ESTIMATING THE PROPORTION OF TRUE NULL HYPOTHESES IN MULTIPLE TESTING PROBLEMS

Oluyemi Oyeniran

A Dissertation

Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

June 2016

Committee:

Hanfeng Chen, Advisor

John Laird,
Graduate Faculty Representative

John Chen

Junfeng Shang
ABSTRACT

Hanfeng Chen, Advisor

The problem of estimating the proportion, \( \pi_0 \), of the true null hypotheses in a multiple testing problem is important in cases where large scale of parallel hypotheses tests are performed independently. While the problem is a quantity of interest in its own right in many applications, a reliable estimate of \( \pi_0 \) is crucial when we want to assess and/ or control the false discovery rate in a multiple testing problem.

In this dissertation, we investigate the estimation problem coupled with assessing/controlling the false discovery rate. The dissertation develops a new estimating procedure under the two-component mixture model. The components of the mixture are the null and alternative distributions with mixing proportions \( \pi_0 \) and \( 1 - \pi_0 \) respectively, where \( \pi_0 \) is the unknown proportion to be estimated. We establish an innovative non-parametric maximum likelihood estimation of the \( p \)-values density, restricting the alternative to multinomial distribution family of \( k \) categories to address this problem.

To apply this approach, we need to settle two things first: (a) select an integer \( k \), and (b) convert the continuous-type observations (\( p \)-values) into discrete data with \( k \) categories. As many authors have noticed, in applications, the \( p \)-values are highly skewed, so we recommend Sturges’ rule modified for skewness in determining \( k \).

We then propose an iterative optimization technique - EM algorithm to characterize the maximum likelihood estimate for an approximation to the maximum likelihood estimate of \( \hat{\pi}_0 \). Simulation studies are conducted to assess the performance of the proposed procedure. The simulation results show that our proposed procedure perform significantly better than the existing procedures. The new procedure is applied to the leukemia gene expression dataset and the inherited breast can-
cer cDNA dataset that were analyzed by many other statisticians. Again, our procedure provides an overall satisfactory performance.
– to abel modupe olujare olufumnilayo oyeniran
First and most importantly, I would like to express my deepest gratitude and respect toward my advisor Professor Hanfeng Chen. His guidance, advice, patience, encouragement and constant support through my PhD study into various statistical methods is the key factor to the success of this dissertation. Hanfeng is a brilliant and passionate statistician, and his insightful ideas in both statistical theories and applications inspire me. It is a great honor and privilege to have the opportunity to work closely and learn from him. In addition, I am grateful to my other committee members: Professor John Chen, Professor Junfeng Shang, and Professor John Laird, for their valued time, advice and encouragement.

My gratitude also goes to all of my professors in the Department of Mathematics and Statistics at Bowling Green State University. My solid background is built from receiving rigid training here in both theories and applications of probability and statistics, which are not only beneficial to my research but also priceless for my career. I am also very thankful to our department staff: Marcia Seubert, Mary Busdeker, and Barbara Berta for their constant help. Thanks to all my colleagues whose friendship has been priceless to my success in graduate school. I am forever grateful to the Department of Mathematics and Statistics at Bowling Green State University for their financial support throughout my PhD study.

Finally, I am mostly grateful to my family - parents, brothers, sister and friends for their support, never-ending love and patience. My special thanks to my wife, damilola, for her love, devotion, support and sacrifice.

Oluyemi Oyeniran
Bowling Green, OH
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER 1 INTRODUCTION</th>
<th>........................</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Multiple Testing</td>
<td>........................</td>
<td>1</td>
</tr>
<tr>
<td>1.2 False Discovery Rate Control</td>
<td>........................</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1 Error Rate Control</td>
<td>........................</td>
<td>5</td>
</tr>
<tr>
<td>1.2.2 Family Wise Error Rate</td>
<td>........................</td>
<td>6</td>
</tr>
<tr>
<td>1.2.3 False Discovery Rate (FDR)</td>
<td>........................</td>
<td>7</td>
</tr>
<tr>
<td>1.2.4 A Multiple Endpoint Analysis Example</td>
<td>........................</td>
<td>10</td>
</tr>
<tr>
<td>1.3 Structure of the Dissertation</td>
<td>........................</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 2 REVIEW OF EXISTING METHODS</th>
<th>........................</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Storey’s Method</td>
<td>........................</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Convest Method - Convex Decreasing Density Estimator</td>
<td>........................</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Average Estimate Approach</td>
<td>........................</td>
<td>19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 3 ESTIMATING $\pi_0$ IN MULTINOMIAL MODELS: AN INNOVATIVE NONPARAMETRIC APPROACH</th>
<th>........................</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Multinomial Model</td>
<td>........................</td>
<td>22</td>
</tr>
<tr>
<td>3.2 Approximating a distribution via Histogram</td>
<td>........................</td>
<td>23</td>
</tr>
<tr>
<td>3.2.1 The $L_2$ Theory of Histograms</td>
<td>........................</td>
<td>25</td>
</tr>
<tr>
<td>3.2.2 Sturges’ Rule</td>
<td>........................</td>
<td>29</td>
</tr>
<tr>
<td>3.3 Expectation Maximization (EM) Algorithm</td>
<td>........................</td>
<td>31</td>
</tr>
<tr>
<td>3.4 New estimating procedure for $\pi_0$</td>
<td>........................</td>
<td>36</td>
</tr>
<tr>
<td>3.4.1 Introduction</td>
<td>........................</td>
<td>36</td>
</tr>
<tr>
<td>3.4.2 Mixture Model Framework</td>
<td>........................</td>
<td>38</td>
</tr>
<tr>
<td>3.4.3 Estimating $\pi_0$ using EM Algorithm</td>
<td>........................</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 4 SIMULATION STUDIES</th>
<th>........................</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Testing Scenario</td>
<td>........................</td>
<td>46</td>
</tr>
</tbody>
</table>
4.2 Generation of simulated data .................................................. 46
4.3 Results of simulation study .................................................. 47

CHAPTER 5 REAL LIFE EXAMPLE: APPLICATION TO DNA MICROARRAY DATA . 56

5.1 Introduction ................................................................. 56
5.2 Microarray ................................................................. 57
   5.2.1 Background on Cells, Genes, DNAs, and RNAs ................. 57
   5.2.2 Gene Expression ..................................................... 59
   5.2.3 Microarray Technology ............................................. 62
5.3 Description of Microarray dataset ...................................... 66
   5.3.1 Microarray Data ..................................................... 67
   5.3.2 Leukemia Gene Expression Data .................................. 70
   5.3.3 Breast Cancer cDNA microarray data ............................ 71
5.4 Microarray Data Results ................................................ 76
   5.4.1 Leukemia Gene Expression Data Results ......................... 76
   5.4.2 Breast Cancer Gene Expression Data Results ................... 76

CHAPTER 6 SUMMARY AND CONCLUSION .................................. 79

6.1 Future Research ........................................................... 82

BIBLIOGRAPHY ................................................................. 83
**LIST OF FIGURES**

3.1 Histogram ............................................. 24

3.2 Graphical interpretation of a single iteration of the EM algorithm ............. 35

4.1 Histogram of simulated mixture dataset with true value $\pi_0 = 0.25$ for 1,000 independent observations ............................................. 48

4.2 Simulated mixture density with true value $\pi_0 = 0.25$ for 1,000 independent observations ............................................. 48

4.3 Histogram of simulated mixture dataset with true value $\pi_0 = 0.5$ for 1,000 independent observations ............................................. 49

4.4 Simulated mixture density with true value $\pi_0 = 0.5$ for 1,000 independent observations ............................................. 49

4.5 Histogram of simulated mixture dataset with true value of $\pi_0 = 0.75$ for 1,000 independent observations ............................................. 50

4.6 Simulated mixture density with true value $\pi_0 = 0.75$ for 1,000 independent observations ............................................. 50

4.7 The estimate, $\hat{\pi}_0$, of the proportion of true null hypotheses is compared for: the proposed new innovative estimating procedure, Storey’s $\hat{\pi}_0^S$, Langaas’s nonparametric maximum likelihood approach (\textit{convex}) $\hat{\pi}_0^C$, and average estimate approach with $B$ chosen via the bootstrapping procedure, $\hat{\pi}_0^A$. There are 1,000 simulated data sets, each with a total of $m=1,000$ hypotheses tests, for each value of $\pi_0$ ............................................. 54

4.8 Simulation results of the False Discovery Rate (FDR) estimates at significance level $\alpha = 0.05$ for the new procedure in comparison to the existing methods ............................................. 55

5.1 An outline of the microarray methodology adapted from Kumar et al. (2012) .............. 64

5.2 Histogram of the $p$-values for the leukemia gene expression data reported in Golub et al. (1999) ............................................. 72
5.3 Density function of the leukemia gene expression data reported in Golub et al. (1999) 72

5.4 Histogram of the $p$-values for the hereditary breast cancer gene expression data reported in Hedenfalk et al. (2001) 75

5.5 Density function of the hereditary breast cancer gene expression data reported in Hedenfalk et al. (2001) 75
LIST OF TABLES

1.1 Classification of $m$ null hypotheses tests ............................................. 4

4.1 The estimate, $\hat{\pi}_0$, of the proportion of true null hypotheses is compared for: the proposed new innovative estimating procedure $\hat{\pi}_0^n$, Storey’s $\hat{\pi}_0^s$, Langaas’s nonparametric maximum likelihood approach (\textit{convest}) $\hat{\pi}_0^c$, and average estimate approach with $B$ chosen via the bootstrapping procedure, $\hat{\pi}_0^a$. There are 1,000 simulated data sets, each with a total of $m=1,000$ hypotheses tests, for each value of $\pi_0$. .................................................. 52

4.2 The standard deviation of $\pi_0$ estimate of the proportion of true null hypotheses is compared for: the proposed new innovative estimating procedure $\hat{\pi}_0^n$, Storey’s $\hat{\pi}_0^s$, Langaas’s nonparametric maximum likelihood approach (\textit{convest}) $\hat{\pi}_0^c$, and average estimate approach with $B$ chosen via the bootstrapping procedure, $\hat{\pi}_0^a$. There are 1,000 simulated data sets, each with a total of $m=1,000$ hypotheses tests, for each value of $\pi_0$. .................................................. 53

5.1 The estimate of the proportion of true null hypotheses and the number of statistically significant genes for the leukemia gene expression data at significance level $\alpha = 0.05$ after applying BH’s adaptive FDR controlling procedure with $\pi_0$ estimated using the new innovative approach in comparison to other three methods. ............. 77

5.2 The estimate of the proportion of true null hypotheses and the number of statistically significant genes for the breast cancer gene expression data at significance level $\alpha = 0.05$ after applying BH’s adaptive FDR controlling procedure with $\pi_0$ estimated using the new innovative approach in comparison to other three methods. ............. 78
# LIST OF ALGORITHMS

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Simes’ procedure</td>
<td>8</td>
</tr>
<tr>
<td>1.2 BH procedure</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Adaptive BH Procedure</td>
<td>10</td>
</tr>
<tr>
<td>3.1 EM Iteration</td>
<td>45</td>
</tr>
</tbody>
</table>
LIST OF THEOREMS

3.1 Theorem ................................................................. 28
3.2 Theorem ................................................................. 37
3.3 Theorem ................................................................. 43
CHAPTER 1 INTRODUCTION

In this chapter, we briefly describe the main concepts involving in this dissertation, such as multiple testing and false discovery rate. The objective of my dissertation research is to develop and present an innovative estimating procedure for assessing the proportion, denoted by \( \pi_0 \), of true null hypothesis, in a multiple hypotheses testing setup.

The most important reason for wanting to estimate \( \pi_0 \) is that it is a quantity of independent interest in many situations. In the case of testing for differential expression in DNA microarrays, for example, it is important to know the proportion of differentially expressed genes is \( 1 - \pi_0 \), even if we cannot identify these genes (Egil Ferkingstad and Lindqvist (2003)). In addition to this, a reliable estimate of \( \pi_0 \) is crucial when we want to assess and/or control the false discovery rate in a multiple testing problems.

1.1 Multiple Testing

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. A common feature in genomes studies is the analysis of a large number of simultaneous measurements. One must decide whether the findings are truly causative correlations or just the byproducts of multiple hypotheses testing (Gyorffy et al. (2005)). If one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected by chance alone may be unduly large (Romano et al. (2010)).

When testing many different parameters in a patient simultaneously and setting the significance threshold at a traditionally acknowledged level, say 5%, on average one parameter will be reported to be different as compared to the healthy reference population even without any biological significance. Not taking into account the possible impact of multiple simultaneous testing can greatly
increase the probability of false positive findings.

In genomes studies, we test many hypotheses simultaneously, thus the overall consequences of this 5% are drastically magnified. For example, when investigating a thousand genes and setting \( \alpha \) to 0.05, 50 genes would be found as significant even without any real correlation. Clearly, appropriate multiple testing corrections are needed for the correct interpretation of the results (Györffy et al. (2005)).

In a multiple hypothesis testing problem, \( m \) null hypotheses are tested simultaneously, that is, we test

\[
H_{0i} \text{ versus } H_{1i},
\]

for \( i = 1, 2, \ldots, m \), simultaneously. Assume that the \( m \) tests are constructed based on the observed \( p \)-values, \( p_1, \ldots, p_m \), respectively. The quantity \( \pi_0 \) to be estimated is the proportion of the true null hypotheses among \( H_{01}, \ldots, H_{0m} \). We introduce the random variables \( H_1, \ldots, H_m \) as:

\[
H_i = \begin{cases} 
0, & \text{if } H_{0i} \text{ is true,} \\
1, & \text{otherwise}
\end{cases}
\]

for \( i = 1, \ldots, m \). We assume that each \( H_i \sim \text{Bernoulli} \left( \pi_0 \right) \), that is, \( P \left( H_i = 0 \right) = \pi_0 \) and \( P \left( H_i = 1 \right) = 1 - \pi_0 \) and that the \( H_i \)'s are independent. The number of true null hypothesis is then the random variable \( M_0 = \sum_{i=1}^{m} (1 - H_i) \). \( M_0 \) is binomially distributed as \( M_0 \sim \text{Bin}(m, \pi_0) \). A realized value of \( M_0 \) is denoted by \( m_0 \). We also assume the \( p \)-values, \( p_1, \ldots, p_m \), are continuous and independent random variables, so that the \( p \)-values are identically and independently distributed as \( \text{Unif}(0, 1) \) when the null hypotheses are all true (Egil Ferkingstad and Lindqvist (2003)).
1.2 False Discovery Rate Control

The concept of false discovery rate can be motivated as follows. In recent years, there has been a substantial increase in the size of data sets collected in a number of scientific fields, including genomics, astrophysics, brain imaging, and spatial epidemiology. This has been due in part to an increase in computational abilities and the invention of various technologies, such as high-throughput biological devices. The analysis of high-dimensional datasets often involves performing simultaneous hypothesis tests on each of thousands or millions of measured variables.

Classical multiple hypothesis testing methods utilizing the family-wise error rate were developed for performing just a few tests, where the goal is to guard against any single false positive occurring. However, in the high-dimensional setting, a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate (FDR) is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets (see Storey (2011)).

Hypothesis testing in high-dimensional genomics datasets has been particularly influential in increasing the popularity of false discovery rates (Storey and Tibshirani (2003)). For example, DNA microarrays measure the expression levels of thousands of genes from a single biological sample. It is often the case that microarrays are applied to samples collected from two or more biological conditions, such as from multiple treatments or over a time course. The use of microarray technology allows investigators to ask important questions such as "which genes expressions are affected by selected treatments?" or "which genes on a chip are undergoing differential expression between target mRNA samples?". These questions can be addressed as a multiple hypothesis testing problem. In other words, one would consider a simultaneous test of the null hypothesis that there is no association between the expression levels within each gene and the target mRNA responses of interest. It is most likely that the number of genes which change will be small, likewise the proportion of genes whose expression levels are unaffected will be large. In addition to
incurring false positives, failing to identify truly differentially expressed genes is a major concern, leading to the false discovery rate being in widespread use in this area (see Pounds and Morris (2003), Efron et al. (2001), Davy (2004)).

This problem has lead to the consideration of the so called false discovery rate (FDR) controlling procedures in microarray analysis as a suitable method of controlling the amount of error when determining genes undergoing significant changes (see Benjamini and Hochberg (1995), Reiner et al. (2001)).

When \( m \) null hypotheses are tested, we choose to reject or fail to reject each null hypothesis based on its corresponding \( p \)-value. The testing situation is summarized in Table 1.1 using the notation of Benjamini and Hochberg (Benjamini and Hochberg (1995)).

