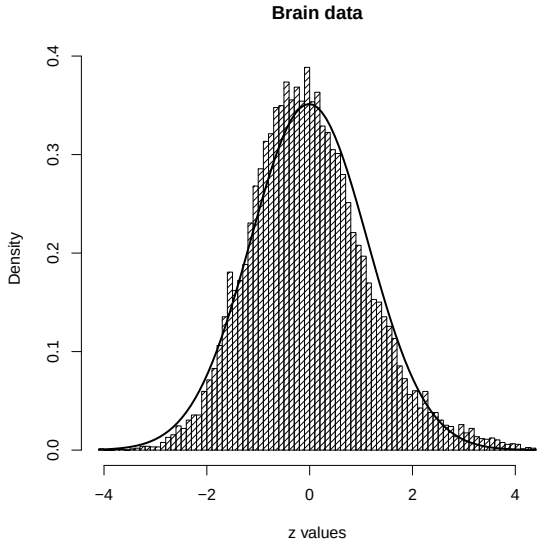



Figure 5.1: Brain data: histogram of all $N = 15,433$ $z$-values.

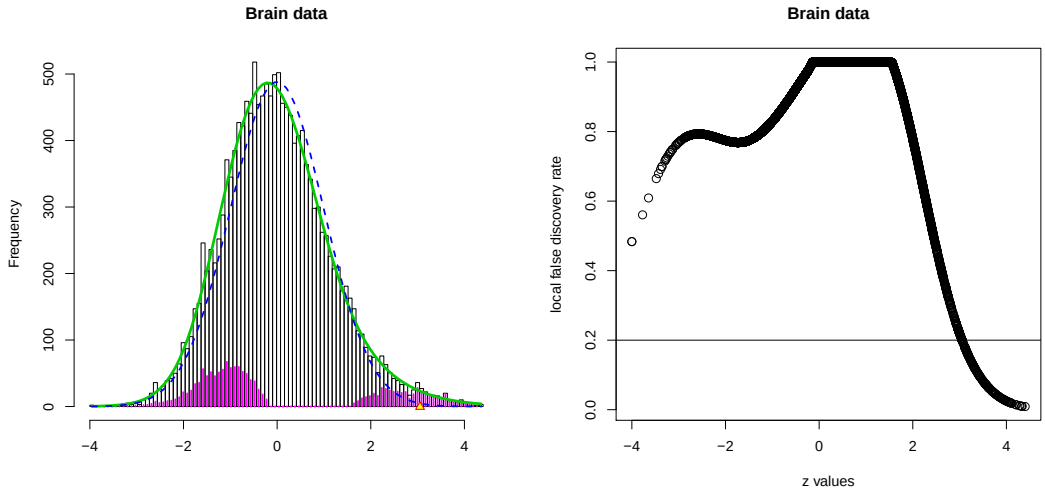



Figure 5.2: Left panel, histograms of $z$-values from Brain data. Short vertical bars are estimated non-null counts, useful for power calculations. Right panel, the LFDR estimate versus $z$-values for $N = 15,443$ measured voxels in Brain data under SMM. The horizontal line represents the threshold 0.2.

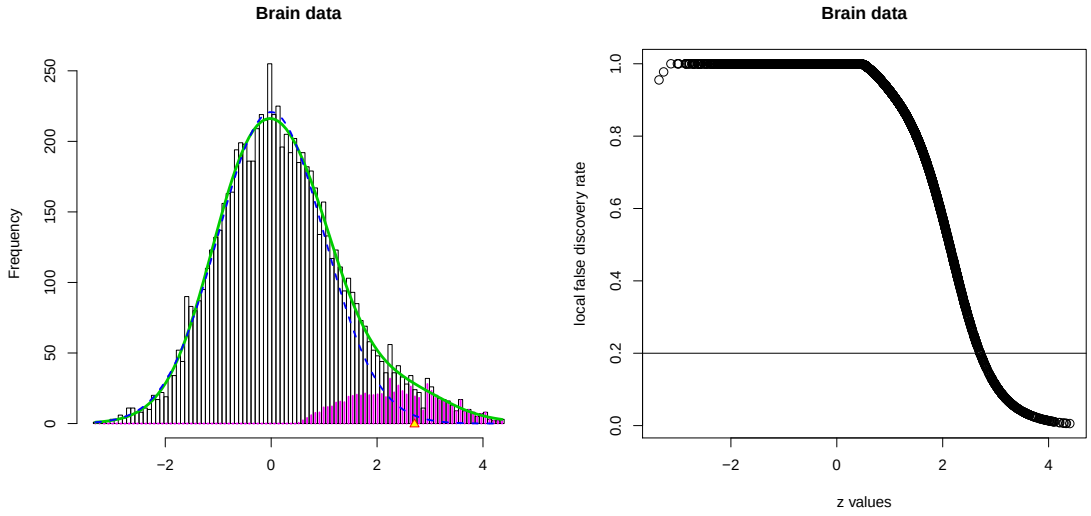



Figure 5.3: Left panel, histograms of $z$-values from half-front Brain data. Short vertical bars are estimated non-null counts, useful for power calculations. Right panel, the LFDR estimate versus $z$-values for $7,782$ half-front voxels in Brain data under SMM. The horizontal line represents the threshold 0.2.

## 5.2 Separate-class Model vs. Combined-class Model

A set of hypotheses (or features) to determine the posterior probability of a null hypothesis is called the *reference class*, and the problem of finding that set is called the *reference class* problem [13]. For example, in genetic association applications, the reference class problem is the problem of deciding which genetic markers should be used to determine the probability that a genetic marker is associated with the disease [13].

Efron [27] proposed a two-class model. The $N$ tests can be divided into two separate classes with specified prior probabilities, such that the two-groups model in Section 3.3 holds separately within each class. As an example, the Brain data can be divided into two classes $U$ and $V$ (i.e. front and back) with prior probabilities $\pi_U$

and $\pi_V$ respectively. Under class $U$, the two-groups model in Section 3.3 is assumed,

$$\pi_0 = \pi_{0U} \ , \ f_0(z_i) = f_{0U}(z_i) \ , \ f_1(z_i) = f_{1U}(z_i), \tag{5.2.1}$$

and the LFDR for the $i^{\text{th}}$ statistic $Z_i = z_i$ belongs to the class $U$ is

$$\psi_U(z_i) \equiv P(A_i = 0 | Z_i = z_i, z_i \in U) = \frac{\pi_{0U} f_{0U}(z_i)}{f_U(z_i)}, \tag{5.2.2}$$

The mixture density of statistic $Z_i$ under class $U$ is

$$f_U(z_i) = \pi_{0U} f_{0U}(z_i) + (1 - \pi_{0U}) f_{1U}(z_i).$$

The LFDR for class $V$ can be defined in the same way. The class label, $U$ or $V$, is specified by the statistician, whereas the null and non-null division must be concluded. Combining the two classes gives the following marginal densities,

$$\begin{aligned} f_0(z_i) &= \pi_U \frac{\pi_{0U}}{\pi_0} f_{0U}(z_i) + \pi_V \frac{\pi_{0V}}{\pi_0} f_{0V}(z_i) \\ f_1(z_i) &= \pi_U \frac{\pi_{1U}}{\pi_1} f_{1U}(z_i) + \pi_V \frac{\pi_{1V}}{\pi_1} f_{1V}(z_i), \end{aligned}$$

with prior probability $\pi_0 = \pi_U \pi_{0U} + \pi_V \pi_{0V}$. The mixture density $f(z_i)$ for statistic $Z_i$ is

$$f(z_i) = \pi_U f_U(z_i) + \pi_V f_V(z_i),$$

where $f_U$ and $f_V$ are the mixture density of statistic under class $U$ and $V$.

Under the separate-class model, each class is considered to identify non-null features (i.e. SNPs, genes, voxels) separately, while under the combined-class model all classes are combined and the $N$ hypotheses are considered together. Efron [27] indicated the connection between the separate-class LFDR in (5.2.2) and the combined-class LFDR in (4.1.1).

**Theorem 5.2.1** *Define $\pi_U(z_i)$ as the conditional probability of class $U$ given $Z_i = z_i$*

$$\pi_U(z_i) = P(Z_i \in U | Z_i = z_i),$$

*and let the conditional probability of class $U$ given $Z_i = z_i$ under the null hypothesis be denoted*

$$\pi_{U0}(z_i) = P_0(Z_i \in U | Z_i = z_i).$$

*Then*

$$\psi_U(z_i) = \psi(z_i)\frac{\pi_{U0}(z_i)}{\pi_U(z_i)},$$

*where $\psi$ denotes the combined-class LFDR in (4.1.1) and $\psi_U$ denotes the separate-class LFDR in (5.2.2).*

**Proof:** The ratio of the combined-class LFDR and the separate-class LFDR in (4.1.1) and (5.2.2) may be written as,

$$\begin{aligned}
\frac{\psi_U(z_i)}{\psi(z_i)} &= \frac{\mathrm{P}(A_i = 0 | Z_i \in U, Z_i = z_i)}{\mathrm{P}(A_i = 0 | Z_i = z_i)} \\
&= \frac{\mathrm{P}(A_i = 0, Z_i \in U | Z_i = z_i)}{\mathrm{P}(Z_i \in U | Z_i = z_i)\mathrm{P}(A_i = 0 | Z_i = z_i)} \\
&= \frac{\mathrm{P}(Z_i \in U | A_i = 0, Z_i = z_i)}{\mathrm{P}(Z_i \in U | Z_i = z_i)} \\
&= \frac{\mathrm{P}_0(Z_i \in U | Z_i = z_i)}{\mathrm{P}(Z_i \in U | Z_i = z_i)} \\
&= \frac{\pi_{U0}(z_i)}{\pi_U(z_i)}.
\end{aligned}$$

∎

Theorem 5.2.1 indicates that it is enough just to estimate two unknown conditional probabilities $\pi_{U0}(z_i)$ and $\pi_U(z_i)$. We show how to obtain such estimates on the Brain data. The Bayes false discovery rate in Section 3.3 also follows Theorem 5.2.1.

In order to estimate $\pi_U(z_i)$, it is easier to use the bin approach for the $z$-values. Suppose the region of $z$-values is partitioned into $K$ bins with equal length $l$. Let $N_k$ be the number of $z$-values in $k^{\text{th}}$ bin, and $N_{Uk}$ be the number of $z$-values from class $U$ in the bin $k$. Let $r_{Uk}$ be the proportion of $z$-values belong to both bin $k$ and class $U$,

$$r_{Uk} = \frac{N_{Uk}}{N_k}. \tag{5.2.3}$$

Let $x_k$ be the center point of the $k^{\text{th}}$ bin. Efron proposed the standard weighted cubic logistic regression approach in estimating $\pi_U(z_i)$. The estimate $\widehat{\pi}_U(z_i)$ is obtained by fitting the logistic regression of proportions $r_{Uk}$ as a cubic function of center point $x_k$ of the $k^{\text{th}}$ bin, where $N_k$ is used as weight. In order to estimate $\pi_{U0}(z_i)$, the null density $f_0(z_i)$ for each class is assumed to have normal densities as

$$f_{0U}(z_i) \sim N(\delta_{0U}, \sigma^2_{0U}) \quad , \quad f_{0V}(z_i) \sim N(\delta_{0V}, \sigma^2_{0V}). \tag{5.2.4}$$

Then by Bayes' Theorem,

$$\begin{aligned} \frac{\pi_{U0}(z_i)}{\pi_{V0}(z_i)} &= \frac{\pi_{U0}(z_i)}{1-\pi_{U0}(z_i)} \\ &= \frac{\pi_U \pi_{0U} f_{0U}(z_i)}{\pi_V \pi_{0V} f_{0V}(z_i)} \\ &= \frac{\pi_U \pi_{0U} \sigma_{0U}}{\pi_V \pi_{0V} \sigma_{0V}} \exp\Big\{ -0.5\Big[\Big(\frac{z_i-\delta_{0U}}{\sigma_{0U}}\Big)^2 - \Big(\frac{z_i-\delta_{0V}}{\sigma_{0V}}\Big)^2\Big]\Big\}, \end{aligned} \tag{5.2.5}$$

which yields $\pi_{U0}(z_i)$.

**Corollary 5.2.1** *For $z$-values close to zero,*

$$\widehat{\pi}_{U0}(z_i) = \widehat{\pi}_U(z_i).$$

**Proof:** From Theorem 5.2.1 and equations (5.2.1) and (5.2.2),

$$\begin{aligned}\frac{\pi_{U0}(z_i)}{1-\pi_{U0}(z_i)} &= \frac{\mathrm{P}(Z_i \in U | Z_i = z_i, A_i = 0)}{1 - \mathrm{P}(Z_i \in U | Z_i = z_i, A_i = 0)} \\ &= \frac{\mathrm{P}(Z_i \in U | Z_i = z_i, A_i = 0)}{\mathrm{P}(Z_i \in V | Z_i = z_i, A_i = 0)} \\ &= \frac{\mathrm{P}(A_i = 0 | z_i \in U, Z_i = z_i)\mathrm{P}(Z_i \in U | Z_i = z_i)}{\mathrm{P}(A_i = 0 | z_i \in V, Z_i = z_i)\mathrm{P}(Z_i \in V | Z_i = z_i)}\end{aligned}$$

$$= \frac{\psi_U(z_i)\pi_U(z_i)}{\psi_V(z_i)\pi_V(z_i)}.$$

Then for the $z$-values close to zero, the LFDR estimates will be approximately equal 1, which means $\psi_U(z_i) \approx 1$ and $\psi_V(z_i) \approx 1$, and for such $z$-values $\widehat{\pi}_{U0}(z_i) = \widehat{\pi}_U(z_i)$. ∎

Under the following corollary, the conditional probability $\pi_{U0}(z_i)$ may be constant as a function of $z_i$ when the null densities under the two-class model are the same.

**Corollary 5.2.2** *If the null densities under the two-class model are the same, i.e.* $f_{0U}(z_i) = f_{0V}(z_i)$*, then the conditional probability* $\pi_{U0}(z_i)$ *is,*

$$\pi_{U0}(z_i) = \frac{\pi_U \pi_{0U}}{\pi_0}.$$

**Proof:** From Theorem 5.2.1 and (5.2.5), when $f_{0U}(z_i) = f_{0V}(z_i)$,

$$\frac{\pi_{U0}(z_i)}{1-\pi_{U0}(z_i)} = \frac{\pi_U \pi_{0U}}{(1-\pi_U)\pi_{0V}}, \tag{5.2.6}$$

and by solving for $\pi_{U0}$,

$$\pi_{U0} = \frac{\pi_U \pi_{0U}}{\pi_U \pi_{0U} + (1-\pi_U)\pi_{0V}} = \frac{\pi_U \pi_{0U}}{\pi_0}.$$

The accuracy of separation analysis in estimating the LFDR is another concern in large-scale hypothesis testing. This separation is dangerous from the frequentist view point while it is applicable from the Bayesian view [27].

