CLASSIFYING DISCOVERIES: IMPLEMENTING A GENERALIZED MULTIPLE TESTING PROTOCOL FOR EXPLORATORY DATA ANALYSIS

By

DAVID D. WATTS

Bachelor of Arts in Mathematics
University of Central Arkansas
Conway, Arkansas
2006

Master of Science in Applied Mathematics
University of Central Arkansas
Conway, Arkansas
2008

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2016

CLASSIFYING DISCOVERIES: IMPLEMENTING A GENERALIZED MULTIPLE TESTING PROTOCOL FOR EXPLORATORY DATA ANALYSIS

Dissertation Approved:

Dr. Joshua D. Habiger

Dissertation Advisor

Dr. Carla L. Goad

Dr. Lan Zhu

Dr. Michael P. Anderson

Name: DAVID D. WATTS

Date of Degree: JULY, 2016

Title of Study: CLASSIFYING DISCOVERIES: IMPLEMENTING A GENERALIZED MULTIPLE TESTING PROTOCOL FOR EXPLORATORY DATA ANALYSIS

Major Field: STATISTICS

Abstract: In large-scale exploratory data analysis one objective is to discover interesting attributes that are worthy of further study. Standard statistical analysis employs a multiple testing procedure which aims to discover as many attributes as possible subject to the constraint that an error rate, such as the false discovery rate (FDR), is controlled at a prespecified level. However, the objective of this statistical protocol need not be in line with the objectives of the study at hand since discovered attributes need not be interesting (worthy of further study), and likewise, interesting attributes need not be discovered. This work provides a new statistical method that allows for the nature of the follow-up analysis to be considered when determining which attributes are discovered. The methodology is illustrated on a dataset in which the objective is to discover bacterial species near the roots of wheat plants that are associated with plant health and to classify discovered species into groups based on the nature (positive or negative) and degree (strong or weak) of their association. This definition of interesting leads to a procedure that ranks attributes according to their local misclassification rates (LMCR). Theoretical and numerical results illustrate that the proposed LMCR procedure outperforms the current standard procedure in that it has a smaller misclassification rate among discoveries and still controls the FDR. The new method also performs favorably over the traditional approach when applied to real-world datasets, including the aforementioned plant health data, where expectation-maximization (EM) algorithms are used to estimate unknown parameters.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

High throughput technology, such as sequencing and imaging technology, is now routinely used to generate so called “high dimensional” (HD) data sets. The first stage in the analysis of HD data is often to identify which among thousands of variables or attributes are worthy of more exploration using some multiple testing procedure. For example, in genome-wide association studies the goal is to determine which among thousands of single nucleotide polymorphisms are associated with a trait (Roeder and Wasserman, 2009) while in fMRI analysis the objective is to identify which voxels or regions of the brain are associated with a stimulus (Lindquist, 2008). In this HD setting, a false discovery rate (FDR) (Benjamini and Hochberg, 1995; Farcomeni, 2008; Dudoit and van der Laan, 2008) controlling multiple testing procedure is typically employed. The FDR is the expected proportion of false discoveries among discoveries, or from a multiple hypothesis testing perspective, is the expected proportion of type I errors among rejected null hypotheses. The FDR is formally defined below and will be used throughout this dissertation. But first we consider an example of a multiple testing problem.

1.1 A Simple Multiple Testing Example

A common situation requiring large scale multiple testing is that of microarray data. A microarray is used in biomedical settings to measure gene expressions simultaneously. Such studies may include thousands, tens of thousands, or even millions of genes.

For example, Efron (2010) discussed an analysis of prostate cancer data consisting of $M = 6033$ genes obtained from 50 healthy men and 52 men with prostate cancer. As with most microarray analyses, the goal was to determine which genes showed statistically different expression levels between the two groups. This information could then be used to guide future study.

Formally, for each $m = 1, 2, ..., 6033$, let x_{mj} be the gene expression level for gene m and patient j. Further, let group 1 be the control (healthy) group with population mean μ_m^1 and corresponding sample mean $\bar{x}_m^1$, and let group 2 be the cancer group with population mean μ_m^2 and corresponding sample mean $\bar{x}_m^2$. A standard two sample t-test can be applied to each gene, with null hypothesis

$$H_m : \mu_m^1 = \mu_m^2.$$

The test statistic is now

$$t_m = \frac{\bar{x}_m^2 - \bar{x}_m^1}{s_m},$$

where

$$s_m = \sqrt{\frac{\sum_{j=1}^{50}(x_{mj} - \bar{x}_m^1)^2 + \sum_{j=51}^{102}(x_{mj} - \bar{x}_m^2)^2}{50 + 52 - 2}\left(\frac{1}{50} + \frac{1}{52}\right)},$$

and has a Student's t distribution with 100 degrees of freedom under H_m. For convenience, Efron then converts each t statistic into a corresponding z statistic using

$$Z_m = \Phi^{-1}(F_{100}(t_m)),$$

where Φ and F_{100} are the cdfs for the standard normal distribution and t distribution with 100 degrees of freedom, respectively.

Each null hypothesis is that Z_m has the standard normal distribution ($Z \sim N(0, 1)$). For more details on the creation of the z-values, as well as a brief discussion on the selection of an alternative hypothesis, consult Efron (2010).

If we control the type I error rate at level $\alpha = 0.05$ by rejecting H_m when $|Z_m| \geq 1.96$, we find 478 statistically significant genes. However, we can reasonably expect that many of these rejected nulls will be incorrectly rejected. These incorrect rejections will be type I errors, also known as "false discoveries."

In multiple hypothesis testing, there are two main forms of error control: control of the familywise error rate (FWER) and control of the false discovery rate (FDR). The notion of the FDR has its roots in Simes (1986) but was popularized by Benjamini and Hochberg (BH) in 1995 as an alternative to the FWER.

To define these error rates, first suppose that we have M hypotheses of interest. The state of each null hypothesis is unknown (it can be either true or false), so the researcher tests each one and claims it as either statistically significant or nonsignificant, which leads to four possibilities: true claims of significance and nonsignificance, as well as, false claims for each. Table 1.1 shows this relation, using notation similar to that presented in Storey et al. (2004). Similar tables can be found in Benjamini and Hochberg (1995), Cai and Sun (2009), Efron (2010), and elsewhere throughout FDR literature.

	Claimed Nonsignificant	Claimed Significant	Total
Null	U	V	M_0
Nonnull	T	S	$M - M_0$
Total	W	R	M

Table 1.1: Possibilities for a tested hypothesis.

The FDR and FWER can be defined as

$$\text{FDR} = E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0) \text{ and FWER} = Pr(V \geq 1).$$

While the FDR is the expected proportion of false discoveries among all discoveries, the FWER is the probability of committing at least one type I error. A good introduction to the FDR can be found in Efron (2010), and a good overview of multiple hypothesis testing methodology in general can be found in Farcomeni (2008).

For the prostate cancer data discussed above, we can control the FWER by applying the Bonferroni procedure (formally defined in Section 2.1), which leads to 6 rejected nulls. FDR control can be achieved by applying the BH procedure (outlined in Subsection 2.3.2) to the p-values associated with the given z-values. Doing so, we now have 21 rejected null hypotheses.

1.2 A Major Deficiency in Standard Methods

Standard FDR methods are aimed at maximizing the expected number of rejections or discoveries subject to the constraint of FDR control. Recent work has shown that, while this approach is reasonable in most exploratory analyses, quite often it may provide misleading inference. For example, Sun and McLain (2012) sought to identify schools in California where student performance was associated with socio-economic status, but standard FDR methods tended to identify the largest schools rather than those schools whose performance was strongly associated with socio-economic status. In Habiger et al. (2015) the goal was to identify species of bacteria living near the roots of wheat plants that are associated with plant health or productivity, but standard procedures tended to identify the most abundant species rather than those with the strongest association. As pointed out by Ruppert et al. (2007), the problem with standard approaches is that a discovery need not be "interesting" scientifically.

In this dissertation, we demonstrate that these procedures fail because they are designed to maximize the expected number of discoveries (interesting or not) or minimize some type II error rate, when this objective may not be appropriate for the study at hand. For example, in Anderson and Habiger (2012), one objective was to discover species of bacteria that are correlated with productivity and to classify bacteria into groups based on the nature (positive or negative) and degree (strong or weak) of their association. The standard, or naive approach, which would first apply an FDR procedure to discover associated species and then classify the discovered species, fails for two reasons. First, discovered species can be classified into the "null" group, i.e. the classification procedure identifies them as not being associated with productivity. Second, classification error among the discovered/classified species can be high. The problem here is that the multiple testing procedure failed to consider the manner in which discovered species would be explored in follow up analysis. For more details on this two-stage protocol see Sun and Wei (2015) and references therein.

1.3 Proposed Research

This work focuses on the first stage of the analysis. Specifically, we provide an FDR method that allows for the manner in which the discovered data are to be explored to be incorporated into the initial discovery process. The idea is to allow the attributes to be ranked from most significant to least significant using arbitrary statistics, which could include but are not limited to p-value statistics, posterior probabilities, or what we call the "local misclassification rate." Then a threshold is defined along the rankings for providing FDR control.

Relevant FDR research will be more fully developed in the literature review of Chapter 2. Chapter 3 provides the general framework and illustrates how the rankings are related to the optimality criteria for two common procedures, while Chapter 4 introduces a local misclassification rate procedure aimed at accomplishing the objectives of Anderson and Habiger (2012), outlined above. Chapter 5 directly compares the new procedure with the naive two stage approach in practice. Chapter 6 considers the new procedure under various test scenarios with known parameters, whereas Chapter 7 examines the performance of the new procedure when parameters are estimated. Chapter 8 provides a brief summary of this work.

Chapter 2

Literature Review

Multiple hypothesis testing has a rich history throughout the literature, one that can be traced back to the late 1950s (Dunn, 1958, 1959), but has roots that stretch back even farther. Since then, many approaches have been developed to try to control the error inflation inherent to the simultaneous testing of hypotheses. Part of the proposed research will look to combine a FDR controlling methodology with a classification based group structure. Some of the work most relevant to this goal will be explored below.

2.1 Decision Functions

Let $\theta_m \in \{0, 1\}$, $m \in \mathcal{M} = \{1, 2, ..., M\}$ index the state of null hypothesis H_m so that $\theta_m = I(H_m \text{ false})$ where $I(\cdot)$ is the indicator function. For short, denote the collection of null hypotheses by $\boldsymbol{H} = (H_m,\ m \in \mathcal{M})$, and the state of $\boldsymbol{H}$ by $\boldsymbol{\theta} = (\theta_m,\ m \in \mathcal{M})$. Consider testing H_m with test statistic T_m where $\boldsymbol{T} = (T_m,\ m \in \mathcal{M})$ is the collection of test statistics.

The decision to reject or retain H_m using T_m is denoted by $\delta_m = I(T_m \leq t_m)$ where t_m represents a threshold which is to be specified by the multiple testing procedure of interest. Here, $\delta_m = 1$ when hypothesis m is rejected and 0 otherwise. For example, if the procedure is based on p-values, for $m \in \mathcal{M}$ we may define $t_m = \alpha$ such that $\delta_m = I(P_m \leq \alpha)$. The well known Bonferroni procedure can be based on p-values and uses threshold $t_m = \alpha/M$ for each m. It is formally defined as

$$\delta_m = I(P_m \leq \alpha/M).$$

The collection of decision functions, also called a multiple decision function (MDF), is denoted $\boldsymbol{\delta} = (\delta_m,\ m \in \mathcal{M})$. The basic multiple testing strategy is to define δ_m so as to control an error rate.

2.2 Error Rates

For a single hypothesis test, there are two main types of error considered: type I errors and type II errors. Type I errors occur when a true null hypothesis is incorrectly deemed nonnull, whereas a type II error occurs when a nonnull hypothesis is incorrectly deemed null. In the context of Table 1.1, the single hypothesis case corresponds to $M = 1$. That is, either M_0 or $M - M_0$ equals 1, with the other quantity necessarily being 0. Here, we can define the standard notions of type I and type II error rates as $Pr(V = 1|M_0 = 1)$ and $Pr(T = 1|M - M_0 = 1)$, respectively. Using decision function notation, these two error rates can be written as $Pr(\delta_m = 1 \mid \theta_m = 0)$ and $Pr(\delta_m = 0 \mid \theta_m = 1)$, respectively.

To define multiple testing error rates, let $V = \sum_{m \in \mathcal{M}} \delta_m[1 - \theta_m]$ denote the number of type I errors (false discoveries) and let $R = \sum_{m \in \mathcal{M}} \delta_m$ denote the number of rejected null hypotheses (discoveries). These quantities correspond to the V and R values in Table 1.1, respectively. Further, let $\mathcal{R} = \{m : \delta_m = 1\}$ be the set of indices corresponding to rejected hypotheses (for convenience, referred to as a rejection set), such that the cardinality of $\mathcal{R}$ is $|\mathcal{R}|= R$.

The familywise error rate (FWER) is now

$$\text{FWER} = \Pr(V \geq 1),$$

and the false discovery proportion is

$$\text{FDP} = \frac{V}{R \vee 1},$$

where $R \vee 1$ denotes the maximum of R and 1. We force the denominator to be positive so as to avoid dividing by 0 since there can be no false discoveries when there are no discoveries. The false discovery rate (FDR) is then defined as $\text{FDR} = E[\text{FDP}]$, where $E[\cdot]$ is the expectation operator.

Note that an alternative form for the FDR is

$$\text{FDR} = E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0) \tag{2.1}$$

where this equality holds because of iterated expectation. Here, we see

$$\begin{aligned} E\left[\frac{V}{R \vee 1}\right] &= E\left[\frac{V}{R \vee 1}\middle| R > 0\right] Pr(R > 0) + E\left[\frac{V}{R \vee 1}\middle| R = 0\right] Pr(R = 0) \\ &= E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0) \end{aligned}$$

since $\frac{V}{R \vee 1} = 0$ when $R = 0$ since $V \leq R$.

2.3 Classic Multiple Testing Procedures

Perhaps the simplest method for controlling the FWER at significance level α in multiple testing is the Bonferroni approach, as defined in Section 2.1. This method ensures that the FWER $\leq \alpha$, as is desired, but when $M = 10,000$ and $\alpha = 0.05$, the threshold for rejection is 0.000005, which is quite restrictive. Thus, there have been many improvements and alternatives to the Bonferroni procedure, including those by Šidák (1967), Holm (1979), Hommel (1988), as well as several others.

2.3.1 The Simes Procedure

With the goal of eventually implementing FDR control, the most relevant methodology developed for FWER control is the Simes (1986) procedure. Simes suggested a modification to the Bonferroni approach based on the ordered p-values $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(M)}$. He defined the global null hypothesis as $H : \boldsymbol{\theta} = \mathbf{0}$, i.e. it assumes that all null hypotheses are true. The Simes procedure is not focused on testing individual null hypotheses, but instead is focused on the rejection of H. It rejects H if

$$P_{(i)} \leq \frac{i\alpha}{M} \text{ for any } i = 1, 2, ..., M.$$

Simes then suggested that decisions could be made for the individual hypotheses $H_{(1)}, H_{(2)}, ..., H_{(M)}$, where $H_{(m)}$ corresponds to $P_{(m)}$, by rejecting those hypotheses $H_{(1)}, H_{(2)}, ..., H_{(j)}$ where

$$j = \max\left\{m : P_{(m)} \leq \frac{m\alpha}{M}\right\},$$

if the global null had already been rejected. He made this suggestion as an exploratory approach only and believed that subsequent studies would need to be performed to provide confirmation.

It must be noted that the Simes procedure only controls FWER weakly at a given significance level, as opposed to strong control. Strong control occurs when FWER $\leq \alpha$, regardless of which (or how many) null hypotheses are true. Weak control, on the other hand, occurs when FWER $\leq \alpha$ only when all null hypotheses are true.

2.3.2 The BH Procedure

In their 1995 article *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*, Benjamini and Hochberg (BH) note that classical FWER control procedures have several practical difficulties. These include: concerns over distributional assumptions, the fact that most analyses yielded much lower power for multiple testing when compared to per comparison procedures, and the general belief that control of FWER may not truly be needed in all circumstances.

As a result, BH focused on the FDR and identified two important properties associated with this error rate:

1. When all null hypotheses are true, FDR control implies weak FWER control.

2. When $M_0 < M$, FDR $\leq$ FWER. Thus, in such cases, control of only the FDR could lead to a gain in power because it is less restrictive.

To provide some intuition on why FDR control is less strict than FWER control, suppose a researcher has $10,000$ hypotheses to test, of which $1,000$ are rejected. Even one incorrect rejection is unacceptable for FWER controlling procedures. However, by focusing on the number of rejections, the FDP is still only 0.05 even if 50 discoveries are false ($50/1000 = 0.05$).

To formally define the BH procedure, reconsider the situation where we are testing $\boldsymbol{H} =$

$(H_m,\ m \in \mathcal{M})$ using the ordered p-values $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(M)}$, and denote $H_{(m)}$ as the hypothesis corresponding to $P_{(m)}$. The BH procedure can be defined (see Storey et al. (2004)) by $\delta_m = I(P_m \leq \alpha j/M)$, where j is defined as

$$j = \begin{cases} 0 & \text{if } P_{(m)} \geq m\alpha/M\ \forall\ m \in \mathcal{M} \\ \max\left\{m : P_{(m)} \leq m\alpha/M\right\} & \text{otherwise} \end{cases}.$$

Note that the first case of $j = 0$ corresponds to the situation when there will be no rejections and the second case leads to j rejections. BH showed that FDR $\leq \alpha M_0/M$ if p-values corresponding to true null hypotheses are independent of one another and independent of p-values from false null hypotheses.

In the years since BH popularized their methodology for controlling the FDR, much work has been done expanding and generalizing the uses of FDR controlling procedures. For instance, adaptive procedures often operate by incorporating an estimate of $\pi_0 = M_0/M$ into the BH procedure. As an example, consider the adaptive procedure in Storey (2002) and Storey et al. (2004) which uses $\alpha/\hat{\pi}_0$ in place of α in the BH procedure, where $\hat{\pi}_0$ is some estimate of π_0. Storey showed that this procedure controls the FDR at α if all p-values are independent and if $\hat{\pi}_0$ is appropriately defined.

Other examples of FDR research include: an empirical Bayes interpretation, as well as the estimation of an empirical null distribution (Efron et al., 2001; Efron, 2004, 2008b, 2010); the definition of a positive false discovery rate (pFDR), as well as exploration of optimality and other concerns (Storey, 2002, 2003, 2007; Storey et al., 2004); the exploration of weighted p-value schemes (Benjamini and Hochberg, 1997; Genovese et al., 2006; Roeder et al., 2006); adaptive procedures for FDR control (Benjamini and Hochberg, 2000; Genovese and Wasserman, 2004; Blanchard and Roquain, 2009); the estimation of the proportion of null and nonnull hypotheses (Efron, 2004; Nettleton et al., 2006; Jin and Cai, 2007); a link between p-value based procedures and decision theoretic approaches (Habiger, 2012); and a compound p-value approach (Habiger and Peña, 2014). There has also been significant work regarding FDR control and dependence (Benjamini and Yekutieli, 2001; Farcomeni, 2007; Finner et al., 2007; Sarkar, 2008; Wu, 2008; Sun and Cai, 2009).

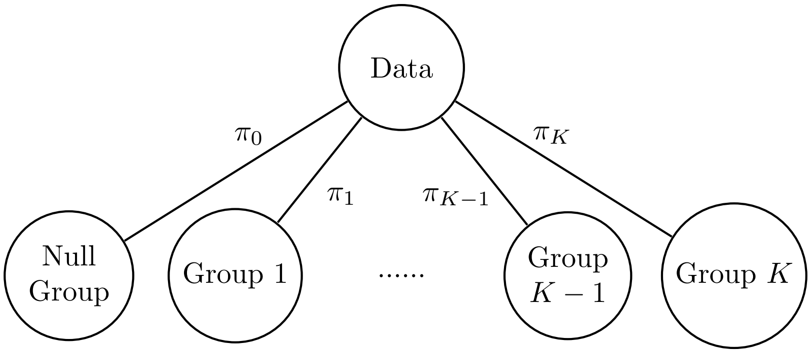

Figure 2.1: Illustration of the multiple group model defined by Model 1.

2.4 Classification Into One of Several Groups

As mentioned above, part of the proposed research will look to combine the notion of FDR control for groups with a data driven procedure for defining those groups. Classification and clustering analysis has been a big topic of research in many disciplines (see Xu and Wunsch (2009), Duda et al. (2001), or Hastie et al. (2009) for examples in computer science and data mining, among others) and this work will use both types of analysis to examine the performance of the procedure developed in Chapter 4.

When the desired group structure is known, classification is often based on a mixture model (see, for example, Scott and Symons (1971), Anderson (1984), or Fraley and Raftery (2002)). In the following mixture model, we let G_m take on values in $\mathcal{K} = \{0, 1, ..., K\}$ to emphasize that $G_m = 0$ indicates that H_m is true (null). If $G_m \neq 0$ then H_m is false (nonnull). Figure 2.1 illustrates this model which will be used throughout the remainder of this manuscript.

Model 1. *Let $(X_m, G_m) \in \mathcal{X} \times \{0, 1, ..., K\}$ be independent and identically distributed random vectors. Assume X_m has mixture*

$$f(x) = \sum_{k \in \mathcal{K}} \pi_k f_k(x) \tag{2.2}$$

where $f_k(x)$ is the probability mass function (pmf) or probability density function (pdf) of x given $G_m = k$, $\pi_k = Pr(G_m = k)$, and $\mathcal{K} = \{0, 1, ..., K\}$.

Classification analysis calculates an estimate for G_m, denoted $\widehat{G}_m$, using data X_m, or classifies

X_m as belonging to group k if $\widehat{G}_m = k$. Anderson (1984) demonstrates that if the cost of incorrectly classifying X_m into group k given that $G_m = j \neq k$ is the same across all $j,\ k$ then the admissible classification rule is defined as

$$\widehat{G}_m = \widehat{G}_m(x_m) = \underset{k \in \mathcal{K}}{\operatorname{argmax}} \{\pi_k f_k(x_m)\}. \tag{2.3}$$

Note that the event $[\widehat{G}_m = k]$ can equivalently be written as $[x_m \in A_k]$ for some $A_k \subset \mathcal{X}$ where $\boldsymbol{A} = \{A_0,\ A_1,\ ...,\ A_K\}$ is a partition of $\mathcal{X}$.

2.5 Cluster Analysis

Formally, classification requires a known number of groups or classes into which a new observation may be placed. That is, classification essentially applies a label to a datum. On the other hand, cluster analysis looks to identify groups of observations that are similar within each group and yet distinct from observations within other groups. In the context of machine learning, classification is a common supervised learning methodology, whereas cluster analysis is an example of unsupervised learning.

In this work, classification analysis will be used throughout Chapter 6 to illustrate the properties of the proposed procedure in simulations where the data structure and corresponding parameters are known. Cluster analysis will then be used in Chapter 7 to consider the performance of the proposed procedure when dealing with data that has unknown parameter values.

2.5.1 General Clustering Techniques

Cluster analysis (and general pattern recognition as a whole) has a rich and varied history in many disciplines, including biology, psychiatry, psychology, archeology, geology, geography, and marketing, as mentioned in Jain et al. (1999). Specifically, there have been many books published on the topic, including those of Anderberg (1973), Hartigan (1975), Jain and Dubes (1988), Duda et al. (2001), and Xu and Wunsch (2009). For a statistical treatment of machine learning in general, including cluster analysis, consult Hastie et al. (2009). There are many hundreds (if not thousands) of techniques for identifying groups of elements based on some form of similarity. For the most part though, there are two main types of clustering, hierarchical and partitional.

Hierarchical clustering is an iterative methodology based on nests of clusters. The nested structure (often presented graphically as a dendrogram) can be defined in one of two main ways, either through agglomeration or division. Agglomerative clustering begins with each data point in its own cluster, and during each step of the process, two clusters are merged based on some criterion (often some notion of distance or variance). For an early yet popular example of agglomerative clustering, see Ward (1963). Divisive clustering (also known as top-down clustering) begins with all of the data contained in a single cluster. Each step of a divisive algorithm looks to divide each cluster into two smaller clusters, again based on some criterion (often distance based). For specific details on the types of criteria used to determine clusters, see Jain et al. (1999) or one of the books mentioned above.

While hierarchical clustering seeks a way of combining or dividing elements of a dataset, iteratively, partitional clustering looks to form the clusters simultaneously (thus forming a partition of the data), often based on some metric. There are a multitude of partitional clustering algorithms, but perhaps one of the most popular techniques is known as k-means clustering (MacQueen, 1967). In k-means clustering, k elements are randomly selected to serve as centers for each cluster, at which point, all other elements are assigned to the nearest cluster center. Once all elements have been assigned, new centers are calculated. This process is then repeated until some convergence criterion is satisfied (often based on a squared error measurement or based on whether or not cluster membership has changed from one iteration to the next). The next subsection will detail a second partitional clustering method that will be the focus of Chapter 7.

2.5.2 Model-Based Clustering

Model-based clustering often uses a finite mixture of probability models to determine clusters (see Bock (1996), Fraley and Raftery (1998), McLachlan and Peel (2000), or Fraley and Raftery (2002), among many others, for details). When performing model-based clustering, we must first establish an appropriate mixture model. For our application, Model 1 defines such a model using convenient notation. Specifically, Equation (2.2) defines a $K+1$ group mixture model where π_k is the mixture proportion for group k with corresponding density $f_k(x)$. In fact, the only thing that should be included notationally is a reference to the parameters within each density. So, let $\boldsymbol{\eta} = (\eta_k,\ k \in \mathcal{K})$ be a collection of parameters where η_k represents the parameters of density f_k. Equation (2.2) can then be written as

$$f(x|\boldsymbol{\eta}) = \sum_{k\in\mathcal{K}} \pi_k f_k(x|\eta_k). \tag{2.4}$$

In this context, given $\boldsymbol{X} = \boldsymbol{x}$, the likelihood of the mixture model is

$$L(\boldsymbol{\eta}; \boldsymbol{\pi} \mid \boldsymbol{x}) = \prod_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} \pi_k f_k(x_m|\eta_k), \tag{2.5}$$

where $\boldsymbol{\pi} = (\pi_k,\ k \in \mathcal{K})$.

To determine a good clustering solution, our objective is to maximize the likelihood of the mixture model in Equation (2.5) by finding estimates for $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$. Unfortunately, in this likelihood, each observation is composed of two pieces of information, the actual data value and the group membership value. To account for both pieces, let $z_m = (x_m, \boldsymbol{y}_m)$ represent the *complete data*, where x_m is observed and $\boldsymbol{y}_m = (y_{m0}, y_{m1}, ..., y_{mK})$ is the unobserved group membership. Here, for $k \in \mathcal{K}$, let $y_{mk} = I(z_m \text{ is from group } k)$.

For a mixture model, we assume that each $\boldsymbol{y}_m$ is a single iid realization from a $K+1$ group multinomial distribution with probabilities defined by $\boldsymbol{\pi}$. Given the group information, $\boldsymbol{y}_m$, the density of x_m is then $\prod_{k\in\mathcal{K}} f_k(x_m|\eta_k)^{y_{mk}}$, so that the complete data likelihood is

$$L(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{y}|\boldsymbol{z}) = \prod_{m\in\mathcal{M}} \prod_{k\in\mathcal{K}} [\pi_k f_k(x_m|\eta_k)]^{y_{mk}} \tag{2.6}$$

where $\boldsymbol{z} = (z_m,\ m \in \mathcal{M})$ and $\boldsymbol{y} = (\boldsymbol{y}_m,\ m \in \mathcal{M})$.

2.5.3 Parameter Estimation and the EM Algorithm

To estimate the parameters of Equation (2.6), we will use a maximum likelihood estimation (MLE) approach coupled with the expectation-maximization (EM) algorithm. MLEs are discussed at length in Cramér (1946), Bain and Engelhardt (2000), Lehmann and Casella (2003), and elsewhere. Among the many desirable properties of MLEs is the fact that, under suitable regularity conditions, MLEs are consistent estimators of the parameters they estimate. That is, the maximum likelihood estimate converges in probability to the value of the parameter of interest as the sample size increases. See Corollary 3.5 in Chapter 6 of Lehmann and Casella (2003) or section 33.3 of Cramér (1946) for details. When determining the MLEs of Equation (2.6), it is often easier to maximize the complete data log-likelihood

$$l(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{y}|\boldsymbol{z}) = \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} y_{mk}\log[\pi_k f_k(x_m|\eta_k)]. \tag{2.7}$$

The EM algorithm for incomplete data, proposed in Dempster et al. (1977) and discussed at length in McLachlan and Peel (2000) and McLachlan and Krishnan (2008), offers us a convenient way to determine the MLEs of $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ in Equation (2.7). In general, the EM algorithm alternates between the expectation of a likelihood function (the E-step) and the subsequent estimates of the parameters that maximize that expectation (the M-step).

More specifically, suppose that we have an initial estimate of group membership for each observation, say $\hat{\boldsymbol{y}}_m = (\hat{y}_{mk},\ k \in \mathcal{K})$, and we want estimates of η_k and π_k, say $\hat{\eta}_k^{(1)}$ and $\hat{\pi}_k^{(1)}$ to signify that this is the first iteration. The next iteration begins by finding the expected value of $\hat{\boldsymbol{y}}_m$ given $\hat{\eta}_k^{(1)}$ and $\hat{\pi}_k^{(1)}$. The updated values $\hat{\eta}_k^{(2)}$ and $\hat{\pi}_k^{(2)}$ are then calculated, and the process repeats until some convergence criterion is satisfied (often either a maximum number of iterations is reached or there is not a significant change in the likelihood value, within some tolerance).

