2012

Assessing differential expression when the distribution of effect sizes is asymmetric and evaluating concordance of differential expression across multiple gene expression experiments

Megan Christina Orr
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation
Orr, Megan Christina, "Assessing differential expression when the distribution of effect sizes is asymmetric and evaluating concordance of differential expression across multiple gene expression experiments" (2012). Graduate Theses and Dissertations. 12765. https://lib.dr.iastate.edu/etd/12765

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Assessing differential expression when the distribution of effect sizes is asymmetric and evaluating concordance of differential expression across multiple gene expression experiments

by

Megan Christina Orr

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Peng Liu, Co-major Professor
Dan Nettleton, Co-major Professor
Alicia L. Carriquiry
Max D. Morris
Daniel J. Nordman

Iowa State University
Ames, Iowa
2012

Copyright © Megan Christina Orr, 2012. All rights reserved.
DEDICATION

I would like to dedicate this work to my parents, John Orr and Jackie Howlett, and my fiancé, Bradley. Their endless emotional support has kept me sane throughout my graduate student career.
# TABLE OF CONTENTS

LIST OF TABLES ........................................ vi
LIST OF FIGURES .......................................... viii
ACKNOWLEDGEMENTS ................................. ix
ABSTRACT ................................................. x

CHAPTER 1. GENERAL INTRODUCTION ................. 1
  1.1 Multiple Testing .................................. 1
  1.2 False Discovery Rate ............................. 2
  1.3 Controlling False Discovery Rate ............... 3
    1.3.1 Benjamini and Hochberg (1995) .............. 3
    1.3.2 The $q$-value Method ....................... 4
    1.3.3 Estimating the Number of Genes that are Equivalently Expressed .......... 4
  1.4 Additional Research ............................. 5
  1.5 Organization ..................................... 6

CHAPTER 2. ESTIMATING THE NUMBER OF GENES THAT ARE DIFFERENTIALLY EXPRESSED IN BOTH OF TWO INDEPENDENT EXPERIMENTS ........................................ 9
  2.1 Introduction ...................................... 10
  2.2 Methods .......................................... 12
    2.2.1 Overview of $m_{11}$ Estimation .............. 12
    2.2.2 Review of the $\lambda$-Estimator and Histogram-Based Method .......... 13
    2.2.3 Estimating the Number of Genes that are Equivalently Expressed in Both Experiments ........................................ 17
2.2.4 Properties of \( \hat{m}_{11} \) .................................................. 17
2.3 Simulation Studies ............................................................ 19
  2.3.1 Simulations Using Independent, Normally Distributed Data .... 21
  2.3.2 Simulations Using Real Microarray Data .......................... 21
  2.3.3 Results ................................................................. 22
2.4 Real Data Analysis ........................................................... 28
2.5 Discussion ................................................................. 30
2.6 Appendix ................................................................. 31

CHAPTER 3. IDENTIFYING GENES THAT ARE DIFFERENTIALLY EXPRESSED IN BOTH OF TWO INDEPENDENT EXPERIMENTS . . . 36
3.1 Introduction ................................................................. 37
3.2 Review of FDR Analysis for One Experiment ......................... 39
  3.2.1 False Discovery Rate .................................................. 39
  3.2.2 Storey’s \( q \)-value .................................................. 40
  3.2.3 Estimation of \( m_0 \) by \( \lambda \) Selection ............................ 41
3.3 Extending False Discovery Rate Analysis to the Two Experiment Case .... 41
  3.3.1 Expression Status in Two Experiments .......................... 42
  3.3.2 Estimating Cross Classification Frequencies Based on the Differential Expression Statuses of Genes .................. 42
  3.3.3 Assessing False Discovery Rate in Two Experiments ......... 43
3.4 Methods for False Discovery Rate Analysis in Two Experiments .... 43
  3.4.1 Intersection Method ............................................... 44
  3.4.2 Proposed Method for Assessing False Discovery Rate using a Rejection Region Based on \( q^{\text{max}} \) ......................... 44
  3.4.3 An Alternative Strategy Based on Local False Discovery Rates .... 47
3.5 Simulation Studies .......................................................... 49
  3.5.1 Simulations Using Independent Normal Data .................... 51
  3.5.2 Simulations Using Gene Expression Data ........................ 51
  3.5.3 Estimation Constraints ............................................. 52
CHAPTER 4. AN IMPROVED METHOD FOR COMPUTING Q-VALUES WHEN THE DISTRIBUTION OF EFFECT SIZES IS ASYMMETRIC .

4.1 Introduction .................................................. 63
4.2 Methods ...................................................... 65
   4.2.1 Review of Storey’s $q$-value Method .................. 65
   4.2.2 False Discovery Rate Estimation Using Two Subsets of $p$-values .... 67
   4.2.3 Advantages of the Proposed Method ................... 68
4.3 Simulation Studies ............................................ 70
   4.3.1 Simulations Using Independent Normal Data ............ 71
   4.3.2 Simulations Using Gene Expression Data ................ 72
   4.3.3 Results .................................................. 72
4.4 Real Data Analysis ........................................... 76
4.5 Discussion .................................................... 78

CHAPTER 5. GENERAL CONCLUSIONS .............................. 81

5.1 Summary ...................................................... 81
5.2 Future Research ............................................... 82
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>Outcomes when testing $m$ null hypotheses.</td>
<td>2</td>
</tr>
<tr>
<td>Table 2.1</td>
<td>Contingency table of frequencies based on cross classification of the expres-</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>sion status (EE or DE) for each of the $m$ genes by experiment.</td>
<td></td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Mean estimates of $m_{11}$ with corresponding RMSEs for simulations using independent normal data</td>
<td>24</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Mean estimates of $m_{11}$ with corresponding RMSEs for simulations using microarray data</td>
<td>25</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Table of random variables corresponding to frequencies associated with all four outcomes when testing $m$ null hypotheses.</td>
<td>39</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Contingency table of frequencies based on cross classification of the expres-</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>sion status (EE or DE) for each of the $m$ genes by experiment.</td>
<td></td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Table of random variables corresponding to frequencies associated with all four outcomes when testing $m$ pairs of hypotheses.</td>
<td>44</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Mean pAUC for both Simulation Studies</td>
<td>54</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Mean $S$ and $V/R$ for simulations using independent normal data.</td>
<td>55</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>Mean $S$ and $V/R$ for simulations using microarray data</td>
<td>56</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>The number of genes declared to be DE in the maize leaf experiments</td>
<td>58</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Table of random variables corresponding to frequencies associated with all four outcomes when testing $m$ hypotheses.</td>
<td>66</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Mean pAUC, mean $S$, and $V/R$ for simulations using independent normal data</td>
<td>74</td>
</tr>
</tbody>
</table>
Table 4.3  Mean pAUC, mean $S$, and $V/R$ for simulations using microarray data 75

Table 5.1  Contingency table of frequencies based on cross classification of the expression status when taking direction of change into account. . . . . 82
LIST OF FIGURES

| Figure 2.1 | Scatterplot of \( p \)-values from two experiments | 16 |
| Figure 2.2 | Boxplots of the \( m_{11} \) estimates for simulations using independent normal data | 26 |
| Figure 2.3 | Boxplots of the \( m_{11} \) estimates for simulations using microarray data | 27 |
| Figure 4.1 | Histogram of test statistics from data with asymmetric effect sizes | 64 |
| Figure 4.2 | Histograms of \( p \)-values based on sign of test statistic | 69 |
| Figure 4.3 | The densities of effect sizes for DE genes for different values of \( \pi_A \) | 71 |
| Figure 4.4 | Scatterplot of the test statistic versus the log ratio of the \( q \)-values | 77 |
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my major professors, Dr. Peng Liu and Dr. Dan Nettleton, for their guidance, support, and encouragement over the last four years of my graduate student career. I have learned so much from them and will forever be indebted to them for everything they have taught me. I would also like to thank my remaining committee members, Dr. Alicia Carriquiry, Dr. Max Morris, and Dr. Dan Nordman, for taking an interest in my research. Finally, I would like to thank my friends and family for their everlasting emotional support.
ABSTRACT

The emergence and development of gene expression technologies has resulted in an ever-increasing number of high-dimensional data sets available for analysis. The availability of these data sets has prompted much research into the development of methods for statistically analyzing gene expression experiments. Many of these methods focus on identifying genes that are differentially expressed (DE), i.e., exhibit changes in mean expression levels between treatments, in a single experiment. This dissertation presents novel methods for detecting differential expression in one experiment and proposes methods for analyzing gene expression data from two independent experiments.

Many methods have been proposed for estimating the number of genes that are equivalently expressed (EE), and thus the number of DE genes, in a single gene expression experiment, but many researchers are interested in comparing the results of two independent experiments. Estimating the number of genes that are DE in two independent experiments is generally performed in two steps. First, data from each experiment are analyzed separately, and a list of genes identified as DE is obtained for each experiment. Each list is generally produced by a method that attempts to control false discovery rate (FDR) at some desired level $\alpha$. Then, the number of genes common to both lists is used as an estimate of the number of genes DE in both experiments. A major flaw of this method is that the resulting estimates can vary greatly depending on the value of $\alpha$. Chapter 2 proposes a new method that estimates the number of genes that are DE in both of two independent experiments, which includes analyzing the $p$-values from each experiment simultaneously, and results in a single estimate that does not depend on $\alpha$. Through simulation studies, we show the advantages of our approach. In Chapter 3, we extend the idea of Chapter 2 by proposing a new method for identifying genes that are DE in both of two independent experiments while controlling FDR and compare this method to two existing methods. These three methods are compared through simulation studies that
show the proposed method controls FDR better as well as provides similar or better power when compared to the existing methods.

Chapter 4 proposes a new method for calculating $q$-values when the distribution of effect sizes in a gene expression experiment is asymmetric. This method first estimates the number of genes that are EE in an experiment based on the distribution of all $p$-values. Then, the $p$-values are split into two subsets based on the signs of their corresponding test statistics, and $q$-values are then calculated separately for each subset. Simulation study results show that the proposed method, when compared to the traditional $q$-value method, generally provides a better ranking for genes as well as a higher number of truly DE genes identified as DE, while still adequately controlling FDR.
CHAPTER 1. GENERAL INTRODUCTION

The prevalence of high-dimensional data analysis has skyrocketed in recent years as advances in technology have allowed for the generation of such data sets. This is especially the case in gene expression experiments, where the abundance of mRNA transcripts, known as gene expressions, are measured for thousands of genes in subjects from different treatment groups. A major goal in the analysis of these experiments is to identify genes that exhibit differential expression, i.e., a change in mean expression levels across treatments. Identifying such genes is typically accomplished by performing a hypothesis test for each gene. The ever-increasing number of available gene expression data sets has prompted much research in the area of multiple testing. Because of the high number of hypotheses that are being tested simultaneously (usually tens of thousands), traditional multiple testing methods are not appropriate for analyzing gene expression data. This chapter briefly describes issues with using traditional multiple testing approaches to analyze high-dimensional data sets, and introduces multiple testing methods that have been developed specifically for analyzing gene expression data with the goal of identifying differentially expressed (DE) genes.

1.1 Multiple Testing

In a typical multiple testing setting, \( m \) \( p \)-values are obtained from testing each of \( m \) hypotheses. Then, tests with \( p \)-values below a chosen cutoff \( c \) are declared to be significant. Table 1.1 presents the frequencies associated with the four possible outcomes when testing \( m \) hypotheses in an experiment.

The cutoff \( c \) is generally chosen in order to control some type of error rate at level \( \alpha \). Traditionally, statisticians have chosen to control family-wise error rate (FWER) in multiple
Table 1.1: Outcomes when testing $m$ null hypotheses.

<table>
<thead>
<tr>
<th>Null true</th>
<th>Null false</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U$</td>
<td>$T$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>$V$</td>
<td>$S$</td>
<td>$m - m_0$</td>
</tr>
<tr>
<td>$m - R$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

testing situations. FWER is defined as $\Pr(V > 0)$, or the probability of committing at least one Type I error.

Many methods for determining an appropriate cutoff $c$ that controls FWER have been proposed. The most well-known and simplest approach is the Bonferroni method (Bland and Altman, 1995) which conservatively controls FWER at $\alpha$ by setting $c = \alpha/m$. This method can be useful in traditional multiple testing situations when only a handful of hypothesis tests are to be performed, but for experiments that test thousands of hypotheses simultaneously, the power of declaring a test significant when the null hypothesis is false is extremely low. For example, if a gene expression data set has $m = 20,000$ genes and we wish to control FWER at $\alpha = 0.05$, then the $p$-value cutoff for declaring a gene DE would be $c = 0.0000025$. This is an extremely stringent cutoff that can lead to many Type II errors, and very few, if any, rejections.

Holm (1979) proposed a less conservative and slightly more powerful method for controlling FWER at level $\alpha$. This approach uses the ordered $p$-values, $p(1), \ldots, p(m)$, to find the largest integer $k$ such that

$$p(i) \leq \frac{\alpha}{(m - i + 1)}$$

for all $i = 1, \ldots, k$, and sets the cutoff to be $c = p(k)$. If no such $k$ exists, then $c = 0$ and no tests are declared significant. Although Holm’s method is more powerful than Bonferroni’s method, it still results in low power for cases in which the null hypothesis is false if thousands of hypotheses are to be simultaneously tested.

1.2 False Discovery Rate

Storey and Tibshirani (2003) argued that FWER is too conservative of an error rate to control when $m$ is large and that false discovery rate (FDR) is a more reasonable measure of
the balance between the number of “false discoveries” \((V)\) and the number of “true discoveries” \((S)\). In terms of gene expression analysis, a “discovery” is a gene that is declared to be DE, a “false discovery” is an EE gene that is declared to be DE (or alternatively, a Type I error), and a “true discovery” is a DE gene that is correctly declared to be DE.

False discovery rate (FDR) was first introduced by Benjamini and Hochberg (1995) and is formally defined as

\[
\text{FDR} = E\left[\frac{V}{\max\{R, 1\}}\right],
\]

the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses.

### 1.3 Controlling False Discovery Rate

Many methods have been proposed for identifying genes that are DE while controlling false discovery rate. This section introduces and describes some of the most well-known and commonly-used methods. For each method, consider the problem of testing \(m\) null hypotheses \(H_1, \ldots, H_m\) based on their corresponding \(p\)-values \(p_1, \ldots, p_m\), where \(H_j\) and \(p_j\) represent the null hypothesis and \(p\)-value, respectively, for gene \(j\). Assume that if \(H_j\) is true (i.e., gene \(j\) is equivalently expressed across treatments denoted by EE) then \(p_j \sim \text{Uniform}(0,1)\) distribution, and if \(H_j\) is false, then \(p_j\) follows a distribution stochastically smaller than the \(\text{Uniform}(0,1)\) distribution. Also, let \(p_{(1)}, \ldots, p_{(m)}\) denote the \(m\) \(p\)-values ordered from smallest to largest.

#### 1.3.1 Benjamini and Hochberg (1995)

Benjamini and Hochberg (1995) proposed a method for controlling FDR at level \(\alpha\) by finding the largest integer \(k\) such that

\[
p_{(k)} \leq \frac{k\alpha}{m},
\]

and setting \(c = p_{(k)}\). Then, all genes with \(p\)-values less than or equal to \(c\) are declared to be DE. If no such \(k\) exists, then \(c = 0\) and no gene are declared DE.

Notice that (1.3) implies that

\[
\alpha \leq \frac{p_{(k)} m}{k}.
\]

(1.4)
The quantity on the right side of this inequality is the expected proportion of false discoveries among all discoveries when all genes with \( p \)-values less than or equal to \( p(k) \) are declared DE if all \( m \) genes are EE, (i.e., \( m_0 = m \)). This follows from the assumption that \( p \)-values from EE genes are uniformly distributed. Thus, it is not surprising that Benjamini and Hochberg proved that their procedure actually controls FDR at level \( \alpha(m_0/m) \) rather than at level \( \alpha \).

If the number of truly EE genes, \( m_0 \), is close to \( m \), this will not affect the results of the experiment greatly. On the other hand, if a large proportion of the \( m \) genes are DE, or if \( m_0 \) is much smaller than \( m \), then many more Type II errors will be committed than necessary, and potentially important information about which genes are DE will be lost.

### 1.3.2 The \( q \)-value Method

An improved method for controlling FDR at level \( \alpha \) would be to replace the quantity \( m \) with \( m_0 \) in (1.3). Unfortunately, \( m_0 \) is unknown, so this quantity must be estimated. Thus, as proposed by Storey (2002), an appropriate approach for identifying genes that are DE while controlling FDR at \( \alpha \) would be to follow the same procedure as described in Section 1.3.1, but replace \( m \) with \( \hat{m}_0 \), the estimator of \( m_0 \). Storey (2002) also defined the quantity called the \( q \)-value as

\[
q(j) = \min \left\{ \frac{p(r)\hat{m}_0}{r} : r = j, \ldots, m \right\},
\]

where \( q(j) \) is the \( q \)-value that corresponds to the gene with the \( j \)th smallest \( p \)-value, and \( q_j \) corresponds to the \( q \)-value for gene \( j \).

The \( q \)-value converts the \( p \)-value from a significance measure of the Type I error rate of a single hypothesis test to a significance measure of the FDR for a family of \( m \) hypothesis tests. Specifically, the \( q \)-value for a given gene represents the estimated FDR if this gene and all genes with smaller \( q \)-values are declared to be DE.

### 1.3.3 Estimating the Number of Genes that are Equally Expressed

Estimating the number of genes that are EE in an experiment, \( m_0 \), is a necessary element in the process of estimating FDR using the \( q \)-value method. Many proposed methods attempt
to do this by estimating the $p$-value density at 1 (which corresponds to estimating $\pi_0 = m_0/m$, the proportion of genes that are EE) and multiplying this value by $m$. Storey (2002) proposed a conservative estimator for $m_0$ as

$$\hat{m}_0(\lambda) = \frac{\sum_{j=1}^{m} 1\{p_j > \lambda\}}{(1 - \lambda)}$$

for a fixed $\lambda \in (0, 1)$.

Although any value of $\lambda$ will result in a conservative estimator of $m_0$, adaptively choosing $\lambda$ based on the data can improve $\hat{m}_0(\lambda)$. Storey (2002) and Storey et al. (2004) discuss how, in most cases, the bias of $\hat{m}_0(\lambda)$ decreases but the variance increases as $\lambda \to 1$, and both papers present similar algorithms that use bootstrapping approaches to choose a $\lambda$ that attempts to minimize the mean square error of $\hat{m}_0(\lambda)$. Nettleton et al. (2006) introduces an algorithm that is slightly modified in Liang and Nettleton (2012), and originally proposed by Mosig et al. (2001), that chooses a $\lambda$ in which the distribution of the $p$-values smaller than $\lambda$ have an approximately decreasing distribution and the distribution of the $p$-values larger than $\lambda$ follow an approximately uniform distribution.

Storey and Tibshirani (2003) introduced another approach for estimating $m_0$. This method involves calculating $\hat{m}_0(\lambda)$ for a series of $\lambda$ values between 0 and 1, fitting a natural cubic spline between the points $(\lambda, \hat{m}_0(\lambda))$ to estimate a functional relationship between $\lambda$ and $\hat{m}_0(\lambda)$. The final estimate of $m_0$ is obtained by evaluating this estimated function at $\lambda = 1$.

### 1.4 Additional Research

Much research has been done in the areas of $m_0$ estimation and FDR estimation. The reader is directed to the following papers for additional references on these areas of research: Benjamini and Yekutieli (2001), Dudoit et al. (2002), Efron (2004), Langaas et al. (2005), Pounds and Cheng (2006), Genovese and Wasserman (2004), Efron (2007), Sun and Cai (2007), Leek and Storey (2008), Blanchard and Roquain (2009), and Gavrilov et al. (2009).
1.5 Organization

The rest of this dissertation focuses on identifying genes that are DE in different multiple testing situations. Chapter 2 extends and alters the idea of estimating $m_0$ to the two experiment case. In this paper, we estimate $m_{11}$, the number of genes that are DE in both of two independent experiments. Chapter 3 uses methods introduced in Chapter 2 to identify genes that are DE in two independent experiments. Chapter 4 introduces a method for analyzing experiments that exhibit asymmetry in the effect sizes which alters Storey (2002)’s $q$-value method and produces a more reliable list of genes declared to be DE.

In each of Chapters 2, 3, and 4, the advantages of the proposed method over traditional methods are demonstrated through two simulation studies, one involving independent normal data and one involving microarray data. Also in each of these chapters, a real microarray data set involving the gene expressions in maize leaves (Covshoff et al., 2008) is analyzed. The paper concludes with Chapter 5, which provides some discussion and possible future work.

Bibliography


CHAPTER 2. ESTIMATING THE NUMBER OF GENES THAT ARE DIFFERENTIALLY EXPRESSED IN BOTH OF TWO INDEPENDENT EXPERIMENTS

A paper submitted to the Journal of Agricultural, Biological, and Environmental Statistics

Megan Orr, Peng Liu, and Dan Nettleton

Abstract

A common procedure for estimating the number of genes that are differentially expressed (DE) in two experiments involves two steps. In the first step, data from the two experiments are separately analyzed to produce a list of genes declared to be DE in each experiment. Usually, each list is produced using a method that attempts to control the false discovery rate (FDR) in each experiment at some desired level $\alpha$. In the second step, the number of genes common to both lists is used as an estimate of the number of genes DE in both experiments. A problem with this approach is that the resulting estimates can vary greatly with $\alpha$, and the value of $\alpha$ that produces the best estimate for any given pair of experiments is difficult to predict. We propose a method that uses the $p$-values from both experiments simultaneously to produce one estimate – which does not depend on FDR level $\alpha$ – for the number of genes that are DE in both experiments. We use two simulation studies (one involving independent, normally distributed data and one involving microarray data) to compare the performances of our proposed method, the commonly used method, and another method proposed in literature to test for consistency of replicate experiments. The results of the simulation studies demonstrate the advantages of our approach. We conclude the article by estimating the number of genes that are DE in both
of two experiments involving gene expressions in maize leaves.