We apply the two-class model for the Brain data. The class $U$ refers to front-half voxels. The points in Figure 5.4 represent $r_{Uk}$ for $K = 42$ bins, with equal length $l = 0.2$, between $-4.2$ and 4.2. We consider a standard weighted logistic regression model to fit $\text{logit}(r_{Uk})$ as a cubic function of center point $x_k$ of bin $k$ with weight $N_k$ which is denoted by solid line in Figure 5.4.

From Figure 5.4, the dashed line displays the estimate of conditional probability $\pi_{U0}$. To obtain such an estimate, the null distribution should be specified. The empirical null hypothesis is assumed for each class. From Table 5.1, the required parameters were estimated by applying the maximum likelihood fitting method in Section 4.3.1. It is assumed $\pi_U = \pi_V$, which means half of the voxels are in the class $U$ and the other half in the class $V$. By applying (5.2.5), the estimate $\widehat{\pi}_{U0}(z_i)$ is determined.

Figure 5.4 shows $\widehat{\pi}_{U0}(z_i) = \widehat{\pi}_U(z_i)$ for $z$-values close to zero (see in Corollary 5.2.1). On the other hand, the estimate $\widehat{\pi}_U(z_i)$ is not a constant function of $z$-values around zero, which means there is not enough evidence to assume $f_{0U}(z_i)$ is the same as $f_{0V}(z_i)$.

In the next section, we apply the two-class model on the CAD data discussed in Sections 3.4.2 and 4.5.

Table 5.1: Parameter estimates for Brain data in two-class modeling

| Estimates | $\hat{\pi}_0$ | $\hat{\delta}_0$ | $\hat{\sigma}$ | $\pi$ |
|---|---|---|---|---|
| front-half ($U$): | 0.97 | 0.041 | 1.09 | 0.5 |
| back-half ($V$): | 0.98 | $-0.319$ | 0.98 | 0.5 |

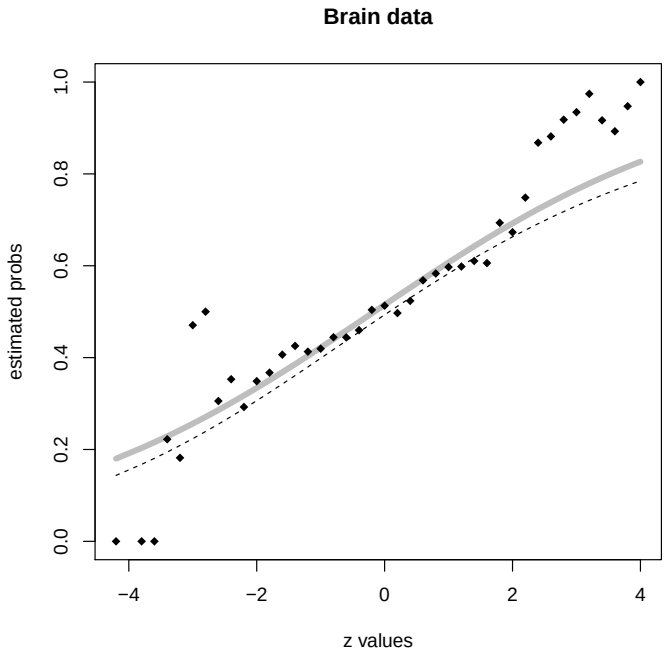



Figure 5.4: Brain data: Points are proportion of front-half voxels $r_{Uk}$. Solid curve is $\widehat{\pi}_U$. Dashed curve is $\widehat{\pi}_{0U}$.

## 5.3 Application

The CAD data introduced in Sections 2.2 is used as an example for determining the disease-associated SNPs under the two-class model.

There are $N = 394,838$ measured SNPs. Under SMM, 44 SNPs are associated with disease, with $\widehat{\psi^\star} = 0.93$ which shows low power in identifying non-null SNPs. We improve the LFDR estimate in identifying the disease-associated SNPs when the MAF for each SNP is incorporated. We used the information from MAFs in classifying $N$ tests into two distinct classes. The two classes are: SNPs with $1\% \leq \text{MAF} \leq 5\%$ are called *low-frequency* SNPs (i.e. class $U$), SNPs with $5\% \leq \text{MAF} \leq 50\%$ are called *common* SNPs (i.e. class $V$). We apply the two-groups model defined in Section 3.3 for each class. For each class, we identify the number of disease-associated SNPs

under SMM given in Section 4.3.1.

The number of SNPs in class $U$ is $37,369$, 6 disease-associated SNPs are in this class when the theoretical null hypothesis is assumed with $\widehat{\pi}_{0U} = 0.98$ and power diagnostic estimate $\widehat{\psi^\star} = 0.39$ shows better power in identifying disease-associated SNPs compared with given results in Section 4.5 when all $N$ SNPs are used in identifying disease-associated SNPs. On the other hand, the number of SNPs in class $V$ is $357,469$, under SMM and the theoretical null hypothesis in Section 4.3.1, 36 SNPs are associated with disease with power diagnostic estimate $\widehat{\psi^\star} = 0.94$, which is still low to identify disease-associated SNPs. We see how the two-class model affects the number of discoveries when analyzing the CAD data.

The points in Figure 5.5 represent $r_{Uk}$ for $K = 60$ bins, with equal length $l = 0.2$ between $-7.5$ and 4.5. We consider a standard weighted logistic regression model to fit $\text{logit}(r_{Uk})$ as a cubic function of center point $x_k$ of bin $k$ with weight $N_k$, which is denoted by a solid line in Figure 5.5.

From Figure 5.5, the dashed line displays the estimate of conditional probability $\pi_{U0}$. To obtain such an estimate, the null distribution should be specified. The empirical null hypothesis is assumed for each class. From Table 5.2, the required parameters were estimated by applying the maximum likelihood fitting method in Section 4.3.1. It is assumed $\pi_U = 0.1$ and $\pi_V = 0.9$, which means 90% of SNPs are common and 10% of SNPs are low-frequency. By applying (5.2.5), the estimates $\widehat{\pi}_{U0}(z_i)$ is determined.

Figure 5.5 shows $\widehat{\pi}_{U0}(z_i) = \widehat{\pi}_U(z_i)$ for $z$-values close to zero (see in Corollary 5.2.1). From Corollary 5.2.2, the estimates $\widehat{\pi}_U(z_i)$ is a constant function of $z$-values around zero, which means there is enough evidence to assume $f_{0U}(z_i) = f_{0V}(z_i)$.

Table 5.2: Parameter estimates for CAD data in two-class modeling

| Estimates | $\hat{\pi}_0$ | $\hat{\delta}_0$ | $\hat{\sigma}$ | $\pi$ |
|---|---|---|---|---|
| low-frequency SNPs ($U$): | 0.99 | $-0.002$ | 1 | 0.1 |
| common SNPs ($V$): | 0.99 | $-0.03$ | 1.01 | 0.9 |

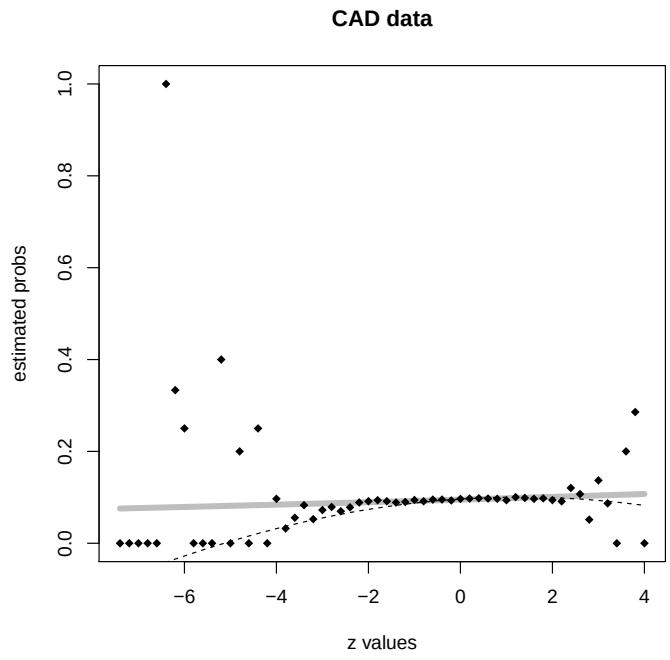



Figure 5.5: CAD data: Points are proportion of minor-allele frequency SNPs $r_{Uk}$. Solid curve is $\widehat{\pi}_U$. Dashed curve is $\widehat{\pi}_{0U}$.

## 5.4 Discussion and Conclusions

The application of Theorem 5.2.1 is for small classes. For both the Brain and CAD data sets, the specified classes contain enough features. Instead of considering Theorem 5.2.1, it is easier to apply each class directly in estimating the LFDR to identify the non-null features (i.e. SNPs, genes, voxels), whereas for small classes, instead of

direct estimation of the LFDR in (5.2.2), it is more appropriate to apply Theorem 5.2.1 to obtain the LFDR estimate. Efron elaborated further on the efficiency of the LFDR estimate regarding to the separate-class model [27].

The Brain data analysis indicates how incorporating the brain location in estimating the LFDR can affect the number of discoveries. Also from the CAD data analysis, it is seen that the combined-class model is misleading in estimating the LFDR for the low-frequency SNPs. By considering low-frequency SNPs to estimate the LFDR for that class, it seems power is increased to identify the disease-associated SNPs. In other words, classifying SNPs by incorporating MAFs can improve the results based on the LFDR estimate.

Following the two motivating examples in this chapter, we propose a novel approach in the next chapter to improve the LFDR estimate when a co-variate for each test is incorporated.

# Chapter 6

# Improving the Local False Discovery Rate Estimate by Incorporating a Co-variate

Previous statistical techniques mentioned in Section 5.1 considered the finite number of groups. When these methods use the continuous co-variates, divide them into the finite number of groups where some information is lost. Our novel approach relaxes such limitation. Our method is adaptive, since it looks at the test statistics to determine the reference classes. In this study, we are concerned with estimating an optimal reference class to improve the LFDR estimate. We will present a novel approach in improving the LFDR estimate. In such a case, we incorporate a co-variate for each test in order to improve the estimation of LFDR. Each null hypothesis test is assigned a distinct prior probability where it is a function of a co-variate. In applications, the co-variate represents the scientific structure of each test. We propose an adaptive reference class (ARC) method to estimate the LFDR, where both a bias-variance trade off and a ootstrap approaches are used in order to determine the optimal reference class. We compare the performance of the ARC method with the

combined reference class (or combined-class) method explained in Chapter 4. Then, we apply the methodology on real data.

## 6.1 Proposed Model

Suppose $N$ hypothesis tests are considered. Each hypothesis test refers to a gene (e.g. in microarry data), a SNP (e.g. in GWA data), or a voxel (e.g. in Brain data). Under the $i^{\text{th}}$ null hypothesis $H_{0i}$, the $p$-value $p_i$ is determined according to the test statistic $T_i$. As discussed in Chapter 3, the $p$-values under the null hypothesis are uniformly distributed. Under the $i^{\text{th}}$ null hypothesis $H_{0i}$, the observed test statistic $t_i$ is transformed to the observed statistic $z_i$. The $z$-value vector $\underline{z} = (z_1, .., z_N)^T$ is considered a realization of $\underline{Z} = (Z_1, ..., Z_N)^T$, where each $Z_i$ under the null hypothesis has a standard normal distribution,

$$\text{under } H_{0i} : Z_i \sim N(0, 1). \tag{6.1.1}$$

We incorporate a co-variate $X$. In applications, the observed co-variate may represent the MAF for SNPs in the CAD data study or the brain location in analyzing the Brain data. Each test has individual information which may be connected to the other tests. As an example, in the CAD data, we may have some SNPs with the same MAF. This means such SNPs scientifically have the same structure. Let $\{X_i\}$ be independent and identically distributed random variables with probability distribution $P_x$. Let $\underline{x} = (x_1, ..., x_N)^T$ be a realization of $\underline{X} = (X_1, ..., X_N)^T$.

Let $A_i$ be an indicator variable for the event that a non-null hypothesis $H_{ai}$ is true,

$$A_i = \begin{cases} 0 & \text{if } H_{0i} \text{ is true} \\ 1 & \text{if } H_{ai} \text{ is true} \end{cases} \tag{6.1.2}$$

Assume that the conditional distribution of $A_i$ given $X_i = x_i$, is Bernoulli$\big(1 - \pi_0(x_i)\big)$

where $\pi_0(x_i)$, the prior probability that the $i^{\text{th}}$ null hypothesis is true, is now a function of $x_i$.

The posterior probability that the $i^{\text{th}}$ null hypothesis is true given the data, is called the *local false discovery rate* (LFDR) is denoted as $\psi(z_i; x_i)$ where

$$\psi(z_i; x_i) \equiv \mathrm{P}(A_i = 0 | Z_i = z_i, X_i = x_i) = \frac{\pi_0(x_i) f_0(z_i)}{f\big(z_i; \pi_0(x_i)\big)} \tag{6.1.3}$$

and $f_0(z_i) \sim N(0,1)$ denotes the null density of the statistic $Z_i$. $f\big(z_i; \pi_0(x_i)\big)$ represents the mixture density of $Z_i$ conditional on the co-variate $X_i = x_i$,

$$f\big(z_i; \pi_0(x_i)\big) = \pi_0(x_i) f_0(z_i) + \big(1 - \pi_0(x_i)\big) f_1(z_i; x_i), \tag{6.1.4}$$

and $f_1(z_i; x_i)$ is an unknown density function of $Z_i$ under the $i^{\text{th}}$ non-null hypothesis $H_{ai}$. The $Z_i$'s are not identically distributed under the non-null hypothesis. Associated with each hypothesis, we have three components,

$$(x_i, z_i, A_i) \quad , \quad i = 1, ..., N, \tag{6.1.5}$$

where $x_i$ represents the observed co-variate, $z_i$ is the observed $z$-value, and $A_i$ is the unobservable indicator variable. The connection between the LFDR $\psi(z_i; x_i)$ and the prior probability $\pi_0(x_i)$ is considered in the following Lemma.