Formally, the E-step is $\hat{y}_{mk} = E[y_{mk}|x_m, \eta_k, \pi_k]$, which is the conditional expectation that observation m is a member of group k given the data and parameters. This expectation is equivalent to the probability that observation m belongs to group k. In the context of Model 1, we have

$$\hat{y}_{mk} = Pr(G_m = k|X_m = x_m) = \frac{Pr(G_m = k, X_m = x_m)}{Pr(X_m = x_m)} = \frac{\hat{\pi}_k f_k(x_m|\hat{\eta}_k)}{\sum_{k\in\mathcal{K}} \hat{\pi}_k f_k(x_m|\hat{\eta}_k)}, \tag{2.8}$$

where $\hat{\eta}_k$ and $\hat{\pi}_k$ are estimates of the corresponding parameters. With a value of $\hat{y}_{mk}$ in hand, the M-step looks to find the MLEs by differentiating Equation (2.7) with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$. A specific derivation of the components of the EM algorithm for a univariate normal mixture model is presented in Section 7.1.

To facilitate much of the analysis in the chapters to come, `R` software (R Core Team, 2015) was used. Two useful `R` functions that implement the EM algorithm are the `Mclust` function from the **mclust** package (Fraley and Raftery, 2002; Fraley et al., 2012) and the `normalmixEM` function from the **mixtools** package (Benaglia et al., 2009). The dataset in Subsection 5.2.1 will be analyzed with the `Mclust` function whereas the simulations in Chapter 7 will utilize the `normalmixEM` function.

In general, the `Mclust` function is much faster than `normalmixEM` and includes an automatic selection of the best variance structure for the model. Thus, some of the parameter estimates may seem unusual (for example, a non-standard normal null group with equal variance across all five groups, as seen in Table 5.3) but they are in fact associated with the best model (for a given number of groups) as determined by the Bayesian information criterion (BIC). The main advantage of using the `normalmixEM` function is its flexibility in restricting the range of possible values on a parameter by parameter basis. This is especially useful when dealing with unusual datasets or when implementing a parameter constraint.

Chapter 3

A Method for Defining Procedures for FDR Control

The analysis to follow is based on a random mixture model which assumes that the states of null hypotheses to be tested are random. This framework was perhaps first introduced within the context of the FDR in Efron et al. (2001); Genovese and Wasserman (2002); Storey (2003).

Let $\mathcal{M} = \{1, 2, ..., M\}$, and let $(\boldsymbol{X}, \boldsymbol{\theta}) = ((X_m, \theta_m),\ m \in \mathcal{M})$ be random vectors with support $\mathcal{X} \times \{0, 1\}^M$ and distribution $F \in \mathcal{F}$, where $\mathcal{F}$ is a model for F. Suppose that each θ_m is an unobservable Bernoulli random variable defined by $\theta_m = 1 - I(H_m \text{ true})$ where H_m is a null hypothesis and $I(\cdot)$ is the indicator function. Here, H_m is of the form $H_m : F \in \mathcal{F}_m \subseteq \mathcal{F}$ for $\mathcal{F}_m$ a submodel of $\mathcal{F}$, and the collection of null hypotheses is denoted $\boldsymbol{H} = (H_m,\ m \in \mathcal{M})$. Denote the decision to reject or retain H_m by $\delta_m : \mathcal{X} \mapsto \{0, 1\}$ or $\delta_m(\boldsymbol{X})$ or δ_m for short, where $\delta_m = 1(0)$ means H_m is rejected (retained). The multiple decision function (MDF) is $\delta : \mathcal{X} \mapsto \{0, 1\}^M$ and is denoted $\boldsymbol{\delta}(\boldsymbol{X}) = (\delta_m(\boldsymbol{X}),\ m \in \mathcal{M})$ or $\boldsymbol{\delta}$ for short.

To define decision rules (and subsequent rejection sets $\mathcal{R} = \{m : \delta_m = 1\}$), we adopt the approach formally outlined in Storey (2007) and Benjamini and Bogomolov (2014), where the basic strategy is to

1. rank the hypotheses from most significant to least significant using statistics $T_m = T_m(\boldsymbol{X})$, say $T_{(1)} \leq T_{(2)} \leq ... \leq T_{(M)}$, where $T_{(1)}$ is the most significant, $T_{(2)}$ the next most significant,

and so on. Then,

2. reject the R null hypotheses corresponding to $T_{(1)}, T_{(2)}, ..., T_{(R)}$.

Formally, we define $\delta_m = I(T_m \leq T_{(R)}),\ m \in \mathcal{M}$ and denote the corresponding rejection index set as

$$\mathcal{R}(\boldsymbol{T}, R) = \left\{m \in \mathcal{M} : T_m \leq T_{(R)}\right\}. \tag{3.1}$$

where we recall that $\boldsymbol{T}$ is the collection of test statistics and R is the cardinality of the set ($|\mathcal{R}|$).

Traditional FDR methods focus on choosing $\boldsymbol{T}$ so as to maximize R or minimize some type II error rate subject to the constraint that the FDR $\leq \alpha$. For example, the BH procedure of the previous chapter uses the ordered p-values $(P_{(1)}, P_{(2)}, ..., P_{(M)})$ to choose $R = \max_j \left\{P_{(j)} \leq \alpha j/M\right\}$ and define $\delta_m = I(P_m \leq P_{(R)})$. The corresponding rejection set is then $\mathcal{R}(\boldsymbol{P}, R) = \{m \in \mathcal{M} : P_m \leq P_{(R)}\}$ where $\boldsymbol{P} = (P_m,\ m \in \mathcal{M})$ is the collection of p-values.

Theorem 3.1 below provides a general route for choosing R such that the FDR $\leq \alpha$.

Theorem 3.1. *Let $\mathcal{R} = \{m : \delta_m = 1\} \subseteq \mathcal{M}$ index the collection of discoveries with $|\mathcal{R}|= R$ and denote posterior probabilities by $Q_m = Pr(\theta_m = 0|\boldsymbol{X} = \boldsymbol{x})$ for $m \in \mathcal{M}$. If*

$$\frac{1}{R}\sum_{m\in\mathcal{R}} \mathrm{Q}_m \leq \alpha \tag{3.2}$$

then FDR $\leq \alpha$*.*

Proof of Theorem 3.1. This proof is a generalized version of the proof for Theorem 1 in Cai and Sun (2009). First note that

$$\begin{aligned} E\left[V|\boldsymbol{X} = \boldsymbol{x}\right] = E\left[\sum_{m\in\mathcal{M}} \delta_m[1-\theta_m] \mid \boldsymbol{X} = \boldsymbol{x}\right] &= \sum_{m\in\mathcal{M}} \delta_m E\left[I(\theta_m = 0)|\boldsymbol{X} = \boldsymbol{x}\right] \\ = \sum_{m\in\mathcal{R}} Pr(\theta_m = 0|\boldsymbol{X} = \boldsymbol{x}) &= \sum_{m\in\mathcal{R}} \mathrm{Q}_m. \end{aligned} \tag{3.3}$$

The first and second equalities are due to the definition of V and the fact that δ_m is a function of $\boldsymbol{X}$ only. For the third equality, note that the set of hypotheses is reduced to $\mathcal{R}$ by δ_m and that the expectation of an indicator function is the probability of its argument. The fourth equality substitutes the definition of Q_m. Thus, we have

$$\begin{aligned} \text{FDR} = E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0) &= E\left[E\left[\frac{V}{R}\middle| R > 0, \boldsymbol{X} = \boldsymbol{x}\right]\right] Pr(R > 0) \\ &= E\left[\frac{1}{R}\, E[V|\boldsymbol{X} = \boldsymbol{x}]\right] Pr(R > 0) = E\left[\frac{1}{R} \sum_{m \in \mathcal{R}} \mathrm{Q}_m\right] Pr(R > 0) \\ &\leq \alpha\ Pr(R > 0) \leq \alpha. \end{aligned}$$

The second equality is a consequence of the law of iterated expectation. With $\boldsymbol{X}$ known, R is known, since R is a function of $\boldsymbol{X}$ through $\boldsymbol{\delta}$. Thus, the third equality holds and the fourth equality substitutes Equation (3.3). The first inequality is due to the supposition of Equation (3.2) and the final inequality is obvious since $Pr(R > 0) \leq 1$. □

The rest of this chapter illustrates how this theorem can be used to verify FDR control in a variety of settings considered in the literature and shows that proposed test statistics for rankings arise out of optimality criteria not originally considered. Chapter 4 develops a specific application of the two step strategy by introducing an alternative optimality criterion. The proposed approach will lead to a different ranking of hypotheses which will then form the basis of the rest of the current work.

3.1 Procedures that Rank the Local FDR

Consider a two group mixture model composed of a single null and a single nonnull group. This model, defined as Model 1 with $K = 1$, has previously been considered within the context of FDR control in Efron et al. (2001), Genovese and Wasserman (2002), Sun and Cai (2007), and Cai and Sun (2009).

When $K = 1$, Equation (2.2) becomes $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$, where we note that the nonnull component ($\pi_1 f_1(x)$) may itself be a mixture of nonnull ($G_m \neq 0$) distributions. Define the local false discovery rate (Lfdr), introduced by Efron et al. (2001), as

$$\mathrm{Lfdr}(x) = Pr(\theta_m = 0|X_m = x) = \frac{\pi_0 f_0(x)}{f(x)}. \tag{3.4}$$

Further, we define the Lfdr statistic for H_m as $\mathrm{Lfdr}_m = \mathrm{Lfdr}(X_m)$ and let $\mathbf{Lfdr} = (\mathrm{Lfdr}_m,\ m \in \mathcal{M})$ be the corresponding collection.

Now consider the Lfdr procedure of Sun and Cai (2007), which in our notation chooses

$$R^{SC} = \max_j \left\{ \frac{1}{j} \sum_{m \in \mathcal{R}^{SC}(\mathbf{Lfdr}, j)} \mathrm{Lfdr}_m \leq \alpha \right\} \tag{3.5}$$

with corresponding decision functions $\delta_m^{SC} = I\left(\mathrm{Lfdr}_m \leq \mathrm{Lfdr}_{(R^{SC})}\right)$ if $j > 0$ (that is, when there is at least one rejection) and defines $R^{SC} = 0$ when $j = 0$ (thus, $\mathcal{R}^{SC} = \emptyset$). Note that the Lfdr procedure ranks the hypotheses using the Lfdr statistic through the definition of $\mathcal{R}^{SC}(\mathbf{Lfdr}, j)$ (as in step 1 of the general protocol defined above). Also note that FDR control is now an immediate consequence of Theorem 3.1. In context, Corollary 3.1 states that FDR control can be achieved for any rejection set $\mathcal{R}$ if the average Lfdr value is controlled within that set.

Corollary 3.1. *The Lfdr procedure in Equation* (3.5) *has FDR $\leq \alpha$ under Model 1.*

Proof of Corollary 3.1. The proof follows from Theorem 3.1 by defining $\theta_m = I(G_m \neq 0)$ and observing that $Q_m = \mathrm{Lfdr}(x_m)$. The inequality in Equation (3.2) is satisfied by construction. □

The next proposition provides a simplified form for the expected number of false discoveries for any rejection set $\mathcal{R}$ given the data $\boldsymbol{X}$. This result is important because it will be used in the following theorem to associate an optimality criterion to the choice of statistic $\boldsymbol{T}$ for use in the two step process defined above.

Proposition 3.1. *For any rejection set $\mathcal{R}$ with cardinality $|\mathcal{R}| = R$ and corresponding decision functions $\boldsymbol{\delta}$, under Model 1,*

$$E\left[\sum_{m \in \mathcal{M}} \delta_m \theta_m | \boldsymbol{X} = \boldsymbol{x}\right] = R - \sum_{m \in \mathcal{R}} \mathrm{Lfdr}_m. \tag{3.6}$$

Proof of Proposition 3.1. First note that

$$\begin{aligned}
E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m \mid \boldsymbol{X}=\boldsymbol{x}\right] &= E\left[\sum_{m\in\mathcal{M}}\delta_m I(\theta_m\neq 0) \mid \boldsymbol{X}=\boldsymbol{x}\right] = \sum_{m\in\mathcal{R}} E\left[I(\theta_m\neq 0) \mid \boldsymbol{X}=\boldsymbol{x}\right] \\
&= \sum_{m\in\mathcal{R}} Pr(\theta_m\neq 0|\boldsymbol{X}=\boldsymbol{x}) = \sum_{m\in\mathcal{R}}[1-Pr(\theta_m=0|\boldsymbol{X}=\boldsymbol{x})] \\
&= \sum_{m\in\mathcal{R}}[1-Pr(\theta_m=0|X_m=x_m)] = R-\sum_{m\in\mathcal{R}}\text{Lfdr}_m.
\end{aligned}$$

The second equality comes from knowing δ_m since it is a function of $\boldsymbol{X}$ (thus, we are restricted to the rejection set $\mathcal{R}$). The third equality is an application of the definition of expectation and the fourth equality rewrites the equation in terms of the probability for the null group. The fifth equality is due to the independence of the model and the sixth equality simplifies the summation. □

Theorem 3.2. *Let $\mathcal{R}$ be any rejection set with corresponding decision functions $\boldsymbol{\delta}$ satisfying Equation* (3.2)*. Further, let $\mathcal{R}^{SC}(\textbf{Lfdr}, R^{SC})$ be the rejection set defined as in Equation* (3.5) *with corresponding decision functions $\boldsymbol{\delta}^{SC}$. Then, under Model 1,*

$$E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m\right] \leq E\left[\sum_{m\in\mathcal{M}}\delta_m^{SC}\theta_m\right]. \tag{3.7}$$

Proof of Theorem 3.2. We first note that, for any $\mathcal{R}$ and corresponding decision functions $\boldsymbol{\delta}$,

$$E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m\right] = E\left[E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m \mid \boldsymbol{X}=\boldsymbol{x}\right]\right] = E\left[|\mathcal{R}| - \sum_{m\in\mathcal{R}}\text{Lfdr}_m\right]$$

as a consequence of the law of iterated expectation and an application of Proposition 3.1.

Thus, it suffices to show that

$$\mathcal{R}^{SC}(\textbf{Lfdr}, R^{SC}) = \underset{\mathcal{R}:\sum_{m\in\mathcal{R}}\text{Lfdr}_m\leq\alpha|\mathcal{R}|}{\text{argmax}}\left\{|\mathcal{R}| - \sum_{m\in\mathcal{R}}\text{Lfdr}_m\right\}.$$

From Equation (3.1), for any $\mathcal{R}$ with $|\mathcal{R}|= r$, we establish the following:

$$\begin{aligned}
&\underset{\mathcal{R}:\sum_{m\in\mathcal{R}}\text{Lfdr}_m\le\alpha|\mathcal{R}|}{\text{argmax}}\left\{|\mathcal{R}|-\sum_{m\in\mathcal{R}}\text{Lfdr}_m\right\}\\
&\qquad\le\underset{\mathcal{R}(\mathbf{Lfdr},r):\sum_{m\in\mathcal{R}(\mathbf{Lfdr},r)}\text{Lfdr}_m\le\alpha|\mathcal{R}(\mathbf{Lfdr},r)|}{\text{argmax}}\left\{|\mathcal{R}(\mathbf{Lfdr},r)|-\sum_{m\in\mathcal{R}(\mathbf{Lfdr},r)}\text{Lfdr}_m\right\}\\
&\qquad=\mathcal{R}^{SC}(\mathbf{Lfdr},R^{SC}).
\end{aligned}$$

The inequality is satisfied because for any $\mathcal{R}$ with cardinality r we have

$$|\mathcal{R}|-\sum_{m\in\mathcal{R}}\text{Lfdr}_m = r-\sum_{m\in\mathcal{R}}\text{Lfdr}_m \le\; r-\sum_{j=1}^{r}\text{Lfdr}_{(j)} =\; r-\sum_{m\in\mathcal{R}(\mathbf{Lfdr},r)}\text{Lfdr}_m.$$

The equality holds because $|\mathcal{R}(\mathbf{Lfdr},r)| - \sum_{m\in\mathcal{R}(\mathbf{Lfdr},r)}\text{Lfdr}_m$ is nondecreasing in r and by the definition of $\mathcal{R}^{SC}(\mathbf{Lfdr},R^{SC})$. □

Theorem 3.2 shows that the Lfdr procedure maximizes the expected number of true positives subject to Equation (3.2). This extends the result of Sun and Cai (2007), where they showed that the Lfdr procedure minimizes the missed discovery rate or MDR under a monotone likelihood ratio (MLR) condition (Cao et al., 2013). The MDR is defined as the proportion of the expected number of retained nonnull hypotheses among the expected number of retained hypotheses, that is,

$$\text{MDR}=\frac{E\left[\sum_{m\in\mathcal{M}}[1-\delta_m]\theta_m\right]}{E\left[\sum_{m\in\mathcal{M}}[1-\delta_m]\right]}.$$

Note that here the MLR condition is not required.

3.2 Procedures that Rank the Conditional Lfdr

This section considers a situation where $\boldsymbol{X}$ is known to come from many heterogeneous groups. Such cases have previously been considered in the context of FDR control in Efron (2008b); Cai and Sun (2009); Hu et al. (2010). The multiple group model used in Cai and Sun (2009) is formally defined by Model 2 and is illustrated in Figure 3.1. Note that a superscript asterisk (*) is used for several

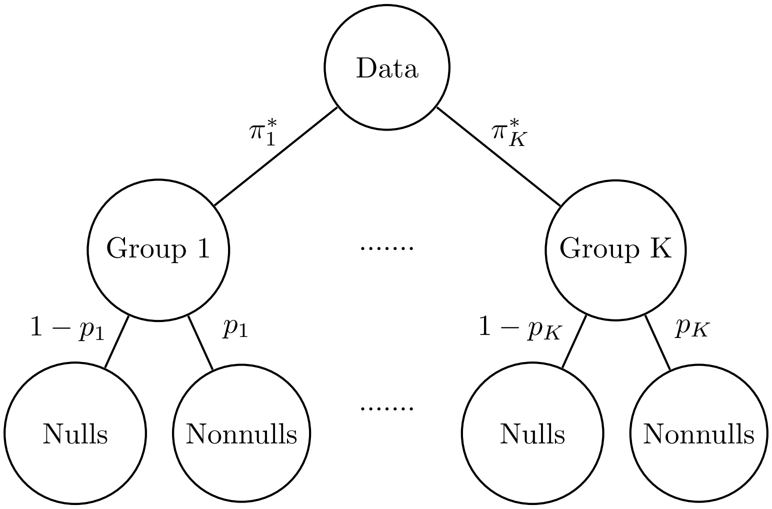

Figure 3.1: Illustration of the multiple group mixture model of Cai and Sun (2009), defined by Model 2.

variables to help distinguish them from similar variables used above.

Model 2. *Let $(X_m, G_m^*, \theta_m) \in \mathcal{X} \times \{1, 2, ..., K\} \times \{0, 1\}$ be independent and identically distributed random vectors. Define $\pi_m^* = Pr(G_m^* = k)$ and assume (X_m, G_m^*) has mixture $f(x, k) = (1 - p_k) f_{k0}(x) + p_k f_{k1}(x)$ where $p_k = Pr(\theta_m = 1 | G_m^* = k)$ and f_{k0} and f_{k1} are the null and nonnull pmfs or pdfs of x given $G_m^* = k$, respectively.*

In conditional analysis, we assume that $\boldsymbol{G}^* = (G_m^*,\ m \in \mathcal{M})$ is observable, with realization $\boldsymbol{g}^*$. Define the conditional Lfdr (CLfdr) by

$$\text{CLfdr}(x, k) = Pr(\theta_m = 0 | G_m^* = k, X_m = x) = \frac{(1 - p_k) f_{k0}(x)}{f(x, k)}. \tag{3.8}$$

We also define $\text{CLfdr}_m = \text{CLfdr}(X_m, G_m^*)$, so that $\textbf{CLfdr} = (\text{CLfdr}_m,\ m \in \mathcal{M})$ is a collection of CLfdr statistics. Here, the terminology “conditional” was introduced by Cai and Sun (2009) to emphasize the fact that we are conditioning on $G_m^* = k$, in addition to $X_m = x$. Defining posterior probabilities “conditionally” will be revisited in Subsection 5.2.3.

In a manner similar to that seen in the previous section, the CLfdr procedure of Cai and Sun

(2009) chooses

$$R^{CSC} = \max_j \left\{ \frac{1}{j} \sum_{m \in \mathcal{R}^{CSC}(\mathbf{CLfdr}, j)} \text{CLfdr}_m \leq \alpha \right\} \tag{3.9}$$

with corresponding decision functions $\delta_m^{CSC} = I\left(\text{CLfdr}_m \leq \text{CLfdr}_{(R^{CSC})}\right)$ if $j > 0$ and defines $R^{CSC} = 0$ when $j = 0$ (thus, $\mathcal{R}^{CSC} = \emptyset$). Again, we see that FDR control is an immediate consequence of Theorem 3.1. In context, we achieve FDR control for any rejection set if the average CLfdr value is controlled within that set.

Corollary 3.2. *The CLfdr procedure in Equation* (3.9) *has FDR* $\leq \alpha$ *under Model 2.*

Proof of Corollary 3.2. The proof follows from Theorem 3.1 by observing that $Q_m = \text{CLfdr}(x_m, g_m^*)$. Again, the inequality in Equation (3.2) is satisfied by construction. □

In a manner similar to that seen in Proposition 3.1, the following proposition will give a simplified form for the expected number of false discoveries for any rejection set $\mathcal{R}$ given the data $\boldsymbol{X}$ and the group information $\boldsymbol{G}^*$. Similarly, the following theorem illustrates that the CLfdr procedure maximizes the expected number of true positives subject to the constraint given by Equation (3.2). Again, this extends the result in their paper, which demonstrated that the CLfdr procedure minimized the overall MDR across all groups. As before, the MLR condition is not needed.

Proposition 3.2. *Under Model 2, for any rejection set* $\mathcal{R}$ *with cardinality* $|\mathcal{R}| = R$ *and corresponding decision functions* $\boldsymbol{\delta}$*,*

$$E\left[\sum_{m \in \mathcal{M}} \delta_m \theta_m | \ \boldsymbol{G}^* = \boldsymbol{g}^*, \boldsymbol{X} = \boldsymbol{x}\right] = R - \sum_{m \in \mathcal{R}} \text{CLfdr}_m. \tag{3.10}$$

Proof of Proposition 3.2. Similar to the process seen in the proof of Proposition 3.1, we have

$$
\begin{aligned}
&E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m |\ \boldsymbol{G}^*=\boldsymbol{g}^*,\boldsymbol{X}=\boldsymbol{x}\right]=E\left[\sum_{m\in\mathcal{M}}\delta_m I(\theta_m\neq 0)\mid \boldsymbol{G}^*=\boldsymbol{g}^*,\ \boldsymbol{X}=\boldsymbol{x}\right]\\
&=\sum_{m\in\mathcal{R}}E\left[I(\theta_m\neq 0)\mid \boldsymbol{G}^*=\boldsymbol{g}^*,\ \boldsymbol{X}=\boldsymbol{x}\right]=\sum_{m\in\mathcal{R}}Pr(\theta_m\neq 0\mid \boldsymbol{G}^*=\boldsymbol{g}^*,\ \boldsymbol{X}=\boldsymbol{x})\\
&=\sum_{m\in\mathcal{R}}[1-Pr(\theta_m=0\mid \boldsymbol{G}^*=\boldsymbol{g}^*,\ \boldsymbol{X}=\boldsymbol{x})]=\sum_{m\in\mathcal{R}}[1-Pr(\theta_m=0\mid G^*_m=g^*_m,\ X_m=x_m)]\\
&=R-\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m.
\end{aligned}
$$

The justification for each step is similar to that seen in the proof of Proposition 3.1. □

Theorem 3.3. *Let $\mathcal{R}$ be any rejection set with corresponding decision functions $\boldsymbol{\delta}$ satisfying Equation* (3.2)*. Further, let $\mathcal{R}^{CSC}(\mathbf{CLfdr}, R^{CSC})$ be the rejection set defined as in Equation* (3.9) *with corresponding decision functions $\boldsymbol{\delta}^{CSC}$. Then, under Model 2,*

$$
E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m\right]\leq E\left[\sum_{m\in\mathcal{M}}\delta^{CSC}_m\theta_m\right]. \tag{3.11}
$$

Proof of Theorem 3.3. As seen in the proof of Theorem 3.2, we note that, for any $\mathcal{R}$ and corresponding decision functions $\boldsymbol{\delta}$,

$$
E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m\right]=E\left[E\left[\sum_{m\in\mathcal{M}}\delta_m\theta_m\mid \boldsymbol{G}^*=\boldsymbol{g}^*,\ \boldsymbol{X}=\boldsymbol{x}\right]\right]=E\left[|\mathcal{R}|-\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m\right]
$$

by the law of iterated expectation and by applying the results of Proposition 3.2.

Hence it suffices to show that

$$
\mathcal{R}^{CSC}(\mathbf{CLfdr}, R^{CSC})=\underset{\mathcal{R}:\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m\leq\alpha|\mathcal{R}|}{\mathrm{argmax}}\left\{|\mathcal{R}|-\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m\right\}.
$$

From Equation (3.1), for any $\mathcal{R}$ with $|\mathcal{R}|= r$, we establish the following:

$$
\begin{aligned}
&\underset{\mathcal{R}:\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m\le\alpha|\mathcal{R}|}{\operatorname{argmax}}\left\{|\mathcal{R}|-\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m\right\}\\
&\qquad\le\underset{\mathcal{R}(\mathbf{CLfdr},r):\sum_{m\in\mathcal{R}(\mathbf{CLfdr},r)}\mathrm{CLfdr}_m\le\alpha|\mathcal{R}(\mathbf{CLfdr},r)|}{\operatorname{argmax}}\left\{|\mathcal{R}(\mathbf{CLfdr},r)|-\sum_{m\in\mathcal{R}(\mathbf{CLfdr},r)}\mathrm{CLfdr}_m\right\}\\
&\qquad=\mathcal{R}^{CSC}(\mathbf{CLfdr},R^{CSC}).
\end{aligned}
$$

Here, the inequality is satisfied since for any $\mathcal{R}$ with cardinality r we have

$$
|\mathcal{R}|-\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m=r-\sum_{m\in\mathcal{R}}\mathrm{CLfdr}_m\le\ r-\sum_{j=1}^{r}\mathrm{CLfdr}_{(j)}=r-\sum_{m\in\mathcal{R}(\mathbf{CLfdr},r)}\mathrm{CLfdr}_m.
$$

The equality holds because $|\mathcal{R}(\mathbf{CLfdr},r)|-\sum_{m\in\mathcal{R}(\mathbf{CLfdr},r)}\mathrm{CLfdr}_m$ is nondecreasing in r and by the definition of $\mathcal{R}^{CSC}(\mathbf{CLfdr},R^{CSC})$. □

Chapter 4

A New Approach

As mentioned in the previous chapter, this chapter will provide an application of the two step strategy for defining procedures that provide FDR control by considering Model 1 coupled with a new optimality criterion. Specifically, the tools of Chapter 3 will be used to define a procedure that accomplishes the objectives of Anderson and Habiger (2012) which recall are to

1. identify nonnull hypotheses, $G_m \neq 0$ under Model 1, subject to FDR control, and
2. classify nonnull hypotheses into one of several groups, $\widehat{G}_m = k \in \{1, 2, ..., K\}$.

The idea is to use Equation (3.2) to ensure that the proposed procedure provides FDR control, and results similar to Theorems 3.2 and 3.3 to ensure that the proposed procedure maximizes the expected number of correctly classified elements.

Before proceeding, recall that Model 1 defines a $K+1$ group mixture model composed of a single null group ($G_m = 0$) and K nonnull groups ($G_m = k \in \mathcal{K}^* = \{1, 2, ..., K\}$) with marginal density $f(x) = \sum_{k \in \mathcal{K}} \pi_k f_k(x)$ where $f_k(x)$ is the pmf or pdf of x given $G_m = k$, $\pi_k = Pr(G_m = k)$, and $\mathcal{K} = \{0, 1, ..., K\}$. The basic structure of Model 1 was shown in Figure 2.1.

4.1 The Local Misclassification Rate Procedure

The proposed procedure looks to maximize the number of *correctly classified* nonnull rejections. Recall from Section 2.4 that, under Model 1, X_m may be classified into group k using the classification rule $\widehat{G}_m = \text{argmax}_{k \in \mathcal{K}}\{\pi_k f_k(X_m)\}$ (or equivalently, $\widehat{G}_m = k$ when $X_m \in A_k$ where $\boldsymbol{A} = \{A_k,\ k \in \mathcal{K}\}$

is a partition of $\mathcal{X}$). We begin by considering the local correct classification rate (LCCR) for each $x \in A_k$, $k \in \mathcal{K}^*$, defined as

$$\text{LCCR}(x,k) = Pr(G_m = k \mid \widehat{G}_m = k, X_m = x) = \frac{\pi_k f_k(x)}{f(x)}. \tag{4.1}$$

Although $Pr(G_m = k|\widehat{G}_m = k, X_m = x) = Pr(G_m = k|X_m = x)$ when $x \in A_k$, we prefer the former to emphasize that any observed X_m can be classified. With the LCCR defined, we may now define the local misclassification rate (LMCR) as

$$\text{LMCR}(x,k) = Pr(G_m \neq k \mid \widehat{G}_m = k, X_m = x) = 1 - \text{LCCR}(x,k). \tag{4.2}$$

Define the LMCR statistic for H_m as $\text{LMCR}_m = \text{LMCR}(X_m, G_m)$. The LMCR procedure is implemented as follows. Given each $X_m = x_m$,

Step 1: calculate the collection of classifications $\widehat{\boldsymbol{G}} = (\widehat{G}_m,\ m \in \mathcal{M})$, the collection of LMCR statistics $\mathbf{LMCR} = (\text{LMCR}_m,\ m \in \mathcal{M})$, and the corresponding collection of posterior probabilities $\mathbf{Q} = (\text{Q}_m,\ m \in \mathcal{M})$, where $Q_m = Pr(\theta_m = 0|X_m = x_m)$.

