NEW APPROACHES TO MULTIPLE TESTING OF GROUPED HYPOTHESES

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Yanping Liu
May, 2016

Examining Committee Members:

Sanat K. Sarkar, Advisory Chair, Statistics
Zhigen Zhao, Statistics
Xu Han, Statistics
Cheng Yong Tang, Statistics
Li He, External Reader, Merck Research Laboratories

ProQuest Number: 10111313

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

ProQuest 10111313

Published by ProQuest LLC (2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

©
by
Yanping Liu
May, 2016
All Rights Reserved

ABSTRACT

NEW APPROACHES TO MULTIPLE TESTING OF GROUPED HYPOTHESES

Yanping Liu
DOCTOR OF PHILOSOPHY
Temple University, 2016
Professor Sanat K. Sarkar, Chair

Testing multiple hypotheses appearing in non-overlapping groups is a common statistical problem in many modern scientific investigations, with this group formation occurring naturally in many of these investigations. The goal of this dissertation is to explore the current state of knowledge in the area of multiple testing of grouped hypotheses and to present newer and improved statistical methodologies.

As the first part of this dissertation, we propose a new Bayesian two-stage multiple testing method controlling false discovery rate (FDR) across all hypotheses. The method decomposes a posterior measure of false discoveries across all hypotheses into within- and between-group components allowing a portion of the overall FDR level to be used to maintain control over within-group false discoveries. Such within-group FDR control effectively captures the group structure as well as the dependence, if any, within the groups. The procedure can maintain a tight control over the overall FDR, as shown numerically under two different model assumptions, independent and Markov dependent

Bernoulli's, for the hidden states of the within-group hypotheses. The proposed method in its oracle form is optimal at both within-and between-group levels of its application. We also present a data driven version of the proposed method whose performance in terms of FDR control and power relative to its relevant competitors is examined through simulations.

We apply this Bayesian method to a real data application, which is the Adequate Yearly Progress (AYP) study data of California elementary schools (2013) comparing the academic performance for socioeconomically advantaged (SEA) versus socioeconomically disadvantaged (SED) students, and our method has more meaningful discoveries than two other competing methods existing in the literature.

The second part of the dissertation is geared towards making contribution to the outstanding problem of developing an FDR controlling frequentist method for multiple testing of grouped hypotheses, which can serve not only as an extension of the classical Benjamini -Hochberg (BH, 1995) method from single to multiple groups but also can be more powerful due to the underlying group structure. We suggest a number of such methods and examine their performances in comparison with the single-group BH method mainly based on simulations.

Some possible future directions of research in the proposed area are discussed at the end of this dissertation.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have contributed to the work in this thesis and who have supported me during my study.

First and foremost, I am deeply indebted to my advisor, Dr. Sanat Sarkar, for his encouragement, support, guidance and insights throughout this research. His extensive knowledge, his enthusiasm, his inspiration, and his endless encouragement help me all the time of my research. And I would like to be thankful to his kindness and support after Dr. Raghavarao passed away. Without his willingness to be my advisor, I cannot finish my Ph.D degree.

My sincere appreciation also goes to Dr. Zhigen Zhao for his inspiring discussions and suggestions during my research. His detailed guidance, encouragement, and patience help me tremendously throughout all the stages of my thesis writing. I also would like to thank my dissertation committee members, Dr. Xu Han, Dr. Cheng Yong Tang, and Dr. Li He, for their helpful suggestions and editing remarks.

I also express my gratitude to Dr. Damaraju Raghavarao who served as my previous advisor but sadly passed away on February 6, 2013. His kindness, encouragement, and wisdom will be remembered in my heart forever.

I am indebted to all the professors in the Statistics Department, Temple University. The assistantships from the Department and NSF Grants DMS-1208735 and DMS-1309273 are gratefully acknowledged.

Finally, I thank my family members especially my husband Zijiang Yang, for all their endless support and love throughout my study at Temple University!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Multiple testing problems arising in modern scientific investigations often involve hypotheses that appear in non-overlapping groups. Such group formation occurs naturally in many of these investigations due to the underlying biological or experimental process, although it can be created in some applications to effectively capture certain specific features of the data. For instance, the hypotheses associated with (i) the locations (voxels) in each region of a brain (either anatomic or functional) in fMRI research (Benjamini and Heller (2007) and Pacifico et al. (2004)), (ii) the genes in each of non-overlapping gene sets determined through Gene Ontology in certain microarray experiments (Subramanian et al. (2005) and Heller et al. (2009)), or (iii) the time-periods corresponding to each gene in a time-course microarray experiment (Arbeitman et al. (2002) and Calvano et al. (2005)) naturally form a group. On the other hand, in the applications of multiple testing methods to detect astronomical transient source from nightly telescopic image consisting of large number of pixels (Clements et al. (2011)) or vegetation fluctuation in a particular region based on satellite remote sensing data collected over large number of grid points (Clements et al. (2014)), the pixels or grid points were grouped into suitably defined blocks or sub-regions to capture local dependency.

The hypotheses under such group structure are correlated with their neighboring hypotheses with some correlations in each group but uncorrelated between groups. Or we can say, if one hypothesis is significant, its nearby hypotheses are more likely to be significant in the same group. Ignoring the group structure when constructing multiple testing methods may result in misleading conclusions (Efron (2008)). With that being the rationale, a considerable amount of research has taken place recently in the development of multiple testing methods designed specifically for grouped hypotheses both from frequentist and Bayesian perspectives (for example, Heller et al. (2009), Clements et al. (2011), Hu et al. (2010), and Sun and Wei (2011)). Of course, these methods have been developed in the context of controlling false discovery rate (FDR), as originally defined by Benjamini and Hochberg (Benjamini and Hochberg (1995)), or its variants such as local FDR (Efron et al. (2001)) and marginal FDR (Sun and Cai (2007) and Sun and Cai (2009)), which is a more acceptable practice in modern large-scale multiple testing than controlling the more traditional familywise error rate. Moreover, controlling FDR across all hypotheses has been the ultimate goal in most of these methods, although in some, the focus has been on controlling FDR only across the groups (Sun and Wei (2011)), or controlling the expected average value of some error rate over the selected groups (Benjamini and Bogomolov (2014)).

1.2 Some Basics of Multiple Hypothesis Testing

Suppose that we have n null hypotheses $H_1, \ldots, H_n$ to be tested simultaneously. Our goal is to determine a rule to specify what decision should be made for each null hypothesis, based on test statistics or p-values associated with the null hypotheses. The rule is determined based on the idea of using a procedure that leads us to the right decisions with high probability and a control of a suitably defined error rate. For single hypothesis, this problem is quite straightforward. A good procedure would be the one that controls Type I error rate, which is the probability of making false rejection of the hypothesis when it is truly null, while minimizing Type II error rate (or maximizing the power), which is the probability of accepting it when it is truly false (or making correct rejection). However, when it comes to multiple testing, unlike in testing of a single hypothesis where the Type I error rate, is being controlled, a multiple testing procedure adopts the error rate that provides an overall measure of Type I errors.

The outcomes of multiple testing can be summarized as in Table (1.1). Different types of error rate could then be defined based on this table. Note that here the total number of hypotheses n is fixed and known, the number of true and false null hypotheses n_0 and n_1, respectively, are fixed but unknown, while random variables A and R are observable. The four variables U, V, S and T are not observable.

Now, let's quickly review these frequently used overall type I error rates in multiple testing procedures.

	H_i is rejected	H_i is accepted	
H_i is true	V	U	n_o
H_i is false	S	T	n_1
Total	R	A	n

Table 1.1: Classification of Tested Hypothesis

1.2.1 Overall Type I Error Rates

- Per-Family Error Rate: The expected number of false rejections, i.e.

$$\text{PFER} = E(V).$$

- Familywise Error Rate: The probability of having at least one false rejection, i.e.

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

- Generalized Family-Wise Error Rate (GFWER): The probability of at least k Type I errors for a user-supplied integer k, i.e.

$$\text{GFWER(k)} = \Pr(V \geq k).$$

When $k = 1$, GFWER(k) reduces to the usual family-wise error rate (FW-ER). The FWER has been the most commonly used among the above error rates. However, with large number of hypotheses, as encountered in many modern statistical investigations, procedures controlling it become too stringent, resulting in conservative procedures with inadequate power. To overcome this problem, the following alternative measures have been introduced.

- False Discovery Rate (FDR): The expected proportion of false discoveries among all rejections. Let

$$\text{FDP} = \begin{cases} \frac{V}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

be the False Discovery Proportion. Then, the FDR is defined as

$$\text{FDR} = E(\text{FDP}) = E\left(\frac{V}{R} \mid R > 0\right) \cdot \Pr(R > 0) = E\left(\frac{V}{R \vee 1}\right). \quad (1.1)$$

Here, $R \vee 1$ means the maximum value of R or 1.

- Positive False Discovery Rate (pFDR): The expected proportion of false discoveries among all rejections given there is at least one rejection, i.e.

  $$\text{pFDR} = E(\text{FDP} \mid R > 0) = E\left(\frac{V}{R} \mid R > 0\right). \quad (1.2)$$

- The Exceedance Probability of False Discovery Proportion (γ-ExFDP): The probability of FDP exceeding $\gamma \in (0, 1)$, i.e.

  $$\text{ExFDP}(\gamma) = \Pr(\text{FDP} > \gamma).$$

- Generalized False Discovery Rate (k-FDR): The expected proportion of k or more false discoveries among all rejections, where k is pre-specified, i.e.

  $$k\text{-FDR} = E(k\text{-FDP}), \quad (1.3)$$

  where

  $$k\text{-FDP} = \begin{cases} \frac{V}{R} & \text{if } V \geq k; \\ 0 & \text{if } V < k. \end{cases}$$

  Note that when $k = 1$, it reduces to the original FDR.

- Marginal FDR (mFDR) is the ratio of expected number of Type I errors to the expected number of rejected hypotheses.

  $$mFDR = \frac{E(V)}{E(R)}$$

The concept of FDR has been introduced by Benjamini and Hochberg (1995), which is more powerful than the concept of the FWER. However, as Storey (2002) argues, there are some difficulties with this notion of FDR. To cope with these difficulties, he [Storey (2002) and Storey (2003)] introduced the notion of positive FDR (pFDR) by deleting the second term ($\Pr(R > 0)$) in the FDR definition (1.1). Of course, the pFDR cannot always be controlled, as it is equal to 1 when all hypotheses are true. Storey (2002) proposed an estimation based approach to control it through a suitable estimate of it for a fixed rejection region.

In many applications, particularly when n is large, one might be willing to tolerate more than one false rejection and seeks to control at least k, rather than at least one, false rejections. This will improve the power of a procedure to detect more false null hypotheses. The concepts of k-FWER, γ-ExFDP and k-FDR have been introduced as appropriate measures of error rates in these situations. Korn et al. (2004) first considered using the k-FWER and the γ-ExFDP. Sarkar (2007) proposed the idea of controlling the k-FDR. It is a less conservative notion than the k-FWER and is a natural generalization of the idea of improving the FWER using the FDR. Guo et al. (2014) further introduced $\gamma - FDP$, the probability of false discovery proportion (FDP) exceeding $\gamma \in [0, 1)$, and provided several procedures with more powerful performances than other existing methods to control the false discovery proportion.

In this dissertation, we will use FDR as the overall Type I error measure..

We will review some FDR controlling procedures with and without group structure in Chapter 2.

1.2.2 Power and Overall Type II Error Rates

Different concepts of power have been used in the literature when comparing procedures controlling an overall Type I error rate. These are presented in the following:

- Probability of rejecting at least one false null hypothesis:

  $$\Pr(S > 0) = \Pr(T \neq n_1).$$

  This power might be too large to compare, especially when the number of hypothesis is large.

- Probability of rejecting all false null hypothesis:

  $$\Pr(S = n_1) = \Pr(T = 0).$$

  This power might be too small to compare, especially when the number of hypotheses is large.

- Average Power, the expected proportion of correct rejections among all false null hypotheses (AvePower):

  $$\text{AvePower} = \begin{cases} \frac{E(S)}{n_1} & \text{if } n_1 > 0 \\ 1 & \text{if } n_1 = 0. \end{cases} \tag{1.4}$$

- Number of rejections: This measure is simple, and is often used when comparing two procedures controlling the same error rate. The procedure with larger total number of rejections is said to be more powerful than the other.

Each of these power measures corresponds to an overall measure of Type II errors. For example, $P(T = n_1)$ and $P(0 < T \leq n_1)$ can be thought of as overall measures of Type II errors obtained from the first two measures of power.

The FNR was defined by Genovese and Wasserman (2001), and independently by Sarkar (2004) (who calls it the False Negative Rate), as an analogue of FDR in terms of Type II errors (false negatives). The FNP is defined as,

$$\text{FNP} = \begin{cases} \frac{T}{A} & \text{if } A > 0 \\ 0 & \text{if } A = 0 \end{cases}$$

and the corresponding error rate, the False Non-Discovery Rate (FNR), is defined as:

$$\text{FNR} = E(FNP) = E\left(\frac{T}{A} \mid A > 0\right) \cdot \Pr(A > 0),$$

The marginal false non-discovery rate (mFNR) is also a criteria to measure the Type II error rate corresponding to mFDR. It is the proportion of the expected number of false non-discoveries to the expected number of acceptances.

$$mFNR = \frac{E(U)}{E(A)}$$

1.3 Mixture and Hidden Markov Models

The random mixture model is convenient and efficient model for large-scale multiple testing and has been widely used in many applications. In this thesis, we also use this model in the grouped multiple testing environment for our procedures.

Definition 1 (Mixture Model) *Given pairs of independent random variables (X_i, θ_i), $i = 1, \ldots, n$, with X_i representing test statistic or p-value associated with hypothesis H_i and $\theta_i = 0$ or 1 indicating whether the hypothesis H_i is true or false, respectively, the following distributional assumptions describe a two-group mixture model for the X_i's: Conditioned on the θ_i's , each X_i is distributed as*

$$X_i|\theta_i \sim (1-\theta_i)F_0 + \theta_i F_1,$$

given two distribution functions F_0 and F_1, and the θ_i's are Bernoulli random variables with $P(\theta_i = 0) = \pi_0$ and $P(\theta_i = 1) = \pi_1$. Marginally, the cumulative distribution function of each X_i is the mixture distribution $F(x) = \pi_0 F_0(x) + \pi_1 F_1(x)$.

Also, the following Hidden Markov Model (HMM) was considered in many papers, such as Sun and Cai (2009) and Sun and Wei (2011), to model the dependence structure among the test statistics, especially for microarray time-course (MTC) experiments. In such MTC experiments, the genes at different time points exhibit Markov dependence.

Definition 2 (Hidden Markov Model) *Let θ_i be the hidden state indicator indicating whether the hypothesis Hi is true $(\theta_i = 0)$ or false $(\theta_i = 1)$. This model is described by assuming that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ form a markov chain with some initial probability and transition probability (p_h^1,a_{kh}^j) , with X_i's having distributions as in Definition (1), conditionally given the θ_i's.*

1.4 Structure

In Chapter 2 of the thesis, we will go over some existing FDR controlling procedures for testing multiple hypotheses appearing in a single as well as multiple groups. Typically, a multiple testing method for grouped hypotheses operates in two stages. Significant groups are identified at the first stage and the final discoveries made within each significant group at the second stage. However, the novel Bayesian method we propose in this thesis takes the reverse route. We consider controlling FDR a posteriori. Given α at which the posterior total FDR is to be controlled, our method screens the hypotheses within each group at the first stage for possible rejections subject to a control over the posterior within-group FDR at a certain level less than or equal to α. At the second stage, it makes the final decision on ultimately rejecting these hypotheses if the groups containing them are identified as significant, given the potential decisions made about them at the first stage, subject to the desired control over the posterior total FDR. Thus, our method enjoys, unlike the other available methods, the added flexibility in adjusting or preserving a pre-chosen level of control over false discoveries within each significant group along with controlling the false discoveries across all hypotheses. Such within-group control of false discoveries is an effective way of capturing the underlying group structure as well as the within-group dependence. We will present the proposed methodology in details and discuss the optimality of the proposed method in Chapter 3.