**Lemma 6.1.1** *If conditional on $X_i = x_i$, the random indicator $A_i$ is Bernoulli$\big(1 - \pi_0(x_i)\big)$, then*

$$E\big(\psi(z_i; x_i) | X_i = x_i\big) = \pi_0(x_i).$$

**Proof:** We have

$$E(A_i | Z_i = z_i, X_i = x_i) = P(A_i = 1 | Z_i = z_i, X_i = x_i) = 1 - \psi(z_i; x_i).$$

Taking expectation with respect to the density of $Z_i$ given $X_i = x_i$,

$$E(A_i|X_i = x_i) = 1 - E\big(\psi(z_i; x_i)|X_i = x_i\big),$$

since $E(A_i|X_i = x_i) = 1 - \pi_0(x_i)$, then implies $E\big(\psi(z_i; x_i)|X_i = x_i\big) = \pi_0(x_i)$. ∎

The LFDR in (6.1.3) is unknown, since both the prior probability $\pi_0(x_i)$ and the non-null density $f_1(z_i; x_i)$ are unknown. It is seen that the total number of unknown parameters is greater than the number of observations.

## 6.2 Methods for Estimation

We consider the adaptive reference class (ARC) method in order to estimate the LFDR $\psi(z_i; x_i)$ in (6.1.3). We also consider the combined reference class (CRC) method (i.e. combined-class method) [27] for this proposed model. In the CRC method the effect of the co-variates is ignored, whereas in the proposed ARC method, we consider some assumptions locally to estimate the LFDR.

### 6.2.1 Combined Reference Class (CRC) Method

The CRC method ignores the co-variate information and we suppose the random variables $A_i$'s are independent and identically distributed,

$$A_i \overset{iid}{\sim} \text{Bernoulli}\big(1 - \pi_0\big) \tag{6.2.1}$$

with common prior probability $\pi_0$ for $i = 1, ..., N$. The posterior probability that the $i^{\text{th}}$ null hypothesis $H_{0i}$ is true given $Z_i = z_i$, denote $\psi(z_i)$, is given by

$$\psi(z_i) \equiv \mathrm{P}(A_i = 0 | Z_i = z_i) = \frac{\pi_0 f_0(z_i)}{f\big(z_i; \pi_0\big)} \tag{6.2.2}$$

which is the special form of the LFDR in (6.1.3), where

$$f\big(z_i; \pi_0\big) = \pi_0 f_0(z_i) + \big(1 - \pi_0\big) f_1(z_i) \tag{6.2.3}$$

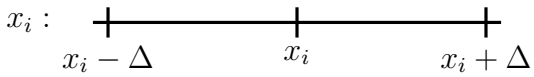
and $f_1(z_i)$ represents the unknown non-null density of $Z_i$. The LFDR in (6.2.2) remains unknown since $\pi_0$ and $f_1(z_i)$ are unknown. Under the CRC method, the observed vector of statistics, $\underline{z}$, is used to estimate the LFDR. Let $\widehat{\psi}_i(\underline{z})$ denote an estimate of the LFDR $\psi(z_i)$ in (6.2.2), where $\widehat{\psi}_i(\bullet)$ is a function such that $\widehat{\psi}_i : \mathbb{R}^d \to [0, 1]$ for $0 < d \leq N$. Assume $\widehat{\psi}_i(\underline{z})$ is a weakly consistent estimator of $\psi(z_i)$,

$$\lim_{N \to \infty} P\big(|\widehat{\psi}_i(\underline{z}) - \psi(z_i)| > \epsilon\big) = 0,$$

for any $\epsilon > 0$.

### 6.2.2 Adaptive Reference Class (ARC) Method

Under the ARC method, we make certain assumptions that hold only locally in a symmetric window around each co-variate. Specifically, we consider a symmetric window around $x_i$ with length $2\Delta$, where the tuning parameter $\Delta \in [\Delta_0, \infty)$, and $\Delta_0 > 0$ represents the smallest value of $\Delta$,




Such a symmetric window provides a sub-vector of $\underline{z}$ is called a *reference class*

and is denoted by $\underline{z}_i^{\Delta}$. When $X_i = x_i$, the reference class $\underline{z}_i^{\Delta}$ is

$$\underline{z}_i^{\Delta} \equiv \underline{z}(\Delta, x_i, \underline{x}) = \{z_j : \ |x_j - x_i| \leq \Delta \ , \ j = 1, ..., N\}. \tag{6.2.4}$$

The reference class $\underline{z}_i^{\Delta}$ contains $z_j$'s such that their co-variates are within $\Delta$ of $x_i$. When $X_i = x_i$, the expected dimension of the reference class $\underline{z}_i^{\Delta}$ denoted by $d_i^{\Delta}$ is given by,

$$d_i^{\Delta} \equiv N\mathrm{P}\big(|X_j - x_i| \leq \Delta, j = 1, ..., N\big). \tag{6.2.5}$$

If the probability on the right side of (6.2.5) is positive, then the expected dimension $d_i^{\Delta}$ will be large for large $N$.

When $X_i = x_i$ for fixed $\Delta$, under the two-groups model in Section 3.3, $\underline{z}_i^{\Delta}$ is used to estimate the LFDR $\psi(z_i; x_i)$, in (6.1.3). We may apply either SMM or PMM to estimate the LFDR in (6.1.3). In other words, the function $\widehat{\psi}_i$ is applied on the reference class $\underline{z}_i^{\Delta}$ to estimate the LFDR which is denoted by $\widehat{\psi}_i(\underline{z}_i^{\Delta})$.

## 6.3 Bias-Variance Tradeoff and Bootstrap Estimation

When $X_i = x_i$, a change in $\Delta$ yields a different reference classes and provides different LFDR estimates. Among those estimates, we would like to determine the one which is closest to the true LFDR in (6.1.3). The estimation of the tuning parameter $\Delta$ is discussed in the next section.

### 6.3.1 Bias-Variance Tradeoff

The described ARC method depends on the tuning parameter $\Delta$ that needs to be determined. The tuning parameter $\Delta$ specifies the symmetric window size. By increasing the window size, the expected dimension of the reference class gets large.

Among such classes, we would like to choose the one which minimizes the errors in estimating the LFDR in (6.1.3). The errors in estimation are the errors due to bias and to variance.

The choice of the optimal $\Delta$ depends on the choice of a loss function to measure errors in estimation. We consider quadratic loss. When $X_i = x_i$, the quadratic loss for the estimator of the LFDR is denoted by

$$L\big(\psi(z_i; x_i), \widehat{\psi}_i(\underline{z}_i^{\Delta})\big) = \big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)\big)^2. \tag{6.3.1}$$

**Definition 6.3.1** *When $X_i = x_i$, the mean and variance of the estimator $\widehat{\psi}_i(\underline{z}_i^{\Delta})$ are defined respectively as,*

$$\mu_{\Delta}(x_i) = E\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) \;\;,\;\; \sigma^2_{\Delta}(x_i) = E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \mu_{\Delta}(x_i)\big)^2|X_i = x_i\big].$$

**Definition 6.3.2** *When $X_i = x_i$, the prediction bias for the estimator $\widehat{\psi}_i(\underline{z}_i^{\Delta})$ is given by,*

$$\mathcal{B}_{\Delta}(x_i) = E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)\big)|X_i = x_i\big].$$

**Lemma 6.3.1** *The prediction bias for the estimator $\widehat{\psi}_i(\underline{z}_i^{\Delta})$ is given by*

$$\mathcal{B}_{\Delta}(x_i) = \mu_{\Delta}(x_i) - \pi_0(x_i).$$

**Proof:** From definition 6.3.1, the prediction bias in definition 6.3.2 can be expanded to,

$$\begin{aligned}
\mathcal{B}_{\Delta}(x_i) &= E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)\big)|X_i = x_i\big] \\
&= E\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) - E\big(\psi(z_i; x_i)|X_i = x_i\big) \\
&= \mu_{\Delta}(x_i) - \pi_0(x_i),
\end{aligned}$$

which follows from Lemma 6.1.1. ∎

**Definition 6.3.3** *Conditional on $X_i = x_i$, the expected prediction squared error for the estimator $\widehat{\psi}_i(\underline{z}_i^{\Delta})$ is defined as*

$$EPE\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) = E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)\big)^2|X_i = x_i\big].$$

**Lemma 6.3.2** *The expected prediction squared error for estimator $\widehat{\psi}_i(\underline{z}_i^{\Delta})$ is expanded to*

$$EPE\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) = \sigma^2_{\Delta}(x_i) + \mathcal{B}^2_{\Delta}(x_i) + Var\big(\psi(z_i; x_i)|X_i = x_i\big).$$

**Proof:** From definitions 6.3.1 and 6.3.2, the expected prediction squared error in definition 6.3.3 can be expressed as,

$$\begin{aligned}
\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) &= E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)\big)^2|X_i = x_i\big] \\
&= E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \mu_{\Delta}(x_i) + \mu_{\Delta}(x_i) - \psi(z_i; x_i)\big)^2|X_i = x_i\big] \\
&= E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \mu_{\Delta}(x_i)\big)^2|X_i = x_i\big] \\
&+ E\big[\big(\mu_{\Delta}(x_i) - \psi(z_i; x_i)\big)^2|X_i = x_i\big] \\
&+ 2E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \mu_{\Delta}(x_i)\big)\big(\mu_{\Delta}(x_i) - \psi(z_i; x_i)\big)|X_i = x_i\big],
\end{aligned}$$

where the first term represents the variance of $\widehat{\psi}_i(\underline{z}_i^{\Delta})$,

$$E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \mu_{\Delta}(x_i)\big)^2|X_i = x_i\big] = \sigma^2_{\Delta}(x_i).$$

The second term becomes,

$$\begin{aligned}
E\big[\big(\mu_{\Delta}(x_i) - \psi(z_i; x_i)\big)^2|X_i = x_i\big] &= E\big(\psi^2(z_i; x_i)|X_i = x_i\big) \\
&+ E\big(\mu^2_{\Delta}(x_i)|X_i = x_i\big)
\end{aligned}$$

$$
\begin{aligned}
&- 2E\big[\mu_\Delta(x_i)\psi(z_i;x_i)|X_i = x_i\big] \\
&= E\big(\psi^2(z_i;x_i)|X_i = x_i\big) + \mu_\Delta^2(x_i) \\
&- 2\mu_\Delta(x_i)E\big[\psi(z_i;x_i)|X_i = x_i\big] \\
&= E\big(\psi^2(z_i;x_i)|X_i = x_i\big) \\
&+ \mu_\Delta^2(x_i) - 2\mu_\Delta(x_i)\pi_0(x_i).
\end{aligned}
$$

Finally the third term is

$$
\begin{aligned}
E\big[\big(\widehat{\psi}_i(\underline{z}_i^\Delta) - \mu_\Delta(x_i)\big)\big(\mu_\Delta(x_i) - \psi(z_i;x_i)\big)|X_i = x_i\big] &= E\big[\big(\widehat{\psi}_i(\underline{z}_i^\Delta) - \mu_\Delta(x_i)\big)|X_i = x_i\big] \\
&\times E\big[\big(\mu_\Delta(x_i) - \psi(z_i;x_i)\big)|X_i = x_i\big] \\
&= 0.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^\Delta)|X_i = x_i\big) &= \sigma_\Delta^2(x_i) + \mu_\Delta^2(x_i) - 2\mu_\Delta(x_i)\pi_0(x_i) \\
&+ E\big(\psi^2(z_i;x_i)|X_i = x_i\big) \\
&= \sigma_\Delta^2(x_i) + \big[\mu_\Delta(x_i) - \pi_0(x_i)\big]^2 - \pi_0^2(x_i) \\
&+ E\big(\psi^2(z_i;x_i)|X_i = x_i\big) \\
&= \sigma_\Delta^2(x_i) + \big[E\big(\widehat{\psi}_i(\underline{z}_i^\Delta)|X_i = x_i\big) - E\big(\psi(z_i;x_i)|X_i = x_i\big)\big]^2 \\
&- E^2\big(\psi(z_i;x_i)|X_i = x_i\big) + E\big(\psi^2(z_i;x_i)|X_i = x_i\big) \\
&= \sigma_\Delta^2(x_i) + \mathcal{B}_\Delta^2(x_i) + Var\big(\psi(z_i;x_i)|X_i = x_i\big).
\end{aligned}
$$

∎

Since the third term in Lemma 6.3.2 is independent of $\Delta$, we shall consider in what

follows only the errors due to bias and to variance,

$$\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) = \sigma^2_{\Delta}(x_i) + \mathcal{B}^2_{\Delta}(x_i). \tag{6.3.2}$$

Typically we would like to choose $\Delta$ to trade bias off with variance in such a way as to minimize the errors in (6.3.2) with respect to all possible value $\Delta \in [\Delta_0, \infty)$.

**Definition 6.3.4** *Conditional on $X_i = x_i$, the optimal $\Delta$ denoted by $\Delta^\star(x_i)$ minimizes the errors in 6.3.2 and is given by*

$$\Delta^\star(x_i) = \arg \inf_{\Delta \geq \Delta_0} err\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big).$$

In order to estimate $\Delta^\star(x_i)$, it is necessary to estimate the unknown terms, the variance $\sigma^2_{\Delta}(x_i)$, the expected value of the LFDR estimate $\mu_{\Delta}(x_i)$, and the prior probability $\pi_0(x_i)$. We consider the *bootstrap* approach to estimate these quantities.

### 6.3.2 Bootstrap Approach

Consider the pair $(z_i, x_i)$ where $z_i \in \underline{z}$ and $x_i \in \underline{x}$. The bootstrap data $(\ddot{\underline{z}}, \ddot{\underline{x}})$ represents a random sample of size $N$ drawn with replacement from $(z_j, x_j)$ for $j = 1, ..., N$. Repeat this process $B$ times to get the bootstrap samples,

$$(\ddot{\underline{z}}_1, \ddot{\underline{x}}_1), (\ddot{\underline{z}}_2, \ddot{\underline{x}}_2), ..., (\ddot{\underline{z}}_B, \ddot{\underline{x}}_B).$$

Among the $B$ bootstrap samples, we retain the samples which contain the pair $(z_i, x_i)$. Let $B'(x_i) = B'_i$ be the total number of bootstrap samples containing $(z_i, x_i)$. Such samples are given by

$$(\underline{z}^\star_1, \underline{x}^\star_1), (\underline{z}^\star_2, \underline{x}^\star_2), ..., (\underline{z}^\star_{B'_i}, \underline{x}^\star_{B'_i}).$$

The $b^{\text{th}}$ bootstrap sample $(\underline{z}_b^\star, \underline{x}_b^\star)$ contains pairs $(z_{bj}^\star, x_{bj}^\star)$,

$$(\underline{z}_b^\star, \underline{x}_b^\star) : (z_{b1}^\star, x_{b1}^\star), ..., (z_{bN}^\star, x_{bN}^\star), \tag{6.3.3}$$

where $b = 1, ..., B_i'$. The bootstrap reference class is defined as

$$\underline{z}_{i,b}^\Delta \equiv \underline{z}_b^\star(\Delta, x_i, \underline{x}_b^\star) = \{z_{bj}^\star : \ |x_{bj}^\star - x_i| \leq \Delta \ , \ j = 1, ..., N\}. \tag{6.3.4}$$

Each $b^{\text{th}}$ bootstrap reference class yields an estimate of $\psi(z_i; x_i)$ in (6.1.3) and is denoted by $\widehat{\psi}_i(\underline{z}_{i,b}^\Delta)$ (see in Section 6.2.2).