Step 2: Define the rejection set $\mathcal{R}^{LMCR}(\mathbf{LMCR}, j) = \{m \in \mathcal{M}_1 : \text{LMCR}_m \leq \text{LMCR}_{(j)}\}$ where j is the cardinality of $\mathcal{R}^{LMCR}$ and $\mathcal{M}_1 = \{m : \widehat{G}_m \neq 0\}$. For $j > 0$, define the number of hypotheses to reject by

$$R^{LMCR} = \max_j \left\{ \frac{1}{j} \sum_{m \in \mathcal{R}^{LMCR}(\mathbf{LMCR}, j)} \text{Q}_m \leq \alpha \right\}, \tag{4.3}$$

and define decision functions $\delta_m^{LMCR} = I(\text{LMCR}_m \leq \text{LMCR}_{(R^{LMCR})})$ and rejection set $\mathcal{R}^{LMCR}(\mathbf{LMCR}, R^{LMCR}) = \{m \in \mathcal{M}_1 : \delta_m^{LMCR} = 1\}$. When $j = 0$, R^{LMCR} is defined to be 0 and $\mathcal{R}^{LMCR}(\mathbf{LMCR}, R^{LMCR}) = \emptyset$.

Step 3: For each $m \in \mathcal{R}^{LMCR}(\mathbf{LMCR}, R^{LMCR})$, report $\widehat{G}_m$.

Observe that for each $m \in \mathcal{R}(\mathbf{LMCR}, R^{\text{LMCR}})$, we have $\delta_m = 1$ and $\widehat{G}_m = k \neq 0$. That is, we have identified X_m as nonnull and classified it into group k. Note that the main difference between

Steps 1 and 2 above and the procedures in Sections 3.1 and 3.2 is that two sets of statistics were used in determining which attributes to discover. The LMCR statistics were used in Step 1 to rank hypotheses while the Q_m's were used in Step 2 to determine how many hypotheses could be rejected. The previous procedures used only Q_m's throughout.

4.2 Characteristics of the LMCR Procedure

This section will follow the basic structure of Sections 3.1 and 3.2 in showing that the LMCR procedure maximizes the expected number of correctly classified nonnull discoveries (CCND), defined as $\text{CCND} = E\left[\sum_{m\in\mathcal{M}} \delta_m I(G_m = \widehat{G}_m \neq 0)\right]$, subject to FDR control. We first verify that it controls the FDR.

Corollary 4.1. *The LMCR procedure has FDR $\leq \alpha$ under Model 1.*

Proof of Corollary 4.1. The proof follows from Theorem 3.1 by defining $\theta_m = I(G_m \neq 0)$. Again, the inequality in Equation (3.2) is satisfied by construction. □

The next proposition provides a simplified form for the expected number of false discoveries for any rejection set $\mathcal{R}$, given the data, when accounting for classification. The following theorem illustrates that the LMCR procedure maximizes the CCND, subject to the constraint of Equation (3.2).

Proposition 4.1. *Under Model 1, for any rejection set $\mathcal{R}$ with cardinality $|\mathcal{R}| = R$ and corresponding decision functions $\boldsymbol{\delta}$,*

$$E\left[\sum_{m\in\mathcal{M}} \delta_m I(G_m = \widehat{G}_m \neq 0) \mid \boldsymbol{X} = \boldsymbol{x}\right] = R - \sum_{m\in\mathcal{R}} \text{LMCR}_m. \tag{4.4}$$

Proof of Proposition 4.1. This proof follows logic similar to that seen in the proofs of Propositions 3.1

and 3.2. Here, we see

$$
\begin{aligned}
&E\left[\sum_{m\in\mathcal{M}} \delta_m I(G_m = \widehat{G}_m \neq 0) \mid \boldsymbol{X} = \boldsymbol{x}\right] = \sum_{m\in\mathcal{R}} E\left[I(G_m = \widehat{G}_m \neq 0) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \sum_{m\in\mathcal{R}} Pr(G_m = \widehat{G}_m \neq 0 \mid \boldsymbol{X} = \boldsymbol{x}) = \sum_{m\in\mathcal{R}} Pr(G_m = k \mid \widehat{G}_m = k, \boldsymbol{X} = \boldsymbol{x}) \\
&= \sum_{m\in\mathcal{R}} [1 - Pr(G_m \neq k \mid \widehat{G}_m = k, \boldsymbol{X} = \boldsymbol{x})] = R - \sum_{m\in\mathcal{R}} Pr(G_m \neq k \mid \widehat{G}_m = k, X_m = x_m) \\
&= R - \sum_{m\in\mathcal{R}} \text{LMCR}_m.
\end{aligned}
$$

In the first equality, we are restricted to the rejection set $\mathcal{R}$ since δ_m is a function of $\boldsymbol{X}$. The second equality is an application of the definition of expectation, and the third equality rewrites the equation in terms of the known information, in that, $\widehat{G}_m$ is known when X_m is known. The fourth equality introduces the notion of misclassification rate $(G_m \neq k)$. The fifth equality is due to the independence of the model, and the sixth equality simplifies the summation. □

Theorem 4.1. *Let $\mathcal{R}$ be any rejection set with corresponding decision functions $\boldsymbol{\delta}$ satisfying Equation* (3.2)*. Further, let $\mathcal{R}^{LMCR}(\boldsymbol{LMCR}, R^{LMCR})$ be the rejection set defined by the LMCR procedure with corresponding decision functions $\boldsymbol{\delta}^{LMCR}$. Then, under Model 1,*

$$
E\left[\sum_{m\in\mathcal{M}} \delta_m I(G_m = \widehat{G}_m \neq 0)\right] \leq E\left[\sum_{m\in\mathcal{M}} \delta_m^{LMCR}\ I(G_m = \widehat{G}_m \neq 0)\right]. \tag{4.5}
$$

Proof of Theorem 4.1. As in the proofs of Theorems 3.2 and 3.3, we observe that, for any $\mathcal{R}$ and corresponding decision functions $\boldsymbol{\delta}$,

$$
\begin{aligned}
\text{CCND} = E\left[\sum_{m\in\mathcal{M}} \delta_m I(G_m = \widehat{G}_m \neq 0)\right] &= E\left[E\left[\sum_{m\in\mathcal{M}} \delta_m I(G_m = \widehat{G}_m \neq 0) \mid \boldsymbol{X} = \boldsymbol{x}\right]\right] \\
&= E\left[|\mathcal{R}| - \sum_{m\in\mathcal{R}} \text{LMCR}_m\right]
\end{aligned}
$$

by the law of iterated expectation and by applying the results of Proposition 4.1.

Thus, it suffices to show that

$$\mathcal{R}^{LMCR}(\mathbf{LMCR}, R^{LMCR}) = \underset{\mathcal{R}:\sum_{m\in\mathcal{R}} \mathrm{Lfdr}_m \leq \alpha|\mathcal{R}|}{\operatorname{argmax}} \left\{ |\mathcal{R}| - \sum_{m\in\mathcal{R}} \mathrm{LMCR}_m \right\}.$$

From Equation (3.1), for any $\mathcal{R}$ where $|\mathcal{R}|= r$, we establish the following:

$$\begin{aligned}
&\underset{\mathcal{R}:\sum_{m\in\mathcal{R}} \mathrm{Lfdr}_m \leq \alpha|\mathcal{R}|}{\operatorname{argmax}} \left\{ |\mathcal{R}| - \sum_{m\in\mathcal{R}} \mathrm{LMCR}_m \right\} \\
&\qquad \leq \underset{\mathcal{R}(\mathbf{LMCR},r):\sum_{m\in\mathcal{R}(\mathbf{LMCR},r)} \mathrm{Lfdr}_m \leq \alpha|\mathcal{R}(\mathbf{LMCR},r)|}{\operatorname{argmax}} \left\{ |\mathcal{R}(\mathbf{LMCR}, r)| - \sum_{m\in\mathcal{R}(\mathbf{LMCR},r)} \mathrm{LMCR}_m \right\} \\
&\qquad = \mathcal{R}^{LMCR}(\mathbf{LMCR}, R^{LMCR}).
\end{aligned}$$

The inequality holds because for any $\mathcal{R}$ with cardinality r we have

$$|\mathcal{R}| - \sum_{m\in\mathcal{R}} \mathrm{LMCR}_m = r - \sum_{m\in\mathcal{R}} \mathrm{LMCR}_m \leq \; r - \sum_{j=1}^{r} \mathrm{LMCR}_{(j)} = r - \sum_{m\in\mathcal{R}(\mathbf{LMCR},r)} \mathrm{LMCR}_m.$$

The equality is satisfied since $|\mathcal{R}(\mathbf{LMCR}, r)| - \sum_{m\in\mathcal{R}(\mathbf{LMCR},r)} \mathrm{LMCR}_m$ is nondecreasing in r and by the definition of $\mathcal{R}^{LMCR}(\mathbf{LMCR}, R^{LMCR})$. □

Chapter 5

Comparing the Lfdr and LMCR Procedures

This chapter considers the Lfdr and LMCR procedures using two illustrative scenarios and three real data applications.

5.1 Illustrating Each Procedure

This section compares the proposed procedure to the standard approach, which seeks to make as many discoveries as possible subject to FDR control and then classifies those discoveries. As mentioned in the Introduction, the main difference is that the proposed procedure considers classification error when ranking the hypotheses from most to least significant, while the standard ranks hypotheses using the Lfdr statistic, which recall is ideal if the objective is to maximize the expected number of true discoveries (see Theorem 3.2), but not necessarily ideal otherwise. The two main advantages of the LMCR procedure are: (1) it provides a built-in safeguard that prevents the rejection of hypotheses that would be classified as null ($\widehat{G}_m = 0$) and (2) rejects hypotheses that have the smallest LMCR. The rest of this section uses Figure 5.1 and Tables 5.1 and 5.2 to further compare the two procedures.

Figure 5.1a presents a two group model where $X_m \sim 0.8N(0,1) + 0.2N(2,1)$, so that $A_0 = \{x < 1.7\}$ and $A_1 = \{x > 1.7\}$. Suppose that $M = 5$ hypotheses are tested with $x_1 = 1.4$, $x_2 = 2.2$,

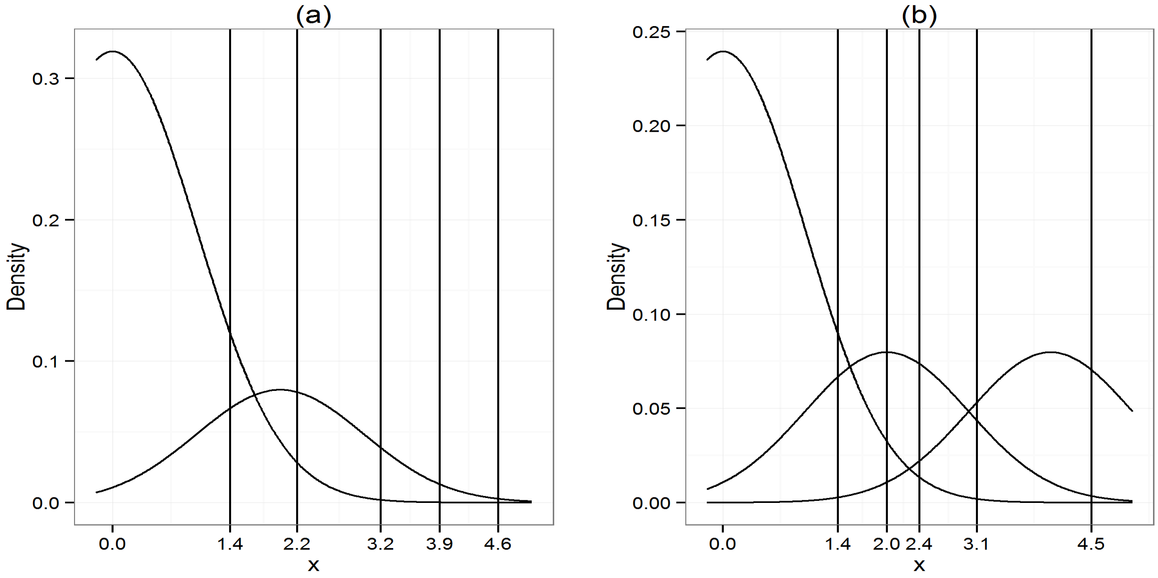

Figure 5.1: Examples of different normal mixture configurations with vertical lines representing realizations of $X_1, X_2, ..., X_5$: (a) two group mixture and (b) non-symmetric three group mixture.

$x_3 = 3.2$, $x_4 = 3.9$, $x_5 = 4.6$. Column one of Table 5.1 lists these values along with the corresponding Lfdr and LMCR ranks (from most significant to least significant).

			Rejected by Lfdr/LMCR Procedure for α value:				
x_m	Lfdr (Rank)	LMCR (Rank)	0.005	0.01	0.05	0.10	0.20
1.4	0.643 (5)	NA	N/N	N/N	N/N	N/N	Y/N
2.2	0.266 (4)	0.266 (4)	N/N	N/N	N/N	Y/Y	Y/Y
3.2	0.047 (3)	0.047 (3)	N/N	N/N	Y/Y	Y/Y	Y/Y
3.9	0.012 (2)	0.012 (2)	N/N	Y/Y	Y/Y	Y/Y	Y/Y
4.6	0.003 (1)	0.003 (1)	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y

Table 5.1: Summary of x values, Lfdr, LMCR, and rejection information for the five points indicated in Figure 5.1a. For the rejection columns, a Y (N) indicates that a given hypothesis is rejected (not rejected) by the Lfdr/LMCR procedure at the specified α level.

Here, both procedures yield the same set of rejected hypotheses for $\alpha \leq 0.10$ since, as you move from left to right in Figure 5.1a, the Lfdr decreases as the LMCR decreases. That is, we are more sure of having correctly classified (and are more likely to reject) hypotheses with higher x values.

Formally, for $k \in \mathcal{K}^* = \{1\}$ and $x \in A_1$,

$$\text{LMCR}(x,k) = \text{LMCR}(x,1) = 1 - \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \text{Lfdr}(x). \tag{5.1}$$

When $\alpha = 0.20$, the LMCR procedure automatically retains the H_m corresponding to $x_m = 1.4 \in A_0$, thus illustrating the rejection safeguard. The Lfdr procedure rejects this hypothesis even though it will be classified into the null group. This behavior will be discussed further in Chapter 6.

Figure 5.1b illustrates a situation where $X_m \sim 0.6N(0,1) + 0.2N(2,1) + 0.2N(4,1)$, so that $A_0 = \{x \leq 1.5\}$, $A_1 = \{1.5 < x \leq 3.0\}$, and $A_2 = \{x > 3.0\}$. Again, suppose that $M = 5$ hypotheses are tested with $x_1 = 1.4$, $x_2 = 2.0$, $x_3 = 2.4$, $x_4 = 3.1$, $x_5 = 4.5$. In Table 5.2, observe that the Lfdr values decrease as x increases, as expected. But, as we can see here, the LMCR values need not follow this pattern. For example, $x_m = 3.1$ is near the boundary for classification into one of the two nonnull groups, so its LMCR is 0.461 which the LMCR procedure ranks as 4th most significant even though it has the 2nd smallest Lfdr. However, $x_m = 2.4$ is further from the boundary of A_1 and A_2, so the LMCR procedure ranks its LMCR = 0.326 as 2nd most significant.

			Rejected by Lfdr/LMCR Procedure for α value:				
x_m	Lfdr (Rank)	LMCR (Rank)	0.005	0.01	0.05	0.10	0.20
1.4	0.564 (5)	NA	N/N	N/N	N/N	N/N	Y/N
2.0	0.263 (4)	0.351 (3)	N/N	N/N	N/N	N/N	Y/Y
2.4	0.123 (3)	0.326 (2)	N/N	N/N	Y/N	Y/Y	Y/Y
3.1	0.020 (2)	0.461 (4)	N/N	Y/N	Y/N	Y/N	Y/Y
4.5	<0.001 (1)	0.048 (1)	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y

Table 5.2: Summary of x values, Lfdr, LMCR, and rejection information for the five points indicated in Figure 5.1b. For the rejection columns, a Y (N) indicates that a given hypothesis is rejected (not rejected) by the Lfdr/LMCR procedure at the specified α level.

In general, the LMCR procedure discovers X_m's that are "most easily" classified, which need not correspond to the largest x values or smallest Lfdr values. For example, when $\alpha = 0.10$ the LMCR procedure rejects the H_m's corresponding to $x_m = 4.5$ and $x_m = 2.4$ since the LMCRs are 0.048 and 0.326, but retains the H_m corresponding to $x_m = 3.1$ since its LMCR = 0.461. In short, the LMCR procedure focuses on discovering attributes that are most easily classifiable, as per the objectives of Anderson and Habiger (2012).

5.2 Applying Both Procedures to Real Data

In this section, we consider three examples that directly apply the LMCR procedure of Section 4.2 and the competing procedure discussed above. For the first two examples, normal mixture models are considered using Z statistics, statistics that have been transformed such that they have a mean of 0 and a variance of 1, for convenience. The third example uses a multinomial mixture model.

5.2.1 Cancer Data

The first dataset is the p53 data discussed in Chapter 9 of Efron (2010). This dataset consists of information for 33 mutated and 17 unmutated cell lines of the p53 gene, as collected by the National Cancer Institute. In total, there are 10,100 gene expressions measured for each cell line.

Formally, for each $m = 1, 2, ..., 10100$, let x_{mj} be the gene expression level for gene m and cell line j. Further, let group 1 be the unmutated group with population mean μ_m^1 and corresponding sample mean $\bar{x}_m^1$, and let group 2 be the mutated group with population mean μ_m^2 and corresponding sample mean $\bar{x}_m^2$. A standard two sample t-test can be applied to each gene, with null hypothesis

$$H_m : \mu_m^1 = \mu_m^2. \tag{5.2}$$

The test statistic is then

$$t_m = \frac{\bar{x}_m^2 - \bar{x}_m^1}{s_m}, \tag{5.3}$$

where

$$s_m = \sqrt{\frac{\sum_{j=1}^{17}(x_{mj} - \bar{x}_m^1)^2 + \sum_{j=18}^{50}(x_{mj} - \bar{x}_m^2)^2}{17 + 33 - 2}\left(\frac{1}{17} + \frac{1}{33}\right)}, \tag{5.4}$$

and has a Student's t distribution with 48 degrees of freedom under H_m. For convenience, we then convert each t statistic into a corresponding z statistic using

$$Z_m = \Phi^{-1}(F_{48}(t_m)), \tag{5.5}$$

where Φ and F_{48} are the cdfs for the standard normal distribution and t distribution with 48 degrees of freedom, respectively. Figure 5.2 presents the histogram of these z-values, the fitted density curves, and the procedure results when $\alpha = 0.05$.

For this dataset, the `Mclust` function (as discussed in Subsection 2.5.3) was used to find a five group normal mixture model whose parameter estimates ($\hat{\pi}$, $\hat{\mu}$, and $\hat{\sigma}$) are given in Table 5.3. The groups were labeled based on their position relative to the assigned null group. Loosely, the lower extreme group ($G_m = 1$) can be interpreted as consisting of those hypotheses that are most likely to be highly significant on the low end, whereas the upper extreme group ($G_m = 4$) consists of those hypotheses that are most likely highly significant on the high end. The intermediate groups are meant to signify those hypotheses that are likely nonnull, but their effects are not as pronounced as those in the extreme groups ($G_m = 2$ for the lower intermediate group and $G_m = 3$ for the upper intermediate group). $G_m = 0$ is reserved for the null group. In the context of this dataset, the distinction between intermediate and extreme effects may be of interest.

					Lfdr/LMCR Rejections			
Group $= k$	$\hat{\pi}_k$	$\hat{\mu}_k$	$\hat{\sigma}_k$	$\lvert A_k \rvert$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$	$\alpha = 0.20$
$G_m = 1$	0.087	-1.556	0.651	554	554/248	554/262	554/385	554/554
$G_m = 2$	0.304	-0.629	0.651	4050	732/0	1741/0	2776/1974	3934/3935
$G_m = 0$	0.219	0.054	0.651	861	0/0	0/0	0/0	0/0
$G_m = 3$	0.316	0.744	0.651	4245	821/195	1881/850	2938/2346	4139/4138
$G_m = 4$	0.074	1.643	0.651	390	390/165	390/169	390/243	390/390
Total number of rejections					2497/608	4566/1281	6658/4948	9017/9017
Average Lfdr among rejections					0.05/0.05	0.10/0.10	0.15/0.15	0.20/0.20
Average LMCR among rejections					0.37/0.23	0.35/0.27	0.36/0.34	0.41/0.41

Table 5.3: Parameters and select results for the five group normal mixture model applied to the p53 dataset. The $|A_k|$ column represents the number of Z_m values that were classified into each group. The Lfdr/LMCR rejection columns show the number of hypotheses that were rejected for each group for different levels of α.

While the null group is not a true standard normal, $N(0, 1)$, this example does illustrate a stark contrast between the rejection sets of the LMCR and the Lfdr procedures, echoing the results of the example in the previous section. Table 5.3 also provides information on the number of hypotheses classified into each group (the cardinality of $A_k = |A_k|$), and the rejection results for each procedure.

Neither procedure rejects all hypotheses classified as nonnull for any given α, thus there are zero rejections for hypotheses classified as null, in each case. For the nonnull groups, we see that the Lfdr procedure rejects more hypotheses per group than the LMCR procedure for each given $\alpha \neq 0.20$. This behavior is expected, in that, the LMCR procedure is focused on rejecting only those hypotheses with the smallest LMCR rather than rejecting the maximum number of hypotheses, as the Lfdr procedure does. This behavior is also seen in the average LMCR among rejections values, in that the LMCR procedure provides rejection sets with lower average LMCR values, especially for $\alpha = 0.05$, as Theorem 4.1 suggests. Because of the orientation of the distributions, when $\alpha = 0.20$ the rejection sets for both procedures are virtually the same (with a single difference on both the high and low ends).

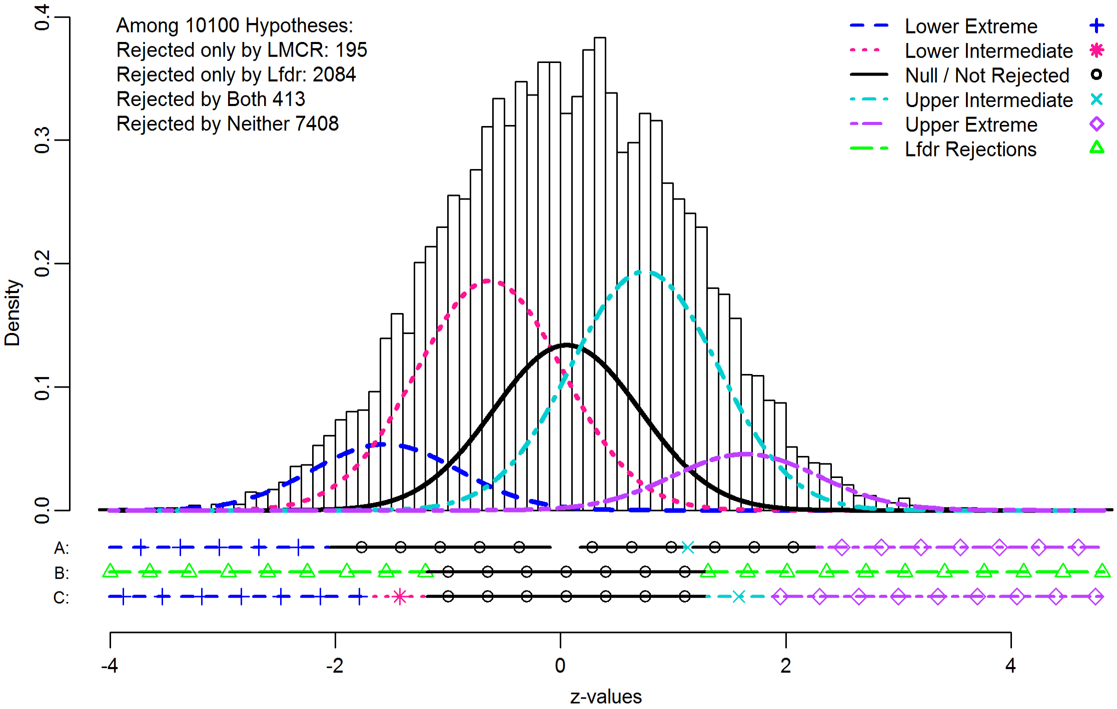

Figure 5.2: The histogram of z-values with fitted densities for the p53 dataset. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C) when $\alpha = 0.05$.

In Figure 5.2, the three rows of points/lines below the histogram, labeled A-C, illustrate the rejection sets for the LMCR, the Lfdr, and the clustered Lfdr methods, respectively. As defined above, the clustered Lfdr results show the current state of FDR analyses when coupled with classification. Namely, a rejection set is defined (as in row B) and then hypotheses are assigned to groups while ignoring the possibility of misclassification.

In row A, we see that there is a very narrow band of rejected hypotheses from the upper intermediate group near 1.1. This band and the corresponding gap in the set of rejections is the behavior described in the discussion of Figure 5.1b. Based on the specifics of this mixture model, we can be fairly certain that the hypotheses in that band are correctly classified as nonnull with an intermediate level of significance. Unfortunately, the close proximity of the null distribution exhausts our ability to reject any more than the 195 hypotheses listed in the table. This closeness also explains why there are no LMCR rejections from the lower intermediate group at all. The B row illustrates the Lfdr procedure's indifference toward classification with 1286 consecutive rejections on the lower end and 1211 consecutive rejections on the upper end. Row C shows the effects of classification for the Lfdr rejection set where we again mention that the Lfdr procedure rejects more hypotheses in each and every nonnull group, as compared to the LMCR procedure.

5.2.2 Productivity Associated Microbiome Data: Normal Mixture Model

For a second example, consider the dataset shown in Table 5.4 and described below. The main goal here is to identify and correctly classify the association between bacterial species abundance and shoot biomass in wheat plants. For more information on the characterization and analysis of this "productivity associated microbiome" see Anderson and Habiger (2012) and Habiger et al. (2015).

For this dataset, we let y_{mn} be the abundance of species m in productivity group n where $m = 1, 2, ..., M$ and $i = 1, 2, ..., N$ where $M = 778$ and $N = 5$. Here, a productivity group represents the shoot biomass measured in grams and is defined to be $\boldsymbol{x} = (x_1, x_2, ..., x_5)^T = (0.86, 1.34, 1.81, 2.37, 3.00)^T$, where superscript T is the transpose operator. For bacterial species m, its abundance in each productivity group is denoted by $\boldsymbol{y}_m = (y_{m1}, y_{m2}, ..., y_{m5})^T$ and its total abundance is defined as $n_m = \sum_{i=1}^{5} y_{mi}$. We then denote the associated random vector and random

	Shoot Biomass Groups					
Bacterial Species m	y_{m1}	y_{m2}	y_{m3}	y_{m4}	y_{m5}	Total (n_m)
1	2	1	3	0	2	8
2	0	0	2	3	7	12
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
777	9	1	4	0	6	20
778	1	0	2	3	1	7

Table 5.4: Overview of the productivity associated microbiome dataset showing the species identification number, prevalence within each biomass group, and row total. The shoot biomass values are measured in grams and are given by $\boldsymbol{x} = (0.86, 1.34, 1.81, 2.37, 3.00)^T$.

sample size by $\boldsymbol{Y}_m$ and N_m, respectively. The matrix of observed data can then be written as $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, ..., \boldsymbol{y}_{778}^T)^T$ and the corresponding random matrix as $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T, ..., \boldsymbol{Y}_{778}^T)^T$.

The analysis begins by assuming that $Y_{mi} \sim \text{Pois}(\mu_{mi})$ where $\log(\mu_{mi}) = \alpha_m + \beta_m x_i$. Here, both α_m and β_m are parameters with values in $\mathbb{R}$ and all Y_{mi} are mutually independent. As we are testing whether or not species m is associated with shoot biomass, the null hypothesis of interest is $H_m : \beta_m = 0$. We say that species m is positively (negatively) associated with productivity if $\beta_m > 0$ (< 0).

To simplify the analysis, estimation of α_m can be avoided by implementing a procedure based on the conditional distribution $\boldsymbol{Y}_m | N_m = n_m$. This distribution has multinomial pmf

$$f(\boldsymbol{y}_m|\beta_m, n_m) = \frac{n_m!}{\prod_{i=1}^5 y_{mi}!} \prod_{i=1}^{5} p_i(\beta_m)^{y_{mi}} \tag{5.6}$$

where

$$p_i(\beta_m) = \frac{\exp(\beta_m x_i)}{\sum_{i=1}^5 \exp(\beta_m x_i)} \text{ for } i = 1, 2, ..., 5. \tag{5.7}$$

For additional details, see McCullagh and Nelder (1989).

If we let $\boldsymbol{p}(\beta_m) = [p_1(\beta_m), p_2(\beta_m), ..., p_5(\beta_m)]^T$ be the multinomial probability vector defined by Equation (5.7), we see that $E[\boldsymbol{Y}_m|\beta_m, n_m] = n_m \boldsymbol{p}(\beta_m)$ and

$$\text{Cov}[\boldsymbol{Y}_m|\beta_m, n_m] = n_m[\boldsymbol{D}(\boldsymbol{p}(\beta_m)) - \boldsymbol{p}(\beta_m)\boldsymbol{p}(\beta_m)^T],$$

where $\boldsymbol{D}(\boldsymbol{p}(\beta_m))$ is a 5×5 diagonal matrix with diagonal elements $p_i(\beta_m),\ i = 1, 2, ..., 5$. Now, if we consider the sufficient statistic $T_m = \boldsymbol{x}^T \boldsymbol{Y}_m$, under Equation (5.6), we can define

$$\mu(\beta_m, n_m) = E[T_m|\beta_m, n_m] = \boldsymbol{x}^T E[\boldsymbol{Y}_m|\beta_m, n_m] = n_m \boldsymbol{x}^T \boldsymbol{p}(\beta_m)$$

and

$$\sigma^2(\beta_m) = \text{Var}[T_m|\beta_m, n_m] = \boldsymbol{x}^T \text{Cov}[\boldsymbol{Y}_m|\beta_m, n_m]\boldsymbol{x} = n_m \boldsymbol{x}^T \left[\boldsymbol{D}(\boldsymbol{p}(\beta_m)) - \boldsymbol{p}(\beta_m)\boldsymbol{p}(\beta_m)^T\right] \boldsymbol{x}.$$

We are now able to define Z-scores under the null hypothesis ($H_m : \beta_m = 0$) as

$$Z_m = \frac{T_m - \mu(0, n_m)}{\sigma(0)}.$$

To consider this situation in the context of a mixture model as defined in Model 1, suppose we are interested in modeling these Z-values using $K + 1$ groups (one null and K nonnull groups). For $k \in \mathcal{K}$, let $\gamma_0, \gamma_1, ..., \gamma_K$ represent the distinct values of β_m where $Pr(\beta_m = \gamma_k) = \pi_k$, and define

$$E[Z_m|\beta_m = \gamma_k, n_m] = E\left[\frac{T_m - \mu(0, n_m)}{\sigma(0)} \,\middle|\, \beta_m = \gamma_k, n_m\right] = \frac{\mu(\gamma_k, n_m) - \mu(0, n_m)}{\sigma(0)}$$

and

$$\text{Var}[Z_m|\beta_m = \gamma_k, n_m] = \text{Var}\left[\frac{T_m - \mu(0, n_m)}{\sigma(0)} \,\middle|\, \beta_m = \gamma_k, n_m\right] = \frac{\text{Var}[T_m|\beta_m = \gamma_k, n_m]}{\sigma^2(0)} = \frac{\sigma^2(\gamma_k)}{\sigma^2(0)}.$$

So that,

$$Z_m|\beta_m = \gamma_k, n_m \sim N\left(\frac{\mu(\gamma_k, n_m) - \mu(0, n_m)}{\sigma(0)},\ \frac{\sigma^2(\gamma_k)}{\sigma^2(0)}\right), \tag{5.8}$$

where it is important to note that the mean of this normal distribution depends on n_m. This reliance on n_m is important because each unique value of n_m will create a unique distribution for each γ_k. While the definition of $\sigma^2(\gamma_k)$ also depends on n_m, the ratio of σ^2 values in Equation (5.8) removes this dependence. See Habiger et al. (2015) for details.