<table>
<thead>
<tr>
<th></th>
<th>Declared non-significant</th>
<th>Declared significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No discovery</td>
<td>No discovery</td>
<td>Declared discovery</td>
<td></td>
</tr>
<tr>
<td>Negative Result</td>
<td></td>
<td>Positive Result</td>
<td></td>
</tr>
<tr>
<td>( H_0 ) true null hypotheses</td>
<td>( U )</td>
<td>( V )</td>
<td>( m_0 )</td>
</tr>
<tr>
<td>( H_0 ) non-true null hypotheses</td>
<td>( T )</td>
<td>( S )</td>
<td>( m - m_0 )</td>
</tr>
<tr>
<td>Total</td>
<td>( m - R )</td>
<td>( R )</td>
<td>( m )</td>
</tr>
</tbody>
</table>

Consider the problem of testing simultaneously \( m \) hypotheses, of which \( m_0 \) are true. \( R \) is the number of hypotheses rejected. The specific \( m \) hypotheses are assumed to be known in advance. \( R \) is an observable random variable; \( U, V, S \) and \( T \) are unobservable random variables. \( V \) is the number of Type I errors, (hypotheses declared significant when they are actually from the null distribution), and \( T \) is the number of Type II errors, (hypotheses declared not significant when they are actually from the alternative distribution). If each individual null hypothesis is tested separately at level \( \alpha \), then \( R = R(\alpha) \) is increasing in \( \alpha \).
This section introduces the definition of FDR and FDR controlling procedure proposed by Benjamini and Hochberg (1995), which controls the expected proportion of false positive in a list of rejected hypotheses. The level of control in this procedure is an expectation, the experimental variance is unknown for any realization of the procedure. To assess the variability of this procedure, the general operating characteristics of FDR control are examined (Davy (2004)). In the following, we will describe the development of the concept of FDR.

1.2.1 Error Rate Control

In multiple testing problems, we would like to control some adequate error measure (Error Rate, ER). The first multiple testing correction tests were set up in line with the statistical hypotheses of Neyman and Pearson in the 1920s (Gyorffy et al. (2005)). Back then, the main goal was to predict the number of defects in industrial production. For example, how can we estimate the overall number of defect light bulbs when we check 20 out of 1000?

In today’s genomic research, the correction protects us from making over-optimistic assumptions after finding few minimally significant parameters when randomly investigating a large set of parameters. Similar scenario arises when the same test is performed on different subgroups, for example in the investigation of the role of gene polymorphisms in several diseases. Furthermore, in studies randomly searching for correlations without any pre-defined hypothesis the use of multiple testing correction is of utmost importance (see Gyorffy et al. (2005)).

In statistical literature, $\alpha$ is used as the standard notation for the probability of committing a Type I error (a false positive), and $\beta$ as the probability of committing a Type II error (a false negative). The power in statistical hypothesis testing is defined as the probability, $1 - \beta$, of correctly identifying non true null hypotheses. As the number of simultaneous hypothesis tests increases, the $\alpha$ significance threshold must be modified to account for the increasing number of expected
rejections due to chance, so as to maintain a specified level of error rate control as discussed in section 1.1.

Multiple comparison procedures (MCPs) are used to determine $\alpha$ based on various criteria. The observed probabilities are conditional on which hypotheses are actually true. Generally speaking, we have weak control and strong control of error rate: the **weak control** of an error rate at level $\alpha$, says, means the $\text{ER} \leq \alpha$, for $m_0 = m$, provided all null hypotheses are true, while the **strong control** of an error rate at level $\alpha$ says, means that $\text{ER} \leq \alpha$ for all values of $m_0$ simultaneously (Davy (2004), Langaas et al. (2005)).

1.2.2 Family Wise Error Rate

Multiple testing adjustments attempt to control quantities related to the unobservable random variable $V$, that is defined in Table 1.1. The classical error rate chosen to control multiple testing problems is the Family Wise Error Rate (FWER), defined as the probability of committing at least one false positive,

$$FWER = P(V \geq 1)$$

$$= 1 - P(V = 0)$$

(1.2.1)

In multiple hypothesis testing under FWER control, as the number of hypothesis tests, $m$, increases, the $p$-value rejection threshold, $\alpha$, decreases toward 0, thus providing a high level of certainty in the rejected null hypotheses at the expense of the rejection threshold being overly conservative.

To control the FWER at level $\alpha$ (that is, $P(V \geq 1) \leq \alpha$) requires that each individual test is conducted at a lower level. For example, using the common Bonferonni procedure to control the
FWER at level $\alpha$, each null hypothesis $H_{0i}$ is rejected when $p_i \leq \frac{\alpha}{m}$, that is,

$$Pr(\text{Reject $H_0$| when $H_0$ is true}) = P(p_1 \leq \frac{\alpha}{m}|H_0) = \frac{\alpha}{m}. $$

In the context of microarray experiments, FWER control against a single false positive is typically too strict, which leads to many missed detections. A strong control of the FWER can result in low power to reject alternative hypotheses (Dudoit et al. (2002)).

1.2.3 False Discovery Rate (FDR)

The FDR introduced by Benjamini and Hochberg (1995), is a new multiple error rate, which can provide a substantial gain in power in situations where control of the FWER is not necessary. In most cases of testing thousands of genes for differential expression using DNA microarrays, the DNA microarray experiment would only be screening to pick out genes for further study, and one would prefer to accept a few false discoveries rather than sacrificing the power of the testing procedure. Furthermore, the problem of estimating the proportion $\pi_0$ has naturally arisen in assessing or controlling an overall false rejection rate in a simultaneous hypotheses testing problems. A reliable estimate of $\pi_0$ is crucial when we want to control and/or assess the false discovery rate (FDR) proposed by Benjamini and Hochberg (BH), as described as follows.

Consider a family of $m$ hypotheses tests, of which $m_0$ are true. The proportion of erroneously rejected null hypotheses among all the rejected null hypotheses can be captured by the random variable $Q = \frac{V}{R}$, where $R$ is the number of rejected hypotheses and $V$ is the number of false rejections (type I errors) as described in the Table 1.1. The FDR is defined as the expected proportion of falsely rejected hypotheses among all rejections.
The FDR is defined as the mean of the false rejection rate if $R > 0$, and zero if $R = 0$, that is,

$$FDR = E \{ Q I(R > 0) \}$$

(1.2.2)

where

$$I(R > 0) = \begin{cases} 
1 & \text{if } R > 0, \\
0 & \text{if } R = 0.
\end{cases}$$

Equivalently, this can be written as

$$FDR = E \{ Q | R > 0 \} \cdot \text{Prob}(R > 0)$$

(1.2.3)

Benjamini and Hochberg (1995), then presented a multiple test procedure designed to control the FDR in the case that the test statistics are independent under the null hypotheses. Benjamini and Hochberg (BH) controlling procedure was based on Simes (1986) procedure for testing multiple hypotheses. Simes’ procedure is as follows:

\underline{Algorithm 1.1:} Simes’ procedure

1. Compute the $p$-values $p_1, \ldots, p_m$ for each individual test;
2. Order the $p$-values as $p_1 \leq \cdots \leq p_m$;
3. Reject $H_0$ if there exists an index $1 \leq k \leq m$, such that $p_{(k)} \leq \alpha \frac{k}{m}$.

It is proved that the algorithm controls the type I error with respect to global hypothesis $H_0$ at level $\alpha$ if the underlying test statistics and the corresponding $p$-values are identically and independently distributed (Dickhaus (2008)). Benjamini and Hochberg employed Simes’ critical values in the context of FDR control and thus, developed the linear step-up procedure which works as follows:
Algorithm 1.2: BH procedure

1. Compute the $p$-values $p_1, \ldots, p_m$ for each individual test;
2. Order the $p$-values as $p_1 \leq \cdots \leq p_m$;
3. Determine $\hat{k} = \max\{k : p_{(k)} \leq \alpha \frac{k}{m}\}$;
4. if $\hat{k}$ exists then
5.     Reject all null hypotheses $H_{0i}$ for $i = 1, \ldots, \hat{k}$;
6. else
7.     no hypotheses are rejected;

Benjamini and Hochberg (1995), showed that Simes’ procedure strongly controls the FDR under independence at level $\frac{m_0}{m} \alpha = \pi_0 \alpha \leq \alpha$, where $\alpha$ is a level to be determined, and $\pi_0$ is the proportion of the true null hypotheses in $m$ hypotheses of interest, regardless of the distributions of the test statistics corresponding to false null hypotheses, when the distributions under the null are continuous.

In situations when $\pi_0$ is small, the expected FDR will be well below $\alpha$. Depending on the power of the hypothesis tests, as $\pi_0$ approaches 1 the number of rejected hypotheses, $R$, will decrease since $m_0$ is large, but the probability of false positive in the rejected hypotheses will increase. If $\pi_0$ is significantly less than one, and $\hat{\pi}_0$ is a sufficiently good estimate, then the utilization of $\hat{\pi}_0$ in this way leads to an increase in power, while still achieving control at approximately level $\alpha$. If $\gamma$ is the FDR level that we wish to achieve, and if $\pi_0$ can be estimated efficiently, say by $\hat{\pi}_0$, then $\alpha$ is determined by $\alpha = \frac{\gamma}{\hat{\pi}_0}$, so that $FDR = \frac{\gamma}{\pi_0}$. This is known as adaptive control of the FDR. FDR control at the level $\alpha$ provides a list of rejected hypotheses where the expected proportion of false discoveries is controlled at $\pi_0 \alpha$ under the non-adaptive control, and at the constant threshold level $\alpha$ under adaptive control.
Algorithm 1.3: Adaptive BH Procedure

1 Compute the BH procedure at level $q$, and:

2 if no hypothesis is rejected then
3    stop;
4 else
5    proceed;

6 Estimate $m_0(k)$ by $\frac{m+1-k}{1-p(k)}$;

7 Starting with $k = 2$, stop when for the first time $m_0(k) > m_0(k - 1)$;

8 Estimate $\hat{m}_0 = \min(m_0(k), m)$ rounding up to the next highest integer;

9 Compute the BH procedure with $q^* = \frac{am}{m_0}$.

Suppose a researcher rejects $H_0$ with a $p$-value less than a pre-specified chosen level $\alpha$. To estimate the corresponding FDR in this situation, Storey (2002) proposed

$$ \overline{FDR}(\alpha) = \frac{m\hat{\pi}_0\alpha}{R(\alpha)} $$

In practice, $\pi_0$ is generally unknown during hypothesis testing, therefore adaptive control relies on good estimation of $\pi_0$.

1.2.4 A Multiple Endpoint Analysis Example

Multiple endpoints analysis in clinical trials is one of the most commonly encountered multiplicity problems in medical research. This example on multiple endpoints illustrates the various controlling procedures discussed earlier in this chapter and shows the increased number of rejections when $m_0$ is estimated and accounted for. Since the data represent multiple measurements on the same individual, an individual’s innate level can be viewed as latent (Benjamini et al. (2006)).
We discuss an example that was adopted from Benjamini and Hochberg (1995). Thrombolysis with recombinant tissue-type plasminogen activator (rt-PA) and anisoylated plasminogen streptokinase activator (APSAC) in myocardial infarction has been proved to reduce mortality. Neuhaus et al. (1992) investigated the effects of a new front-loaded administration of rt-PA versus those obtained with a standard regimen of APSAC, in a randomized multi center trial in 421 patients with acute myocardial infarction. Four families of hypotheses can be identified in the study:

1. base-line comparisons (eleven hypotheses), where the problem is of showing equivalence;
2. patency of infarct-related artery (eight hypotheses);
3. reocclusion rates of patent infarct-related artery (six hypotheses);
4. cardiac and other events after the start of thrombolitic treatment (fifteen hypotheses).

In this, the last family FDR control may be desired. According to Benjamini and Hochberg (1995), we do not wish to conclude that the front-loaded treatment is better if it is merely equivalent to the previous treatment in all respects. Neuhaus et al. (1992), however, did not pay attention to the problem of multiplicity (the only exception being the division of the end points into primary and secondary). The individual $p$-values are reported as they are, with no word of warning regarding their interpretation.

Neuhaus et al. (1992) concluded that "compared to APSAC treatment, despite more early reocclusions, the clinical course with rt-PA treatment is more favorable with fewer bleeding complications and a substantially lower in-hospital mortality rate, presumably due to improved early patency of the infarct-related artery".

The statement about the mortality was based on a $p$-value of 0.0095. If we consider the fourth family, which contains the comparison of mortality and 14 other comparisons. The ordered $p(i)$'s for the 15 comparisons made are
Controlling the FWER at 0.05, the Bonferroni approach, using \( \frac{0.05}{15} = 0.0033 \). Hence we reject the three hypotheses corresponding to the smallest \( p \)-values. These hypotheses correspond to reduced allergic reaction, and to two different aspects of bleeding; they do not include the comparison of mortality.

Using the BH FDR controlling procedure with \( \alpha = 0.05 \), we now compare sequentially each \( p(i) \)'s with \( \frac{0.05i}{15} \), starting with \( p_{(15)} \). The first \( p \)-value to satisfy the constraint is \( p_{(4)} \) as

\[
p_{(4)} = 0.0095 \leq \frac{4}{15} 0.05 = 0.013
\]

Thus we reject the four hypotheses having \( p \)-values which are less than or equal to 0.013.

Four hypotheses were also rejected at the first stage of the adaptive BH controlling procedure at level 0.05. At the second stage of the adaptive BH controlling procedure, we now compute sequentially \( m_0(i) \) by \( \frac{15 + 1 - i}{1 - p(i)} \) starting with \( p_{(5)} \) and stop when \( m_0(i) > m_0(i - 1) \). The first \( m_0 \) value to satisfy the constraint is \( m_k(10) \) as

\[
m_0(10) = \frac{6}{0.676} = 8.875 > m_0(9) = \frac{7}{0.9541} = 7.336
\]

Therefore, the adaptive BH procedure will yield \( q^* = \frac{0.05 \times 15}{9} = 0.08333 \), resulting in the rejection of the nine hypotheses whose \( p \)-values are less than or equal to 0.08333. We may support now with appropriate confidence the statements about mortality decrease, of which we did not have sufficiently strong evidence before.
1.3 Structure of the Dissertation

The core ideas of estimating the proportion of true null hypotheses in multiple testing problems and false discovery rate are seen throughout this dissertation. In chapter 2 we review the existing estimating methods for estimating the proportion of the true null hypothesis $\pi_0$. In the third chapter, we present the multinomial model and discuss density estimation via histogram. We also discuss an optimization iterative method - Expectation Maximization (EM) Algorithm.

Using all these tools together, we investigate a new estimation method of the proportion of the true null hypotheses by approximating the alternative $p$-values with a multinomial model and computing the EM iterative step for finding $\pi_0$. The fourth chapter of this dissertation applies the ideas of the third chapter in simulation studies and then compare the results with the existing methods discussed in chapter 2.

An application of newly develop estimate is applied to the real life data - DNA microarray data of the leukemia gene expression dataset and the breast cancer gene expression dataset, is given in chapter 5. Conclusions, recommendations and outlook on further studies are discuss in chapter 6.
CHAPTER 2  REVIEW OF EXISTING METHODS

In this chapter, we will review and discuss the details of some of the existing estimating methods for estimating the proportion of true null hypothesis, $\pi_0$, in a multiple testing problem. Without loss of generality, we define $H = 1$ as the alternative hypothesis, which will be used throughout this section. Three non-parametric estimators proposed recently are described and discussed as follows and they are:

1. Storey’s Method (Storey (2002))

2. Convest Method (Langaas et al. (2005))

3. Average Estimate Approach (Jiang and Doerge (2008))

2.1 Storey’s Method

Consider the marginal mixture density of the $p$-values, for any $\lambda \in (0,1)$:

$$P(p > \lambda) = \pi(1 - \lambda) + (1 - \pi)P(p > \lambda|H = 1)$$  \hspace{1cm} (2.1.1)

Let $W(\lambda) = \#\{p_i > \lambda\}$, the number of $p$-values greater than some value $\lambda$. On the basis that $P(p > \lambda|H = 1)$ is typically small, a large majority of the $p$-values in the interval $[\lambda, 1]$ should be corresponding to the true null hypothesis and thus uniformly distributed on the interval $(0, 1)$, of which most should be close to 1, so that $(1 - \pi)P(p > \lambda|H = 1)$ is approximately zero. This implies that the expected value of $W(\lambda)$ should be approximately equal to the product of $m\pi_0$ and the length of the interval $[\lambda, 1]$, that is,

$$E(W(\lambda)) \approx m\pi_0(1 - \lambda)$$
Hence, Storey (2002) proposed that the proportion of true null hypotheses, $\pi_0$, is estimated by (with an appropriately chosen $\lambda$):

$$\hat{\pi}_0(\lambda) = \frac{W(\lambda)}{(1 - \lambda)m} \quad (2.1.2)$$

where $W(\lambda) = \# \{ p_i : p_i > \lambda \}$ and $0 \leq \lambda < 1$ is a tuning parameter.