The $B_i'$ random variables $\widehat{\psi}_i(\underline{z}_{i,1}^\Delta), ..., \widehat{\psi}_i(\underline{z}_{i,B_i'}^\Delta)$ provide the estimators $\widehat{\mu}(\Delta, B_i')$ and $\widehat{\sigma}^2(\Delta, B_i')$ for $\mu_\Delta(x_i)$ and $\sigma_\Delta^2(x_i)$ respectively,

$$\begin{aligned} \widehat{\mu}(\Delta, B_i') &= \frac{1}{B_i'} \sum_{b=1}^{B_i'} \widehat{\psi}_i(\underline{z}_{i,b}^\Delta), \\ \widehat{\sigma}^2(\Delta, B_i') &= \frac{1}{B_i' - 1} \sum_{b=1}^{B_i'} \big(\widehat{\psi}_i(\underline{z}_{i,b}^\Delta) - \widehat{\mu}(\Delta, B_i')\big)^2. \end{aligned} \tag{6.3.5}$$

In order to estimate the errors of $\widehat{\psi}_i(\underline{z}_i^\Delta)$ in (6.3.2), the prior probability $\pi_0(x_i)$ in Lemma 6.1.1 has to be estimated. We propose a reference class $\underline{z}_{i,b}^{\Delta_0}$ which contains observed statistics $z_j$'s such that their co-variates are within $\Delta_0$ of $x_i$. By applying (6.3.5) when $\Delta = \Delta_0$, the estimator $\widehat{\mu}(\Delta_0, B_i')$ is given by

$$\widehat{\mu}(\Delta_0, B_i') = \frac{1}{B_i'} \sum_{b=1}^{B_i'} \widehat{\psi}_i(\underline{z}_{i,b}^{\Delta_0}), \tag{6.3.6}$$

that is the bootstrap estimator of the prior probability $\pi_0(x_i)$. So the bootstrap estimator of the prediction bias in Lemma 6.3.1 denoted by $\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i')$ is given by

$$\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i') = \widehat{\mu}(\Delta, B_i') - \widehat{\mu}(\Delta_0, B_i'). \tag{6.3.7}$$

The estimator of $\text{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big)$, denoted by $\widehat{\text{err}}(\Delta, \Delta_0, B_i')$, is computed as

$$\widehat{\text{err}}(\Delta, \Delta_0, B_i') = \widehat{\sigma}^2(\Delta, B_i') + \widehat{\mathcal{B}}^2(\Delta, \Delta_0, B_i'). \tag{6.3.8}$$

From definition 6.3.1, the estimators $\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta})$, $b = 1, ..., B_i'$ are independent and identically distributed with the same mean and variance as

$$\begin{aligned} \mu_{\Delta}(x_i) &= E\big(\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta})|X_i = x_i\big), \\ \sigma_{\Delta}^2(x_i) &= E\big[\big(\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta}) - \mu_{\Delta}(x_i)\big)^2|X_i = x_i\big] \end{aligned} \tag{6.3.9}$$

where $\sigma_{\Delta}^2(x_i) < \infty$.

**Definition 6.3.5** *Conditional on $X_i = x_i$, the bootstrap estimator of $\Delta^{\star}(x_i)$ denoted by $\widehat{\Delta}_{0i}^{\star}$, is given by*

$$\widehat{\Delta}_{0i}^{\star} = \arg \inf_{\Delta \geq \Delta_0} \widehat{err}(\Delta, \Delta_0, B_i').$$

The optimal reference class is determined by $\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}}$,

$$\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}} \equiv \underline{z}(\widehat{\Delta}_{0i}^{\star}, x_i, \underline{x}) = \{z_j : \ |x_j - x_i| \leq \widehat{\Delta}_{0i}^{\star} \ , \ j = 1, ..., N\}, \tag{6.3.10}$$

which contains $z_j$'s such that their co-variates are within $\widehat{\Delta}^{\star}(x_i)$ of $x_i$. The reference class $\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}}$ is used to estimate $\psi(z_i; x_i)$ by applying SMM or PMM (see Sections 4.3.1 and 4.3.2). Such an estimate is denoted by $\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}})$.

## 6.4 Combined Reference Class (CRC) Method vs. Adaptive Reference Class (ARC) Method

In this section we compare the two estimators of the LFDR in (6.1.3) related to the CRC method's estimator, $\widehat{\psi}_i(\underline{z})$, and the ARC method's estimator; $\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}})$. In the

following section, we assess their performances.

The following two lemmas show the weak consistency of the bootstrap estimators in (6.3.5). The bootstrap sample mean $\widehat{\mu}(\Delta, B_i')$ and the bootstrap sample variance $\widehat{\sigma}^2(\Delta, B_i')$ converge to $\mu_\Delta(x_i)$ and $\sigma_\Delta^2(x_i)$ respectively, when the number of bootstrap samples $B_i'$ gets large [33].

**Lemma 6.4.1** *The bootstrap sample mean $\widehat{\mu}(\Delta, B_i')$, is a weakly consistent estimator of $\mu_\Delta(x_i)$, that is*

$$\lim_{B_i' \to \infty} P\big(|\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)| > \epsilon | X_i = x_i\big) = 0$$

*for any $\epsilon > 0$.*

**Proof:** The bootstrap sample mean $\widehat{\mu}(\Delta, B_i')$ in (6.3.5) is an unbiased estimator of $\mu_\Delta(x_i)$, since

$$\begin{aligned} E\big(\widehat{\mu}(\Delta, B_i')|X_i = x_i\big) &= \frac{1}{B_i'} \sum_{b=1}^{B_i'} E\big(\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta})|X_i = x_i\big) \\ &= \frac{1}{B_i'} \sum_{b=1}^{B_i'} E\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) \\ &= \mu_\Delta(x_i). \end{aligned}$$

It has finite variance

$$E\big[\big(\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)\big)^2 | X_i = x_i\big] = \frac{\sigma_\Delta^2(x_i)}{B_i'}.$$

By applying Chebyshev's inequality to $\widehat{\mu}_\Delta(x_i)$, for any $\epsilon > 0$ we have

$$\mathrm{P}\big(|\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)| > \epsilon | X_i = x_i\big) \leq \frac{E\big[\big(\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)\big)^2 | X_i = x_i\big]}{\epsilon^2}$$

$$= \frac{\sigma^2_\Delta(x_i)}{B'_i\epsilon^2}.$$

Then

$$\lim_{B'_i\to\infty} \mathrm{P}\big(|\widehat{\mu}(\Delta, B'_i) - \mu_\Delta(x_i)| > \epsilon | X_i = x_i\big) \leq \lim_{B'_i\to\infty} \frac{\sigma^2_\Delta(x_i)}{B'_i\epsilon^2}$$
$$= 0.$$

∎

**Lemma 6.4.2** *The bootstrap sample variance $\widehat{\sigma}^2(\Delta, B'_i)$, is a weakly consistent estimator of $\sigma^2_\Delta(x_i)$, that is*

$$\lim_{B'_i\to\infty} P\big(|\widehat{\sigma}^2(\Delta, B'_i) - \sigma^2_\Delta(x_i)| > \epsilon | X_i = x_i\big) = 0$$

*for any $\epsilon > 0$.*

**Proof:** The bootstrap sample variance $\widehat{\sigma}^2(\Delta, B'_i)$ in (6.3.5) is

$$\begin{aligned}
\widehat{\sigma}^2(\Delta, B'_i) &= \frac{1}{B'_i - 1}\sum_{b=1}^{B'_i}\big(\widehat{\psi}_i(\underline{z}^\Delta_{i,b}) - \widehat{\mu}(\Delta, B'_i)\big)^2 \\
&= \frac{1}{B'_i - 1}\sum_{b=1}^{B'_i}\big(\widehat{\psi}_i(\underline{z}^\Delta_{i,b}) - \mu_\Delta(x_i)\big)^2 \\
&- \frac{B'_i}{B'_i - 1}\big(\widehat{\mu}_\Delta(x_i, B) - \mu_\Delta(x_i)\big)^2 \qquad (6.4.1) \\
&= \frac{B'_i}{B'_i - 1}\big[\frac{1}{B'_i}\sum_{b=1}^{B'_i}\big(\widehat{\psi}_i(\underline{z}^\Delta_{i,b}) - \mu_\Delta(x_i)\big)^2\big] \\
&- \frac{B'_i}{B'_i - 1}\big(\widehat{\mu}(\Delta, B'_i) - \mu_\Delta(x_i)\big)^2.
\end{aligned}$$

Then $\widehat{\sigma}^2(\Delta, B_i')$ is an unbiased estimator of $\sigma_\Delta^2(x_i)$ since

$$\begin{aligned} E\big(\widehat{\sigma}^2(\Delta, B_i')|X_i = x_i\big) &= \frac{B_i'}{B_i'-1}\frac{1}{B_i'}\sum_{b=1}^{B_i'} E\big[\big(\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta}) - \mu_\Delta(x_i)\big)^2|X_i = x_i\big] \\ &\quad - \frac{B_i'}{B_i'-1}E\big[\big(\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)\big)^2|X_i = x_i\big] \\ &= \frac{B_i'}{B_i'-1}\sigma_\Delta^2(x_i) - \frac{B_i'}{B_i'-1}\frac{\sigma_\Delta^2(x_i)}{B_i'} \\ &= \sigma_\Delta^2(x_i). \end{aligned}$$

For any $\epsilon > 0$, from Lemma 6.4.1, the second term in (6.4.1) converges to zero by applying Markov's inequality,

$$\begin{aligned} \lim_{B_i'\to\infty} \mathrm{P}\big((\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i))^2 > \epsilon|X_i = x_i\big) \leq & \\ \lim_{B_i'\to\infty} \frac{E\big[(\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i))^2|X_i = x_i\big]}{\epsilon} & \\ = \lim_{B_i'\to\infty} \frac{\sigma_\Delta^2(x_i)}{B_i'\epsilon} & \\ = 0. & \end{aligned}$$

On the other hand, since random variables $\{\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta})\}$, $b = 1, ..., B_i'$ are independent and identically distributed, we get

$$E\big[\big(\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta}) - \mu_\Delta(x_i)\big)^2|X_i = x_i\big] = \sigma_\Delta^2(x_i) < \infty.$$

Applying Khinchine's Theorem to the first term in (6.4.1),

$$\lim_{B_i'\to\infty} \mathrm{P}(|\frac{1}{B_i'}\sum_{b=1}^{B_i'}\big(\widehat{\psi}_i(\underline{z}_{i,b}^{\Delta}) - \mu_\Delta(x_i)\big)^2 - \sigma_\Delta^2(x_i)| > \epsilon|X_i = x_i) = 0.$$

Therefore

$$\lim_{B_i' \to \infty} \mathrm{P}\big(|\widehat{\sigma}^2(\Delta, B_i') - \sigma_\Delta^2(x_i)| > \epsilon | X_i = x_i\big) = 0.$$

∎

The preceding lemmas demonstrate the weak consistency of the bootstrap estimators. The main concern now is related to the prior probability estimator, $\widehat{\mu}(\Delta_0, B_i')$ in (6.3.6) which depends on $\Delta_0$. An appropriate value of $\Delta_0$ may lead us to get a weakly consistent estimator for $\pi_0(x_i)$. First we assume a biological meaningful function for the prior probability $\pi_0(X_i)$. Then, we concentrate on situations where the prior probability estimator can be a weakly consistent estimator. Suppose that $\pi_0(X_i)$, $i = 1, ..., N$ is a step function defined,

$$\pi_0(X_i) = \begin{cases} \pi_{01} & \text{if } X_i \leq x_0 \\ \pi_{02} & \text{if } X_i > x_0 \end{cases} \tag{6.4.2}$$

where $x_0 \in [x_{(1)}, x_{(N)}]$, $x_{(1)} = \min\{x_1, ..., x_N\}$ , and $x_{(N)} = \max\{x_1, ..., x_N\}$. The parameters $\pi_{01}$ and $\pi_{02}$ are both unknown but we assume $0 \leq \pi_{01} < \pi_{02} \leq 1$. Such a function divides the $N$ tests into two distinct classes such that in each class, the test statistics are identically distributed. Consequently, the mixture density of statistic $Z_i$ conditional on $X_i = x_i$ is given by,

$$f(z_i; \pi_0(x_i)) = \begin{cases} f(z_i; \pi_{01}) & \text{if } X_i \leq x_0 \\ f(z_i; \pi_{02}) & \text{if } X_i > x_0 \end{cases}$$

For given $x_0$ and $\Delta_0$, the observed co-variate vector $\underline{x}$ may be partitioned into

three regions; $\mathcal{R}_1(x_0, \Delta_0)$, $\mathcal{R}_2(x_0, \Delta_0)$, and $\mathcal{R}_3(x_0, \Delta_0)$ given by

$$\begin{aligned}
\mathcal{R}_1(x_0, \Delta_0) &= \{x_i;\ x_i \leq x_0 - \Delta_0\ ,\ i = 1, ..., N\}, \\
\mathcal{R}_2(x_0, \Delta_0) &= \{x_i;\ x_0 - \Delta_0 < x_i < x_0 + \Delta_0\ ,\ i = 1, ..., N\}, \\
\mathcal{R}_3(x_0, \Delta_0) &= \{x_i;\ x_i \geq x_0 + \Delta_0\ ,\ i = 1, ..., N\}.
\end{aligned} \tag{6.4.3}$$

Hence, the expectation of the LFDR is

$$E\big(\psi(z_i; x_i)|X_i = x_i\big) = \begin{cases} \pi_{01} & \text{if } x_i \in \mathcal{R}_1(x_0, \Delta_0) \\ \pi_{01} \text{ or } \pi_{02} & \text{if } x_i \in \mathcal{R}_2(x_0, \Delta_0) \\ \pi_{02} & \text{if } x_i \in \mathcal{R}_3(x_0, \Delta_0) \end{cases} \tag{6.4.4}$$

Under regions $\mathcal{R}_i(x_0, \Delta_0)$ for $i = 1, 3$, the bootstrap estimator $\widehat{\mu}(\Delta_0, B_i')$ is a weakly consistent estimator of the prior probability $\pi_0(x_i)$. The following lemma demonstrates the consistency of the bootstrap estimator $\widehat{\mu}(\Delta_0, B_i')$ when both $B_i'$ and $N$ get large.