Suppose we are interested in fitting a four group (one null and three nonnull) mixture model to the data presented in Table 5.4. Under Model 1, the mixture density is given by

$$f(z) = \sum_{k \in \mathcal{K}} \pi_k \phi_k(z|\gamma_k, n) \tag{5.9}$$

as in Equation (2.2), where $\mathcal{K} = \{0, 1, 2, 3\}$ and $\phi_k(z|\gamma_k, n)$ is the normal pdf associated with Equation (5.8).

An EM algorithm was developed to get maximum likelihood estimates for π_k and γ_k, the details of which are partially discussed in Subsection 2.5.3 and Section 7.1. To initialize the EM algorithm, the `ga` function was used from the **GA** package (Scrucca, 2013) in `R`. The `ga` function is an implementation of a genetic algorithm which is a type of evolutionary stochastic search. For details of and multiple references for genetic algorithms, see Scrucca (2013).

Table 5.5 shows the results of the convergent EM algorithm (initially a five group model was considered, but the algorithm failed to converge). The group labels are assigned based on the values of $\hat{\gamma}_k$, relative to the null group (where $\gamma_k = 0$ is fixed). We note that there are two groups that show negative association with shoot biomass, an extreme group ($\hat{\gamma}_1 = -1.216$) and an intermediate group ($\hat{\gamma}_2 = -0.821$), and a single group with a positive association ($\hat{\gamma}_3 = 0.705$).

				Lfdr/LMCR Rejections			
Group $= k$	$\hat{\pi}_k$	$\hat{\gamma}_k$	$\lvert A_k \rvert$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$	$\alpha = 0.20$
$G_m = 1$	0.120	-1.216	68	48/29	57/39	68/54	68/68
$G_m = 2$	0.054	-0.821	21	13/9	18/12	21/14	21/21
$G_m = 0$	0.650	0	597	0/0	0/0	3/0	15/0
$G_m = 3$	0.176	0.705	92	32/37	56/62	72/82	92/92
Total number of rejections				93/75	131/113	164/150	196/181
Average Lfdr among rejections				0.05/0.05	0.10/0.10	0.15/0.15	0.20/0.18
Average LMCR among rejections				0.14/0.08	0.19/0.15	0.24/0.21	0.28/0.26

Table 5.5: Parameters and select results for the four group normal mixture model applied to the bacteria dataset. The $|A_k|$ column represents the number of bacterial species that were classified into each group. The Lfdr/LMCR rejection columns show the number of hypotheses that were rejected for each group for different levels of α.

Table 5.5 also presents the results of the classification analysis where the $|A_k|$ column signifies how many bacterial species are assigned to each group. The Lfdr/LMCR rejection columns present the number of rejections per group when the Lfdr/LMCR procedures are applied for $\alpha = 0.05,\ 0.10,\ 0.15,$ and 0.20. First, note that when $\alpha = 0.20$, the LMCR procedure's safeguard effect prevents the discovery of species that are most likely not associated with productivity. Thus, the average Lfdr among rejections for the LMCR procedure is $0.18 < \alpha = 0.20$.

Next, observe that the Lfdr procedure makes more discoveries for all values of α, as suggested by Theorem 3.2. For example, when $\alpha = 0.10$, the LMCR procedure discovers 113 species while the Lfdr procedure discovers 131. However, as Theorem 4.1 suggests, the rejection set defined by the LMCR procedure has smaller average LMCR. For example, when $\alpha = 0.10$, the average LMCR among rejections for the LMCR procedure is 0.15 whereas the average LMCR among rejections is 0.19 for the Lfdr procedure. This means that the 39 species identified as strongly negatively associated with plant health by the LMCR procedure are more likely to be correctly identified as such.

5.2.3 Productivity Associated Microbiome Data: Multinomial Mixture Model

Here, consider the dataset shown in Table 5.4 and described above. Rather than analyzing the data using an approximate normal mixture model, we now implement an exact multinomial mixture model developed using Equation (5.6).

As above, to fit into the context of Model 1, let $\gamma_0, \gamma_1, ..., \gamma_K$ represent the distinct possible values of β_m, so that for $k \in \mathcal{K} = \{0, 1, ..., K\}$, when $\beta_m = \gamma_k$ we say $G_m = k$ and thus $Pr(\beta_m = \gamma_k) = \pi_k$. Now, our mixture pmf is

$$f(\boldsymbol{y}_m|\boldsymbol{\gamma}, \boldsymbol{\pi}, n_m) = \sum_{k \in \mathcal{K}} \pi_k f(\boldsymbol{y}_m|\gamma_k, n_m)$$

where $\boldsymbol{\gamma} = (\gamma_k,\ k \in \mathcal{K})$ and $\boldsymbol{\pi} = (\pi_k,\ k \in \mathcal{K})$. Here, the appropriate posterior probability is

$$Q_m = Pr(\theta_m = 0|\boldsymbol{y}_m, \boldsymbol{\gamma}, \boldsymbol{\pi}, n_m) = \frac{\pi_0 f(\boldsymbol{y}_m|\gamma_0, n_m)}{f(\boldsymbol{y}_m|\boldsymbol{\gamma}, \boldsymbol{\pi}, n_m)}. \tag{5.10}$$

Another EM algorithm was used to obtain maximum likelihood estimates for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ (see Habiger et al. (2015) for the specifics of implementation). Table 5.6 shows the results of a five group model (representing strongly negative, weakly negative, null, weakly positive, and strongly positive associations). Table 5.6 also presents the results of the classification analysis, where the $|A_k|$ column signifies how many bacterial species are assigned to each group as per Equation (2.3). The Lfdr/LMCR rejection columns present the number of rejections per group when the Lfdr/LMCR procedures are applied for $\alpha = 0.05,\ 0.10,\ 0.15,\ 0.20$, and 0.25. The $\alpha = 0.25$ case is included here to again illustrate the LMCR procedure's safeguard effect. Note that the LMCR procedure rejects all bacteria classified into one of the nonnull groups before reaching the threshold (the average Lfdr among rejections is $0.24 < \alpha = 0.25$).

				Lfdr/LMCR Rejections				
Group $= k$	$\hat{\pi}_k$	$\hat{\gamma}_k$	$\|A_k\|$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$	$\alpha = 0.20$	$\alpha = 0.25$
$G_m = 1$	0.028	-2.686	23	23/13	23/18	23/20	23/21	23/23
$G_m = 2$	0.146	-1.018	102	50/46	74/76	87/83	100/90	102/102
$G_m = 0$	0.480	0	483	0/0	0/0	0/0	0/0	17/0
$G_m = 3$	0.292	0.246	127	9/9	21/16	49/40	84/75	124/127
$G_m = 4$	0.054	1.344	43	32/19	43/24	43/27	43/32	43/43
Total number of rejections				114/87	161/134	202/170	250/218	309/295
Average Lfdr among rejections				0.05/0.05	0.10/0.10	0.15/0.15	0.20/0.20	0.25/0.24
Average LMCR among rejections				0.18/0.10	0.24/0.18	0.28/0.22	0.32/0.28	0.35/0.34

Table 5.6: Parameters and select results for the five group multinomial mixture model applied to the bacteria dataset. The $|A_k|$ column represents the number of bacterial species that were classified into each group. The Lfdr/LMCR rejection columns show the number of hypotheses that were rejected for each group for different levels of α.

First, observe that the Lfdr procedure makes more discoveries for all values of α, which is consistent with Theorem 3.2. For example, when $\alpha = 0.10$, the LMCR procedure discovers 134 bacteria while the Lfdr procedures discovers 161. However, the LMCR procedure has smaller average LMCR, as Theorem 4.1 suggests. For example, when $\alpha = 0.10$, the average LMCR among rejections is 0.18 for the LMCR procedure while the average LMCR among rejections is 0.24 for the Lfdr procedure. This means, for example, that the 18 bacteria declared to be strongly negatively associated with productivity by the LMCR procedure are more likely correctly declared as such, thereby allowing for more rigorous soil science theories to be posited and tested in future studies.

Next, observe that both the Lfdr and LMCR procedures make more total discoveries when using the multinomial mixture model as compared to the normal mixture model seen in the previous subsection. For example, when $\alpha = 0.05$, the Lfdr procedure under the normal mixture discovers 93 bacteria whereas the Lfdr procedure under the multinomial mixture discovers 114 bacteria. The corresponding number of LMCR discoveries is 75 and 87, respectively. However, we note that the change in models increases the average LMCR among rejections. In going from the normal mixture to the multinomial mixture, the average LMCR among rejections increases from 0.14 to 0.18 for the Lfdr procedure and from 0.08 to 0.10 for the LMCR procedure when $\alpha = 0.05$.

Chapter 6

Simulation

This chapter compares the numerical performance of the LMCR procedure with the competing procedure mentioned in Sections 5.1 and 5.2. Recall that the current standard method for incorporating classification information involves applying the Lfdr procedure first and then classifying the resulting rejections.

Consider a three group normal mixture model defined by:

$$X_m \sim \pi_0 N(0, \sigma_0^2) + \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2). \tag{6.1}$$

Note that these three groups are oriented non-symmetrically as in Figure 5.1b and that the mean of the null group (μ_0) will always be 0. Also note that the upper intermediate and upper extreme groups are labeled as group 1 and group 2, respectively (that is, $0 < \mu_1 < \mu_2$).

The following sections focus on adjustments to the α value; the value of M; the locations of μ_1 and μ_2; the mixing proportions π_0, π_1, and π_2; and the standard deviations σ_0, σ_1, and σ_2. Specifically, the two procedures will be compared with respect to the average number of rejected hypotheses, the calculated false discovery proportion (FDP), the false nondiscovery proportion (FNP), and the misclassification proportion among rejected hypotheses (MP) for $1,000$ repetitions of the specified simulation settings. Here, the FNP is defined as the number of nonnull hypotheses that are claimed nonsignificant divided by the total number of hypotheses claimed nonsignificant and the MP is defined as the number of hypotheses that are incorrectly classified ($\widehat{G}_m \neq G$) and rejected divided by the total number of hypotheses rejected. That is,

$$\text{FNP} = \frac{\sum_{m \in \mathcal{M}} (1-\delta_m) I(G_m \neq 0)}{\sum_{m \in \mathcal{M}} (1-\delta_m)} \qquad \text{and} \qquad \text{MP} = \frac{\sum_{m \in \mathcal{R}} \delta_m I(\widehat{G}_m \neq G_m)}{\sum_{m \in \mathcal{R}} \delta_m}.$$

6.1 Illustrating the LMCR Procedure's Safeguard Behavior: The Effect of α

In this section, we consider Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $M = 10,000$ hypotheses with varying α levels. The simulation results of $1,000$ repetitions are presented in Figure 6.1.

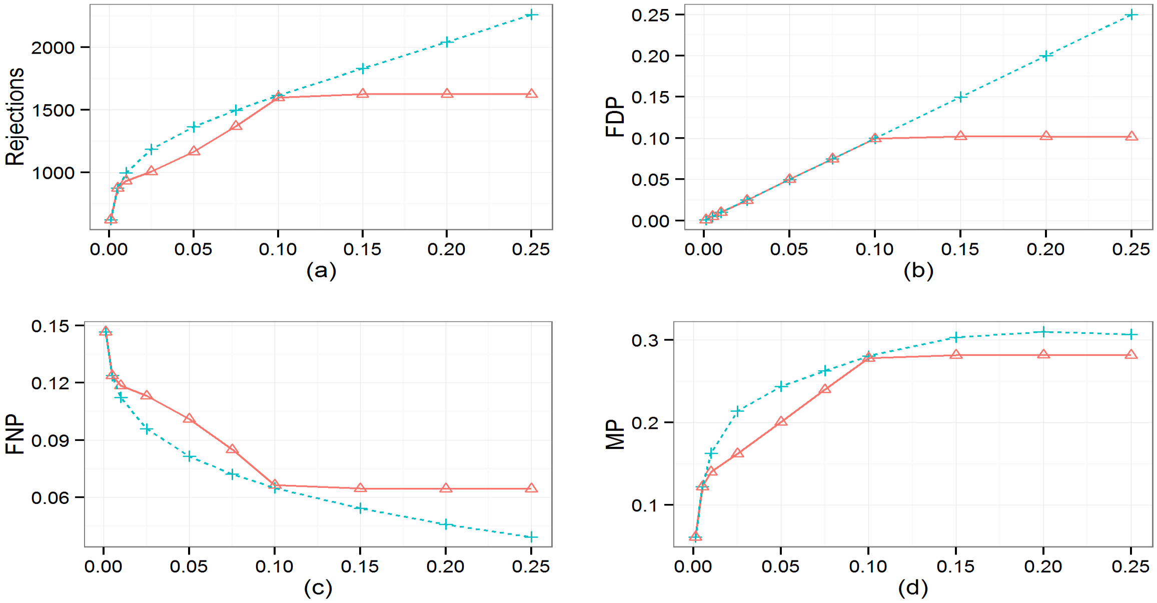

Figure 6.1: The Effect of α: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

In each plot we see that the LMCR procedure exhibits the previously mentioned safeguard effect when $\alpha \geq 0.10$. Recall that the LMCR procedure, by design, only tests those hypotheses that have been classified as nonnull (contrary to the Lfdr procedure). So, for α approximately greater than or equal to 0.10, the LMCR procedure rejects all such hypotheses and automatically retains all remaining hypotheses. At this point, the rejection set is fixed so that the average number of rejections, FDP, FNP, and MP are also fixed.

As an example, consider the case when $\alpha = 0.15$, as presented in Figure 6.2. Here, the LMCR procedure rejects every hypothesis classified as either upper intermediate or upper extreme, but must then stop. The Lfdr procedure, on the other hand, is able to continue rejecting hypotheses (an additional 205, in fact). Hence, we now have a new symbol in row C signifying those rejected hypotheses that were classified as null. This figure highlights one of the main advantages of the LMCR procedure, as compared to the standard approach, by directly illustrating the built-in safeguard that prevents researchers from rejecting hypotheses that are more likely null.

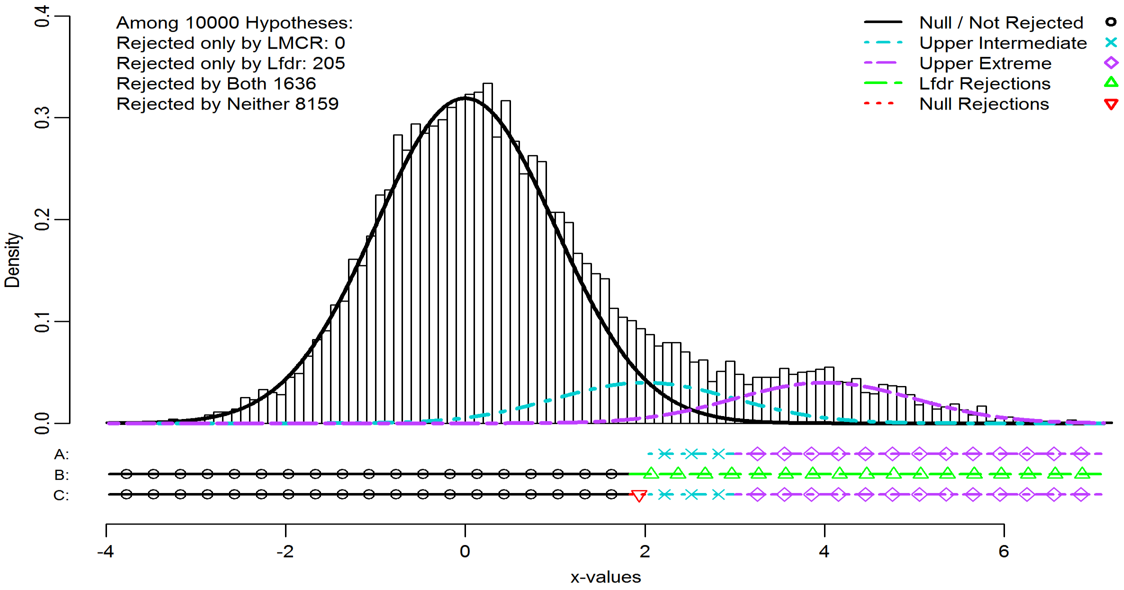

Figure 6.2: The histogram of 10,000 simulated x-values for the α simulation settings when $\alpha = 0.15$, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

To understand this behavior mathematically, we first recall that $A_k,\ k \in \{0, 1, 2\}$ represents the classification region for group k as defined in Section 2.4. Thus, the probability that a hypothesis is classified as nonnull is

$$Pr(X_m \in A_1 \bigcup A_2, G_m = 0) + Pr(X_m \in A_1 \bigcup A_2, G_m = 1) + Pr(X_m \in A_1 \bigcup A_2, G_m = 2) = 0.162.$$

Now, the FDP becomes a measure of the misclassification rate for the null group ($Pr(X_m \in A_1 \bigcup A_2, G_m = 0) = 0.017$). So, the probability that a null hypothesis is classified as nonnull (and thus rejected) is approximately 0.102.

This logic also explains the behavior of the LMCR procedure in plots (c) and (d). Again, when all hypotheses classified as nonnull are then rejected, the FNP becomes

$$\frac{Pr(X_m \in A_0, G_m = 1) + Pr(X_m \in A_0, G_m = 2))}{Pr(X_m \in A_0, G_m = 0) + Pr(X_m \in A_0, G_m = 1) + Pr(X_m \in A_0, G_m = 2)} = \frac{0.054}{0.838} = 0.065.$$

As for the MP for the LMCR procedure, it becomes a measure of the misclassification rate in the rejection region. Here, the probability of misclassification is

$$Pr(X_m \in A_1 \bigcup A_2, G_m = 0) + Pr(X_m \in A_2, G_m = 1) + Pr(X_m \in A_1, G_m = 2) = 0.046.$$

So, the probability of misclassification given rejection becomes $0.046/0.162 = 0.282$.

The separation between the two curves in plots (a), (c), and (d), for $\alpha < 0.1$, arises from the different rejection sets selected by the two procedures. When the α threshold is achieved before rejecting all hypotheses classified as nonnull, the LMCR procedure rejects hypotheses in the right tail of group 2 and/or in a band centered in group 1, thus creating a gap in the rejection set near the intersection of the densities for groups 1 and 2 where misclassification is likely to occur. This type of behavior can be seen in Figure 5.2 above. Since the Lfdr procedure has no such gap in the rejection set, it introduces that misclassification likelihood. Therefore, we see a greater number of rejections for the Lfdr procedure, as well as a greater MP, as compared to the LMCR procedure (as Theorem 4.1 suggests). The separation between the FNP curves can also be explained by the larger number of rejections. As more hypotheses are rejected, the likelihood of claiming a nonnull hypothesis as nonsignificant decreases.

To reiterate, the band of hypotheses in group 1, rejected by the LMCR procedure, and the corresponding gap in rejections centered around the intersection of groups 1 and 2 is important in understanding the real difference between the two procedures (accounting for misclassification versus ignoring it). More specifically, the Lfdr procedure was developed to minimize the FNP while controlling the FDP, thus it outperforms the LMCR procedure in this regard. Likewise, the LMCR procedure was developed to minimize the misclassification rate of rejected hypotheses (thus making them more interpretable) so it dominates.

6.2 The Effect of the Number of Hypotheses

In this section, we consider Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$ with varying M levels. The simulation results of $1,000$ repetitions are presented in Figure 6.3.

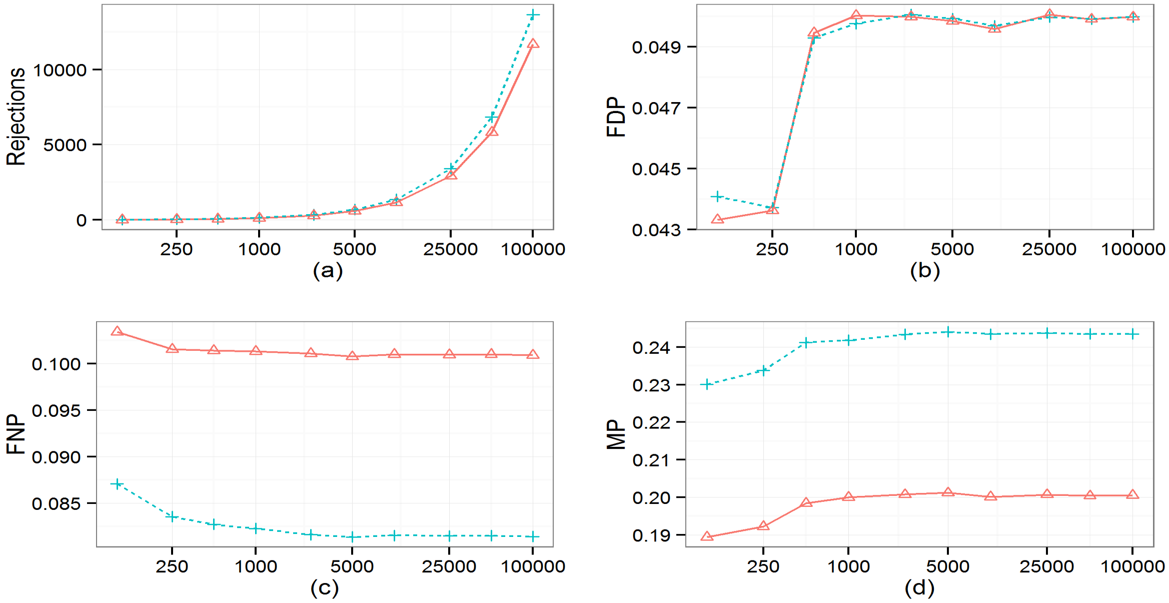

Figure 6.3: The Effect of M: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

In Figure 6.3 we note that the number of rejections increases for both procedures, as expected, but again, the number of LMCR rejections is less than or equal to those for the Lfdr procedure. As for the FDP plot, we see that neither procedure reaches the α threshold among the first couple of M levels, but it seems to stabilize after that. Despite this, it seems that the LMCR procedure's performance is comparable to that of the Lfdr procedure in controlling the FDP. The initial dip for both procedures can be explained by the relatively small number of hypotheses for the $M = 100$

and $M = 250$ cases.

In plots (c) and (d), we see the asymptotic effects of increasing M, in that, both the FNP and MP curves seem to approach a limiting value defined by their respective probability regions. In the case of the LMCR, the region of rejection is not easily defined because of the banding and gapping behavior discussed previously. Here, the band of rejections for group 1 will be a window centered around the intersection of the density curves for the null group and group 2 (approximately 2.040). Thus, a simple probability argument cannot be used to verify the limiting behavior at this time.

The banding and gapping also explain the significant difference in the magnitude for the FNP and MP plots. This simulation setting corresponds to the $\alpha = 0.05$ level of the previous section, thus as seen in Figure 6.1, we should expect separation between the two curves. The scale of each plot explains the sizable nature of this difference. Regardless, we see that the Lfdr procedure performs best in terms of the FNP, whereas the LMCR procedure dominates with respect to the MP, as expected.

6.3 The Effect of Differing Alternative Means

In this section, we consider Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, $M = 10,000$ and $\alpha = 0.05$. The alternative means μ_1 and μ_2 are defined such that $(\mu_1 + \mu_2)/2 = 2$, 3, or 4 with $\mu_1 < \mu_2$. For instance, the simulation results presented in Figure 6.4 are plotted such that the x-axis represents $\mu_2 - 2$. That is, a value of 0.5 on the x-axis corresponds to a mixture model where $\mu_1 = 1.5$ and $\mu_2 = 2.5$. Similar x-axes represent $\mu_2 - 3$ and $\mu_2 - 4$ in Figures 6.6 and 6.8, respectively.

6.3.1 Alternative Means Centered at 2

Each plot of Figure 6.4 indicates that the LMCR and Lfdr procedures behave similarly, for each value of $\mu_2 - 2$. In fact, for separation values 0.1 through 1.7, they produce the exact same set of rejections. This is because $A_1 = \emptyset$ since $0.1\phi(x; \mu_1, 1) < \max\{0.8\phi(x; 0, 1),\ 0.1\phi(x; \mu_2, 1)\}$ where $\phi(x; a, b)$ denotes the normal probability density function with mean a and variance b. The equivalence of rejection sets in the two group scenario was verified in Equation (5.1). As an example, consider the

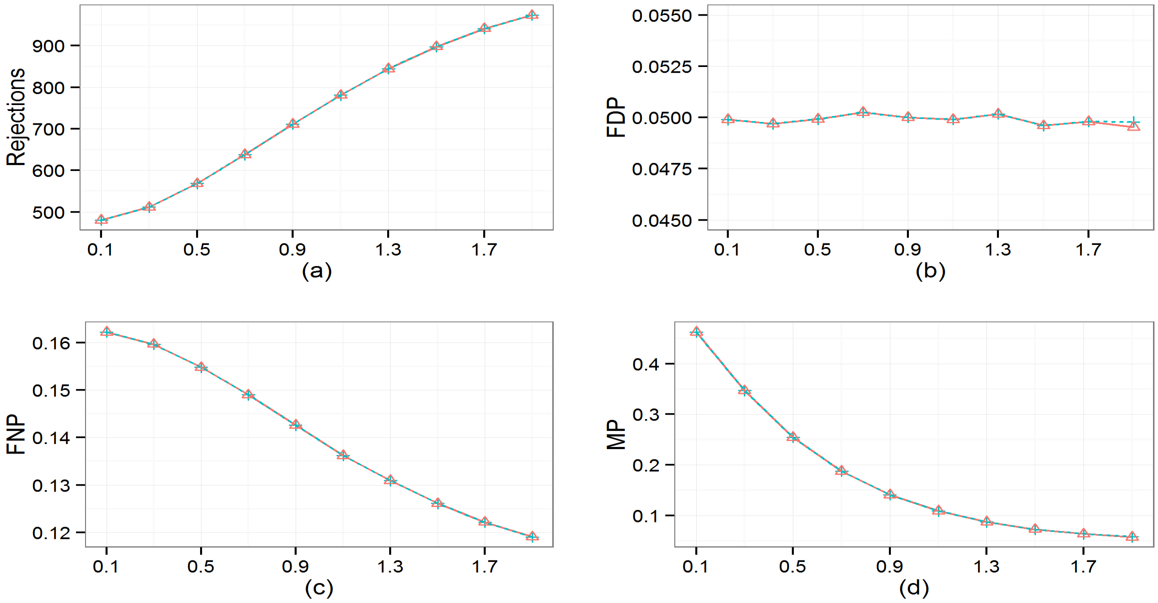

Figure 6.4: The Effect of Alternative Mean Placement when $(\mu_1 + \mu_2)/2 = 2$: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d). In each plot, the x-axis represents $\mu_2 - 2$.

single repetition shown in Figure 6.5 where we have a separation of 0.5. Note that both procedures reject the same 580 hypotheses.

However, there is a caveat that must be mentioned. Upon close examination, we see that there is a slight discrepancy in the FDP values in plot (b) when the separation is 1.9. The Lfdr and LMCR procedures only behave identically when they are acting on the same rejection set. As mentioned above, in cases where the LMCR procedure rejects all hypotheses classified as nonnull without reaching the desired α threshold, the Lfdr procedure may begin rejecting hypotheses classified as null until it reaches that threshold (as seen in Figure 6.2). Though here the difference is quite small, this behavior will be seen repeatedly in the simulations that follow.

The FNP values are again behaving as expected, with greater separation leading to lower false nondiscovery proportions. We see a similar trend in the MP values also. More separation between the null group and the one remaining nonnull group leads to lower misclassification proportions.