**Key Words:** False discovery rate; Microarray data analysis; Multiple testing; $\lambda$-estimator.

### 2.1 Introduction

Comparing the results of two independent experiments is of common interest to many researchers. This becomes a difficult problem when large data sets are analyzed and hundreds or thousands of hypothesis tests are performed for each experiment. This problem is often encountered in the analysis of gene expression experiments where mRNA expression levels are compared between two or more treatment groups for each of thousands of genes. In many cases, one of the primary interests is to estimate how many genes exhibit differential expression (i.e., a difference in mean expression levels) in both experiments. For example, Covshoff et al. (2008) performed experiments for each of two cell types in maize leaves: bundle sheath (BS) and mesophyll (M). In each experiment, expression levels were measured in wild-type and mutant cells for a set of genes. The mutant cells lacked the PSII activity of the wild-type cells, and researchers were interested in observing the effects of this lack of activity on gene expressions. To understand whether PSII activity plays a similar role in both the BS and M cell types, the researchers were specifically interested in determining whether the impact of the mutation on gene expression was similar in both BS and M cell types. Thus, the researchers attempted to estimate the number of genes differentially expressed in both BS and M cell types, which is a numerical quantity essential for understanding the extent to which the genes differentially expressed in the BS cell type overlap with the genes differentially expressed in the M cell type.

Many methods have been proposed to estimate the number of equivalently expressed (EE) genes, and thus the number of differentially expressed (DE) genes, when performing a hypothesis test for each gene in one gene expression data set. The estimation problem becomes more complicated when there are two independent data sets to be compared and the number of DE genes in both experiments, a quantity we call $m_{11}$, is to be estimated.

In practice, $m_{11}$ is typically estimated by creating a list of genes that are declared to be DE (i.e., the null hypotheses for these genes are rejected) separately for each experiment and then counting the number of genes that appear on both lists. We will call this method of $m_{11}$
estimation the “intersection method”. The intersection method is commonly used throughout the scientific literature to compare results from multiple gene expression experiments. We focus on the paper of Covshoff et al. (2008) as one representative example out of many similar examples. Other examples include Ianculescu et al. (2012), Voineagu et al. (2011), Buchanan-Wollaston et al. (2005), Wang et al. (2004), and Akopyants et al. (2004). Covshoff et al. (2008) used the intersection method to estimate $m_{11}$ at two different levels of false discovery rate (FDR) control ($\alpha = 0.01$ and $\alpha = 0.05$). We performed the intersection method at an additional level of FDR control ($\alpha = 0.10$), as well as in conjunction with an $\alpha$ selection algorithm, and found that the four $m_{11}$ estimates ranged from 168 to 2107.

As observed in Covshoff et al. (2008), an obvious flaw of the intersection method is that the estimate of $m_{11}$ highly depends on what level $\alpha$ is chosen to control FDR in each experiment. The estimate of $m_{11}$ using the intersection method is a non-decreasing function of $\alpha$, and there is no way to know which value of $\alpha$ will produce the most accurate estimate. A low value for $\alpha$ can lead to underestimation of $m_{11}$, especially when the effect sizes are relatively small for many of the truly differentially expressed genes. This is due to the high number of Type II errors that can occur when controlling FDR. On the other hand, a large value for $\alpha$ may result in many Type I errors in each experiment, which can lead to overestimation of $m_{11}$ by the intersection method.

The purpose of this paper is to introduce an improved method for estimating $m_{11}$ which does not depend on FDR control. We first analyze the $p$-values from each data set separately to estimate $m_0^{(1)}$ and $m_0^{(2)}$, the number of EE genes in the first experiment and second experiment, respectively, using the methods described in Liang and Nettleton (2012). Then we pair the $p$-values from both experiments by gene to estimate $m_{00}$, the number of genes that are EE in both experiments. We propose a bivariate extension of the $\lambda$-estimator (Storey, 2002) in order to estimate $m_{00}$. Finally, from these three estimates we obtain an estimate for $m_{11}$.

The rest of the paper is organized as follows. In Section 2.2, we describe the proposed method for estimating $m_{11}$. In Section 2.3, two simulation studies are described and the results of these studies are presented. The results of the proposed method are compared to the intersection method when controlling FDR at 5%, 10%, and levels chosen based on the
data, respectively. In addition, we also use a method proposed by Lai et al. (2007) – originally proposed with a different goal in mind – to estimate $m_{11}$ and compare its results to those of the other methods. The results of the simulation studies show that the proposed method, when compared to both the intersection method and Lai’s method, results in lower root mean squared error (RMSE) when estimating $m_{11}$ for most simulation settings. In Section 2.4, the proposed method is used to analyze the data from the experiments described in Covshoff et al. (2008), and these results are compared to those of the intersection method and Lai’s method. Finally, we provide some discussion in Section 2.5.

Please note that R code for the estimation of $m_{11}$ is available by request.

2.2 Methods

This section describes the proposed method for estimating $m_{11}$, the number of genes that are DE in both of two gene expression experiments. In Section 2.2.1, we illustrate how $m_{11}$ can be estimated as a linear combination of $m$ and the null counts $m_{01}^{(1)}$, $m_{02}^{(2)}$, and $m_{00}$. In Section 2.2.2, we review a method for estimating the number of EE genes in a single experiment, which we subsequently use to estimate $m_{01}^{(1)}$ and $m_{02}^{(2)}$. Section 2.2.3 describes our proposed method for estimating $m_{00}$. Finally, Section 2.2.4 discusses properties of $\hat{m}_{11}$, the estimator of $m_{11}$.

2.2.1 Overview of $m_{11}$ Estimation

Consider the problem of testing $m$ pairs of null hypotheses $(H_{11}, H_{21}), (H_{12}, H_{22}), \ldots, (H_{1m}, H_{2m})$, where $H_{ij}$ is the null hypothesis for experiment $i$ and gene $j$ ($i = 1, 2; j = 1, \ldots, m$). Each hypothesis $H_{ij}$ is either true, meaning that gene $j$ in experiment $i$ is EE, or false, meaning that this gene is DE. Table 2.1 is a contingency table which cross classifies the expression status (EE or DE) for each of the $m$ genes by experiment and presents their frequencies.

The counts with two digits in the subscript represent interesting frequencies when looking at the experiments simultaneously. For example, $m_{00}$ is the number of genes that are EE in both Experiment 1 and Experiment 2, while $m_{01}$ is the number of genes that are EE in Experiment 1 but DE in Experiment 2. The marginal totals in Table 2.1 represent the number
Table 2.1: Contingency table of frequencies based on cross classification of the expression status (EE or DE) for each of the \( m \) genes by experiment.

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Gene EE</th>
<th>Gene DE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene EE</td>
<td>( m_{00} )</td>
<td>( m_{01} )</td>
<td>( m_{0}^{(1)} )</td>
</tr>
<tr>
<td>Gene DE</td>
<td>( m_{10} )</td>
<td>( m_{11} )</td>
<td>( m_{1}^{(1)} )</td>
</tr>
<tr>
<td>Total</td>
<td>( m_{0}^{(2)} )</td>
<td>( m_{1}^{(2)} )</td>
<td>( m )</td>
</tr>
</tbody>
</table>

of EE genes, \( m_{0}^{(i)} \), and the number of DE genes, \( m_{1}^{(i)} \), for experiment \( i \) \((i = 1, 2)\). We are ultimately interested in estimating \( m_{11} \), the number of genes that are DE in both experiments. From Table 2.1, it is easy to see that

\[
m_{11} = m - m_{0}^{(1)} - m_{0}^{(2)} + m_{00}.
\]

The process of estimating \( m_{11} \) begins by estimating the marginal counts \( m_{0}^{(i)} \) for each experiment. We will call these estimates \( \hat{m}_{0}^{(i)} \) for \( i = 1, 2 \). Next, estimation of \( m_{00} \), resulting in \( \hat{m}_{00} \), is performed. Finally, we estimate the number of genes that are DE in both experiments by replacing the unknown counts in (2.1) by their estimates to get

\[
\hat{m}_{11} = m - \hat{m}_{0}^{(1)} - \hat{m}_{0}^{(2)} + \hat{m}_{00}.
\]

2.2.2 Review of the \( \lambda \)-Estimator and Histogram-Based Method

Now consider the problem of simultaneously testing null hypotheses \( H_{i1}, \ldots, H_{im} \) for experiment \( i \) based on corresponding \( p \)-values \( p_{i1}, \ldots, p_{im} \). For \( j = 1, \ldots, m \), we assume that \( p_{ij} \sim \text{Uniform}(0,1) \) when \( H_{ij} \) is true and that \( p_{ij} \) has a distribution stochastically smaller than uniform when \( H_{ij} \) is false. These are standard assumptions which imply that an unbiased size \( \alpha \) test can be obtained for each \( j \) by rejecting \( H_{ij} \) if and only if \( p_{ij} \leq \alpha \).

For any fixed \( \lambda_i \in [0, 1) \) in experiment \( i \), Storey (2002) proposed

\[
\hat{m}_{0}^{(i)}(\lambda_i) = \frac{\sum_{j=1}^{m} 1\{p_{ij} > \lambda_i\}}{1 - \lambda_i}
\]

as an estimate for \( m_{0}^{(i)} \), the number of true null hypotheses among \( H_{i1}, \ldots, H_{im} \).
It follows from our uniformity assumption that
\[
E(\hat{m}_0^{(i)}(\lambda_i)) = \frac{1}{1 - \lambda_i} \sum_{j=1}^{m} E(1\{p_{ij} > \lambda_i\})
\]
\[
= \frac{1}{1 - \lambda_i} \sum_{j=1}^{m} \Pr(p_{ij} > \lambda_i)
\]
\[
= \frac{1}{1 - \lambda_i} \left( \sum_{\{j: H_{ij} \text{true}\}} \Pr(p_{ij} > \lambda_i) + \sum_{\{j: H_{ij} \text{false}\}} \Pr(p_{ij} > \lambda_i) \right)
\]
\[
= \frac{1}{1 - \lambda_i} \left( \sum_{\{j: H_{ij} \text{true}\}} (1 - \lambda_i) \right) + \frac{1}{1 - \lambda_i} \left( \sum_{\{j: H_{ij} \text{false}\}} \Pr(p_{ij} > \lambda_i) \right)
\]
\[
= m_0^{(i)} + \frac{1}{1 - \lambda_i} \sum_{\{j: H_{ij} \text{false}\}} \Pr(p_{ij} > \lambda_i).
\]

Clearly, \(\hat{m}_0^{(i)}(\lambda_i)\) is a conservatively biased estimator of \(m_0^{(i)}\) for all \(\lambda_i \in [0, 1)\), where the bias is the second term in equation (2.4). The degree of bias depends on the probabilities that \(p\)-values from tests with false null hypotheses are larger than \(\lambda_i\). This directly relates to the power of the test for DE genes. The more powerful the tests for DE genes, the smaller the \(\Pr(p_{ij} > \lambda_i)\) for DE genes and the smaller the bias.

The value of \(\lambda_i\) plays an important role in the estimation of \(m_0^{(i)}\) as well as \(m_{00}\), and ultimately \(m_{11}\), which is described in Section 2.2.3. Storey (2002) investigated how the value of \(\lambda_i\) affects the bias and variance of \(\hat{m}_0^{(i)}(\lambda_i)\). He concluded that as \(\lambda_i\) increases, the bias of \(\hat{m}_0^{(i)}(\lambda_i)\) tends to decrease while the variance of \(\hat{m}_0^{(i)}(\lambda_i)\) tends to increase. Thus, it is important to determine a \(\lambda_i\) with an appropriate trade-off between bias and variance.

There are many methods proposed for determining an appropriate value of \(\lambda_i\) (see Storey (2002); Storey and Tibshirani (2003); Mosig et al. (2001); Nettleton et al. (2006); or Liang and Nettleton (2012), for example), and it is important to note that the method we propose in Section 2.2.3 for estimating \(m_{11}\) can use any of these methods. However, to illustrate our method for \(m_{11}\) estimation and to evaluate its performance relative to the intersection method and Lai’s method, we use the \(\lambda_i\) selection strategy recently proposed by Liang and Nettleton (2012). This is a “histogram-based” method that is closely related to a procedure originally proposed by Mosig et al. (2001) and studied in detail by Nettleton et al. (2006). Liang and Nettleton (2012)’s version of this procedure performs well relative to competing approaches.
in simulation studies and has desirable theoretical properties, as demonstrated by Liang and Nettleton (2012).

The idea behind the histogram-based method is to select a value of $\lambda_i$ from a set of candidates so that a histogram of $p$-values less than $\lambda_i$ is approximately decreasing while a histogram of $p$-values greater than or equal to $\lambda_i$ is approximately uniform. It makes sense to select such a $\lambda_i$ because choosing a smaller value would lead to higher bias while choosing a larger value would lead to higher variance without an appreciable reduction in bias. A limitation of the histogram-based method is that targets $m$ times the height of the $p$-value density at 1 as its estimand. As discussed by Genovese and Wasserman (2004) and Langaas et al. (2005), this estimand is an upper bound on $m_0^{(i)}$ that is chosen because $m_0^{(i)}$ is not identifiable without additional parametric assumptions. Because the histogram-based estimator targets an identifiable upper bound on $m_0^{(i)}$, it – like most other competing estimators – tends to be conservatively biased as an estimator of $m_0^{(i)}$, especially when the average power is low due to small sample sizes, large measurement error, and high variation in biological replicates, all of which are common in gene expression experiments.

The algorithm of Liang and Nettleton (2012) can be formally described as follows:

1. Partition the interval $[0,1]$ into $B$ bins of equal width. Let $c_b = \left(\frac{b-1}{B}, \frac{b}{B}\right]$ for $b = 1, 2, \ldots, B$.

2. Denote the number of $p$-values in the interval $c_b$ as $n_b$ for $b = 1, 2, \ldots, B$.

3. For each $b = 1, 2, \ldots, B$, calculate
   \[
   \bar{n}_b = \frac{\sum_{k=b}^{B} n_b}{B - b + 1}. \tag{2.5}
   \]

4. Let $b^* = \min\left\{ \min\{b : n_b \leq \bar{n}_b\}, B - 1 \right\}$.

5. Select $\lambda_i = \frac{b^*}{B}$.

Throughout this paper, we set $B = 20$ when using the this algorithm, in accordance with the recommendations of Nettleton et al. (2006) and Liang and Nettleton (2012).
The values of $\lambda_i$ that result from applying this algorithm to the $p$-values from the experiments of Covshoff et al. (2008) are depicted in the marginal histograms of Figure 2.1. For the bundle sheath data, 890 $p$-values exceeded 0.70, the selected value of $\lambda_1$. Using (2.3), this yields $890/(1 - 0.70) = 2967$ as an estimate of $m_0^{(1)}$ for the bundle sheath data. Similarly, the estimate of $m_0^{(2)}$ for the mesophyll data is $1162/(1 - 0.55) = 2582$. 
2.2.3 Estimating the Number of Genes that are Equivalently Expressed in Both Experiments

Let \((p_{11}, p_{21}), (p_{12}, p_{22}), \ldots, (p_{1m}, p_{2m})\) represent \(m\) pairs of \(p\)-values from testing the \(m\) pairs of null hypotheses mentioned in Section 2.2.1, \((H_{11}, H_{21}), (H_{12}, H_{22}), \ldots, (H_{1m}, H_{2m})\), where \(p_{ij}\) is the \(p\)-value for testing \(H_{ij}\), the null hypothesis in experiment \(i\) for gene \(j\) \((i = 1, 2; j = 1, \ldots, m)\).

We begin by estimating the number of true null hypotheses (EE genes) for each experiment using the histogram-based method as described in Section 2.2.2. Let \(\lambda_i\) denote the value selected by the algorithm in Section 2.2.2 for experiment \(i\), and let \(\hat{m}_0^{(i)}(\lambda_i)\) denote the estimated number of true null hypotheses in experiment \(i\) \((i = 1, 2)\).

Next, notice that if a pair of tests corresponding to gene \(j\) both have true null hypotheses, then the pair \((p_{1j}, p_{2j})\) is assumed to follow a product uniform distribution given by

\[
\Pr\left((p_{1j}, p_{2j}) \in [a, b] \times [c, d]\right) = (b - a)(d - c),
\]

for all \(a < b, c < d,\) and \(a, b, c, d \in [0, 1]\). This follows from the assumption that \(p\)-values from tests with true null hypotheses are uniform and from the assumption that Experiments 1 and 2 are independent.

The next step is to estimate \(m_{00}\), the number of genes with true null hypotheses in both experiments (i.e., the number of genes that are EE in both experiments). Define

\[
n_{00} = \sum_{i=1}^{m} \mathbf{1}\left((p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1]\right).
\]

From (2.6), we see that the probability that a \(p\)-value pair falls in \([\lambda_1, 1] \times [\lambda_2, 1]\) is \((1 - \lambda_1)(1 - \lambda_2)\) if its corresponding gene is EE in both experiments. Thus, a conservative estimate of \(m_{00}\) is

\[
\hat{m}_{00} = \frac{n_{00}}{(1 - \lambda_1)(1 - \lambda_2)}.
\]

This is a bivariate analog of Storey’s \(\lambda\)-estimator.

2.2.4 Properties of \(\hat{m}_{11}\)

Once the estimates \(\hat{m}_0^{(1)}, \hat{m}_0^{(2)}\), and \(\hat{m}_{00}\) have been obtained, we use (2.2) to estimate \(m_{11}\).
Performing steps analogous to those used in (2.4), we derive the expected value of \( \hat{m}_{11} \), for fixed \( \lambda_1, \lambda_2 \in [0, 1) \), as

\[
E(\hat{m}_{11}) = m_{11} - \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} \sum_{F} A_j,
\]

(2.9)

where \( F = \{ j : H_{1j} \text{ false and } H_{2j} \text{ false} \} \) and

\[
A_j = (1 - \lambda_1) \Pr(p_{2j} > \lambda_2) + (1 - \lambda_2) \Pr(p_{1j} > \lambda_1)

- \Pr(p_{1j} > \lambda_1) \Pr(p_{2j} > \lambda_2).
\]

(2.10)

For the complete derivation of \( m_{11} \), see Appendix 2.6.

Because the distribution of a \( p \)-value from a DE gene is stochastically smaller than uniform, \( \Pr(p_{ij} > \lambda_i) < (1 - \lambda_i) \) for \( i = 1, 2 \) when gene \( j \) is DE in both experiments, making \( A_j \) positive for genes that are DE in both experiments. Furthermore, as the power of each test increases for all genes that are DE in both experiments, the values of \( A_j \) decrease, decreasing the bias of \( \hat{m}_{11} \). As the power of each test decreases, both \( \Pr(p_{1i} > \lambda_1) \) and \( \Pr(p_{2i} > \lambda_2) \) increase toward \( (1 - \lambda_1) \) and \( (1 - \lambda_2) \), respectively, decreasing \( E(\hat{m}_{11}) \) toward zero. Thus, \( \hat{m}_{11} \) is biased downward, and \( E(\hat{m}_{11}) \) is bounded between zero and \( m_{11} \).

Ideally, we would also like to assess the uncertainty of the estimator \( \hat{m}_{11} \) by providing a standard error or a confidence interval for \( m_{11} \). Unfortunately, it is extremely difficult to evaluate the uncertainty through either analytical or numerical methods due to both the unknown complex correlation structure of the data as well as the potential for low average power of tests corresponding to DE genes. Our method for estimating \( m_{11} \) is a more complicated extension of the problem of estimating \( m_0 \). Many methods have been proposed for estimating \( m_0 \) including Storey (2002), Storey and Tibshirani (2003), Storey et al. (2004), Langaas et al. (2005), Nettleton et al. (2006), and Liang and Nettleton (2012). Langaas et al. (2005) is the only one of these papers that contains a formula for the variance of \( \hat{m}_0 \). Unfortunately, this expression for the variance applies only to the unrealistic case of independent \( p \)-values. To our knowledge, no one has developed a method for accurately evaluating the uncertainty associated with estimators of
Assessing the uncertainty associated with our estimator of $m_{11}$ is an even more challenging open question.

### 2.3 Simulation Studies

In order to evaluate the performance of our proposed method for estimating $m_{11}$, simulation studies were performed. For each simulated data set, two null hypotheses

$$H_{1j} : \mu_{1j1} = \mu_{1j2} \quad \text{and} \quad H_{2j} : \mu_{2j1} = \mu_{2j2}$$

were tested (each against the two-sided alternative) for each gene $j$ of $m = 10000$ genes, where $\mu_{ijk}$ represents the population mean expression for treatment $k$ of gene $j$ in experiment $i$ ($i = 1, 2; j = 1, \ldots, m; k = 1, 2$). These tests were performed by calculating a test statistic and its corresponding $p$-value, $p_{ij}$, for each gene $j$ in experiment $i$ using the moderated $t$-test approach proposed by Smyth (2004). This method of testing for DE genes was specifically designed for microarray experiments and involves borrowing information across all genes in order to better estimate the error variance of each gene and to obtain a $t$-distributed test statistic that performs better than the regular $t$-test with respect to ranking genes for differential expression. These $p$-values calculated using the moderated $t$-test were used to estimate $m_{11}$ using Lai’s method, the intersection method, and the proposed method.

The intersection method was performed by first using the histogram-based method to calculate $\hat{m}_0(i)$ for each experiment $i$. Then $\hat{m}_0(i)$ was used to convert the $p$-values to $q$-values (Storey, 2002) separately for each experiment. A list of genes declared to be DE at a specific level $\alpha$ of FDR control was created for each experiment. Finally, $m_{11}$ was estimated as the number of genes common to both lists. In addition to the proposed and intersection methods, a method described in Lai et al. (2007) was used to estimate $m_{11}$. The original purpose of Lai’s method was to determine if the results of two large replicate experiments are consistent enough for their data to be combined for a more powerful analysis. Although this purpose is not to estimate $m_{11}$ directly, the most general mixture model proposed in Lai et al. (2007) can be used to estimate $m_{11}$ by summing the estimated mixing proportions for components in the mixture model that correspond to genes that are DE in both experiments and multi-
plying the resulting proportion by $m$. Lai’s method was performed using R code available at http://home.gwu.edu/~ylai/research/Concordance.

Our proposed method was implemented as described in Section 2.2. It should be noted that the estimators $\hat{m}_0^{(1)}$, $\hat{m}_0^{(2)}$, $\hat{m}_{00}$, and $\hat{m}_{11}$ are all random variables that can result in values outside the range of their estimands. We estimated $m_{11}$ under many possible constraint combinations. For example, in one combination, we estimated $m_{11}$ by constraining all estimates to values within the range of their estimands. Thus, any estimates of $m_0^{(1)}$, $m_0^{(2)}$, or $m_{00}$ above $m$ were replaced by $m$, and any negative estimates of $m_{11}$ were replaced by 0. After considering many different constraint methods and examining the performance of the resulting $m_{11}$ estimators, we concluded that constraining only the final estimate of $m_{11}$ (and not constraining $\hat{m}_0^{(1)}$, $\hat{m}_0^{(2)}$, or $\hat{m}_{00}$ in intermediate calculations) produced slightly better results. Thus, this approach was used to estimate $m_{11}$ in the following simulation studies.