**Lemma 6.4.3** *For $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, bootstrap estimator $\widehat{\mu}(\Delta_0, B_i')$ is a weakly consistent estimator of $\pi_{01}$ that is*

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} P\big(|\widehat{\mu}(\Delta_0, B_i') - \pi_{01}| > \epsilon|X_i = x_i\big) = 0$$

*for any $\epsilon > 0$.*

**Proof:** By applying *Markov*'s inequality, for any $\epsilon > 0$ we have

$$\begin{aligned}
\mathrm{P}\big(|\widehat{\mu}(\Delta_0, B_i') - \pi_{01}| > \epsilon|X_i = x_i\big) &\leq \frac{E\big[|\widehat{\mu}(\Delta_0, B_i') - \pi_{01}|\big|X_i = x_i\big]}{\epsilon} \\
&\leq \frac{E\big[|\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i)|\big|X_i = x_i\big]}{\epsilon} \\
&+ \frac{E\big[|\mu_{\Delta_0}(x_i) - \pi_{01}|\big|X_i = x_i\big]}{\epsilon}.
\end{aligned} \tag{6.4.5}$$

For given $\Delta_0$, $\widehat{\mu}(\Delta_0, B_i')$ is an unbiased estimator of $\mu_{\Delta_0}(x_i)$ with variance

$$\lim_{B_i' \to \infty} E\big[(\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i))^2 | X_i = x_i\big] = \lim_{B_i' \to \infty} \frac{\sigma^2_{\Delta_0}(x_i)}{B_i'} = 0.$$

By *Holder*'s inequality,

$$E\big[|\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i)|\big|X_i = x_i\big] \leq \big[E(|\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i)|^2\big|X_i = x_i)\big]^{\frac{1}{2}}$$

which implies,

$$\lim_{B_i' \to \infty} E\big[|\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i)|\big|X_i = x_i\big] = 0.$$

On the other hand, when $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, the probability $\mathrm{P}(|X_j - x_i| \leq \Delta_0; j = 1, ..., N) > 0$ and hence, the expected dimension of the reference class $\underline{z}_i^{\Delta_0}$ in (6.2.5) as $N \to \infty$ is

$$\lim_{N \to \infty} d_i^{\Delta_0} = \lim_{N \to \infty} N\mathrm{P}(|X_j - x_i| \leq \Delta_0; j = 1, ..., N) = \infty.$$

By applying the consistency assumption of $\widehat{\psi}_i(\bullet)$ in Section 6.2.1 on the reference class $\underline{z}_i^{\Delta_0}$,

$$\lim_{N \to \infty} \mathrm{P}\big(|\widehat{\psi}_i(\underline{z}_i^{\Delta_0}) - \psi(z_i; x_i)| > \epsilon | X_i = x_i\big) = 0.$$

Also $|\widehat{\psi}_i(\underline{z}_i^{\Delta_0}) - \psi(z_i; x_i)| \leq 1$, and by the *dominated convergence* Theorem,

$$\lim_{N \to \infty} E\big[\widehat{\psi}_i(\underline{z}_i^{\Delta_0}) - \psi(z_i; x_i)\big|X_i = x_i\big] = 0.$$

Then the second term in (6.4.5) converges to

$$\lim_{N \to \infty} \frac{E\big[|\mu_{\Delta_0}(x_i) - \pi_{01}|\big|X_i = x_i\big]}{\epsilon} = \lim_{N \to \infty} \frac{|\mu_{\Delta_0}(x_i) - \pi_{01}|}{\epsilon} = 0$$

when from (6.4.4), $E(\psi(z_i, x_i)|X_i = x_i) = \pi_{01}$. ∎

Similarly for $x_i \in \mathcal{R}_3(x_0, \Delta_0)$ and any $\epsilon > 0$,

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \mathrm{P}\big(|\widehat{\mu}(\Delta_0, B_i') - \pi_{02}| > \epsilon | X_i = x_i\big) = 0. \tag{6.4.6}$$

Lemma 6.4.3 and equation (6.4.6) can be used next to prove the weak consistency of the bootstrap estimator $\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i')$.

**Lemma 6.4.4** *If $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, the bootstrap estimator $\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i')$ is a weakly consistent estimator of the prediction bias $\mathcal{B}_\Delta(x_i)$, that is*

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} P(|\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i') - \mathcal{B}_\Delta(x_i)| > \epsilon | X_i = x_i) = 0$$

*for any $\epsilon > 0$.*

**Proof:** By Markov's inequality, for any $\epsilon > 0$ we have

$$\begin{aligned}
\mathrm{P}(|\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i') - \mathcal{B}_\Delta(x_i)| > \epsilon | X_i = x_i) &\le \frac{E\big[|\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i') - \mathcal{B}_\Delta(x_i)| \big| X_i = x_i\big]}{\epsilon} \\
&\le \frac{E\big[|\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)| \big| X_i = x_i\big]}{\epsilon} \\
&+ \frac{E\big[|\widehat{\mu}(\Delta_0, B_i') - \pi_{01}| \big| X_i = x_i\big]}{\epsilon}.
\end{aligned}$$

For given $\Delta_0$, from Lemma 6.4.1 the bootstrap estimator $\widehat{\mu}(\Delta, B_i')$ is an unbiased estimator of $\mu_\Delta(x_i)$ with variance

$$\lim_{B_i'\to\infty} E\big[(\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i))^2 | \big| X_i = x_i\big] = \lim_{B_i'\to\infty} \frac{\sigma_\Delta^2(x_i)}{B_i'} = 0.$$

By Holder's inequality,

$$E\big[|\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i)| \big| X_i = x_i\big] \le \big[E(|\widehat{\mu}(\Delta_0, B_i') - \mu_{\Delta_0}(x_i)|^2 \big| X_i = x_i)\big]^{\frac{1}{2}},$$

Thus it is concluded,

$$\lim_{B'\to\infty} E\big[|\widehat{\mu}(\Delta, B_i') - \mu_\Delta(x_i)|\big|X_i = x_i\big] = 0.$$

On the other hand, from Lemma 6.4.3 when $x_i \in \mathcal{R}_1(x_0, \Delta_0)$,

$$\lim_{N\to\infty}\lim_{B_i'\to\infty} \mathrm{P}\big(|\widehat{\mu}(\Delta_0, B_i') - \pi_{01}| > \epsilon|X_i = x_i\big) = 0,$$

and since $|\widehat{\mu}(\Delta_0, B_i') - \pi_0(x_i)| \leq 1$, by the dominated convergence Theorem,

$$\lim_{N\to\infty}\lim_{B_i'\to\infty} E\big[|\widehat{\mu}(\Delta_0, B_i') - \pi_0(x_i)|X_i = x_i\big] = 0.$$

∎

Similarly, a weakly consistent result can also be concluded for $x_i \in \mathcal{R}_3(x_0, \Delta_0)$. The weak consistency of $\widehat{\Delta}_{0i}^\star$ as the bootstrap estimator of $\Delta^\star(x_i)$ is presented in the following lemma.

**Lemma 6.4.5** *For $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, the bootstrap estimator $\widehat{\Delta}_{0i}^\star$ is a weakly consistent estimator of $\Delta^\star(x_i)$, that is*

$$\lim_{N\to\infty}\lim_{B_i'\to\infty} P\big(|\widehat{\Delta}_{0i}^\star - \Delta^\star(x_i)| > \epsilon|X_i = x_i\big) = 0$$

*for any $\epsilon > 0$.*

**Proof:** Recall from Lemma 6.4.2,

$$\lim_{B_i'\to\infty} \mathrm{P}\big(|\widehat{\sigma}^2(\Delta, B_i') - \sigma_\Delta^2(x_i)| > \epsilon|X_i = x_i\big) = 0,$$

and from Lemma 6.4.4, when $x_i \in \mathcal{R}_1(x_0, \Delta_0)$,

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \mathrm{P}(|\widehat{\mathcal{B}}(\Delta, \Delta_0, B_i') - \mathcal{B}_\Delta(x_i)| > \epsilon | X_i = x_i) = 0.$$

Thus the bootstrap estimator $\widehat{\mathrm{err}}(\Delta, \Delta_0, B_i')$ is a weakly consistent estimator of $\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big)$ in (6.3.2),

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \mathrm{P}\big(|\widehat{\mathrm{err}}(\Delta, \Delta_0, B_i') - \mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big)| > \epsilon | X_i = x_i\big) = 0.$$

By the continuous mapping Theorem,

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \mathrm{P}\big(|\arg\inf_{\Delta\geq\Delta_0} \widehat{\mathrm{err}}(\Delta, \Delta_0, B_i') - \arg\inf_{\Delta\geq\delta_0} \mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big)| > \epsilon | X_i = x_i\big) = 0$$

which implies

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} P\big(|\widehat{\Delta}_{0i}^{\star} - \Delta^{\star}(x_i)| > \epsilon | X_i = x_i\big) = 0.$$

∎

A similar result holds when $x_i \in \mathcal{R}_3(x_0, \Delta_0)$.

**Corollary 6.4.1** *If $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, then*

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \big[err\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}})|X_i = x_i\big) - err\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^{\star}(x_i)})|X_i = x_i\big)\big] = 0.$$

**Proof:** From Lemma 6.4.5, we have

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \mathrm{P}\big(|\widehat{\Delta}_{0i}^{\star} - \Delta^{\star}(x_i)| > \epsilon | X_i = x_i\big) = 0 \tag{6.4.7}$$

for any $\epsilon > 0$. Then by the continuous mapping Theorem,

$$\lim_{N\to\infty}\lim_{B'\to\infty} \mathrm{P}\big(|\widehat{\psi}(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}}) - \widehat{\psi}(\underline{z}_i^{\Delta^{\star}(x_i)})| > \epsilon \big| X_i = x_i\big) = 0. \tag{6.4.8}$$

By dominated convergence Theorem,

$$\lim_{N\to\infty}\lim_{B'\to\infty} E\big[|\widehat{\psi}(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}}) - \widehat{\psi}(\underline{z}_i^{\Delta^{\star}(x_i)})|\big| X_i = x_i\big] = 0 \tag{6.4.9}$$

which implies

$$\lim_{N\to\infty}\lim_{B'_i\to\infty} \big[\mathrm{Var}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i = x_i\big) - \mathrm{Var}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^{\star}(x_i)})|X_i = x_i\big)\big] = 0. \tag{6.4.10}$$

For $x_i \in \mathcal{R}_1(x_0, \Delta_0)$,

$$\begin{aligned}
&\lim_{N\to\infty}\lim_{B'_i\to\infty} \big[\mathrm{Var}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i = x_i\big) - \mathrm{Var}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^{\star}(x_i)})|X_i = x_i\big)\big] = \\
&\lim_{N\to\infty}\lim_{B'_i\to\infty} \big[E\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i = x_i\big) - E\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^{\star}(x_i)})|X_i = x_i\big)\big] \\
&\qquad\qquad \times \\
&\lim_{N\to\infty}\lim_{B'_i\to\infty} \big[E\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i = x_i\big) + E\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^{\star}(x_i)})|X_i = x_i\big) \\
&\qquad -2E\big(\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i = x_i\big)\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^{\star}(x_i)})|X_i = x_i\big)\big)\big].
\end{aligned} \tag{6.4.11}$$

From (6.4.9). ∎

A similar result holds when $x_i \in \mathcal{R}_3(x_0, \Delta_0)$.

**Lemma 6.4.6** *If* $x_i \in \mathcal{R}_1(x_0, \Delta_0)$,

$$\lim_{N\to\infty} EPE\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) = 0.$$

**Proof:** We have seem that the expected dimension of $\underline{z}_i^{\Delta}$ in (6.2.5) is $\lim_{N\to\infty} d_i^{\Delta} = \infty$. The consistency assumptions in Section 6.2.1 can now be applied to indicate the weak consistency of $\widehat{\psi}_i(\underline{z}_i^{\Delta})$,

$$\lim_{N\to\infty} \mathrm{P}(|\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)| > \epsilon | X_i = x_i) = 0 \tag{6.4.12}$$

for any $\epsilon > 0$. From (6.4.12), by applying the continuous mapping Theorem,

$$\lim_{N\to\infty} \mathrm{P}\big((\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i))^2 > \epsilon | X_i = x_i\big) = 0. \tag{6.4.13}$$

Since $|\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)| \leq 1$, by dominated convergence Theorem,

$$\lim_{N\to\infty} E\big[\big(\widehat{\psi}_i(\underline{z}_i^{\Delta}) - \psi(z_i; x_i)\big)^2 | X_i = x_i\big] = 0. \tag{6.4.14}$$

From Lemma 6.3.2, (6.4.14) implies

$$\lim_{N\to\infty} \mathrm{EPE}(\widehat{\psi}_i(\underline{z}_i^{\Delta}) | X_i = x_i) = 0.$$

∎

The same result can be obtained for $x_i \in \mathcal{R}_3(x_0, \Delta_0)$. The next theorem compares the performances of the two estimators, $\widehat{\psi}_i(\underline{z})$ and $\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}})$. Such comparisons depend on the regions $\mathcal{R}_1(x_0, \Delta_0)$ and $\mathcal{R}_3(x_0, \Delta_0)$.