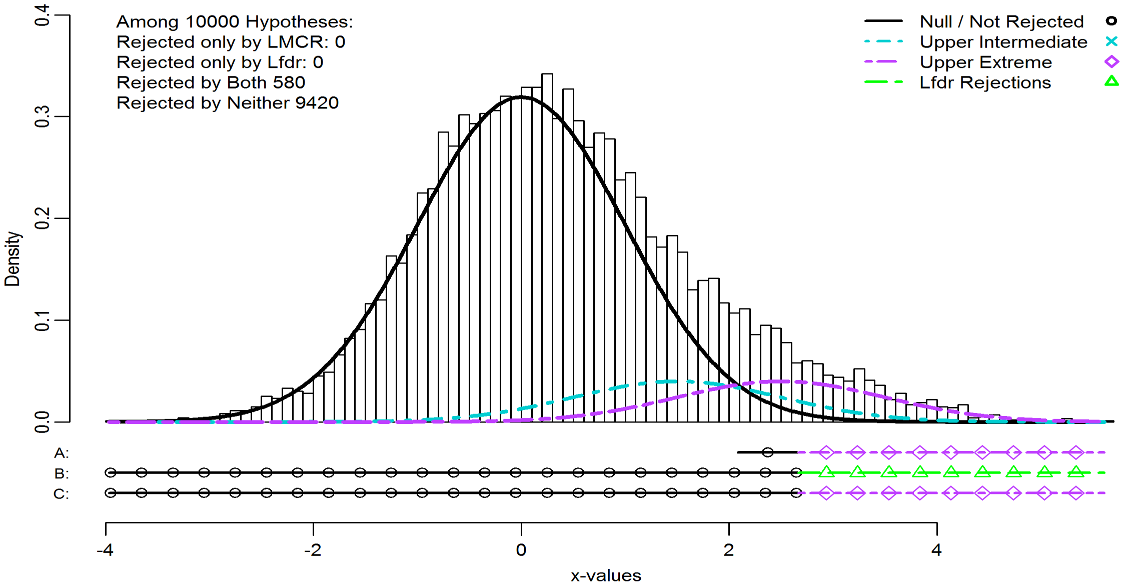

Figure 6.5: The histogram of 10,000 simulated x-values when we have a nonnull mean separation of 0.5 when centered at 2, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

6.3.2 Alternative Means Centered at 3

For Figure 6.6, the top row indicates that the LMCR and Lfdr procedures behave similarly for the first two separation values. In fact, for separation values 0.1 and 0.3, they are almost identical. This can be explained by noting that groups 1 and 2 are very close to one another at those settings. They are close not only to each other, but also to the null group. This situation leads each procedure to select the same rejection set (as seen in Figure 6.7, where the separation is 0.3). That is, despite there being three distinct groups and added misclassification likelihood, only the right tail values are rejected with no gap in the set of rejections for the LMCR procedure. This may seem strange with regard to the LMCR since one may expect the uncertainty of classification near the intersection of the densities for groups 1 and 2 to exclude certain data points. However, because of the orientation of the three groups, these values are more likely to be classified correctly (about 50-50) than points that are closer to the null (where there are three potentially correct groups to choose from rather than just two). With more separation, this is no longer the case. The band of rejected hypotheses

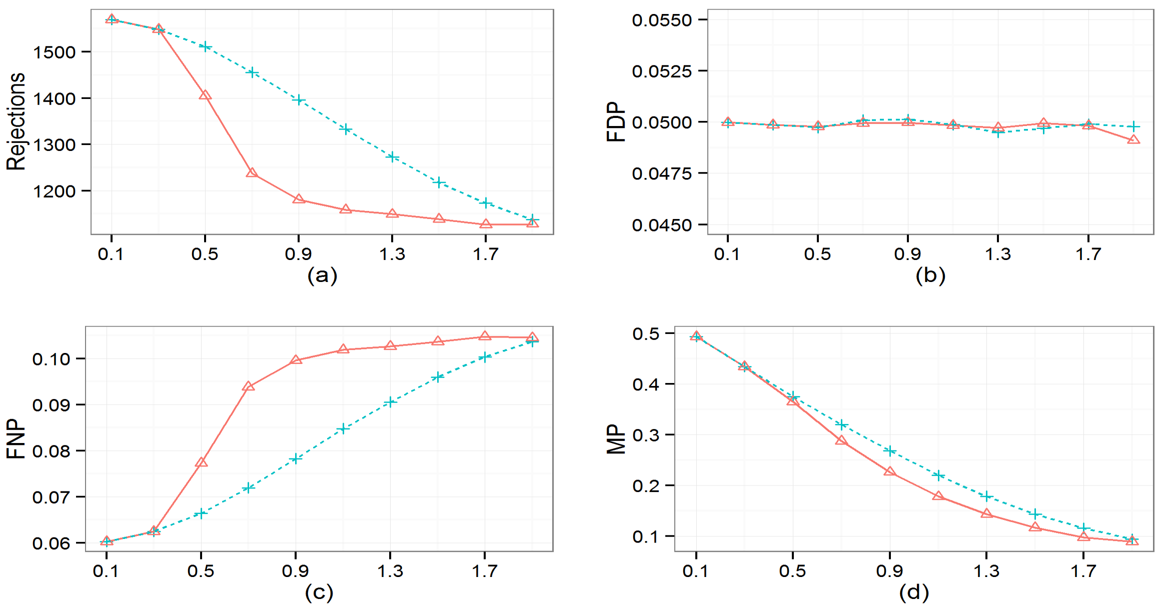

Figure 6.6: The Effect of Alternative Mean Placement when $(\mu_1 + \mu_2)/2 = 3$: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d). In each plot, the x-axis represents $\mu_2 - 3$.

contained in group 1 becomes more prominent, thus, we see a dip in the number of rejections, as well as pronounced separation between the curves of FNP and MP.

Plot (b) shows that both procedures achieve the desired threshold in most instances, though we again see a slight discrepancy for the final separation value. Plot (c) shows that the FNP values have actually changed direction from those in Figure 6.4. Here, as the separation between groups 1 and 2 gets larger, there is more overlap between group 1 and the null group (group 1 is again being absorbed by the other two groups). Thus, we have an increased likelihood of false nondiscoveries for both procedures. In contrast, the FNP of Figure 6.4 decreased, not because group 1 increased its overlap with the null group, but because group 2 started very close to the null and withdrew (note the difference in the magnitude of the FNP for each). Plot (d) decreases with separation, as expected, for the same reasons as discussed above. As group 2 moves further away from group 1, the likelihood of misclassification between those two groups goes down, just as the pool of potential rejections shifts more and more toward group 2.

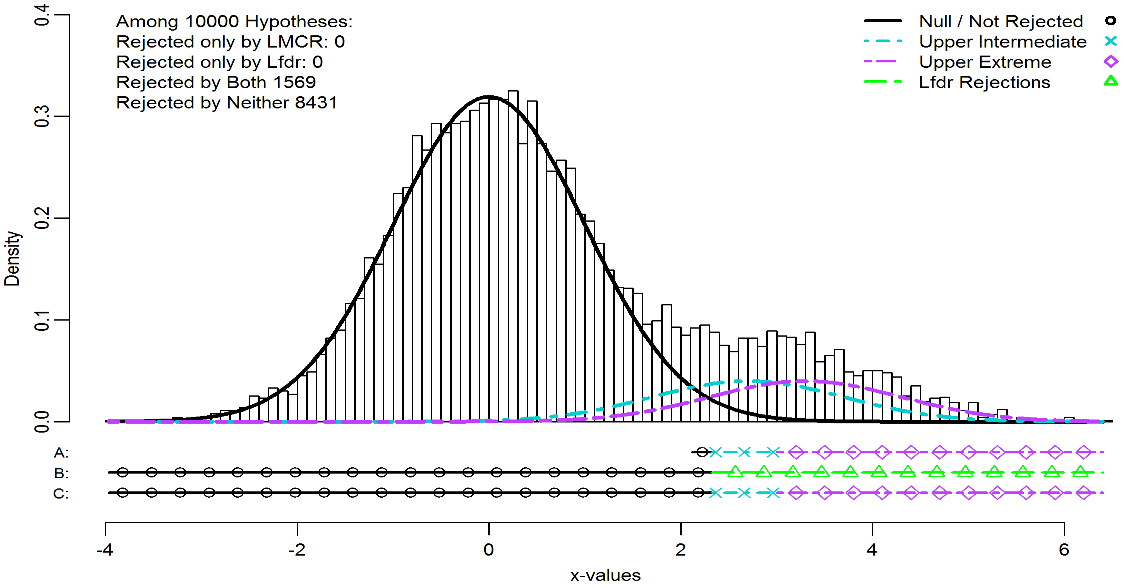

Figure 6.7: The histogram of 10,000 simulated x-values when we have a nonnull mean separation of 0.3 when centered at 3, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

6.3.3 Alternative Means Centered at 4

In Figure 6.8, we see results in plots (a), (c), and (d) similar to those in Figure 6.6, just to a lesser degree. In plot (a) we note that the first few separation settings yield comparable rejection values, but once enough separation is achieved, we begin to see the familiar dip. Again, the effect of the gap in the set of rejected hypotheses for the LMCR procedure is seen in plots (a), (c), and (d).

In plot (b), we see a significant difference between the FDP values for the two procedures. Here, the first four LMCR values are actually exhibiting the safeguard behavior as discussed in Section 6.1, only to a much larger degree. As described above, the Lfdr procedure is not restricted to the set of hypotheses classified as nonnull, so it is more likely to reach the desired α level (as illustrated by Figure 6.2). Once the LMCR procedure exhausts the set of nonnulls, it must stop rejecting hypotheses. Thus, not only can the LMCR procedure fail to reach the α threshold, at times it may not even get particularly close.

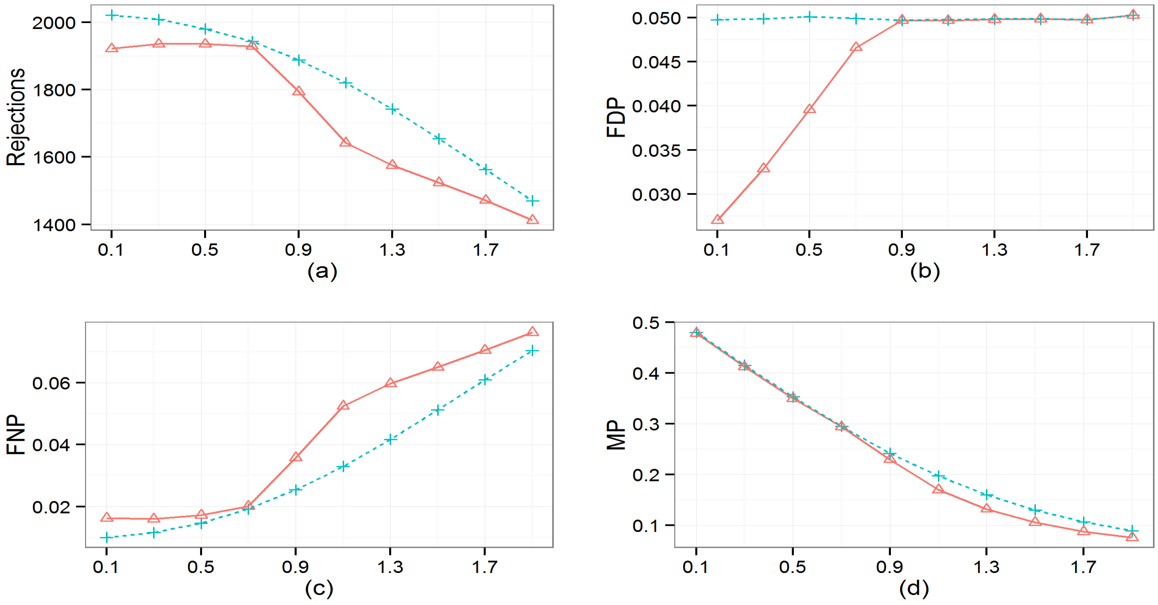

Figure 6.8: The Effect of Alternative Mean Placement when $(\mu_1+\mu_2)/2 = 4$: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d). In each plot, the x-axis represents $\mu_2 - 4$.

6.4 The Effect of Differing Alternative Proportions

In this section, we consider Equation (6.1) where $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, $M = 10,000$ and $\alpha = 0.05$. In this section, the nonnull proportions are defined such that $\pi_1 + \pi_2 = 0.1$, 0.2, and 0.3. For instance, in Figure 6.9 the x-axis represents the proportion of group 2 (our π_2 value) when $\pi_1 + \pi_2 = 0.1$, so that a value of 0.02 on the x-axis corresponds to a mixture model where $\pi_0 = 0.90$, $\pi_1 = 0.08$, and $\pi_2 = 0.02$.

6.4.1 Total Nonnull Contribution = 0.1

Plot (a) of Figure 6.9 shows that, as the value of π_2 increases, the number of rejections for the Lfdr and LMCR procedures begin to converge to one another. This makes intuitive sense, in that, as π_2 increases and π_1 decreases, the contribution of group 1 lessens. Thus, as π_2 approaches 0.1, we are left with the two group case discussed above. Note that when $\pi_2 = 0.005$, there were times when no hypotheses were rejected. Out of the 1000 repetitions, five left us with an empty rejection set. Those trials have been ignored in this analysis.

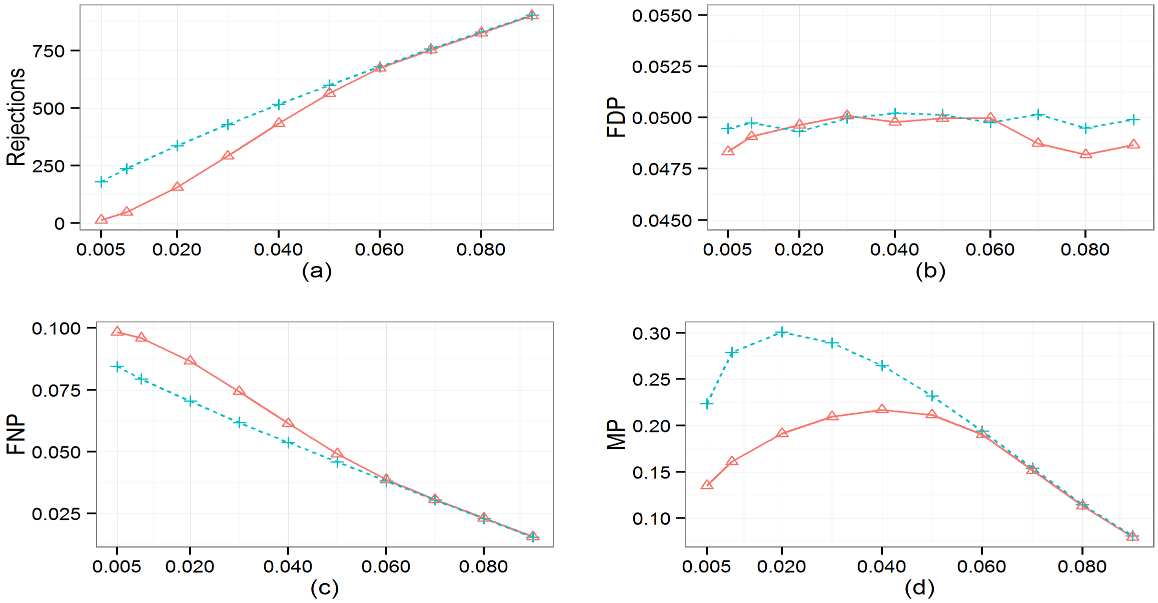

Figure 6.9: The Effect of π_2 when $\pi_1 + \pi_2 = 0.1$: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

Plot (b) seems to show significant disparity between the two procedures with respect to the FDP. However, for small π_2, the differences are again due to the small rejection sets mentioned above. With such a small set of rejections on average (around 14 rejected hypotheses when $\pi_2 = 0.005$), there were many trials where there were no false discoveries at all. When π_2 nears 0.1, the two group case has effectively taken over and the α safeguard behavior becomes apparent. Here, A_1 is empty for large π_2 because $\pi_1\phi(x; 2, 1) < \max\{\pi_0\phi(x; 0, 1),\ \pi_2\phi(x; 4, 1)\}$.

Plot (c) is behaving as expected, in that, as π_1 decreases (since π_2 is increasing), the likelihood of a false nondiscovery should decrease as well, regardless of the procedure used. On the other hand, plot (d) exhibits a somewhat parabolic behavior that we have yet to see. But, if we think about the effect of shifting proportional weight from group 1 to group 2, it is intuitive that the misclassification proportion increases as the groups become more equal. When π_2 is small and π_1 is large, the LMCR is large for larger x-values, say $x > \mu_2$, whereas the Lfdr is small for such values. Consequently, the Lfdr procedure tends to reject hypotheses when x_m is large, despite possible misclassification. Thus, the LMCR procedure has smaller MP, as suggested by Theorem 4.1. As π_2

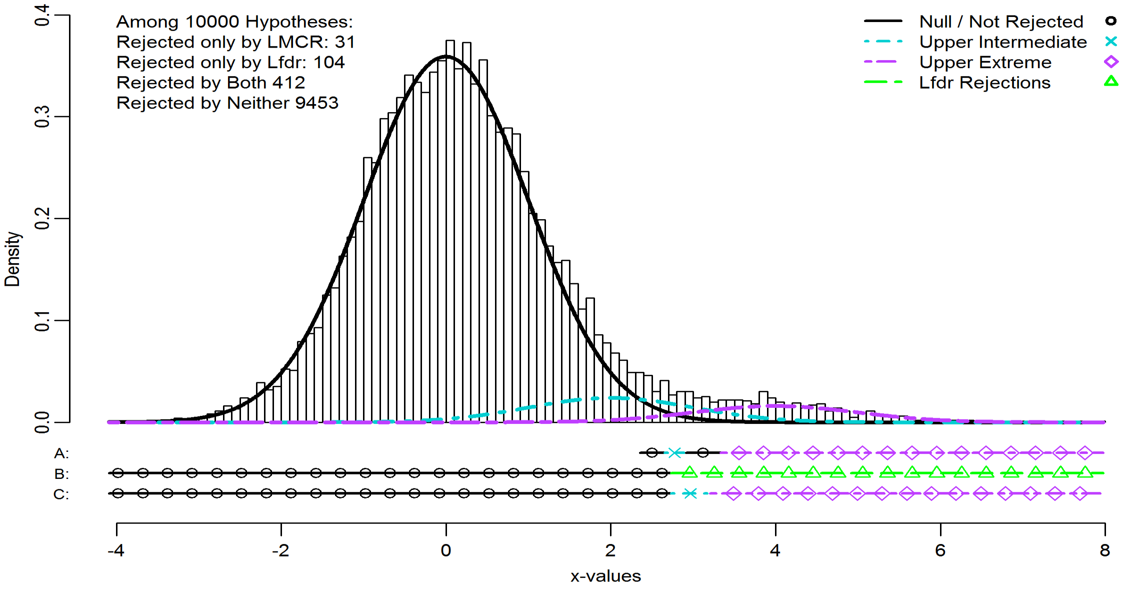

Figure 6.10: The histogram of 10,000 simulated x-values when we have a cumulative nonnull proportion of 0.1 with a π_2 contribution of 0.04, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

increases, the misclassification proportion should decrease as group 2 dominates and the influence of group 1 becomes negligible. The fact that the highest MP value for the LMCR procedure is near $\pi_2 = 0.04$, rather than a 50-50 split near 0.05 is due to the orientation of the groups and the close proximity of group 1 to the null group, as can be seen in Figure 6.10 (where $\pi_2 = 0.04$). Again, recall that the MP contains contributions from misclassified hypotheses that are rejected, regardless of whether $G_m = 0, 1$, or 2.

6.4.2 Total Nonnull Contribution = 0.2

For Figure 6.11, plots (a) and (c) behave just as they did in the previous case. The main difference between these results and the corresponding results of Figure 6.9 is the scale of the y-axis. Note that in each case, the number of rejections and the FNP values have increased as $\pi_1 + \pi_2$ has increased. This is to be expected, since the shift of proportional weight from the null to the set of nonnull groups increases the number of hypotheses likely to be classified as nonnull. Thus, the corresponding number of rejections and likelihood of a false nondiscovery increase.

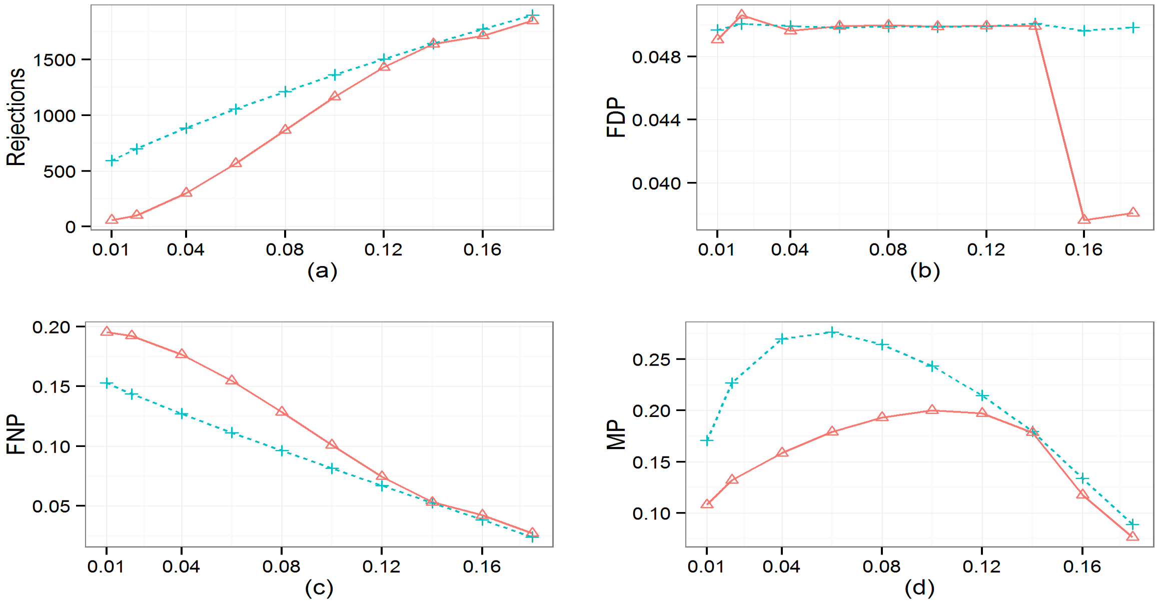

Figure 6.11: The Effect of π_2 when $\pi_1 + \pi_2 = 0.2$: The top row compares the LMCR (△) and Lfdr (+) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

Plot (b) shows us that, as we have seen above, the FDP for both procedures is very similar until we reach the last two separation values. As shown, when $\pi_2 \geq 0.16$ (though likely somewhere between 0.14 and 0.16) we see a familiar dip in the FDP value for the LMCR procedure indicating the α safeguard effect. For instance, when $\pi_2 = 0.16$ (and $\pi_1 = 0.04$), A_1 is empty, as seen in Figure 6.12. We are again left with the two group situation where the set of rejection candidates is exhausted before reaching our desired α threshold (and thus, the Lfdr procedure begins rejecting hypotheses classified as null, 68 this time). As noted in the discussion of Figure 6.8, we see that this problem can be quite severe, in that, the FDP does not reach 0.04, let alone $\alpha = 0.05$. In plot (d), we see the same parabolic shapes as were discussed previously. Again we note that the LMCR procedure dominates the Lfdr procedure, as Theorem 4.1 suggests.

6.4.3 Total Nonnull Contribution = 0.3

Figure 6.13 shows the same general trends as those discussed in the previous subsection. One noticeable difference is the dip in the number of rejections when $\pi_2 = 0.03$. This can be explained by noting that, when $\pi_2 = 0.015$, the contribution of group 2 is largely negligible as compared to

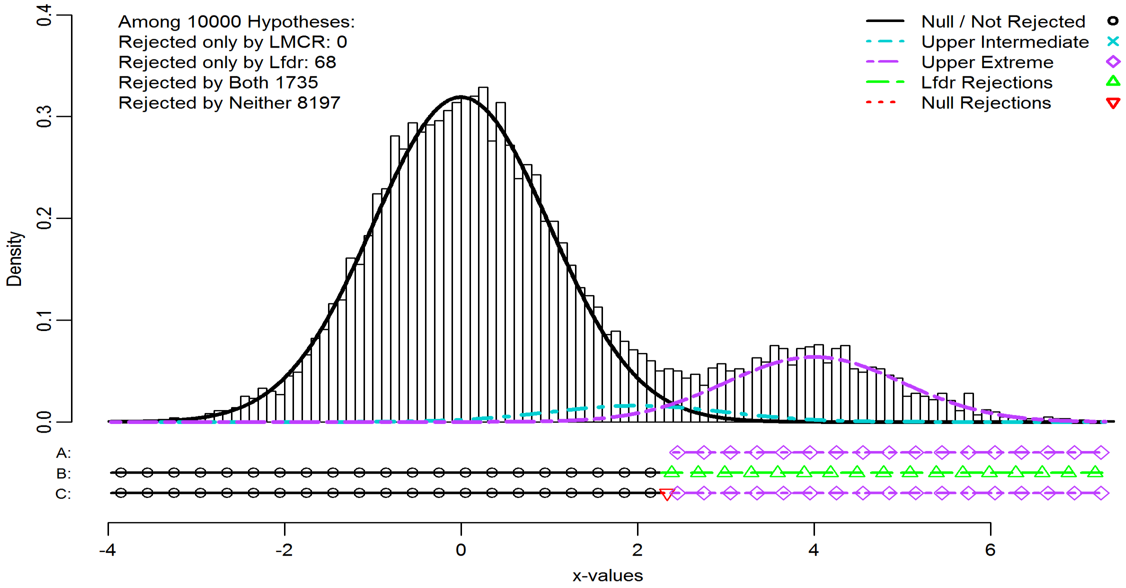

Figure 6.12: The histogram of 10,000 simulated x-values when we have a cumulative nonnull proportion of 0.2 with a π_2 contribution of 0.16, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

the contributions of the null group ($\pi_0 = 0.70$) and group 1 ($\pi_1 = 0.285$). Thus, most of the LMCR rejections will be those hypotheses that are correctly classified as coming from group 1. We see this in Figure 6.14. As π_2 increases to 0.03, as in Figure 6.15, the proportions shift so that, while there still aren't many hypotheses coming from group 2, there are now far fewer rejected hypotheses coming from group 1 (a much narrower band), thus we see a dip. As π_2 increases beyond 0.03, the proportional weight shifts more and more to group 2, so that we see a generally increasing trend similar to the one we have seen in the previous sections. Plots (b), (c), and (d) are very much like those of Figure 6.11 and can be explained in much the same way.

6.5 The Effect of σ

In this section, we consider Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $M = 10,000$ and $\alpha = 0.05$. The σ value will be varied in one of three ways: where $\sigma_0 = \sigma_1 = \sigma_2 = \sigma$, where $\sigma_0 = \sigma_2 = 1$ and σ_1 varies, and where $\sigma_0 = \sigma_1 = 1$ and σ_2 varies. For example, in Figure 6.16, an x-axis value of 0.7 indicates that $\sigma_0 = \sigma_1 = \sigma_2 = 0.7$ whereas in Figure 6.18, an x-axis value of

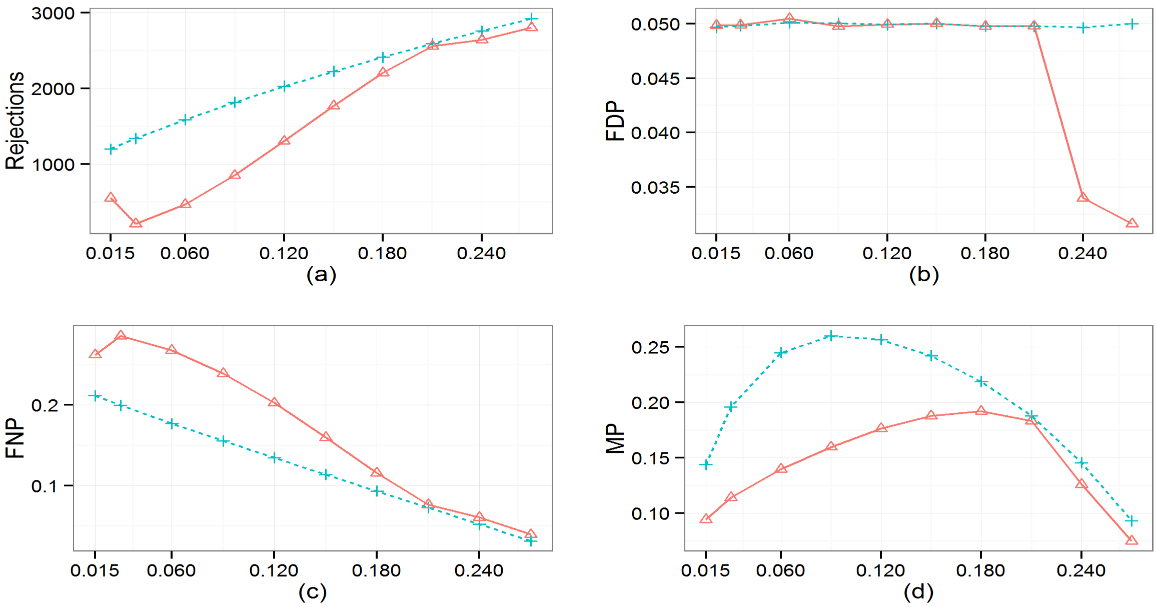

Figure 6.13: The Effect of π_2 when $\pi_1 + \pi_2 = 0.3$: The top row compares the LMCR (△) and Lfdr (+) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

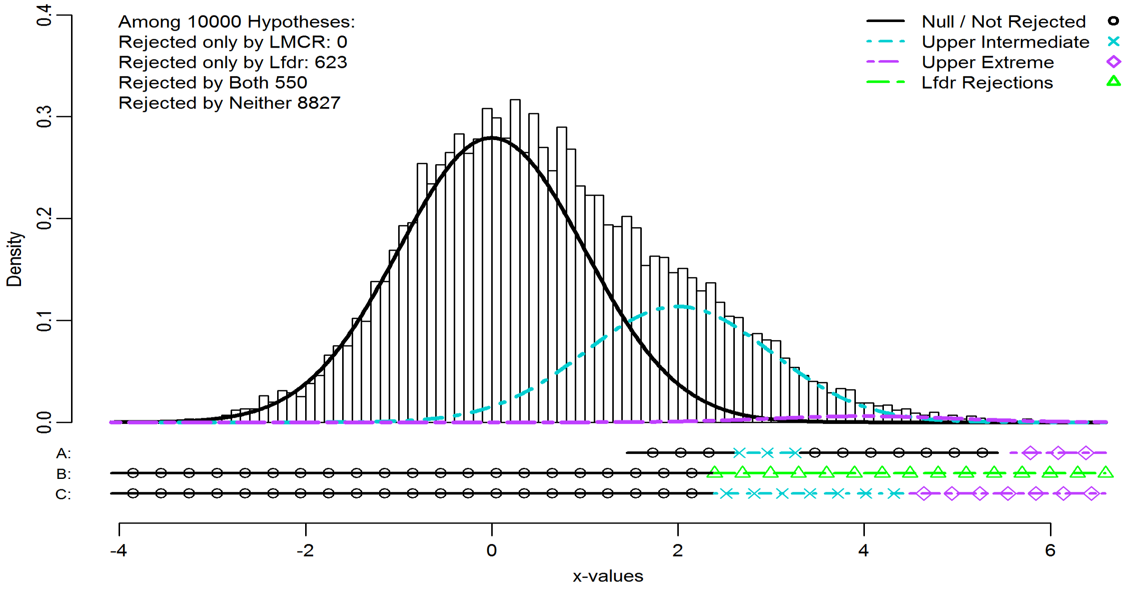

Figure 6.14: The histogram of 10,000 simulated x-values when we have a cumulative nonnull proportion of 0.3 with a π_2 contribution of 0.015, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

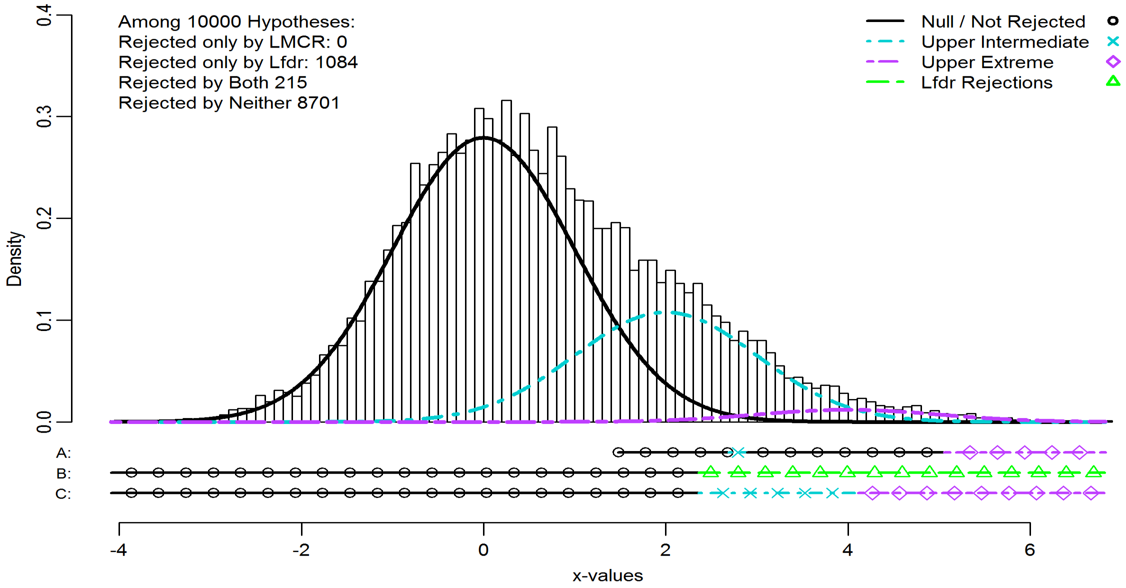

Figure 6.15: The histogram of 10,000 simulated x-values when we have a cumulative nonnull proportion of 0.3 with a π_2 contribution of 0.03, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

0.7 indicates that $\sigma_1 = 0.7$ while $\sigma_0 = \sigma_2 = 1$.