Two simulation studies were performed. The first set of simulations used independent, normally distributed data. This allowed us to evaluate our method under ideal conditions that are consistent with the assumptions used to derive the moderated $t$-test (Smyth, 2004). The second set of simulations used real microarray data in order to evaluate how our method performed when data have a distribution and correlation structure that we can not model precisely but will encounter in practice. For each simulation setting, 100 data sets were randomly generated.

Aside from the distribution of the data, we also varied the sample size for each treatment ($n$), the magnitude of effect sizes for differentially expressed genes (controlled by a parameter $\mu_\delta$ defined in the Section 2.3.1), and the quantities defined in Table 2.1. Sample sizes of $n = 4$, 10, and 20 were chosen for the simulation studies because typical gene expression experiments have small sample size, usually due to the high cost of experimentation. For example, approximately 90% of all data sets available on the Gene Expression Omnibus (Edgar et al., 2002) have a total sample size of $N = 40$ or less, with 40% of available data sets having at most $N = 8$. These sample sizes correspond to $n = 20$ and $n = 4$, respectively, in our simulation studies. Also, due to the high variation in both measurement error and biological replicates in most gene expression experiments, the average power for detecting DE genes, and thus the average relative effect size, is generally small. Because of this, we chose small mean relative effect sizes of $\mu_\delta = 1$
and 2.

We compare the results of the proposed method to those of Lai’s and the intersection methods when controlling FDR at various levels. Three different estimates are obtained using the intersection method. The first two estimates control FDR at an \( \alpha \)-level determined \textit{a priori} for each experiment. Motivated by a reviewer’s comment, the final estimate for \( m_{11} \) is obtained by choosing a separate \( \alpha \) \textit{a posteriori} for each experiment. For this intersection method estimator, \( \alpha \) is chosen so that the number of genes declared to be DE is equal to the estimated number of DE genes for the given experiment.

All methods used to estimate \( m_{11} \) are evaluated both visually and using the root mean squared error (RMSE).

2.3.1 Simulations Using Independent, Normally Distributed Data

In the first simulation study, data consisting of \( m = 10000 \) genes were simulated from independent normal distributions with gene-specific variances. The gene-specific variances \( \{\sigma_{ij}^2 : i = 1, 2; j = 1, \ldots, m\} \) were drawn independently from an inverse gamma distribution. The parameters of this distribution were estimated from a microarray data set using the methods of Smyth (2004). This data set consists of gene expressions from patients suffering from different types cardiomyopathy and can be obtained from the Gene Expression Omnibus (GEO, \url{http://www.ncbi.nlm.nih.gov/geo/}) under accession number GSE5406. For a description of the experiment performed to obtain these data, see Hannenhalli et al. (2006). Conditional on \( \{\sigma_{ij}^2 : i = 1, 2; j = 1, \ldots, m\} \), treatment mean values \( \{\mu_{ijk} : i = 1, 2; j = 1, \ldots, m; k = 1, 2\} \) were determined as follows. For an EE gene, \( \mu_{ij1} = \mu_{ij2} = 0 \). For a DE gene, \( \mu_{ij1} \) was set to zero, and \( \mu_{ij2} \) was drawn from a \( \text{N}(\mu_\delta \sigma_{ij}, \sigma_{ij}^2) \) distribution. Given the value for \( \mu_{ijk} \) and \( \sigma_{ij}^2 \), \( n \) observation for experiment \( i \), gene \( j \), and treatment \( k \) were independently drawn from the \( \text{N}(\mu_{ijk}, \sigma_{ij}^2) \) distribution.

2.3.2 Simulations Using Real Microarray Data

For the second simulation study, a microarray data set was used that consists of gene expressions from the bone marrow or peripheral blood of subjects with cytogenetically normal
acute myeloid leukemia (CN-AML) and is described in Metzeler et al. (2008). This data set is also available on GEO under the accession number GSE12417. Only data from subjects in the training cohort were used. In this experiment, the tissues were prepared and then hybridized with individual Affymetrix HU133A arrays and the raw expression values were then transformed and normalized. There are 22284 genes on the Affymetrix HU133A array, but \( m = 10000 \) genes were randomly selected to be included in the simulations. The \( N = 163 \) total subjects were randomly split into two subsets representing two independent experiments with total sample sizes \( N_1 = 82 \) and \( N_2 = 81 \).

In order to simulate a data set for the first experiment using the microarray data, we started with the original data from the 10000 randomly selected genes and the \( N_1 = 82 \) randomly selected subjects. For the \( j^{th} \) gene, the standard deviation, \( s_{1,j} \), was calculated from the expression values of all \( N_1 = 82 \) subjects. Then, \( 2n \) subjects were randomly chosen as the subjects for the experiment. This group was further split, randomly, into two groups of size \( n \), each representing a different treatment group. Differentially expressed genes were then created by adding a randomly generated treatment effect to the data from the second treatment group. Treatment effects, \( \mu_{1j2} \), were generated as in Section 2.3.1, except the calculated \( s_{1j} \) values were used in place of the simulated \( \sigma_{1j} \) values. Data were simulated for the second experiment in a similar manner.

### 2.3.3 Results

Tables 2.2 and 2.3 give the mean estimates of \( m_{11} \) along with their RMSEs for each simulation setting and estimation method. Also, for settings with \( m_{11} = 0 \), the number of times \( m_{11} \) was correctly estimated as 0 is given in square brackets under the mean and RMSE. In each table for each setting, the estimate with the smallest RMSE is presented in bold font.

Although the values of \( m_{01} \) and \( m_{10} \) as shown in Table 2.1 are not provided, \( m_{01} = m_{10} \) for each simulation setting, and their common value can be obtained by taking half the value of \( m - m_{00} - m_{11} \). Figures 2.2 and 2.3 illustrate the estimation results with boxplots for simulation settings with \( n = 4 \) and \( n = 10 \). Boxplots for settings with \( n = 20 \) look similar to those with \( n = 10 \) presented in Figures 2.2 and 2.3. For both sets of simulations, the proposed method
outperforms the other methods for most simulation settings.

When \( m_{11} = 0 \), our proposed method either estimates \( m_{11} \) correctly with \( \hat{m}_{11} = 0 \), or overestimates it, due to the lower bound of zero placed on the estimate. As a consequence, the bounded estimator is positively biased. In this case, our method is outperformed in the simulations using both normal data and microarray data (see Tables 2.2 and 2.3) when we evaluate the estimation methods using RMSE. In general, the intersection method, when FDR is controlled at 5%, outperforms the other estimation methods for settings with \( m_{11} = 0 \). This is not surprising as this method uses the most stringent cutoff for declaring genes to be DE. For each setting with \( m_{11} = 0 \), we also evaluated the methods by counting the number of data sets (out of 100) whose analysis resulted in \( \hat{m}_{11} = 0 \). In half of these 12 simulation settings, the proposed method correctly estimates \( m_{11} \) for more data sets than any other method, including the intersection method with the lowest level of FDR control. This suggests that our method remains effective when \( m_{11} = 0 \).

In practice, researchers are more interested in estimating \( m_{11} \) when there are some genes DE in both experiments under study, i.e., when \( m_{11} > 0 \). For simulations using independent, normally distributed data, in the 24 settings with \( m_{11} > 0 \) (see Table 2.2), the proposed method outperforms all other methods in 16 of the 24 simulation settings, including 13 of the 16 settings with \( n = 4 \) or \( n = 10 \). Lai’s method estimates \( m_{11} \) best in five of the 24 settings, mostly for higher values of \( m_{11} \). The intersection method (when FDR is controlled at either 5% or 10%) performs best in 3 of the 24 settings. See Figure 2.2 for a visual representation of the \( m_{11} \) estimation results.

For simulations using microarray data (see Table 2.3 and Figure 2.3), the proposed method outperforms the other methods in 13 of the 24 simulation settings with \( m_{11} > 0 \), and generally performs best when the average power of the test is low. When \( n = 4 \), the proposed method is best in all 8 settings. Lai’s method has the lowest RMSE in five of the 24 settings, four of which when \( n = 20 \). When FDR is controlled at 10%, the intersection method performs best in the remaining six settings, all with \( n = 10 \) or \( n = 20 \). The intersection method never performs best when \( \alpha \) is chosen \( a \ posteriori \) for each experiment, although it tends to outperform the intersection method when FDR is controlled at 5% and 10% for simulation settings with lower
Table 2.2: Mean estimates of $m_{11}$ and the RMSE (rounded to the nearest integer) for each simulation setting and estimation method when data are independent and normally distributed. Columns with “FDR” in their name refer to the intersection method with level of control given after the “≤” sign. Here, $\hat{\alpha}_i$ indicates the $\alpha$ level was estimated separately for each experiment using the method described in Section 2.3. Within each row, the results for the method with the lowest RMSE are presented in bold font. For simulation settings with $m_{11} = 0$, the number of times $m_{11}$ was correctly estimated as 0 is given in square brackets below the mean and RMSE, and the results with the highest number is given in bold font.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu_3$</th>
<th>$m_{00}$</th>
<th>$m_{11}$</th>
<th>Proposed</th>
<th>Lai</th>
<th>$\text{FDR} \leq 0.05$</th>
<th>$\text{FDR} \leq 0.10$</th>
<th>$\text{FDR} \leq \hat{\alpha}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>9000</td>
<td>0</td>
<td>17 (29)</td>
<td>11 (12)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>5 (6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[45/100]</td>
<td>[0/100]</td>
<td>[100/100]</td>
<td>[100/100]</td>
<td>[2/100]</td>
</tr>
<tr>
<td>9000</td>
<td>500</td>
<td>116 (387)</td>
<td>50 (451)</td>
<td>0 (500)</td>
<td>0 (500)</td>
<td>17 (1983)</td>
<td>67 (1933)</td>
<td>469 (1532)</td>
</tr>
<tr>
<td>7000</td>
<td>1000</td>
<td>291 (715)</td>
<td>208 (793)</td>
<td>2 (998)</td>
<td>10 (990)</td>
<td>172 (828)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2000</td>
<td>606 (1400)</td>
<td>563 (1437)</td>
<td>17 (1983)</td>
<td>67 (1933)</td>
<td>469 (1532)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>943 (2065)</td>
<td>64 (2936)</td>
<td>201 (2799)</td>
<td>861 (2140)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9000</td>
<td>500</td>
<td>303 (201)</td>
<td>226 (274)</td>
<td>16 (484)</td>
<td>42 (458)</td>
<td>147 (354)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>1000</td>
<td>646 (362)</td>
<td>548 (543)</td>
<td>121 (789)</td>
<td>231 (769)</td>
<td>464 (537)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2000</td>
<td>1349 (660)</td>
<td>1208 (793)</td>
<td>446 (1554)</td>
<td>747 (1253)</td>
<td>1122 (880)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>2066 (944)</td>
<td>1950 (1051)</td>
<td>981 (2019)</td>
<td>1538 (1463)</td>
<td>1918 (1085)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>9000</td>
<td>0</td>
<td>10 (21)</td>
<td>7 (8)</td>
<td>0 (1)</td>
<td>1 (1)</td>
<td>6 (7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[48/100]</td>
<td>[0/100]</td>
<td>[97/100]</td>
<td>[70/100]</td>
<td>[0/100]</td>
</tr>
<tr>
<td>9000</td>
<td>500</td>
<td>303 (201)</td>
<td>226 (274)</td>
<td>16 (484)</td>
<td>42 (458)</td>
<td>147 (354)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>1000</td>
<td>646 (362)</td>
<td>548 (543)</td>
<td>121 (789)</td>
<td>231 (769)</td>
<td>464 (537)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2000</td>
<td>1349 (660)</td>
<td>1208 (793)</td>
<td>446 (1554)</td>
<td>747 (1253)</td>
<td>1122 (880)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>2066 (944)</td>
<td>1950 (1051)</td>
<td>981 (2019)</td>
<td>1538 (1463)</td>
<td>1918 (1085)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9000</td>
<td>0</td>
<td>9 (17)</td>
<td>2 (2)</td>
<td>1 (2)</td>
<td>3 (3)</td>
<td>7 (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[50/100]</td>
<td>[0/100]</td>
<td>[34/100]</td>
<td>[6/100]</td>
<td>[2/100]</td>
</tr>
<tr>
<td>9000</td>
<td>500</td>
<td>303 (201)</td>
<td>226 (274)</td>
<td>16 (484)</td>
<td>42 (458)</td>
<td>147 (354)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>1000</td>
<td>646 (362)</td>
<td>548 (543)</td>
<td>121 (789)</td>
<td>231 (769)</td>
<td>464 (537)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2000</td>
<td>1349 (660)</td>
<td>1208 (793)</td>
<td>446 (1554)</td>
<td>747 (1253)</td>
<td>1122 (880)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>2066 (944)</td>
<td>1950 (1051)</td>
<td>981 (2019)</td>
<td>1538 (1463)</td>
<td>1918 (1085)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>9000</td>
<td>0</td>
<td>9 (19)</td>
<td>5 (6)</td>
<td>0 (1)</td>
<td>2 (2)</td>
<td>6 (7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[55/100]</td>
<td>[0/100]</td>
<td>[49/100]</td>
<td>[15/100]</td>
<td>[1/100]</td>
</tr>
<tr>
<td>9000</td>
<td>500</td>
<td>290 (214)</td>
<td>271 (229)</td>
<td>135 (365)</td>
<td>159 (341)</td>
<td>190 (311)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>1000</td>
<td>601 (404)</td>
<td>600 (401)</td>
<td>344 (656)</td>
<td>413 (587)</td>
<td>481 (520)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2000</td>
<td>1223 (782)</td>
<td>1278 (723)</td>
<td>805 (1196)</td>
<td>983 (1017)</td>
<td>1075 (926)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>1864 (1142)</td>
<td>1736 (1025)</td>
<td>1519 (685)</td>
<td>1533 (468)</td>
<td>1485 (517)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9000</td>
<td>0</td>
<td>10 (19)</td>
<td>0 (1)</td>
<td>2 (2)</td>
<td>4 (5)</td>
<td>4 (5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[55/100]</td>
<td>[0/100]</td>
<td>[14/100]</td>
<td>[1/100]</td>
<td>[7/100]</td>
</tr>
<tr>
<td>9000</td>
<td>500</td>
<td>430 (74)</td>
<td>411 (90)</td>
<td>353 (148)</td>
<td>373 (128)</td>
<td>363 (137)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>1000</td>
<td>871 (136)</td>
<td>818 (182)</td>
<td>776 (225)</td>
<td>839 (162)</td>
<td>792 (209)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>2000</td>
<td>1755 (250)</td>
<td>1660 (341)</td>
<td>1654 (347)</td>
<td>1815 (187)</td>
<td>1655 (346)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>2645 (361)</td>
<td>2511 (489)</td>
<td>2635 (366)</td>
<td>2961 (50)</td>
<td>2580 (423)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3: Mean estimates of $m_{11}$ and the RMSE (rounded to the nearest integer) for each simulation setting and estimation method for simulations using microarray data. Columns with “FDR” in their name refer to the intersection method with level of control given after the “≤” sign. Here, $\hat{\alpha}_i$ indicates the $\alpha$ level was estimated separately for each experiment using the method described in Section 2.3. Within each row, the results for the method with the lowest RMSE are presented in bold font. For simulation settings with $m_{11} = 0$, the number of times $m_{11}$ was correctly estimated as 0 is given in square brackets below the mean and RMSE, and the results with the highest number is given in bold font.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu_3$</th>
<th>$m_{00}$</th>
<th>$m_{11}$</th>
<th>Proposed</th>
<th>Lai</th>
<th>FDR ≤ 0.05</th>
<th>FDR ≤ 0.10</th>
<th>FDR ≤ $\hat{\alpha}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>9000</td>
<td>0</td>
<td>68 (220)</td>
<td>1144 (2515)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>30 (106)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[39/100]</td>
<td>0/100</td>
<td>[100/100]</td>
<td>[98/100]</td>
<td>[83/100]</td>
</tr>
<tr>
<td></td>
<td>9000</td>
<td>500</td>
<td>181 (432)</td>
<td>1381 (2816)</td>
<td>0 (500)</td>
<td>1 (493)</td>
<td>75 (462)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>1000</td>
<td>381 (726)</td>
<td>957 (2066)</td>
<td>2 (998)</td>
<td>10 (990)</td>
<td>201 (840)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>2000</td>
<td>708 (1350)</td>
<td>1173 (1727)</td>
<td>16 (1984)</td>
<td>63 (1938)</td>
<td>518 (1522)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9000</td>
<td>0</td>
<td>155 (366)</td>
<td>657 (1561)</td>
<td>0 (3)</td>
<td>1 (7)</td>
<td>85 (229)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[49/100]</td>
<td>0/100</td>
<td>[94/100]</td>
<td>[77/100]</td>
<td>[59/100]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9000</td>
<td>500</td>
<td>371 (343)</td>
<td>670 (1490)</td>
<td>14 (486)</td>
<td>39 (462)</td>
<td>145 (425)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>1000</td>
<td>723 (501)</td>
<td>785 (1015)</td>
<td>109 (892)</td>
<td>224 (781)</td>
<td>514 (589)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>2000</td>
<td>1506 (688)</td>
<td>1500 (944)</td>
<td>412 (1595)</td>
<td>737 (1282)</td>
<td>1239 (900)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3000</td>
<td>2258 (887)</td>
<td>2153 (966)</td>
<td>962 (2049)</td>
<td>1560 (1470)</td>
<td>2069 (1026)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>9000</td>
<td>0</td>
<td>50 (152)</td>
<td>325 (1404)</td>
<td>0 (1)</td>
<td>1 (4)</td>
<td>41 (144)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[68/100]</td>
<td>[0/80]</td>
<td>[87/100]</td>
<td>[68/100]</td>
<td>[69/100]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9000</td>
<td>500</td>
<td>228 (365)</td>
<td>412 (864)</td>
<td>43 (457)</td>
<td>63 (437)</td>
<td>112 (442)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>1000</td>
<td>510 (572)</td>
<td>788 (774)</td>
<td>146 (854)</td>
<td>215 (787)</td>
<td>429 (641)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>2000</td>
<td>1097 (985)</td>
<td>1362 (1053)</td>
<td>406 (1596)</td>
<td>593 (1415)</td>
<td>953 (1123)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3000</td>
<td>1628 (1431)</td>
<td>795 (2210)</td>
<td>1157 (1860)</td>
<td>1548 (1514)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9000</td>
<td>0</td>
<td>116 (313)</td>
<td>320 (1440)</td>
<td>2 (4)</td>
<td>7 (19)</td>
<td>84 (246)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[60/100]</td>
<td>[0/100]</td>
<td>[45/100]</td>
<td>[26/100]</td>
<td>[49/100]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9000</td>
<td>500</td>
<td>415 (264)</td>
<td>559 (1170)</td>
<td>228 (273)</td>
<td>265 (236)</td>
<td>206 (381)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>1000</td>
<td>871 (319)</td>
<td>884 (387)</td>
<td>576 (426)</td>
<td>675 (333)</td>
<td>728 (421)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>2000</td>
<td>1739 (485)</td>
<td>1672 (549)</td>
<td>1325 (683)</td>
<td>1561 (476)</td>
<td>15865 (578)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3000</td>
<td>2681 (640)</td>
<td>2529 (665)</td>
<td>2247 (779)</td>
<td>2710 (460)</td>
<td>2575 (680)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>9000</td>
<td>0</td>
<td>138 (429)</td>
<td>177 (551)</td>
<td>1 (3)</td>
<td>6 (19)</td>
<td>97 (308)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[69/100]</td>
<td>[0/100]</td>
<td>[63/100]</td>
<td>[40/100]</td>
<td>[59/100]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9000</td>
<td>500</td>
<td>277 (268)</td>
<td>309 (257)</td>
<td>134 (367)</td>
<td>158 (343)</td>
<td>110 (416)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>1000</td>
<td>632 (483)</td>
<td>740 (545)</td>
<td>345 (656)</td>
<td>424 (581)</td>
<td>529 (570)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>2000</td>
<td>1320 (773)</td>
<td>1478 (696)</td>
<td>817 (1185)</td>
<td>1010 (998)</td>
<td>1195 (879)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3000</td>
<td>1954 (1129)</td>
<td>2241 (1018)</td>
<td>1386 (1619)</td>
<td>1750 (1269)</td>
<td>1841 (1239)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9000</td>
<td>0</td>
<td>40 (146)</td>
<td>35 (111)</td>
<td>2 (4)</td>
<td>5 (11)</td>
<td>34 (110)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[70/100]</td>
<td>[0/100]</td>
<td>[41/100]</td>
<td>[16/100]</td>
<td>[59/100]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9000</td>
<td>500</td>
<td>521 (302)</td>
<td>465 (116)</td>
<td>356 (145)</td>
<td>379 (123)</td>
<td>295 (346)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7000</td>
<td>1000</td>
<td>990 (375)</td>
<td>940 (304)</td>
<td>785 (221)</td>
<td>862 (179)</td>
<td>845 (405)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>2000</td>
<td>1857 (344)</td>
<td>1762 (381)</td>
<td>1666 (346)</td>
<td>1846 (224)</td>
<td>1740 (425)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3000</td>
<td>2817 (510)</td>
<td>2631 (558)</td>
<td>2680 (382)</td>
<td>3058 (372)</td>
<td>2761 (544)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.2: Boxplots of the $m_{11}$ estimates for simulations involving independent, normally distributed data. The true value of $m_{11}$ is given above each plot and is also represented by the horizontal dashed line in each plot. Plots (2.2a)-(2.2e) and (2.2k)-(2.2o) show the results for simulations with $\mu_0 = 1$, and plots (2.2f)-(2.2j) and (2.2p)-(2.2t) show the results for simulations with $\mu_0 = 2$. The first and second boxplots in each plot represent the estimates for the proposed and Lai’s methods, respectively. The third and fourth boxplots represent the estimates for the intersection method when FDR was controlled at 0.05 and 0.10, respectively. The fifth boxplot represents the estimates for the intersection method when $\alpha$ was chosen \textit{a posteriori}. 