**Theorem 6.4.1** *If* $x_i \in \mathcal{R}_1(x_0, \Delta_0)$,

$$\lim_{N\to\infty} \lim_{B_i'\to\infty} \big[EPE\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}_{0i}^{\star}}) | \mathcal{R}_1(x_0, \Delta_0)\big) - EPE\big(\widehat{\psi}_i(\underline{z}) | \mathcal{R}_1(x_0, \Delta_0)\big)\big] \leq 0.$$

**Proof:**

The difference between $\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big)$ and $\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big)$ can be written as,

$$\begin{aligned}\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) - \mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big) = \big[&\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) \\ &- \mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^\star(x_i)})|X_i = x_i\big)\big] \\ &+ \big[\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^\star(x_i)})|X_i = x_i\big) \\ &- \mathrm{err}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big)\big].\end{aligned} \tag{6.4.15}$$

From Corollary 6.4.1, the weak consistency of the bootstrap estimator $\widehat{\Delta}^\star_{0i}$ implies,

$$\lim_{N\to\infty}\lim_{B_i'\to\infty}\big[\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) - \mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^\star(x_i)})|X_i = x_i\big)\big] = 0. \tag{6.4.16}$$

Therefore, both (6.4.15) and (6.4.16) indicate,

$$\begin{aligned}\lim_{N\to\infty}\lim_{B'\to\infty}\big[\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) - \mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big)\big] = \\ \lim_{N\to\infty}\lim_{B'\to\infty}\big[\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^\star(x_i)})|X_i = x_i\big) - \mathrm{err}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big)\big].\end{aligned} \tag{6.4.17}$$

From definition 6.3.4,

$$\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta^\star(x_i)})|X_i = x_i\big) \leq \mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big), \tag{6.4.18}$$

for all $\Delta \geq \Delta_0$, which implies (6.4.15) as

$$\begin{aligned}\lim_{N\to\infty}\lim_{B'\to\infty}\big[\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) - \mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big)\big] \leq \\ \lim_{N\to\infty}\lim_{B'\to\infty}\big[\mathrm{err}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i = x_i\big) - \mathrm{err}\big(\widehat{\psi}_i(\underline{z})|X_i = x_i\big)\big],\end{aligned} \tag{6.4.19}$$

for all $\Delta \geq \Delta_0$. By applying both Lemma 6.3.2 and (6.3.2) we can get,

$$\begin{aligned}\lim_{N\to\infty}\lim_{B'\to\infty}\big[\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star(x_i)})|X_i=x_i\big)-\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i=x_i\big)\big]\leq\\ \lim_{N\to\infty}\lim_{B'\to\infty}\big[\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\Delta})|X_i=x_i\big)-\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i=x_i\big)\big].\end{aligned} \tag{6.4.20}$$

the term $\mathrm{Var}(\psi(z_i;x_i))$ was added and subtracted at the right side of equation (6.4.19)). From Lemma 6.4.6, the right side in (6.4.20) becomes zero,

$$\lim_{N\to\infty}\lim_{B'\to\infty}\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i=x_i\big)-\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i=x_i\big)\leq 0. \tag{6.4.21}$$

The expected prediction error of the estimators $\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})$ and $\widehat{\psi}_i(\underline{z})$ conditional on the region $\mathcal{R}_1(x_0,\Delta_0)$ are given by,

$$\begin{aligned}\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|\mathcal{R}_1(x_0,\Delta_0)\big)&=\int_{x_i\in\mathcal{R}_1(x_0,\Delta_0)}\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i=x_i\big)\,dP_x(x_i),\\ \mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|\mathcal{R}_1(x_0,\Delta_0)\big)&=\int_{x_i\in\mathcal{R}_1(x_0,\Delta_0)}\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|X_i=x_i\big)\,dP_x(x_i).\end{aligned} \tag{6.4.22}$$

Both (6.4.21) and (6.4.22) show,

$$\lim_{N\to\infty}\lim_{B'_i\to\infty}\big[\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|\mathcal{R}_1(x_0,\Delta_0)\big)-\mathrm{EPE}\big(\widehat{\psi}_i(\underline{z})|\mathcal{R}_1(x_0,\Delta_0)\big)\big]\leq 0.$$

∎

Following Theorem 6.4.1, for $x_i\in\mathcal{R}_3(x_0,\Delta_0)$, a similar result is concluded.

## 6.5 Application

We now use the CAD data described in Sections 2.2 and apply our methods to identify the disease-associated SNPs. We consider the CRC and ARC methods in estimating

the LFDR in (6.1.3). We computed the Wald $\chi^2$ test statistics for all $N = 394,838$ SNPs. From Section 5.3, we incorporate MAF (i.e. the scientific structure of each test) as a co-variate. The histogram of MAF in Figure 6.1 shows that MAF does not follow a uniform distribution. Thus, instead of working with MAF, we consider the empirical distribution of MAF. We transformed the observed Wald test statistics into $z$-values. Each measured SNP is denoted by a pair $(z_i, \text{MAF}_i)$. From the results in Section 4.5, the central region of the histogram of $z$-values matches the standard normal distribution very well. For each co-variate $x_i$, we apply SMM as the function $\widehat{\psi}_i(\bullet)$. We consider all measured SNPs and compute $\widehat{\psi}_i(\underline{z})$ under the CRC method. For the ARC method, set $\Delta_0 = 0.001$. From Figure 6.2, the total number of the disease-associated SNPs under the CRC method is 44, while under the ARC method 160 SNPs are associated with disease.

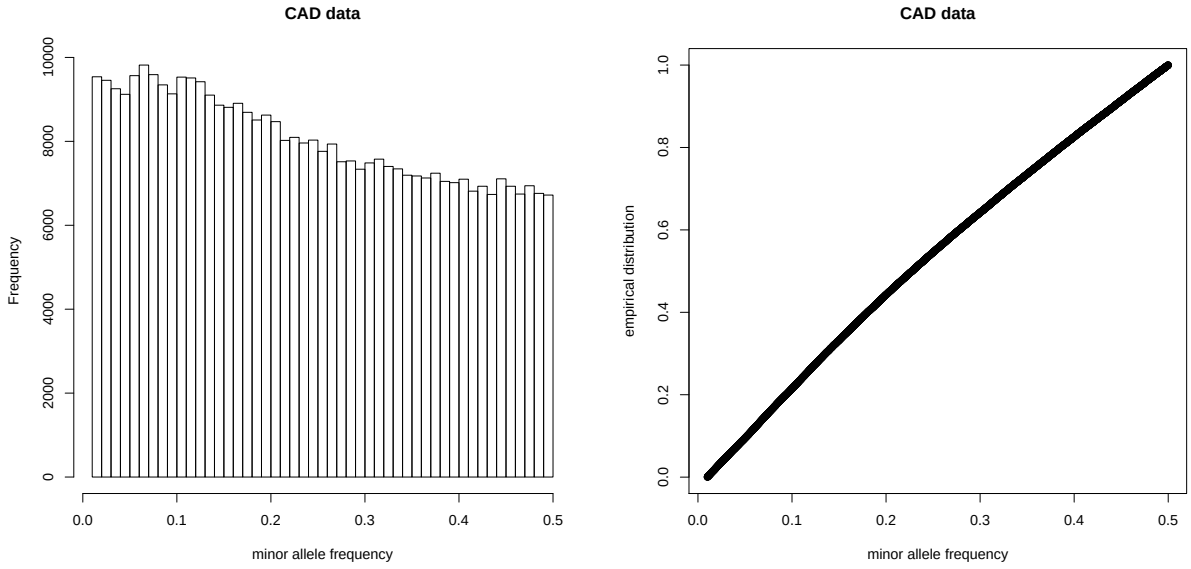



Figure 6.1: Distribution of $N = 394,838$ minor allele frequencies, the left panel shows the histogram and the right panel shows the empirical distribution.

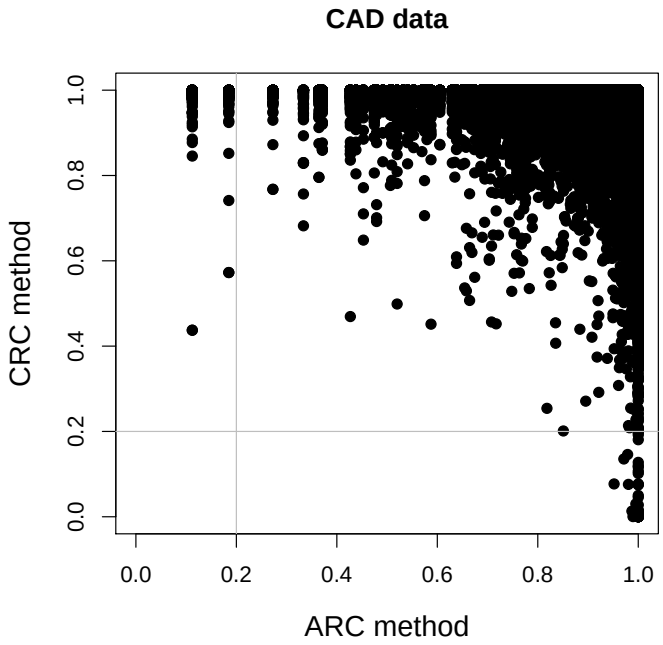



Figure 6.2: CAD data: the LFDR estimate for $N = 394,838$ measured SNPs according to the CRC method versus the ARC method. The vertical and horizontal lines show threshold 0.2.

The following process motivates the specification of the non-centrality $\delta$ for each simulation study to generate observed Wald $\chi^2$ test statistics. We considered the CAD data and computed the Wald $\chi^2$ test statistics for all $N = 394,838$ SNPs. Under PMM, we used constraints on the MLE of the non-centrality $\delta$ in (4.3.3). Thus, for given prior probability, the log-likelihood function in (4.3.16) can be simplified,

$$l(\delta) = \sum_{i=1}^{N} \log(\pi_0 g_0(t_i) + (1-\pi_0) g_{\delta_i}(t_i)), \tag{6.5.1}$$

is called the *profile likelihood*. Hence the prior probability $\pi_0$ can take any assigned true prior probabilities ranging from 0.60 to 0.999. From (6.5.1), we may derive numerically the MLE of $\widehat{\delta}$. Figure 6.3 shows the log likelihood values from (6.5.1) for true prior probabilities range from 0.60 to 0.999 versus non-centrality values range

$\delta \in (0, 40)$.

Table 6.1: Estimation of Parameter $\delta$ in PMM for the CAD data

| $\pi_0$ | 0.6 | 0.8 | 0.9 | 0.95 |
|---|---|---|---|---|
| $\widehat{\delta}$ | 0.4 | 0.7 | 1 | 1.32 |

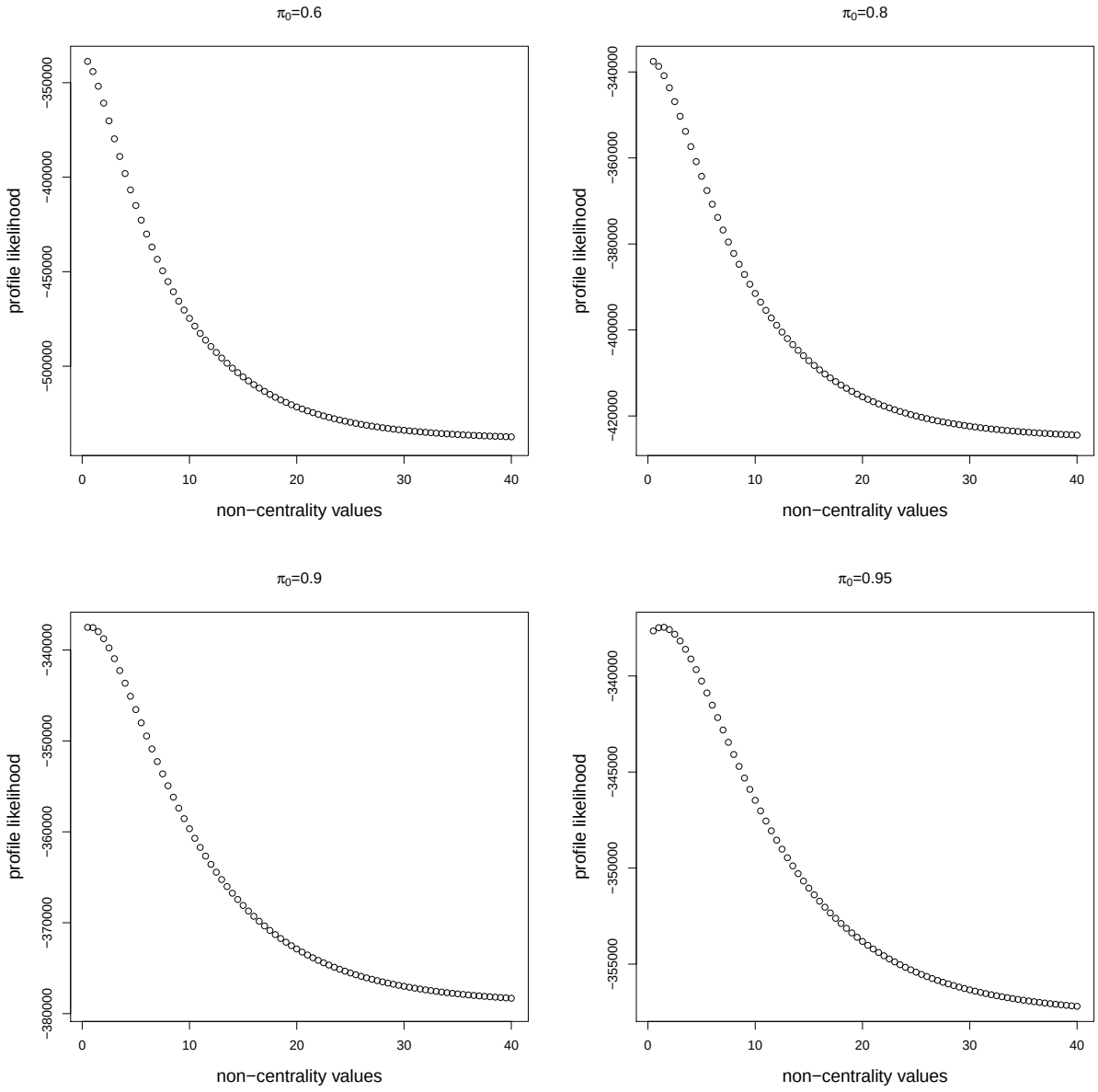



Figure 6.3: The profile likelihood values in (6.5.1) versus non-centrality values for true prior probability range from 0.60 to 0.95.

## 6.6 Simulation Study

The aim of the following simulation studies is to compare the abilities of the two estimation methods, the CRC and ARC methods, to estimate the LFDR. Such estimates lead us to determine the proportion of not disease-associated SNPs $p_0$ and the true association indicator $a_i$ defined in Section 4.3.