6.5.1 Equal Standard Deviations: $\sigma_0 = \sigma_1 = \sigma_2$

Plots (a) and (c) of Figure 6.16 exhibit patterns similar to those seen in Figures 6.6 and 6.8 with their separation in the middle. Though here we see that in both plots, as the value of σ increases, the number of rejections for the Lfdr and LMCR procedures begin to converge to one another. In fact, for $\sigma \geq 1.3$, the rejection sets are identical and are composed only of hypotheses classified as coming from group 2. This makes intuitive sense, in that, as σ increases, the influence of group 2 increases relative to the contribution of group 1 which is largely being absorbed by the other two groups.

Plot (b) again exhibits the safeguard behavior discussed previously. Here, when $\sigma \leq 0.7$, the LMCR procedure rejects all hypotheses classified as nonnull before reaching the desired α. Figure 6.17 illustrates this issue when $\sigma = 0.5$. Note that the classification regions A_1 and A_2 are well defined because of the orientation of the means. Plot (d), shows a parabolic pattern similar to that

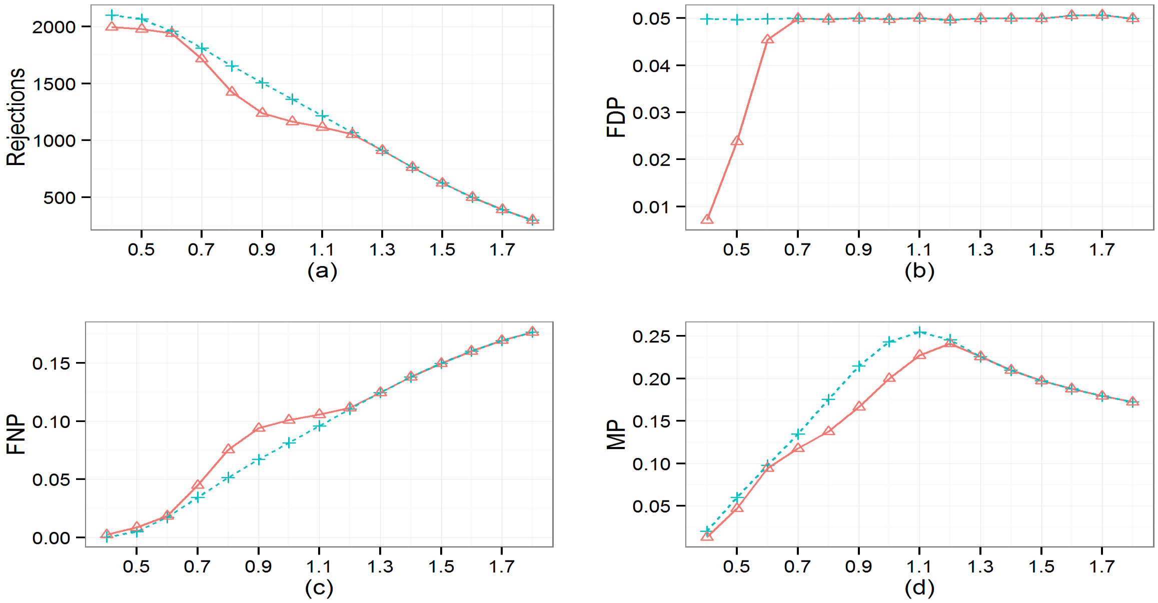

Figure 6.16: The Effect of Equal Standard Deviations: The top row compares the LMCR (△) and Lfdr (+) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

seen in the previous section. Here, when σ is small, the groups are separated enough to avoid major misclassification (as seen in Figure 6.17), but as σ increases the overlap between the groups becomes more severe. When σ is large, A_1 is small since $0.1\phi(x;2,\sigma) < \max\{0.8\phi(x;0,\sigma),\ 0.1\phi(x;4,\sigma)\}$. Once A_1 is empty, the contribution of group 1 to the MP value becomes fixed and the increased spread of the null group and group 2 causes a slight decrease in MP.

6.5.2 The Effect of σ_1

In Figure 6.18, plots (a), (c), and (d) look similar to those of Figure 6.6 only now they are trending in the opposite direction (the number of rejections and the MP are increasing from left to right, while the FNP is decreasing). Plot (b) shows that the desired threshold was achieved for both procedures, as Theorem 3.1 suggests.

Here, the distinct separation in the rejection and FNP curves can be attributed to the location of the means, relative to one another. When σ_1 is less than one, the group 1 density is narrower and more compact, and has a significant influence on the histogram of the test statistics because of μ_1's placement between μ_0 and μ_2. This contribution can be seen in Figure 6.19 where $\sigma_1 = 0.5$.

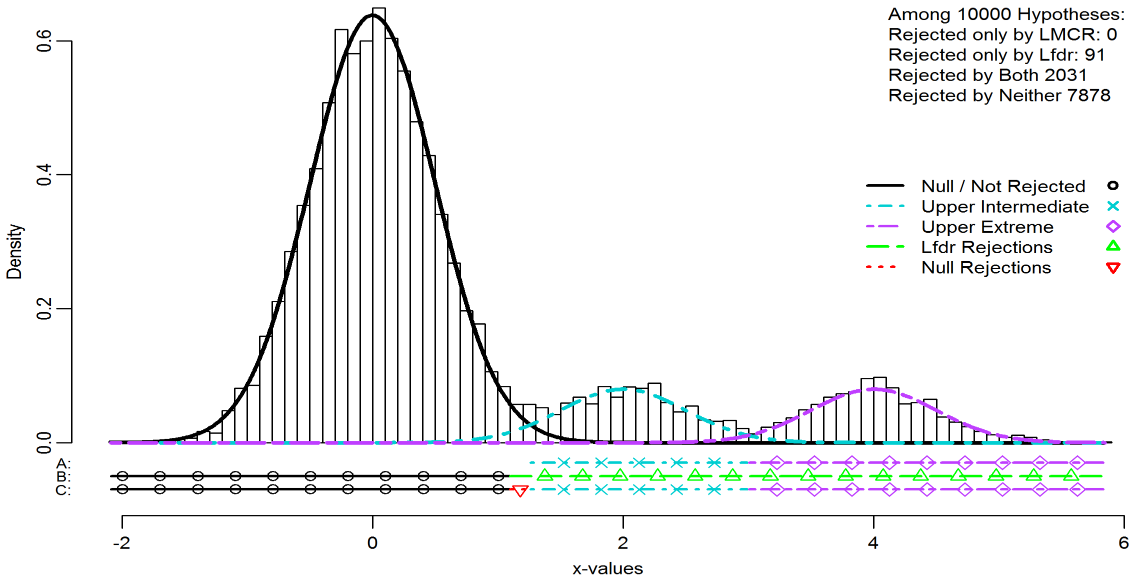

Figure 6.17: The histogram of 10,000 simulated x-values when we have $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

Observe that the intersection of the densities for groups 1 and 2 is near the x-axis (and near the null group density), thus causing the large gap in the LMCR rejection set shown in line A. As σ_1 increases, A_1 shrinks which reduces the gap in the LMCR rejection set, thus increasing the number of rejections and decreasing the FNP.

The MP increases as σ_1 increases because, as the group 1 density flattens (and A_1 shrinks), the overlap between group 1 and the other two groups increases which increases the likelihood of misclassification. For $\sigma_1 > 1.6$, A_1 is essentially empty which leads to the same rejection sets for both the LMCR and Lfdr procedures. This explains the convergence we see in the right tail of each plot.

6.5.3 The Effect of σ_2

The most striking part of Figure 6.20 is the large separation between the LMCR and Lfdr curves in plots (a), (c), and (d). This behavior can be explained by the placement of μ_2 relative to the null group. Regardless of the value of σ_2, since $\mu_2 = 4$ is extreme relative to a null group that is N(0,1),

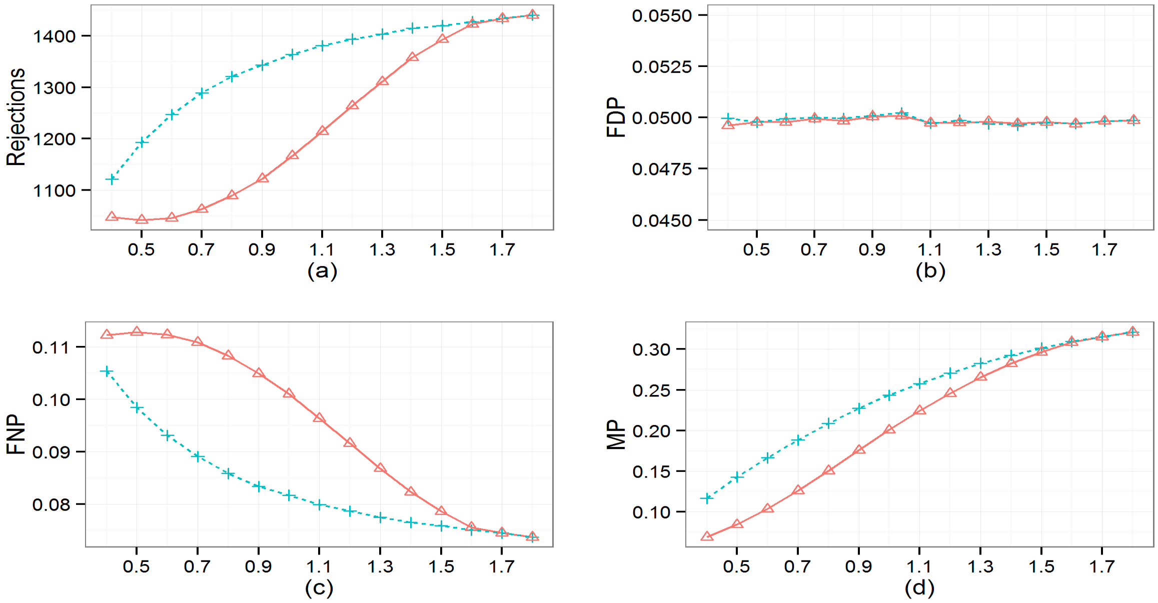

Figure 6.18: The Effect of σ_1: The top row compares the LMCR ($\triangle$) and Lfdr ($+$) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

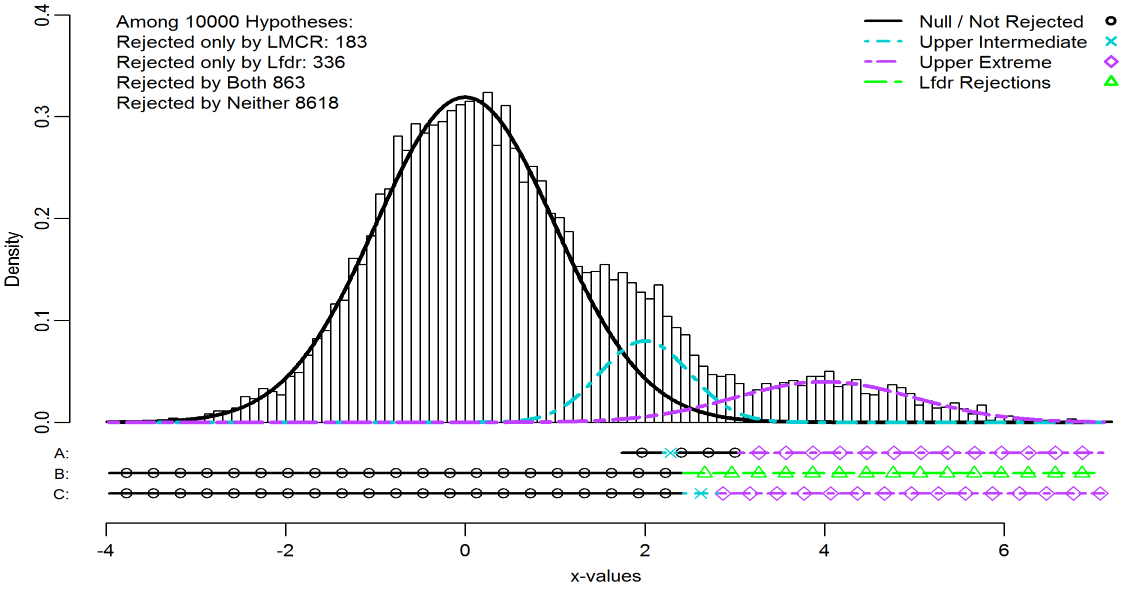

Figure 6.19: The histogram of 10,000 simulated x-values when $\sigma_1 = 0.5$, along with fitted densities. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

the Lfdr procedure will reject all hypotheses classified as coming from group 2 and many that are classified as coming from group 1. However, the rejection set for the LMCR procedure will contain a gap where the group 1 and group 2 densities intersect. Again, the FDP plot shows that the desired α threshold was achieved for each procedure.

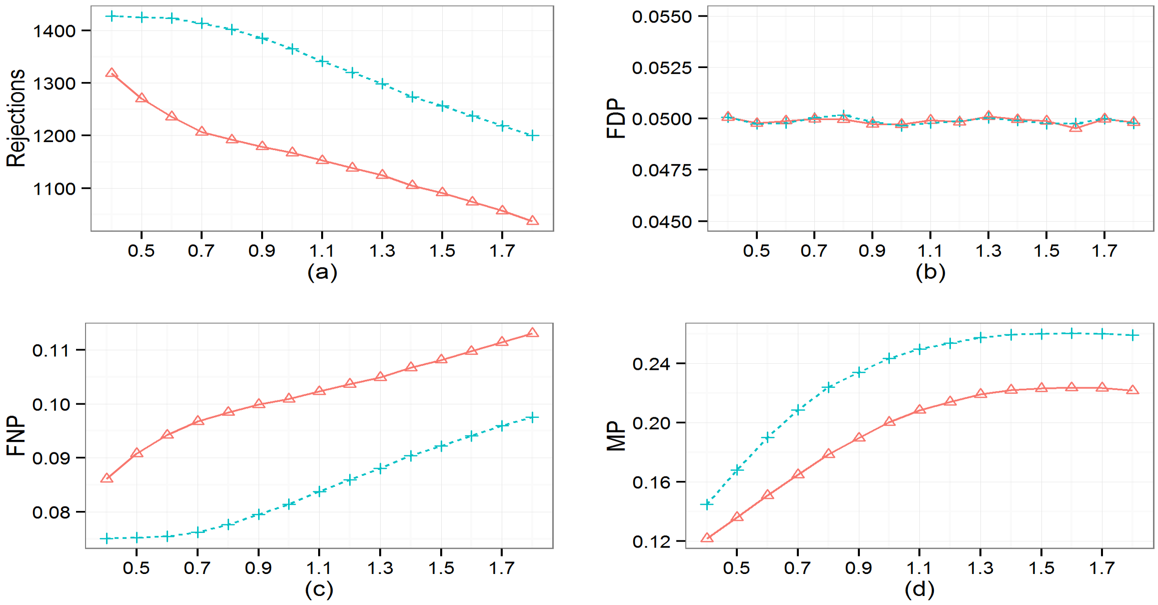

Figure 6.20: The Effect of σ_2: The top row compares the LMCR (△) and Lfdr (+) procedures with respect to the number of rejections (a) and the FDP (b). The bottom row shows the FNP (c) and the MP (d).

As for the trend in plots (a) and (c), as σ_2 increases, A_2 shrinks and the overlap between group 2 and the null group increases, thus reducing the likelihood that a hypothesis is rejected (so there are fewer rejections) despite it being nonnull (so the FNP increases). The MP trend can be explained similarly, in that, as the group 2 density widens, its overlap with the other two groups increases, thus increasing the likelihood of misclassification.

Chapter 7

Performance of the Adaptive LMCR Procedure

Recall that the procedure defined in Section 4.2 is based on classifying data into one of $K + 1$ (a null plus K nonnull) groups. However, in practice, we are often unaware of the true distribution parameters in a particular dataset. Thus, as discussed in Sections 2.4 and 2.5, we are not technically interested in classification, but are really interested in cluster analysis.

Section 7.1 considers the specifics of an EM algorithm applied to a univariate normal mixture. The remainder of this chapter will use the EM algorithm to examine the performance of two forms of the adaptive LMCR procedure (that is, when the distribution parameters are estimated from the data) as applied to Equation (6.1). The first comparison considers the relative performance of the oracle LMCR procedure (when all parameters are known, as examined throughout Chapter 6) versus the theoretical LMCR procedure where $\mu_0 = 0$ and $\sigma_0 = 1$ are defined but all other parameters are estimated. The second comparison considers the relative performance of the theoretical LMCR procedure versus the empirical LMCR procedure where the EM algorithm is used to estimate all necessary parameters. The oracle and adaptive Lfdr procedures are also considered.

For more information on defining adaptive procedures for FDR control see Benjamini and Hochberg (2000); Genovese and Wasserman (2004); Blanchard and Roquain (2009). For more on empirical null distributions see Efron (2004, 2008b, 2010) and for a nonparametric perspective see

Habiger and Peña (2011).

7.1 The EM Algorithm for a Univariate Normal Mixture

In many applications, the mixture model of interest is comprised of univariate (or multivariate) normal distributions, as in Efron (2008a), Cai and Sun (2009), or Efron (2010). For our work, we focus on a univariate mixture where f_k in Equation (2.4) is ϕ_k and $\eta_k = (\mu_k, \sigma_k^2)$ for $k \in \mathcal{K} = \{0, 1, ..., K\}$. Here ϕ_k is

$$\phi_k(x_m|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x_m - \mu_k)^2}{2\sigma_k^2}\right\}.$$

(For details of the multivariate case see Fraley and Raftery (2002)).

Now, from Equation (2.8), the E-step of the EM algorithm considers the group membership $y_{mk} = I(z_m \text{ is from group } k)$ and is given by

$$\hat{y}_{mk} = \frac{\hat{\pi}_k \phi_k(x_m|\hat{\mu}_k, \hat{\sigma}_k^2)}{\sum_{k\in\mathcal{K}} \hat{\pi}_k \phi_k(x_m|\hat{\mu}_k, \hat{\sigma}_k^2)}.$$

The M-step is determined by the complete data log-likelihood, which recall from Equation (2.7) is

$$\begin{aligned}
l(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{y}|\boldsymbol{z}) &= \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \log[\pi_k \phi_k(x_m|\mu_k, \sigma_k^2)] \\
&= \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \log\left[\frac{\pi_k}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x_m - \mu_k)^2}{2\sigma_k^2}\right\}\right] \\
&= \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \left[\log\left(\frac{\pi_k}{\sqrt{2\pi\sigma_k^2}}\right) - \frac{(x_m - \mu_k)^2}{2\sigma_k^2}\right] \qquad (7.1) \\
&= \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \log\left(\frac{\pi_k}{\sqrt{2\pi\sigma_k^2}}\right) - \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \frac{(x_m - \mu_k)^2}{2\sigma_k^2} \\
&= \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \log(\pi_k) - \frac{1}{2} \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \log(2\pi\sigma_k^2) - \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \frac{(x_m - \mu_k)^2}{2\sigma_k^2},
\end{aligned}$$

where $\boldsymbol{\eta} = (\eta_k,\ k \in \mathcal{K})$, $\boldsymbol{\pi} = (\pi_k,\ k \in \mathcal{K})$ is the vector of mixing proportions, $\boldsymbol{y} = (\boldsymbol{y}_m,\ m \in \mathcal{M})$ for $\boldsymbol{y}_m = (y_{mk},\ k \in \mathcal{K})$, and $\boldsymbol{z} = (z_m,\ m \in \mathcal{M})$ for $z_m = (x_m, \boldsymbol{y}_m)$.

Differentiating Equation (7.1) with respect to μ_k, for each $k \in \mathcal{K}$, gives

$$\begin{aligned}
\frac{\partial \; l(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{y}|\boldsymbol{z})}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k}\left[-\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} y_{mk}\frac{(x_m-\mu_k)^2}{2\sigma_k^2}\right] \\
&= \frac{\partial}{\partial \mu_k}\left[-\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} y_{mk}\frac{x_m^2-2x_m\mu_k+\mu_k^2}{2\sigma_k^2}\right] \\
&= -\frac{1}{2\sigma_k^2}\sum_{m\in\mathcal{M}} y_{mk}\frac{\partial}{\partial \mu_k}[x_m^2-2x_m\mu_k+\mu_k^2] \\
&= -\frac{1}{2\sigma_k^2}\sum_{m\in\mathcal{M}} y_{mk}[-2x_m+2\mu_k] \\
&= -\frac{1}{\sigma_k^2}\sum_{m\in\mathcal{M}} y_{mk}[-x_m+\mu_k].
\end{aligned}$$

Setting this equal to 0 gives

$$\begin{aligned}
\hat{\mu}_k &= \frac{\sum_{m\in\mathcal{M}} y_{mk}x_m}{\sum_{m\in\mathcal{M}} y_{mk}} \\
&= \frac{\sum_{m\in\mathcal{M}} y_{mk}x_m}{M_k}
\end{aligned}$$

where M_k is the estimated number of observations in group k.

Differentiating Equation (7.1) with respect to σ_k^2, for each $k \in \mathcal{K}$, gives

$$\begin{aligned}
\frac{\partial \; l(\boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{y}|\boldsymbol{z})}{\partial \sigma_k^2} &= \frac{\partial}{\partial \sigma_k^2}\left[-\frac{1}{2}\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} y_{mk}\log(2\pi\sigma_k^2) - \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} y_{mk}\frac{(x_m-\mu_k)^2}{2\sigma_k^2}\right] \\
&= \frac{\partial}{\partial \sigma_k^2}\left[-\frac{1}{2}\sum_{m\in\mathcal{M}} y_{mk}\log(2\pi\sigma_k^2) - \sum_{m\in\mathcal{M}} y_{mk}\frac{(x_m-\mu_k)^2}{2\sigma_k^2}\right] \\
&= -\frac{1}{2}\sum_{m\in\mathcal{M}} y_{mk}\frac{\partial}{\partial \sigma_k^2}\left[\log(\sigma_k^2) + \frac{(x_m-\mu_k)^2}{\sigma_k^2}\right] \\
&= -\frac{1}{2}\sum_{m\in\mathcal{M}} y_{mk}\left[\frac{1}{\sigma_k^2} + \frac{-(x_m-\mu_k)^2}{(\sigma_k^2)^2}\right] \\
&= -\frac{1}{2(\sigma_k^2)^2}\sum_{m\in\mathcal{M}} y_{mk}\left[\sigma_k^2 - (x_m-\mu_k)^2\right].
\end{aligned}$$

Setting this equal to 0 gives

$$\begin{aligned}\hat{\sigma}_k^2 &= \frac{\sum_{m\in\mathcal{M}} y_{mk}(x_m-\mu_k)^2}{\sum_{m\in\mathcal{M}} y_{mk}} \\ &= \frac{\sum_{m\in\mathcal{M}} y_{mk}(x_m-\mu_k)^2}{M_k}.\end{aligned}$$

Differentiating Equation (7.1) with respect to π_k is a little more difficult because of the added constraint that $\sum_{k\in\mathcal{K}} \pi_k = 1$. For this part, we need to apply the method of Lagrange Multipliers. That is, we look to simultaneously solve the system of $K+1$ equations defined by

$$\begin{gathered}\bigtriangledown l(\boldsymbol{\eta},\boldsymbol{\pi},\boldsymbol{y}|\boldsymbol{z}) = \lambda \bigtriangledown g(\boldsymbol{\pi}) \\ g(\boldsymbol{\pi}) = 1,\end{gathered} \tag{7.2}$$

where $\bigtriangledown l(\boldsymbol{\eta},\boldsymbol{\pi},\boldsymbol{y}|\boldsymbol{z})$ is from Equation (7.1) and $g(\boldsymbol{\pi}) = \sum_{k\in\mathcal{K}} \pi_k$. For each $k \in \mathcal{K}$, the first portion of Equation (7.2) becomes

$$\begin{aligned} & \frac{\partial}{\partial \pi_k} \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} y_{mk} \log(\pi_k) && = \lambda \\ \Rightarrow \quad & \sum_{m\in\mathcal{M}} y_{mk} \frac{\partial}{\partial \pi_k} \log(\pi_k) && = \lambda \\ \Rightarrow \quad & \sum_{m\in\mathcal{M}} \frac{y_{mk}}{\pi_k} && = \lambda \\ \Rightarrow \quad & \sum_{m\in\mathcal{M}} \frac{y_{mk}}{\lambda} && = \pi_k. \end{aligned}$$

Substituting these values into the constraint (the second portion of Equation (7.2)), we have

$$\begin{aligned} & \sum_{k\in\mathcal{K}} \sum_{m\in\mathcal{M}} \frac{y_{mk}}{\lambda} && = 1 \\ \Rightarrow \quad & \sum_{k\in\mathcal{K}} \sum_{m\in\mathcal{M}} y_{mk} && = \lambda \\ \Rightarrow \quad & M && = \lambda. \end{aligned}$$

Substitution of the λ value into each π_k gives the estimates

$$\hat{\pi}_k = \frac{\sum_{m \in \mathcal{M}} y_{mk}}{M}$$

for each k. These estimates make intuitive sense because they represent the proportion of attributes in group k.

Putting this all together, the EM algorithm is composed of the following:

1. Initialize $\hat{y}_{mk}$ (perhaps by using an agglomerative clustering technique).

2. Use the current estimates of $\hat{y}_{mk}$ to compute the MLEs of $\hat{\pi}_k$, $\hat{\mu}_k$, and $\hat{\sigma}_k^2$. (the M-step)

3. Use the estimates from the M-step to compute an updated $\hat{y}_{mk}$. (the E-step)

4. Repeat steps 2 and 3 until some convergence criterion (often a maximum number of iterations or a specified tolerance) is satisfied.

7.2 Numerical Comparisons of Adaptive Procedures

This section examines the asymptotic performance of the adaptive LMCR and Lfdr procedures as the number of hypotheses increases. The results below follow the same structure as those seen in Chapter 6, in that, data are generated from a non-symmetric three group normal mixture model as defined in Equation (6.1). While $1,000$ repetitions of each scenario were generated, in some situations the `normalmixEM` function (discussed in Section 5.2) failed to converge. The following results present the average number of rejections, FDP, FNP, and MP for the remaining iterations.

7.2.1 Examples of Rapid Convergence

We begin by considering Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$ and with $\alpha = 0.05$. The comparison of the oracle LMCR and Lfdr procedures versus the theoretical LMCR and Lfdr procedures are presented in Figure 7.1. Figure 7.2 presents the results of the theoretical versus empirical comparisons.

Table 7.1 shows the number of usable repetitions represented in each figure. Here, a repetition is deemed unusable if the EM algorithm did not converge or if the convergent algorithm attempted to fit a model where $\mu_2 < \mu_1$. Note that, as M increases, the number of usable repetitions steadily increases for the theoretical case whereas the empirical scenario shows much more variability.