Storey’s estimator $\hat{\pi}_0(\lambda)$ will tend to overestimate $\pi_0$ because

$$E\hat{\pi}_0(\lambda) = \frac{E\left(\sum_{i=1}^{m} I\{p_i > \lambda\}\right)}{(1 - \lambda)m}$$

$$= \frac{E(U(\lambda)) + E(T(\lambda))}{(1 - \lambda)m}$$

$$= \frac{E(E(U(\lambda)|M_0)) + E(E(T(\lambda)|M_0))}{(1 - \lambda)m}$$

$$= \frac{E((1 - \lambda)M_0) + E(1 - H(\lambda)(m - M_0))}{(1 - \lambda)m}$$

$$= \frac{(1 - \lambda)m\pi_0 + (1 - H(\lambda))m(1 - \pi_0)}{(1 - \lambda)m}$$

$$= \pi_0 + \frac{1 - H(\lambda)}{1 - \lambda}$$

$$\geq \pi_0 \quad (2.1.3)$$

But Storey’s estimator has a large bias and small variance when $\lambda$ is small and a small bias and large variance when $\lambda$ is big respectively. Since both extremes of $\lambda$ have a bias-variance trade off, Storey et al. (2004) proposed a bootstrapping, which is a resampling technique, to choose $\lambda$ when estimating $\hat{\pi}_0(\lambda)$ so as to minimize the mean square error of $\hat{\pi}_0(\lambda)$. The rationale behind this approach is to approximate

$$MSE(\lambda) = E[(\hat{\pi}_0(\lambda) - \pi_0)^2] \quad (2.1.4)$$
by
\[
\hat{MSE}(\lambda) = \frac{1}{K} \sum_{i=1}^{K} (\hat{\pi}_0^{*i}(\lambda) - \min_{\lambda} \hat{\pi}_0(\lambda))^2
\] (2.1.5)

where \(\hat{\pi}_0^{*i}(\lambda)\) represents the \(i\)th of \(K\) bootstraps replications of size \(m\) from the observed \(p\)-values and \(\min_{\lambda} \hat{\pi}_0(\lambda)\) is a plug in estimator of \(\pi_0\), denotes the minimum of \(\hat{\pi}_0(\lambda)\) over \(\lambda \in [0, 1)\). The optimal \(\lambda\) is
\[
\lambda = \arg\min_{\lambda} \hat{MSE}(\lambda).
\] (2.1.6)

However, bootstrap versions of \(\hat{\pi}_0(\lambda)\) behave poorly as an estimator of \(\pi_0\). \(\hat{\pi}_0(\lambda)\) was a conservative estimate and the size of the bias depends on the amount of the tail probability
\[
P(p > \lambda|H = 1) = \int_{1}^{\lambda} x^2 dx h(p) dp.
\]

2.2 Convex Method - Convex Decreasing Density Estimator

Consider a marginal probability density function \(f(p)\) of the \(p\)-values
\[
f(p) = \pi + (1 - \pi) h(p)
\] (2.2.1)

where \(h(p)\) is the conditional probability density function (pdf) of the \(p\)-value given \(H=1\).

[Langaaas et al. (2005)] proposed a convex decreasing estimate on a vector of \(p\)-values. Now, if we examine the empirical distribution function, \(\hat{F}(\cdot)\) for the observed \(p\)-values \(p_1, p_2, \ldots, p_m\) as
\[
\hat{F}(\lambda) = \frac{\{p_i \leq \lambda\}}{m} = \frac{m - \{p_i > \lambda\}}{m}
\]

where \(0 \leq \lambda \leq 1\).

Let \(F\) be the true (unknown) cumulative distribution function and \(f\) the density of the \(p\)-values.
Note that \( \hat{\pi}_0(\lambda) \) can be written as

\[
\hat{\pi}_0(\lambda) = \frac{1 - \hat{F}(\lambda)}{(1 - \lambda)}.
\]

Therefore, we can view \( \hat{\pi}_0(\lambda) \) as a plug-in estimator for the quantity \( \pi_0(\lambda) \), defined by

\[
\pi_0(\lambda) = \frac{1 - F(\lambda)}{(1 - \lambda)}.
\]

The assumption by Langass and Lindqvist is that the \( p \)-values has distribution \( F(\lambda) \) with smooth decreasing convex density \( f(\lambda) \). Applying Taylor series expansion to \( F(\lambda) \) yields

\[
1 = F(1) = F(\lambda) + (1 - \lambda)f(\lambda) + \frac{1}{2}(1 - \lambda)^2 f'(\lambda) + \cdots,
\]

and from this, we obtain,

\[
\pi_0(\lambda) = f(\lambda) + \frac{1}{2}(1 - \lambda)^2 f'(\lambda) + \cdots.
\]

So when \( \lambda \) is close to 1, and \( f(\cdot) \) is known, then \( f(\lambda) \) can be regarded as an estimator of \( \pi_0(\lambda) \). But in fact \( f(\cdot) \) is unknown and Langass and Lindqvist decided to find an estimate for \( f(\cdot) \).

Langaas et al. (2005) estimated \( f \) defined in the mixture model in equation 2.2.1 under the assumption that it is convex and decreasing on \([0,1]\) (\( f \) vanishes outside the interval). For this assumption to hold, it is necessary and sufficient that \( h \) is convex and decreasing on the convex set \([0,1]\). Langaas and Lindqvist proposed fitting the \( p \)-values using a nonparametric mixture model:

\[
f(p) = \int_0^1 f_\theta(p) \mu(d\theta)
\]

where the mixing density is

\[
f_\theta(p) = \begin{cases} 
\frac{2(\theta - p)}{\sigma^2 I_{(0,\sigma)}(p)} & \text{if } \theta \in (0,1], \\
I_{[0,1]}(p) & \theta = 0
\end{cases}
\]
Here $I_A$ is the indicator function on the set $A$, that is, $I_A(p) = 1$ if $p \in A$ and equals zero otherwise. $\mu$ is any probability measure on $[0,1]$ and each density $f_{\theta}$ is a very simple triangular density. Thus they were able to characterize the maximum likelihood estimate for $f$ using the nonparametric maximum likelihood estimates (NPMLE) for the density $f(p)$ as:

$$
\hat{f}(p) = \int_0^1 f_{\theta}(p) \hat{\mu}(d\theta).
$$

(2.2.2)

Given the ordered observed $p$-values $p(1), \ldots, p(m)$ from $m$ hypotheses tests. Thus, the nonparametric maximum likelihood estimator of $f(\cdot)$ is thus, given by

$$
\hat{f} = \arg\max_m \prod_{i=1}^m f(p(i))
$$

(2.2.3)

Then the proportion $\pi_0$ is proposed to be estimated by

$$
\hat{\pi}_0^c = \hat{f}(1) = \hat{f}_m
$$

(2.2.4)

where

$$
\hat{f}_m = \min_{i \leq m-1} \frac{\hat{F}(p(m)) - \hat{F}(p(i))}{p(m) - p(i)}
= \min_{i \leq m-1} \frac{1 - l/m}{p(m) - p(i)}
$$

(2.2.5)

The estimator $\hat{\pi}_0^c$ was constructed via a density estimate $\hat{f}(p)$ at or about the upper bound of support, namely $\hat{f}(1)$. Lindqvist and Langass remarked that the estimator can be conservative when the assumption $h(1) = 0$ is questionable or $h(p) \to 0$ slowly as $p \to 1$. The overestimating problem would be even more severe when the true value of $\pi_0$ is not large, i.e., $1 - \pi_0$ is not so small. If $\pi_0$ is close to 1, the factor $1 - \pi_0$ helps the excessive term $(1 - \pi_0)h(1)$ get closer to zero.
so that the bias reduces.

2.3 Average Estimate Approach

Jiang and Doerge (2008) approach was motivated by the Storey (2002) estimate. They observed (and many other authors also pointed out) that Storey’s estimator has a large bias and small variance when \( \lambda \) is small, and a small bias and large variance when \( \lambda \) is big respectively. Since both extremes of \( \lambda \) have a bias-variance trade off, Jiang and Doerge, JD proposed to combine Storey’s estimates with different \( \lambda \)-values that vary from small extreme to large extreme.

Let \( 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n < 1 \), and suppose that for each \( \lambda_i \), we have Storey (2002) estimate

\[
\hat{\pi}_0(\lambda_i) = \frac{W(\lambda_i)}{(1-\lambda_i)m} \tag{2.3.1}
\]

Let

\[
E[\hat{\pi}_0(\lambda_i)] = \pi_0 + \delta_i \tag{2.3.2}
\]

where \( \delta_i \) is the expected bias of the estimate, and \( \delta_i \geq \delta_{i+1}, Var[\delta_i] = \sigma_i^2 \) and \( \sigma_i^2 \leq \sigma_{i+1}^2 \). JD’s new estimate of \( \hat{\pi}_0(\lambda) \) is to consider the average of \( \hat{\pi}_0(\lambda) \) over the values of \( \lambda_i \),

\[
\hat{\pi}_0(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_0(\lambda_i) \tag{2.3.3}
\]

The bias of \( \hat{\pi}_0, \frac{1}{n} \sum_{i=1}^{n} \delta_i \) is smaller than that of \( \hat{\pi}_0(\lambda_1) \) (the bias of the estimate of \( \pi_0 \) at \( \lambda = \lambda_1 \)) and \( \hat{\pi}_0 \) has a smaller variance than \( \hat{\pi}_0(\lambda_n) \). Considering the average of \( \hat{\pi}_0(\lambda) \) reduces the problem to choosing the range of \( \lambda \).

Lets define \( 0 = t_1 < t_2 < \cdots < t_B < t_{B+1} = 1 \) as equally spaced points in the intervals \([0, 1]\), such that the interval \([0, 1]\), is divided into \( B \) intervals with equal length \( \frac{1}{B} \), that is, \( t_i = \frac{(i-1)}{B} \).
For each \( t_i \), \( \hat{\pi}_0 (t_i) \) is an estimate of \( \pi_0 \) via equation 3.1.2 with \( \lambda = t_i \). The goal is then to find a subset of \( t_i \)'s such that a new estimate of \( \pi_0 \) is obtained by taking the average of the corresponding values of \( \hat{\pi}_0 (t_i) \). Let

\[
NB_i = \# \{ p_k : p_k \geq t_i \},
\]

(2.3.4)

that is, \( NB_i \) denotes the number of \( p \)-values which are greater than or equal to \( t_i \), and let

\[
NS_i = \# \{ p_k : t_i \leq p_k \leq t_{i+1} \},
\]

(2.3.5)

where \( NS_i \) the number of the \( p \)-values in the interval of \([t_i, t_{i+1})\). Then we have, for \( 1 \leq i \leq B \),

\[
\hat{\pi}_0 (t_i) = \frac{NB_i}{(1 - t_i) m}
\]

(2.3.6)

If the \( NB_i \) \( p \)-values come from the null distribution, then on average, there are \( \frac{NB_i}{B - i + 1} \) \( p \)-values in each of the \( (B - i + 1) \) small intervals on \([t_i, 1]\). For small \( i \), \( NS_i \) is usually greater than \( \frac{NB_i}{B - i + 1} \). Therefore, initiating from \( i = 1 \), we find the first \( i \) such that \( NS_i \leq \frac{NB_i}{B - i + 1} \). If such \( i \) exists, \( t_i \) can be considered as the change point and we assume all the \( p \)-values bigger than \( t_i \) come from the true null hypotheses. From this \( \pi_0 \) can be estimated by

\[
\hat{\pi}_0 (B) = \frac{1}{B - i + 1} \sum_{j=1}^{B} \hat{\pi}_0 (t_j)
\]

(2.3.7)

\[
= \frac{1}{B - i + 1} \sum_{j=1}^{B} \frac{NB_j}{(1 - t_j) m},
\]

where \( i = \min \{ i : NS_i \leq \frac{NB_i}{B - i + 1} \} \).

A bootstrapping algorithm to pick the optimal \( B \) is proposed in [Jiang and Doerge] (2008) to minimize the mean-squared error, (mean-squared error is used to measure both the bias and variance). For each \( B \in I, I = \{5, 10, 20, 50, 100\} \), they estimated its respective mean squared error
as

$$\widehat{MSE}(B) = \frac{1}{N} \sum_{b=1}^{N} (\hat{\pi}_b^*(B) - \bar{\pi}_0)^2$$ \hspace{1cm} (2.3.8)

where $\bar{\pi}_0 = \text{average}_{B' \in I} \{\hat{\pi}_0(B')\}$. Let $\hat{B} = \arg \min_{B \in I} \widehat{MSE}(B)$, then Jiang and Doerge estimate’s of $\pi_0$ is $\hat{\pi}_0 = \hat{\pi}_0(\hat{B})$.

In the next chapter, we introduce our newly proposed method with discussion in multinomial models, histogram estimation and an iterative optimization technique.
CHAPTER 3  ESTIMATING \( \pi_0 \) IN MULTINOMIAL MODELS: AN INNOVATIVE NON
PARAMETRIC APPROACH

3.1 Multinomial Model

In this section, we briefly introduce the multinomial model. Consider a random variable \( Y \) that
may take one of several discrete values, which we index \( 1, 2, \ldots, k \). Define

\[
\pi_j = P\{Y = j\} \tag{3.1.1}
\]

denoted as the probability that the response \( Y \) falls in the \( j \)-th category. If we assume that the
response categories are mutually exclusive and exhaustive, we have \( \sum_{j=1}^{k} \pi_j = 1 \), that is, the prob-
abilities add up to one for each individual, and we have only \( k - 1 \) parameters.

Suppose \( m \) individuals are observed independently on \( Y \) and let \( Y_{ij} \) be an indicator (or dummy)
variable that takes the value 1 if the \( i \)-th individual falls in the \( j \)-th category and 0 otherwise, and
\( \sum_j Y_{ij} = 1 \), since one and only one of the indicators \( Y_{ij} \) can be on for each case, that is,

\[
Y_{ij} = \begin{cases} 
1, & \text{if the } i^{th} \text{ individual’s } Y \text{ value is equal to } j, \\
0, & \text{otherwise.}
\end{cases}
\]

For grouped data, let \( m_i \) denote the number of cases in the \( i \)-th group and let \( Y_{ij} \) denote the
number of responses from the \( i \)-th group that fall in the \( j \)-th category, with observed value \( y_{ij} \). Note
that the counts in the various response categories add up to the number of cases in each group, that is,

\[
\sum_j y_{ij} = m_i.
\]
The probability distribution of the counts $Y_{ij}$ given the total $m_i$ is given by the multinomial distribution,

$$P\{Y_{i1} = y_{i1}, \ldots, Y_{ik} = y_{ik}\} = \left(\begin{array}{c} m_i \\ y_{i1}, \ldots, y_{ik} \end{array}\right) \pi_1^{y_{i1}} \cdots \pi_k^{y_{ik}} \quad (3.1.2)$$

where

$$\left(\begin{array}{c} m_i \\ y_{i1}, \ldots, y_{ik} \end{array}\right) = \frac{m_i!}{y_{i1}! \cdots y_{ik}!}$$

is the multinomial coefficient.

When $k = 2$ and $m_i = 1$, we have only two response categories, which is the Bernoulli ($\pi$) distribution, that is,

$$P\{Y_{i1} = y_{i1}\} = \pi^{y_{i1}}(1 - \pi)^{1-y_{i1}} \quad (3.1.3)$$

Then the probability distribution of $\sum_j Y_{i1}$ is given by

$$P\{\sum_j Y_{i1}\} = \binom{m}{y} \pi^y(1 - \pi)^{m-y}, \quad y = 0, 1, \ldots, m \quad (3.1.4)$$

A common example of a multinomial is a histogram with $m$ independent events distributed into in $k$ bins (see German (2007)).

3.2 Approximating a distribution via Histogram

A density estimate is a way to estimate the probability density as a function of an observed sample of data. A histogram is a type of a density estimator. A density estimation problem requires a non-parametric approach if we have no information about the target distribution other than the observed data. In other cases, we may have incomplete information about the distribution, so that traditional estimation methods are not directly applicable.

The histogram is a piecewise constant estimate of the density. The classical frequency histogram is formed by constructing a complete set of non-overlapping intervals, called bins, and counting the number of points in each bin. In order for the bin counts to be comparable, the bins should all have the same width. The choice of bin width for a histogram is a choice of smoothing
parameter. There is no best number of bins, and different bin sizes can reveal different features of the data. Grouping data is at least as old as Graunt’s work in 1662, but no systematic guidelines were given until Herbert Sturges’s work in 1926.

A narrow bin width may undersmooth the data, presenting too much detail while a wider bin width may oversmooth the data, obscuring important features as shown in figure 3.1. Using wider bins where the density is low reduces noise due to sampling randomness; using narrower bins where the density is high (so the signal drowns the noise) gives greater precision to the density estimation. The choice of smoothing parameter and bin center is a challenging problem that continues to attract much attention in research. In determining an optimal number of bins, there have been various assumptions made about the shape of the distribution, which is an important factor in determining the optimal bin (see Scott (2015)).
3.2.1 The $L_2$ Theory of Histograms

In this section, we present the mean square error (MSE) properties of a density histogram (see Scott (1979), Scott (2015)). As we have seen and observed in chapter 2, a good estimate would minimize the mean square error. The mean square error of an estimator measures the average of the squares of the errors, that is, the difference between the estimator and what is estimated.