In this section, each SNP is assigned its prior probability which is a function of the scientific structure of each SNP, that is MAF. Thus $\pi_0(x_i)$ denotes the true prior probability that the $i^{\text{th}}$ SNP is not associated with disease. In application, the proportion of SNPs that are not associated with disease tends to be very large. Therefore, in simulated data we assumed the proportion of disease-associated SNPs to be small. In Section 4.3, according to (4.3.3), each non-centrality $\delta_i$ depends on the logarithm of the odds ratio, and SNPs with different values of the odds ratio do not necessarily have the same value of the non-centrality $\delta_i$. But in simulation studies, we assume that the observed Wald $\chi^2$ test statistics for the disease-associated SNPs are generated from the same distribution with the same value of $\delta_i = \delta$, whereas in Section 4.6, Yang et al. assumed that the observed Wald $\chi^2$ test statistics for the disease-associated SNPs were generated from different distributions with different values of $\delta_i$ [84]. For the non-associated SNPs, we assume that the observed Wald $\chi^2$ test statistics are from the same distribution with $\delta_i = 0$.

We considered several simulation studies each with a different value of $p_0$ ranging from 0.60 to 0.95. In each simulation study, we generated randomly 2000 data sets, each corresponding to an artificial case-control study. For each data set, we generated both co-variates and observed Wald $\chi^2$ test statistics simultaneously. Each data set has $N$ pairs $(t_i, x_i)$ with $N = 300000$. Each observed co-variate $x_i$ is generated randomly from the uniform distribution between 0 and 1.

In each simulation study, according to the assigned $p_0$, the true prior probability

$\pi_0(x_i)$ is determined as the function of observed co-variates,

$$\pi_0(x_i) = \begin{cases} 0 & \text{if } x_i \leq x_0 \\ 1 & \text{if } x_i > x_0 \end{cases} \tag{6.6.1}$$

where $x_0 = 1 - p_0$. Hence, $x_0$ of SNPs are associated with disease, while $(1 - x_0)$ of SNPs are not disease-associated.

For each of the measured SNPs, we compared a value that is generated from the uniform distribution between 0 and 1 with $\pi_0(x_i)$. If the generated value is greater than $\pi_0(x_i)$, let $A_i = 1$, and otherwise $A_i = 0$. To generate the observed $\chi^2$ test statistics, if $A_i = 1$, the measured SNP is assumed to be associated with disease. Then, the Wald $\chi^2$ test statistic for this measured SNP is sampled from a non-central $\chi^2$ distribution with one degree of freedom and an assigned non-centrality parameter $\delta$. From Table 6.1, an appropriate value for $\delta$ can be specified under (6.5.1). The Wald $\chi^2$ test statistics for the non-associated SNPs, when $A_i = 0$, were sampled from a central $\chi^2$ distribution with one degree of freedom. We transformed the observed Wald $\chi^2$ test statistics into $z$-values. The expected prediction error was considered to compare the performances of these two estimators.

In each simulation study, each data set contain $N$ pairs $(z_i, x_i)$. One pair is selected randomly from each data set. For given co-variate $x_i$, two estimators are computed. Under the ARC method, the $\Delta_0$ was specified in advance to determine $\widehat{\Delta}^{\star}_{0i}$. We considered the range $\Delta_0 \in (0, x_0)$ and $B = 1,000$. The estimators $\widehat{\psi}_i(\underline{z})$ and $\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^{\star}_{0i}})$ are computed.

Following (6.4.3), the number of pairs in the $r^{\text{th}}$ region, denoted $\mathcal{I}_r$, is given by

$$\mathcal{I}_r = \{x_i; x_i \in \mathcal{R}_r(x_0, \Delta_0)\} \quad , \quad r = 1, 2, 3$$

where $\mathcal{I} = \bigcup_{r=1}^{3} \mathcal{I}_r$ denotes the set of all chosen co-variates. According to (6.4.2), the

true LFDR $\psi(z_i; x_i)$ can be computed by direct substitution in (6.6.1),

$$\psi(z_i; x_i) = \begin{cases} 0 & \text{if } x_i \in \mathcal{R}_1(x_0, \Delta_0) \\ 0 \text{ or } 1 & \text{if } x_i \in \mathcal{R}_2(x_0, \Delta_0) \\ 1 & \text{if } x_i \in \mathcal{R}_3(x_0, \Delta_0) \end{cases} \tag{6.6.2}$$

We define the approximation of the expected prediction error given in Theorem 6.4.1 as

$$\begin{aligned} \widehat{\mathrm{EPE}}\big(\widehat{\psi}_i(\underline{z})|\mathcal{R}_r(x_0, \Delta_0)\big) &= \frac{1}{\#\{x_i \in \mathcal{I}_r\}} \sum_{x_i \in \mathcal{I}_r} \big(\widehat{\psi}_i(\underline{z}) - \psi(z_i; x_i)\big)^2, \\ \widehat{\mathrm{EPE}}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|\mathcal{R}_r(x_0, \Delta_0)\big) &= \frac{1}{\#\{x_i \in \mathcal{I}_r\}} \sum_{x_i \in \mathcal{I}_r} \big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}}) - \psi(z_i; x_i)\big)^2 \end{aligned} \tag{6.6.3}$$

for $r = 1, 2, 3$. The approximations of the expected prediction squared error conditional on each region are defined in (6.4.3) based on two estimation methods. The expected prediction squared error conditional on region $\mathcal{R}_1(x_0, \Delta_0)$ denotes the expected prediction squared error for disease-associated SNPs, while conditioning on region $\mathcal{R}_3(x_0, \Delta_0)$ represents the expected prediction squared error for non-associated SNPs. The following approximations indicate the approximation of the marginal expected prediction squared error according to two estimation methods,

$$\begin{aligned} \widehat{\mathrm{EPE}}\big(\widehat{\psi}_i(\underline{z})\big) &= \sum_{x_i \in \mathcal{I}} \widehat{\mathrm{EPE}}\big(\widehat{\psi}_i(\underline{z})|x_i \in \mathcal{I}\big), \\ \widehat{\mathrm{EPE}}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})\big) &= \sum_{x_i \in \mathcal{I}} \widehat{\mathrm{EPE}}\big(\widehat{\psi}_i(\underline{z}_i^{\widehat{\Delta}^\star_{0i}})|x_i \in \mathcal{I}\big). \end{aligned} \tag{6.6.4}$$

From Figures 6.4 to 6.7, we see the performances of the ARC method and the CRC method depend on the true values of $p_0$.

We indicate the situation where the results from Theorem 6.4.1 do not hold. Instead of considering different true prior probabilities $\pi_0(x_i)$ for each SNP, we assume all measured SNPs have the same prior probability $p_0$. For each of the measured

SNPs, we compared a value that is generated from the uniform distribution between 0 and 1 with $p_0$. If the generated value is greater than $p_0$, we assigned $A_i = 1$, and otherwise $A_i = 0$. To generate the observed $\chi^2$ test statistics, if $A_i = 1$, the measured SNP is assumed to be associated with disease. Then, the Wald $\chi^2$ test statistic for this measured SNP is sampled from a non-central $\chi^2$ distribution with one degree of freedom and an assigned non-centrality parameter $\delta$. From Table 6.1, an appropriate value for non-centrality $\delta$ can be specified. The value of this assigned non-centrality parameter $\delta$ is determined under the log-likelihood function in (6.5.1). The Wald $\chi^2$ test statistics for the non-associated SNPs, when $A_i = 0$, were sampled from a central $\chi^2$ distribution with one degree of freedom. We transformed the observed Wald $\chi^2$ test statistics into $z$-values under (3.1.3).

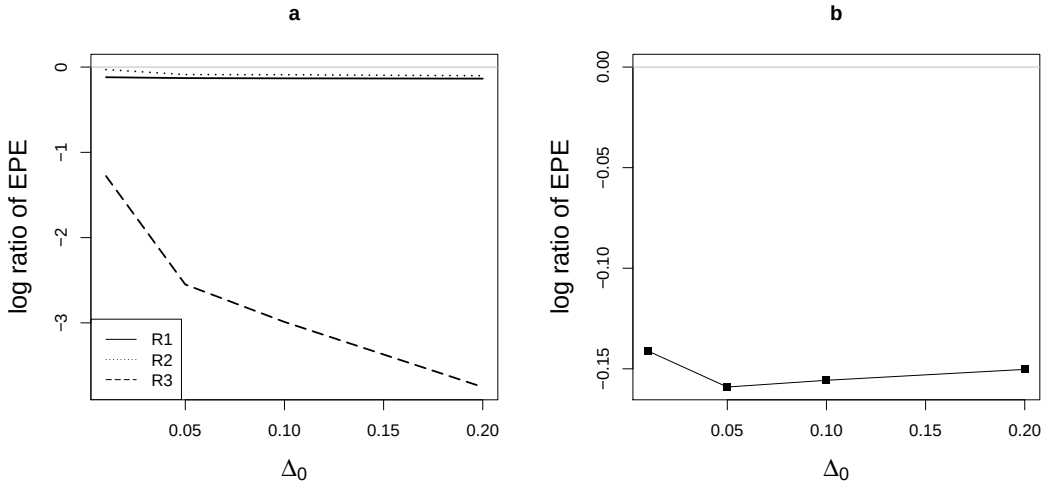



Figure 6.4: Binary logarithm ($\log_2$) of the ratio of the expected prediction squared error, the ARC method in numerator and the CRC method in denominator, versus $\Delta_0$ values for $p_0 = 0.60$. Left panel shows conditional on region in (6.6.3), and right panel shows marginal in (6.6.4).

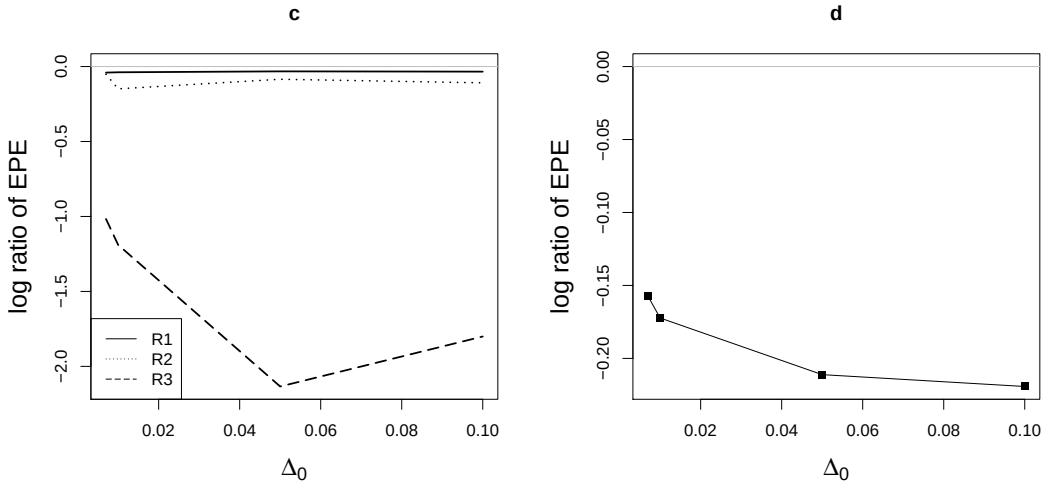



Figure 6.5: Binary logarithm ($\log_2$) of the ratio of the expected prediction squared error, the ARC method in numerator and the CRC method in denominator, versus $\Delta_0$ values for $p_0 = 0.80$. Left panel shows conditional on region in (6.6.3), and right panel shows marginal in (6.6.4)

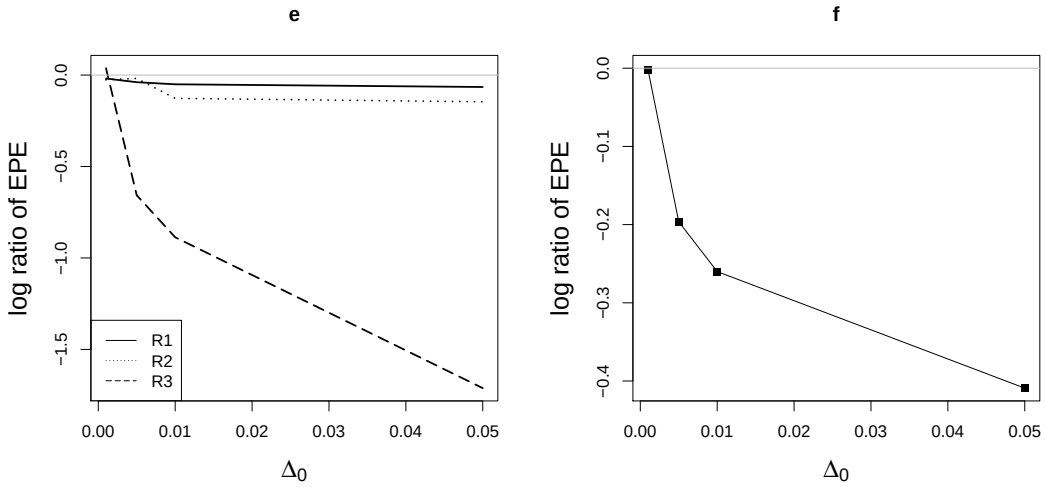



Figure 6.6: Binary logarithm ($\log_2$) of the ratio of the expected prediction squared error, the ARC method in numerator and the CRC method in denominator, versus $\Delta_0$ values for $p_0 = 0.90$. Left panel shows conditional on region in (6.6.3), and right panel shows marginal in (6.6.4).

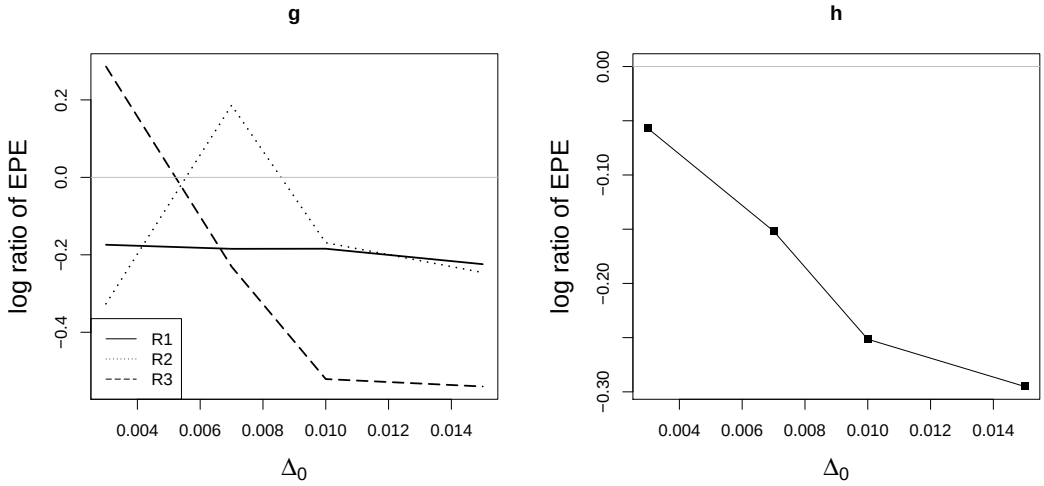



Figure 6.7: Binary logarithm ($\log_2$) of the ratio of the expected prediction squared error, the ARC method in numerator and the CRC method in denominator, versus $\Delta_0$ values for $p_0 = 0.95$. Left panel shows conditional on region in (6.6.3), and right panel shows marginal in (6.6.4).