Procedure	Number of Usable Repetitions								
Theoretical	744	819	867	864	921	954	982	999	998
Empirical	897	930	930	937	924	926	893	864	840
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.1: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

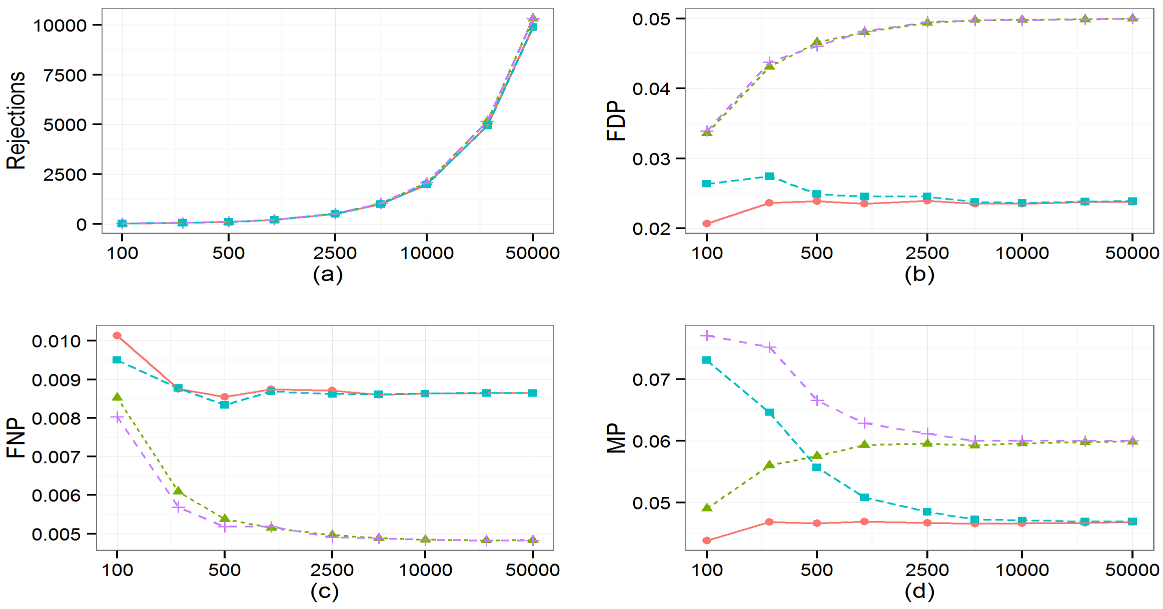

Figure 7.1: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The oracle LMCR (○), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

In Figure 7.1, we see that each theoretical procedure's values converge to the corresponding oracle values quite quickly. By the time $M = 5,000$, there is a slight difference between the theoretical and oracle results for the MP, but virtually no difference in the number of rejections, FDP, or FNP. We see similar results in Figure 7.2, only now the empirical procedures' values seem to have converged

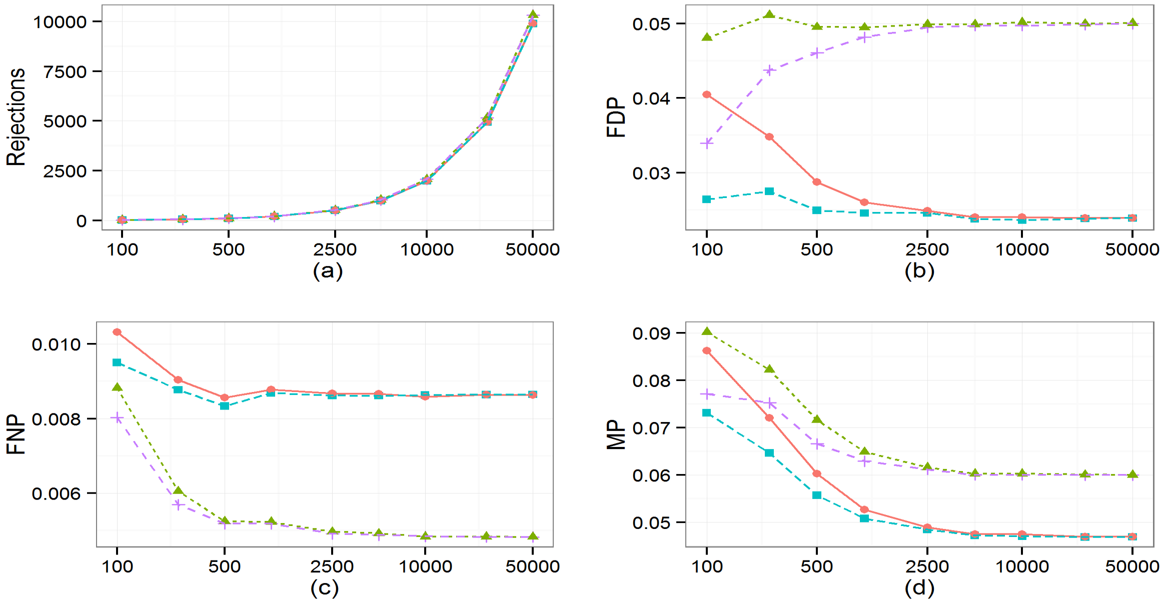

Figure 7.2: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

in each plot by the time $M = 2,500$. As seen in Table 7.1, at least 745 repetitions were used to calculate the average number of rejections, FDP, FNP, and MP.

Now consider Equation (6.1) with $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 4$, $\mu_2 = 8$, $\sigma_0 = \sigma_2 = 1$, $\sigma_1 = 0.5$, and with $\alpha = 0.05$. The comparison of the oracle LMCR and Lfdr procedures versus the theoretical LMCR and Lfdr procedures are presented in Figure 7.3. Figure 7.4 shows the results of the theoretical versus empirical comparisons. Table 7.2 provides the number of usable repetitions represented in each figure, where we see that as M increases, the number of usable repetitions increases for both cases.

Procedure	Number of Usable Repetitions								
Theoretical	914	955	977	988	1000	1000	1000	1000	1000
Empirical	981	992	997	1000	1000	1000	1000	1000	1000
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.2: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_2 = 1$, $\sigma_1 = 0.5$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

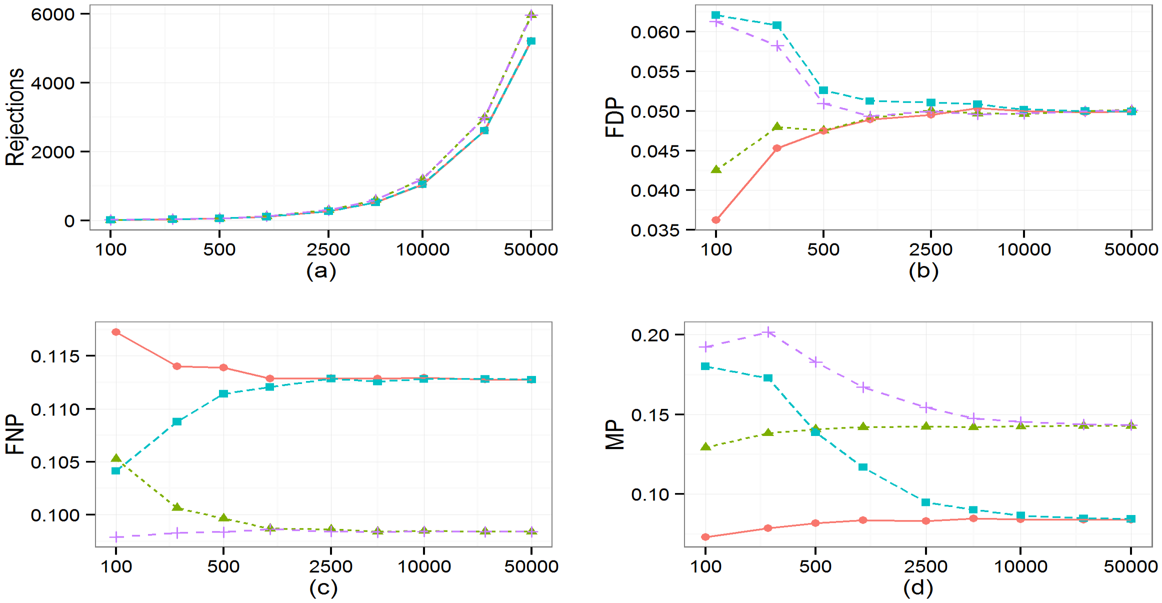

Figure 7.3: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_2 = 1$, $\sigma_1 = 0.5$, and $\alpha = 0.05$: The oracle LMCR (◯), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

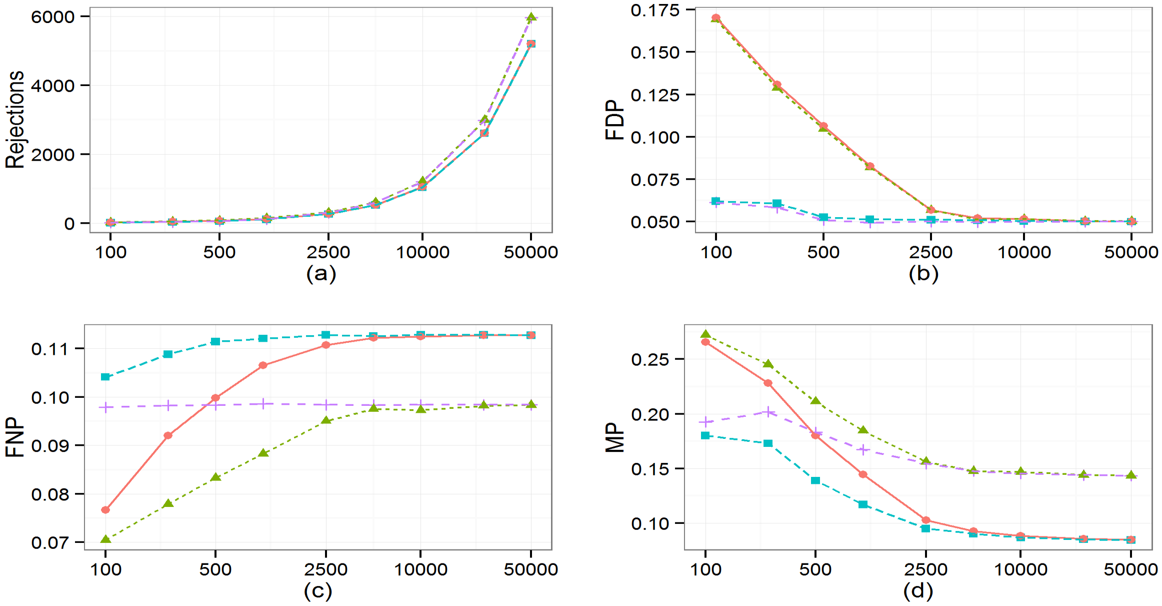

Figure 7.4: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_2 = 1$, $\sigma_1 = 0.5$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

As seen in Figures 7.1 and 7.2 above, Figures 7.3 and 7.4 show that convergence occurs by the time $M = 10,000$. Again, from Table 7.2 we see that at least 914 repetitions were used to calculate the average number of rejections, FDP, FNP, and MP in each plot.

7.2.2 Examples of Slow Convergence

Consider Equation (6.1) with $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 1.5$, $\mu_2 = 2.5$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. The comparison of the oracle LMCR and Lfdr procedures versus the theoretical LMCR and Lfdr procedures are presented in Figure 7.5. Figure 7.6 shows the results of the theoretical versus empirical comparison. Table 7.3 provides the number of usable repetitions represented in each figure.

Unlike the previous two simulations, which showed relatively quick convergence, this simulation setting exhibits interesting behavior, particularly in Table 7.3 and Figure 7.5. Observe that the number of usable repetitions increases then decreases for both the theoretical and empirical cases. Of particular interest are the respective lows of 459 and 333 for the $M = 50,000$ setting. Such comparatively small numbers stem from the difficulty of classification when μ_1 is so close to μ_0 and M is so large. For reference, see Figure 6.5 which presents a histogram for this simulation setting when $M = 10,000$.

Procedure	Number of Usable Repetitions								
Theoretical	494	583	597	623	641	669	641	546	459
Empirical	788	784	790	797	743	721	623	450	333
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.3: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 1.5$, $\mu_2 = 2.5$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

In Figure 7.5 we see that the number of rejections is fairly close for each of the four procedures presented, though the theoretical LMCR procedure seems to have the fewest. The FDP plot shows that all procedures converge to the desired $\alpha = 0.05$ level by the time $M = 2,500$. However, this behavior is not surprising since FDR control is incorporated into each procedure's definition. The FNP plot is much more interesting, particularly for the theoretical LMCR procedure. For $M = 250$

to $M = 10,000$ the theoretical LMCR procedure's FNP values are distinctly larger than those of the other three procedures (though the FNP scale mitigates this difference somewhat). These differences can be attributed to the relatively small number of rejections for many of the repetitions.

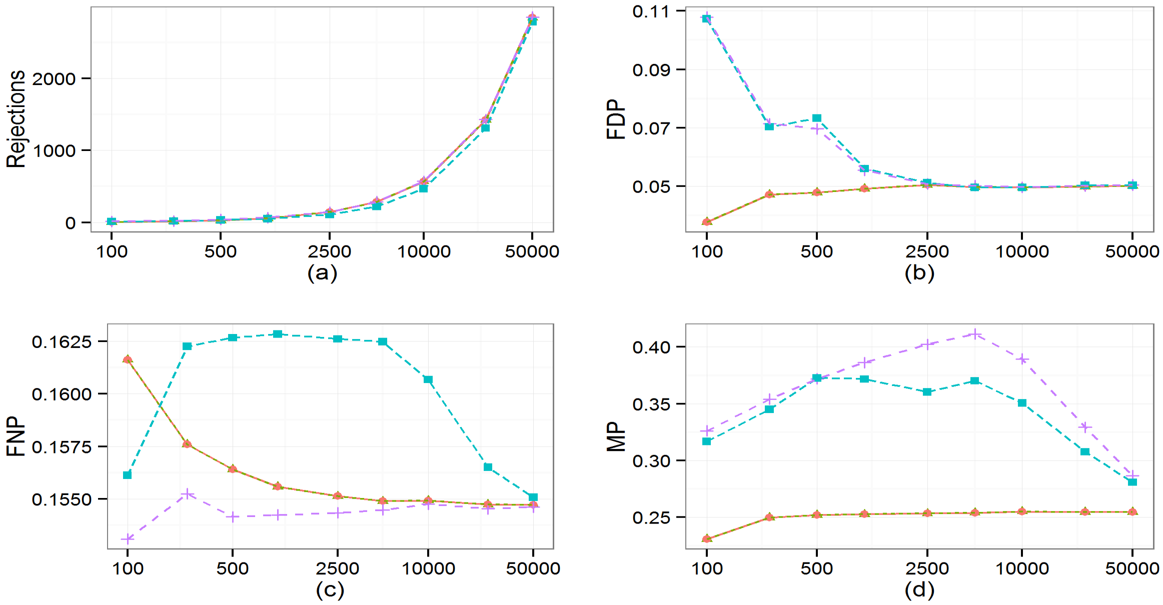

Figure 7.5: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 1.5$, $\mu_2 = 2.5$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The oracle LMCR (◯), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

For example, one particular run of the $M = 2,500$ setting had the following estimates for the theoretical case: $\hat{\pi}_0 = 0.78$, $\hat{\pi}_1 = 0.21$, $\hat{\pi}_2 = 0.01$, $\hat{\mu}_1 = 1.87$, $\hat{\mu}_2 = 3.41$, $\hat{\sigma}_1 = 1.06$, and $\hat{\sigma}_2 = 0.36$. Based on these values, there were only eight LMCR rejections (and only 169 Lfdr rejections), which led to an inflated FNR. To more clearly see the influence of parameter estimation, contrast these results with those of Figure 6.4, which showed that the oracle LMCR and oracle Lfdr procedures behaved identically for this particular scenario since $A_1 = \emptyset$. By the time $M = 50,000$, convergence is very close for the FNP, but still in progress for the MP. However, observe that the theoretical LMCR procedure's MP is less than the theoretical Lfdr's MP for most M values, as suggested by Theorem 4.1.

As for Figure 7.6, the theoretical and empirical procedures are converging, though there is still some separation even when $M = 50,000$. In practice, situations like these may require very large

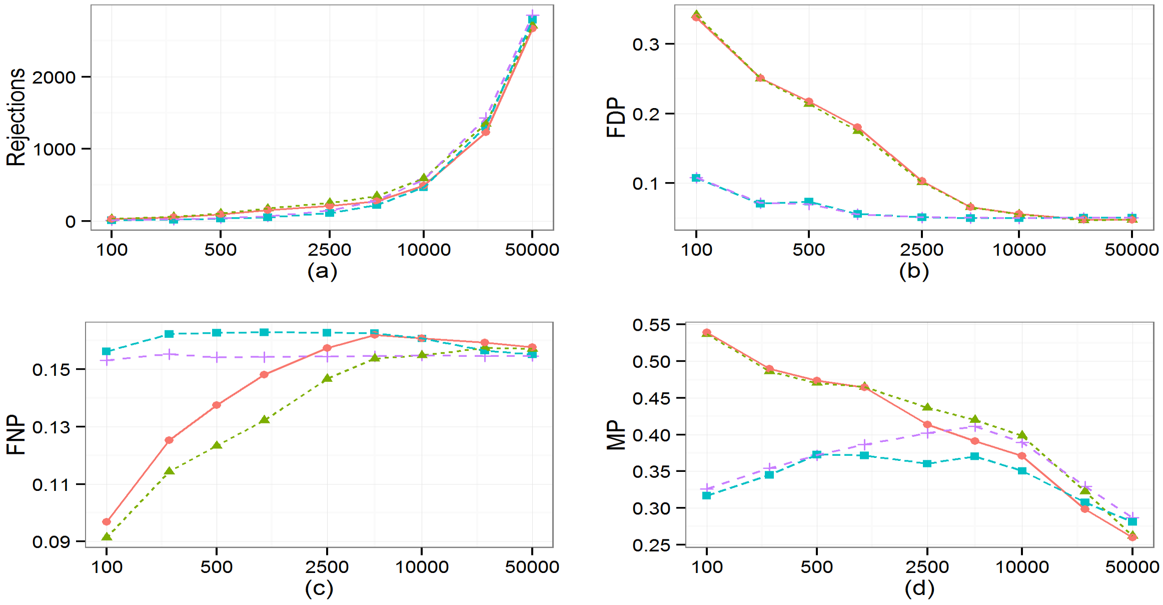

Figure 7.6: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 1.5$, $\mu_2 = 2.5$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

samples to get accurate results. When such large samples ($M = 50,000+$) are not possible, a different model may need to be considered.

Now consider Equation (6.1) with $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2.7$, $\mu_2 = 3.3$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. Table 7.4 gives the number of usable repetitions for this setting while Figures 7.7 and 7.8 provide the results of the theoretical versus oracle and theoretical versus empirical comparisons, respectively.

Procedure	Number of Usable Repetitions								
Theoretical	752	780	797	768	770	719	672	502	427
Empirical	910	899	923	874	841	786	696	545	460
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.4: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2.7$, $\mu_2 = 3.3$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

Table 7.4 shows results similar to those presented in Table 7.3, only now the number of usable repetitions is somewhat larger for each M level. That said, observe that these values still trend downward as M increases. As before, these relatively low numbers (less than half of the original $1,000$ repetitions) can be attributed to the difficulty in estimating parameters for distributions that are so close together when M is large. For reference, consider the histogram in Figure 6.7 when $M = 10,000$.

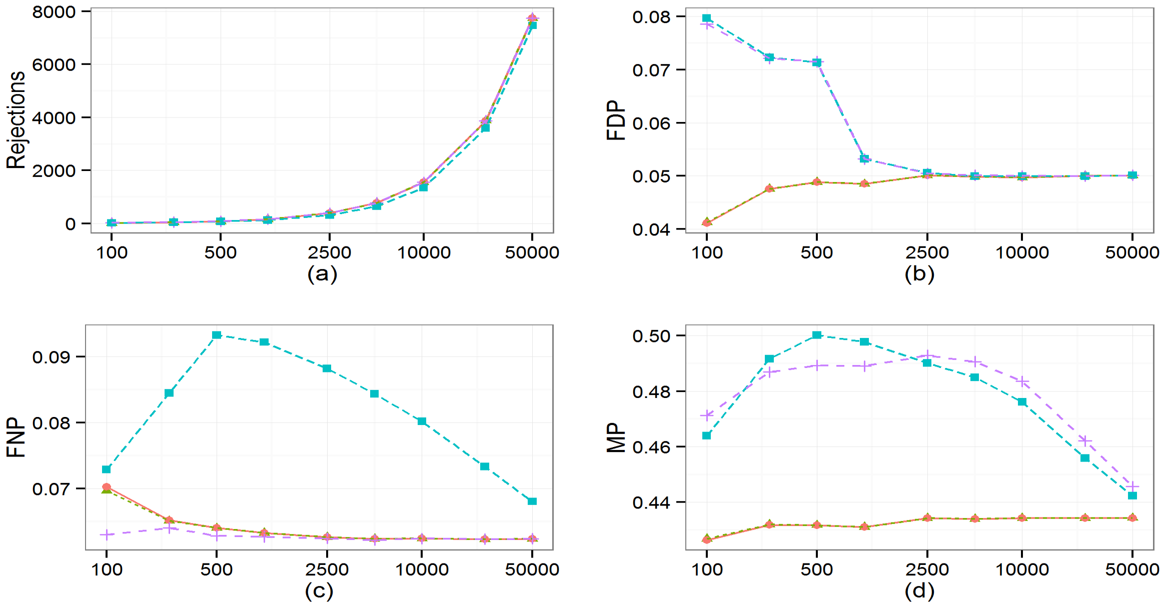

Figure 7.7: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2.7$, $\mu_2 = 3.3$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The oracle LMCR (◯), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

Figure 7.7 is also similar to Figure 7.5, only now the unusual theoretical LMCR procedure's FNP values are not as close to the oracle FNP when $M = 50,000$. Here, the large FNP values can be attributed to the same behavior as before. For example, for one run with $M = 500$ we had only three LMCR rejections with the following estimates: $\hat{\pi}_0 = 0.77$, $\hat{\pi}_1 = 0.18$, $\hat{\pi}_2 = 0.05$, $\hat{\mu}_1 = 2.60$, $\hat{\mu}_2 = 3.54$, $\hat{\sigma}_1 = 0.97$, and $\hat{\sigma}_2 = 0.39$. Multiple repetitions with similar values again resulted in an inflated FNP. As for Figure 7.8, the MP values have mostly converged by the time $M = 25,000$, whereas the FNP values are still in the process of converging. As mentioned above, in practice, reliable results may require very large samples or, in some instances, adjustments to the model being considered.

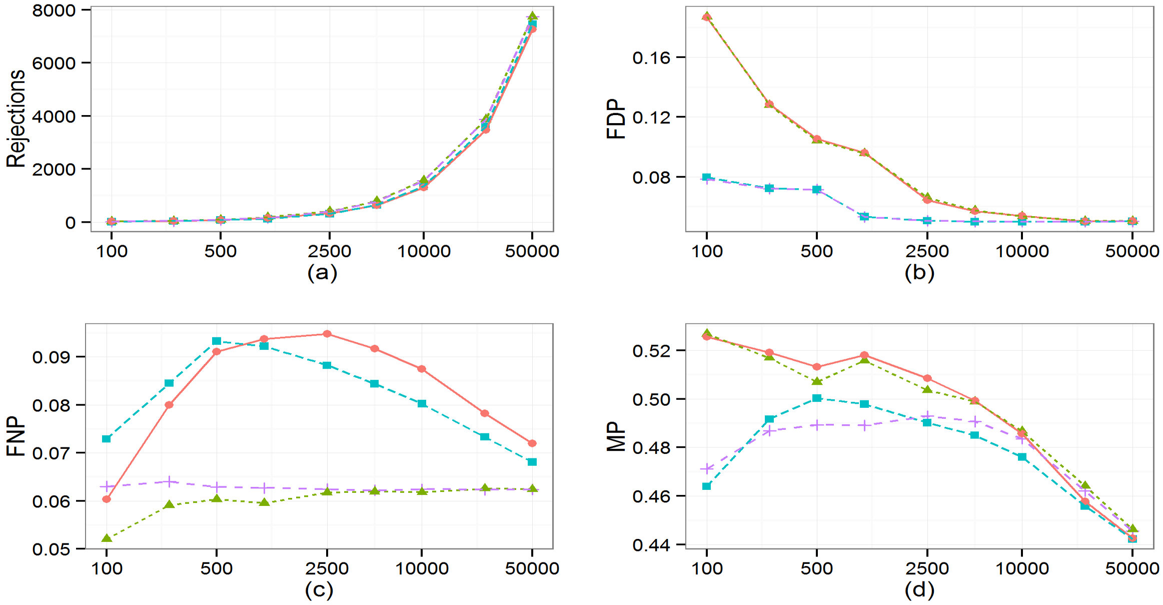

Figure 7.8: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2.7$, $\mu_2 = 3.3$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

7.2.3 Other Difficulties

Subsection 7.2.1 considered simulation settings that illustrated relatively fast convergence between the theoretical LMCR and Lfdr procedures' values and the corresponding oracle procedure values. Subsection 7.2.2, on the other hand, presented simulation scenarios that showed very slow convergence that required very large M values. This subsection will consider three cases where convergence is so slow that it is almost imperceptible, even when $M = 50,000$. Suggestions for handling such cases will also be considered.

We begin by considering Equation (6.1) with $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, and $\alpha = 0.05$. The comparison of the oracle LMCR and Lfdr procedures versus the theoretical LMCR and Lfdr procedures are found in Figure 7.9. Figure 7.10 shows the results of the theoretical versus empirical comparison. Table 7.5 provides the number of usable repetitions represented in each figure.

Based solely on Table 7.5, there are no indications of EM algorithm issues, which is to be expected

Procedure	Number of Usable Repetitions								
Theoretical	998	1000	1000	1000	1000	1000	1000	1000	1000
Empirical	1000	1000	1000	1000	1000	1000	1000	1000	1000
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.5: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

due to the nice separation between the three groups. For reference, see the histogram in Figure 6.17 for $M = 10,000$. With the relative ease of parameter estimation, one might expect good convergence results, however as we see in Figure 7.3, this is not the case. Neither the theoretical LMCR nor the theoretical Lfdr procedure's values show any indication of converging to the corresponding oracle values. However, when comparing Figures 7.9 and 7.10, we see a similar gap pattern between them, especially for the FNP.

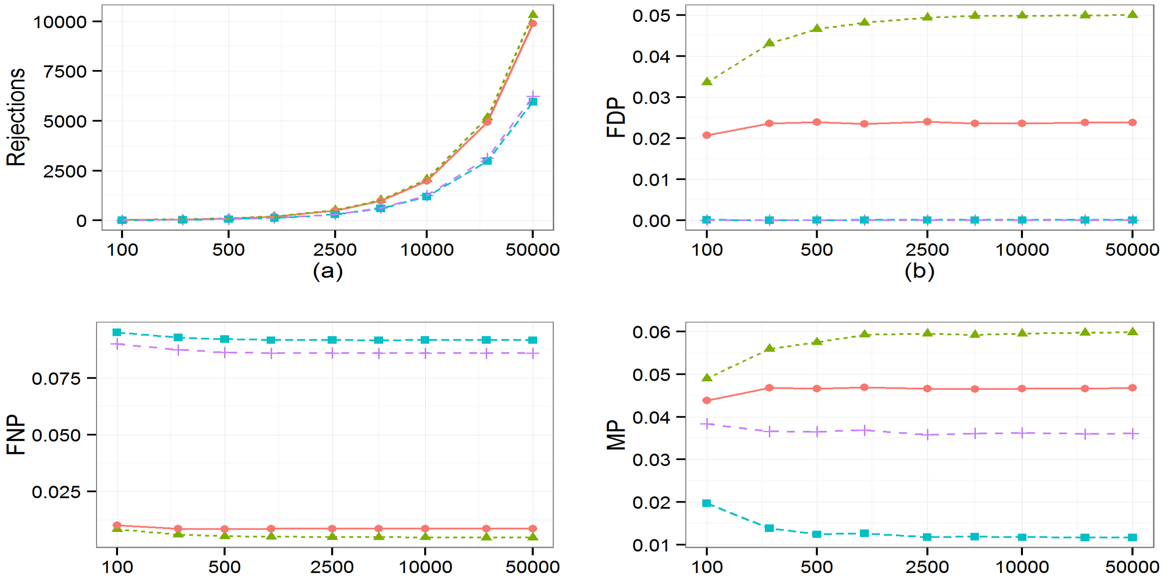

Figure 7.9: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, and $\alpha = 0.05$: The oracle LMCR (◯), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

Inspired by this similarity, Figure 7.11 considers the empirical procedures as compared to the oracle procedures. Here, we see very good convergence results for $M \geq 2,500$. This good performance

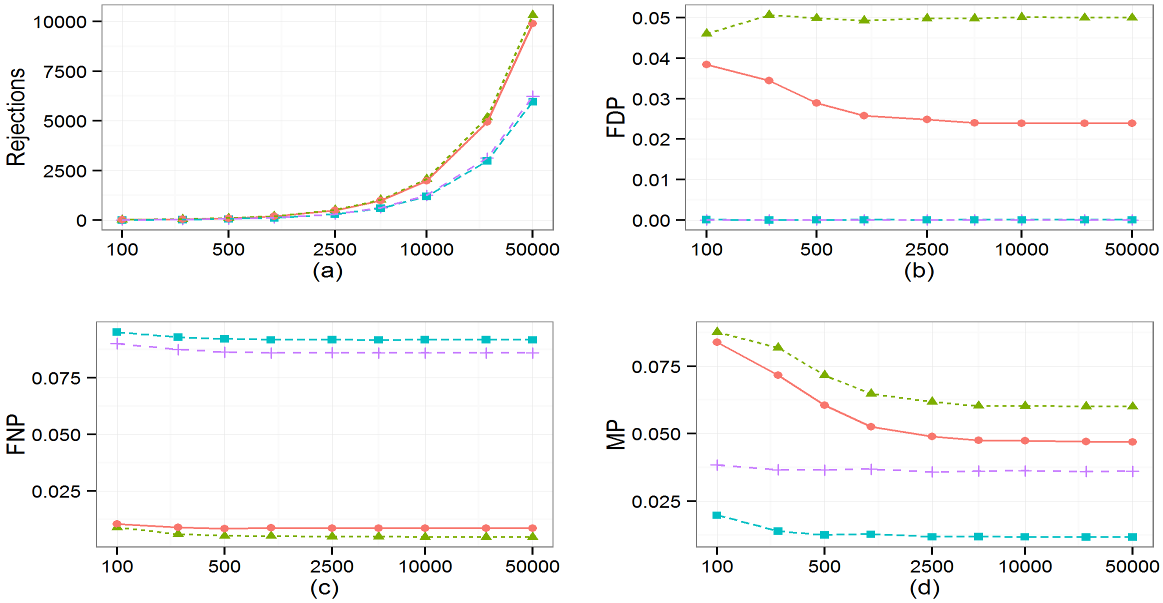

Figure 7.10: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

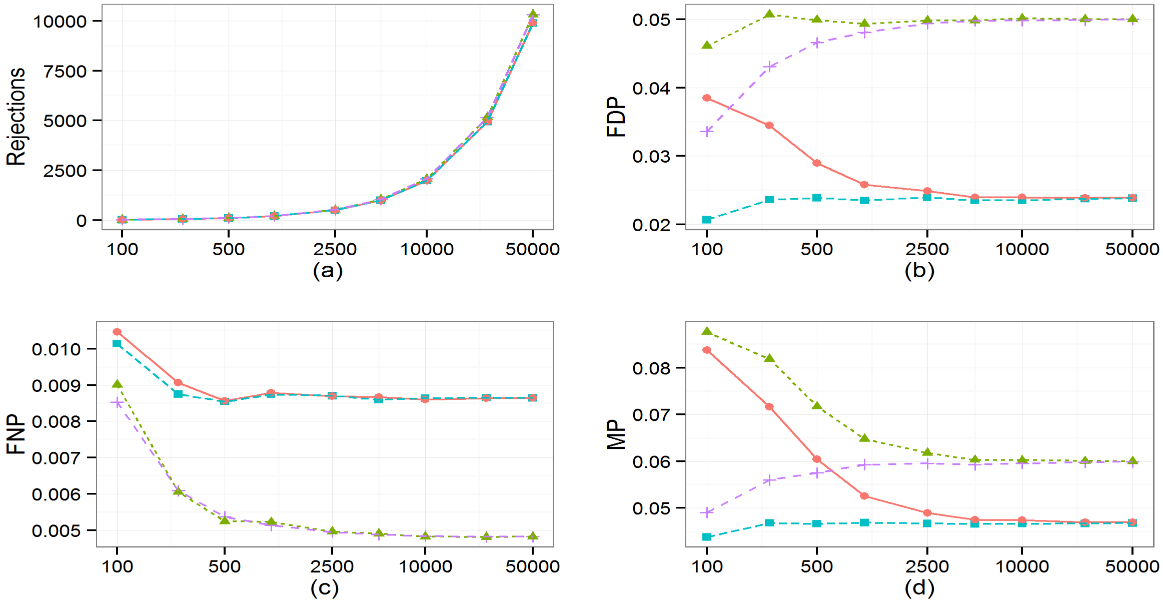

Figure 7.11: The effect of M on empirical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), oracle LMCR (□), and oracle Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

is to be expected because of the nice distribution separation discussed above. The initial poor performance is due to the misspecification of the null distribution. There are zero false discoveries for either theoretical procedure because, since $N(0,1)$ is a much wider distribution than the actual $N(0,0.5)$ null, the Lfdr values increase quickly which depresses the number of rejections to the point that no true nulls are discovered and increases the FNP. When the $N(0,1)$ null constraint is lifted, the added flexibility allows the EM algorithm to identify more appropriate parameter estimates. In practice, model misspecification may make analysis quite difficult, however in some instances, relaxing the constraints of a given model may provide useful insights.