The mean square error is a risk function corresponding to the expected value of the squared error loss. Consider constructing a probability histogram of a observed sample $x_1, \ldots x_m$ from $f(x)$ and a choice of mesh $\{t_k, -\infty < k < \infty\}$ requires the data to be sorted into bins, and the binning operation is determined by the boundaries of the class intervals. Define $B_k = [t_k, t_{k+1})$ denote the $k$th bin. Given a class intervals of equal width $h$, where $h = t_{k+1} - t_k$ for all $k$; then the histogram is said to have a fixed bin width $h$. Thus, the histogram density estimate based on a sample size $m$ is

$$
\hat{f}(x) = \frac{v_k}{mh} = \frac{1}{mh} \sum_{i=1}^{m} I_{[t_k, t_{k+1})}(x_i), \quad \text{where } t_k \leq x < t_{k+1}
$$

(3.2.1)

where $v_k$ is the number of sample points in the class interval $[t_k, t_{k+1})$.

The bias of a histogram density estimator (3.2.1) is proportional to the bin width $h$. The bias in a histogram density estimate is determined by $f'$, the first order derivative of the density. The analysis of the histogram random variable, $\hat{f}(x)$, is quite simple, once it recognized that the bin counts, $\{v_k\}$ are binomial random variables:

$$
v_k \sim B(m, p_k), \quad \text{where } p_k = \int_{B_k} f(t) dt
$$

(3.2.2)

Now, let us consider the MSE of the histogram density estimate $\hat{f}(x)$ for $x \in B_k$. Now,
\[ E[v_k] = mp_k \]

and

\[ \text{Var}[v_k] = mp_k(1 - p_k). \]

Thus

\[
\begin{align*}
\text{Var}\hat{f}(x) &= \frac{\text{Var}[v_k]}{(mh)^2} \\
&= \frac{mp_k(1 - p_k)}{m^2h^2} \\
&= \frac{p_k(1 - p_k)}{mh^2}
\end{align*}
\]

and

\[
\text{Bias}\hat{f}(x) = E\hat{f}(x) - f(x) \\
&= \frac{1}{mh}Ev_k - f(x) \\
&= \frac{mp_k}{mh} - f(x) \\
&= \frac{p_k}{h} - f(x)
\]

Let's assume that \( f(x) \) is Lipschitz continuous over the bin \( B_k \).

**Definition 3.1.** A function is said to be Lipschitz continuous over an interval \( B_k \) if there exists a positive constant \( \varphi_k \) such that \( |f(x) - f(y)| < \varphi_k|x - y| \) for all \( x, y \in B_k \).

By applying the Mean Value Theorem (MVT) to \( p_k \) in 3.2.2 it follows that

\[
p_k = \int_{B_k} f(t)dt = hf(\psi_k), \text{ for some } \psi_k \in B_k
\]
Thus, the variance and bias of the histogram random variable \( \hat{f}(x) \) becomes

\[
Var \hat{f}(x) = \frac{p_k(1-p_k)}{m h^2} \leq \frac{p_k}{m h^2} = \frac{f(\psi_k)}{m h}
\]

and

\[
|Bias \hat{f}(x)| = \left| \frac{p_k}{\hat{h}} - f(x) \right| = \left| \frac{h f(\psi_k)}{\hat{h}} - f(x) \right| = \left| f(\psi_k) - f(x) \right| 
\leq \varphi_k |\psi_k - x| 
\leq \varphi_k h
\]

respectively. Hence,

\[
Bias^2 \hat{f}(x) \leq \varphi^2_k h^2
\]

and the MSE of \( \hat{f}(x) \) is

\[
MSE(\hat{f}(x)) \equiv Var(\hat{f}(x)) + Bias^2(\hat{f}) 
\leq \frac{f(\psi_k)}{m h} + \varphi^2_k h^2
\]

The bin width \( h \) is referred to as a smoothing parameter, since it controls the amount of smoothness in the estimator for a given sample of size \( m \). Equation 3.2.8 summarizes the recurring trade-off between bias and variance as determined by the choice of the smoothing parameter. The variance may be controlled by making \( h \) large so that the bins are wide and of relatively stable height; however, the bias is large.
On the other hand, the bias may be reduced by making $h$ small so that the bins are narrow; however, the variance is large. Note that the bias can be eliminated by choosing $h = 0$, but this very rough histogram is exactly the empirical probability density function $f_m(x)$, which has infinite (vertical) variance. The bias and variance may be controlled simultaneously by choosing an intermediate value of the bin width, and allowing the bin width to slowly decrease as the sample size increases.

**Definition 3.2.** A density estimator is said to be consistent in the mean square if the $MSE\{\hat{f}(x)\} \rightarrow 0$ as $m \rightarrow \infty$.

In order to minimize the asymptotic MSE, an optimal smoothing parameter $h^*$ is defined. This follows as a consequence of equation 3.2.8.

**Theorem 3.1.** Assume that $x \in B_k$ is a fixed point and that $f$ is Lipschitz continuous in this bin with $\varphi_k$. Then the histogram estimate $\hat{f}(x)$ is mean square consistent if as $m \rightarrow \infty$, then $h \rightarrow 0$ and $mh \rightarrow \infty$.

The first condition ensures that the bias vanishes asymptotically, while the second condition ensures that the variance goes to zero.

**Corollary 3.1.** The $MSE(\hat{f}(x))$ bound in Equation 3.2.8 is minimized when

$$h^*(x) = \left[ \frac{f(\psi_k)}{2\varphi_k^2 m} \right]^{1/3} \quad (3.2.9)$$

the resulting $MSE(\hat{f}(x))$ is $O(m^{-2/3})$.

The optimal bin width, thus, decreases at a rate proportional to $m^{-1/3}$. The noise inherent in the histogram varies directly with the square root of its height, since $Var\hat{f}(x) \approx \frac{f(x)}{mh}$ from Equation 3.2.6. This heteroscedasticity (unequal variance) across the histogram estimate may be eliminated
by using a variance-stabilizing transformation.

Among the commonly applied rules for determining the boundaries of class intervals of a histogram are Sturges’ rule, Scott’s reference rule, the Freedman-Diaconis rule among many other rules. There have been modifications of these rules. In most real life applications, data rarely follows a normality and skewed data characterized the probability values in multiple hypotheses testings.

3.2.2 Sturges’ Rule

Sturges’ rule is a rule for determining how wide to choose the bins. Sturges’ rule is based on the implicit assumption that the sampled population is normally distributed. Sturges (1926a) considered a frequency histogram with $k$ bins, each of width 1 and centered on the points $i = 0, 1, ..., k - 1$, and the chose the bin count of the $i$th bin to be the binomial coefficient $\binom{k-1}{i}$. As $k$ increases, the frequency histogram assumes the shape of a normal density with mean $(k - 1)/2$ and variance $(k - 1)/4$.

The total sample size is

$$m = \sum_{i=0}^{k-1} \binom{k-1}{i} = (1 + 1)^{k-1} = 2^{k-1}$$  

by the binomial expansion. So Sturges’ number of bins rule is

$$k = 1 + \log_2 m.$$  

Note that the boundary condition $m = 1$ gives $k = 1$ in equation (3.2.11) For non-normal dataset, additional classes maybe required. In practice, Sturges’ rule is applied by dividing the sample range of the data into the prescribed number of equal-width bins (see Scott (2015)).
As discussed in section 1.2, the use of microarray technology allows investigators to simultaneously measure the expression of thousands of genes, and this can be termed as a statistical hypothesis test for each gene placing a null hypothesis versus and alternative hypothesis. Thousands of hypotheses are tested resulting in a complex multiple testing problem. The multiple testing problem concerns the occurrence of erroneous conclusions among such a large set of hypothesis tests. In classical statistics, \( p \)-values have unified the determination of significance for a wide variety of experimental designs and hypotheses.

In microarray studies, a technique based upon the analysis of the set of \( p \)-values could unify the determination of significance across a wide variety of experimental designs and adequately address the multiple testing issue. Under the null hypotheses, the \( p \)-values will have a uniform density corresponding to a flat horizontal line. Under the alternative hypothesis, the \( p \)-values will have a distribution that has high density for small \( p \)-values and the density will decrease as the \( p \)-values increases (Pounds and Morris (2003)). In other words, the \( p \)-values are highly skewed.

To obtain better density estimates for skewness, Doane (1976) suggested a modification to Sturges’ rule to allow for skewness. Doane (1976) proposed increasing the numbers of bins in equation 3.2.11 by

\[
\log_2(1 + \hat{\gamma})
\]

where \( \hat{\gamma} \) is an estimate of the skewness coefficient. If \( \hat{\gamma} = 0 \), no extra classes are added. As departure from the symmetric normal distribution becomes more obvious, classes are added, but at a decreasing rate.

The skewness statistic is often denoted by \( \sqrt{b_1} \) and is defined as

\[
\sqrt{b_1} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^3}{[\sum_{i=1}^{n} (X_i - \overline{X})^2]^{3/2}}
\]
and the standard deviation of $\sqrt{b_1}$ depends on the sample size, becoming smaller as the sample size increases:

$$\sigma(\sqrt{b_1}) = \sqrt{\frac{6(m - 1)}{(m + 1)(m + 3)}}.$$ 

So that the estimate of the skewness coefficient is $\hat{\gamma} = \frac{\sqrt{b_1}}{\sigma(\sqrt{b_1})}$. Hence, equation 3.2.12 can be written as

$$\log_2\left(1 + \frac{\sqrt{b_1}}{\sigma(\sqrt{b_1})}\right).$$

(3.2.13)

Doane (1976) explained that the formula is motivated by the theory of information coding. The entropy of a message (in bits) is given by the formula $-\log_2 \frac{1}{M}$ where $M$ is the number of different equiprobable symbols which may occur. As more symbols are used to encode a message, entropy increases at a decreasing rate. A frequency classification is appropriately viewed as a mapping from the real line to a finite set of symbols. Therefore, classification is a coding process subject to diminishing returns. The proposed modification of Sturges’ rule is consistent with this view.

3.3 Expectation Maximization (EM) Algorithm

The EM algorithm is an iterative optimization algorithm for approximating the maximum likelihood estimates when some of the random variables involved are not observed, that is, considered missing or incomplete. The main idea of EM algorithm is to obtain parameter estimates when some of the data are missing, via two steps, so-called the expectation step and maximization step.

In the iterative approximations, one is to:

1° replace missing values by estimated values,

2° estimate parameters,

3° repeat

- 1° using previous approximation values as true values, to obtain a new/better approximation and
• $2^n$ using estimated values as observed values for the missing values, iterating until convergence.

The EM idea has been in use for many years before Orchard et al. (1972) in their missing information principle provided the theoretical foundation of the underlying idea. The term EM was introduced in Dempster and Rubin (1977) where they prove the general results about the behavior of the algorithm and showed its wide scope of applications to various statistical models (Yuille (2006)). The characteristics of the EM algorithm are well documented (see Krishnan and McLachlan (1997)). It leads in general to simple equations, has the nice property of increasing the log-likelihood at each iteration until stationary, and in many circumstances, it derives the maximum likelihood estimation (Biernacki et al. (2003)).

According to Dempster and Rubin (1977), Krishnan and McLachlan (1997), Yuille (2006), and Borman (2004), we will discuss the EM algorithm as follows. Suppose we have a random variable $y$ whose joint density $f(y; \theta)$ is indexed by a $p$-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. If the complete data vector $y$ were observed, it is of interest to compute the maximum likelihood estimate of $\theta$ based on the distribution of $y$. The log-likelihood function of $y$ is

$$
\log L(\theta; y) = l(\theta; y) = \log f(\theta; y),
$$

is to be maximized.

In a missing data situation, however, only a function of the complete data vector $y$ is observed. Let us denote this by expressing $y$ as $(y_{\text{obs}}, y_{\text{mis}})$, where $y_{\text{obs}}$ denote the observed but incomplete data and $y_{\text{mis}}$ denotes the unobserved or missing data. For simplicity, let us assume that the missing
data are missing at random, so that

\[ f(\mathbf{y}; \theta) = f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}; \theta) \]

\[ = f_1(\mathbf{y}_{\text{obs}}; \theta) \cdot f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}; \theta), \]

(3.3.1)

where \( f_1 \) is the joint density of \( \mathbf{y}_{\text{obs}} \) and \( f_2 \) is the joint density of \( \mathbf{y}_{\text{mis}} \) given the observed data \( \mathbf{y}_{\text{obs}} \), respectively. Thus the log-likelihood of the observed-data of \( \mathbf{y} \):

\[ l(\theta; \mathbf{y}_{\text{obs}}) = l(\theta; \mathbf{y}) - \log f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}; \theta) \]

(3.3.2)

The EM algorithm attempts to maximize \( l(\theta; \mathbf{y}_{\text{obs}}) \) iteratively, by replacing it by its conditional expectation given the observed data \( \mathbf{y}_{\text{obs}} \). This expectation is computed with respect to the distribution of the complete-data evaluated at the current estimate of \( \theta \).

Now, if we let \( \theta^{(0)} \) be the initial value \( \theta \), then on the first iteration, we compute

\[ Q(\theta; \theta^{(0)}) = E_{\theta^{(0)}}[l(\theta; \mathbf{y})|\mathbf{y}_{\text{obs}}] \]

(3.3.3)

\( Q(\theta; \theta^{(0)}) \) is now maximized with respect to \( \theta \), that is, \( \theta^{(1)} \) is found such that

\[ Q(\theta^{(1)}; \theta^{(0)}) \geq Q(\theta; \theta^{(0)}) \] for all \( \theta \in \Theta \).

Thus the EM algorithm consists of an Expectation step (E-step) followed by a Maximization step (M-step) defined as follows:

**E – step**: Compute \( Q(\theta; \theta^{(t)}) \) where

\[ Q(\theta; \theta^{(t)}) = E_{\theta^{(t)}}[l(\theta; \mathbf{y})|\mathbf{y}_{\text{obs}}] \]

(3.3.3)
M - step : Find $\theta^{(t+1)} \in \Theta$ such that

$$Q(\theta^{(t+1)}; \theta^{(t)}) \geq Q(\theta; \theta^{(t)}) \text{ for all } \theta \in \Theta$$ (3.3.4)

The E-step and the M-step makes up the Expectation and Maximization EM algorithm, hence the name EM. The E-step and the M-step are repeated alternately until

$$l(\theta^{(t+1)}) - l(\theta^{(t)}) \leq \epsilon$$ (3.3.5)

where $\epsilon$ is a prescribed small quantity.

The computation of these two steps simplify a great deal when it can be shown that the log-likelihood is linear in the sufficient statistic for $\theta$. In particular, this turns out to be the case when the distribution of the complete-data vector, that is, $y$ belongs to the exponential family. In this case, the E-step reduces to computing the expectation of the complete data sufficient statistic given the observed data.

When the complete-data are from the exponential family, the M-step also simplifies. The M-step involves maximizing the expected log-likelihood computed in the E-step. In the exponential family case, actually maximizing the expected log-likelihood to obtain the next iterate can be avoided. Instead, the conditional expectations of the sufficient statistics computed in the E-step can be directly substituted for the sufficient statistics that occur in the expressions obtained for the complete-data maximum likelihood estimators of $\theta$ to obtain the next iterate.

In figure 3.2, the function $Q(\theta; \theta^{(t)})$ is bounded above by the function $Q(\theta)$. The functions are equal at $\theta = \theta^{(t)}$. The EM algorithm chooses $\theta^{(t+1)}$ as the value of $\theta$ for which $Q(\theta; \theta^{(t)})$ is a maximum. Since $Q(\theta^{(t+1)}; \theta^{(t)}) \geq Q(\theta; \theta^{(t)})$, increasing $Q(\theta; \theta^{(t)})$ ensures that the value of the function $Q(\theta)$ is increased at each step.
The EM algorithm has several appealing properties. The EM algorithm is useful for several reasons: conceptual simplicity, ease of implementation, and the fact that each iteration improves $l(\theta)$.

- Firstly, it is typically easily implemented because it relies on complete data computations: the E-step of each iteration only involves taking expectations over complete-data conditional distributions. The M-step of each iteration only requires complete data maximum likelihood estimation, for which simple closed form expressions are already available.

- Secondly, it is numerically stable: each iteration is required to increase the log-likelihood $l(\theta; y_{\text{obs}})$ in each iteration, and if $l(\theta; y_{\text{obs}})$ is bounded, the sequence $l(\theta^{(t)}; y_{\text{obs}})$ converges to a stationery value.