$n = 4$

(a) \hspace{1cm} (b) \hspace{1cm} (c) \hspace{1cm} (d) \hspace{1cm} (e) 

(f) \hspace{1cm} (g) \hspace{1cm} (h) \hspace{1cm} (i) \hspace{1cm} (j) 

$n = 10$

(k) \hspace{1cm} (l) \hspace{1cm} (m) \hspace{1cm} (n) \hspace{1cm} (o) 

(p) \hspace{1cm} (q) \hspace{1cm} (r) \hspace{1cm} (s) \hspace{1cm} (t)
$n = 4$

![Boxplots](image)

Figure 2.3: Boxplots of the $m_{11}$ estimates for simulations involving microarray data. The true value of $m_{11}$ is given above each plot and is also represented by the horizontal dashed line in each plot. Plots (2.3a)-(2.3e) and (2.3k)-(2.3o) show the results for simulations with $\mu_\delta = 1$, and plots (2.3f)-(2.3i) and (2.3p)-(2.3t) show the results for simulations with $\mu_\delta = 2$. The first and second boxplots in each plot represent the estimates for the proposed and Lai’s methods, respectively. The third and fourth boxplots represent the estimates for the intersection method when FDR was controlled at 0.05 and 0.10, respectively. The fifth boxplot represents the estimates for the intersection method when $\alpha$ was chosen a posteriori.
power. Also, in many settings, even though the intersection method outperforms the proposed method at one level of FDR control, it does not perform as well as the intersection method for the other levels of FDR control.

Unsurprisingly, $m_{11}$ is underestimated when using our method in settings with lower power (except when $m_{11}=0$ as explained earlier). This underestimation is unavoidable, as illustrated in equations (2.9) and (2.10). However, the other estimation methods usually result in higher degrees of underestimation, except in the microarray simulations when Lai’s method results in highly skewed estimates and high RMSEs, as illustrated in Figure 2.3.

### 2.4 Real Data Analysis

In this section, we analyze the data described in Covshoff et al. (2008) using the proposed method and compare the results to those of Lai’s method and the intersection method. The data come from two independent experiments in which the same two-color microarray platform was used to measure gene expressions in maize leaves. One experiment was performed on mesophyll (M) cells in the maize leaves and the other on bundle sheath (BS) cells. Each experiment had two treatments, wild type and mutant, with $n = 6$ two-color slides. Maize leaves in the mutant treatment had cells that lacked the PSII activity of the maize leaves in the wild-type treatment. As discussed in Section 2.1 researchers were specifically interested in estimating $m_{11}$, the number of genes that are DE expressed, due to lack of PSII activity, in both the BS and M cells.

Although the same platform was used for both experiments, only 7377 and 8463 genes were detected above background level for the M and BS experiments, respectively. For each experiment, the difference in mean expressions between the wild-type and mutant cells was tested against a null value of zero for each detected gene using the moderated $t$-test (Smyth, 2004). The $p$-value from each test was converted to a $q$-value, and the intersection method was performed for each of three FDR levels. When controlling FDR at 1%, 5%, and 10%, $m_{11}$ was estimated to be 168, 573, and 1012, respectively. When the $\alpha$ selection algorithm described in Section 2.3 is performed, $q$-value cutoffs of 0.188 and 0.239 were used for the M and BS experiments, respectively. This resulted in an $m_{11}$ estimate of 2107 genes that are DE in both
experiments. The most important observation from these results is that the estimates of $m_{11}$ vary drastically for different levels of FDR control, and we have no way of knowing which estimate is the most appropriate.

One complication of this real data example is that different numbers of genes were analyzed in the BS and M experiments due to different number of genes detected above background in the two experiments. Because a gene that is not detected above background level in one experiment cannot be determined to be DE in this experiment and thus cannot be determined to be DE in both experiments, we only used the 5670 genes that were detected in both experiments for the purpose of estimating $m_{11}$ using the proposed method.

Figure 2.1 shows histograms of the $p$-values for the BS and M experiments individually as well as the scatterplot of the $p$-values paired by gene. These $p$-values are available in Supplemental Tables S5 and S6 in Covshoff et al. (2008) in the columns labeled “pvalue_limma(FDR)”.

As discussed in Section 2.2.2, the dashed line in each histogram represents the value of $\lambda_i$ ($i = 1, 2$) selected using the algorithm in Section 2.2.2. For the BS experiment, $\lambda_1 = 0.70$, and there are 890 $p$-values larger than $\lambda_1$. For the M experiment, $\lambda_2 = 0.55$, and there are 1162 $p$-values larger than $\lambda_2$. Thus $\hat{m}_0^{(1)} = 2967$ and $\hat{m}_0^{(2)} = 2582$ as calculated in Section 2.2.2.

We now use the method described in Section 2.2.3 to estimate $m_{00}$. We have selected the value of $\lambda_i$ for each experiment, so we can now count the $p$-value pairs in the region $[\lambda_1, 1] \times [\lambda_2, 1]$, as illustrated by the box in the upper right corner of the scatterplot in Figure 2.1. There are 216 $p$-values that fall within this region, so we estimate $m_{00}$ as

$$\hat{m}_{00} = \frac{216}{(1 - 0.70)(1 - 0.55)} = 1600. \quad (2.12)$$

Using (2.2), we can now estimate $m_{11}$, or the number of genes that are differentially expressed between wild type and mutant cells in both the BS and M experiments, as

$$\hat{m}_{11} = m - \hat{m}_0^{(1)} - \hat{m}_0^{(2)} + \hat{m}_{00}$$

$$= 5670 - 2967 - 2582 + 1600$$

$$= 1721. \quad (2.13)$$

Thus, using our method, we estimate that 1721 genes are DE from the lack of PSII activity
in both the M and BS experiments using the proposed method. This is approximately 30% of the genes that were detected above background level in both experiments.

We also used Lai’s method to obtain an estimate of \( m_{11} \) as 5223. This is an unrealistically high value relative to the \( m = 5670 \) genes that were analyzed. As discussed in Section 2.3.3 and observed in Figure 2.3, the use of Lai’s method to estimate \( m_{11} \) can result in severe overestimation for microarray data, which may be the case for the data from the maize experiments.

### 2.5 Discussion

As illustrated by the results of the simulation studies, the effectiveness of the intersection method can depend highly on the \( \alpha \)-level chosen. Even when \( \alpha \) is chosen \( a \ posteriori \), it rarely outperforms the proposed method, although it tends to improve the performance of the intersection method when average power is low. The proposed method has a clear advantage over the intersection method because it does not depend on choosing a level of FDR control \( \alpha \), and we recommend using the proposed method over the intersection method for this reason. The proposed method also outperforms Lai’s method in most simulation settings. Furthermore, we developed the proposed method for the purpose of analyzing gene expression experiments, the large majority of which have low sample size and low average power for detecting differential expression. Thus, we are mostly interested in how our method performs in such settings, and the simulation studies suggest that our method performs best in most of these settings. When sample sizes grow large, the differences among the methods diminish. However, we still recommend using the proposed method in experiments with larger \( n \) because it performs similarly when compared to the other methods and does not require choosing \( \alpha \) as the intersection method does. Overall, our proposed method is effective, especially for experiments with low average power, but still performs adequately in experiments with larger sample sizes.

Additionally, the proposed method does not require that the two experiments use the same platform because it only requires two sets of p-values that correspond to a common group of genes. However, our method will likely produce the most biologically meaningful results if the same platform and experimental conditions are used for both of the two independent
experiments being compared, and also if the same conditions or closely related conditions are compared. If experiments using different platforms to measure gene expression are compared, the platform could act as a confounding factor, and thus the results might be more of a comparison of platforms and less of biological effects. Therefore, when different experimental platforms are used, researchers should exercise caution when interpreting the results. This applies not only to our proposed method but also to all other methods used to estimate $m_{11}$.

### 2.6 Appendix

Section 2.2.2 derived $E(\hat{m}_0^{(i)})$ as

$$m_0^{(i)} + \frac{1}{1 - \lambda_i} \sum_{j: H_j^{false}} \Pr(p_{ij} > \lambda_i). \quad (2.14)$$

Performing similar steps, we can calculate $E(\hat{m}_{00})$ as

$$E(\hat{m}_{00}) = \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} \sum_{j=1}^{m} E(1 \{ (p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1] \})$$

$$= \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} \sum_{j=1}^{m} \Pr((p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1])$$

$$= \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} \sum_{H_{00}} Pr \{ (p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1] \}$$

$$+ \sum_{H_{01}} Pr \{ (p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1] \} + \sum_{H_{10}} Pr \{ (p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1] \}$$

$$\quad + \sum_{H_{11}} Pr \{ (p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1] \}$$

$$= \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} \left[ \sum_{H_{00}} (1 - \lambda_1)(1 - \lambda_2) \Pr(p_{2j} > \lambda_2) 

+ \sum_{H_{01}} (1 - \lambda_2) \Pr(p_{1j} > \lambda_1) + \sum_{H_{10}} \Pr(p_{1j} > \lambda_1) \Pr(p_{2j} > \lambda_2) \right]$$

$$= m_{00} + \frac{1}{1 - \lambda_2} \sum_{H_{01}} \Pr(p_{2j} > \lambda_2) + \frac{1}{1 - \lambda_1} \sum_{H_{10}} \Pr(p_{1j} > \lambda_1)$$

$$+ \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} \sum_{H_{11}} \Pr(p_{1j} > \lambda_1) \Pr(p_{2j} > \lambda_2). \quad (2.15)$$
where

\[ H_{00} = \{ j : H_{1j} \text{ true and } H_{2j} \text{ true} \} , \]
\[ H_{01} = \{ j : H_{1j} \text{ true and } H_{2j} \text{ false} \} , \]
\[ H_{10} = \{ j : H_{1j} \text{ false and } H_{2j} \text{ true} \} , \text{ and} \]
\[ H_{11} = \{ j : H_{1j} \text{ false and } H_{2j} \text{ false} \} . \]

This leads to \( E(\hat{m}_{11}) \) as

\[
E(\hat{m}_{11}) = E(m - \hat{m}_{0}^{(1)} - \hat{m}_{0}^{(2)} + \hat{m}_{00}) \\
= m - E(\hat{m}_{0}^{(1)}) - E(\hat{m}_{0}^{(2)}) + E(\hat{m}_{00}) \\
= m - \left( \frac{m_{0}(1)}{1 - \lambda_{1}} \sum_{\{ j: H_{1j} \text{ false} \}} \Pr(p_{1j} > \lambda_{1}) \right) \\
- \left( \frac{m_{0}(2)}{1 - \lambda_{2}} \sum_{\{ j: H_{2j} \text{ false} \}} \Pr(p_{2j} > \lambda_{2}) \right) \\
+ \left( m_{00} + \frac{1}{1 - \lambda_{2}} \sum_{H_{01}} \Pr(p_{2j} > \lambda_{2}) + \frac{1}{1 - \lambda_{1}} \sum_{H_{10}} \Pr(p_{1j} > \lambda_{1}) \right) \\
+ \frac{1}{(1 - \lambda_{1})(1 - \lambda_{2})} \sum_{H_{11}} \Pr(p_{1j} > \lambda_{1}) \Pr(p_{2j} > \lambda_{2}) \\
= \left( m - m_{0}(1) - m_{0}(2) + m_{00} \right) - \frac{1}{1 - \lambda_{1}} \sum_{H_{10}} \Pr(p_{1j} > \lambda_{1}) - \frac{1}{1 - \lambda_{2}} \sum_{H_{11}} \Pr(p_{1j} > \lambda_{1}) \\
- \frac{1}{1 - \lambda_{1}} \sum_{H_{01}} \Pr(p_{2j} > \lambda_{2}) - \frac{1}{1 - \lambda_{2}} \sum_{H_{11}} \Pr(p_{2j} > \lambda_{2}) \\
+ \frac{1}{1 - \lambda_{2}} \sum_{H_{01}} \Pr(p_{2j} > \lambda_{2}) + \frac{1}{1 - \lambda_{1}} \sum_{H_{10}} \Pr(p_{1j} > \lambda_{1}) \\
+ \frac{1}{(1 - \lambda_{1})(1 - \lambda_{2})} \sum_{H_{11}} \Pr(p_{1j} > \lambda_{1}) \Pr(p_{2j} > \lambda_{2}) \\
= m_{11} - \sum_{H_{11}} \left( \frac{1}{1 - \lambda_{1}} \Pr(p_{1j} > \lambda_{1}) + \frac{1}{1 - \lambda_{2}} \Pr(p_{2j} > \lambda_{2}) \right) \\
- \frac{1}{(1 - \lambda_{1})(1 - \lambda_{2})} \Pr(p_{1j} > \lambda_{1}) \Pr(p_{2j} > \lambda_{2}) \right) \\
= m_{11} - \frac{1}{(1 - \lambda_{1})(1 - \lambda_{2})} \sum_{H_{11}} \left( (1 - \lambda_{2}) \Pr(p_{1j} > \lambda_{1}) + (1 - \lambda_{1}) \Pr(p_{2j} > \lambda_{2}) \\
- \Pr(p_{1j} > \lambda_{1}) \Pr(p_{2j} > \lambda_{2}) \right) \\
= m_{11} - \frac{1}{(1 - \lambda_{1})(1 - \lambda_{2})} \sum_{H_{11}} A_{j} , \tag{2.16}
\]
\[ A_j = (1 - \lambda_2) \Pr(p_{1j} > \lambda_1) + (1 - \lambda_1) \Pr(p_{2j} > \lambda_2) - \Pr(p_{1j} > \lambda_1) \Pr(p_{2j} > \lambda_2). \]  

(2.17)

This result directly corresponds to (2.9) in Section 2.2.4.

**Bibliography**


CHAPTER 3. IDENTIFYING GENES THAT ARE DIFFERENTIALLY EXPRESSED IN BOTH OF TWO INDEPENDENT EXPERIMENTS

A paper to be submitted to the Annals of Applied Statistics

Megan Orr, Peng Liu, and Dan Nettleton

Abstract

Identifying genes that are differentially expressed (DE) in two independent experiments generally involves two steps. In the first step, gene expressions from each experiment are analyzed separately to produce a list of genes that are declared DE in each experiment while controlling false discovery rate at some desired level $\alpha$. Then, genes common to both lists are declared to be DE in both experiments. We call this approach the “intersection method”. Little, if any, research has been done to evaluate how well this method controls FDR or ranks genes based on significance. In addition to exploring these questions, we also propose a new method for estimating FDR for the intersection method. These two methods, as well as another method developed with a different goal in mind, are compared through two simulation studies, one involving independent normal data and one involving real gene expression data. These simulation studies demonstrate the advantages of the proposed method. We conclude the paper by providing an analysis of data from two experiments involving gene expressions in maize leaves.

Key Words: Bivariate analysis; false discovery rate; multiple testing; $q$-values
3.1 Introduction

With the emergence and advancement of gene expression technologies, including microarray and the more recent RNA-seq (see Brown and Botstein, 1999, or Metzker, 2010, for reviews of these technologies), the number of gene expression data sets available for analysis as well as the number of methods proposed for analyzing these data sets are increasing exponentially. In a typical gene expression experiment consisting of subjects from two different treatments groups, thousands of expression levels, one corresponding to each gene, are measured for each subject. A major goal of these experiments is to identify genes that are differentially expressed (DE), i.e., have a difference in treatment expression means. These groups of “important” genes are then used to draw potentially important biological conclusions, so it is critical for the results of a gene expression experiment to be reliable. This is generally accomplished by identifying genes using a method that controls false discovery rate (FDR) (Benjamini and Hochberg, 1995) at some level $\alpha$.

FDR is the expected proportion of equivalently expressed (EE) genes (i.e., genes that have equal treatment expression means) among all genes declared to be DE. FDR is generally the preferred error rate to control in gene expression experiments. Many traditional multiple testing situations control family-wise error rate (FWER) using approaches such as the Bonferroni method or Holm’s (1979) method, but these methods provide extremely low power for detecting DE genes due to the high number (usually tens of thousands) of hypotheses being tested simultaneously. When controlling FWER and FDR at the same numeric level $\alpha$, FDR provides much more power for detecting DE genes, albeit at the cost of a larger number of Type I errors. Researchers are generally willing to allow for more Type I errors as long the number of errors is small relative to number of genes correctly identified as DE. Thus, the concept of FDR control is well-suited for gene expression studies.

Although most gene expression research has focused on the analysis of a single experiment, many researchers are interested in comparing the results of two or more related experiments. More specifically, many researchers are interested in identifying genes that are DE in both of two experiments. The only method that we are aware of with this specific purpose involves
two steps. In the first step, each gene expression data set is analyzed separately to produce a list of genes declared to be DE, while controlling FDR at some level $\alpha$. Then, genes that are common to both lists are declared DE in both experiments. We will call this method of gene identification the “intersection method”. For examples of where the intersection method is implemented, see Ianculescu et al. (2012), Voineagu et al. (2011), Covshoff et al. (2008), Buchanan-Wollaston et al. (2005), Wang et al. (2004), and Akopyants et al. (2004).

As a specific example, consider the experiments described in Covshoff et al. (2008). Two experiments were performed, one in the mesophyll (M) cell type of maize leaves and one in the bundle sheath (B) cell type of maize leaves. In each experiment, gene expression levels were compared between two genotypes: wild-type and mutant. The mutant cells lacked the PSII activity of the wild-type cells, and a major goal of the researchers was to identify genes that were differentially expressed, due to the lack of PSII activity, in both the M and B experiments. In order to accomplish this, researchers used the intersection method to produce a list of genes declared to be DE in both experiments.

Little, if any, research has been done to evaluate the performance of the intersection method. Thus, a goal of this paper is to investigate how well the intersection method controls FDR. Additionally, we propose a new method for estimating FDR for a given list of genes identified by the intersection method. We compare our newly proposed method to the intersection method as well as another existing method with respect to ranking genes based on evidence of differential expression, the number of truly DE genes declared to be DE, and FDR control. The comparisons among methods are carried out through two simulation studies, one involving independent normally distributed data and one involving real microarray data. We will also use each method to analyze the data described in Covshoff et al. (2008) and compare the results.

The rest of the paper is organized as follows. Section 3.2 reviews FDR and the $q$-value method, which identifies DE genes while controlling FDR in a single experiment. Section 3.3 discusses extending FDR analysis from one experiment to two independent experiments. Section 3.4 introduces existing methods and the proposed method for identifying genes that are DE in two experiments while controlling FDR. Section 3.5 describes the simulation studies that were performed and presents the results of these studies. Section 3.6 analyzes the data
Table 3.1: Table of random variables corresponding to frequencies associated with all four outcomes when testing $m$ null hypotheses.

from Covshoff et al. (2008). The paper concludes with some discussion in Section 3.7.

### 3.2 Review of FDR Analysis for One Experiment

This section begins by reviewing false discovery rate (FDR) in Section 3.2.1. Then, a common approach for identifying DE genes while controlling FDR in a single experiment proposed by Storey (2002) will be described in Section 3.2.2. Finally, in Section 3.2.3, methods for estimating the number of genes that are EE in one experiment, a quantity essential for estimating FDR using Storey’s (2002) approach, will be discussed.

#### 3.2.1 False Discovery Rate

Consider simultaneously testing $m$ null hypotheses $H_1, \ldots, H_m$ corresponding to genes $j = 1, \ldots, m$, for a single experiment, and let $p_j$ be the $p$-value that corresponds to testing the null hypothesis $H_j$. If gene $j$ is EE, then we will assume $p_j$ follows a Uniform(0,1) distribution. If gene $j$ is DE, then we will assume $p_j$ follows a distribution that is stochastically smaller than uniform. Each gene $j$ will then be declared DE by rejecting $H_j$ or EE by accepting (failing to reject) $H_j$ based on some significance measure. Table 3.1 presents a frequency table of the four possible outcomes of testing these $m$ null hypotheses.

<table>
<thead>
<tr>
<th></th>
<th>Declared EE</th>
<th>Declared DE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truly EE</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Truly DE</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>
duced by Benjamini and Hochberg (1995), provides a much better balance between the number “false discoveries” (Type I errors) and the number of “true discoveries” (correctly rejected null hypotheses). Formally, FDR is defined as

\[ \text{FDR} = E \left[ \frac{V}{\max\{R, 1\}} \right], \]  

(3.1)

the expected proportion of false discoveries among all “discoveries”, i.e., rejected null hypotheses.

### 3.2.2 Storey’s q-value

For a set of \( m \) \( p \)-values and rejection region \([0, \gamma]\), Storey (2002) proposed an estimator of FDR as

\[ \hat{\text{FDR}}(\gamma) = \frac{\hat{m}_0 \gamma}{\max\{R(\gamma), 1\}}, \]  

(3.2)

where

\[ R(\gamma) = \sum_{k=1}^{m} 1 \{p_k \leq \gamma\} \]  

(3.3)

and \( \hat{m}_0 \) is an estimator of \( m_0 \), the number of genes that are EE in the experiment (see Table 3.1). Estimating \( m_0 \) to obtain \( \hat{m}_0 \) is discussed in Section 3.2.3.

The numerator in (3.2) is the estimator of the expected number of EE genes that are declared DE (i.e., the number of Type I errors) if \( \gamma \) is used as a significance cutoff for the \( p \)-values. This is simply an estimator of the expectation of the numerator in (3.1). This follows from the assumption that \( p \)-values corresponding to EE genes are uniformly distributed. \( R(\gamma) \) in the denominator of (3.2) is simply the number of genes that are declared significant for the rejection region \([0, \gamma]\), i.e., the number of \( p \)-values less than or equal to \( \gamma \), and directly corresponds to \( R \) in (3.1). Thus, (3.2) is a reasonable estimator of FDR.

Whereas the \( p \)-value is a significance measure for the Type I error rate for a single test, Storey (2002) proposed an analogous significance measure for FDR called the “q-value”. For the set of \( p \)-values from a single experiment, the q-value for gene \( j \) is defined as

\[ q_j = \min_{k:p_k \geq p_j} \left\{ \frac{\hat{\text{FDR}}(p_k)}{1} \right\}. \]  

(3.4)
Specifically, $q_j$ is the estimated FDR if gene $j$ and all other genes with $q$-values smaller than or equal to $q_j$ are declared to be DE.

### 3.2.3 Estimation of $m_0$ by $\lambda$ Selection

Obtaining a reliable estimate of $m_0$, the number of genes that are EE in an experiment, is a key step in estimating false discovery rate. Most methods aim to do this by estimating the value of the $p$-value density at 1 for a given set of $p$-values and multiplying this value by $m$.

For a single experiment Storey (2002) proposed a method for doing this by estimating $m_0$ as

$$\hat{m}_0(\lambda) = \sum_{j=1}^{m} \frac{1 \{ p_j > \lambda \}}{(1 - \lambda)}$$

for $\lambda \in (0, 1)$. The reasoning behind this estimator is that $p$-values corresponding to DE genes are generally small, so if $\lambda$ is large enough, then most genes with $p$-values greater than $\lambda$ should be EE. Then, by momentarily assuming that all genes with $p$-values greater than $\lambda$ are EE, the uniformity assumption can be used to conservatively estimate the total number of EE genes in the experiment.