In Chapter 4, we carried out extensive simulations and numerical studies assessing the performance of our method, both in terms of its oracle and data-driven versions, against its relevant competitors under two different dependence structures, with the distributional parameters being estimated using the EM algorithm for the data-driven version. With θ_{gj} representing the hidden state for the jth hypothesis in the gth group, two types of distribution have been considered for the θ_{gj}'s to reflect these dependence structures, each under the assumption of independence between but not within groups that seems justified in most of the aforementioned applications. These are (i) iid Bernoullis and (ii) dependent Bernoullis with Markovian dependence as often used in practice (Rabiner (1989), Churchill (1992), Krogh et al. (1994) and Ephraim and Merhav (2002)). The first one was considered with the idea of capturing only the group structure while the other pays attention to capturing both group and dependence structures. These studies have revealed superior performance of our method in terms of FDR control and power (measured using false non-discoveries and the expected proportion of correctly rejected false nulls) over its competitors in many practical scenarios. A real data application (The data from the Adequate Yearly Progress (AYP) study of California elementary schools (California (2013)) comparing the academic performance for socioeconomically advantaged (SEA) versus socioeconomically disadvantaged (SED) students) under the truncated Independent Bernoulli Model is also illustrated.

We then further investigate some alternative frequentists' methods for such grouped multiple hypotheses testing in Chapter 5. We try several methods which utilizes the group structure and group information to exam the total FDR controlling and also we compare these methods with the regular BH method which ignores the group structure. We show the simulation results for these methods, try to explain the reasons if the validity is invalid for some method, and we would like to have some contributions for the further research in this area. The performance of AYP application analyzed by the proposed frequentist's method 1 is also compared with the regular BH method.

Finally, some summaries and future work are discussed in Chapter 6, and most of the proofs and technical details are given in Appendix.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we will review some frequently used FDR controlling approaches that are related to the procedures we discuss in this dissertation and grouped hypotheses testing approaches. FDR controlling approaches include frequentist and Bayes approaches. Frequentist approaches are those that do not utilize any prior information on the parameters are utilized, while Bayes approaches take into account such prior information. We also review some currently available approaches to grouped hypotheses testing.

2.1 FDR Controlling Approaches

2.1.1 The Benjamini and Hochberg Procedure

Benjamini and Hochberg (1995) first proposed this procedure, which is now referred to as the BH method in the literature. Given n null hypotheses $H_1, \dots, H_n$ that are to be tested simultaneously using their respective p-values $P_1, \dots, P_n$, the BH method is a step-up procedure based on the Simes critical values $\alpha_i = \frac{i\alpha}{n}$, $i = 1, \ldots, n$ (Sarkar and Chang, 1997; Sarkar, 1998). More specifically, with $P_{1:n} \leq \cdots \leq P_{n:n}$ denoting the ordered p-values and $H_{(i)}$ being the hypothesis corresponding to $P_{i:n}$, the BH method rejects $H_{(i)}$ for all $i \leq R$, where

$$R = \max\left\{1 \leq i \leq n : P_{i:n} \leq \frac{i\alpha}{n}\right\},$$

provided this maximum exists; otherwise, it rejects none.

Benjamini and Hochberg (1995) proved that this method can control the FDR at $n_0\alpha/n$ conservatively when the p-values are independent. Later, Benjamini and Yekutieli (2001) showed that the BH procedure in fact controls the FDR exactly at $n_0\alpha/n$ under independence, and conservatively if the test statistics are positive regression dependent on subset (PRDS) of the test statistics corresponding to the null hypotheses; see also Sarkar (2002). Sarkar (2002) further proved that Simes' critical values can be adopted in a generalized step-up-down procedure (SUDP) and the FDR can still be controlled under similar dependency.

2.1.2 Adaptive BH Procedure

The BH procedure, when $n_0 < n$, it is conservative, especially when n_0 is much less than n. Therefore, improving the performance of this procedure by incorporating into it an estimate of n_0 is a problem of importance.

Storey (2002) proposed a different approach to control the FDR. His approach leads to a (random) rejection threshold for each P_i given by $t_\alpha = \sup\{t : \widehat{\mathrm{FDR}}(t) \leq \alpha\}$, with $\widehat{\mathrm{FDR}}(t)$ representing an estimate of the FDR of the single-step test rejecting each H_i if $P_i \leq t$, for some fixed t, based on the available p-values. It was developed under the aforementioned mixture model 1 with independent p-values and using a conservatively biased point estimate of the FDR.

A number of conservatively biased point estimates of the FDR(t) were obtained through different estimates $\hat{\pi}_0$ of π_0, each producing a step-up procedure with critical constants of the form $i\alpha/n\hat{\pi}_0$, that is, a type of adaptive BH method (2000).

Definition 3 (Adaptive BH procedure) *Let $P = (P_1, \ldots, P_n)$ be the vector of* p*-values corresponding to n hypotheses. A step-up procedure with the set of critical values $\{i\alpha G(\mathbf{P})/n, i = 1, 2, \ldots, n\}$ is called an adaptive BH procedure, where $G(\mathbf{P}) = 1/\hat{\pi}_0$ and $\hat{\pi}_0$ is an estimator for π_0.*

Storey considered the following estimates of π_0.

1. The estimator of π_0 in Storey (2002) is

$$\hat{\pi}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)n}$$

2. The estimator of π_0 in Storey (2004) is

$$\hat{\pi}_0{}^*(\lambda) = \frac{W(\lambda)+1}{(1-\lambda)n}$$

where $W(\lambda)$ is the number of accepted hypotheses with significance threshold λ applied to p-values.

The estimates of FDR(t) corresponding to the above estimates of π_0 are:

$$\widehat{FDR}_\lambda(t)(2002) = \frac{\hat{\pi}_0(\lambda)t}{[R(t)\bigvee 1]/n}$$

and

$$\widehat{FDR}^*_\lambda(t)(2004) = \begin{cases} \frac{\hat{\pi}^*_0(\lambda)t}{[R(t)\bigvee 1]/n} & \text{if } t \leq \lambda, \\ 1 & \text{if } t > \lambda, \end{cases} \textit{respectively.}$$

Benjamini et al. (2006) suggested another estimator of π_0 as $\frac{n-R_0}{(1-\lambda)n}$, where $\lambda = \frac{\alpha}{\alpha+1}$ and R_0 is the number of rejection by a BH procedure at level λ. This procedure controls the FDR under independence.

In addition to the adaptive methods described above, the general idea of improving the BH method has led to many other adaptive methods (Benjamini and Hochberg (2000); Blanchard and Roquain (2009); Gavrilov et al. (2009); and Sarkar (2008)). Other ways to improve the BH method include incorporating information about correlations or utilizing the dependence structure into the BH method (Efron (2007); Romano and Wolf (2008); and Yekutieli and Benjamini (1999)), or generalizing the notion of FDR to k-FDR by relaxing control over at most $k-1$ false rejections (Sarkar (2007); Sarkar and Guo (2009); and Sarkar and Guo (2009)).

The key point for adaptvie BH procedure is to use data information to get a better estimate of π_0 to improve the performance of the procedure. In this dissertation, we will use the information of our group model setting to estimate such π_0 in the BH-Bonferroni procedure and also Grouped-BH procedure which we will discuss later in the following chapters.

2.1.3 Efron's Empirical Bayes Approach

Efron et al. (2001) introduced an empirical Bayes procedure in the gene microarray expression study. Based on the mixture model (Definition 1), they proposed the concept of local FDR and connect it to estimated posterior probabilities.

With the direct application of Bayes' theorem, the *posteriori* probabilities are given by:

$$p_0(x) = Prob\{Not\ Different|X_i = x\} = p_0 f_0(x)/f(x), \tag{2.1}$$

and

$$p_1(x) = Prob\{Different|X_i = x\} = 1 - p_0 f_0(x)/f(x), \tag{2.2}$$

The $p_0(x)$ has been defined as the local False Discovery Rate in their paper. Their local FDR controlling procedures are as follows:

1. Compute the *local* FDR $p_0(x)$ for each hypothesis H_i based on (2.1).
2. Compare the *local* FDR with α to make the decision. If the *local* FDR is less than α, reject the corresponding hypothesis, and verse visa.

In Efron and Tibshirani (2002), the local FDR is estimated by using the massively parallel structure of microarray data. The ratio $f_0(x)/f_1(x)$ is estimated from the empirical distributions. The estimation of p_0 and p_1 needs strong parameteric assumptions. To make sure $\hat{p_1}(x)$ to be always nonnegative, the following condition needs to be satisfied: $p_0 \leq \hat{p}_{0,max} = min_x\{\hat{f}(x)/f_0(x)\}$.

2.1.4 Sarkar and Zhou's BFDR Approach

Sarkar et al. (2008) proposed a general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. Randomized as well as non-randomized procedures controlling the Bayes false discovery rate (BFDR) and Bayes false non-discovery rate (BFNR) are developed. The control of the BFDR or BFNR is achieved through local FDR or local FNR.

Let $d = (d_1, \ldots, d_n)$, with $d_i = 0$ or 1 according as H_i is accepted or rejected, represent the decision vector. Let $h = (h_1, \ldots, h_n)$, with $h_i = 0$ or 1 according as $\theta_i \in \Theta_{i0}$ or $\theta_i \in \Theta_{i1}$, represent the unknown configuration of true or false null hypotheses. The vector $\delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \ldots, \delta_n(\mathbf{X}))$ is referred to as a multiple decision rule or multiple testing procedure. If $0 < \delta_i(\mathbf{X}) < 1$, for at least one i, then $\delta(\mathbf{X})$ is randomized; otherwise, it is non-randomized. The Bayes FDR (BFDR) and Bayes FNR (BFNR) are defined as:

$$\begin{aligned} BFDR &= E_\theta E_{\mathbf{X}|\theta}\left[\sum_d \frac{\sum_{i=1}^n d_i(1-h_i)}{\{\sum_{i=1}^n d_i\} \bigvee 1}\delta(d|\mathbf{X})\right] \\ or &= E_{\mathbf{X}} E_{\theta|\mathbf{X}}\left[\sum_d \frac{\sum_{i=1}^n d_i(1-h_i)}{\{\sum_{i=1}^n d_i\} \bigvee 1}\delta(d|\mathbf{X})\right] \end{aligned}$$

$$\begin{aligned} BFNR &= E_\theta E_{\mathbf{X}|\theta}\left[\sum_d \frac{\sum_{i=1}^n (1-d_i)h_i}{\{\sum_{i=1}^n (1-d_i)\} \bigvee 1}\delta(d|\mathbf{X})\right] \\ or &= E_{\mathbf{X}} E_{\theta|\mathbf{X}}\left[\sum_d \frac{\sum_{i=1}^n (1-d_i)h_i}{\{\sum_{i=1}^n (1-d_i)\} \bigvee 1}\delta(d|\mathbf{X})\right] \end{aligned}$$

The BFDR controlling procedures are as follows:

Theorem 2.1 *Let $K(\mathbf{X}) = max\{0 \le j \le n : A_j(\mathbf{X}) \le \alpha\}$. Then, the one-step randomized procedure δ defined as follows given $K(\mathbf{X}) = k$*

$$\delta_{i:n}(\mathbf{X}) = \begin{cases} 1 & \textit{if } i \le k \\ \frac{\alpha - A_k(\mathbf{X})}{A_{k+1}(\mathbf{X}) - A_k(\mathbf{X})} & \textit{if } i = k+1 \\ 0 & \textit{otherwise,} \end{cases} \tag{2.3}$$

with $\delta_{i:n} = 1 \forall i$ when $k = n$, controls the BFDR at α.

The test statistic they used is also local FDR proposed by Efron et al. (2001), where $r_i(\mathbf{X}) = \frac{\pi_0 f_0(\mathbf{X_i})}{f(\mathbf{X_i})} = \frac{\pi_0 f_0(\mathbf{X_i})}{\pi_0 f_0(\mathbf{X_i})+(1-\pi_0)f_1(\mathbf{X_i})}$. However, different from Efron's Bayesian approach through controlling ordered $r_i(\mathbf{X_{i:n}})$ to be less than or equal to α, BFDR procedure controls $max\{j : A_j(\mathbf{X}) \leq \alpha\}$, where $A_k(\mathbf{X}) = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{X_{i:n}})$for $k = 1, \ldots, n$. since $max\{j : A_j(\mathbf{X}) \leq \alpha\} \geq max\{j : r(\mathbf{X_{j:n}}) \leq \alpha\}$, BFDR procedure is more powerful than Efron's Bayesian procedure.

The BFNR controlling procedures are as follows:

Theorem 2.2 *Let $K(\mathbf{X}) = max\{0 \leq j \leq n : B_j(\mathbf{X}) \leq \beta\} - 1$. Then, the one-step randomized procedure δ defined as follows given $K(\mathbf{X}) = k$*

$$\delta_{i:n}(\mathbf{X}) = \begin{cases} 1 & \textit{if } i \leq k \\ \frac{B_{k-1}(\mathbf{X})-\beta}{B_{k-1}(\mathbf{X})-B_k(\mathbf{X})} & \textit{if } i = k \\ 0 & \textit{otherwise,} \end{cases} \tag{2.4}$$

with $\delta_{i:n} = 0 \forall i$ when $k = -1$, controls the BFNR at β.

Sarkar and Zhou (2008) give a procedure controlling the Bayesian directional false discovery rate (BDFDR) as an alternative to Lewis and Thayer (2004). By starting with a decision theoretic formulation of multiple testing null hypotheses against two-sided alternatives, the procedure is developed through controlling the posterior directional false discovery rate (PDFDR). They also propose a corresponding empirical Bayes method in the context of one-way random effects model.

Our proposed novel grouped hypotheses testing procedure adopts similar BFDR controlling procedure while using the test statistic of local FDR. The detailed will be discussed in Chapter 3.

2.1.5 mFDR Controlling Approaches

Sun and Cai (2007) and Sun and Cai (2009) developed a Bayesian decision theoretic approach which can yield a powerful multiple testing method not only incorporating costs of false and missed discoveries but also simultaneously addressing dependency, optimality and multiplicity. Based on the conclusion derived by Genovese and Wasserman (2001) that, the marginal false discovery rate $mFDR = E(V)/E(R)$ is asymptotically equivalent to the FDR measure in the sense that $mFDR = FDR + O(m^{-1/2})$ under weak conditions, Sun and Cai (2007) derived an oracle rule based on the z values that minimizes the FNR ($FNR = E(T/A|A > 0)Pr(A > 0)$,Genovese and Wasserman (2001) and Sarkar (2004)) subject to a constraint on the FDR and also proposed an adaptive procedure based on the z values.

Under the mixture model (Definition 1), they started with the weighted 0-1 loss function:

$$L_\lambda(\delta(\mathbf{X}, \theta) = \sum_{i=1}^{n} \{\lambda(1-\theta_i)\delta_i(\mathbf{X}) + \theta_i(1-\delta_i(\mathbf{X}))\} \tag{2.5}$$

for a decision rules $\delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \ldots, \delta_n(\mathbf{X})) \in \{0,1\}^n$, where λ is the relative cost of make a false discovery (type I error) to that of missing a true discovery(type II error) and assumed to be constant over all the hypotheses. They considered the Bayes rule associated with this loss function and showed that it is also optimal from a multiple testing view that controlling the mFDR at level α while minimizing the mFNR, where

$$mFDR = \frac{E[\sum_{i=1}^{n} \delta_i(\mathbf{X})(1-\theta_i)]}{E[\sum_{i=1}^{n} \delta_i(\mathbf{X})]} \tag{2.6}$$

$$mFNR = \frac{E[\sum_{i=1}^{n} \{1-\delta_i(\mathbf{X})\}\theta_i]}{E[\sum_{i=1}^{n} \{1-\delta_i(\mathbf{X})\}]} \tag{2.7}$$

They achieved this optimality using the test statistics $T_i(x) = Pr(\theta_i = 0|x)$ in terms of local FDR measures, and called it the oracle procedure.