A disadvantage of the EM algorithm is that its rate of convergence can be extremely slow if a lot of data are missing. The rate of convergence on the first few steps is typically quite good, but can become excruciatingly slow as you approach a local optima. Generally, the EM algorithm works best when the fraction of missing information is small and the dimensionality of the data is not too large. EM algorithm can require many iterations, and higher dimensionality can dramatically slow
down the E-step.

Dempster and Rubin (1977) show that convergence is linear with rate proportional to the fraction of information about $\theta$ in $(\theta; y)$ that is observed. The choice of initial values is of great importance in the algorithm-based literature as might heavily influence the speed of convergence of the algorithm and its ability to locate the global maximum.

Both the drawbacks, slow convergence and dependence on the initial positions, can be regarded as linked in practical situations. Actually, it is possible that starting from some position leads to a slower convergence rate for the EM algorithm and that the EM algorithm is stopped before reaching a sensible value of the likelihood. There are optimal stopping rules for the EM algorithm which help to accelerate convergence, one of which is Aitken’s acceleration rule will be discuss later in this chapter.

3.4 New estimating procedure for $\pi_0$

3.4.1 Introduction

In this section, we develop the new estimator using details discussed in sections 3.1 - 3.3. A number of methods have been proposed for estimating $\pi_0$ or $m_0$ equivalently, some of which were reviewed in chapter 2. All of these methods are based on the fact that $p$-values under the null hypotheses are uniform $U(0, 1)$ random variables. There are two fundamental properties that provide the basis of the analysis of a large set of $p$-values as follows:

1. $p$-values arising from the null hypothesis are distributed uniformly on the interval $(0,1)$ (see Casella and Berger (2002)),

2. the general distribution of the set of $p$-values can be expressed as a mixture consisting of a uniform $(0,1)$ component and another component that represents the distribution of the set of $p$-values under the alternative (Lindsay (1995)).
The \( p \)-values from the alternative hypotheses in some cases, are not explicitly modeled. Any \( p \)-value that differs significantly from the null \( U(0, 1) \) distribution is rejected as non-null and excluded from the estimation process.

**Theorem 3.2.** Let \( X \) be any continuous-type random variable with pdf \( f(x) \) and \( Y \) be Bernoulli(\( \pi \)).

Then \( f \) can be expressed as

\[
f(x) = \pi_0 h_0(x) + (1 - \pi_0) h_1(x)
\]

(3.4.1)

where \( \pi_0 = P(Y = 1) \) and \( h_0(x) \) is the conditional pdf of \( X \) given \( Y = 0 \) and \( h_1(x) \) is the condition pdf of \( X \) given \( Y = 1 \).

**Proof.** For any Borel-measurable function \( g \),

\[
Eg(X) = E[E\{g(X)|Y\}]
\]

\[= \pi_0 E\{g(X)|Y = 0\} + (1 - \pi_0) E\{g(X)|Y = 1\}\]

\[= \pi_0 \int g(x)h_0(x)dx + (1 - \pi_0) \int g(x)h_1(x)dx\]

(3.4.2)

\[= \int g(x)\{\pi_0 h_0(x) + (1 - \pi_0) h_1(x)\}dx\]

It has thus shown that \( \pi_0 h_0(x) + (1 - \pi_0) h_1(x) \) is pdf of \( X \). The proof is complete. \( \square \)

Applying Theorem 3.2, we can express the general pdf of the \( p \)-values in a multiple testing problem as a mixture of Uniform (0,1) and another pdf (pdf under alternative) as follows:

\[
f(x) = \pi_0 + (1 - \pi_0) h(x)
\]

(3.4.3)

Here \( \pi_0 \) is interpreted as the proportion of true null-hypotheses, and \( p \)-values, \( p_1, \cdots, p_m \) for testing \( H_{0j} \) versus \( H_{1j} \), \( j = 1, \cdots, m \) respectively can be viewed as a random sample of the mixture pdf 3.4.3 as we described in chapter 2.
3.4.2 Mixture Model Framework

Using this mixture representation established in previous section [3.4.1], we are able to characterize the maximum likelihood estimate for $\pi_0$. This characterization is then the basis for the EM algorithm to calculate an approximate maximum likelihood estimate. To motivate this development, the estimation method is based on the common mixture model of the $p$-values:

$$ f = \pi_0 + (1 - \pi_0)h $$  \hspace{1cm} (3.4.4)

We propose to use a multinomial distribution $(k, q_1, \ldots, q_k)$ to fit the $p$-values under alternative, that is, to approximate $h$, in order to approximate the maximum likelihood estimator of $\pi_0$. Denote $\text{UM}(\pi_0; k, q_1, \ldots, q_k)$ for the finite multinomial mixture distribution. By this approach, the alternative distribution $h$ as defined in equation [3.4.4] is restricted to the multinomial distribution family, $\text{UM}(k, q_1, \ldots, q_k)$. Alternatively, the multinomial distribution can be viewed as a parametric approximation to the non-parametric unknown density $h$, similarly to the idea of the empirical likelihood method. So we consider this approach to be non-parametric. Thus, consider $h$ being as follows:

$$ h(p) = \frac{k!}{\xi_1! \cdots \xi_k!} q_1^{\xi_1} \cdots q_k^{\xi_k} $$  \hspace{1cm} (3.4.5)

Note that

$$ \sum_{i=1}^{k} \xi_i = k $$

To apply the approach, we need to settle two things first:

1. select an integer $k$,

2. convert the continuous-type observations $p_1, \ldots, p_m$ into discrete data with $k$ categories.
Selection of \(k\) can be data-based. As it happens often in applications that the \(p\)-values are highly skewed (see Storey and Tibshirani (2003), Zhao et al. (2012), Markitsis and Lai (2010)), we recommend Doane’s modification of Sturges’ rule that is proposed for skewed distribution (Doane (1976), Sturges (1926b)), which was discussed in section 3.2.2. The suggested correction to Sturges’ rule is to add \(K_e = \log_2(1 + \hat{\gamma})\) classes as discussed in section 3.2.2.

To transform \(p_i\)’s, partition the unit interval into \(k\) subintervals with equal width \(1/k\). Define

\[
\xi_{ij} = \begin{cases} 
1, & \text{if } \frac{i-1}{k} \leq p_i < \frac{i}{k}, \\
0, & \text{otherwise}
\end{cases}
\]

for \(i = 1, \ldots, m, j = 1, \ldots, k\). Note that

\[
\sum_{i=1}^{m} \sum_{j=1}^{k} \xi_{ij} = m, \quad \sum_{j=1}^{k} \xi_{ij} = 1, \quad i = 1, \ldots, m.
\]

From the Bayesian interpretation for the multiple testing problem, given that the alternative is true with a probability of \(1 - \pi_0\), the \(p\)-value follows the distribution \(h\). In the same way, the transformed data \(\xi_{ij}\)’s can be interpreted as follows. Given that the alternative is true with a probability of \(1 - \pi_0\), \(\eta_i = (\xi_{i1}, \ldots, \xi_{ik})'\) is a multinomial random vector with distribution multinomial \((k, q_1, \ldots, q_k)\), for \(i = 1, \ldots, m\). Therefore, \(\eta_1, \ldots, \eta_m\) are independently and identically distributed as the finite mixture distribution \(\text{UM}(\pi_0; k, q_1, \ldots, q_k)\) and the maximum likelihood estimate \(\hat{\pi}_0\) for \(\pi_0\) thus results from the transformed data \(\xi_{ij}\)’s.

We should note that the multinomial random vector is a member of the exponential family, hence equation 3.4.5 can be rewritten as

\[
h(p) = e^{x}(\xi_1 \log q_1 + \cdots + \xi_k \log q_k)g(p) \quad (3.4.6)
\]
Let $G$ be a collection of distributions that are continuous at $\theta = 0$. Then, the density $f$ in equation 3.4.4 of the $p$-values can be represented by the following mixture model

$$
\begin{align*}
    f(p, \pi_0, G) &= \pi_0 + (1 - \pi_0) \int h(\theta) dG(\theta) \\
    &= \pi_0 + (1 - \pi_0) \int \exp(\eta(p, \theta)) dG^*(\theta), \\n    &\quad 0 \leq p \leq 1
\end{align*}
$$

With regards to mixture model in 3.4.7, we consider the question of identifiability of $\pi_0$. The identifiability question must be settled before one can meaningfully discuss the problem of estimating the mixing cumulative distribution function $G$ on the basis of observations from the mixture. Identifiability is a property of parameters in statistical models. It is beneficial to know which parameters we may be able to estimate from the data and which parameters cannot be estimated from the data. Identifiability helps make this determination. The idea is to study the mapping between parameters and statistical models. From previous estimation methods discussed in chapter 2, for $\pi_0$ to be identifiable, the density $h$ of the alternative $p$ values equals zero when $p$ gets large. Assuming $f$ is decreasing and $h(1) = 0$, then $\pi_0$ is identifiable. The non-null distribution of the $p$-values are right-skewed and are generally decreasing in shape. The non null $p$-values are expected to be small, thus, $f$ can be assume to a decreasing function. In most applications, the density is convex. Thus the minimum of the density is achieved at 1 and $f(1) = \pi_0$ if $h(1) = 0$.

According to Teicher (1961), the model in equation 3.4.7 is identifiable in $\pi_0$, and $G$ for $0 < \pi_0 < 1$ and $G \in \mathcal{G}$. Let $G^*$ be the cumulative distribution function. Let the random variable $\zeta$ have the distribution $G^*$ and let $\pi = G^*\{0\} = P(\zeta = 0)$. Define $G(0) = P(\zeta \geq \mid \zeta \neq 0)$ to be the conditional distribution of $\zeta$ given $\zeta = 0$, so that

$$
G^*(\theta) = \pi_0 I(\theta \geq 0) + (1 - \pi_0)G(\theta)
$$

with $G \in \mathcal{G}$. 

Define
\[ G_i^*(\theta) = \pi_0 i(\theta \geq 0) + (1 - \pi_0)G_i(\theta) \]
with \(0 < \pi_0 < 1, G_i \in \mathcal{G}, i = 1, 2\) and \(G_1^*(\theta) = G_2^*(\theta)\) for all \(\theta\); then for \(\pi_{01} = \pi_{02}\),
\[ G_1^*\{0\} = \pi_{01} = G_2^*\{0\} = \pi_{02} \]
and so \(G_1(\theta) = G_2(\theta)\) for all \(\theta\). Hence \(f(p, \pi_0, G)\) is identifiable in \(\pi_0\) and \(G\) for \(0 < \pi_0 < 1\) and \(G \in \mathcal{G}\).

3.4.3 Estimating \(\pi_0\) using EM Algorithm

The EM algorithm can be used to obtain an approximation to \(\hat{\pi}_0\). The number of partition, \(k\) is pre-specified using Doane’s modification of Sturges’ rule described in section 3.2.2 In order to use the EM algorithm, we introduce a latent Bernoulli variable \(w\) that indicates the component membership in the finite mixture distribution \(\text{UM}(\pi_0; k, q_1, \ldots, q_k)\), that is, \(w=1\) if and only if the null hypothesis is true. Then \(\text{UM}(\pi_0; k, q_1, \ldots, q_k)\) can be viewed as a missing value model with the density function for the complete data as follows:
\[ g(w, z | \pi_0, q_1, \ldots, q_k) = \pi_0^w \left(1 - \pi_0\right)^{1-w} q_1^{z_1} q_2^{z_2} \cdots q_k^{z_k} \]
where \(w = 0 \text{ or } 1\), and \(z_j = 0 \text{ or } 1, j = 1, \ldots, k\). Let \(w_i\) be the missing value on \(w, i = 1, \ldots, k\). Now the data available for analysis are \(\xi_{ij}\)’s.
Thus, the likelihood with the complete data is

\[ L(\pi_0, q_1, \cdots, q_k) = \prod_{i=1}^{m} \left\{ \pi_0^{w_i} \left[ (1 - \pi_0) \prod_{j=1}^{k} q_j^{(i)} \right]^{1-w_i} \right\} \]

\[ = \prod_{w_i=1} \{ \pi_0 \}^{w_i} \cdot \prod_{w_i=0} \{ (1 - \pi_0) \prod_{j=1}^{k} q_j^{(i)} \}^{1-w_i} \]

\[ = \pi_0^T \cdot (1 - \pi_0)^{m-T} \left\{ \prod_{j=1}^{k} q_j^{(i)} \right\}^{m-T} \tag{3.4.8} \]

where \( T = \sum_{i=1}^{m} w_i \), \( \xi_j = \sum_{i=1}^{m} \xi_{ij} \), for \( 1 \leq j \leq k \)

The log likelihood function with complete data is

\[ l(\pi_0, q_1, \ldots, q_k) = T \log \pi_0 + (m - T) \log(1 - \pi_0) + \sum_{j=1}^{k} (m - T) \xi_j \log q_j \tag{3.4.9} \]

With the log likelihood function with the complete data, we can now characterize the likelihood estimate for \( \pi_0 \) using the EM algorithm discussed in section 3.3.

**E-Step:** Let \( \pi_0^{(s)} \) and \( q^{(s)} \) be the current approximation to the mixture model \( 3.4.9 \). For the next approximation, the E-step establishes the expected log-likelihood function as:

\[ Q(\pi_0, q) = E_{\pi_0^{(s)}, q^{(s)}} \{ l(\pi_0, q_1, \ldots, q_k | n_1, \ldots, n_m) \} \]

\[ = E_{\pi_0^{(s)}, q^{(s)}} (T | \eta) \log \pi_0 + (m - E_{\pi_0^{(s)}, q^{(s)}} (T | \eta)) \log(1 - \pi_0) \]

\[ + \sum_{j=1}^{k} (m - E_{\pi_0^{(s)}, q^{(s)}} (T | \eta)) \xi_j \log q_j \tag{3.4.10} \]

**M-Step:** In the M-step, \( Q(\pi_0, q) \) is maximized to obtain the next approximation \( \pi_0^{(s+1)}, q^{(s+1)} \), that is,
\[
\frac{\partial Q}{\partial \pi_0} = 0 \implies \pi_0^{(s+1)} = \frac{E_{\pi_0^{(s)}, q_0^{(s)}}(T|\eta)}{m}
\]

\[
E_{\pi_0^{(s)}, q_0^{(s)}}(w_i|\eta) = P_{\pi_0^{(s)}, q_0^{(s)}}(w_i = 1|\eta)
= \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1 - \pi_0^{(s)}) \prod_{j=1}^k \left[q_j^{(s)}\right]^{x_{ij}}}
\]

\[
E_{\pi_0^{(s)}, q_0^{(s)}}(T|\eta) = \sum_{i=1}^m E_{\pi_0^{(s)}, q_0^{(s)}}(w_i|\eta)
\]

\[
\frac{\partial Q}{\partial q} = 0 \implies q_j^{(s+1)} = \frac{\xi_j}{m - E_{\pi_0^{(s)}, q_0^{(s)}}(T|\eta)} = \frac{\xi_j}{m(1 - \pi_0^{(s+1)})}
\]

The results are summarized in the following theorem.

**Theorem 3.3.** Let \(\pi_0^{(s)}, q_1^{(s)}, \cdots, q_k^{(s)}\) be the \(s^{th}\) approximation to \(\hat{\pi}_0, \hat{q}_1, \cdots, \hat{q}_k\) that maximize

\[
\prod_{i=1}^m \left\{\pi_0 + (1 - \pi_0) h(p_i|q_1, \cdots, q_m)\right\},
\]

where \(h\) is a multinomial model. Then the \((s+1)^{th}\) approximation with the EM algorithm is given by

\[
\pi_0^{(s+1)} = \frac{1}{m} \sum_{i=1}^m \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1 - \pi_0^{(s)}) \prod_{j=1}^k \left[q_j^{(s)}\right]^{x_{ij}}}
\]

and

\[
q_j^{(s+1)} = \frac{\xi_j}{m(1 - \pi_0^{(s+1)})}
\]

It is well known that each EM iteration gets closer to the maximum of log-likelihood but only
in a linear convergency rate. If the components are similar in their densities, then the convergence
is extremely slow. The convergence will also be slow when the maximum likelihood solution re-
quires some of the weight parameters to be zero, because the algorithm can never reach such a
boundary point. An additional and related problem is that of deciding when to stop the algorithm.
One risk to a naive user is the natural tendency to use a stopping rule for the algorithm based on
the changes in the parameters or the likelihood being sufficiently small. Taking smalls steps does
not indicate that we are close to the solution.

To combat this problem, Lindsay (1995) exploited the regularity of the EM algorithm process to
predict via the device known as Aitken acceleration, the value of the log likelihood at the maximum
likelihood solution. The Aitken’s acceleration rule is usually recommended to predict to maximum
efficiently and suitable whenever one is using a linearly convergent algorithm with a slow rate
of convergence. If we let $l_{i-2}, l_{i-1}$, and $l_i$ be the log-likelihood values for the three consecutive
iterations, then the predicted final value is

$$l_i^\infty = l_{i-2} + \frac{l_{i-1} - l_{i-2}}{1 - c_i}$$

where $c_i = \frac{l_{i-1} - l_{i-2}}{l_{i-1} - l_{i-2}}$ and terminate the EM iteration when $l_i^\infty - l_i$ is sufficiently small.