If we consider the same true prior probability for all measured SNPs, that is the same as the two-groups model in Section 3.3, then from Figure 6.8 the CRC method has smaller expected prediction squared error compared with the ARC method. Such a result happens when all SNPs have the same mixture densities. Our proposed model does not seem to work well under that assumption. In such a case, instead of using the ARC method, the CRC method should be applied.

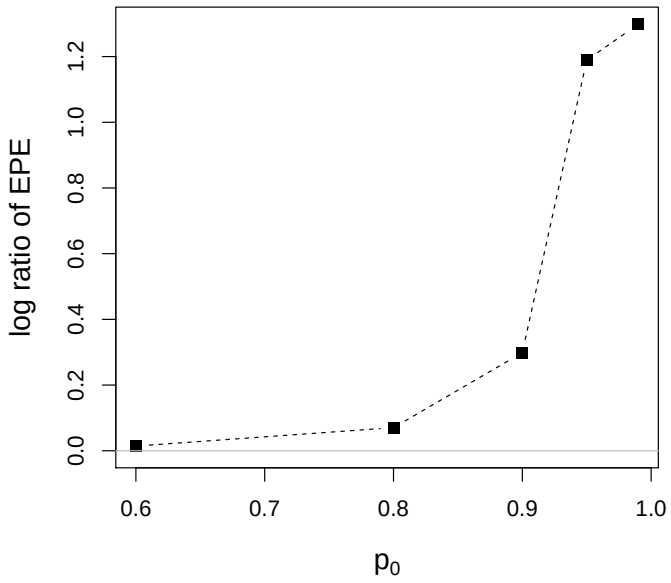



Figure 6.8: Binary logarithm ($\log_2$) of the ratio of the marginal expected prediction squared error in (6.6.4), the ARC method (in numerator) and the CRC method (in denominator), versus $p_0$ values.

## 6.7 Discussion and Conclusions

In the situation where the prior probability $\pi_0(X_i)$ is the step function in (6.4.2), the LFDR estimator under the ARC method performs better than the CRC method when N gets large. It would be interesting to demonstrate this result for a general prior $\pi_0(X_i)$. Our simulation results, confirm that, under regions $\mathcal{R}_1(x_0, \Delta_0)$ and $\mathcal{R}_3(x_0, \Delta_0)$, the LFDR estimator under the ARC method has smaller expected prediction squared error than under the CRC method. Under $\mathcal{R}_2(x_0, \Delta_0)$ the weak consistency of $\widehat{\mu}(\Delta_0, B)$ as an estimator of $\pi_0(x_i)$ is not clear. The ARC method was applied on the CAD data, by increasing the tuning parameter $\Delta_0$, the proportion of disease-associated SNPs decreases, approaching the proportion based on the CRC

method. This suggests that further investigations may be necessary to control the tuning parameter $\Delta_0$ to improve results.

# Chapter 7

# Future Works

Based on the research we have done, the following issues will be considered for further research.

In Chapter 6, the weak consistency of the LFDR estimator $\widehat{\psi}(\bullet)$ was assumed, while we can prove such consistency for the LFDR estimator under parametric mixture model, or semi-parametric mixture model.

Under the ARC method explained in Chapter 6, we would like to define any diagnostic statistics to assess the power of such method to identify the non-null features (i.e. SNPs, genes, voxels). In other words, we would like to examine the effect of the ARC method on power.

Under the independency of co-variates $X_1, X_2, ..., X_N$, it is easier to indicate that the ARC method has smaller error than the CRC method in Chapter 6, when the total number of tests gets large. If we ignore the independence assumption of the co-variates, then we would like to demonstrate similar results under the ARC method.

Suppose the prior probability $\pi_0(X_i)$ is the function of the co-variate $X_i$ given

by

$$\pi_0(X_i) = \begin{cases} h_1(X_i) & \text{if } X_i \leq x_0 \\ h_2(X_i) & \text{if } X_i > x_0 \end{cases} \tag{7.0.1}$$

where $x_0 \in [x_{(1)}, x_{(N)}]$, $x_{(1)} = \min\{x_1, ..., x_N\}$ and $x_{(N)} = \max\{x_1, ..., x_N\}$. The functions $h_1$ and $h_2$ are defined as

$$h_j(X_i) = \frac{\exp(\beta_{0j} + \beta_{1j}X_j)}{1 + \exp(\beta_{0j} + \beta_{1j}X_j)}\ , \quad j = 1, 2$$

where both parameters $\beta_{0j}$ and $\beta_{1j}$ are unknown. We would like to show that the ARC method has asymptotically smaller error than the CRC method under such an assumption for the prior probability.

We incorporated a co-variate for each test to improve the LFDR estimate. Then, we proposed the ARC method such that under some assumptions yields the LFDR estimator with less error compared with the CRC method. Now suppose $m > 1$ co-variates are incorporated for each test,

$$\begin{aligned} \underline{X}_1 &= (X_{11}, ..., X_{1N})^T \\ \underline{X}_2 &= (X_{21}, ..., X_{2N})^T \\ &\vdots \\ \underline{X}_m &= (X_{m1}, ..., X_{mN})^T \end{aligned} \tag{7.0.2}$$

Then the LFDR is the posterior probability that the $i^{\text{th}}$ null hypothesis is true given the data,

$$\psi(z_i; x_{1i}, x_{2i}, ..., x_{mi}) \equiv \mathrm{P}(A_i = 0 | Z_i = z_i, X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{mi} = x_{mi})$$

where $A_i$'s are identically distributed indicator random variables with Bernoulli distribution $A_i \sim \text{Bernoulli}(1 - \pi_0(x_{1i}, x_{2i}, ..., x_{mi}))$. We would like to improve the ARC

method in order to estimate the LFDR when more than one co-variate is incorporated.

# Bibliography


[1] D. B. Allison, G. L. Gadbury, M. Heo, J. R. Fernádez, C. Lee, T. A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 38:1–20, 2002.

[2] J. Aubert, A. Bar-Hen, J. Daudin, and S. Robin. Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, 5:125, 2004.

[3] Julie Aubert, Avner Bar-Hen, Jean-Jacques Daudin, and Stéphane Robin. Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics*, 5(1):125, 2004.

[4] M Madan Babu. Introduction to microarray data analysis. *Computational Genomics: Theory and Application*, pages 225–249, 2004.

[5] Jeffrey C Barrett and Lon R Cardon. Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6):659–662, 2006.

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

[7] Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[8] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[9] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188, 2001.

[10] José M Bernardo and Adrian FM Smith. *Bayesian Theory*, volume 405. John Wiley & Sons, 2009.

[11] D. R. Bickel. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. *Canadian Journal of Statistics*, 39:610–631, 2011.

[12] D. R. Bickel. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. *Canadian Journal of Statistics*, 39(4):610–631, 2011.

[13] D. R. Bickel. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. *International Statistical Review*, 81:188–206, 2013.

[14] D. R. Bickel. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. *International Statistical Review*, 81(2):188–206, 2013.

[15] D. R. Bickel. Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison. *International Statistical Review*, 82:457–476, 2014.

[16] Jeffrey D Blume. Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, 21(17):2563–2599, 2002.

[17] J. Bukszár, J. L. McClay, and E. J. C. G. van den Oord. Estimating the posterior probability that genome-wide association findings are true or false. *Bioinformatics*, 25:1807–1813, 2009.

[18] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, 2012.

[19] William S. Cleveland and Susan J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[20] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[21] Gayle K Deutsch, Robert F Dougherty, Roland Bammer, Wai Ting Siok, John DE Gabrieli, and Brian Wandell. Children's reading performance is correlated with white matter structure measured by diffusion tensor imaging. *Cortex*, 41(3):354–363, 2005.

[22] Anthony WF Edwards. Statistical methods in scientific inference. *Nature*, 222:1233–1237, 1969.

[23] Bradley Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465), 2004.

[24] Bradley Efron. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100(469):1–5, 2005.

[25] Bradley Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477), 2007.

[26] Bradley Efron. Size, power and false discovery rates. *The Annals of Statistics*, pages 1351–1377, 2007.

[27] Bradley Efron. Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics*, pages 197–223, 2008.

[28] Bradley Efron. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge University Press, 2010.

[29] Bradley Efron et al. Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23(1):1–22, 2008.

[30] Bradley Efron, John D Storey, and Robert Tibshirani. *Microarrays Empirical Bayes Methods, and False Discovery Rates*. Department of Statistics, Stanford University, 2001.

[31] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

[32] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

[33] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC Press, 1994.

[34] Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[35] Greg Gibson. Hints of hidden heritability in gwas. *Nature Genetics*, 42(7):558–560, 2010.

[36] Raphael Gottardo, Adrian E Raftery, KA Yee Yeung, and Roger E Bumgarner. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, 62(1):10–18, 2006.

[37] C. M. T. Greenwood, J. Rangrej, and L. Sun. Optimal selection of markers for validation or replication from genome-wide association studies. *Genetic Epidemiology*, 31:395–407, 2007.

[38] Ian Hacking. *Logic of Statistical Inference*. CUP Archive, 1965.

[39] Leland Hartwell, Leroy Hood, and Michael L Goldberg. *Genetics: from Genes to Genomes*. Granite Hill Publishers, 2008.

[40] Yi He, Wei Pan, and Jizhen Lin. Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational Statistics & Data Analysis*, 51(2):641–658, 2006.

[41] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[42] Yosef Hochberg and Ajit C Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., 1987.

[43] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, New York, 2000.

[44] James X Hu, Hongyu Zhao, and Harrison H Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491), 2010.

[45] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[46] William S Klug, Michael R Cummings, et al. *Concepts of Genetics*. Number Ed. 7. Pearson Education, Inc, 2003.

[47] L. Kruglyak. The road to genome-wide association studies. *Nature Reviews Genetics*, 9(4):314–318, 2008.

[48] Solomon Kullback. *Information Theory and Statistics*. Courier Corporation, 1997.

[49] Cathryn M Lewis. Genetic association studies: design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2):146–153, 2002.

[50] J. G. Liao, Y. Lin, Z. E. Selvanayagam, and W. J. Shih. A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, 20:2694–2701, 2004.

[51] Rupert G Miller. *Simultaneous Statistical Inference*. Springer, 1981.

[52] Giovanni Montana. Statistical methods in genetics. *Briefings in Bioinformatics*, 7(3):297–308, 2006.

[53] Omkar Muralidharan et al. An empirical bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 4(1):422–438, 2010.

[54] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*, volume 4. Irwin Chicago, 1996.

[55] Marta Padilla and D. R. Bickel. Estimators of the local false discovery rate designed for small numbers of tests. *Statistical Applications in Genetics and Molecular Biology*, 11(5):art. 4, 2012.

[56] W. Pan, J. Lin, and C. T. Le. A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics*, 3:117–124, 2003.

[57] Ju-Hyun Park, Sholom Wacholder, Mitchell H Gail, Ulrike Peters, Kevin B Jacobs, Stephen J Chanock, and Nilanjan Chatterjee. Estimation of effect size

distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, 42(7):570–575, 2010.

[58] S. Pounds and S. W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19:1236–1242, 2003.

[59] C Radhakrishna Rao. Some observations on multivariate statistical methods in anthropological research. *Bulletin of the International Statistical Institute*, 37(4):99–109, 1961.

[60] C Radhakrishna Rao. Problems of selection with restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 401–405, 1962.

[61] C Radhakrishna Rao. Use of discriminant and allied functions in multivariate analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 149–154, 1962.

[62] C Radhakrishna Rao. *Linear Statistical Inference and Its Applications*, volume 22. John Wiley & Sons, 2009.

[63] Kathryn Roeder and Larry Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 24(4):398, 2009.

[64] Richard Royall. *Statistical Evidence: A Likelihood Paradigm*, volume 71. CRC Press, 1997.

[65] Armin Schwartzman, Robert F Dougherty, and Jonathan E Taylor. Cross-subject comparison of principal diffusion direction maps. *Magnetic Resonance in Medicine*, 53(6):1423–1431, 2005.

[66] G. Shieh. On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference*, 128(1):43–59, 2005.

[67] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[68] Joan G Staniswalis. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283, 1989.

[69] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[70] John D Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 2013–2035, 2003.

[71] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[72] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[73] L. Sun, R. V. Craiu, A. D. Paterson, and S. B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30:519–530, 2006.

[74] Lei Sun, Radu V Craiu, Andrew D Paterson, and Shelley B Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, 2006.

[75] PC Sundgren, Q Dong, D Gomez-Hassan, SK Mukherji, P Maly, and R Welsh. Diffusion tensor imaging of the brain: review of clinical applications. *Neuroradiology*, 46(5):339–350, 2004.

[76] Robert Tibshirani and Trevor Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

[77] Angélique B van't Wout, Ginger K Lehrman, Svetlana A Mikheeva, Gemma C O'Keeffe, Michael G Katze, Roger E Bumgarner, Gary K Geiss, and James I Mullins. Cellular gene expression upon human immunodeficiency virus type 1 infection of cd4+-t-cell lines. *Journal of Virology*, 77(2):1392–1402, 2003.

[78] J. Wakefield. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, 81:208–227, 2007.

[79] Larry Wasserman and Kathryn Roeder. Weighted hypothesis testing. *arXiv preprint math/0604172*, 2006.

[80] Y. Wei, S. Wen, P. Chen, C. Wang, and C. K. Hsiao. A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies. *European Journal of Human Genetics*, pages 942–947, 2010.

[81] Yu-Chung Wei, Shu-Hui Wen, Pei-Chun Chen, Chih-Hao Wang, and Chuhsing K Hsiao. A simple bayesian mixture model with a hybrid procedure for genome-wide association studies. *European Journal of Human Genetics*, 18(8):942–947, 2010.

[82] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[83] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

[84] Ye Yang, Farnoosh Abbas Aghababazadeh, and David R Bickel. Parametric estimation of the local false discovery rate for identifying genetic associations. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 10(1):98–108, 2013.