According to the Bayes rule, $\delta_i(x) = 1$ when $T_i(x) \leq \frac{1}{1+\lambda}$. Then they show the oracle procedure how to control the mFDR at α in the following steps:

1. Order $T_i(x)$ increasingly by $T_{(1)}(x) \leq T_{(2)}(x) \leq \ldots T_{(R)}(x) \ldots \leq T_{(n)}(x)$.

2. Choose $R = max\{j : \frac{1}{j} \sum_{i=1}^{j} T_{(i)}(x) \leq \alpha\}$.

3. Set $\frac{1}{1+\lambda} = T_{(R)}(x)$. Reject $H_{(1)}, H_{(2)}, \ldots, H_{(R)}$.

Assuming a monotone likelihood ratio condition on the test statistic $T_i(x)$, they also constructed a data-driven version of it under independence (Sun and Cai (2007)) or Markov dependence (Sun and Cai (2009)) for the θ_i's. The numerical results show their oracle procedure outperforms other competitors, such as BH procedure (1995) and Genovese and Wasserman (2001). They also proved their data-driven procedure is asymptotically equivalent to the corresponding oracle procedure.

Sun and Cai (2009) further discussed this mFDR controlling procedure under dependence. The observed data are assumed to be generated from an underlying two-state Hidden Markov model. It is shown the procedures proposed control mFDR therefore control FDR at the desired level, enjoy certain optimality properties and are especially powerful in identifying clustered non-null cases.

He et al. (2015) reformulate Sun and Cai's compound decision theoretic framework for multiple testing by using a general loss function incorporating the type II error severity and under a general model allowing arbitrary dependence. They derive an oracle procedure taking the type II error severity into account without requiring any specific distributional properties for the underlying test statistics, and also a data-driven procedure is provided and proved to be asymptotic equivalence to the oracle procedure under independence.

2.2 Grouped Hypotheses Testing Approaches

For modern large-scale multiple hypotheses testing, we usually have thousands of hypotheses to test at the same time. In many of these cases, there is some prior information that some group structure exists among these hypotheses, or the hypotheses can be divided into subgroups based on the characteristics of the problem. Ignoring the group structures in such data analysis can be inappropriate. Testing such grouped hypotheses has been receiving more and more research interests. In this section, we will discuss some of these existing procedures, which will be recalled later in the next few chapters for the purpose of comparing them with our newly proposed procedures.

2.2.1 Heller et al. Procedure

Heller et al. (2009) posed the problem of discovering differentially expressed genes with gene sets as a multiple testing problem of grouped hypotheses. They consider selecting promising gene sets before looking for differentially expressed genes within these gene sets. They define an erroneous discovery set if the set is selected while actually there is no significant gene in the set, or if the set is appropriately selected but one of the genes in the set is erroneously discovered. They define the Overall FDR (OFDR) as the expected proportion of erroneous discoveries of gene sets out of all the selected gene sets.

2.2.2 Grouped BH Procedure

Hu et al. (2010) considered the FDR control for testing groups of hypotheses from a different perspective and named it the Grouped Benjamini-Hochberg (GBH) procedure. This procedure utilizes a weighting scheme based on a simple Bayesian argument using prior information on the proportion of true nulls among the hypotheses within each group. The weighted p- values are calculated by group based on each groups' known ratio of null-to-false null hypotheses before pulling all the weighted p-values into a single group and applying the BH procedure on them at an adjusted level $\alpha^w = \frac{\alpha}{1-\pi_0}$. Here, the group structure of the p-values are assumed to be known, but if it is unknown, the authors considered two grouping strategies, one using Gene Ontology (GO) and the other using the k-means clustering.

In the oracle case of the GBH procedure, they set the weighted p-values in a way that assumes $P^w_{g,i} = \infty$ if $\pi_{g,0} = 1$, accepting the corresponding hypotheses for these groups g. This seems to be sophisticated way to do the test and does not allow making a fair comparison with other relevant procedures in this context. When $\pi_{g,0}$ is unknown, this advantage is unachievable. And if $\pi_{g,0}$ is unknown, they propose an adaptive method to estimate it, with the corresponding adaptive GBH offering an asymptotic control of the FDR under weak dependence.. The asymptotic notion assumes $N \to \infty$(the total number of hypotheses), with the number of groups k being finite, which means that the at least one group size approaches to ∞. This is not a valid assumption in many applications. Furthermore, when $\pi_{g,0}$ for each group is not estimatable, as in a precise model setting, the weight is meaningless. Moreover, this procedure doesną́t take advantage of evidence gathered from the data for a particular group being significant.

2.2.3 CLfdr Procedure

Cai and Sun (2009) discussed the simultaneous testing of grouped hypotheses for independent data collected from heterogeneous sources from a Bayesian perspective. The multiple group model they assume is as follows: the M hypotheses are divided into G groups with prior probability π_g; the random mixture model $X \sim (1-p_g)F_{g0}(x) + p_g F_{g1}(x)$ holds separately within each group, with possibly different p_g, f_{g0} (the density of F_{g0}), and f_{g1} (the density of F_{g1}). The CLfdr procedure involves the following three steps:

1. Calculate the CLfdr values

   $$CLfdr^g(x_{gj}) = \frac{(1-p_g)f_{g0}(x_{gj})}{f_g(x_{gj})}$$

   where $g = 1, 2, \ldots, G; j = 1, 2, \ldots, m_g$.

2. Pull the CLfdr values into a single group and rank them. Denote by $CLfdr_{(1)}, \cdots, CLfdr_{(N)}$ the ranked CLfdr values with $H_{(1)}, \ldots, H_{(N)}$the corresponding hypotheses.

3. Reject all $H_{(i)}$,$i = 1, \ldots, l$, where $l = max\{i : (1/i) * \sum_{j=1}^{i} CLfdr_j \leq \alpha\}$.

Similar to GBH procedure, CLfdr procedure utilizes the external group information to calculate the CLfdr values before pulling them into a single group to to make the final decision. However, when the specific group information for each group is ambiguous, or all the groups share the same information, the CLfdr procedure would become the procedure without group information (we refer it as PLfdr procedure later).

2.2.4 Benjamini and Bogomolov's Average Error Rate Controlling Procedure

Benjamini and Bogomolov (2014) considered the adjustment for selection bias in testing multiple families of hypotheses. They consider selecting groups first and look into the selected families. However, they are controlling the expected average value of some error rate C over the selected families. They apply the selection rule S to the ensemble of sets P, identifying the selected set of families $S(P)$. Let R be the number of selected families and apply E(C)-controlling procedure in each selected family separately at new adjusted level $R\alpha/G$ (G is the number of group). This is a continued work of Benjamini et al. (2009) about selective inference in the context of confidence intervals (CIs) for the selected parameters.

2.2.5 BH-Bonferroni Procedure

Similar to Benjamini and Bogomolov (2014)'s grouped hypotheses testing idea which select significant families first then make discoveries within each selected family, Clements et al. (2011) developed two new statistical procedures to control the total FDR for group-dependent data, a two-stage BH method and an adaptive two-stage BH method. The similarity is that they both use the number of significant group to adjust for the significance level of each group; while Benjamini and Bogomolov (2014)'s method controls the expected average value of some error rate, Clements et al. (2011)'s method controls the overall FDR. In this dissertation, we refer to Clements et al. (2011)'s procedure as the BH-Bonferroni procedure. The detail of this procedure in the context of a rectangular form of data is as follows:

1. Divide the data rectangle into D by D mutually exclusive groups. The group size is $S = D^2$ and the total number of groups is $G = N/S$, where N is the total number of hypotheses.

2. Find the minimum p-value of each group, $P^{(g)}_{min}$. Using Bonferroni Combination method to find the grouped p-values for each group, $Q_g = SP^{(g)}_{min}$, for $g = 1, 2, \ldots, G$.

3. Apply the BH procedure to these grouped p-values to detect the potential significant groups.. Record the number of these significant groups as $k*_{BH}$.

4. Identify the jth individual component within the gth potential group as being significant if the corresponding p-value, say P_{gj}, is such that $SP_{gj} \leq k*_{BH}\, \alpha/G$.

They proved this BH-Bonferroni procedure controls the FDR at α if the groups are independent or positively dependent in a certain sense.

In the adaptive two-stage BH method, they use $\hat{S}_g$ instead of S when using Bonferroni Combination method to find the grouped p-values for each group, where $\hat{S}_g = min\{\frac{\sum_{j=1}^{S} I(P_{gj}>\lambda)+1}{1-\lambda}, S\}$ for some tuning parameter λ.

2.2.6 Sun and Wei's Procedure

Sun and Wei (2011)considered multiple testing of groups of hypotheses in the context of pattern identification of gene sets in microarray time-course (MTC) experiments. They formulated their problem as a set-wise multiple testing problem using a decision theoretic framework, before proposing a data-driven procedure that aims to minimize the missed set rate subject to a constraint on the false set rate. Their primary focus has been to identify gene sets that are significant in the sense of exhibiting overall temporal patterns across time-points, not to carry forward this identification process at each of the time-points for a significant gene.

They discussed that one feature of the time-course experiments is that if a gene is differentially expressed at one time point, it is very likely to remain differentially expressed at the next time point. So, the local dependency structure can be approximated by a hidden Markov model (HMM). Specifically, an HMM assumes that the temporal sequence of the underlying states (Differentially Expressed or Equally Expressed) of a particular gene form a Markov chain and the observed gene expression data are independent conditioning on the hidden states. In their analysis, they considered inhomogeneous HMMs and the hierarchical Gamma-Gamma model.

They proposed the false set rate (FSR) and missed set rate (MSR) to combine the errors in set-wise testing.

$$FSR = E\{\frac{\sum_{i=1}^{G}(1-\vartheta_i)\delta_i}{(\sum_{i=1}^{G})\bigvee 1}\}$$

and

$$MSR = E\{\frac{\sum_{i=1}^{G}\vartheta_i(1-\delta_i)}{(\sum_{i=1}^{G})\bigvee 1}\}$$

Here, a binary vector $\vartheta = (\vartheta_1, \ldots, \vartheta_m) \in \{0,1\}^m$, where $\vartheta_i = o$, if $\theta_i \in \Theta_0$ and $\vartheta_i = 1$ otherwise.

And the test statistic they used is the generalized local index of significance

(GLIS),

$$\begin{aligned} GLIS_i &= P_{\widehat{\Psi}}(\vartheta_i = 0|e_i) \\ &= \sum_{s=\{s_1,\ldots,s_K\}\in\Theta_0} P_{\widehat{\Psi}}(\theta_i = s|e_i) \end{aligned}$$

where $\widehat{\Psi}$ is the collection of all estimated parameters, ϑ_i is a binary random variable indicating whether a gene is interesting or not, s is a K-dimensional binary vector and Θ_0 is the null parameter space. Compared to the combined p-value which can only be used for testing the total number of nonnulls in a set, the GLIS statistic can be used to screen for various complicated patterns. And the maximum likelihood estimate (MLE) of $\widehat{\Psi}$ can be obtained by EM algorithm.

Our proposed Bayesian procedure will be a new multiple testing method for grouped hypotheses controlling total false discoveries (across all hypotheses) and within-group false discoveries. The aforementioned existing procedures identify the significant groups first and then make the final decisions within each significant group. Our proposed procedure, however, takes the reverse route. It screens the hypotheses within each group for potential discoveries at a significant level less than or equal to α, then at the second stage, it makes the final decision by selecting the significance groups with significant within-group discoveries by controlling the FDR at level α. We will discuss our procedure in details in the next chapter.

CHAPTER 3

THE PROPOSED METHODOLOGY AND ITS OPTIMALITY

This chapter provides our proposed new methodology for multiple testing of grouped hypotheses controlling false discoveries from Bayesian perspective. The methodology is developed by starting with the definition of false discovery rate in its posterior form across all hypotheses and then representing it in terms of posterior measures of within-group false discoveries across all significant groups. In addition to providing the validity of our proposed procedure as a false discovery rate controlling procedure, we present some optimality properties for the decisions made both within and between groups.

3.1 The Proposed Methodology

Let us first define the following with the only model assumption made at this point that the hidden states are binary random variables:

$$fdr_{j|g}(\boldsymbol{X}) = P(\theta_{j|g} = 0|\theta_g = 1, \boldsymbol{X}) \text{ and } fdr_g(\boldsymbol{X}) = P(\theta_g = 0|\boldsymbol{X}). \tag{3.1}$$

These are the local FDR scores, respectively, for the jth hypothesis in the gth group given that the group is truly significant and for the gth group itself. Then, since the posterior total FDR, denoted by $\mathrm{PFDR}_T(\boldsymbol{X})$,

$$\begin{aligned}
& \frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}(1-\theta_{gj})\delta_{gj}(\boldsymbol{X})}{\left\{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X})\right\}\vee 1} \\
= \quad & I(\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X}) > 0) - \frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\theta_{gj}\delta_{gj}(\boldsymbol{X})}{\left\{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X})\right\}\vee 1},
\end{aligned}$$

can also be expressed as follows:

$$\begin{aligned}
& I(\sum_{g}\sum_{j}\delta_{gj}(\boldsymbol{X}) > 0) - PFDR_T(\boldsymbol{X}) \\
= \quad & \frac{\sum_g\left\{\delta_g(\boldsymbol{X})P(\theta_g = 1|\boldsymbol{X})\sum_j\delta_{j|g}(\boldsymbol{X})P(\theta_{j|g} = 1|\theta_g = 1, \boldsymbol{X})\right\}}{\left(\sum_g\delta_g(\boldsymbol{X})\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\right)\vee 1} \\
= \quad & \frac{\sum_g\left\{\delta_g(\boldsymbol{X})(1 - fdr_g(\boldsymbol{X}))\sum_j\delta_{j|g}(\boldsymbol{X})(1 - fdr_{j|g})(\boldsymbol{X})\right\}}{\left(\sum_g\delta_g(\boldsymbol{X})\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\right)\vee 1}, \\
= \quad & \frac{\sum_g\left\{\delta_g(\boldsymbol{X})(1 - fdr_g(\boldsymbol{X}))[I(\sum_j\delta_{j|g}(\boldsymbol{X}) > 0) - PFDR_{W|g}(\boldsymbol{X})]\sum_j\delta_{j|g}(\boldsymbol{X})\right\}}{\left(\sum_g\delta_g(\boldsymbol{X})\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\right)\vee 1},
\end{aligned}$$

where

$$PFDR_{W|g} = \frac{\sum_j\delta_{j|g}(\boldsymbol{X})fdr_{j|g}(\boldsymbol{X})}{\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\vee 1}$$

is the posterior FDR within the gth group that is truly significant. Thus,

$$PFDR_T(\boldsymbol{X}) = \frac{\sum_g\delta_g(\boldsymbol{X})\left\{1 - (1 - fdr_g(\boldsymbol{X}))(1 - PFDR_{W|g}(\boldsymbol{X}))\right\}\sum_j\delta_{j|g}(\boldsymbol{X})}{\left(\sum_g\delta_g(\boldsymbol{X})\{\sum_j\delta_{j|g}(\boldsymbol{X})\}\right)\vee 1} \tag{3.2}$$

This leads us to our proposed method in its oracle form, as stated in the following. For notational convenience, from this point onwards we will often suppress the symbol $\boldsymbol{X}$ in the quantities that obviously depend on the data.

The Proposed Method

Step 1. For each g, let $fdr_{(1)|g} \leq fdr_{(2)|g} \cdots \leq fdr_{(m_g)|g}$ be the ordered $fdr_{j|g}$, with $H_{g(1)}, \ldots, H_{g(m_g)}$ being the corresponding hypotheses, and find

$$R_g = \max\left\{k_g : \frac{1}{k_g}\sum_{j=1}^{k_g} fdr_{(j)|g} \leq \eta\right\},$$

given $0 < \eta \leq \alpha < 1$. Mark the hypotheses $H_{g(1)}, \ldots, H_{g(R_g)}$ for possible rejection and go to the next step.

Step 2. Calculate $\eta_g = \frac{1}{R_g}\sum_{j=1}^{R_g} fdr_{(j)|g}$, and define $fdr_g^* = 1-(1-\eta_g)(1-fdr_g)$, for each g. Order these fdr_g^* values as $fdr_{(1)}^* \leq \cdots \leq fdr_{(G)}^*$, and find

$$l = \max\left\{k : \frac{\sum_{g=1}^{k} R_{(g)} fdr_{(g)}^*}{\sum_{g=1}^{k} R_{(g)}} \leq \alpha\right\},$$

with $R_{(g)}$ being the value of R for the group that corresponds to $fdr_{(g)}^*$. The hypotheses that were marked for possible rejection in the groups $(1), \ldots, (l)$ are ultimately rejected.