If the algorithm is moving slowly, then $c_i$, an estimate of the rate will be close to 1, and $l_i^\infty$
will be substantially larger than $l_i$. The Aitken accelerated value can also be used to construct a
stopping rule that more adequately captures the desired numerical accuracy than the usual "lack of
progress" stopping rule criterion

$$\text{stop if } l_i - l_i^\infty < tol$$

where $tol$ is the prespecified tolerance level. Provided that $l_i^\infty$ is a good estimator of the final
likelihood, the rule

$$\text{stop if } l_i - l_i^\infty < tol$$
will cause the algorithm to be terminated only when the solution is near, and the $tol$ will more meaningfully represent the actual accuracy attained. Below is the EM algorithm to compute $\hat{\pi}_0$.

**Algorithm 3.1: EM Iteration**

**Input:** $\xi_{ij}, i = 1, \cdots, k, j = 1, \cdots, m$ transformed from the observed $p$-values $p_1, \cdots, p_m$

**Output:** Estimate of $\pi_0$: $\hat{\pi}_0$

1 begin
2    Initialization: Set $\pi_0 = \pi_0^{(o)}, q_j = q_j^{(o)}, 1 \leq j \leq k$.
3 repeat
4    Set $\pi_0 = \pi_0^{(s)}, q_j = q_j^{(s)}$ be the $s^{th}$ (current) approximation.
5     Compute
6     $$\pi_0^{(s+1)} = \frac{1}{m} \sum_{i=1}^{m} \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1 - \pi_0^{(s)}) \prod_{j=1}^{k} [q_j^{(s)}]^{\xi_{ij}}}$$
7     and
8     $$q_j^{(s+1)} = \frac{\xi_j}{m(1 - \pi_0^{(s+1)})}$$
9 until $l(\pi_0^{(s+1)}, q_1^{(s+1)}, \cdots, q_k^{(s+1)}) - l(\pi_0^{(s)}, q_1^{(s)}, \cdots, q_k^{(s)}) \leq \epsilon$;
10 then $\hat{\pi}_0 = \pi_0^{s+1}$. 

CHAPTER 4 SIMULATION STUDIES

In order to investigate the properties of the estimators described in chapter 2 and compare the performance to the newly developed estimator described in section 3.4, we carried out a simulation experiment. The generation of simulated data and the calculation of the estimates were both done in the language $R$. The simulation study examines the case of randomly generated independent normally distributed data.

4.1 Testing Scenario

Simulation studies are conducted with the $p$-values based on a one-sided $z$-test in the finite normal mixture models for performance comparisons. For each $i = 1, \ldots, m$, we tested

$$H_{0i} : \mu_i = 0 \text{ vs } H_{1i} : \mu_i \neq 0.$$  

Monte Carlo data are simulated independently from

$$z \sim \pi N(0, 1) + (1 - \pi) N(1, 1)$$

and for each $i$, the $p$-value is computed by $p_i = 1 - \phi(z)$.

4.2 Generation of simulated data

For the generation of simulated data, the distributions were estimated by $N = 1000$ simulation runs with $m = 200, 500, 1000$ independent observations, at each true value of $\pi_0 = 0.25, 0.5, 0.75$, generated from a mixture of two normal distributions as follows,

(i) generate $z_1, \ldots, z_m$ from Bernoulli ($\pi_0$), and then
(ii) generate $x_1, \cdots, x_m$ according to

$$x_i \sim N(z_i, 1), \quad i = 1, \ldots, m,$$

In each case, 1000 Monte Carlo trials were performed. In implementation of the EM algorithm to compute $\hat{\pi}_0$, the determination rule of the EM algorithm is $l^\infty_i - l_i \leq 10^{-4}$. The EM algorithm converges for all the simulated data sets. Figures 4.1 - 4.6 show the histograms of simulated mixture dataset for $m = 1,000$ independent observations and simulated mixture densities with true value $\pi_0 = 0.25, 0.5, 0.75$ for $m = 1,000$ independent observations respectively.

4.3 Results of simulation study

The results of the simulation study are summarized in Table 4.1. The performance of the new estimator is compared with several existing procedures. Specifically,

1. the estimate $\hat{\pi}_0^s$ which is the Storey’s estimator described in section 2.1

2. the estimate $\hat{\pi}_0^c$ which is the Convex decreasing estimator described in section 2.2

3. the estimate $\hat{\pi}_0^a$ which is the Average estimate approach described in section 2.3

For procedure 1, Storey’s estimator was computed using the $R$ package qvalue, with the function qvalue. For procedure 2, the $R$ package limma, with the function convest, was employed in the computation of the estimator. Procedure 3 was computed using the $R$ language.

The newly innovative estimating procedure for $\pi_0$ shows satisfactory performance in comparison to other known methods. Table 4.1 displays the estimates of $\pi_0$ compared for the new procedure, Storey’s $\hat{\pi}_0^s$, Langaas’s non-parametric maximum likelihood approach (convest) $\hat{\pi}_0^c$, and average estimate $\hat{\pi}_0^a$. All estimators of $\pi_0$ are developed under the assumption of independence between the $p$-values.
Figure 4.1: Histogram of simulated mixture dataset with true value $\pi_0 = 0.25$ for 1,000 independent observations

Figure 4.2: Simulated mixture density with true value $\pi_0 = 0.25$ for 1,000 independent observations
Figure 4.3: Histogram of simulated mixture dataset with true value $\pi_0 = 0.5$ for 1,000 independent observations

Figure 4.4: Simulated mixture density with true value $\pi_0 = 0.5$ for 1,000 independent observations
Figure 4.5: Histogram of simulated mixture dataset with true value of $\pi_0 = 0.75$ for 1,000 independent observations.

Figure 4.6: Simulated mixture density with true value $\pi_0 = 0.75$ for 1,000 independent observations.
In our simulation study, the new estimating procedure performance is consistent with respect to the true values of $\pi_0$. The average estimate $\hat{\pi}_0$ is a conservative estimator as seen in table 4.1. In other words, it gives positively biased $\pi_0$ estimates on average. The Langaas’s non-parametric maximum likelihood approach ($\text{convest}$) $\hat{\pi}_0$ and Storey’s $\hat{\pi}_0^s$ are also considered to be a conservative estimate. However, $\text{convest}$ underestimated $\pi_0$ for true value of $\pi_0 = 0.75$ for 1,000 hypotheses tests.

Table 4.2 summarizes the standard deviation results. The standard error of the estimates are estimated by

$$
\hat{\text{se}} = \sqrt{\frac{1}{l-1} \sum_{i=1}^{l} (\hat{\pi}_{0i} - \frac{1}{l} \sum_{i=1}^{l} \hat{\pi}_{0i})^2}
$$

where $\hat{\pi}_{0i}$ estimates $\pi_0$ for the $i$th simulation and $\pi_0$ is the true value.

The average of the true false discovery rate (FDR) from 1000 simulations is also compared in this simulation study by applying Benjamini and Hochberg’s adaptive FDR controlling procedure with $\pi_0$ estimated using the above mentioned existing methods and the newly innovative method. The FDR significance level was chosen as $\alpha = 0.05$. For the purpose of comparison, the original BH FDR controlling procedure ([Benjamini and Hochberg (1995)](1995)) with the incorporation of the true value of $\pi_0$ were also applied to the $p$-values as seen in figure 4.8.

Storey’s $\hat{\pi}_0^s$, Langaas’s non-parametric maximum likelihood approach ($\text{convest}$) $\hat{\pi}_0$, and average estimate $\hat{\pi}_0$ produce higher FDRs estimate than the pre-chosen level because all three methods are conservative. The newly innovative method estimates $\pi_0$ and its FDR is below but very close to the pre-chosen significance level $\alpha = 0.05$. Figure 4.8 demonstrates that the FDR for the proposed new estimate has the relatively small variation.
Table 4.1: The estimate, $\hat{\pi}_0$, of the proportion of true null hypotheses is compared for: the proposed new innovative estimating procedure $\hat{\pi}_0^n$, Storey’s $\hat{\pi}_0^s$, Langaas’s nonparametric maximum likelihood approach (convest) $\hat{\pi}_0^c$, and average estimate approach with $B$ chosen via the bootstrapping procedure, $\hat{\pi}_0^a$. There are 1,000 simulated data sets, each with a total of $m=1,000$ hypotheses tests, for each value of $\pi_0$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$\pi_0$</th>
<th>$\hat{\pi}_0^n$</th>
<th>$\hat{\pi}_0^s$</th>
<th>$\hat{\pi}_0^c$</th>
<th>$\hat{\pi}_0^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.25</td>
<td>0.235</td>
<td>0.30</td>
<td>0.314</td>
<td>0.430</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>0.489</td>
<td>0.613</td>
<td>0.584</td>
<td>0.626</td>
</tr>
<tr>
<td>200</td>
<td>0.75</td>
<td>0.78</td>
<td>0.717</td>
<td>0.685</td>
<td>0.773</td>
</tr>
<tr>
<td>500</td>
<td>0.25</td>
<td>0.249</td>
<td>0.34</td>
<td>0.33</td>
<td>0.491</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>0.52</td>
<td>0.646</td>
<td>0.623</td>
<td>0.704</td>
</tr>
<tr>
<td>500</td>
<td>0.75</td>
<td>0.72</td>
<td>0.806</td>
<td>0.785</td>
<td>0.819</td>
</tr>
<tr>
<td>1000</td>
<td>0.25</td>
<td>0.242</td>
<td>0.393</td>
<td>0.349</td>
<td>0.394</td>
</tr>
<tr>
<td>1000</td>
<td>0.5</td>
<td>0.51</td>
<td>0.61</td>
<td>0.61</td>
<td>0.63</td>
</tr>
<tr>
<td>1000</td>
<td>0.75</td>
<td>0.72</td>
<td>0.79</td>
<td>0.67</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Table 4.2: The standard deviation of $\pi_0$ estimate of the proportion of true null hypotheses is compared for: the proposed new innovative estimating procedure $\hat{\pi}_0^n$, Storey’s $\hat{\pi}_0^s$, Langaas’s non-parametric maximum likelihood approach ($\hat{\pi}_0^c$), and average estimate approach with $B$ chosen via the bootstrapping procedure, $\hat{\pi}_0^a$. There are 1,000 simulated data sets, each with a total of $m=1,000$ hypotheses tests, for each value of $\pi_0$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$\pi_0$</th>
<th>$\hat{\pi}_0^n$</th>
<th>$\hat{\pi}_0^s$</th>
<th>$\hat{\pi}_0^c$</th>
<th>$\hat{\pi}_0^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.25</td>
<td>0.016</td>
<td>0.021</td>
<td>0.022</td>
<td>0.030</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>0.035</td>
<td>0.043</td>
<td>0.041</td>
<td>0.044</td>
</tr>
<tr>
<td>200</td>
<td>0.75</td>
<td>0.055</td>
<td>0.051</td>
<td>0.048</td>
<td>0.054</td>
</tr>
<tr>
<td>500</td>
<td>0.25</td>
<td>0.011</td>
<td>0.015</td>
<td>0.014</td>
<td>0.022</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>0.023</td>
<td>0.028</td>
<td>0.027</td>
<td>0.031</td>
</tr>
<tr>
<td>500</td>
<td>0.75</td>
<td>0.032</td>
<td>0.036</td>
<td>0.035</td>
<td>0.035</td>
</tr>
<tr>
<td>1000</td>
<td>0.25</td>
<td>0.005</td>
<td>0.012</td>
<td>0.011</td>
<td>0.013</td>
</tr>
<tr>
<td>1000</td>
<td>0.5</td>
<td>0.016</td>
<td>0.019</td>
<td>0.019</td>
<td>0.019</td>
</tr>
<tr>
<td>1000</td>
<td>0.75</td>
<td>0.023</td>
<td>0.025</td>
<td>0.021</td>
<td>0.026</td>
</tr>
</tbody>
</table>
Estimates of proportion of true null hypotheses

Figure 4.7: The estimate, $\hat{\pi}_0$, of the proportion of true null hypotheses is compared for: the proposed new innovative estimating procedure, Storey’s $\hat{\pi}_0^s$, Langaas’s nonparametric maximum likelihood approach (convest) $\hat{\pi}_0^c$, and average estimate approach with $B$ chosen via the bootstrapping procedure, $\hat{\pi}_0^a$. There are 1,000 simulated data sets, each with a total of $m=1,000$ hypotheses tests, for each value of $\pi_0$. 
Figure 4.8: Simulation results of the False Discovery Rate (FDR) estimates at significance level $\alpha=0.05$ for the new procedure in comparison to the existing methods.
CHAPTER 5  REAL LIFE EXAMPLE: APPLICATION TO DNA MICROARRAY DATA

5.1 Introduction

To further evaluate the performance of the newly developed method in comparison to the other methods for real life application datasets, we consider applications in microarray studies, specifically gene expressions. Microarray data have two basic qualities, biological significance and statistical significance. The biological significance tells how much the expression of a gene is influenced by the condition under study, while the statistical significance tells how trustworthy the biological significance is, as described in Lee (2007).

Deciding whether a particular gene is differentially expressed across experimental conditions frequently employs classical hypothesis testing. The test alternatives for a gene may be stated as follows:

\[ H_0 : \text{Gene is similarly expressed} \]

\[ vs \]

\[ H_1 : \text{Gene is differentially expressed} \]

The key objective of any microarray study is to have a high probability of declaring a gene to be differentially expressed if it is truly differentially expressed, while keeping the probability of making false declaration of differential expression acceptably low. As microarray studies typically involves the simultaneous testing of hundreds or thousands of genes for differential expression, the probabilities of producing incorrect test conclusions must be controlled for the whole gene set and this was discussed in section [1.2].

Wit and McClure (2004) discussed that differential expression indicates the changing of tran-
scription levels across different phenotypes or conditions. The idea is that these transcriptions changes might be responsible for or caused by the change in phenotype. For example, the genes responsible for the presence of a certain disease X will be transcribed at a different rate than when the disease is absent. We consider two microarray dataset which are:

1. Leukemia Gene Expression dataset in [Golub et al. (1999)].

2. Breast Cancer Gene Expression dataset in [Hedenfalk et al. (2001)].

In the next sections, we will give a brief introduction to microarray technology, the Leukemia Gene Expression dataset, Breast Cancer Gene Expression dataset, and discuss the estimates of $\pi_0$ for the each data sets.

5.2 Microarray

According to [Allison et al. (2005)], microarray technology allows measurement of the levels of thousands of different RNA molecules at a given point in the life of an organism, tissue or cell. Microarray analysis is often considered a discovery based rather than hypothesis driven approach, largely due to the potential for discovering altered expression of novel genes for which little or no prior information was available to suggest a role in the disease or experimental condition examined. The comparison of healthy and diseased cells can yield vital information on the causes of diseases. In order to analyze any experimental data correctly, it is fundamental to understand the experiments that generated the data. In this section, we discuss fundamentals of gene expressions and microarray technologies.

5.2.1 Background on Cells, Genes, DNAs, and RNAs

The ever increasing rate at which genomes are being sequenced has opened a new area of genome research, functional genomics, which is concerned with assigning biological function to DNA sequences. With the complete DNA sequences of many genomes already known and the development of the human genome, an essential and formidable task is to define the role of each gene
and understand how the genome functions as a whole. Innovative approaches, such as the cDNA and oligonucleotide microarray technologies, have been developed to exploit DNA sequence data and yield information about gene expression levels for entire genomes (Dudoit et al. (2002)). Basic genetic notions useful for understanding microarray experiments are reviewed next.

According to Benjamin (2000) and Lee (2007) and many other authors, we present the genetic notions. A cell is the minimal unit of life. A life process involves a wide array of molecules ranging from water to small organic compounds (like fatty acids and sugars) and macromolecules (DNA, proteins, and polysaccharides), that define the structure of the cells. *Deoxyribonucleic acid* (DNA) molecules store information about the structure of macromolecules, allowing them to be made precisely according to cells’ specification and needs.

DNA is a very stable molecule that forms the "blueprint" of an organism or a cell. The DNA structure encodes information as a sequence of chemically linked molecules that can be read by the cellular machinery and guides the construction of the linear arrangements of protein building blocks, which eventually fold to form functional proteins. Each cell contains a complete copy of its genetic material in the form of DNA molecules. The DNA can be copied and passed on to cell’s progeny through a mechanism called replication. The genetic information can be copied as a transportable working copy composed of *ribonucleic acid* (RNA) molecules, which are closely related to DNA.