In other instances, imposing additional constraints on a model may provide improved results. For example, consider Equation (6.1) with $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. Table 7.6 gives the number of usable repetitions for this setting. Figure 7.12 and Figure 7.13 provide the results of the theoretical versus oracle and theoretical versus empirical comparisons, respectively.

Procedure	Number of Usable Repetitions								
Theoretical	777	844	855	842	904	929	945	959	983
Empirical	912	930	932	946	929	905	881	870	825
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.6: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 0.5$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

The results in Table 7.6 are very similar to those found in Table 7.1, in that the number of usable repetitions tends to increase for the theoretical scenario as M increases, whereas the empirical case shows more variability. Additionally, all values are relatively large, thus indicating few EM algorithm convergence issues. Despite these nice results, Figure 7.12 shows multiple convergence issues, particularly in terms of the theoretical LMCR procedure's FNP and MP values. In both instances, there are sizable gaps between the theoretical and corresponding oracle procedure values, for every value of M. In Figure 7.13, we see that the empirical results seem to mostly converge to the theoretical results, especially for the MP values.

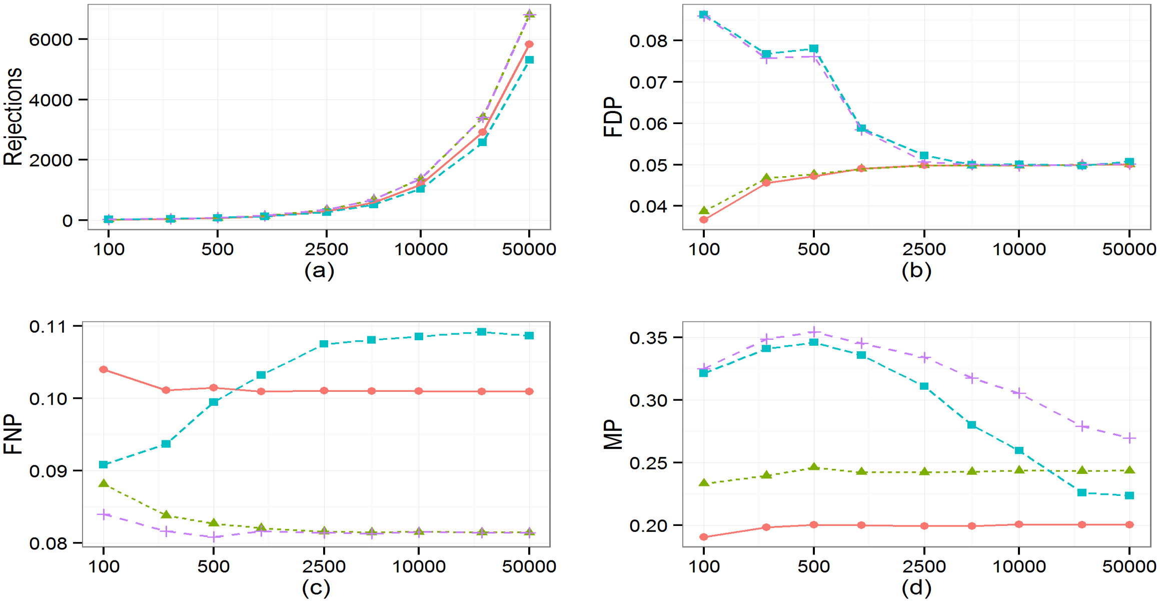

Figure 7.12: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$ and $\alpha = 0.05$: The oracle LMCR (◯), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

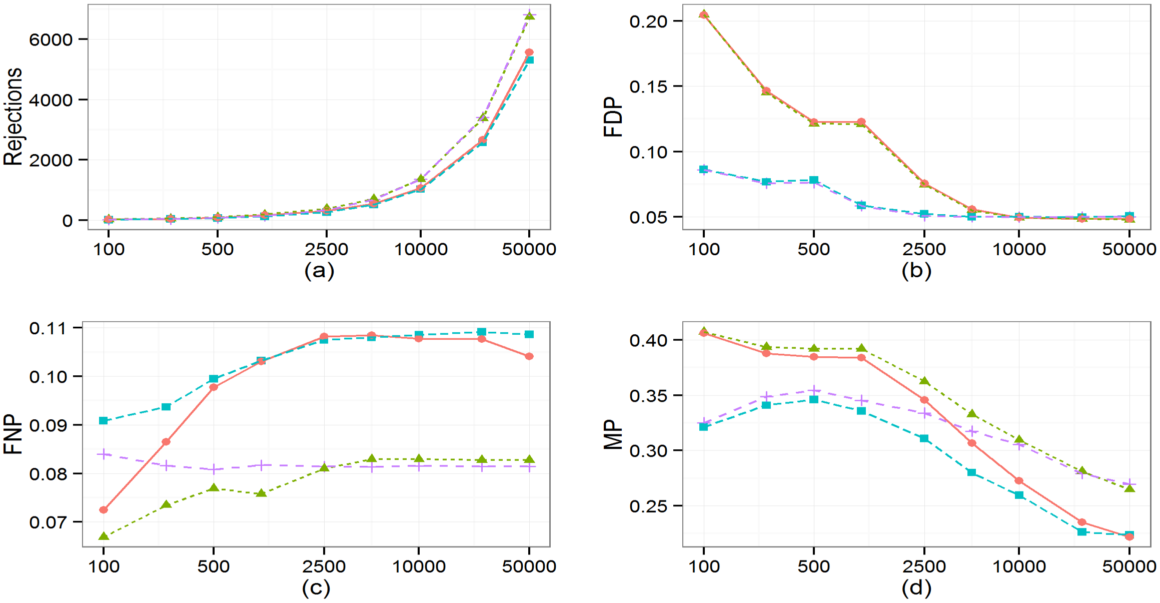

Figure 7.13: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.8$, $\pi_1 = \pi_2 = 0.1$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$ and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

Unfortunately, this means that relaxing the theoretical null constraint will be ineffective for this simulation setting. Further analysis of the results suggest that the gaps between the LMCR curves seen in Figure 7.12 can be attributed to relatively small rejection sets, as mentioned above. For example, consider the situation in Figure 7.14 where the parameter estimates are: $\hat{\pi}_0 = 0.79$, $\hat{\pi}_1 = 0.16$, $\hat{\pi}_2 = 0.05$, $\hat{\mu}_1 = 2.40$, $\hat{\mu}_2 = 4.32$, $\hat{\sigma}_1 = 1.40$, and $\hat{\sigma}_2 = 0.86$. We note that the theoretical LMCR procedure rejects only two out of a possible $50,000$ hypotheses in this case. Since $\hat{\sigma}_1$ is relatively large, misclassification is much more likely in the right tail of group 2, thus reducing the number of rejections. This, in turn, increases the FNP value to 0.202 since a large majority of nonnull hypotheses are not being discovered. The corresponding MP value is 0.5.

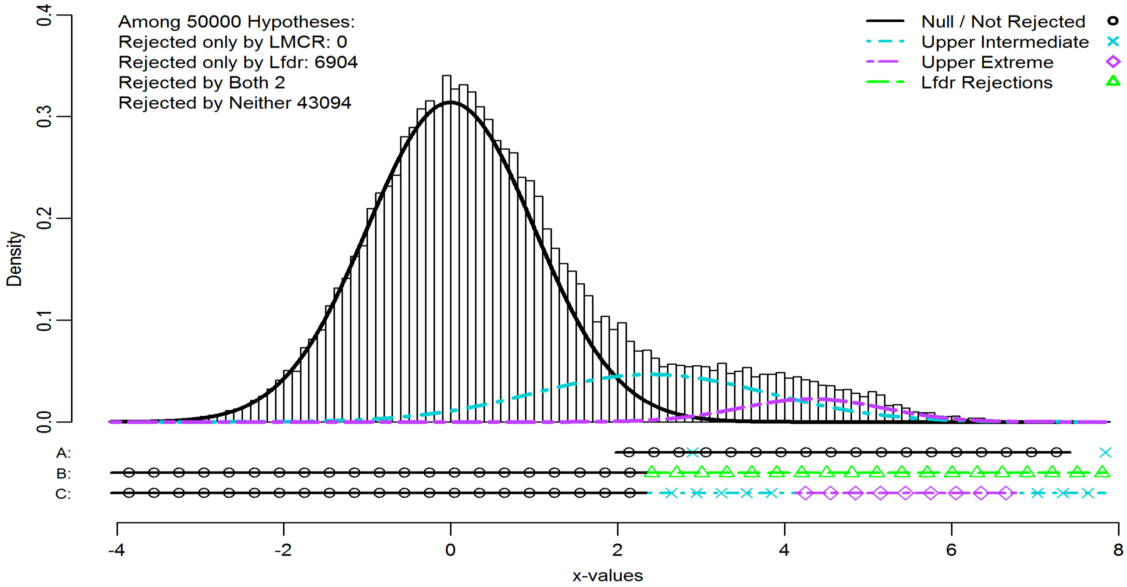

Figure 7.14: The histogram of 50,000 simulated x-values, along with fitted densities defined by a convergent EM algorithm. The rows of points near the bottom represent the pattern of rejections for the LMCR procedure (A), the original Lfdr procedure (B), and the Lfdr procedure with clustering information included (C).

In practice, in situations like this it is important to inspect parameter estimates to see if the groups are separable. For the example presented in Figure 7.14, $\hat{\mu}_1$ and $\hat{\mu}_2$ have good separation and, while $\hat{\pi}_2$ is relatively small, both $\hat{\pi}_1$ and $\hat{\pi}_2$ are bounded away from zero. So here, the issue seems to be due to the fact that $\hat{\sigma}_1$ is much larger than $\hat{\sigma}_2$. To handle this, one may need to consider fitting a model with equal alternative standard deviations. For example, suppose that a series of models are considered with $\hat{\pi}_0 = 0.79$, $\hat{\pi}_1 = 0.16$, $\hat{\pi}_2 = 0.05$, $\hat{\mu}_1 = 2.40$, and $\hat{\mu}_2 = 4.32$, as before, but now with fixed values for $\hat{\sigma}_1 = \hat{\sigma}_2$. Table 7.7 presents the number of rejections, FNP, and MP values for the theoretical LMCR procedure with various $\hat{\sigma}_1 = \hat{\sigma}_2$.

$\hat{\sigma}_1 = \hat{\sigma}_2 =$	0.8	0.9	1.0	1.1	1.2	1.3	1.4
Rejections	4790	4018	3271	2465	1678	1153	857
FNP	0.126	0.139	0.151	0.164	0.177	0.185	0.190
MP	0.282	0.290	0.304	0.320	0.355	0.390	0.448

Table 7.7: The number of rejections, FNP, and MP for the theoretical LMCR procedure when applied to the dataset presented in Figure 7.14 with $\hat{\pi}_0 = 0.79$, $\hat{\pi}_1 = 0.16$, $\hat{\pi}_2 = 0.05$, $\hat{\mu}_1 = 2.40$, $\hat{\mu}_2 = 4.32$, $M = 50,000$, and $\alpha = 0.05$.

In Table 7.7 we see that the $\hat{\sigma}_1 = \hat{\sigma}_2$ constraint greatly increases the number of rejections and reduces the FNP and MP values for each standard deviation considered. Though these values are still not quite comparable to the corresponding oracle LMCR values, they are much better than the original theoretical LMCR values. Similar improvements would also be expected for the empirical LMCR procedure.

Finally, consider Equation (6.1) with $\pi_0 = 0.9$, $\pi_1 = 0.06$, $\pi_2 = 0.04$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. The comparison of the oracle LMCR and Lfdr procedures versus the theoretical LMCR and Lfdr procedures are found in Figure 7.15. Figure 7.16 shows the results of the theoretical versus empirical comparisons. Table 7.8 provides the number of usable repetitions represented in each figure.

Procedure	Number of Usable Repetitions								
Theoretical	694	796	850	847	895	916	939	948	966
Empirical	865	911	922	925	908	910	859	830	821
$M =$	100	250	500	1000	2500	5000	10000	25000	50000

Table 7.8: The number of usable repetitions for the three group model of Equation (6.1) where $\pi_0 = 0.9$, $\pi_1 = 0.06$, $\pi_2 = 0.04$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$. The oracle procedures used all $1,000$ repetitions.

The results of Table 7.8 are similar to those of Tables 7.1 and 7.6 in showing that there are relatively few EM algorithm convergence issues for this simulation setting for large M. However, as in Figure 7.12, these results do not prevent issues in the convergence of theoretical LMCR values to the corresponding oracle LMCR values. In fact, while the theoretical LMCR procedure's MP value seems to be slowly converging in Figure 7.15, its FNP value is actually diverging from the oracle FNP when $10,000 \leq M \leq 50,000$. Again, Figure 7.16 shows that the empirical LMCR values converge to those of the theoretical LMCR, so we will focus on analyzing the theoretical situation.

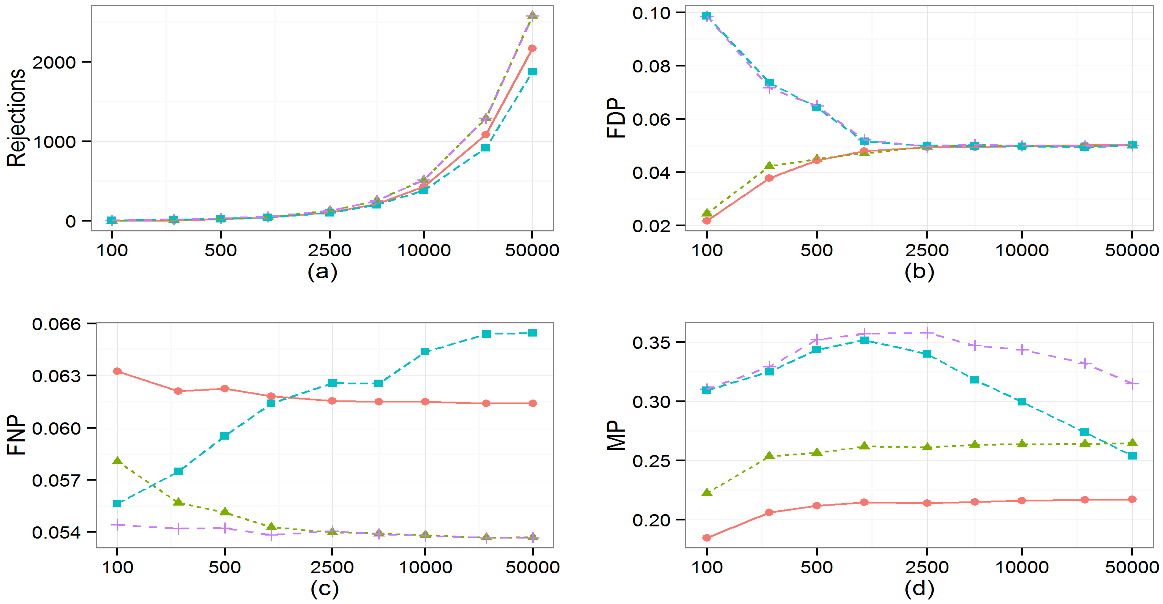

Figure 7.15: The effect of M on theoretical and oracle convergence when $\pi_0 = 0.9$, $\pi_1 = 0.06$, $\pi_2 = 0.04$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The oracle LMCR (◯), oracle Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c), and the MP (d).

As mentioned above, the difficulties in this simulation are analogous to those of the previous simulation, only now the clustering problem is more difficult because of the smaller alternative proportions. For reference, consider the histogram in Figure 6.10, where we note that the identification of two nonnull distributions seems to be nontrivial since there are relatively few data points in the right tail. Further, for the $M = 50,000$ case, 32 repetitions resulted in zero rejections (with 35 resulting in two or fewer rejections) whereas the previous simulation setting resulted in only 14 such repetitions (with 23 resulting in two or fewer rejections). For the repetitions with zero rejections, the estimated standard deviations are quite similar for both simulation settings, with $\hat{\sigma}_2 < 1 < \hat{\sigma}_1$. However here, for the repetitions with zero rejections, the average $\hat{\pi}_2$ value is 0.017 (as compared to 0.051 for the previous analysis), which indicates that the EM algorithm had difficulty identifying

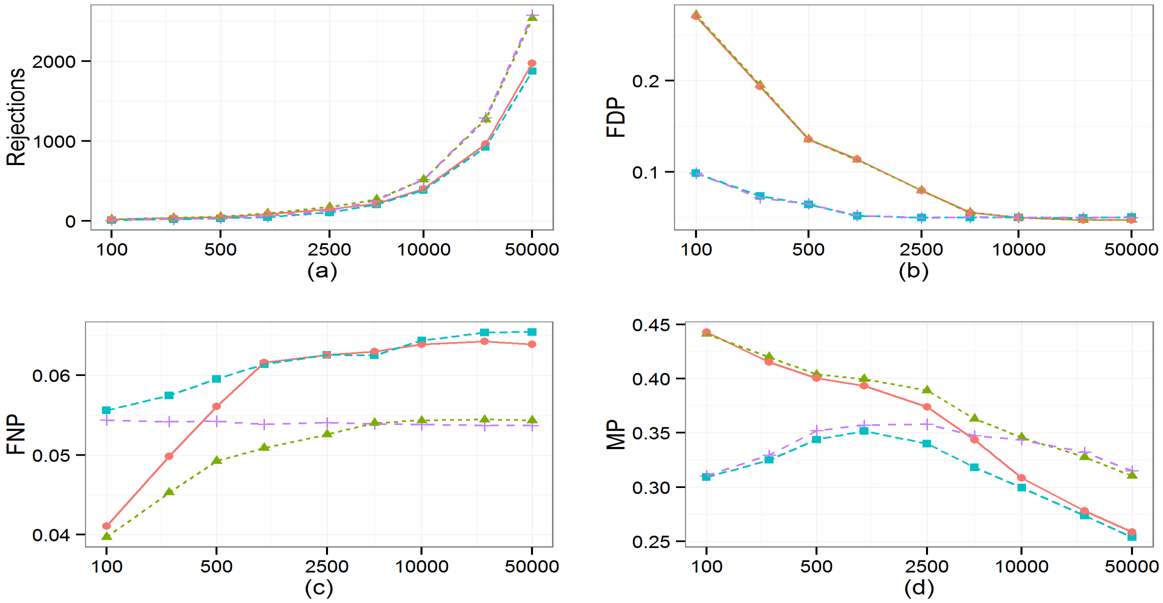

Figure 7.16: The effect of M on theoretical and empirical convergence when $\pi_0 = 0.9$, $\pi_1 = 0.06$, $\pi_2 = 0.04$, $\mu_1 = 2$, $\mu_2 = 4$, $\sigma_0 = \sigma_1 = \sigma_2 = 1$, and $\alpha = 0.05$: The empirical LMCR (◯), empirical Lfdr (△), theoretical LMCR (□), and theoretical Lfdr (+) procedures are compared with respect to the number of rejections (a), the FDP (b), the FNP (c) and the MP (d).

a second nonnull group. In practice, once the parameter estimates have been examined, one may consider applying a model with the $\sigma_1 = \sigma_2$ constraint, as discussed above, or perhaps even consider a simplified two group model, if necessary.

Chapter 8

Summary

In this dissertation, Chapter 3 provided a new statistical protocol for large-scale exploratory analysis that allows for the nature of the follow-up analysis to be included in the discovery process. Contrary to standard procedures, which are designed to maximize the expected number of discoveries or minimize some type II error rate, the proposed protocol creates FDR controlling procedures that are more appropriate for the study at hand by incorporating the specific statistic of interest. Chapter 4 used the protocol to develop the LMCR procedure which controls the FDR while incorporating classification. Chapter 5 compared the LMCR procedure with the standard Lfdr based procedure using illustrative and real data. It was shown that the LMCR procedure had two main advantages: (1) it provides a built-in safeguard that prevents the rejection of hypotheses that are most likely null and (2) rejects hypotheses that have the smallest LMCR. Chapters 6 and 7 then provided extensive simulation results for the oracle and adaptive LMCR procedures, respectively.

While this work provided three specific examples (the Lfdr, CLfdr, and LMCR procedures) of how the two step protocol of Chapter 3 could be used to create procedures that control the FDR, the framework is general enough to incorporate any arbitrary statistic or criterion of interest. For example, the LMCR procedure could be extended by introducing a non-uniform misclassification penalty. In other studies, attributes may be ranked based on information criteria, effects sizes, or some other study specific definition of "interesting." Traditional procedures have been developed to maximize the number of discoveries perhaps because this approach is reasonable in most exploratory analyses, despite not necessarily being the best approach. The proposed method allows for the inclusion of the statistic best suited for the current study.

Bibliography

Anderberg, M. R. (1973). *Cluster Analysis for Applications.* New York: Academic Press.

Anderson, M. and J. D. Habiger (2012). Characterization and identification of productivity-associated rhizobacteria in wheat. *Applied and Environmental Microbiology 78*(12), 4434–4446.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.). New York: Wiley.

Bain, L. J. and M. Engelhardt (2000). *Introduction to Probability and Mathematical Statistics* (2nd ed.). Duxbury Press.

Benaglia, T., D. Chauveau, D. R. Hunter, and D. Young (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software 32*(6), 1–29.

Benjamini, Y. and M. Bogomolov (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society. Series B 76*(1), 297–318.

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

Benjamini, Y. and Y. Hochberg (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics 24*(3), 407–418.

Benjamini, Y. and Y. Hochberg (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics 25*(1), 60–83.

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

Blanchard, G. and Étienne. Roquain (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research 10*, 2837–2871.

Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis 23*(1), 5–28.

Cai, T. T. and W. Sun (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association 104*(488), 1467–1481.

Cao, H., W. Sun, and M. R. Kosorok (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika*, 495–502.

Cramér, H. (1946). *Mathematical Methods of Statistics*, Volume 9. Princeton University Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B 39*(1), 1–38.

Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern Classification* (2nd ed.). New York: Wiley.

Dudoit, S. and M. J. van der Laan (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Dunn, O. J. (1958). Estimation of the means of dependent variables. *The Annals of Mathematical Statistics 29*(4), 1095–1111.

Dunn, O. J. (1959). Confidence intervals for the means of dependent, normally distributed variables. *Journal of the American Statistical Association 54*(287), 613–621.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association 99*(465), 96–104.

Efron, B. (2008a). Microarrays, empirical bayes, and the two-groups model. *Statistical Science 23*(1), 1–22.

Efron, B. (2008b). Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics 2*(1), 197–223.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge: Cambridge University Press.

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

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics 34*(2), 275–297.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research 17*(4), 347–388.

Finner, H., T. Dickhaus, and M. Roters (2007). Dependency and false discovery rate: Asymptotics. *The Annals of Statistics 35*(4), 1432–1455.

Fraley, C. and A. E. Raftery (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal 41*(8), 578–588.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.*

Genovese, C., K. Roeder, and L. Wasserman (2006). False discovery control with p-value weighting. *Biometrika 93*(3), 509–524.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series B 64*(3), 499–517.

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

Habiger, J., D. Watts, and M. Anderson (2015). Multiple testing with heterogeneous multinomial distributions. *arXiv preprint arXiv:1511.01400*.

Habiger, J. D. (2012). A method for modifying multiple testing procedures. *Journal of Statistical Planning and Inference 142*(7), 2227–2231.

Habiger, J. D. and E. A. Peña (2011). Randomised p-values and nonparametric procedures in multiple testing. *Journal of nonparametric statistics 23*(3), 583–604.

Habiger, J. D. and E. A. Peña (2014). Compound p-value statistics for multiple testing procedures. *Journal of Multivariate Analysis 126*, 153–166.

Hartigan, J. A. (1975). *Clustering Algorithms.* New York: Wiley.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika 75*(2), 383–386.

Hu, J. X., H. Zhao, and H. H. Zhou (2010). False discovery rate control with groups. *Journal of the American Statistical Association 105*(491), 1215–1227.

Jain, A., M. Murty, and P. Flynn (1999). Data clustering: A review. *ACM Computing Surveys 31*(3), 264–323.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for Clustering Data.* New Jersey: Prentice-Hall.

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

Lehmann, E. L. and G. Casella (2003). *Theory of Point Estimation* (2nd ed.). Springer.

Lindquist, M. A. (2008). The statistical analysis of fmri data. *Statistical Science 23*(4), 439–464.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (2nd ed.), Volume 37.; 37. New York; London: Chapman and Hall.

McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions*. Hoboken, NJ: Wiley-Interscience.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

Nettleton, D., J. G. Hwang, R. Caldo, and R. p. Wise (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics 11*(3), 337–356.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roeder, K., S.-A. Bacanu, L. Wasserman, and B. Devlin (2006). Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics 78*(2), 243–252.

Roeder, K. and L. Wasserman (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical Science 24*(4), 398–413.

Ruppert, D., D. Nettleton, and J. T. G. Hwang (2007). Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics 63*(2), 483–495.

Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A 70*(2), 135–168.

Scott, A. J. and M. J. Symons (1971). Clustering methods based on likelihood ratio criteria. *Biometrics 27*(2), 387–397.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software 53*(4), 1–37.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika 73*(3), 751–754.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B 64*(3), 479–498.

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics 31*(6), 2013–2035.

Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society. Series B 69*(3), 347–368.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B 66*(1), 187–205.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association 102*(479), 901–912.

Sun, W. and T. T. Cai (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society. Series B 71*(2), 393–424.

Sun, W. and A. C. McLain (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association 107*(498), 673–687.

Sun, W. and Z. Wei (2015). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika 102*(2), 267–280.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Wu, W. B. (2008). On false discovery control under dependence. *The Annals of Statistics 36*(1), 364–380.

Xu, R. and D. C. Wunsch (2009). *Clustering*. New Jersey: Wiley.

VITA

David D. Watts

Candidate for the Degree of

Doctor of Philosophy

Thesis: CLASSIFYING DISCOVERIES: IMPLEMENTING A GENERALIZED MULTIPLE TESTING PROTOCOL FOR EXPLORATORY DATA ANALYSIS

Major Field: Statistics

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Statistics at Oklahoma State University, Stillwater, Oklahoma in July, 2016.

Completed the requirements for the Master of Science in Applied Mathematics at the University of Central Arkansas, Conway, Arkansas in August, 2008.

Completed the requirements for the Bachelor of Arts in Mathematics at the University of Central Arkansas, Conway, Arkansas in May, 2006.

Experience:

January, 2014 - May, 2016: Graduate Research Assistant for the Department of Computer Science at Oklahoma State University, Stillwater, Oklahoma.

August, 2011 - December, 2013: Graduate Teaching Assistant for the Department of Statistics at Oklahoma State University, Stillwater, Oklahoma.

August, 2008 - December, 2010: Instructor for the Department of Mathematics at the University of Central Arkansas, Conway, Arkansas.

August, 2006 - July, 2008: Graduate Teaching Assistant for the Department of Mathematics at the University of Central Arkansas, Conway, Arkansas.

Professional Memberships:

American Statistical Association