Theorem 3.1 *Given any $0 < \eta \leq \alpha < 1$, the proposed two-stage multiple testing method for grouped hypotheses controls the PFDR$_T$ at level α.*

Proof. A proof of this theorem is immediate, since PFDR$_T$ of the proposed method is $\sum_{g=1}^{l} R_{(g)} fdr_{(g)}^* / \sum_{g=1}^{l} R_{(g)}$, which is less than or equal to α.

Remark 1 The proposed Bayesian method is developed with special attention given to that the hypotheses are grouped, with each group having a significance probability of its own. It also takes into account the dependency, which could be naturally present or caused inherently due to grouping, between the significance of a hypothesis and that of the group containing it. More specifically, given data, (i) it measures strength of evidence towards rejection for each hypotheses within a group conditional on that for the group itself by using $1 - fdr_{j|g}$, (ii) pulls up the hypotheses with the highest average measure of conditional evidence exceeding $1 - \eta$ from each group, and (iii) then sets up a rejection rule for these selected sub-groups of hypotheses by taking into account the measures of evidence towards rejection for the respective groups subject to a control over total false discoveries at the desired level.

Let us consider testing a single group of hypotheses, say $H_{1(1)}, \ldots, H_{1(m_1)}$, assuming that $G = 1$. Here, since $\eta_1 = \frac{1}{R_1}\sum_{j=1}^{R_1} fdr_{(j)|1}$,
where $R_1 = \max\left\{k_1 : \frac{1}{k_1}\sum_{j=1}^{k_1} fdr_{(j)|1} \leq \eta\right\}$, and $fdr_1^* = fdr_1 + (1 - fdr_1)\frac{1}{R_1}\sum_{j=1}^{R_1} fdr_{(j)|1}$,
our method rejects the first R_1 of these hypotheses if

$$R_1 = \max\left\{k : \frac{1}{k}\sum_{j=1}^{k} fdr_{(j)|1} \leq \min\left(\eta, [\alpha - fdr_1]/[1 - fdr_1]\right)\right\}. \qquad (3.3)$$

Our assumption that η should be restricted within the interval $(0, \alpha]$ can be justified from (2.3), since outside this interval η has no effect on R_1. Although our method works for any $\eta \leq \alpha$, it works the best when $\eta = \alpha$ with $R_1 = \max\left\{k : \frac{1}{k}\sum_{j=1}^{k} fdr_{(j)|1} \leq [\alpha - fdr_1]/[1 - fdr_1]\right\}$. This method with $\eta = \alpha$ is slightly different from the SC (Sun and Cai (2007)) method for testing a single group of hypotheses. It actually modifies the SC method by incorporating into the method the strength of significance of the group measured using its local fdr, and thus lets the SC method to adapt itself according to the group's own significance. Our method, of course, reduces to the SC method (irrespective of η) when $m_g = 1$.

Going back to testing multiple groups of hypotheses, although our method allows η to be chosen differently for the different groups, each within the interval $(0, \alpha]$, we consider keeping the η's same. Our reason is that it allows us to use a certain portion of the overall FDR level to maintain control over within-group false discoveries. This may be desirable in some applications where one would like to attach some measure of reliability to decisions made within each group. Of course, the choice of η is subjective and can be made judiciously based on ones prior knowledge or expertise.

Nevertheless, we will be choosing $\eta = \alpha$ in the following sections on simulation studies and real-data application of our method.

3.2 Optimality of the Proposed Method

The proposed two-stage method is a composition of two multiple testing methods - one applied in Step 1 to the hypotheses within each group to discover those that are potentially significant, and the other applied in Step 2 to the groups containing these potentially significant hypotheses to discover those groups that are significant before ultimately rejecting the potentially significant hypotheses within them. Each of these methods controls a measure of false (hypothesis- or group-specific) discoveries at its own specified level. Therefore, when establishing an optimality for our two-stage method, it seems worthwhile that we do so separately for these two methods.

Our optimality results for the within-group decisions made in Step 1 and the between-group decisions made in Step 2 (given the within-group decisions in Step 1), consist of showing that these decisions are based on a Bayes rule when ranking the hypotheses or groups in terms of a measure of significance and minimize appropriate type II errors subject to controlling an appropriate measure of (hypothesis- or group-specific) false discovery rate aposteriori.

Towards formulating the loss functions for the within- and between-group decisions, we first note that $\sum_g \delta_g \theta_g \sum_j \delta_{j|g}(1-\theta_{j|g})$ and $\sum_g \delta_g \theta_g \sum_j (1-\delta_{j|g})\theta_{j|g}$ are the total numbers of type I and type II errors, respectively, associated with the decisions made within the (truly significant) groups in Step 1. For the decisions made across the groups in Step 2, given the values of $\delta_{j|g}$'s for the hypotheses within these groups as determined in Step 1, the type I and type II errors, respectively, are $\sum_g \delta_g(1-\theta_g)\sum_j \delta_{j|g} + \sum_g \delta_g \theta_g \sum_j \delta_{j|g}(1-\theta_{j|g})$ and $\sum_g (1-\delta_g)\theta_g \sum_j \delta_{j|g}\theta_{j|g}$. In other words,

$$L_1(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_g \delta_g \theta_g \left\{ \sum_j \left[\delta_{j|g}(1-\theta_{j|g}) + \lambda_g (1-\delta_{j|g})\theta_{j|g} \right] \right\} \tag{3.4}$$

is the loss function for the within-group decisions in Step 1; whereas,

$$L_2(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_g \delta_g (1-\theta_g) \sum_j \delta_{j|g} + \sum_g \delta_g \theta_g \sum_j \delta_{j|g}(1-\theta_{j|g}) + \lambda \sum_g (1-\delta_g)\theta_g \sum_j \delta_{j|g}\theta_{j|g} \tag{3.5}$$

is the conditional loss function for the between-group decisions given the decisions made within the groups.

Since

$$\begin{aligned} E(L_1|\boldsymbol{x}) \;\; = \;\; & \sum_g \delta_g (1-fdr_g) \sum_j \delta_{j|g}[fdr_{j|g} - \lambda_g(1-fdr_{j|g})] + \\ & \qquad \sum_g \delta_g(1-fdr_g)\sum_j \lambda_g(1-fdr_{j|g}), \end{aligned}$$

the within-group risk is minimized aposteriori when the $\delta_{j|g}$'s are such that $\delta_{j|g} = 1$ if and only if $fdr_{j|g} \le \lambda_g(1-fdr_{j|g})$. Thus, we note the following:

Theorem 3.2 *For the problem of deciding between $\theta_{j|g} = 0$ and $\theta_{j|g} = 1$ simultaneously for $j = 1, \ldots, m_g; g = 1, \ldots, G$ under the loss function (3.4), the following are the Bayes rules:*

$$\delta_{j|g} = \begin{cases} 1 & \textit{if} \quad fdr_{j|g} \leq \lambda_g(1 - fdr_{j|g}) \\ 0 & \textit{otherwise,} \end{cases} \tag{3.6}$$

for $j = 1, \ldots, m_g; g = 1, \ldots, G$.

For between-group decisions, it is noted that

$$\begin{aligned} E(L_2|\boldsymbol{x}) &= \sum_g \delta_g \left[1 - (1 - fdr_g)(1 - \eta_g) - \lambda(1 - fdr_g)(1 - \eta_g)\right] R_g + \\ &\qquad \lambda \sum_g (1 - fdr_g)(1 - \eta_g) R_g. \end{aligned}$$

Given (η_g, R_g), for $g = 1, \ldots, G$, it is minimized by the δ_g's for which $\delta_g = 1$ if and only if $1 - (1 - fdr_g)(1 - \eta_g) \leq \lambda(1 - fdr_g)(1 - \eta_g)$. In other words, we have the following:

Theorem 3.3 *For the problem of deciding between $\theta_g = 0$ and $\theta_g = 1$, given (η_g, R_g), simultaneously for $g = 1, \ldots, G$ under the loss function (3.5), the following are the Bayes decision rules:*

$$\delta_g = \begin{cases} 1 & \textit{if}\ 1 - (1 - fdr_g)(1 - \eta_g) \leq \lambda(1 - fdr_g)(1 - \eta_g) \\ 0 & \textit{otherwise,} \end{cases} \tag{3.7}$$

for $g = 1, \ldots, G$.

Remark 2 Theorems 3.2 and 3.3 provide optimal paths for developing a two-stage multiple testing procedure for grouped hypotheses which we talked about in Section 2. More specifically, Theorem 3.2 suggests that when determining the significance of the hypotheses within each group subject to a control over a within-group measure of false discoveries, the hypotheses within that group should be ranked according to the values of $fdr_{j|g}/(1-fdr_{j|g})$, or equivalently the values of $fdr_{j|g}$, before determining a threshold guaranteeing the desired control over that measure of false discoveries. When it comes to determining the significance of the groups, given the information about the hypotheses within them obtained in Step 1, subject to controlling a measure of false discoveries across the groups, Theorem 3.3 suggests that the groups should be ranked according to the values of $[1-(1-fdr_g)(1-\eta_g)]/(1-fdr_g)(1-\eta_g)$, or equivalently the values of $1-(1-fdr_g)(1-\eta_g)$ before a threshold is determined subject to controlling that measure of false discoveries.

Now, with these within- and between- group level decision rules, we have the following two theorems.

Theorem 3.4 *For a two-stage decision rule $\boldsymbol{\delta}$, let the posterior false discovery rate and the expected type II errors for the multiple testing method applied in Step 1 to identify the significant hypotheses in group g be defined, respectively, by $PFDR_{W|g}(\boldsymbol{\delta})$ (as in (3.2)) and as follows:*

$$\begin{aligned} TypeII_{W|g}(\boldsymbol{\delta}) &= E\left\{\sum_j (1-\delta_{j|g})\theta_g\theta_{j|g}\middle|\boldsymbol{x}\right\} \\ &= (1-fdr_g)\sum_j (1-\delta_{j|g})(1-fdr_{j|g}). \end{aligned} \tag{3.8}$$

Then, with $\boldsymbol{\delta}$ denoting our two-stage rule, we have $TypeII_{W|g}(\boldsymbol{\delta}) \leq TypeII_{W|g}(\boldsymbol{\delta}')$ for any other two-stage rule $\boldsymbol{\delta}'$ satisfying $PFDR_{W|g}(\boldsymbol{\delta}') \leq PFDR_{W|g}(\boldsymbol{\delta})$.

Theorem 3.5 *For a two-stage decision rule $\boldsymbol{\delta}$, let*

$$PFDR_{T|W}(\boldsymbol{\delta}) = \frac{\sum_g \delta_g \left[1-(1-fdr_g)(1-\eta_g)\right] R_g}{\left(\sum_g \delta_g R_g\right) \vee 1}, \tag{3.9}$$

and

$$\begin{aligned} TypeII_{T|W}(\boldsymbol{\delta}) &= E\left\{\sum_g (1-\delta_g)\theta_g(1-\eta_g)R_g \middle| \boldsymbol{x}\right\} \\ &= \sum_g (1-\delta_g)(1-fdr_g)(1-\eta_g)R_g, \end{aligned} \tag{3.10}$$

be respectively the false discovery rate and the expected type II errors, a posteriori, for the multiple testing method applied in Step 2 to the groups containing the hypotheses marked as potentially significant in Step 1 to identify those groups that are significant. Then, given that Step 2 is restricted to the same set of groups containing the non-zero potentially significant hypotheses, we have, with δ denoting our two-stage rule, $TypeII_{T|W}(\boldsymbol{\delta}) \leq TypeII_{T|W}(\boldsymbol{\delta}')$ for any other two-stage decision rule $\boldsymbol{\delta}'$ for which $PFDR_{T|W}(\boldsymbol{\delta}') \leq PFDR_{T|W}(\boldsymbol{\delta})$.

CHAPTER 4

NUMERICAL STUDIES AND A REAL DATA APPLICATION

We conducted numerical studies to examine how well our proposed Bayesian method performs in comparison with its relevant competitors in their oracle and data-driven forms under certain model assumptions. In this chapter, we present these model assumptions, the other competing methods, and the results of the simulation studies. We also apply this method to a real data application under the BSG model assumption.

Our simulations involve EM algorithm, which were run through the Owl's nest, a Linux cluster for high-performance computing at Temple University. High-performance computing allows us to run simulations parallelly across different scenarios, with each scenario taking an exclusive processor and bandwidth to increase the speed. The simulations were done on the ”highmem” subsection, the most powerful section within the Owl's nest. The running time varies for different scenarios. Calculations with independent Bernoulli models took less time while those with Hidden Markov Model took much more. The actual running time varied from 5 to 48 hours with 500 replications.

4.1 Model Settings

Consider the following model for $(X_{gj}, \theta_g, \theta_{j|g})$, $g = 1, \ldots, G$; $j = 1, \ldots, m_g$,

$$\begin{aligned}
&\theta_g \overset{i.i.d.}{\sim} Bernoulli(\pi_1), \\
&\theta_{1|g}, \ldots, \theta_{m_g|g}|\theta_g = 0 \overset{i.i.d}{\sim} Bernoulli\ (0)\ (i.e.,\ P(\theta_{j|g} = 0|\theta_g = 0) = 1), \\
&\theta_{1|g}, \ldots, \theta_{m_g|g}|\theta_g = 1 \sim \frac{\prod_{j=1}^{m_g} \left\{(1-\pi_{2|1})^{1-\theta_{j|g}}\pi_{2|1}^{\theta_{j|g}}\right\} I(\sum_j \theta_{j|g} > 0)}{1-(1-\pi_{2|1})^{m_g}} \\
&[Truncated\ Bernoulli\ (\pi_{2|1})] \\
&or \\
&\theta_{j|g}|\theta_g = 1 \sim Hidden\ Markov\ Model\ (p_h^1, a_{kh}^j)\ where\ k, h \in \{0, 1\}, \\
&X_{gj} \mid \theta_{gj} \overset{ind}{\sim} (1-\theta_{gj})f_0(x_{gj}) + \theta_{gj}f_1(x_{gj}),\ for\ some\ given\ densities\ f_0\ and\ f_1\ .
\end{aligned} \tag{4.1}$$

For simplicity, we call the model with Truncated **B**ernoulli hidden states within each **S**ignificant **G**roup as **BSG** model and the model with **M**arkov hidden states within each **S**ignificant **G**roup as **MSG** model. For the BSG model, the $\pi_{2|1}$'s are assumed to be the same across all the significant groups. In the MSG model, $p_h^1 = P(\theta_{g1} = h)$ is the initial probability, and $a_{kh}^j = P(\theta_{j|g} = h|\theta_{j-1|g} = k)$ is the transition probability that doesn't depend on g. If a_{kh}^j is the same across all j's, it is called the homogeneous MSG model; otherwise, it is called the in-homogeneous MSG model.

These models reflect group-dependence, that is, the underlying dependence exists within but not between groups, which is often referred to as clumpy or weak dependence and assumed in many multiple testing applications. The BSG model would be adequate if one would like to focus only on the group structure, whereas the MSG model pays attention to group as well as dependence structures.