The core biochemical flow of genetic information can be summarized as the process of RNA synthesis. The RNAs that take part in the process of decoding genes into proteins are referred to as "informational RNAs" called messenger RNA, mRNA. Functional RNAs (cases where RNA itself is the final functional product) are transfer RNAs (tRNA) and the ribosomal RNA (rRNA), which are both part of the intricate protein synthesis machinery that translates the informational mRNA into protein.
Genes are units of the DNA sequence that control the identifiable hereditary traits of an organism. A gene is a segment of DNA that specifies a functional RNA which codes for a particular protein, the ultimate expression of the genetic information. The total set of genes carried by an individual or a cell is called its *genome*. The genome defines the genetic construction of a cell or the *genotype*. The *phenotype* is the total set of characteristics displayed by a cell under a particular set of environmental factors. The outward appearance of an organism (phenotype) may or may not directly reflect the genes that are present (genotype). Today, genome sequences of several species are known including bacteria, yeasts, and humans. Microarray technology allows us to study the expression of all genes in an organism simultaneously (Benjamin (2000) and Lee (2007)).

5.2.2 Gene Expression

Gene expression is the process by which mRNA, and protein is synthesized from the DNA template of each gene. In other words, the gene expression takes care of the conversion of the information from the gene into mRNA via transcription and then to protein via translation resulting in the phenotypic manifestation of the gene (See section 5.2.1).

Gene expression could also be characterized as the process in which the information stored in DNA is used to produce a functional gene product. Gene products are either proteins or non-coding RNAs, such as tRNA and rRNA, which play essential roles in protein synthesis, but do not code for proteins. Gene expression is regulated throughout the lifespan of an individual cell to control the cell’s functions, such as its metabolic activity. Gene expression plays a critical role in the morphological changes that take place in a developing embryo and fetus and in the differentiation of stem cells to form specialized cells.

DNA are made up of chains of chemical building blocks called *nucleotides*. Each nucleotides consists of a phosphate group, a deoxyribose sugar molecules, and one of four different nitroge-
nous bases usually referred to by their initial letters: guanine G, cytosine C, adenine A, or thymine T. Genetic information is encoded in DNA by the sequence of these nucleotides. The information stored in the sequence of nucleotides in terms of the four nitrogenous bases is analogous to a long word in a four-letter alphabet.

The specific base pairing of DNA is the mechanism by which encoded information can be transferred from generation to generation with very little alteration. The two chains of the DNA molecule are held together by hydrogen bonds between nitrogen bases, with base-pairing occurring according to the following rule: G pairs with C, and A pairs with T. While a DNA molecule is built from a four-letter alphabet, proteins are sequences of twenty different types of amino acids. The expression of the genetic information stored in the DNA molecule occurs in the following stages:

1. transcription, during which DNA is transcribed into messenger ribonucleic acid mRNA, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil U replacing thymine T;

2. processing of the RNA product.

3. translation, during which mRNA is translated to produce a protein.

4. post translational modification of the protein product.

The process sequence is illustrated by the following chart:

\[
\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{protein}
\]

The correspondence between DNA’s four-letter alphabet and a protein’s twenty-letter alphabet is specified by the genetic code, which relates nucleotide triplets to amino acids. The genome of a cell or an organism is made up of one or more long molecules of DNA that are organized into chromosomes. A chromosome consists of an uninterrupted length of double-stranded DNA that
contains many genes. The unit of replication is the chromosome. When a cell divides, all the chromosomes are replicated. When a chromosome is replicated, all its genes are replicated (see [Dudoit et al. (2003), Dudoit et al. (2002) and Lee (2007)].

Protein polymers of amino acids are the workhorse molecules of the cell, responsible, for example, serve as the building blocks of cellular structures, producing energy and important bio-molecules like DNA and proteins, and for reproducing the cell chromosomes. Proteins are the building blocks of muscles, skin, hair and as well as the enzymes that catalyze and control all chemical reactions in an organism, ranging from food digestion to nerve impulses (Efron et al. (2001)).

According to Efron et al. (2001), all living cells contain chromosomes, large pieces of DNA containing hundreds or thousands of genes, each of which specifies the composition and structure of a protein (or sometimes several related proteins). Protein polymers of amino acids are the workhorse molecules of the cell, responsible, for example, for cellular structure, producing energy and important bio-molecules like DNA and proteins, and for reproducing the cell chromosomes. Every cell in an organism has nearly the same set of chromosomes, and thus contains the same repertoire of proteins. However, cells have remarkably distinct properties, such as the differences between human eye cells, hair cells and liver cells, distinctions that are the result of differences in the abundance, distribution, and state of the cell proteins.

One of the seminal discoveries of molecular biology was that these changes in protein abundance are determined in part by changes in the levels of messenger RNA (mRNA), small and relatively unstable nucleic acid polymers that shuttle information from chromosomes to the cellular machines that synthesize new proteins. Thus, there is a logical connection between the state of a cell and the details of its protein and mRNA composition. Whereas it remains difficult to measure the abundances of a cell’s proteins, the recently developed DNA microarray makes it possible
to quickly and efficiently measure the relative representation of each mRNA species in the total cellular mRNA population, or in more familiar terms to measure gene expression levels.

Different aspects of gene expression can be studied using microarrays, such as expression at the transcription or translation level, and sub cellular localization of gene products. Most gene expression procedure are primarily focused on expression at the transcription stage, that is, on mRNA or transcript levels. Microarrays derive their power and universality from a key property of DNA molecules described above, complementary base-pairing, and the term hybridization is used to refer to the annealing of nucleic acid strands from different sources according to the base-pairing rules (Dudoit et al. (2002)).

5.2.3 Microarray Technology

Microarray technology is a powerful tool for studying complex diseases (Mootha et al. (2003)) and for assessing the effects of drugs (Salvatore et al. (2008)) at the molecular level. Microarrays are simply measurement assays. Just as one can measure the amount of insulin which is a product of gene in blood, we can measure the products of genes with microarray in any tissue. What distinguishes microarrays is their capacity to measure all gene transcripts at one (Allison et al. (2005)). Microarray is an experimental method by which thousands of genes can be printed on a small chip and their expression can be measured simultaneously (Schena et al. (1995)). It can be used to detect changes in gene expression between normal and abnormal cells, which enables scientists to detect novel disease-related genes (Singh et al. (2002)). Many statistical methods have been developed for this purpose (Cui et al. (2003)). Although other advanced genomics technologies, such as RNA sequencing (Nagalakshmi et al. (2008); Wilhelm et al. (2008)), have been developed, microarrays have been continuously used for broad biomedical studies (McLendon et al. (2008)).

Furthermore, since the structures of data from different genomics technologies are basically
similar, methods for analyzing microarray data can also be useful for analyzing other similar genomics data. Microarrays have become a central tool used in modern biological and biomedical research (Markitsis and Lai (2010)).

Microarrays allow large numbers of DNA clones with known sequences to be immobilized as an array of detection units known as probes, while the pool of RNAs to be examined (targets) is fluorescently labeled and then hybridized to the detectors. There are several types of microarray systems, including the cDNA microarrays developed in the Brown and Botstein labs at Stanford (Hughes et al. (2001)) and the high-density oligonucleotide chips from the Affymetrix company (Lockhart et al. (1996)); the brief description below focuses on two major kinds of microarrays - the oligonucleotide array and the cDNA microarray.

In an oligonucleotide array, the probes are short sequences designed to match parts of the sequence of known or predicted open reading frames. Although oligonucleotide probes are often used in "spotted" microarrays, the term "oligonucleotide array" most often refers to a specific technique. Oligonucleotide arrays are produced by printing short oligonucleotide sequences designed to represent a single gene or family of gene splice-variants by synthesizing this sequence directly onto the array surface instead of depositing intact sequences. Sequences may be longer or shorter probes depending on the desired purpose; longer probes are more specific to individual target genes, and shorter probes may be spotted across the array in higher densities and are cheaper to manufacture. An experimental sample is hybridized on the microarray, and the RNA expression of the gene is estimated by the difference in signal. There is some concern that subtracting mismatch numbers may actually degrade the inferences (Efron et al. (2001)).

A spotted cDNA microarray consist of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic arrayer. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring
the differential hybridization (hybridization is the biochemical method on which DNA microarray technology is based) of the two samples to the sequences onto glass substrate to form an array. Each probe contains a different, characteristic sequence that is specific to a different group of genes under study (Dudoit et al. (2002), Lockhart et al. (1996)).

For mRNA samples, the two samples or targets are reverse-transcribed into cDNA, labeled using different fluorescent dyes (usually a red-fluorescent dye, Cyanine 5 or Cy5, and a green-fluorescent dye, Cyanine 3 or Cy3), then mixed in equal proportions and hybridized with the arrayed DNA sequences or probes. The experimental sample is labeled with red dye and hybridized on the slide. As a control, a reference sample is labeled with green dye and hybridized on the same slide. After this competitive hybridization, the slides are imaged using a scanner and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples (Phimister (1999), Lee et al. (2000)).

Figure 5.1: An outline of the microarray methodology adapted from Kumar et al. (2012)

From either type of microarray, we obtain several thousand expression values, one or many for each gene. Microarrays in current use measure anywhere from 1,000 to 25,000 genes. In a typical study, a number of experimental samples are each hybridized to a different microarray to
learn about gene expression differences across different conditions (Efron et al. (2001). For example Alizadeh et al. (2000) studied gene expression patterns from tissue samples from a number of lymphoma patients and related gene expression to patient survival.

According to Dudoit et al. (2002), DNA microarrays are a new and promising biotechnology which allows the monitoring of expression levels in cells for thousands of genes simultaneously. Microarrays are being applied increasingly in biological and medical research to address a wide range of problems, such as the classification of tumors or the gene expression response of yeast to different environmental stress conditions ( Golub et al. (1999)).

An important and common question in microarray experiments is the identification of differentially expressed genes, that is, genes whose expression levels are associated with a response or co-variate of interest. The co-variates could be either

- polytomous, for example, treatment/control status, cell type, drug type or
- continuous, for example, dose of a drug, time,

and the responses could be, for example, censored survival times or other clinical outcomes (see Dudoit et al. (2003)).

Microarray technologies have found applications in many different areas such as gene discovery, disease diagnosis, and drug discovery. Given such huge amount of data (tens or hundreds of data points for thousands or tens of thousands of genes), an important part in microarray studies is to make sense of the data and to draw biologically meaningful conclusions. This requires to remove genes irrelevant to the learning problems at hand and to select a small number of genes expressed in biological samples under specific conditions. The problem is known in the machine learning community as the feature subset selection problem (where each gene is considered as a feature). Feature selection is essential to reduce the test errors in microarray data processing as the
number of genes is much larger than the number of available samples. Selecting expressed genes is particularly important in genomics studies. For example, in cancer classification and diagnosis, knowing when certain genes are expressed, which genes are suitable as marker genes, which genes are responsible for the change from normal to cancerous cells etc. can help understand the underlying molecular mechanisms and identify therapies targeted to different varieties of cancers (Chen and Mckee (2010)).

The advent of microarray technology has made possible genome-wide expression profiling of diseases such as cancer. It has been proposed that a distinct cancer taxonomy can be identified by thus analyzing global gene expression patterns, which includes sporadic breast cancer (Perou et al. (1999)), leukemia (Golub et al. (1999)), lymphoma (Alizadeh et al. (2000)), melanoma (Bittner et al. (2000)), as well as hereditary breast cancer (Hedenfalk et al. (2001)), and different forms of childhood cancer (Khan et al. (2001)) has been performed. In this dissertation, we consider two of the above examples which are leukemia (Golub et al. (1999)) and the hereditary breast cancer (Hedenfalk et al. (2001)).

5.3 Description of Microarray dataset

Microarray data from an individual array basically form just a high dimension multivariate observation of gene expressions. The array corresponds to an experimental unit and genes correspond to the variables measured on the unit. As discussed in section (5.2.3), in a two color systems, the red and green dyes often correspond to paired specimens from the same or different experimental units. In a typically microarray, there is only a single array for each experimental unit, with assumed relationships involving hundreds and thousands of genes on each array (Allison et al. (2005)). We discuss in the following sections: microarray datasets, and apply the new estimating procedure with other methods discussed in chapter 2 to the microarray datasets.
5.3.1 Microarray Data

Consider a DNA microarray experiment which produces expression data on \( m \) genes, that is, variables or features, for \( n \) mRNA samples, that is, observations, and further suppose that a response or co-variate of interest is recorded for each sample. Such data may arise, for example,

1. from a study of gene expression in tumor biopsy specimens from leukemia patients \( \text{Golub et al. (1999)} \): in this case, the response is the tumor type and the goal is to identify genes that are differentially expressed in the different types of tumors.

2. from a study of gene expression profiles in inherited breast cancer via BRCA mutation tumor samples \( \text{Hedenfalk et al. (2001)} \): in this case, the response is the mutation tumor and the goal is to identify the mutated genes that are differentially expressed.

According to \( \text{Dudoit et al. (2003)} \), the data for sample \( i \) consist of a response or co-variate \( y_i \) and a gene expression profile \( x_i = (x_{i1}, \ldots, x_{mi}) \), where \( x_{ji} \) denotes the expression measure of gene \( j \) in sample \( i \), \( i = 1, \ldots, n \), \( j = 1, \ldots, m \). The expression levels \( x_{ji} \) might be either absolute (oligonucleotide), for example, Affymetrix chips discussed in \( \text{Lockhart et al. (1996)} \) or relative with respect to the expression levels of a suitably defined common reference sample (spotted cDNA microarrays), for example, Stanford two-color discussed in \( \text{Eisen et al. (1998)} \). Note that the expression measures \( x_{ji} \) are in general highly processed data.

The raw data in a microarray experiment consist of image files, and important pre-processing steps include image analysis of these scanned images and normalization \( \text{Yang et al. (2001)} \). The gene expression data are conventionally stored in an \( m \times n \) matrix \( X = (x_{ji}) \), with rows corresponding to genes and columns corresponding to individual mRNA samples. In a typical experiment, the total number \( n \) of samples is anywhere between around 10 and a few hundreds, and the number \( m \) of genes is several thousands. The gene expression measures, \( x \), are generally continuous variables, while the responses or covariates, \( y \), can be either polytomous or continuous, and possibly censored, as discussed in section 5.2.3.
The pairs \( \{(x_i, y_i)\} \ i = 1, \cdots, n \), formed by the expression profiles \( x_i \) and responses or covariates \( y_i \), are viewed as a random sample from a population of interest. The population and sampling mechanism will depend on the particular application, for example, designed factorial experiment in yeast, retrospective study of human tumor gene expression. Let \( X_j \) and \( Y \) denote, respectively, the random variables that correspond to the expression measure for gene \( j, j = 1, \cdots, m \), and the response or covariate.

The goal is to use the sample data \( \{(x_i, y_i)\} \ i = 1, \cdots, n \) to make inference about the population of interest, specifically, test hypotheses concerning the joint distribution of the expression measures \( X = (X_1, \cdots, X_m) \) and response or covariate \( Y \). If a single test is considered for each gene (variable), the null hypothesis for testing that the gene is not differentially expressed between the treatment and the control can be stated as:

\[
H_i: \text{There is no association between } X_i \text{ and } Y
\]

Many microarray experiments are aimed at finding active gene. In the simplest case, one compares two different situations, for example, treatment versus control or two different conditions. The aims of such experiments is to

1. to find which genes behave differently under conditions;

2. to determine a measure of confidence for this different behavior for each gene.

In a typical microarray experiment, the gene expression in two groups of cells can be compared. For each gene, let \( \mu_1 \) and \( \mu_2 \) be the true mean intensities, in groups 1 and 2, respectively. To determine whether the gene is differentially expressed, one would consider the null and alternative hypotheses to be:

\[
H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2.
\]
On a microarray chip, a large number of genes can be monitored simultaneously, which provides researchers with measurement for each gene in each group. For example, to assess genes’ involvement in tumor growth, the expression of tens of thousands of genes can be measured in normal and cancerous cells. Depending on the number of microarray chips available, multiple measurements for the expression of each gene are obtained (Pounds and Morris (2003); Tamhane and Shi (2007)).

A standard approach to the multiple testing problem consists of two aspects:

1. computing a test statistic $T_j$ for each gene $j$, and

2. applying a multiple testing procedure to determine which hypotheses to reject while controlling a suitably defined Type I error rate,

In general, the appropriate test statistic will depend on the experimental design and the type of response or covariate. For example,

1. for binary covariates, one might consider a $t$-statistic or a Mann-Whitney statistic;

2. for polytomous responses, one might use an $F$-statistic; and

3. for survival data one might rely on the score statistic for the Cox proportional hazards model.

For each gene $j$, the null hypothesis $H_j$ is tested based on a statistic $T_j$ which is a function of $X_j$ and $Y$. The lower case $t_j$ denotes a realization of the random variable $T_j$. Although multiple testing is by no means a new subject in the statistical literature, DNA microarray experiments present a new and challenging area of application for multiple testing procedures because of the sheer number of tests.