Although any value of $\lambda$ will result in a conservative estimator of $m_0$, adaptively choosing $\lambda$ based on the $p$-values can result in an improved estimator. Storey (2002) explained that the bias of $\hat{m}_0(\lambda)$ decreases but the variance of $\hat{m}_0(\lambda)$ increases as $\lambda \to 1$, in most cases. Thus, it is important to choose a $\lambda$ that provides a reasonable trade-off between bias and variance. Storey (2002) and Storey et al. (2004) both used similar bootstrapping methods which attempted to choose a $\lambda$ that minimized the mean square error of $\hat{m}_0(\lambda)$. Nettleton et al. (2006) and Liang and Nettleton (2012) use a “histogram-based” method that selects a value of $\lambda$ where the distribution of the $p$-values smaller than $\lambda$ is approximately decreasing and the distribution of the $p$-values larger than $\lambda$ is approximately uniform.

### 3.3 Extending False Discovery Rate Analysis to the Two Experiment Case

Section 3.2 reviewed current methods for identifying genes that are DE in one experiment while controlling FDR. This section explains how we extend the idea of assessing differential expression in one experiment to doing so in two experiments. Specifically, we are
Table 3.2: Contingency table of frequencies based on cross classification of the expression status (EE or DE) for each of the \( m \) genes by experiment.

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene EE</td>
<td>Gene DE</td>
</tr>
<tr>
<td>( m_{00} )</td>
<td>( m_{10} )</td>
</tr>
<tr>
<td>( m_{01} )</td>
<td>( m_{11} )</td>
</tr>
<tr>
<td>( m^{(2)}_0 )</td>
<td>( m^{(2)}_1 )</td>
</tr>
</tbody>
</table>

interested in identifying genes that are DE in both of two independent experiments \( i = 1, 2 \) while controlling false discovery rate. Hence, now consider testing \( m \) pairs of null hypotheses \((H_1^{(1)}, H_1^{(2)}), \ldots, (H_m^{(1)}, H_m^{(2)})\) that correspond to \( m \) pairs of \( p \)-values \((p_1^{(1)}, p_1^{(2)}), \ldots, (p_m^{(1)}, p_m^{(2)})\), where \( H_j^{(i)} \) and \( p_j^{(i)} \) are the null hypothesis and \( p \)-value, respectively, for gene \( j \) in experiment \( i \). Thus, we are now interested in identifying genes in which both \( H_j^{(1)} \) and \( H_j^{(2)} \) are false.

### 3.3.1 Expression Status in Two Experiments

For a set of two independent experiments with \( m \) genes in common, each gene can be cross classified based on its expression status (EE or DE) by experiment. The frequencies based on these cross classifications are summarized in Table 3.2. As an example, the quantity \( m_{00} \) is the number of genes that are EE in both experiment 1 and experiment 2, whereas \( m_{10} \) is the number of genes that are DE in experiment 1 but EE in experiment 2. The estimators of the values in Table 3.2 using methods developed in Orr et al. (2012) will be discussed in Section 3.3.2, and the estimates themselves will be used to estimate FDR using the method discussed in Section 3.4.2.

### 3.3.2 Estimating Cross Classification Frequencies Based on the Differential Expression Statuses of Genes

The method presented in Sections 3.4.2 require the estimation of the quantities in Table 3.2. The number of EE genes (and thus the number of DE genes) in experiment \( i \), \( m_0^{(i)} \) (and \( m_1^{(i)} = m - m_0^{(i)} \), can be estimated as described in Section 3.2.3.
Orr et al. (2012) proposed a method for conservatively estimating $m_{00}$ as

$$\hat{m}_{00} = \sum_{j=1}^{m} \mathbf{1}\left\{ \left( p_j^{(1)}, p_j^{(2)} \right) \in [\lambda_1, 1] \times [\lambda_2, 1] \right\} \frac{1}{(1 - \lambda_1)(1 - \lambda_2)},$$

(3.6)

where $\lambda_i$ is the $\lambda$ value chosen for experiment $i$ (see Section 3.2.3). After estimating $m_0^{(1)}$, $m_0^{(2)}$, and $m_{00}$, Table 3.2 easily allows us to see that $m_{01}$ can be estimated as

$$\hat{m}_{01} = \hat{m}_0^{(1)} - \hat{m}_{00},$$

(3.7)

and $m_{10}$ can similarly be estimated as

$$\hat{m}_{10} = \hat{m}_0^{(2)} - \hat{m}_{00}.$$  

(3.8)

The estimate of $m_{11}$ is not directly required for any of the methods described in Section 3.4. Also note that for the purposes of this paper, the histogram-based method of Liang and Nettleton (2012) will be used to select $\lambda_1$ and $\lambda_2$, which is required for the estimation of all quantities in Table 3.2.

### 3.3.3 Assessing False Discovery Rate in Two Experiments

Because we are now analyzing two experiments and identifying genes that are DE in both, a gene is considered to be null if it is EE in at least one experiment. Thus, a “discovery” now refers to a gene that we declare DE in both experiments, a “true discovery” refers to a gene that is correctly declared to be DE in both experiments, and a “false discovery” refers to a gene that is EE in at least one experiment, but is declared to be DE in both. Table 3.3 alters Table 3.1 in order to reflect this new situation. From this table, it is clear that we can still define FDR as in (3.1), although $V$ and $R$ correspond to different quantities than those defined in Section 3.2.1.

### 3.4 Methods for False Discovery Rate Analysis in Two Experiments

This section presents potential significance measures for identifying genes that are DE in both of two independent experiments. Section 3.4.1 describes the intersection method, the most widely-used method. Section 3.4.2 proposes an alternative approach to FDR estimation which
Table 3.3: Table of random variables corresponding to frequencies associated with all four outcomes when testing $m$ pairs of hypotheses.

<table>
<thead>
<tr>
<th></th>
<th>Declared EE in $\geq$ one experiment</th>
<th>Declared DE in both experiments</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE in $\geq$ one experiment</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_{00} + m_{01} + m_{10}$</td>
</tr>
<tr>
<td>DE in both experiments</td>
<td>$T$</td>
<td>$S$</td>
<td>$m_{11}$</td>
</tr>
<tr>
<td>Total</td>
<td>$m - R$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

extends Storey’s (2002) FDR estimator from the one experiment case to the two experiment case and also uses information based on the intersection method. Finally, we conclude with Section 3.4.3, which describes a method, originally proposed with a different goal in mind, that uses local false discovery rates (lfdrs) to identify genes that are DE in both of two experiments.

### 3.4.1 Intersection Method

The “intersection method” is widely used for identifying genes that are DE in two experiments and is generally performed as follows. First, each experiment is analyzed separately, and a list of genes declared to be DE while controlling FDR at level $\alpha$ is produced for each experiment. Then, genes that are common to both lists are declared DE in both experiments. When the $q$-value method is used to control FDR in each experiment, the intersection method is equivalent to declaring a gene DE in both experiments if and only if

$$q^{\max} \equiv \max \left\{ q^{(1)}, q^{(2)} \right\}$$

is less than or equal to $\alpha$. As far as we know, there has been no extensive research on the FDR level achieved by this procedure. Thus we will investigate how well the statistic $q^{\max}$ controls FDR in the simulation studies of Section 3.5.

### 3.4.2 Proposed Method for Assessing False Discovery Rate using a Rejection Region Based on $q^{\max}$

This section introduces a bivariate extension of Storey’s (2002) FDR estimation approach by proposing a method for estimating FDR given a rectangular rejection $p$-value region in Section 3.4.2.1. Then in Section 3.4.2.2, we describe how a rectangular $p$-value rejection region
can be obtained for a given value of $q^{\text{max}}$ and how to calculate a corresponding FDR estimate based on this region.

### 3.4.2.1 Bivariate Analog of Storey’s False Discovery Rate Estimator

Let $\bar{\gamma}^{(i)}$ denote the probability that the $p$-value for a randomly selected DE gene in experiment $i$ is less than or equal to $\gamma^{(i)}$. Using properties of conditional probability, we can estimate $\bar{\gamma}^{(i)}$ as

$$\hat{\bar{\gamma}}^{(i)} = \frac{R^{(i)}(\gamma^{(i)}) - \hat{m}_0^{(i)} \gamma^{(i)}}{m - \hat{m}_0^{(i)}},$$  

where

$$R^{(i)}(\gamma^{(i)}) = \sum_{k=1}^{m} 1 \left\{ p_k^{(i)} \leq \gamma^{(i)} \right\}.$$  

For a rectangular rejection region $[0, \gamma^{(1)}] \times [0, \gamma^{(2)}]$ and $m$ pairs of $p$-values $(p_1^{(1)}, p_1^{(2)}), \ldots, (p_m^{(1)}, p_m^{(2)})$, we propose an estimator of FDR as

$$\hat{\text{FDR}}(\gamma^{(1)}, \gamma^{(2)}) = \frac{\hat{m}_{00} \gamma^{(1)} \gamma^{(2)} + \hat{m}_{01} \hat{\gamma}^{(1)} \gamma^{(2)} + \hat{m}_{10} \hat{\gamma}^{(1)} \gamma^{(2)}}{\max \left\{ R(\gamma^{(1)}, \gamma^{(2)}), 1 \right\}},$$  

where

$$R(\gamma^{(1)}, \gamma^{(2)}) = \sum_{k=1}^{m} 1 \left\{ \left( p_k^{(1)}, p_k^{(2)} \right) \in [0, \gamma^{(1)}] \times [0, \gamma^{(2)}] \right\}$$  

and the formulas for the estimators $\hat{m}_{00}$, $\hat{m}_{01}$, and $\hat{m}_{10}$ are given in (3.6), (3.7), and (3.8), respectively. Similar to Storey’s (2002) estimator of the expected number of Type I errors given in the numerator of (3.2), the numerator of (3.12) is also an estimator of the expected number of Type I errors in the two experiment case. In the two experiment case, there are three circumstances that constitute the occurrence of a Type I error: a gene that is EE in both experiments declared to be DE in both experiments (call this C1), a gene that is EE in the first experiment and DE in the second experiment declared to be DE in both experiments (call this C2), and a gene that is DE in the first experiment and EE in the second experiment declared to be DE in both experiments (call this C3). The first, second, and third quantities in the numerator of (3.12) correspond to estimators of the number of occurrences of C1, C2, and C3,
respectively. Summing these three estimators results in an estimator of the total number of Type I errors. Furthermore, $R(\gamma^{(1)}, \gamma^{(2)})$ in equation (3.12) directly corresponds to the quantity $R$ in the denominator of (3.1) for the two experiment case. Thus, the estimator of FDR in (3.12) is a bivariate analog to Storey’s 2002 FDR estimator as shown in (3.2).

**3.4.2.2 Proposed Estimator of FDR for a Rectangular Rejection Region Based on $q_{j}^{\text{max}}$**

Suppose that we are given $p$-values $(p_{1}^{(1)}, p_{1}^{(2)}), \ldots, (p_{m}^{(1)}, p_{m}^{(2)})$ and corresponding $q$-values $(q_{1}^{(1)}, q_{1}^{(2)}), \ldots, (q_{m}^{(1)}, q_{m}^{(2)})$. For each gene $j = 1, \ldots, m$, we calculate

$$q_{j}^{\text{max}} = \max \left\{ q_{j}^{(1)}, q_{j}^{(2)} \right\}. \quad (3.14)$$

Now we can find the largest $p$-value in each experiment that corresponds to a $q$-value less than or equal to $q_{j}^{\text{max}}$. More specifically, for each experiment $i = 1, 2$ and each gene $j = 1, \ldots, m$, we define $\gamma_{j}^{(i)}$ as

$$\gamma_{j}^{(i)} = \max \left\{ p_{k}^{(i)} : k \in Q(i, j) \right\}, \quad (3.15)$$

where $Q(i, j) = \left\{ k : q_{k}^{(i)} \leq q_{j}^{\text{max}} \right\}$. Thus, for each gene $j$, we can find a rectangular rejection region $[0, \gamma_{j}^{(1)}] \times [0, \gamma_{j}^{(2)}]$ corresponding to the $q_{j}^{\text{max}}$ rejection region $[0, q_{j}^{\text{max}}]$. Therefore, we propose to estimate the FDR associated with the $q_{j}^{\text{max}}$ rejection region $[0, q_{j}^{\text{max}}]$ for each gene $j = 1, \ldots, m$ as

$$\hat{\text{FDR}}(q_{j}^{\text{max}}) = \hat{\text{FDR}}(\gamma_{j}^{(1)}, \gamma_{j}^{(2)}) = \frac{\hat{m}_{00}\gamma_{j}^{(1)}\gamma_{j}^{(2)} + \hat{m}_{01}\gamma_{j}^{(1)}\hat{\gamma}_{j}^{(2)} + \hat{m}_{10}\hat{\gamma}_{j}^{(1)}\gamma_{j}^{(2)}}{R(\gamma_{j}^{(1)}, \gamma_{j}^{(2)})}. \quad (3.16)$$

This estimator is the same as that of equation (3.12), but we replace $\gamma^{(1)}$ and $\gamma^{(2)}$ with the observed values, $\gamma_{j}^{(1)}$ and $\gamma_{j}^{(2)}$, based on the observed $p$-values.

In order to ensure that $\hat{\text{FDR}}(q_{j}^{\text{max}})$ increases monotonically with increasing $q_{j}^{\text{max}}$, we generalize Storey’s 2002 single-experiment $q$-value to the two experiment case to obtain

$$q_{j}^{(i)} = \min \left\{ \hat{\text{FDR}}(q_{k}^{\text{max}}) : k \in K(j) \right\}, \quad (3.17)$$

where $K(j) = \left\{ k : q_{k}^{\text{max}} \geq q_{j}^{\text{max}} \right\}$, as a $q$-value for the intersection method.
3.4.3 An Alternative Strategy Based on Local False Discovery Rates

Let $t^{(i)}_j$ represent the $t$-statistic for gene $j$ in experiment $i$. Then the $z$-value corresponding to $t^{(i)}_j$ is calculated as

$$z^{(i)}_j = \Phi^{-1} \left( F_r \left( t^{(i)}_j \right) \right), \tag{3.18}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative density function (CDF) and $F_r(\cdot)$ is the CDF of a $t$-distribution with $r$ degrees of freedom.

Lai et al. (2007) proposed a method to determine if the results of two large replicate experiments are consistent (or concordant) enough to combine their data for the purpose of a more powerful analysis. Because the two data sets are collected for the same study, we would expect the expressions in one experiment to act similarly to expressions in the second experiment. More specifically, we would expect genes that exhibit “up-regulation” – i.e., a higher mean expression in one treatment than the other – in one experiment to exhibit up-regulation in the second experiment, and genes that exhibit down-regulation in one experiment to exhibit down-regulation in the second experiment. In order to determine how concordant two experiments are, Lai et al.’s (2007) method includes fitting multiple bivariate normal mixture models to the paired $z$-values, $(z^{(1)}_1, z^{(2)}_1), \ldots, (z^{(1)}_m, z^{(2)}_m)$, and performing likelihood ratio tests to determine which model most appropriately fits the $z$-values. For a given gene, the most general of these models is

$$f \left( z^{(1)}, z^{(2)} \right) = \sum_{s=0}^{2} \sum_{t=0}^{2} \pi_{st} \phi \left( z^{(1)}; \beta_s, \tau^2_s \right) \phi \left( z^{(2)}; \gamma_t, \nu^2_t \right), \tag{3.19}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the probability density function (pdf) of a normal distribution with mean $\mu$ and variance $\sigma^2$. The parameters in (3.19) are constrained as follows: $\beta_0 = \gamma_0 = 0$; $\tau^2_0 = \nu^2_0 = 1$; $\beta_1, \gamma_1 \leq 0$; $\beta_2, \gamma_2 \geq 0$; and $\sum_{st} \pi_{st} = 1$. Parameters were estimated using an expectation-maximization algorithm, and the R code used for this estimation is available at http://home.gwu.edu/~ylai/research/Concordance.

In model (3.19), $(z^{(1)}, z^{(2)})$ comes from component $(s, t)$ with probability $\pi_{st}$. Also note that if $s = 0$, this implies that $z^{(1)}$ follows a standard normal distribution and the gene corresponding to this $z$-value is EE in experiment 1. Similarly, if $t = 0$, then this gene is EE in experiment 2.
As an extension to simply testing for concordance between two replicate experiments, Lai et al. (2009) proposed a method for ranking genes with respect to concordance. This was done by calculating the statistic, $S_j$, for each gene $j = 1, \ldots, m$ as

$$S_j = \frac{\sum_{s=1}^{2} \sum_{t=0}^{2} \hat{\pi}_{st} \phi \left( z_{j}^{(1)}; \hat{\beta}_s, \hat{\tau}_s^2 \right) \phi \left( z_{j}^{(2)}; \hat{\gamma}_t, \hat{\nu}_t^2 \right)}{\sum_{s=0}^{2} \sum_{t=0}^{2} \hat{\pi}_{st} \phi \left( z_{j}^{(1)}; \hat{\beta}_s, \hat{\tau}_s^2 \right) \phi \left( z_{j}^{(2)}; \hat{\gamma}_t, \hat{\nu}_t^2 \right)},$$

(3.20)

where $\hat{\theta}$ denotes the maximum likelihood estimate of $\theta$ for any parameter $\theta$.

The numerator in (3.20) is the contribution to $f \left( z^{(1)}, z^{(2)} \right)$ from components corresponding to complete concordance, i.e. from genes that exhibit either up-regulation in both experiments ($s = 2$ and $t = 2$) or down-regulation in both experiments ($s = 1$ and $t = 1$). Thus, the gene with the highest $S_j$ is considered the most concordant and is ranked first while the gene with the smallest $S_j$ is declared least concordant and is ranked last.

Note that $S_j$ estimates the probability $\Pr(\text{gene } j \text{ is concordant} | z^{(1)} = z_j^{(1)}, z^{(2)} = z_j^{(2)})$. The quantity $S_j$ is closely related to the local false discovery rate (lfdr) (Efron and Tibshirani, 2002), which, in a general situation, estimates the probability that case $j$ is null given the value of the test statistic(s). Thus, in the case of testing for concordance, the lfdr for gene $j$ would be $1 - S_j$ or $\Pr(\text{gene } j \text{ is not concordant} | z^{(1)} = z_j^{(1)}, z^{(2)} = z_j^{(2)})$.

Our purpose is different than the purpose in Lai et al. (2009) because we wish to compare two related experiments, not two replicate experiments. Thus, we are interested in identifying genes that are DE in both experiments and are not concerned if a gene is up-regulated in experiment 1 and down-regulated in experiment 2, or vice versa. Therefore, we define a “null case” as a gene that is EE in at least one experiment and a “non-null case” as a gene that is DE in both experiments. Thus, we can consider

$$f_0 \left( z^{(1)}, z^{(2)} \right) = \sum_{(s,t) \in C_0} \pi_{st} \phi \left( z^{(1)}; \beta_s, \tau_s^2 \right) \phi \left( z^{(2)}; \gamma_t, \nu_t^2 \right),$$

(3.21)

where $C_0 = \{(s, t); s = 0 \text{ or } t = 0\}$, the contribution to $f \left( z^{(1)}, z^{(2)} \right)$ from components that corresponds to genes that are EE in at least one experiment. Then, the estimated lfdr for gene
\( j \) is

\[
\text{lfdr}_j = \Pr(\text{gene } j \text{ is EE in at least one experiment}\mid z^{(1)}_j, z^{(2)}_j) = \frac{\hat{f}_0(\hat{z}^{(1)}_j, z^{(2)}_j)}{\hat{f}(\hat{z}^{(1)}_j, z^{(2)}_j)},
\]

where \( \hat{f} \) and \( \hat{f}_0 \) represent the densities in (3.19) and (3.21) evaluated using the parameter estimates.

The estimated FDR associated with rejecting the null hypotheses for the \( j \) genes with the smallest estimated \text{lfdr} can then be estimated by

\[
\hat{\text{FDR}}(j) = \frac{1}{j} \sum_{k=1}^{j} \text{lfdr}(k),
\]

where \( \text{lfdr}(1) \leq \text{lfdr}(2) \leq \cdots \leq \text{lfdr}(m) \) are the ordered \text{lfdr} values.

### 3.5 Simulation Studies

To evaluate the performance of the methods described in Section 3.4, two simulation studies were performed. For each simulated data set consisting of \( m = 10,000 \) genes for both experiment 1 and experiment 2, the data were analyzed as follows. For each gene \( j = 1, \ldots, m \), in each experiment \( i = 1, 2 \), the null hypothesis

\[
H^{(i)}_j : \mu^{(i)}_{j1} = \mu^{(i)}_{j2}
\]

was tested against a two-sided alternative, where \( \mu^{(i)}_{jk} \) is the population mean expression value for gene \( j \) in treatment \( k \) of experiment \( i \). The test statistic, \( t^{(i)}_j \), and corresponding \( p \)-value \( p^{(i)}_j \) were obtained using the moderated \( t \)-test approach of Smyth (2004). This approach was designed specifically for analyzing microarray data sets and uses information from all gene in order to better estimate the error variance of each individual gene. The \( p \)-values were then converted to their corresponding \( q \)-values using the histogram-based method of Liang and Nettleton (2012). The pairs of \( t \)-statistics \( (t^{(1)}_1, t^{(2)}_1, \ldots, (t^{(1)}_m, t^{(2)}_m) \), pairs of \( p \)-values, \( (p^{(1)}_1, p^{(2)}_1), \ldots, (p^{(1)}_m, p^{(2)}_m) \), pairs of \( q \)-values, \( (q^{(1)}_1, q^{(2)}_1), \ldots, (q^{(1)}_m, q^{(2)}_m) \) were then used, when necessary, to estimate FDR using the methods described in Section 3.4.
The first of the two simulation studies used independent normally distributed data (generated as described in Section 3.5.1) in order to evaluate the methods under the most ideal conditions, i.e., under conditions that were consistent with the assumptions used to derive the moderated $t$-test (Smyth, 2004). The second simulation study (generated as described in Section 3.5.2) used real microarray data in order to evaluate the methods under more realistic conditions, namely when the correlation structure of the data is complex and cannot be precisely modeled using the available data. The data set used for the second simulation study is described in Metzeler et al. (2008) and is available on the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE12417. In each simulation study, three variables were manipulated to create different simulation settings. These variables are as follows.

1. $n = 4, 10, 20$: the sample size of each treatment in each experiment
2. $\mu_\delta = 1, 2$: the relative mean effect size for DE genes
3. $m_{11} = 500, 1000, 2000, 3000$: the number of genes DE in both experiments.

Furthermore, for $m_{11} = 500, 1000, 2000, 3000$, $m_{00}$ was set to 9000, 7000, 5000, and 3000, respectively, and for a setting with given $m_{11}$ and $m_{00}$ values, $m_{01} = m_{10}$ was set to half the value of $m - m_{00} - m_{11}$. Refer to Table 3.2 for more information on these quantities.