When all the parameters are known, it is easy to calculate $fdr_{j|g}$ and fdr_g for the BSG model. As shown in Appendix A.1, with $\boldsymbol{x} = (x_{11}, \ldots, x_{Gm_G})$ and $\boldsymbol{x}_g = (x_{g1}, \cdots, x_{gm_g})$, these are, respectively,

$$\begin{aligned} & fdr_{j|g} && (4.2) \\ = \ & P(\theta_{j|g} = 0 | \theta_g = 1, \boldsymbol{x}) \\ = \ & \frac{f(\theta_{j|g} = 0, \boldsymbol{x} | \theta_g = 1)}{f(\theta_{j|g} = 0, \boldsymbol{x} | \theta_g = 1) + f(\theta_{j|g} = 1, \boldsymbol{x} | \theta_g = 1)} \\ = \ & \frac{(1-\pi_{2|1}) f_0(x_j) \prod_{k(\neq j)=1}^{m_g} \{(1-\pi_{2|1}) f_0(x_k) + \pi_{2|1} f_1(x_k)\} - (1-\pi_{2|1})^{m_g} \prod_{k=1}^{m_g} f_0(x_k)}{\prod_{k=1}^{m_g} \{(1-\pi_{2|1}) f_0(x_k) + \pi_{2|1} f_1(x_k)\} - (1-\pi_{2|1})^{m_g} \prod_{k=1}^{m_g} f_0(x_k)} \\ & && (4.3) \end{aligned}$$

and

$$fdr_g = P(\theta_g = 0 | \boldsymbol{x}_g) = \frac{(1-\pi_1) f(\boldsymbol{x}_g | \theta_g = 0)}{(1-\pi_1) f(\boldsymbol{x}_g | \theta_g = 0) + \pi_1 f(\boldsymbol{x}_g | \theta_g = 1)}, \qquad (4.4)$$

where

$$f(\boldsymbol{x}_g | \theta_g = 0) = \prod_{j=1}^{m_g} f_0(x_{gj} | \theta_{gj} = 0)$$

and

$$f(\boldsymbol{x}_g | \theta_g = 1) = \frac{\prod_{j=1}^{m_g} \{(1-\pi_{2|1}) f_0(x_j) + \pi_{2|1} f_1(x_j)\} - (1-\pi_{2|1})^{m_g} \prod_{j=1}^{m_g} f_0(x_j)}{1 - (1-\pi_{2|1})^{m_g}}.$$

For the MSG model, we use forward-backward algorithm to calculate the local FDR scores (Bilmes (1998)). Define $\alpha_{gj}(l) = P(x_{g1}, x_{g2}, \ldots, x_{gj}, \theta_{j|g} =$

$l|\theta_g = 1)$ and $\beta_{gj}(l) = P(x_{gj+1}, \dots, x_{gm_g}|\theta_{j|g} = l, \theta_g = 1)$ (for $l = 0, 1$). Then, the local FDR values for the MSG model are given by

$$P(\theta_{j|g} = 0|\boldsymbol{x}, \theta_g = 1) = \frac{\alpha_{gj}(0)\beta_{gj}(0)}{\alpha_{gj}(0)\beta_{gj}(0) + \alpha_{gj}(1)\beta_{gj}(1)}$$

and

$$P(\theta_g = 0|\boldsymbol{x}_g) = \frac{(1-\pi_1)\prod_{j=1}^{m_g} f_0(x_{gj})}{(1-\pi_1)\prod_{j=1}^{m_g} f_0(x_{gj}) + \pi_1 f(\boldsymbol{x}_g|\theta_g = 1)},$$

where $f(\boldsymbol{x}_g|\theta_g = 1) = \alpha_{gj}(0)\beta_{gj}(0) + \alpha_{gj}(1)\beta_{gj}(1)$.

4.2 Competing Methods

The most relevant, alternative approaches against which ours should be compared are the ones that ignore the group structure and perform multiple testing by pulling the hypotheses in a single group. With that in mind, we consider the following two methods and present them in their oracle forms under the above model setting.

The SC Method (Sun and Cai (2007)). This is the most naive decision theoretic approach one could take in the present context. Here, the hypotheses are pulled into a single group before applying a multiple testing method developed from a decision theoretic point of view. Since it ignores the group structure of the hypotheses, it should be the most relevant one to compare with for any decision theoretic approach specifically designed to capture the group and/or the associated dependence structure. Let

$$PLfdr_{gj}(\boldsymbol{x}) \quad = \quad Pr(\theta_{gj} = 0|\boldsymbol{x}) = \frac{P(\theta_{gj} = 0)f_0(x_{gj})}{f(x_{gj})},$$

the pulled local FDR score for each hypothesis. Let $PLfdr_{(1)} \leq \cdots \leq PLfdr_{(N)}$ ($N = \sum_g m_g$) be the ordered $PLfdr$ values, with $H_{(1)}, \ldots, H_{(N)}$ being the corresponding hypotheses. Find

$$k = max\left\{i : \frac{1}{i}\sum_{i=1}^{k} PLfdr_{(i)} \leq \alpha\right\},$$

and reject $H_{(i)}$ for all $i = 1, \ldots, k$.

Under the BSG model, we have

$$PLfdr_{gj}(\boldsymbol{x}) = \frac{(1-\pi_1\pi_{2|1})f_0(x_{gj})}{(1-\pi_1\pi_{2|1})f_0(x_{gj}) + \pi_1\pi_{2|1}f_1(x_{gj})}.$$

Under the MSG model, we can use the forward-backward algorithm to calculate $P(\theta_{gj} = 0)$ and $P(\theta_{gj} = 1)$ for each g and j, and then get the PLfdr.

The Adaptive BH Method. (Benjamini and Hochberg (2000)). Let each X_{gj} be transformed to its p-value P_{gj}. Let $P_{(1)} \leq \cdots \leq P_{(N)}$ be the ordered versions of the these p-values when they are pulled into a single group. Compute

$$k = max\left\{i : (1-\pi_1\pi_{2|1})P_{(i)} \leq \frac{i\alpha}{N}\right\}.$$

If such a k exists, then reject the hypotheses associated with $P_{(1)}, \ldots, P_{(k)}$; otherwise, do not reject any hypotheses.

Under the BSG model, $P(\theta_{gj} = 0) = 1 - \pi_1\pi_{2|1}$; under the MSG model, $P(\theta_{gj} = 0)$ is calculated using the forward-backward algorithm.

Remark 3 It should be noted that Sun and Cai (2009) introduced a multiple-group version of the SC method. However, it was proposed under a model setting that is different from ours. Similarly, Hu et al. (2010) developed an adaptive BH method for grouped hypotheses, but again it relies on different model assumptions. In particular, in these two approaches, π_1 has been assumed to be one.

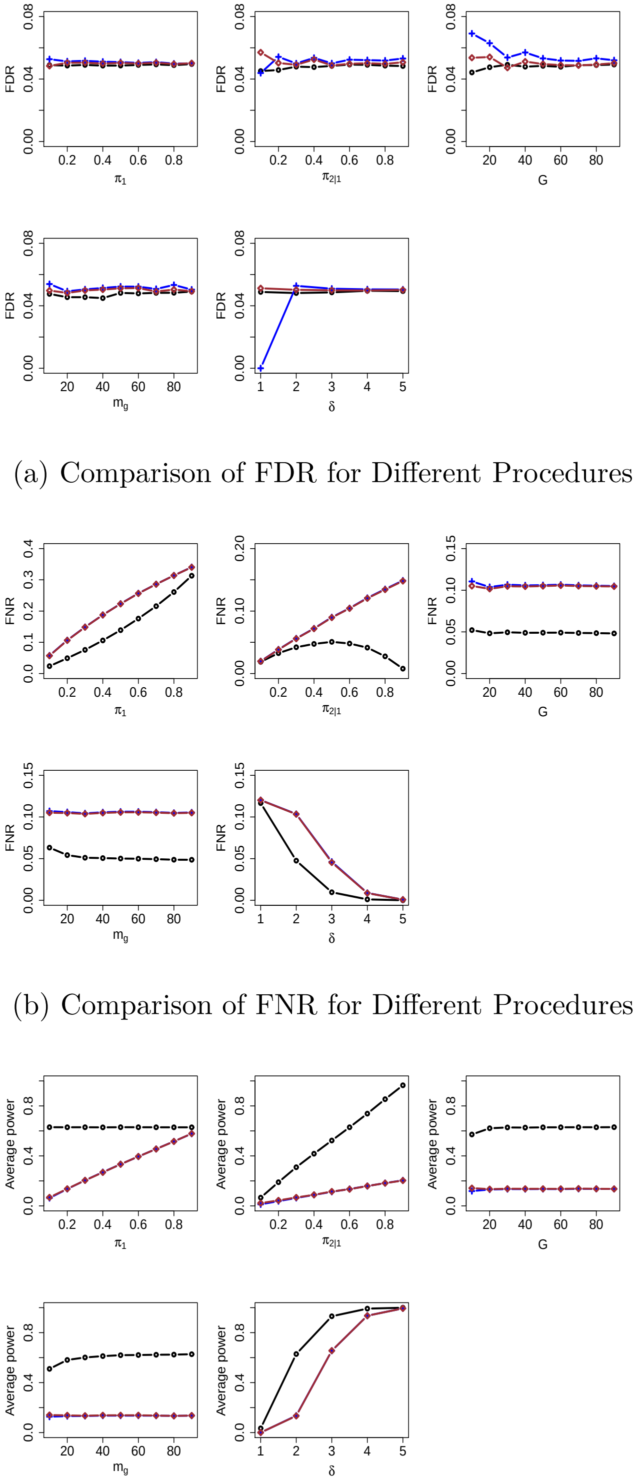
4.3 Oracle Comparison

We now present the results of our numerical studies conducted under both the BSG and MSG models. We examine the performance of our method relative to its aforementioned competitors in terms of FDR control and power as defined through the FNR (the expected proportion of false acceptances among all the accepted hypothesis) as well as the Average Power (the expected proportion of truly rejected hypothesis).

BSG Model. Set $f_0(x) \equiv \phi(x)$, the pdf of $N(0,1)$, $f_1(x) = \phi(x-\delta)$, and $\alpha = 0.05$. There are five unknown quantities, $\pi_1, \pi_{2|1}$, δ, G, and m_g. The value of $\eta = \alpha$ was set at 0.05. The simulated values of FDR, FNR and Average Power were calculated based on 500 runs for each of these three methods having chosen some values for these quantities.

Figure 4.1 shows the comparison of the four methods in terms of simulated FDR, FNR and Average Power. The basic values chosen for the unknown quantities are $\pi_1 = 0.2, \pi_{2|1} = 0.6, G = 100, m_g = 100$, and $\delta = 2$. In each figure, we allow the value of one of these quantities to vary, holding the other quantities at the aforementioned values. As seen from the graphs, our proposed procedure can perform significantly better than the other two methods in all cases.

MSG Model. For the reason of simplicity, we assume the homogeneous MSG model with the initial probability of p_1^1 and the transition matrix $(1 - a_{01}, a_{01}; 1 - a_{11}, a_{11})$ with $\theta_g = 1$. The simulated values of FDR, FNR and Average Power were calculated based on 500 runs for each of the three methods and some chosen values of the unknown quantities $G, m_g, p_1^1, \pi_1, a_{01}, a_{11}, \delta$, and



(a) Comparison of FDR for Different Procedures

(b) Comparison of FNR for Different Procedures

(c) Comparison of Average Power for Different Procedures

Figure 4.1: Simulation results for the proposed(—○—), the SC(—+—), and the Adaptive BH(—◇—) procedures in their oracle forms under the BSG model.

η. We chose $G = 50, m_g = 100, p_1^1 = 0.5, \pi_1 = 0.2, a_{01} = 0.4, a_{11} = 0.8, \delta = 2$, and $\eta = 0.025$ as the basic values for these variables. The FDR, FNR and Average Power of the four methods are plotted in Figure 4.2 where we allow just one quantity to vary, holding the others at the aforementioned values, in each graph. From Figure 4.2, we can see that in most cases, our procedure has superior performance over the other two methods.

It is clearly demonstrated in this section that our proposed Bayesian method can successfully capture the group structure and outperform its competitors when all the model parameters are known. To see how this method performs when the parameters are unknown, we derive a data-driven version of it in the next section.

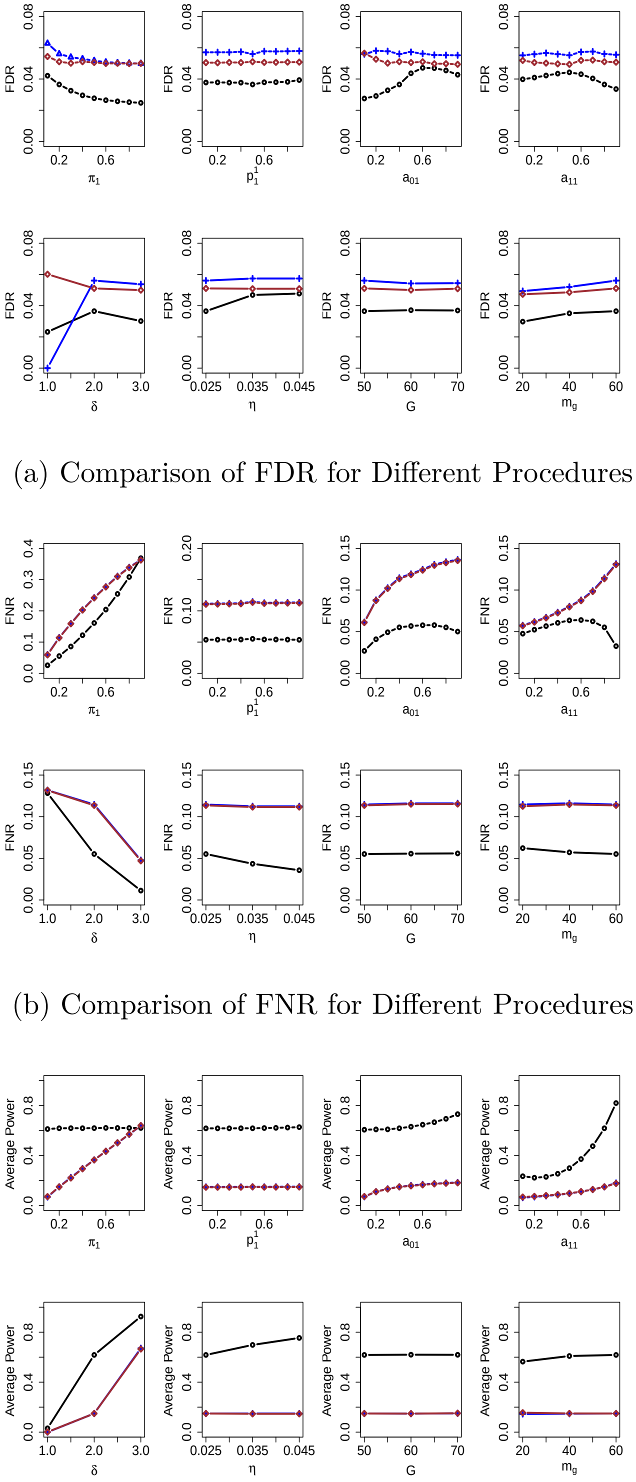
4.4 Data-driven method

In the oracle version of our proposed method, the distribution of the data under the alternative hypothesis is assumed to have a canonical form $f_1(x_{gj})$. In this section, however, we model it as an L-component mixture of normal distributions, that is, we assume $X_{gj}|\theta_{gj} \sim (1-\theta_{gj})f_0(x_{gj}) + \theta_{gj}f_1(x_{gj})$, with $f_0(x) \equiv \phi(x)$ and

$$f_1(x) = \sum_{l=1}^{L} c_l \frac{1}{\sigma_l} \phi\left(\frac{x-\mu_l}{\sigma_l}\right),$$

while proposing a data-driven version of the method. The data-driven method is derived by replacing the parameters in our proposed oracle method under this distributional setting by their estimates that are obtained by the EM algorithm.

In the following, we outline the EM algorithm steps for estimating the



(a) Comparison of FDR for Different Procedures

(b) Comparison of FNR for Different Procedures

(c) Comparison of Average Power for Different Procedures

Figure 4.2: Simulation Results for proposed(—○—), the SC(—+—), and the Adaptive BH(—◇—) procedures in their oracle form under the MSG model.

parameters and present the results of simulation studies assessing the performance of the resulting data-driven method against its competitors under each of the BSG and MSG models.

4.4.1 BSG Model

Parameter Estimation

The within- and between-group local FDRs are given by (4.2) and (4.4), respectively. Let $(\boldsymbol{x}, \boldsymbol{\theta})$ be the complete observation, where $\boldsymbol{\theta} = (\theta_g, \theta_{j|g}, g = 1, 2, \ldots, G, j = 1, 2, \ldots, m_g)$. Let $\boldsymbol{\beta} = (\pi_1, \pi_{2|1}, c_l, \mu_l, \sigma_l^2)$ be the parameter and $\boldsymbol{\beta}' = (\pi_1', \pi_{2|g}', c_l', \mu_l', \sigma_l^{2'})$ be the value of the parameter in the current iteration.