A commonly used test statistic is the Student’s $t$-test (assuming equal variances) for testing the difference between two means. The $t$-statistic allows for simultaneous consideration of differential expression and variability (Dudoit et al. (2002)). A positive is claimed when $H_0$ is rejected in favor of $H_1$, and a negative when $H_0$ is not rejected. A positive means that the gene is declared
differentially expressed; a negative means that the gene is declared similarly expressed.

If we knew the true state of each gene, that is, whether it is truly differentially expressed or similarly expressed, then the results of testing $m$ genes simultaneously could be classified into four categories, each denoted by the random variable in parentheses: true positives $S$, false positives $V$, true negatives $U$ and false negatives $T$ as illustrated in Table 1.1. Ideally, one would like to minimize $V$ and $T$, and maximize $S$ and $U$ as described in Table 1.1.

As earlier mentioned, microarray studies typically involve the simultaneous testing of hundred or thousands of genes for differential expression, the probabilities of producing incorrect test conclusions (false positives and false negatives) must be controlled (see section 1.2).

To further evaluate the performance of the new method in comparison to the other methods for real life application data. The datasets are from the studies where each $m$ genes were tested for differential expression. The estimates were calculated using the same $R$ code as was used in the simulation experiment described in chapter 4.

5.3.2 Leukemia Gene Expression Data

Golub et al. (1999) was interested in the molecular classification of cancer. The cancer classification was divided into two challenges:

- identification of new cancer classes (class discovery) and

- assigning tumors to known already-defined classes (class prediction), which could reflect current states or future outcomes.

Golub et al. (1999) described a generic approach to cancer classification based on gene expression monitoring by DNA microarrays and applied to human acute leukemia as a test case. A class discovery procedure discovered the distinction between acute myeloid leukemia (AML) and
acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. A derived class predictor was able to determine the class of new leukemia cases. In summary, the discovery and prediction of cancer classes was feasible via gene expression monitoring to classify the cancerous cells.

The leukemia gene expression reported in Golub et al. (1999) consisted of 38 bone marrow samples, 27 acute lymphoblastic leukemia ALL, and 11 acute myeloid leukemia AML obtained from acute leukemia patients at the time of diagnosis. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6,817 human genes. But after pre-processing, there were only 3,051 genes accurately readable resulting in a microarray $3051 \times 38$ matrix. The data can be obtained from the R package multtest. A two-sample Welch $t$-statistic is computed for each gene and its two-sided $p$-value is computed under a central $t$-student distribution with 36 degrees of freedom (see Chen et al. (2012)).

Figure 5.2 displays the histogram of the $p$-values of the leukemia gene expression data. Figure 5.3 shows the density function of the leukemia gene expression data. The two figures indicate that the multinomial mixture model is appropriate.

5.3.3 Breast Cancer cDNA microarray data

Hedenfalk et al. (2001) considered gene expression in inherited BReast CAncer (BRCA). Hereditary breast cancers are due to mutations in the two breast cancer susceptibility genes BRCA1 and BRCA2. The histopathological changes in these cancers are often characteristic of the mutant gene. They hypothesized that the genes expressed by these two types of tumors are also distinctive, perhaps allowing to identify cases of hereditary breast cancer on the basis of gene-expression profiles. Both BRCA1 and BRCA2 encode large, multifunctional proteins, and both function as tumor suppressor genes. BRCA1 and BRCA2 proteins are thought to be involved in two main
Figure 5.2: Histogram of the $p$-values for the leukemia gene expression data reported in Golub et al. (1999)

Figure 5.3: Density function of the leukemia gene expression data reported in Golub et al. (1999)
fundamental cellular processes - DNA damage repair and transcriptional regulation, and therefore, play a role in ensuring the stability of the cell’s genetic material. When either of these genes is mutated, or altered, such that its protein product either is not made or does not function correctly, DNA damage may not be repaired properly. As a result, cells are more likely to develop additional genetic alterations that can lead to cancer.

BRCA1 is a large gene spread over 80 kilobytes (kb) of genomic DNA composed of 22 coding exons (Exons are parts of DNA that are converted into mature messenger RNA) that are transcribed into a 7.8 kb mRNA encoding a protein containing 1863 amino acids. The approximate molecular mass of the BRCA1 protein is 220 Kilodalton (kDa). BRCA2, is a large gene, consisting of 27 exons that encode a transcript of approximately 12 kb, contained within 70 kb of genomic sequence. The BRCA2 protein consists of 3418 amino acids, with an estimated molecular mass of 384 kDa.

The aim of the study was to determine whether they could detect differences in gene-expression profiles among tumors with BRCA1 mutations, tumors with BRCA2 mutations, and sporadic cases of breast cancer. They discovered that differential expression of hormone receptors in BRCA1-mutation-positive and BRCA2-mutation-positive breast cancers affects the gene-expression profiles of these tumors. Significantly different groups of genes are expressed by breast cancers with BRCA1 mutations and breast cancers with BRCA2 mutations (see [Hedenfalk et al. (2001)]).

Further studies (see [Hedenfalk et al. (2002)]) show that genomic approaches to classification of hereditary breast cancers are possible, and that further studies will likely identify the most significant genes that discriminate between subgroups and may influence prognosis and treatment. They also discovered that a large number of hereditary breast cancer cases cannot be accounted for by mutations in these two genes, and are believed to be due to as yet unidentified breast cancer predisposition genes (BRCA$x$). In summary, gene expression-based analysis of hereditary breast cancer can potentially be used for classification purposes, as well as to expand upon our knowledge of dif-
ferences between different forms of hereditary breast cancer. Their results suggest that a heritable mutation influences the gene-expression profile of the cancer, and that a patient’s genotype does in fact leave an identifiable trace on her/his cancer’s gene expression profile.

Hedenfalk et al. (2001) obtained samples of complementary DNA (cDNA) with 6512 cDNAs of which 5361 represent unique genes: 2905 are known and 2456 are unknown genes. The microarrays were hybridized and scanned, and image analysis was performed as described in section 5.2.3.

The data include information of the microarray study of hereditary breast cancer of Hedenfalk et al. (2001). Many cases of hereditary breast cancer are due to mutations in either the BRCA1 or the BRCA2 gene. The patients consisted of 23 with BRCA1 mutations (7 arrays), 17 with BRCA2 mutations (8 arrays), 20 with familial breast cancer (familial breast cancer is defined as a history of breast or ovarian cancer in at least one first-degree relative) but no BRCA1 or BRCA2 mutations, 19 with possibly familial breast cancer (defined as a history of breast or ovarian cancer in at least one second-degree relative) but no BRCA1 or BRCA2 mutations, and 34 with sporadic breast cancer to determine whether there are distinctive patterns of global gene expression in these three kinds of tumors.

Since one of the goals of this study was to find genes differentially expressed between BRCA1- and BRCA2-mutation positive tumors, the sporadic data and non BRCA mutations data were excluded from the study. If a gene had a measurement exceeding 20, then the gene is also eliminated. Thus after pre processing, there were only 3,170 genes accurately readable resulting in a microarray 3170x15 matrix (see Storey and Tibshirani (2003)). A two sample $t$-statistic is computed for each gene and its two sided $p$-value is computed under a central $t$-student distribution with 13 degrees of freedom.
Figure 5.4: Histogram of the $p$-values for the hereditary breast cancer gene expression data reported in Hedenfalk et al. (2001).

Figure 5.5: Density function of the hereditary breast cancer gene expression data reported in Hedenfalk et al. (2001).
5.4 Microarray Data Results

In implementation of the EM algorithm to compute $\hat{\pi}_0$, we use the termination rule $l_i^{\infty} - l_i \leq 10^{-4}$, owing to the use of the Aitken’s acceleration stopping rule to predict the maximum. The EM algorithm converges for all the gene expression datasets. As discussed in section 4.3, the EM algorithm is insensitive to the choice of the initial values. The proposed technique implicitly assumes that the $p$-values for the genes are independently and identically distributed according to the multinomial model.

5.4.1 Leukemia Gene Expression Data Results

For completion, the performance of the proposed estimate is compared with several existing procedures. The estimates for $\hat{\pi}_0$ by $\hat{\pi}^b_0$, $\hat{\pi}^c_0$ and $\hat{\pi}^a_0$ are also evaluated: $\hat{\pi}^b_0 = 0.46$, $\hat{\pi}^c_0 = 0.47$, and $\hat{\pi}^b_0 = 0.49$, which are about 10% – 15% higher than the estimate $\hat{\pi}^n_0 = 0.41$. Recalling that the estimates $\hat{\pi}^b_0$, $\hat{\pi}^c_0$ and $\hat{\pi}^a_0$ are usually conservative, the lower estimate of 0.41 for $\pi_0$ appears to be expected. The results are summarized in table 5.1.

5.4.2 Breast Cancer Gene Expression Data Results

For completion, the performance of the proposed estimate is compared with several existing procedures. The estimates for $\hat{\pi}_0$ by $\hat{\pi}^b_0$, $\hat{\pi}^c_0$ and $\hat{\pi}^a_0$ are also evaluated: $\hat{\pi}^b_0 = 0.67$, $\hat{\pi}^c_0 = 0.67$, and $\hat{\pi}^b_0 = 0.68$, which are about 10% – 15% higher than the estimate $\hat{\pi}^n_0 = 0.58$. Recalling that the estimates $\hat{\pi}^b_0$, $\hat{\pi}^c_0$ and $\hat{\pi}^a_0$ are usually conservative, the lower estimate of 0.58 for $\pi_0$ appears to be expected. The results are summarized in table 5.2. The newly estimated procedure provides an overall satisfactory performance. According to Hedenfalk et al. (2001), 176 genes were differentially expressed in tumors with BRCA1 mutations, and tumors with BRCA2 mutations. The newly estimated procedure’s result was 174 significant genes which is less than 2% (1.15% to be exact) lower that the exact value 176 significant genes using weighted gene analysis. Hence the proposed method for estimating $\pi_0$ provides an effective solution. The overall performance is satisfactory.
Table 5.1: The estimate of the proportion of true null hypotheses and the number of statistically significant genes for the leukemia gene expression data at significance level $\alpha = 0.05$ after applying BH’s adaptive FDR controlling procedure with $\pi_0$ estimated using the new innovative approach in comparison to other three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimates of $\pi_0$</th>
<th>Number of Significant genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\pi}_0^n$: New Estimator</td>
<td>0.4150 (0.0075)</td>
<td>1136</td>
</tr>
<tr>
<td>$\hat{\pi}_0^b$: Storey$^{\text{boot}}$</td>
<td>0.4643 (0.0084)</td>
<td>1098</td>
</tr>
<tr>
<td>$\hat{\pi}_0^c$: convest</td>
<td>0.4701 (0.0085)</td>
<td>1092</td>
</tr>
<tr>
<td>$\hat{\pi}_0^a$: Average$^{\text{boot}}$</td>
<td>0.4913 (0.0089)</td>
<td>1073</td>
</tr>
</tbody>
</table>
Table 5.2: The estimate of the proportion of true null hypotheses and the number of statistically significant genes for the breast cancer gene expression data at significance level $\alpha = 0.05$ after applying BH’s adaptive FDR controlling procedure with $\pi_0$ estimated using the new innovative approach in comparison to other three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimates of $\pi_0$</th>
<th>Number of Significant genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\pi}_0^0$: New Estimator</td>
<td>0.5776 (0.0106)</td>
<td>174</td>
</tr>
<tr>
<td>$\hat{\pi}<em>0^b$: Storey$</em>{boot}$</td>
<td>0.6763 (0.0122)</td>
<td>159</td>
</tr>
<tr>
<td>$\hat{\pi}_0^c$: convest</td>
<td>0.6707 (0.0121)</td>
<td>161</td>
</tr>
<tr>
<td>$\hat{\pi}<em>0^a$: Average$</em>{boot}$</td>
<td>0.6818 (0.0123)</td>
<td>159</td>
</tr>
</tbody>
</table>
CHAPTER 6 SUMMARY AND CONCLUSION

In this dissertation, we introduce the problem of estimating the proportion of the true null hypothesis, $\pi_0$, in a multiple testing problem setup. A new estimating procedure for estimating the proportion of the true null hypothesis $\pi_0$ is propose based on a multinomial mixture model.

We propose a finite mixture model of the uniform distribution and a multinomial distribution with a mixing proportion $\pi_0$ is used to fit the $p$-values to estimate $\pi_0$. The continuous-type observations $p_1, \cdots, p_m$ were converted to discrete data with $k$ categories. Selection of $k$ is data-based and we recommended Sturges’ rule modified for skewness. We later characterized an approximate maximum likelihood estimate for the multinomial mixture model via the expectation and maximization algorithm (EM algorithm).

Simulation study was conducted to ascertain the performance of the new estimating procedure which was discussed in chapter 4. In our simulation study, results in section 4.3 shows that our new estimating procedure outperforms the other methods considered in this study. We further employed the new estimating procedure in controlling the false discovery rate at a significance level of $\alpha = 0.05$ as shown in figure 4.8. The new estimating procedure presents satisfying performance in controlling the false discovery rate. When our proposed estimated value of $\pi_0$ was incorporated into Benjamini and Hochberg’s adaptive FDR controlling procedure, more statistical power is gained such that the FDR can be controlled below, yet extremely close to a desired level $\alpha$. The positive outlook of the simulation results encouraged a real life data application. In chapter 5, we applied the new estimating procedure to microarray datasets - Gene expression dataset in Golub et al. (1999) and the breast cancer cDNA dataset in Hedenfalk et al. (2001). The new estimating procedure provides an overall satisfactory results.

The comparison of the performance of the new estimating procedure with various other meth-
ods led to the discussion of other known methods for estimating the proportion of the true null hypothesis, \( \pi_0 \), which includes the Storey’s method by Storey (2002), convest method by Langaas et al. (2005) and the average estimate by Jiang and Doerge (2008). These estimating methods have been shown to be conservative by many authors. Simulation results and real life data application results in this dissertation confirm their conservative nature.

In our proposed method, the choice of partitioning in multinomial model is dependent on the shape of histogram shape. As we have observed that the histogram of the \( p \)-values are highly skewed, hence a modified sturges’ rule (see section 3.2.2) was used. Therefore in practice, we suggest to check the histogram shape before applying any statistical methods for estimating \( \pi_0 \). Histogram density estimation should be based on the histogram shape for selecting the optimal partition \( k \). If the histogram shape is highly skewed then we expect satisfactory estimation performance from our method.

Convergence in EM algorithm has always been a problem. However, Aitken’s acceleration rule was employed in choosing a suitable tolerance level for convergence. The details of the EM algorithm applied to a multinomial mixture model was discussed in section 3.3. The iterative steps for both the expectation and maximization steps were outlined in algorithm 4. While some authors and other estimating methods consider the non null hypothesis \( p \)-values to be approximately zero (see Storey (2002)), we decided to approximately those \( p \)-values using the multinomial model.

The newly proposed estimating procedure implicitly assumes that the \( p \)-values for the genes are independently and identically distributed. This assumption is made primarily for computational ease, but also could imply that genes behave independently, which is plausibly not true. The effect of gene networks and interactions, dependence among genes is still a difficult issue in microarray data analysis and in genomics (Efron and Tibshirani (2007)). However, as investigated by Benjamini and Yekutieli (2001), methods that are based on the independence assumption perform
quite well in general situations of weak positive dependence, and a positive dependency structure is common in many situations. Nonetheless, the newly estimating procedure appears to be a very useful and provides reasonable approximations for the simulation studies and the real life gene expression datasets.

Most of the existing procedures for utilizing false discovery rates in practice involve assumptions about the \( p \)-values being independent or weakly dependent. An area of current research is aimed at performing multiple hypothesis tests when there is dependence among the hypothesis tests, specifically at the level of the data collected for each test or the \( p \)-values calculated for each test. Recent proposals suggest modifying FDR controlling algorithms or extending their theoretical characterizations (Benjamini and Yekutieli (2001)), modifying the null distribution utilized in calculating \( p \)-values (Efron (2012)), or accounting for dependence at the level of the originally observed data in the model fitting (Leek and Storey (2008)).

As array technology improves, it is certain that the number of features per array will only increase, hence multiple testing problems will continue to be a challenging problem. Specific to microarrays, the false discovery rate (FDR) is preferred to family-wise error rate (FWER) because the FDR controlling procedures have more statistical power than the FWER controlling procedures, even at the cost of a few more false positives.

Microarrays have been widely used in biological and medical studies. They are useful in gene discovery, disease diagnosis, drug discovery, and toxicological research. A precise estimate of the proportion of differentially expressed genes is important in controlling the false positive and experiment design. Therefore, the improvement of existing estimation methods, in estimating the proportion of true null hypotheses \( \pi_0 \), still remains very important. The newly proposed innovative procedure for estimating \( \pi_0 \) provides an overall effective solution.
6.1 Future Research

In future, the new estimating procedure will be extended to weak positive dependence and a positive dependence structure. Genes interaction, correlation and network are other factor we wish to consider for future studies in estimating the proportion of true null hypothesis in multiple testing problem.
BIBLIOGRAPHY