The reason that relatively small sample sizes ($n = 4, 10$ and $20$) were used is because most gene expression experiments have low sample size due to the cost of experimentation. In fact, about 40% of available gene expression data sets on the GEO (Edgar et al., 2002) have a total sample size (i.e., the sample size of all subjects in all treatments) of at most 8 and about 90% have a total sample size of at most 40, which corresponds to $n = 4$ and $n = 20$ in our simulation studies. Additionally, because many gene expression data sets have low average power for the detection of DE genes, the relative effect size was also chosen to be relatively small ($\mu_\delta = 1$ or 2).

For each simulation setting in each simulation study, 100 data sets were randomly generated. The methods were then compared based on three criteria that will be described in Section 3.5.4.
3.5.1 Simulations Using Independent Normal Data

For gene $j$ in experiment $i$, data were randomly generated as follows. The variance, $\sigma_{(i)}^{2.}$ was randomly drawn from an inverse gamma distribution. The parameters of this distribution were estimated from a real microarray data set described in Hannenhalli et al. (2006) that is available on the GEO under accession number GSE5406. If gene $j$ in experiment $i$ was EE, then $n$ values were randomly drawn from a Normal($0, \sigma_{(i)}^{2.}$) for both treatment 1 and treatment 2. If gene $j$ in experiment $i$ was DE, then a relative effect size, $\delta_{(i)}^{j}$ was randomly drawn from a Normal($\mu_\delta, 1$) or Normal($-\mu_\delta, 1$) distribution, each with equal probability. Then, $n$ values were randomly drawn from a Normal($0, \sigma_{(i)}^{2.}$) distribution for treatment 1 and $n$ values were randomly drawn from a Normal($\sigma_{(i)}^{j}\delta_{(i)}^{j}, \sigma_{(i)}^{2.}$) distribution for treatment 2.

3.5.2 Simulations Using Gene Expression Data

Although the data set from the experiment described in Metzeler et al. (2008) contained expression from over 20,000 genes, $m = 10,000$ genes were randomly selected to be included for analysis in the microarray simulation study. Also, in order to simulate two separate experiments, $N_1 = 82$ of the total $N = 163$ subjects in this experiment were chosen to be the “population” of subjects from experiment 1 and the remaining $N_2 = 81$ subjects were the “population” from experiment 2. For each gene in each experiment, the variance $s_{(i)}^{2.}$ was calculated from all $N_i$ subjects.

Each data set for this simulation study was randomly generated as follows. For the set of $m$ genes, $m_{00}$ genes were randomly selected to be EE in both experiments, $m_{01}$ of the remaining genes were then randomly selected to be EE in experiment 1 but DE in experiment 2, $m_{10}$ of the remaining genes were then randomly selected to be DE in experiment 1 but EE in experiment 2, and the remaining $m_{11}$ genes were DE in both experiments. For each experiment, the expressions from $n$ subjects were randomly chosen, without replacement, to be in each treatment group. If gene $j$ was selected to be EE in experiment $i$, then the data from the selected subjects were not altered in any way for either treatment. Note that the subjects in each treatment group were randomly selected from the same population, and therefore had the
same mean expression value (and were thus EE). If gene \( j \) was selected to be DE in experiment \( i \), then the relative effect size \( \delta_j^{(i)} \) was randomly generated as described in Section 3.5.1, but replacing \( \sigma_j^{2,(i)} \) with \( s_j^{2,(i)} \). Finally, the expressions from gene \( j \) in treatment 1 were not altered in any way, but \( s_j^{(i)} \delta_j^{(i)} \) was added to the expressions in treatment 2. Note that simulating data using this approach does not alter the correlation structure of the data in any way, and only shifts the treatment means in experiment 2 for DE genes.

### 3.5.3 Estimation Constraints

When analyzing a given data set, it is possible for \( \hat{m}_0^{(i)} \) to be larger than or equal to \( m \) for \( i = 1, 2 \). This is a problem when estimating \( \hat{\gamma}^{(i)} \) in (3.10), as the estimated probability will be either undefined (if \( \hat{m}_0^{(i)} = m \)) or negative (if \( \hat{m}_0^{(i)} > m \)). Thus, the following steps were taken to rectify this situation when it was encountered in the data analysis of Sections 3.5 and 3.6. If \( m_0^{(1)} = m \), then this implies that \( m_{10} = 0 \). Similarly, \( m_0^{(2)} = m \) implies that \( m_{01} = 0 \). Therefore, when \( \hat{m}_0^{(1)} \geq m \), we set \( \hat{m}_{10} = 0 \) as well as the quantity \( p^{(2)} \hat{\gamma}^{(i)} \hat{m}_{10} \). The analogous procedure was performed when \( \hat{m}_0^{(2)} \geq m \).

Additionally, even in cases when \( \hat{m}_0^{(i)} < m \) for \( i = 1 \) and \( i = 2 \), the estimate of \( \hat{\gamma}^{(i)} \) in (3.10) was sometimes less than \( \gamma^{(i)} \). Because this is inconsistent with the assumption that \( p \)-values from DE genes follow a stochastically smaller distribution than \( p \)-values from EE genes, \( \hat{\gamma}^{(i)} \) was set to \( \gamma^{(i)} \) in such cases.

### 3.5.4 Simulation Results

For each setting in each simulation study, three quantities were used to evaluate and compare the performance of each of the significance measures. First, Table 3.4 gives the mean partial area under the receiver operating characteristic curve (based on the 100 simulated data sets), pAUC, with its corresponding standard error in parentheses, for each simulation setting in both simulation studies. We calculated partial areas because we were only interested in the most relevant region of the receiver operating characteristic (ROC) curve which included values of the false positive rate (FPR) no larger than 0.10. The method that ranks the genes best with respect to detection of differential expression in both experiments will have the highest pAUC,
which is given in bold font for each simulation setting. Next, Tables 3.5 and 3.6 present the mean \( S \) and \( \overline{V/R} \), with their corresponding standard errors in parentheses, for each setting in the simulations using normal data and simulations using microarray data, respectively. In both of these tables, mean \( S \) corresponds to the mean number of genes that were truly DE in both experiments and also declared to be DE in both experiments. Thus, as in mean pAUC, higher values are preferred, and the highest value for each simulation setting are presented in bold font. The quantity \( \overline{V/R} \) corresponds to the empirical approximation of FDR when each method attempts to control FDR at 5%. This quantity is obtained by averaging the observed values of \( V/\max\{R, 1\} \) over the 100 simulated data sets for each simulation setting. Ideally for each method, these values will be less than or equal to 5% in each simulation setting, suggesting that the method adequately controls FDR. See Table 3.3 for the definitions of \( S \), \( V \), and \( R \).

Note that because of the approach used to obtain \( q^I \) (as described in Section 3.4.2.2), \( q^{\text{max}} \) and \( q^I \) will produce the same ranking of genes, so there is one column devoted to both methods for the mean pAUC for each simulation study in Table 3.4. For the simulations involving independent normal data, the \( \text{lfdr} \) method outperforms \( q^{\text{max}} \) in mean pAUC in all simulation settings. This is likely due to the process used to simulate the data. More specifically, for each setting in the simulation studies using normal data, expressions from each gene in a given data set were randomly drawn independently from a three density normal mixture model for each experiment (see Section 3.4.3 for details). Thus, when the \( t \)-statistics were converted to \( z \)-values for each gene, these \( z \)-values also followed a three density normal mixture model, which is the model that the \( \text{lfdr} \) method (described in Section 3.4.3) fit to the \( z \)-values. For the microarray simulations, in which data for each gene were not normally distributed and data among genes likely had a high degree of correlation, \( q^{\text{max}} \) outperformed \( \text{lfdr} \) in 18 of the 24 simulation settings. The data used in the microarray simulations are more similar to the data we would encounter in practice than independent normal data, suggesting that the \( \text{lfdr} \) method might not be robust enough to produce reliable results for real gene expression data, and thus \( q^{\text{max}} \) should be used in such situations.

The results for mean \( S \) presented in Tables 3.5 and 3.6 show that \( q^I \) outperformed both \( q^{\text{max}} \) and \( \text{lfdr} \) in 17 of the 24 settings in the simulations using normal data and 13 of the 24
<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu_s$</th>
<th>$m_{11}$</th>
<th>Normal Simulations</th>
<th>Microarray Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{\text{max}}/q^I$</td>
<td>$q_{\text{max}}/q^I$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>500</td>
<td>32.8 (0.2) 40.1 (0.2)</td>
<td>69.6 (2.5) 30.1 (1.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>29.4 (0.2) 33.1 (0.2)</td>
<td>33.0 (0.8) 29.3 (0.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>27.3 (0.1) 29.3 (0.1)</td>
<td>27.2 (0.4) 26.4 (0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3000</td>
<td>24.6 (0.1) 25.8 (0.1)</td>
<td>24.7 (0.4) 23.8 (0.5)</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>63.7 (0.2) 71.4 (0.2)</td>
<td>68.6 (0.9) 64.6 (1.2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>58.5 (0.1) 61.6 (0.1)</td>
<td>58.3 (0.6) 57.0 (0.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>54.7 (0.1) 56.2 (0.1)</td>
<td>53.5 (0.7) 52.7 (0.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>50.8 (0.1) 51.5 (0.1)</td>
<td>49.4 (0.7) 49.0 (0.7)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>500</td>
<td>50.9 (0.2) 65.8 (0.2)</td>
<td>62.7 (1.3) 62.6 (0.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>47.0 (0.1) 52.0 (0.1)</td>
<td>47.4 (0.4) 48.5 (0.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>44.0 (0.1) 46.3 (0.1)</td>
<td>44.0 (0.4) 43.7 (0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3000</td>
<td>40.9 (0.1) 42.0 (0.1)</td>
<td>40.9 (0.5) 39.6 (0.6)</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>81.5 (0.1) 89.8 (0.1)</td>
<td>83.5 (0.4) 88.0 (0.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>78.4 (0.1) 80.7 (0.1)</td>
<td>78.8 (0.3) 78.4 (0.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>76.2 (0.1) 77.2 (0.1)</td>
<td>76.5 (0.4) 75.6 (0.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>73.4 (0.1) 73.7 (0.1)</td>
<td>72.9 (0.5) 72.1 (0.5)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>500</td>
<td>63.2 (0.2) 79.5 (0.1)</td>
<td>70.6 (0.8) 78.3 (0.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>59.6 (0.2) 64.4 (0.2)</td>
<td>60.2 (0.4) 61.9 (0.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>56.9 (0.1) 58.9 (0.1)</td>
<td>57.1 (0.3) 56.8 (0.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3000</td>
<td>54.1 (0.1) 54.9 (0.1)</td>
<td>54.8 (0.4) 54.0 (0.5)</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>88.5 (0.1) 95.0 (0.1)</td>
<td>89.6 (0.3) 94.0 (0.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>86.8 (0.1) 88.4 (0.1)</td>
<td>86.8 (0.2) 87.0 (0.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>85.4 (0.1) 86.0 (0.1)</td>
<td>85.5 (0.2) 85.1 (0.3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>84.0 (0.1) 84.2 (0.1)</td>
<td>84.0 (0.3) 83.5 (0.3)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4: The mean pAUCs (in %) for each method with corresponding standard errors in parentheses for both simulations studies. Note that $q_{\text{max}}$ and $q^I$ methods will produce the same ranking of genes. For each study, each setting with the highest mean values are presented in bold font.
<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu_i$</th>
<th>$m_{11}$</th>
<th>$q_{\text{max}}^\text{max}$</th>
<th>$q_l$</th>
<th>$l_{\text{fdr}}$</th>
<th>$q_{\text{max}}^\text{V/R}$ (in %)</th>
<th>$q_l$</th>
<th>$l_{\text{fdr}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>500</td>
<td>4.9 (0.4)</td>
<td>0.1</td>
<td>(0.0)</td>
<td>0.00 (0.00)</td>
<td>2.24 (0.56)</td>
<td>1.00 (1.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>7.0 (0.5)</td>
<td>0.5</td>
<td>(0.1)</td>
<td>0.84 (0.55)</td>
<td>3.68 (0.67)</td>
<td>0.33 (0.33)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>17.9 (1.1)</td>
<td>6.0</td>
<td>(0.4)</td>
<td>3.65 (0.51)</td>
<td>3.32 (0.49)</td>
<td>0.58 (0.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3000</td>
<td>60.9 (0.8)</td>
<td>30.6</td>
<td>(1.4)</td>
<td>4.85 (0.27)</td>
<td>3.15 (0.34)</td>
<td>2.66 (0.38)</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>16.1 (0.5)</td>
<td>115.4 (1.6)</td>
<td>85.8</td>
<td>(1.7)</td>
<td>0.28 (0.12)</td>
<td>3.84 (0.19)</td>
<td>1.96 (0.15)</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>118.3 (1.2)</td>
<td>183.7 (2.2)</td>
<td>158.2</td>
<td>(1.9)</td>
<td>2.43 (0.13)</td>
<td>4.24 (0.15)</td>
<td>3.34 (0.14)</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>428.1 (2.4)</td>
<td>443.2 (3.7)</td>
<td>415.3</td>
<td>(2.7)</td>
<td>4.08 (0.09)</td>
<td>4.24 (0.10)</td>
<td>3.83 (0.09)</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>924.5 (3.5)</td>
<td>729.8 (4.7)</td>
<td>721.9</td>
<td>(3.5)</td>
<td>5.80 (0.08)</td>
<td>4.09 (0.07)</td>
<td>3.92 (0.06)</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>500</td>
<td>43.6 (0.6)</td>
<td>98.8</td>
<td>(1.1)</td>
<td>0.30 (0.07)</td>
<td>3.41 (0.18)</td>
<td>2.22 (0.15)</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>142.4 (1.2)</td>
<td>172.8 (1.8)</td>
<td>162.4</td>
<td>(1.6)</td>
<td>2.62 (0.12)</td>
<td>4.23 (0.14)</td>
<td>3.43 (0.12)</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>382.3 (1.7)</td>
<td>383.0 (2.6)</td>
<td>371.6</td>
<td>(2.1)</td>
<td>3.78 (0.09)</td>
<td>3.85 (0.10)</td>
<td>3.49 (0.09)</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>714.9 (2.7)</td>
<td>602.7 (3.5)</td>
<td>607.6</td>
<td>(2.8)</td>
<td>5.34 (0.08)</td>
<td>3.65 (0.08)</td>
<td>3.60 (0.08)</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>228.6 (1.2)</td>
<td>320.9 (1.2)</td>
<td>321.1</td>
<td>(1.2)</td>
<td>0.44 (0.04)</td>
<td>4.54 (0.11)</td>
<td>3.61 (0.10)</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>561.5 (1.5)</td>
<td>613.6 (1.5)</td>
<td>607.6</td>
<td>(1.6)</td>
<td>2.53 (0.07)</td>
<td>4.45 (0.09)</td>
<td>4.08 (0.08)</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1261.8 (2.1)</td>
<td>1279.7 (2.4)</td>
<td>1266.5</td>
<td>(2.4)</td>
<td>4.04 (0.06)</td>
<td>4.48 (0.06)</td>
<td>4.04 (0.06)</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>2051.3 (2.9)</td>
<td>1950.7 (3.3)</td>
<td>1931.4</td>
<td>(3.4)</td>
<td>6.29 (0.05)</td>
<td>4.45 (0.05)</td>
<td>4.08 (0.04)</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>500</td>
<td>134.6 (1.1)</td>
<td>198.0</td>
<td>(1.3)</td>
<td>193.7 (1.4)</td>
<td>0.38 (0.05)</td>
<td>3.85 (0.14)</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>335.1 (1.6)</td>
<td>370.0 (2.0)</td>
<td>362.3</td>
<td>(1.9)</td>
<td>2.63 (0.09)</td>
<td>4.31 (0.13)</td>
<td>3.72 (0.11)</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>774.9 (2.1)</td>
<td>779.4 (2.4)</td>
<td>771.2</td>
<td>(2.3)</td>
<td>3.67 (0.07)</td>
<td>3.80 (0.07)</td>
<td>3.51 (0.06)</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>1299.0 (2.8)</td>
<td>1195.8 (3.2)</td>
<td>1197.8</td>
<td>(3.1)</td>
<td>5.56 (0.06)</td>
<td>3.84 (0.05)</td>
<td>3.73 (0.05)</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>350.9 (1.1)</td>
<td>398.5 (1.0)</td>
<td>397.6</td>
<td>(1.0)</td>
<td>0.49 (0.05)</td>
<td>4.53 (0.10)</td>
<td>3.08 (0.08)</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>755.4 (1.2)</td>
<td>783.1 (1.3)</td>
<td>771.2</td>
<td>(1.3)</td>
<td>2.63 (0.05)</td>
<td>4.64 (0.07)</td>
<td>3.52 (0.05)</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1584.7 (1.8)</td>
<td>1595.8 (1.9)</td>
<td>1570.8</td>
<td>(2.0)</td>
<td>4.16 (0.05)</td>
<td>4.68 (0.05)</td>
<td>3.56 (0.05)</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>2465.7 (2.2)</td>
<td>2416.0 (2.4)</td>
<td>2375.8</td>
<td>(2.7)</td>
<td>6.44 (0.04)</td>
<td>4.64 (0.04)</td>
<td>3.52 (0.03)</td>
</tr>
</tbody>
</table>

Table 3.5: The mean $S$ (middle three columns) and $V/R$ for estimated FDR ≤ 5% (last three columns) for each method and simulation setting with corresponding standard errors in parentheses for the simulation study using independent normal data. For each setting the highest value of mean $S$ are presented in bold font.
<table>
<thead>
<tr>
<th>n</th>
<th>μₜ</th>
<th>m₁₁</th>
<th>qₘₐₓ</th>
<th>q²</th>
<th>lfdₚ</th>
<th>qₘₐₓ</th>
<th>V/R (in %)</th>
<th>q²</th>
<th>lfdₚ</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>500</td>
<td>0.2 (0.1)</td>
<td>8.5 (0.7)</td>
<td>32.6 (11.9)</td>
<td>0.00 (0.00)</td>
<td>5.03 (0.97)</td>
<td>9.41 (2.42)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>1.8 (0.3)</td>
<td>13.1 (1.1)</td>
<td>36.9 (17.3)</td>
<td>3.08 (1.28)</td>
<td>4.74 (0.79)</td>
<td>8.04 (2.22)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>14.8 (1.5)</td>
<td>25.8 (2.1)</td>
<td>37.2 (20.3)</td>
<td>4.59 (0.81)</td>
<td>5.35 (0.67)</td>
<td>7.21 (1.41)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>59.6 (5.2)</td>
<td>43.9 (4.4)</td>
<td>83.0 (24.6)</td>
<td>5.14 (0.60)</td>
<td>4.36 (0.59)</td>
<td>7.42 (1.19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>13.6 (1.2)</td>
<td>121.2 (3.8)</td>
<td>77.3 (7.0)</td>
<td>0.76 (0.29)</td>
<td>5.99 (0.47)</td>
<td>7.00 (1.59)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>104.2 (4.5)</td>
<td>186.2 (5.7)</td>
<td>152.3 (9.8)</td>
<td>3.42 (0.45)</td>
<td>5.86 (0.56)</td>
<td>6.04 (1.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>385.6 (11.8)</td>
<td>428.2 (10.5)</td>
<td>385.4 (10.9)</td>
<td>5.10 (0.54)</td>
<td>5.38 (0.48)</td>
<td>5.95 (0.73)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>886.7 (16.8)</td>
<td>725.1 (18.9)</td>
<td>706.7 (16.3)</td>
<td>7.13 (0.50)</td>
<td>5.46 (0.41)</td>
<td>5.67 (0.50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>500</td>
<td>42.9 (1.1)</td>
<td>108.0 (2.0)</td>
<td>90.0 (2.4)</td>
<td>0.66 (0.18)</td>
<td>4.09 (0.40)</td>
<td>6.11 (1.03)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>141.8 (2.5)</td>
<td>171.8 (2.5)</td>
<td>181.7 (6.8)</td>
<td>2.49 (0.32)</td>
<td>3.62 (0.34)</td>
<td>8.96 (1.50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>384.4 (5.4)</td>
<td>380.8 (4.5)</td>
<td>395.2 (10.9)</td>
<td>4.40 (0.51)</td>
<td>4.03 (0.44)</td>
<td>6.21 (0.95)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>734.1 (8.4)</td>
<td>615.5 (7.2)</td>
<td>685.0 (18.9)</td>
<td>6.70 (0.60)</td>
<td>4.53 (0.45)</td>
<td>7.68 (1.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>226.7 (1.8)</td>
<td>324.3 (2.4)</td>
<td>316.1 (2.5)</td>
<td>0.42 (0.08)</td>
<td>4.40 (0.33)</td>
<td>7.65 (1.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>559.2 (3.0)</td>
<td>613.3 (2.9)</td>
<td>600.5 (3.0)</td>
<td>2.71 (0.31)</td>
<td>4.27 (0.32)</td>
<td>6.45 (0.77)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>1265.5 (4.7)</td>
<td>1281.2 (4.7)</td>
<td>1266.5 (4.9)</td>
<td>4.21 (0.41)</td>
<td>4.35 (0.37)</td>
<td>5.87 (0.70)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>2062.9 (5.8)</td>
<td>1959.9 (6.0)</td>
<td>1936.8 (5.8)</td>
<td>7.70 (0.57)</td>
<td>5.46 (0.45)</td>
<td>6.04 (0.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>500</td>
<td>132.9 (1.3)</td>
<td>206.3 (2.1)</td>
<td>197.9 (3.0)</td>
<td>0.56 (0.15)</td>
<td>4.15 (0.28)</td>
<td>6.36 (1.06)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>332.8 (2.2)</td>
<td>369.4 (2.3)</td>
<td>363.2 (4.5)</td>
<td>3.06 (0.42)</td>
<td>4.29 (0.39)</td>
<td>6.46 (0.94)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>782.0 (4.4)</td>
<td>783.5 (4.2)</td>
<td>775.2 (4.7)</td>
<td>4.00 (0.35)</td>
<td>3.84 (0.28)</td>
<td>4.99 (0.44)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>1304.2 (5.8)</td>
<td>1201.8 (5.2)</td>
<td>1210.0 (7.6)</td>
<td>5.53 (0.41)</td>
<td>3.70 (0.30)</td>
<td>5.02 (0.52)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>353.0 (1.2)</td>
<td>401.1 (1.4)</td>
<td>398.6 (1.2)</td>
<td>0.71 (0.13)</td>
<td>5.06 (0.28)</td>
<td>7.42 (0.82)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>758.5 (1.5)</td>
<td>785.9 (1.4)</td>
<td>771.0 (1.5)</td>
<td>3.12 (0.43)</td>
<td>5.00 (0.47)</td>
<td>6.15 (0.73)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>1586.1 (2.8)</td>
<td>1595.0 (2.7)</td>
<td>1568.0 (2.9)</td>
<td>4.58 (0.39)</td>
<td>4.76 (0.34)</td>
<td>5.34 (0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>2469.9 (3.3)</td>
<td>2416.5 (3.1)</td>
<td>2377.4 (3.8)</td>
<td>7.41 (0.55)</td>
<td>5.20 (0.43)</td>
<td>5.07 (0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6: The mean S (middle three columns) and V/R for estimated FDR ≤ 5% (last three columns) for each method and simulation setting with corresponding standard errors in parentheses for the simulation study using microarray data. For each setting the highest value of mean S are presented in bold font.
settings in the simulations using microarray data. In many cases when $q^I$ did not perform the best, the method that did perform the best did not adequately control FDR. One example of this can be seen in the last row of Table 3.6. Although $q^{\text{max}}$ has the highest mean $S$, its estimated FDR is 7.41%, which is much higher than the desired 5% even when the standard error is taken into account.