By maximizing the Q function, the expected value of the complete-data log-likelihood $l(\boldsymbol{x}, \boldsymbol{\theta})$, with respect to the unknown $\boldsymbol{\theta}$ given the observed data $\boldsymbol{x}$ and $\boldsymbol{\beta}'$, we can update the parameter $\boldsymbol{\beta}$ as in the following (see Appendix A.3 for more details):

$$
\begin{aligned}
\pi_1^{new} &= 1 - \frac{\sum_g P(\theta_g = 0|\boldsymbol{x}, \boldsymbol{\beta}')}{G} = 1 - \frac{\sum_g fdr_g(\boldsymbol{\beta}')}{G} \\
\pi_{2|1}^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1|\boldsymbol{x}, \boldsymbol{\beta}')} \\
c_l^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_{j|g} = 1, \theta_g = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1|\boldsymbol{x}, \boldsymbol{\beta}')} \\
\mu_l^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} x_{gj} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')} \\
\sigma_l^{2new} &= \frac{\sum_g \sum_{j=1}^{m_g} (x_{gj} - \mu_l^{new})^2 P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')},
\end{aligned}
$$

where

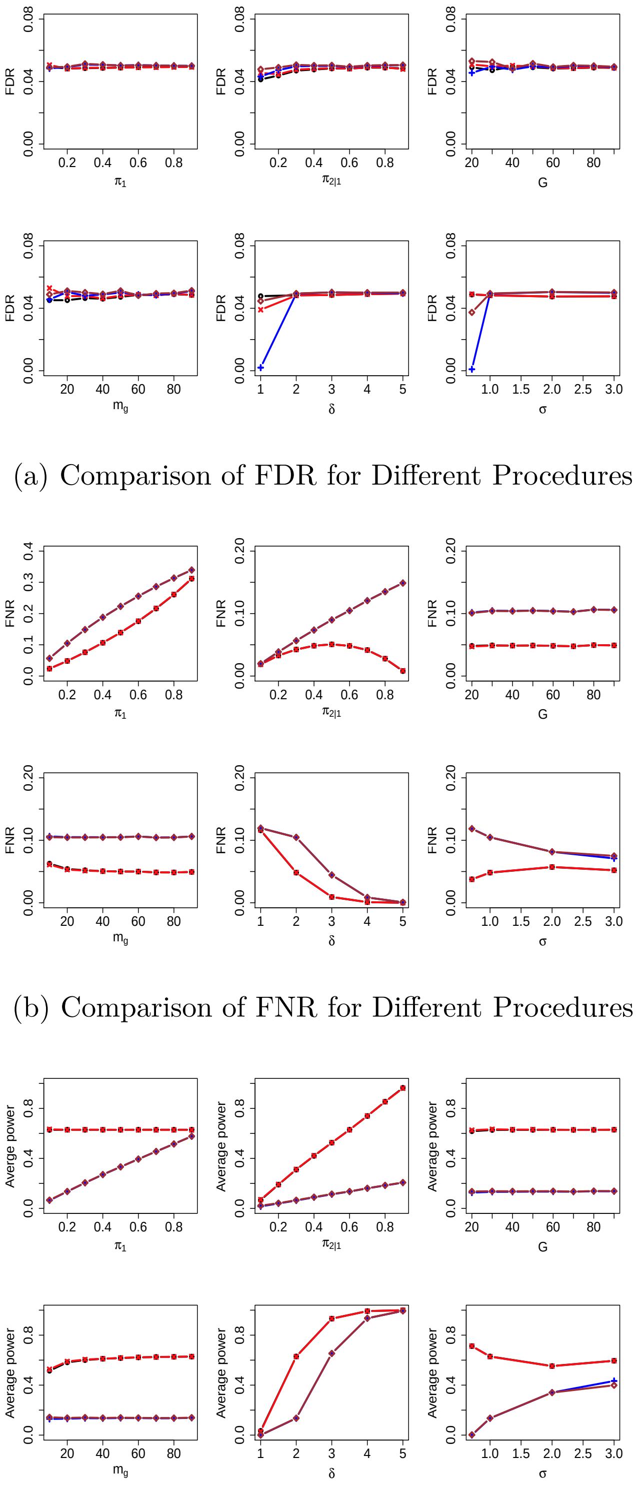
$$
\begin{aligned}
& P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\
= \; & P(\theta_{j|g} = 1, m_{j|g} = l|\theta_g = 1, \boldsymbol{x}, \boldsymbol{\beta}')P(\theta_g = 1|\boldsymbol{x}, \boldsymbol{\beta}') \\
= \; & \frac{c_l f_l(x_{gj}|\boldsymbol{\beta}')}{(1-\pi_{2|1})f_0(x_{gj}|\boldsymbol{\beta}') + \pi_{2|1}f_1(x_{gj}|\boldsymbol{\beta}')}\left[1 - fdr_g(\boldsymbol{\beta}')\right],
\end{aligned}
$$

and $fdr_g(\beta') = p(\theta_g = 0|\boldsymbol{x}, \boldsymbol{\beta}')$. Here $m_{j|g} = l$ means that x_{gj} is generated from $N(\mu_l, \sigma_l^2)$.

Simulation Results

We simulated the FDR, FNR and Average Power, as in the oracle comparison, for each of the two methods with different values of the unknown quantities π_1, $\pi_{2|1}$, G, m_g, δ, σ, and L. When $L = 1$, the basic values chosen for these quantities were: $\pi_1 = 0.2$, $\pi_{2|1} = 0.6$, $G = 100$, $m_g = 100$, $\delta = 2$, and $\sigma = 1$. When $L = 2$, we set $c_1 = c_2 = 0.5$, $\delta_2 = -2$, and $\sigma_2 = 1$. The values of α and η were set at 0.05 and 0.025, respectively. For each setting, the simulated values were based on 500 runs.

The simulation results are displayed in Figures 4.3 (L=1) and 4.4 (L=2). In each graph, we allow the value of one of the above quantities to vary while holding the others at the aforementioned values. As seen from these figures: (i) The performance of the data-driven method is very close to that of the oracle version, (ii) the overall FDR of all the procedures are controlled at the desired level α in all the scenarios considered, and (iii) the average power of the proposed method is the highest in most cases (as seen from Figures 4.3c and 4.4c).

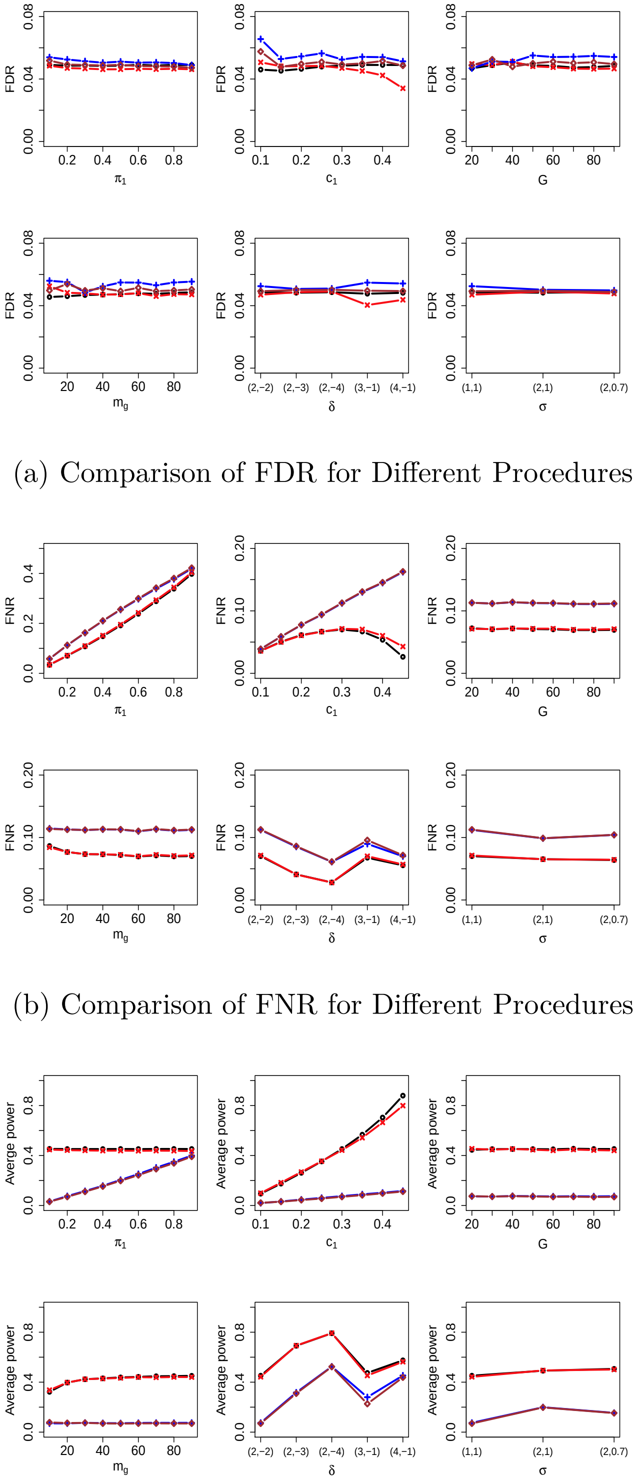


(a) Comparison of FDR for Different Procedures

(b) Comparison of FNR for Different Procedures

(c) Comparison of Average Power for Different Procedures

Figure 4.3: Simulation results for our data-driven(—×—), the SC(—+—), and the Adaptive BH(—◇—) procedures with parameters estimated by EM algorithm and our oracle procedure(—○—) with $L = 1$ under the BSG model.



(a) Comparison of FDR for Different Procedures

(b) Comparison of FNR for Different Procedures

(c) Comparison of Average Power for Different Procedures

Figure 4.4: Simulation results for our data-driven(—×—), the SC(—+—), and the Adaptive BH(—◇—) procedures with parameters estimated by EM algorithm and our oracle procedure(—○—) with $L = 2$ under the BSG model.

4.4.2 MSG Model

Parameter Estimation

Under the MSG model, the parameter to be estimated is $\boldsymbol{\beta} = (\pi_1, p_h^1, a_{k,h}^j, c_l, \mu_l, \sigma_l^2)$. Let $\boldsymbol{\beta}'$ be the value of the parameter at the current iteration. As shown in Appendix A.3, the parameter $\boldsymbol{\beta}$ can be updated iteratively as follows:

1.
$$\pi_1^{new} = 1 - \frac{\sum_g P(\theta_g = 0|\boldsymbol{x}, \boldsymbol{\beta}')}{G} = 1 - \frac{\sum_g fdr_g(\boldsymbol{\beta}')}{G}.$$

2.
$$p_0^{1new} = \frac{\sum_g P(\theta_g = 1, \theta_{1|g} = 0|x, \boldsymbol{\beta}')}{\sum_g P(\theta_g = 1|x, \boldsymbol{\beta}')} = \frac{\sum_g fdr_{1|g}(\boldsymbol{\beta}')\,[1 - fdr_g(\boldsymbol{\beta}')]}{\sum_g (1 - fdr_g(\boldsymbol{\beta}'))}.$$

3. For homogeneous transition probabilities,
$$a_{k,h}^{new} = \frac{\sum_g \sum_{j=2}^{m_g} P(\theta_g = 1, \theta_{j-1|g} = k, \theta_{j|g} = h|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=2}^{m_g} P(\theta_g = 1, \theta_{j-1|g} = k|\boldsymbol{x}, \boldsymbol{\beta}')},$$
and for inhomogeneous transition probabilities,
$$a_{k,h}^{j,new} = \frac{\sum_g P(\theta_g = 1, \theta_{j-1|g} = k, \theta_{j|g} = h|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g P(\theta_g = 1, \theta_{j-1|g} = k|\boldsymbol{x}, \boldsymbol{\beta}')}.$$

4. For $l = 1, 2, \ldots, L$,
$$c_l^{new} = \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1|\boldsymbol{x}, \boldsymbol{\beta}')}.$$

5. For $l = 1, 2, \ldots, L$,
$$\mu_l^{new} = \frac{\sum_g \sum_{j=1}^{m_g} x_{gj} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')},$$
$$\sigma_l^{2new} = \frac{\sum_g \sum_{j=1}^{m_g} (x_{gj} - \mu_l^{new})^2 P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{gj1} = l|\boldsymbol{x}, \boldsymbol{\beta}')},$$

where

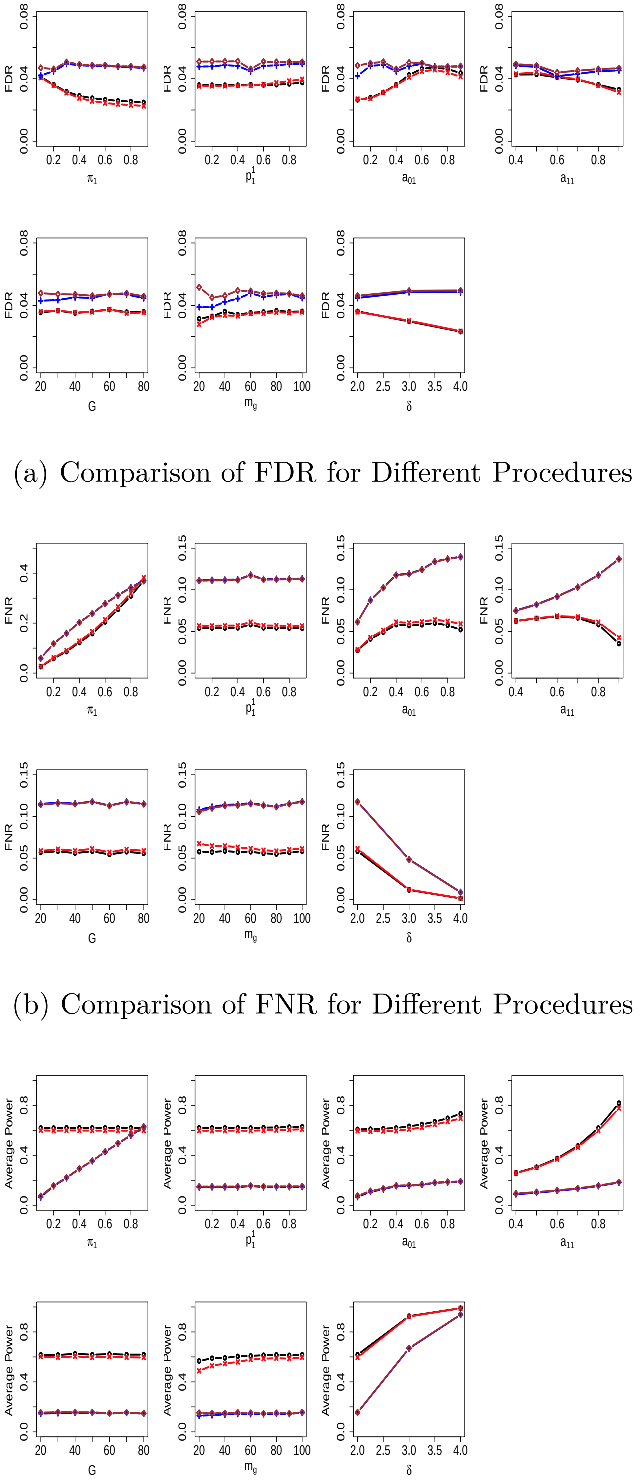
$$
\begin{aligned}
&P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\
=&P(\theta_{j|g} = 1, m_{j|g} = l|\theta_g = 1, \boldsymbol{x}, \beta')P(\theta_g = 1|\boldsymbol{x}, \boldsymbol{\beta}') \\
=&\left[1 - fdr_{j|g}(\boldsymbol{\beta}')\right] \frac{c_l^{new}\frac{1}{\sigma_l^{new}}\phi(\frac{x-\mu_l^{new}}{\sigma_l^{new}})}{\sum_l c_l^{new}\frac{1}{\sigma_l^{new}}\phi(\frac{x-\mu_l^{new}}{\sigma_l^{new}})} \left[1 - fdr_g(\boldsymbol{\beta}')\right].
\end{aligned}
$$

Simulation Results

We simulated values of FDR, FNR and Average Power for the four different procedures obtained by estimating the parameters using the aforementioned EM algorithm. The basic values of the unknown quantities were chosen as follows: (i) when $L = 1$, $\pi_1 = 0.2$, $p_1^1 = 0.5$, $G = 50$, $m_g = 100$, $a_{01} = 0.4$, $a_{11} = 0.8$, $\delta = 2$, and $\sigma = 1$, and (ii) when $L = 2$, $m_g = 50$, $c_1 = c_2 = 0.5$, $\delta_2 = -2$, and $\sigma_2 = 1$, in addition to the values in (i). We used 500 repetitions for each simulated value.