The results in the right halves of 3.5 and 3.6 show that $q^I$ tended to be the most consistent in its control of FDR. When combining the results of both simulation studies, $q^I$ had the $V/R$ closest to 5% in 37 of the 48 total settings when all three methods were compared. It also had the smallest range of $V/R$ values when compared to the other methods. Both $q^{\text{max}}$ and lfdr failed to control FDR at 5% for multiple simulation settings.

### 3.6 Real Data Analysis

In this section, we analyze the data from two independent microarray experiments described in Covshoff et al. (2008) using each of the methods described in Section 3.4. One experiment was performed on mesophyll (M) cells in maize leaves and the other experiment was performed on the bundle sheath (B) cells in maize leaves. In each experiment, the gene expressions of maize leaves from two treatments, wild type and mutant, were measured using $n = 6$ two-color slides. Maize leaves with mutant cells lacked the PSII activity of maize leaves with wild-type cells, and researchers were interested in identifying genes that were DE in both the M and B experiments due to the lack of PSII activity.

Although the same platform was used to measure gene expressions in each experiment, a different number of genes were detected above background levels for each experiment. Therefore, the 5670 genes that were detected above background levels in both experiments were analyzed. Table 3.7 summarizes the results of analyzing these two experiments using the methods described in Section 3.4. The values on the diagonal represent the number of genes that were declared to be DE in both the M and B experiments for each method. All other values represent the overlap of genes declared to be DE when comparing two methods. The $q^{\text{max}}$ and $q^I$ performed similarly, declaring 553 and 537 genes to be DE in both the M and B experiments, respectively. Furthermore, all 537 of the genes declared to be DE in both experiments
Table 3.7: The number of genes declared to be DE in both the M and B experiments by each method as well as the overlap of genes declared to be DE when comparing two methods.

<table>
<thead>
<tr>
<th></th>
<th>$q^\text{max}$</th>
<th>$q^I$</th>
<th>lfdr</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q^\text{max}$</td>
<td>553</td>
<td>537</td>
<td>553</td>
</tr>
<tr>
<td>$q^I$</td>
<td>537</td>
<td>537</td>
<td></td>
</tr>
<tr>
<td>lfdr</td>
<td></td>
<td></td>
<td>3757</td>
</tr>
</tbody>
</table>

by $q^I$ were also declared to be DE in both experiments by $q^\text{max}$. Based on the results of the microarray simulation study, we believe it is likely that the set of genes identified by $q^I$ is more reliable than the set identified by $q^\text{max}$.

Additionally, the lfdr method declared an unusually large number (3757) of genes to be DE in both the M and B experiments. This number is higher than the estimated number of DE genes in both the M or B experiments (2703 and 3088, respectively) when they are analyzed separately using the histogram-based method. Because the histogram-based method has been shown to be a reliable method for estimating the number of EE genes (and thus, the number of DE genes) in a single experiment, we do not recommend using the lfdr method to identify genes for this experiment due to the possibility of a high Type I error rate.

### 3.7 Conclusions

In this paper, three methods for estimating FDR were compared. The proposed method for calculating $q^I$ was the only method that consistently controlled FDR and, along with $q^\text{max}$, usually resulted in a better significance ranking of genes than the lfdr method when real microarray data were used in simulation. The $q^I$ method also declared a larger number of genes truly DE in both experiments to be DE in both experiments, on average, than the $q^\text{max}$ or lfdr methods in the majority of simulation settings. Additionally, in most of the simulation settings when $q^I$ did not perform the best, the method that did perform best did not adequately control FDR. For these reasons, we recommend using $q^I$ to estimate FDR for the purpose of identifying genes that are DE in both of two independent gene expression experiments.
Bibliography


CHAPTER 4. AN IMPROVED METHOD FOR COMPUTING Q-VALUES WHEN THE DISTRIBUTION OF EFFECT SIZES IS ASYMMETRIC

A paper to be submitted to *Bioinformatics*

Megan Orr, Peng Liu, and Dan Nettleton

**Abstract**

Asymmetry is frequently observed in the empirical distribution of test statistics that results from the analysis of gene expression experiments. This asymmetry indicates an asymmetry in the distribution of effect sizes. A common method for identifying differentially expressed (DE) genes in a gene expression experiment while controlling false discovery rate (FDR) is Storey’s $q$-value method. This method ranks genes based solely on the $p$-values from each gene in the experiment. We propose a method that alters and improves upon the $q$-value method by taking the sign of the test statistics, in addition to the $p$-values, into account. Through two simulation studies (one involving independent normal data and one involving microarray data), we show that the proposed method, when compared to the traditional $q$-value method, generally provides a better ranking for genes as well as a higher number of truly DE genes declared to be DE, while still adequately controlling FDR. We illustrate the proposed method by analyzing data from an experiment involving gene expressions in maize leaves.

**Key Words:** False discovery rate; multiple testing
4.1 Introduction

Performing tens of thousands of hypothesis tests for one experiment has become commonplace as technologies for producing high-dimensional data are becoming more prominent. This is especially the case in the field of statistical genomics, where technologies such as microarray and RNA-seq (see Brown and Botstein (1999) or Metzker (2010) for reviews of these technologies) measure the abundance of mRNA transcripts for thousands of genes in each subject of a given gene expression study. Oftentimes, researchers are interested in comparing the gene expressions of subjects from two treatment groups. A major objective of such experiments is to identify genes that exhibit differential expression (DE), i.e., a difference in the population treatment mean expression levels. Meaningful biological results depend on reliable detection of differentially expressed (DE) genes. Hence, the power of detecting differential expression for genes that are truly DE should be as high as possible while genes that are equivalently expressed (EE) – i.e., genes that have no difference in the population treatment mean expression levels – should have a minimal chance of being declared DE. Thus, we must find an appropriate method for testing multiple hypotheses that provides good power while controlling some multiple testing error.

When considering a traditional multiple testing problem, family-wise error rate (FWER) is often the preferred error rate used to control multiple testing error. The Bonferroni method and Holm’s (1979) method are commonly used for this purpose. These methods, however, are not appropriate for high-dimensional gene expression data, as controlling FWER results in extremely low power for detection of differential expression. As a consequence, an alternative error rate known as false discovery rate (FDR) (Benjamini and Hochberg, 1995) is usually used. FDR is simply the expected proportion of EE genes that are declared to be differentially expressed among all genes declared to be DE (where this proportion is defined to be zero when no genes are declared to be DE). Although FDR allows for more Type I errors than FWER when controlling these different error rates at the same numeric level $\alpha$, the power for detecting differential expression is greatly increased, as discussed in Storey and Tibshirani (2003).

A common method for identifying DE genes while controlling FDR in a gene expression
experiment is the $q$-value method, first proposed by Storey (2002). The $q$-value for a given gene represents the estimated false discovery rate if the given gene and all genes with smaller $q$-values are declared to be DE. This method estimates FDR based on a set of $p$-values corresponding to $m$ hypothesis tests from an experiment. As previously mentioned, researchers are often interested in studying gene expression differences in subjects from different groups. This is most often accomplished by performing a hypothesis test for a difference in treatment expression means for each gene, converting the $p$-values from the resulting tests to $q$-values, and then declaring genes with a $q$-value less than a previously chosen FDR to be differentially expressed. Potentially relevant information that this method does not take into account are the signs of the test statistics. Figure 4.1 shows a histogram of $t$-test statistics from a microarray experiment described in Covshoff et al. (2008). In this experiment, gene expressions from wild-type cells in maize leaves were compared to those from mutant cells. Figure 4.1 shows clear asymmetry in the distribution of test statistics with more negative than positive test statistics, which indicates asymmetry in the distribution of effect sizes (i.e., the true difference in gene expression treatment means) as well.

Figure 4.1: A histogram of $t$-statistics from an experiment described in Covshoff et al. (2008) in which gene expressions from wild-type cells were compared to those in mutant cells in the mesophyll cells of maize leaves. There are clearly more negative test statistics than positive test statistics which indicates that the distribution of effect sizes is asymmetric.
We propose a new method for FDR estimation that alters the traditional $q$-value method by separating the two-sided $p$-values from an experiment into two subsets of $p$-values based on the sign of the test statistics, and then computing the $q$-values separately for each subset in order to create a better ranking of the genes with respect to differential expression. Through simulation studies using both independent normally distributed data and real microarray data, we demonstrate how the proposed method results in an improved ranking of genes with respect to differential expression over the traditional $q$-value method while still adequately controlling FDR.

The rest of the paper is organized as follows. Section 4.2 reviews Storey’s (2002) $q$-value method and introduces the proposed method for FDR estimation. Section 4.3 describes two simulation studies and used the results of these studies to compare the performances of the proposed method and traditional $q$-value method by looking at the ranking of genes with respect to differential expression, the number of truly DE genes declared to be DE, and also how well each method controls FDR. Section 4.4 analyzes a real microarray data set. Section 4.5 concludes the paper with some discussion.

4.2 Methods

This section describes the proposed method for estimating FDR when effect sizes in an experiment are asymmetric. Section 4.2.1 reviews Storey et al.’s (2004) $q$-value method and Section 4.2.2 describes how the proposed method alters Storey’s method to obtain a better FDR estimate. Finally Section 4.2.3 discusses the advantages of the proposed method.

4.2.1 Review of Storey’s $q$-value Method

Suppose we wish to test $m$ null hypotheses $H_1, \ldots, H_m$, where $H_j$ is true if gene $j$ is EE and false if gene $j$ is DE. We will assume that $p_j$, the $p$-value that corresponds to $H_j$, follows a Uniform(0,1) distribution if gene $j$ is EE and a distribution stochastically smaller than uniform if gene $j$ is DE. These are standard assumptions made so that an unbiased size $\alpha$ test is obtained by rejecting $H_j$ if and only if $p_j \leq \alpha$. Table 4.1 defines the random variables corresponding to frequencies associated with all four outcomes when testing $m$ hypotheses.
Benjamini and Hochberg (1995) defined false discovery rate as
\[ E(\operatorname{max}\{V, 1\}) \]
where \( V \) is the number of EE genes declared to be DE (or the number of Type I errors) and \( R \) is the total number of genes declared to be DE. Many methods have been proposed for estimating FDR, but the \( q \)-value method (Storey, 2002) is likely the most commonly used approach for gene expression experiments.

The formal definition of the \( q \)-value is given as
\[
q(j) = \min \left\{ \frac{p(r) \hat{m}_0}{r} : r = j, \ldots, m \right\},
\]
where \( q(j) \) is the \( q \)-value corresponding to \( p(j) \), the \( j \)th smallest \( p \)-value, and \( \hat{m}_0 \) is the estimated number of EE genes among all \( m \) genes in the experiment. Specifically, \( q_j \) represents the expected false discovery rate if we declare gene \( j \) to be DE along with all other genes with \( q \)-values smaller than \( q_j \).

Many approaches have been proposed for estimating \( m_0 \) by estimating density of the \( p \)-values at \( p_j = 1 \) and multiplying this by \( m \). Storey (2002), Storey et al. (2004), Nettleton et al. (2006), and Liang and Nettleton (2012) have proposed approaches for doing this by developing methods for selecting a \( \lambda \in (0, 1) \) and estimating \( m_0 \) as
\[
\hat{m}_0(\lambda) = \frac{\sum_{j=1}^{m} 1\{p_j > \lambda\}}{(1 - \lambda)}.
\]

Storey and Tibshirani (2003) developed an alternative method for estimating \( m_0 \) by first calculating \( \hat{m}_0(\lambda) \) for a series of \( \lambda \) values between 0 and 1 using (4.2). Then the relationship between \( \lambda \) and \( \hat{m}_0(\lambda) \) is estimated by fitting a natural cubic spline through the points \((\lambda, \hat{m}_0(\lambda))\). Finally, \( m_0 \) is estimated by evaluating this function at \( \lambda = 1 \). This approach will be used to estimate \( m_0 \) in the simulation studies and real data analysis in Sections 4.3 and 4.4.
4.2.2 False Discovery Rate Estimation Using Two Subsets of p-values

Consider the problem of identifying genes that are DE in an experiment. To do this, a test statistic \( t_j \) and corresponding two-sided p-value \( p_j \) is obtained for each gene \( j = 1, \ldots, m \) by testing the null hypothesis

\[
H_j : \mu_{j1} = \mu_{j2}
\]  

(4.3)

against a two-sided alternative, where \( \mu_{jt} \) is the population treatment mean expression for gene \( j \) in treatment \( t \) for \( i = 1, 2 \).

Our proposed method begins by estimating \( m_0 \) for the entire set of \( m \) p-values from a gene expression experiment. This can be done using any of the methods cited in Section 4.2.1. Then the p-values are divided into two subsets based on the sign of the corresponding test statistics. Let \( \{ p_k^{(1)} : k = 1, \ldots, m_1 \} \) represent the subset of p-values corresponding to genes with negative test statistics, and let \( \{ p_k^{(2)} : l = 1, \ldots, m_2 \} \) represent the remaining p-values, which correspond to genes with positive test statistics. Then the q-value method is applied separately to each subset of p-values. Therefore, for each gene \( k \) in the first subset, the q-value is

\[
q_{(1)}^{(k)} = \min \left\{ \frac{p_{(r)}^{(1)} \hat{m}_0 / 2}{r} : r = k, \ldots, m_1 \right\}.
\]  

(4.4)

Similarly for the second subset, the q-value for gene \( k \) is calculated as

\[
q_{(2)}^{(k)} = \min \left\{ \frac{p_{(r)}^{(2)} \hat{m}_0 / 2}{r} : r = k, \ldots, m_2 \right\},
\]  

(4.5)

where \( p_{(r)}^{(i)} \) is the \( r^{th} \) smallest p-value in the \( i^{th} \) subset.

The estimates of FDR in (4.4) and (4.5) are based on the assumption that the asymmetry in the test statistics is due to the asymmetry in the effect sizes (i.e. \( \mu_{j1} - \mu_{j2} \)) for DE genes, and the expectation that there are an equal number, \( m_0 / 2 \), of positive and negative test statistics corresponding to EE genes. This expectation follows from the uniformity assumption of the distribution of null p-values and is the reason for using \( \hat{m}_0 / 2 \) in (4.4) and (4.5). The numerator in each equation is an estimate of the number of EE genes with p-values less than or equal to \( p_{(r)}^{(i)} \) among genes whose test statistics have sign \( (-1)^{t} \). Thus, \( q_{(k)}^{(i)} \) is a natural expression for the q-value associated with \( p_{(k)}^{(i)} \).
4.2.3 Advantages of the Proposed Method

Sun and Cai (2007) showed that multiple testing methods that rank the significance of hypothesis tests solely on the resulting \( p \)-values, such as the traditional \( q \)-value method, are often inefficient in terms of minimizing the “false nondiscovery rate” (i.e., the expected proportion of DE genes declared to be EE). In many cases, additional information can be used to improve this ranking. When the distribution of effects sizes is asymmetric in an experiment, dividing the set of \( p \)-values from a gene expression experiment into two subsets based on the sign of the test statistics and calculating \( q \)-values separately for each subset, as described in Section 4.2.2, improves efficiency. Figure 4.2 helps illustrate this idea. The histogram on the left plots the two-sided \( p \)-values for genes that have negative test statistics from the microarray experiment in the mesophyll cells of maize leaves described in Covshoff et al. (2008). The histogram on the right plots the two-sided \( p \)-values for genes with positive test statistics from the same experiment. A horizontal dashed line is plotted at the estimated proportion of EE genes, \( \hat{\pi}_0^{(i)} \), for each subset, and represents the estimated density for a \( p \)-value from an EE gene in subset \( i \). This estimate is calculated as

\[
\hat{\pi}_0^{(i)} = \frac{\hat{m}_0/2}{m_i}
\]  

for \( i = 1, 2 \), and again is based on the expectation of an equal number of positive and negative test statistics from EE genes.

For each plot, the area of the left-most bar in each histogram represents the proportion of \( p \)-values that are less than 0.05, and the proportion of the area of this bar below the dashed line represents the estimated proportion of EE genes among genes with test statistics of the appropriate sign and \( p \)-values less than 0.05. Because the area of this bar is larger for the histogram that corresponds to negative test statistics than the one that corresponds to positive test statistics, and the area of the bar below the dashed line is relatively smaller, the estimated proportion of EE genes among genes with test statistics of the appropriate sign and \( p \)-values less than 0.05 will be lower for the first histogram than the second. Thus, a gene with a \( p \)-value close to 0.05 will have a smaller \( q \)-value if this gene has a negative test statistic than if it has
Figure 4.2: A histogram of \( p \)-values for the maize data corresponding to (a) negative \( t \)-test statistics and (b) positive \( t \)-test statistics. The estimated proportion of EE genes, \( \hat{\pi}_0(i), i = 1, 2 \), is plotted as the dashed horizontal line in each plot.

a positive test statistic. More generally, a gene with a small \( p \)-value will be more likely to be declared DE if it corresponds to negative test statistic than if it corresponds to a positive test statistic, and it is possible for a gene with a higher \( p \)-value that corresponds to a negative test statistic to be ranked more significant (i.e. have a lower \( q \)-value) than a gene with a smaller \( p \)-value that corresponds to a positive test statistic. This reasoning agrees with equations (4.4) and (4.5), as the denominator in equation (4.4) will be larger than the denominator in equation (4.5) for the same two-side \( p \)-value. Thus, two genes that have the same \( p \)-value but different signs of their corresponding test statistics will have different \( q \)-values, and the gene with the negative test statistic will have the lower \( q \)-value. We show via simulation in the next section that this strategy results in a better significance ranking of genes.
4.3 Simulation Studies

In order to evaluate the performance of the proposed method and compare it to that of the traditional q-value method, we performed two sets of simulation studies. For each simulated data set, each gene \( j \) of the \( m = 10,000 \) total genes was tested for differential expression by testing \( H_j \), given in (4.3), against a two-sided alternative. For each test, a corresponding test statistic, \( t_j \), and \( p \)-value, \( p_j \), were computed using the moderated \( t \)-test proposed by Smyth (2004). This method was developed specifically for analyzing data from microarray experiments and borrows information across all genes in order to more accurately estimate the error variances for each individual gene. The resulting set of \( p \)-values were then analyzed by the traditional \( q \)-value method and the proposed method as described in Section 4.2.2. Note that both methods use the same estimate of \( \pi_0 \), and this was done using the natural cubic spline approach of Storey and Tibshirani (2003), briefly described in Section 4.2.1.

Two sets of simulations studies were performed. The first simulation study involved generating data sets with independent normally distributed data in order to evaluate the methods with data that are consistent with assumptions used to derive the moderated \( t \)-test. The second set of simulation studies used real gene expression data in order to evaluate the methods under conditions that are not ideal, but are generally observed in gene expression data, namely data with a complex correlation structure that can not be adequately modeled using the available data.

For each simulation study, three variables were changed in order to evaluate the methods under different situations. Sample sizes of \( n = 4 \), \( n = 10 \), and \( n = 20 \) for each gene in each treatment were used. These sample sizes were chosen because gene expression experiments tend to have small sample sizes; about 90% of gene expression data sets on the Gene Expression Omnibus (Edgar et al., 2002) have a total sample size (i.e. the number of subjects across all treatments for each gene) no larger than 40. Thus, a maximum total sample size of \( 2n = 40 \) was chosen for the simulations performed. The number of EE genes was also varied from \( m_0 = 5000 \) to \( m_0 = 9000 \) out of \( m = 10,000 \) total genes. Finally, four levels of an asymmetry parameter \( (\pi_A) \) were considered to examine performance as the distribution of effect sizes ranged from
symmetric about zero to highly asymmetric about zero. This asymmetry parameter is explained in more detail in the Section 4.3.1, but its impact on the effect size distribution can be see in Figure 4.3, which plots the density for each value of $\pi_A$ from which effect sizes, relative to the standard deviations of the genes, were randomly drawn for DE genes.

For each simulation setting in each simulation study, 500 data sets were randomly generated and analyzed.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.3.png}
\caption{The densities of effect sizes for DE genes for different values of $\pi_A$.}
\end{figure}

### 4.3.1 Simulations Using Independent Normal Data

In the first simulation study, data were randomly generated from independent normal distributions. Data for each gene $j = 1, \ldots, 10000$ for each data set were generated as follows. The variance $\sigma_j^2$ was randomly selected from an inverse gamma distribution. The parameters of the inverse gamma distribution were calculated from the data set of an experiment described in Hannenhalli et al. (2006) using methods proposed by Smyth (2004). If gene $j$ was EE, then $\mu_{j1} = \mu_{j2} = 0$. If gene $j$ was DE, then $\mu_{j1} = 0$ and the effect size, $\mu_{j2} = \delta_j$ was randomly drawn from the mixture distribution

\begin{equation}
    h_j(\delta) = \pi_A \phi\left(\delta; \sigma_j, \sigma_j^2\right) + (1 - \pi_A) \phi\left(\delta; -\sigma_j, \sigma_j^2\right),
\end{equation}
where $\phi(\delta; \gamma, \tau^2)$ is the normal density with mean $\gamma$ and variance $\tau^2$ evaluated at $\delta$. Finally, $n$ values were randomly drawn from a Normal($\mu_{jt}, \sigma_j^2$) distribution for each treatment $t = 1, 2$.

### 4.3.2 Simulations Using Gene Expression Data

The second simulation study used microarray data from an experiment in which gene expressions were measured from the heart tissue of $N = 108$ human subjects suffering from idiopathic dilated cardiomyopathy. This experiment is described in detail in Hannenhalli et al. (2006). Although this data set contains data from over 20,000 genes, $m = 10,000$ genes were randomly selected for analysis in the simulation study. The expression values for each data set were generated as follows. For each gene $j = 1, \ldots, m$, the variance $s_j^2$ was calculated from all $N$ subjects. Then, data from $2n$ subjects were randomly drawn, without replacement, from the microarray data set. From this subset of data, the data from $n$ subjects were randomly chosen to be in the first treatment, and the data from the remaining $n$ subjects were assigned to the second treatment. Note that, at this point, because the data from both treatments were randomly drawn from the same population, the population treatment means were equal (i.e. $\mu_{j1} = \mu_{j2}$) for each gene. If gene $j$ was EE, the data from this gene was not altered in any way. If gene $j$ was DE, then the effect size, $\delta_j$, was randomly chosen from mixture model (4.7), but by replacing $\sigma_j$ with $s_j$. Then, $\delta_j$ was added to the data from gene $j$ in the second treatment. Note that this method of data generation did not change the correlation structure of the data in any way but only shifted the mean of the data between the two treatments for DE genes.

### 4.3.3 Results

Tables 4.2 and 4.3 present the results of the simulation studies. For each setting in each simulation study, the mean (based on the 500 simulated data sets) partial area under the receiver operator characteristic (ROC) curve (pAUC) is given with its corresponding standard error in parentheses. These are partial areas because we considered only the most relevant region of the ROC curve where the false positive rate, or FPR, $\leq 0.10$. The method that ranks the genes better with regard to differential expression will have a higher pAUC. For each simulation setting, a traditional paired $t$-test was performed to test for a difference in
the mean pAUCs of the proposed method and the traditional \( q \)-value method. If this test was significant at 5%, then the higher mean pAUC is presented in bold font in Tables 4.2 and 4.3. In addition, mean \( S \), the mean number of DE genes that were declared to be DE is also given for each setting in order to observe the number of correctly identified DE genes, on average. Similar to the pAUCs, \( t \)-tests were used to determine if one method produced a larger mean \( S \). Also, to verify that each method adequately controls FDR, the empirical approximation of FDR was calculated for each method while controlling FDR at 5%. We will call this quantity \( \overline{V/R} \), and for a single data set this is defined as the proportion of EE genes among all genes with \( \hat{FDR} \leq 0.05 \) or 0 if \( \hat{FDR} \leq 0.05 \) for no genes. Refer to Table 4.1 for definitions of the quantities \( S \), \( V \), and \( R \).

The most important observation to be made from the results of the simulation studies is that the proposed method outperformed the traditional \( q \)-value method in terms of both mean pAUC and mean \( S \) for every simulation setting with asymmetric effect sizes, i.e., for settings with \( \pi_A > 0.5 \). Also, in settings with symmetric effect sizes (i.e. \( \pi_A = 0.5 \)), the proposed method is only outperformed by the traditional \( q \)-value method with respect to pAUC, and is never outperformed with respect to mean \( S \). These results apply to both simulation studies.

Additionally, in simulation settings with symmetric effect sizes, although one method might be determined to be better based on the paired \( t \)-test, the results of the two methods tend to be very similar when comparing the actual mean values. For example, refer to the first two columns in Table 4.3 which summarizes the results of the simulation study involving microarray data with respect to mean pAUC. For the setting with \( n = 20 \), \( m_0 = 5000 \), and \( \pi_A = 0.5 \), the traditional method is determined to be significantly better than the proposed method by a paired \( t \)-test even though, when rounded to one decimal, the mean pAUCs are the same for both methods.

Also note that both methods control FDR adequately at 5%. The \( \overline{V/R} \) values tend to be smaller than 5%, indicating a conservative control of FDR, and although some \( \overline{V/R} \) values are larger than 5%, this is expected by chance.

The results of the simulation studies suggest that the proposed method for estimating FDR can be used in place of the traditional \( q \)-value method in any situation without risk of worsening
<table>
<thead>
<tr>
<th>n</th>
<th>m₀</th>
<th>πₐ</th>
<th>Mean pAUC (%)</th>
<th>V/R (%) for ( \hat{FDR} \leq 0.05 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Proposed</td>
<td>Proposed</td>
</tr>
<tr>
<td>4</td>
<td>9000</td>
<td>1.0</td>
<td>39.4 (0.06)</td>
<td>38.7 (0.6)</td>
</tr>
<tr>
<td>0.9</td>
<td>39.6 (0.07)</td>
<td>31.1 (0.5)</td>
<td>20.1 (0.5)</td>
<td>15.9 (0.4)</td>
</tr>
<tr>
<td>0.7</td>
<td>34.4 (0.07)</td>
<td>33.6 (0.06)</td>
<td>36.9 (0.07)</td>
<td>33.5 (0.07)</td>
</tr>
<tr>
<td>0.5</td>
<td>33.6 (0.06)</td>
<td>33.6 (0.06)</td>
<td>38.8 (0.7)</td>
<td>35.9 (0.6)</td>
</tr>
<tr>
<td>0.9</td>
<td>36.8 (0.04)</td>
<td>34.9 (0.6)</td>
<td>279 (1.4)</td>
<td>227.9 (1.4)</td>
</tr>
<tr>
<td>0.7</td>
<td>34.2 (0.04)</td>
<td>33.6 (0.07)</td>
<td>227.7 (1.5)</td>
<td>226.4 (1.5)</td>
</tr>
<tr>
<td>0.5</td>
<td>33.5 (0.04)</td>
<td>33.5 (0.04)</td>
<td>227.7 (1.5)</td>
<td>226.4 (1.5)</td>
</tr>
<tr>
<td>7000</td>
<td>1.0</td>
<td>16.5 (0.03)</td>
<td>14.5 (0.04)</td>
<td>14.3 (0.4)</td>
</tr>
<tr>
<td>0.9</td>
<td>14.3 (0.4)</td>
<td>3.65 (0.15)</td>
<td>5.76 (0.41)</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>14.3 (0.4)</td>
<td>4.47 (0.25)</td>
<td>6.35 (0.61)</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>14.3 (0.4)</td>
<td>5.67 (0.39)</td>
<td>5.90 (0.57)</td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>1.0</td>
<td>11.9 (0.02)</td>
<td>10.7 (0.02)</td>
<td>10.7 (0.03)</td>
</tr>
<tr>
<td>0.9</td>
<td>10.7 (0.03)</td>
<td>3.65 (0.12)</td>
<td>2.99 (0.41)</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>10.7 (0.03)</td>
<td>4.13 (0.06)</td>
<td>4.32 (0.06)</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>10.7 (0.03)</td>
<td>4.89 (0.06)</td>
<td>4.88 (0.05)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9000</td>
<td>1.0</td>
<td>57.4 (0.07)</td>
<td>54.9 (0.07)</td>
</tr>
<tr>
<td>0.9</td>
<td>56.4 (0.07)</td>
<td>33.7 (0.8)</td>
<td>312.4 (0.8)</td>
<td>316 (0.8)</td>
</tr>
<tr>
<td>0.7</td>
<td>55.2 (0.07)</td>
<td>33.6 (0.07)</td>
<td>316 (0.8)</td>
<td>316 (0.8)</td>
</tr>
<tr>
<td>0.5</td>
<td>54.8 (0.07)</td>
<td>316 (0.8)</td>
<td>316 (0.8)</td>
<td>316 (0.8)</td>
</tr>
<tr>
<td>7000</td>
<td>1.0</td>
<td>22.5 (0.03)</td>
<td>21.0 (0.03)</td>
<td>21.0 (0.03)</td>
</tr>
<tr>
<td>0.9</td>
<td>21.0 (0.03)</td>
<td>359.0 (0.8)</td>
<td>359.0 (0.8)</td>
<td>359.0 (0.8)</td>
</tr>
<tr>
<td>0.7</td>
<td>21.0 (0.03)</td>
<td>1326.9 (1.5)</td>
<td>1326.9 (1.5)</td>
<td>1326.9 (1.5)</td>
</tr>
<tr>
<td>0.5</td>
<td>21.0 (0.03)</td>
<td>1326.9 (1.5)</td>
<td>1326.9 (1.5)</td>
<td>1326.9 (1.5)</td>
</tr>
<tr>
<td>5000</td>
<td>1.0</td>
<td>15.5 (0.02)</td>
<td>14.9 (0.02)</td>
<td>14.9 (0.02)</td>
</tr>
<tr>
<td>0.9</td>
<td>14.9 (0.02)</td>
<td>362.0 (0.9)</td>
<td>362.0 (0.9)</td>
<td>362.0 (0.9)</td>
</tr>
<tr>
<td>0.7</td>
<td>14.9 (0.02)</td>
<td>2554.2 (2.1)</td>
<td>2554.2 (2.1)</td>
<td>2554.2 (2.1)</td>
</tr>
<tr>
<td>0.5</td>
<td>14.9 (0.02)</td>
<td>2554.2 (2.1)</td>
<td>2554.2 (2.1)</td>
<td>2554.2 (2.1)</td>
</tr>
<tr>
<td>20</td>
<td>9000</td>
<td>1.0</td>
<td>68.8 (0.06)</td>
<td>67.9 (0.06)</td>
</tr>
<tr>
<td>0.9</td>
<td>68.0 (0.06)</td>
<td>533.6 (0.8)</td>
<td>533.6 (0.8)</td>
<td>533.6 (0.8)</td>
</tr>
<tr>
<td>0.7</td>
<td>67.9 (0.06)</td>
<td>533.6 (0.8)</td>
<td>533.6 (0.8)</td>
<td>533.6 (0.8)</td>
</tr>
<tr>
<td>0.5</td>
<td>67.9 (0.06)</td>
<td>533.6 (0.8)</td>
<td>533.6 (0.8)</td>
<td>533.6 (0.8)</td>
</tr>
<tr>
<td>7000</td>
<td>1.0</td>
<td>26.3 (0.03)</td>
<td>26.0 (0.03)</td>
<td>26.0 (0.03)</td>
</tr>
<tr>
<td>0.9</td>
<td>26.0 (0.03)</td>
<td>556.0 (0.8)</td>
<td>556.0 (0.8)</td>
<td>556.0 (0.8)</td>
</tr>
<tr>
<td>0.7</td>
<td>26.0 (0.03)</td>
<td>1856.1 (1.4)</td>
<td>1856.1 (1.4)</td>
<td>1856.1 (1.4)</td>
</tr>
<tr>
<td>0.5</td>
<td>26.0 (0.03)</td>
<td>1856.1 (1.4)</td>
<td>1856.1 (1.4)</td>
<td>1856.1 (1.4)</td>
</tr>
<tr>
<td>5000</td>
<td>1.0</td>
<td>17.8 (0.02)</td>
<td>17.6 (0.02)</td>
<td>17.6 (0.02)</td>
</tr>
<tr>
<td>0.9</td>
<td>17.6 (0.02)</td>
<td>563.0 (0.8)</td>
<td>563.0 (0.8)</td>
<td>563.0 (0.8)</td>
</tr>
<tr>
<td>0.7</td>
<td>17.6 (0.02)</td>
<td>3338.2 (1.8)</td>
<td>3338.2 (1.8)</td>
<td>3338.2 (1.8)</td>
</tr>
<tr>
<td>0.5</td>
<td>17.6 (0.02)</td>
<td>3338.2 (1.8)</td>
<td>3338.2 (1.8)</td>
<td>3338.2 (1.8)</td>
</tr>
</tbody>
</table>

Table 4.2: The mean pAUC, mean S, and V/R with corresponding standard errors in parentheses for the proposed and traditional q-value methods for each setting in the simulation study using independent normal data. For each setting, for both the mean pAUC and mean S, the highest values are given in bold font if a t-test has determined that there is a difference in the means of the proposed and traditional methods at 5% significance. If the t-test was not significant, bold font is not used.
Table 4.3: The mean pAUC, mean $S$, and $V/R$ with corresponding standard errors in parentheses for the proposed and traditional $q$-value methods for each setting in the simulation study using microarray data. For each setting, for both the mean pAUC and mean $S$, the highest values are given in bold font if a $t$-test has determined that there is a difference in the means of the proposed and traditional methods at 5% significance. If the $t$-test was not significant, bold font is not used.
the ranking of genes with respect to differential expression, or declaring a smaller number of truly DE genes to be DE. Moreover, we highly recommend using the proposed method over the traditional \( q \)-value method when asymmetry is observed in the test statistics, such as in Figure 4.1, or when there is a clear difference in the distribution of \( p \)-values corresponding to negative test statistics and \( p \)-values corresponding to positive test statistics, as observed in Figure 4.2.

### 4.4 Real Data Analysis

In this section, we analyze data from a study described in Covshoff et al. (2008) using both the proposed and traditional \( q \)-value methods. In this study, expressions from \( m = 7377 \) genes in the mesophyll cells of maize leaves were compared between two genotypes, wild-type and mutant, using \( n = 6 \) two-color microarray slides. Mutant plants lacked the PSII activity of wild-type plants, and researchers were interested in identifying genes that have different mean expressions due to this lack of activity.

The test statistics from this experiment are shown in the histogram in Figure 4.1. There is a clear asymmetry in this histogram as there are more negative test statistics than positive test statistics, suggesting asymmetry in the effect sizes. More specifically, there are \( m_1 = 4141 \) genes with negative test statistics and \( m_2 = 3236 \) genes corresponding to positive test statistics. The asymmetry becomes more evident in Figure 4.2, where \( p \)-values corresponding to negative test statistics follow a distribution that is stochastically smaller than the distribution of \( p \)-values corresponding to positive test statistics.

Using Storey and Tibshirani’s (2003) natural cubic spline method, the estimated number of EE genes in this experiment is \( \hat{m}_0 = 2907.11 \). Because we expect there to be the same number of negative and positive test statistics among the EE genes, we estimate that there are \( \hat{m}_0/2 = 1453.56 \) EE genes with negative test statistics and \( \hat{m}_0/2 = 1453.56 \) EE genes with positive test statistics. Therefore, we can use equation (4.4) to estimate the FDR for each gene \( k = 1, \ldots, 4141 \) with a negative test statistic as

\[
q^{(1)}_{(k)} = \min \left\{ \frac{P^{(1)}(1453.56)}{r} : r = k, \ldots, 4141 \right\}.
\]

Similarly, we can use equation (4.5) to estimate FDR for each gene \( k = 1, \ldots, 3236 \) with a
positive test statistic as

\[ q^{(2)}_{(k)} = \min \left\{ \frac{p^{(2)}_{(r)}}{r} : r = k, \ldots, 3236 \right\}. \tag{4.9} \]

Figure 4.4 plots the test statistics versus the log ratio of the \( q \)-values from the proposed and traditional method for the gene expression experiment in maize leaves. Negative log ratios correspond to cases where the \( q \)-value from the proposed method was less than the \( q \)-value from the traditional method. Positive log ratios correspond to larger \( q \)-values for the proposed method. From this scatterplot we can clearly see that the proposed method produces smaller \( q \)-values than the traditional method for genes with negative test statistics and larger \( q \)-values for genes with positive test statistics, with a clear separation when the test statistic is 0. This is not surprising based on Figure 4.2 and the discussion in Section 4.2.3.

Figure 4.4: Scatterplot of the test statistics versus the log ratios of the \( q \)-values for the experiment in maize leaves. Negative log ratios correspond to cases where the \( q \)-value produced by the proposed method is less than the \( q \)-value from traditional method. Positive log ratios correspond to larger \( q \)-values for the proposed method.

Using the traditional \( q \)-value method, 2480 genes were declared to be DE when controlling FDR at 5%. The proposed method declared 2446 genes DE. There were 2260 genes that both methods declared to be DE. All 186 of the genes that were only declared DE by the proposed method had negative test statistics. Similarly, all of the 220 genes that were only declared DE
by the traditional method had positive test statistics. This is what we would expect based on
the higher number of negative test statistics than positive test statistics in this experiment.

Although the traditional method resulted in slightly more genes declared to be DE than the
proposed method, we believe that the genes declared to be DE by the proposed method result
in a more reliable set of genes. This belief is based on the results of the simulation studies in
Section 4.3. First of all, using the proposed method never resulted in a worse ranking of the
genes, on average, compared to the traditional method for settings with asymmetric effect sizes.
Additionally, the number of truly DE genes declared to be DE was never lower, on average,
when using the proposed method than when using the traditional method, and in many cases
was much higher. Thus, although the proposed method declared fewer genes to be DE than the
traditional method, we expect the proposed method to identify at least as many truly DE genes
as the traditional method. This also implies that there are fewer truly EE genes declared to
be DE using the proposed method than the traditional method, and thus a smaller proportion
of false discoveries and a more reliable set of genes on which to base biological conclusions.

4.5 Discussion

The proposed method for estimating FDR by first estimating $\pi_0$ using all $p$-values and then
analyzing two subsets of $p$-values separately based on the sign of the test statistics has clear
advantages over the traditional $q$-value method, especially when effect sizes are asymmetric. In
simulation settings with asymmetric effect sizes, the proposed method was never outperformed
by the traditional method in the two simulation studies, and generally ranks genes better
with respect to differential expression while adequately controlling false discovery rate. The
proposed method also declared at least as many, if not more, truly DE genes to be DE as the
traditional method, on average. Also, in simulation settings with symmetric effect sizes, the
proposed method performed similarly to the traditional method.

Future research might include generalizing the proposed method by partitioning the $p$-values
based on information other than the signs of the test statistics. For example, consider the
ANOVA model in which we wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ for each gene, where $\mu_t$
represents the population mean expression level at time point $t$. After performing these tests, we might
observe smaller $p$-values for genes with $\bar{y}_1 < \bar{y}_2 < \bar{y}_3$ (call this scenario 1) and $\bar{y}_1 > \bar{y}_2 > \bar{y}_3$ (call this scenario 2), where $\bar{y}_t$ is the sample mean expression level for time point $t$. Thus, we might partition the genes to include a subset of $p$-values corresponding to genes with scenarios 1 or 2 and a subset of $p$-values corresponding to all other genes. We might also consider creating a partition of three subsets including one with $p$-values corresponding to genes with scenario 1, one with $p$-values corresponding to genes with scenario 2, and one with $p$-values corresponding to all other genes. Then, after estimating $m_0$ using the entire set of $p$-values, we may compute $q$-values using a method analogous to that described in this paper. Such a method would be expected to perform better than the traditional approach, especially when genes truly DE across times tend to have mean expression levels that change monotonically.

**Bibliography**


CHAPTER 5. GENERAL CONCLUSIONS

5.1 Summary

This dissertation presented novel methods for extending traditional gene expression analysis in a single experiment to gene expression analysis in two independent experiments (Chapters 2 and 3) as well as a method that improves upon Storey’s $q$-value method for identifying DE genes while controlling FDR in a single experiment when the distribution of effect sizes is asymmetric (Chapter 4).

In Chapter 2, we extended the problem of estimating, $m_0$, or the number of genes that are EE in a single experiment to the more complicated problem of estimating, $m_{11}$, or the number of genes that are DE in both of two independent experiments. This method involves analyzing the $p$-values paired by gene simultaneously, and shows much improvement over the common approach of the “intersection method” in simulation studies. Chapter 3 extended the ideas of Chapter 2 and proposed a method for identifying genes that are differentially expressed in both of two independent experiments while controlling FDR. The performance of this method was compared to the performance of two existing methods through two simulation studies. The results of these simulation studies demonstrated an advantage of our proposed method over the other two methods in that it adequately and consistently controlled FDR. The proposed method also resulted in sufficient power when compared to the other methods.

Chapter 4 focused on FDR analysis in one experiment by proposing a method that improves upon Storey’s $q$-value method when the distribution of effect sizes is asymmetric. This was done by estimating $m_0$ using the entire set of $p$-values from an experiment, then splitting the $p$-values into two subsets based on the signs of their corresponding test statistics, and calculating $q$-values individually for each subset. The proposed method shows better performance over the
traditional \( q \)-value method in terms of the significance ranking of genes as well as the number of true discoveries, while still adequately controlling FDR.

### 5.2 Future Research

Few methods have been developed to analyze the data from two independent gene expression experiments, and more research of this topic is encouraged. Although our estimator of \( m_{11} \) shows substantial improvement over that of the intersection method in terms of mean square error, it still tends have a large conservative bias when the average power for detecting DE genes is low. Thus, the development of other methods for estimating \( m_{11} \) that result in smaller biases would be very useful, as this value can be important biologically. Improvement of the estimation of \( m_{11} \) as well as the other quantities in Table 2.1 could also be a key step in the development of methods that adequately control FDR in the two experiment case.

Chapters 2 and 3 focused on estimating the number of and identifying genes that are differentially expressed in both of two independent experiments. These methods could be extended by incorporating the direction of change into the analysis. For example, we might be interested in estimating the quantities in Table 5.1. This table is an extension of Table 2.1, in which the estimate of the number of DE genes is divided into four separate estimates corresponding to each possible two-way combination of DE status (DE and down-regulated denoted by “(−)” and DE and up-regulated denoted by “(+))”. This could be done using the lfdr method described in detail in Chapter 3, but other methods should be explored.

<table>
<thead>
<tr>
<th>Test 1</th>
<th>Gene DE (−)</th>
<th>Gene EE</th>
<th>Gene DE (+)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>( m_{−−} )</td>
<td>( m_{0} )</td>
<td>( m_{−+} )</td>
<td>( m_{1} )</td>
</tr>
<tr>
<td>Gene EE</td>
<td>( m_{0−} )</td>
<td>( m_{00} )</td>
<td>( m_{0+} )</td>
<td>( m_{01} )</td>
</tr>
<tr>
<td>Gene DE (+)</td>
<td>( m_{+−} )</td>
<td>( m_{+0} )</td>
<td>( m_{++} )</td>
<td>( m_{1−} )</td>
</tr>
<tr>
<td>Total</td>
<td>( m_{2−} )</td>
<td>( m_{20} )</td>
<td>( m_{2+} )</td>
<td>( m )</td>
</tr>
</tbody>
</table>

Table 5.1: Contingency table of frequencies based on cross classification of the expression status when taking direction of change into account.

Furthermore, the methods of Chapter 4 could be incorporated into methods involving \( q \)-values discussed in Chapters 2 and 3. For example, suppose we are interested in identifying
genes that are DE in both of two experiments using either the $q^{\text{max}}$ method of Section 3.4.1 or the $q^I$ method of Section 3.4.2. If we observe asymmetry in the test statistics of either experiment 1 or experiment 2, then we might replace traditional $q$-values with $q$-values computed using the methods of Section 4.2 for these experiments.

Finally, the methods described in Chapters 2 and 3 could be extended to more than two experiments. This would be relatively straightforward for estimating a multivariate extension of $m_{11}$, but could be complicated for actually identifying genes as significance measures that accurately estimate FDR might be difficult to derive.