The simulated values for all the procedures, including ours in its oracle form, are displayed in Figures 4.5 and 4.6. Similar to the BSG model, the estimation of the parameters for MSG model using EM algorithm also yields a good data-driven procedure, producing similar results as its oracle version. Similar to previous simulation results, we can see that: (i) The performance of the data-driven method is very close to that of the oracle version, (ii) the overall FDR of all the procedures are controlled at the desired level α in all the scenarios considered, and (iii) the average power of the proposed method is the highest in most cases.

In summary, as demonstrated through extensive simulation studies, the proposed method seems to outperform its competitors by effectively capturing

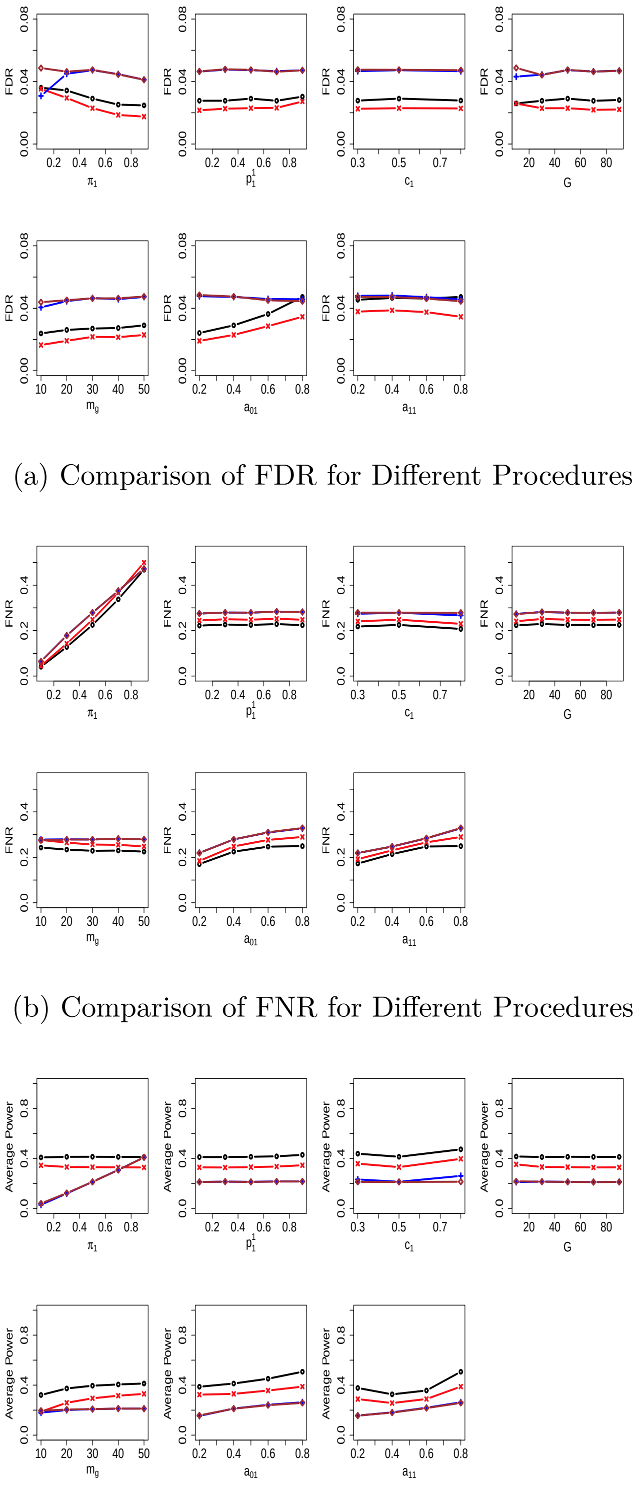


(a) Comparison of FDR for Different Procedures

(b) Comparison of FNR for Different Procedures

(c) Comparison of Average Power for Different Procedures

Figure 4.5: Simulation results for our data-driven(—×—), the SC(—+—), and the Adaptive BH(—◇—) procedures with parameters estimated by EM algorithm and our oracle procedure(—○—) with $L = 1$ under the MSG model.



(a) Comparison of FDR for Different Procedures

(b) Comparison of FNR for Different Procedures

(c) Comparison of Average Power for Different Procedures

Figure 4.6: Simulation results for our data-driven(—×—), the SC(—+—), and the Adaptive BH(—◇—) procedures with parameters estimated by EM algorithm and our oracle procedure(—○—) with $L = 2$ under the MSG model.

the group structure and also the dependence within the groups. We, therefore, strongly recommend it for application to multiple testing of grouped hypotheses.

4.5 AYP Study Under Truncated Independent Bernoulli Model

We apply our proposed Bayesian multiple testing method for grouped hypotheses to a real data set assuming the BSG model. We take up the adequate yearly progress (AYP) study of California elementary schools in 2013 (http://www.cde.ca.gov/ta/ac/ay/aypdatafiles.asp) comparing the academic performance for socioeconomically advantaged (SEA) against socioeconomically disadvantaged (SED) students in the elementary schools. We compare the success rates in Math exams of SEA versus SED students. Although it is generally the case that the average success rate of SEA students is higher than SED students, our focus is in discovering the schools with unusually small or large advantaged-disadvantaged performance differences, and also to identify the school districts with such schools. Such identifications could be of use to policy makers towards implementing social policies that can promote overall performance of students.

Let p_{1i} and p_{2i} be the success rates and n_{1i} and n_{2i} be the numbers of students in the groups of SEA and SED students, respectively, in the ith school, $i = 1, \ldots, N$. Similar to Sun and Cai (2009) and Efron (2008), a

α=0.05 and η=0.05			
Procedures	School Discoveries	Group Decisions	School District Discoveries
Proposed	736	Yes	224
SC	471	No	
Adaptive BH	410	No	
α=0.1 and η=0.1			
Proposed	1085	Yes	284
SC	668	No	
Adaptive BH	629	No	

Table 4.1: Number of Discoveries Made by the Three Procedures

z-value for school i is computed according to

$$z_i = \frac{p_{1i} - p_{2i} - \tau}{\sqrt{p_{1i}(1-p_{1i})/n_{1i} + p_{2i}(1-p_{2i})/n_{2i}}},$$

where τ is the overall difference, median (p_{1i})- median (p_{2i}), which 18.4% in this AYP study. There are 4118 ($= N$) elementary schools and 701 qualified school districts (defined as having at least 20 students in each category and $|z| < 10$ for each school). We consider these school districts as the groups in our application.

We apply the data-driven versions of the proposed, the SC and the adaptive BH methods, assuming that $f_0(x) = \phi(x)$ and $f_1(x)$ is a mixture of two normal distributions each with variance 1. Using the EM algorithm discussed in Section 5, the estimated proportion of group significance $\widehat{\pi_1}$ is seen to be about 0.53, the estimated proportion of within-group significance $\widehat{\pi_{2|1}}$ is about 0.59, and the estimated f_1 is $\widehat{f_1(x)} \sim 0.21N(2.64, 1) + 0.79N(-1.88, 1)$. We chose two different values, 0.05 and 0.10, for $\eta = \alpha$.

The numbers of discoveries made by the three methods are shown in Table 4.1. As seen from this table, our proposed method can identify more unusual schools having extremely small or large academic performance difference between SEA and SED students than the other methods. Our discoveries of schools within each district seem statistically more informative than those made by the other methods that, unlike ours, don't attempt to control within-group false discoveries, and so could potentially be of value to district level education policy makers.

CHAPTER 5

SOME ALTERNATIVE FREQUENTIST'S METHODS

Benjamini and Hochberg (1995) proposed the BH procedure as a widely used False Discovery Rate (FDR) controlling method in large-scale multiple testings arising in modern scientific investigations. It controls the FDR at the desired level α, when the p-values are independent or positively dependent in a certain sense. More specifically, the FDR of the BH procedure equals to $\pi_0\alpha$ when the p-values are independent, and is less than $\pi_0\alpha$ when the p-values are positively regression dependent on subset of null p-values (Benjamini and Yekutieli (2001); Sarkar (2002)), where π_0 is the (true) proportion of null hypotheses. The difference between $\pi_0\alpha$ and the FDR gets larger and larger with increasing (positive) dependence among p-values.

Sharpening the FDR control of the BH procedure and thereby improving its performance has been one of the most attractive research topics in modern multiple testing. An important path of research in that direction is adapting the BH procedure to data through estimating parameters, such as π_0 and/or correlation or a measure of dependence that directly affects the FDR control, and appropriately incorporating them into the BH procedure (Benjamini and Hochberg (2000); Benjamini et al. (2006); Blanchard and Roquain (2009); Gavrilov et al. (2009); Sarkar (2008); Storey (2002); Storey et al. (2004); Yekutieli and Benjamini (1999); Efron (2007); and Romano and Wolf (2008)). Some researchers have taken another path, which is to relax control over a few more than one false rejection, based on the argument that such a measure seems more relevant than the FDR in presence of dependence (Sarkar (2007); Sarkar and Guo (2009); and Sarkar and Guo (2010)).

Extending the BH procedure from single to multiple groups of hypotheses, when such groups are available or can be created, and maintaining control over the overall FDR across all hypotheses falls in the general domain of improving the BH procedure through its adaptive version, as it leads to adapting the BH procedure to the underlying group structure. However, even though a few of such extensions have been put forward in the literature (for instance, Guo et al. (2009); Clements et al. (2011); Clements et al. (2014); and Guo and Sarkar (2012)), they do not seem to present natural extensions of the BH procedure from single to multiple groups. Moreover, they do not provide much improvements over the BH procedure.

The Bayesian perspective taken in the last chapter towards developing an FDR controlling procedure for grouped hypotheses yields an approach that appears to be a natural extension of the method (for instance, Cai and Sun (2009)) from single to multiple groups, but operates differently from the aforementioned procedures developed from a frequentist perspective. Motivated by this, we consider switching our attention in this chapter on developing FDR controlling procedures for grouped hypotheses from Bayesian to frequentist perspective, and attempt to propose some newer procedures extending the BH procedure from single to multiple, or at least two, groups.

5.1 Investigating the FDR control of BB's Expected Average Error Rate Control Method

Before we discuss some new frequentist methods for multiple testing of grouped hypotheses, let us first review Benjamini and Bogomolov (2014)'s Expected Average Error Rate control method and and investigate the possibility of using at as an FDR controlling procedure. This procedure, we refer to as BB's procedure, works as follows:

Step 1 : Apply a selection rule S to the ensemble of sets $\mathbf{P}$, identifying the selected set of families $S(\mathbf{P})$. Let R be the number of selected families (i.e. $R = |S(\mathbf{P})|$).

Step 2 : Given an error rate C and an E(C) controlling procedure being used while testing the hypotheses in each family, apply this E(C) controlling procedure in each selected family separately at level $R\alpha/m$.

Then, we have,

Theorem : For any error rate $\mathrm{E}(C)$ such that C takes values in a countable set, suppose that we have a testing procedure that can control $\mathrm{E}(C)$ at any desired level α under the dependence structure of the p-value, within a family. If the p-values in each family are independent of the p-values in any other family, then for any simple selection rule $S(\mathbf{P})$, the sBB's procedure guarantees $E(C_S) \leq \alpha$.

Remark : $E(C_S) = E\left[\sum_{i \in S(\mathbf{P})} C_i / max\{|S(P)|, 1\}\right]$.

In the general procedure, they didn't specify the selection rule S, however, in their application analysis of voxelwise genome-wide association study by using the study of Stein et al. (2010), they used Sime's combination p-value($min \frac{P_{(j)} m_g}{j}$) as the group p-value, and applied the BH procedure at level 0.05 to select potentially significant families. Let R be the number of such selected significant families. Then within each family, they applied the BH procedure at level $R * 0.05/G$.

What interests us is to investigate if this procedure can control the overall FDR across all hypotheses, and so we performed a simulation study. Figure (5.1) presenting the results of this study does seem to indicate that the FDR may be controlled by this procedure. We do not make any attempt to theoretical validate this assertion, but take it as a numerical support for a promising future research direction.

5.2 Proposed Procedure 1

Now, let us propose some frequentist methods that would be meaningful alternatives to the BH method to control the overall FDR in grouped hypothesis setting. The first proposed procedure is as follows.

1. Apply the BH method to each group at level α. Find $B = \sum_{g=1}^{G} I(R_g > 0)$.

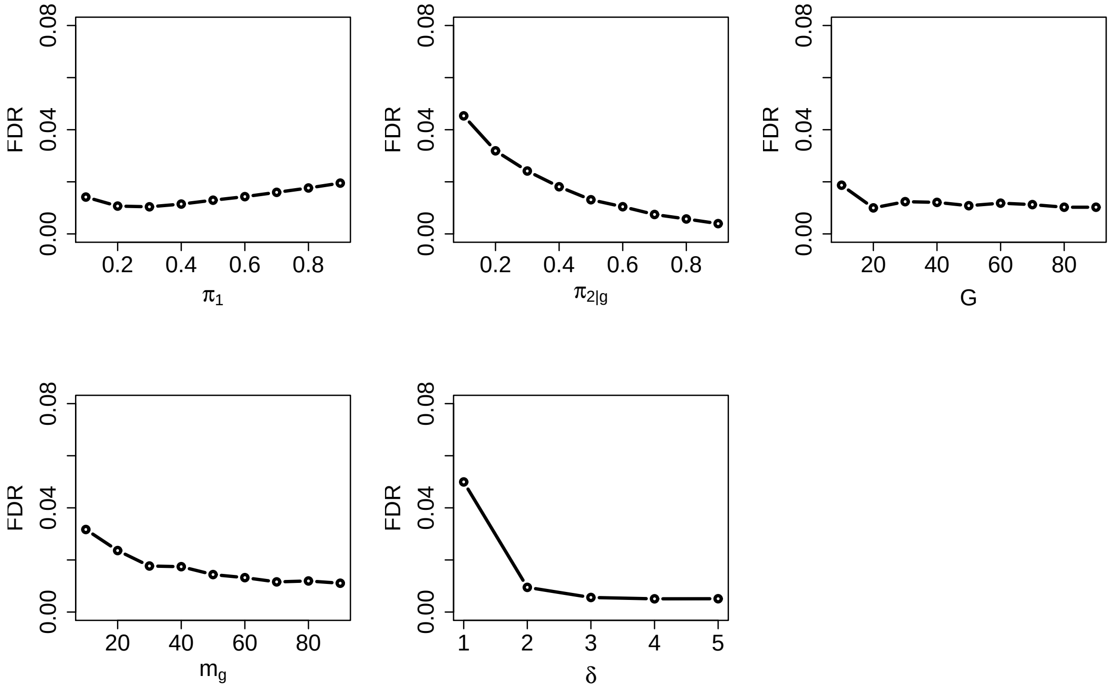


Figure 5.1: Simulation results for BB's method on overall FDR control.

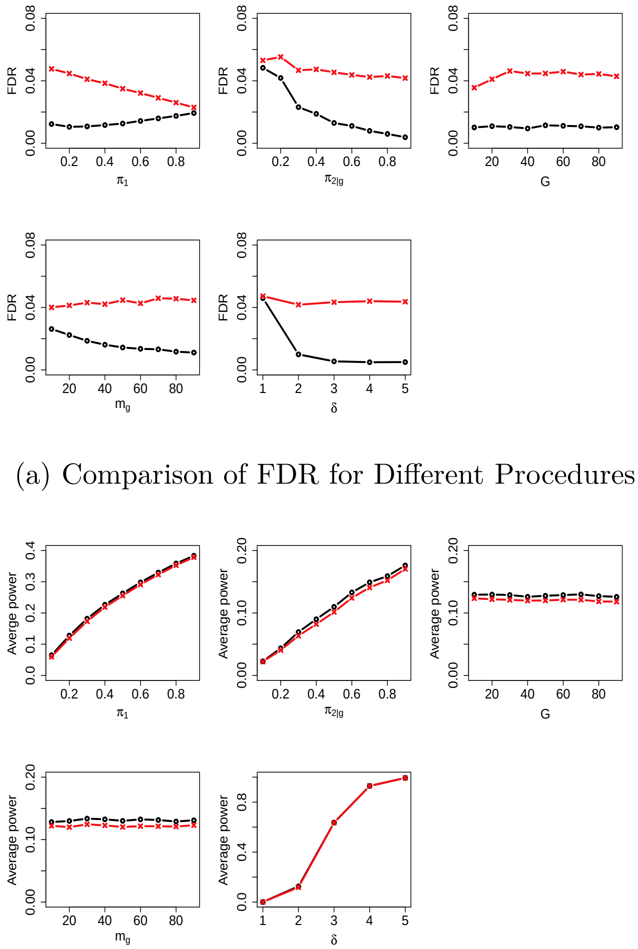
2. Apply the BH method again to each group at adjusted level $\frac{B}{G}\alpha$. Update $B = \sum_{g=1}^{G} I(R_g > 0)$.

3. Repeat Step 2 until B converges.

We conduct numerical studies when the within group hypotheses are independent, i.e. the BSG model in the Bayesian method. The simulation result-s(5.2) show this proposed method could control the overall FDR at the desired α level, and its performance is a little better than the regular BH method. The basic values for this simulation are the same as in the BSG model, which are $L = 1$, $\pi_1 = 0.2$, $\pi_{2|1} = 0.6$, $G = 100$, $m_g = 100$, $\delta = 2$, $\sigma = 1$ and $\rho = 0$.

We further investigate the performance of our proposed procedure when the within-group hypotheses are correlated. We change the correlation from 0.1 to 0.9, and the simulation results (5.3) show the overall FDR in all scenarios are still controlled at the desired α level. However, as the FDR control for this procedure is more conservative than the regular BH procedure, our power is not as good as the regular BH procedure when $\rho > 0.2$ and the discrepancy is increasing when the correlation increases.

It would be interesting to check whether this proposed procedure can control the expected average error rate as defined in BB's paper. As we notice, both of our proposed Procedure 1 and BB's method control within-group at adjusted level $B\alpha/G$ although how to select significant groups are different between the two methods. Our method update B as the number of groups that contains positive R_g, and we find the final B through iteration. Whereas, while applying their method, Benjamini and Mogomolv use Simes combination of p-values to represent the group p-values before applying the BH method to them to obtain this B as the number of significant groups. It is calculated



(a) Comparison of FDR for Different Procedures

(b) Comparison of Average Power for Different Procedures

Figure 5.2: Simulation results for our 1st proposed(—○—), and BH method(—×—) procedures with $L = 1$ under the BSG model.

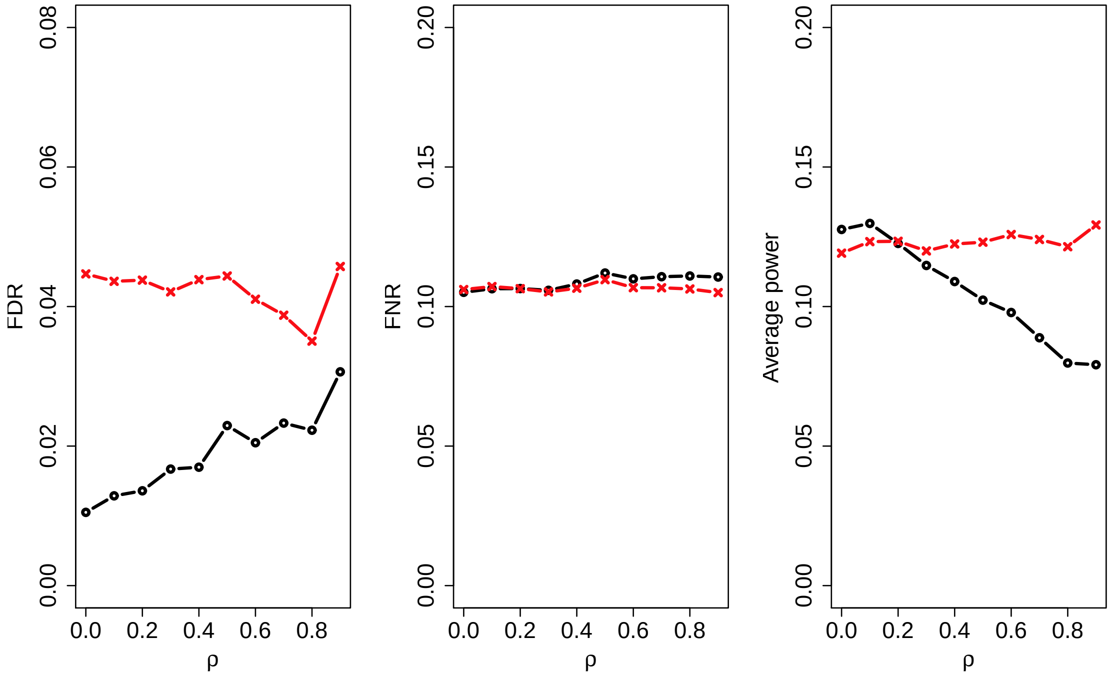


Figure 5.3: Simulation results at different correlation ρ for the 1st proposed(—○—), the regular BH(—×—) methods.

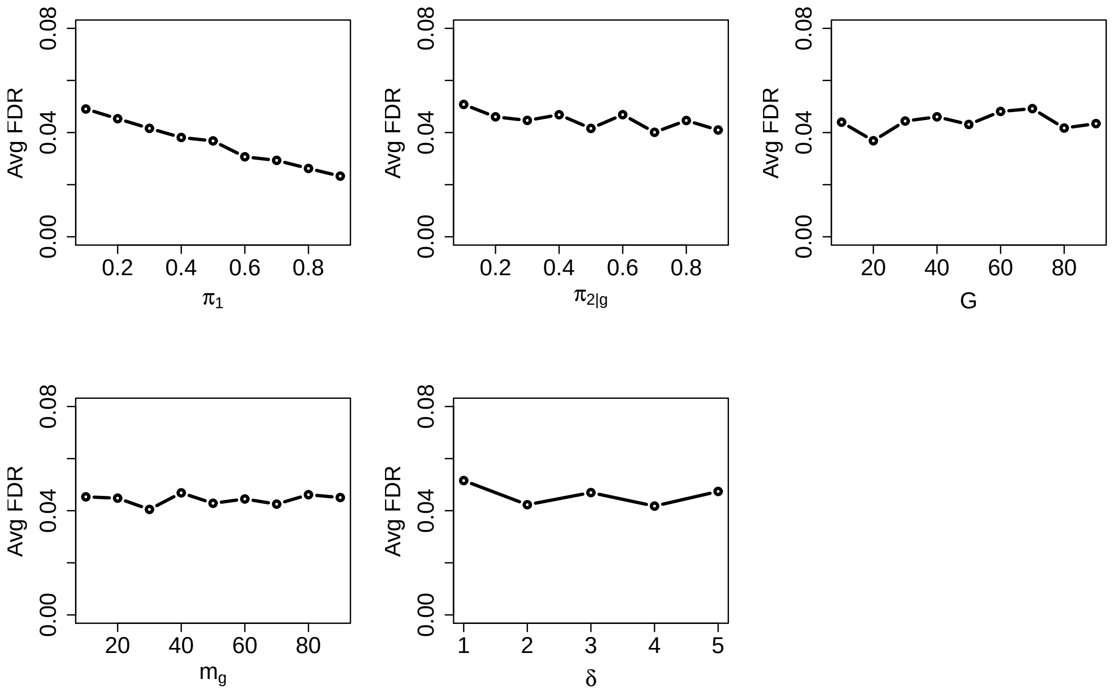


Figure 5.4: Simulation results for Proposed Procedure 1 on Expected Average FDR Control.

and fixed at the first stage. Nevertheless, our proposed Procedure 1 seems to qualify as an E(C) controlling procedure as well, which is confirmed by our simulation results(5.4).

5.3 Proposed Procedure 2

When we extend the BH method from a single group to multiple groups, it is natural for us to think about the following procedure. We consider it when $G = 2$. Let $\tilde{P}_g = m_g min_{1 \le j \le m_g} \frac{P_{g(j)}}{j}$, where $P_{g(1)} \le \cdots \le P_{g(m_g)}$ are the ordered p-values in the gth group, for $g = 1, \ldots, G$. Let $N = \sum_{g=1}^{G} m_g$, and the Proposed Procedure 2 is defined as follows:

Case 1. $\tilde{P}_1 \le \frac{m_1\alpha}{N}$ and $\tilde{P}_2 \ge \alpha$,

Reject $H_{1(j)}$ for all $j \le R_1 = \max\left\{1 \le j \le m_1 : P_{1(j)} \le \frac{j\alpha}{N}\right\}$.

Case 2. $\tilde{P}_2 \le \frac{m_2\alpha}{N}$ and $\tilde{P}_1 \ge \alpha$,

Reject $H_{2(j)}$ for all $j \le R_2 = \max\left\{1 \le j \le m_2 : P_{2(j)} \le \frac{j\alpha}{N}\right\}$.

Case 3. $\tilde{P}_1$,$\tilde{P}_2 \le \alpha$,

Reject $H_{1(j)}$,$H_{2(k)}$ for all $(j, k) \le (R_1, R_2)$, where

$(R_1, R_2) = \max\left\{(1, 1) \le (j, k) \le (m_1, m_2) : P_{1(j)}, P_{2(k)} \le \frac{(j+k)\alpha}{N}\right\}$.

Note: The (R_1, R_2) in Case 3 are related to the R_1 and R_2 defined in Case 1 and 2 as follows: Find R_1 as in Case 1, then find $R_2(R_1) = \max\left\{1 \le j \le m_2 : P_{2(j)} \le \frac{(R_1+j)\alpha}{N}\right\}$; or find R_2 as in Case 2, and then find $R_1(R_2) = \max\left\{1 \le j \le m_1 : P_{1(j)} \le \frac{(j+R_2)\alpha}{N}\right\}$.

By applying this procedure, when there are two groups, we can see from the simulation results that the total FDR can be controlled at the desired α level in the all scenarios. However, as the FDR control(5.5) of this procedure is more conservative than the regular BH method in many scenarios, the power performance(5.6) is close to or a little worse than the regular BH method in these cases. We will explore and improve its performance when there are more groups in the future.

5.4 AYP Application by Proposed Method 1

Since our simulation shows the proposed frequentist's method 1 might control the overall FDR at the desired α level, we apply this method to

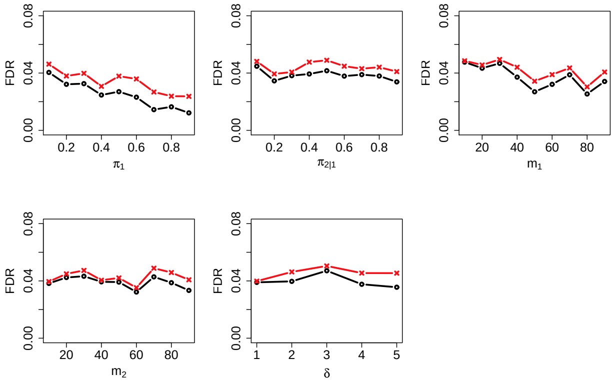


Figure 5.5: Simulation results of FDR control for our 2nd proposed(—○—), and the regular BH method(—×—) procedures with $L = 1$ under the BSG model.

our AYP data to see the performance of this method and the regular BH method. We use the estimates from EM algorithm we got from Chapter 4 ($\widehat{\pi_1} = 0.53, \widehat{\pi_{2|1}} = 0.59$), and $(1 - \pi_1 * \pi_{2|1})$ is used to derive the adaptive version of our proposed method and regular BH method.

After we apply our proposed adaptive frequentist's method 1 to the AYP data, we have 634 school discoveries (the adaptive BH method has 410 school discoveries) when $\alpha = 0.05$ and 1052 school discoveries (the adaptive BH method has 629 discoveries) when $\alpha = 0.1$. This performance is consistent with our numerical studies when within-group hypotheses are independent.

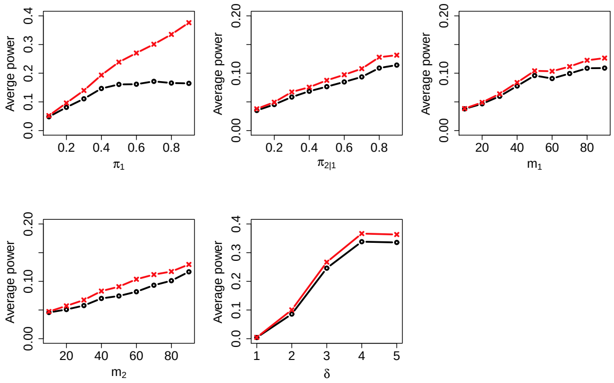


Figure 5.6: Simulation results of Average Power for our 2nd proposed(—○—), and the regular BH method(—×—) procedures with $L = 1$ under the BSG model.

CHAPTER 6

SUMMARY AND FUTURE WORK

When testing grouped hypotheses, how overall false discoveries across all hypotheses is intertwined with false discoveries of hypotheses within each group seems fundamental to deeper understanding towards effectively capturing the underlying group structure. This is an important issue that has not been answered in the literature, as far as we know. This dissertation presents a theoretical framework built on this fundamental understanding from a Bayesian viewpoint, and develops a new approach to multiple testing of grouped hypotheses that allows one to maintain some specific control over within-group false discoveries while controlling the overall false discoveries across all hypotheses. Having a separate control over within-group false discoveries, we argue, is often an effective way of capturing the underlying group structure when testing grouped hypotheses, particularly when there is high positive dependencies within groups. Moreover, this is often desired in some applications, such as in analyzing the AYP data in Chapter 4 where discovering schools within a school district controlling a district specific false discovery rate seems practically more useful than discovering these schools through a global discovery process controlling a global false discovery rate. It allows making statistically more reliable district level decisions for policy makers. Of course, the choice of the level η at which within-group false discoveries is to be controlled is subjective and can be made judiciously based on ones prior knowledge in terms of how stringent that control should be. We also make some attempts on some alternative frequentists' methods to provide a multiple-group and improved version of the single-group Benjamini and Hochberg (1995) method. Although these attempts are not fully developed, they will be useful for the future research in this area.

For future work, we can devote to the universal optimality for the proposed Bayesian method based on current conditional optimality. Working on developing frequentistąŕs multiple-group multiple testing methods would be our next and important goal.

An R-package, called "GroupTest", which is developed to carry out the numerical calculations associated with our proposed Bayesian method in this dissertation is made available at

`http://astro.temple.edu/~zhaozhg/software.html`.

REFERENCES

Arbeitman, M., E. Furlong, F. Imam, E. Johnson, B. Null, B. Baker, M. Krasnow, M. Scott, R. Davis, and K. White (2002). Gene expression during the life cycle of drosophila melanogaster. *Science 297*, 2270–2275.

Benjamini, Y. and M. Bogomolov (2014). Adjusting for selection bias in testing multiple families of hypotheses. *Journal of the Royal Statistical Society.B 76*, 297–318.

Benjamini, Y. and R. Heller (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association 102*, 1272–1281.

Benjamini, Y., R. Heller, and D. Yekutieli (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A 367*, 1–17.

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

Benjamini, Y., K. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*, 491–507.

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

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute 4*(510), 126.

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

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

California (2013). `http://www.cde.ca.gov/ta/ac/ay/aypdatafiles.asp`.

Calvano, S., W. Xiao, D. Richards, R. Felciano, H. Baker, R. Cho, R. Chen, B. Brownstein, J. Cobb, S. Tschoeke, C. Miller-Graziano, L. Moldawer, M. Mindrinos, R. Davis, R. Tompkins, and S. Lowry (2005). A network-based analysis of systemic inflammation in humans. *Nature 437*, 1032–1037.

Churchill, G. (1992). Hidden markov chains and the analysis of genome structure. *Computers and Chemistry 16(2)*, 107–115.

Clements, N., S. K. Sarkar, and W. Guo (2011). Astronomical transient detection using grouped p-values and controlling the false discovery rate. *Statistical Challengers in Modern Astronomy V*, 383ÍC396.

Clements, N., S. K. Sarkar, Z. Zhao, and D. Kim (2014). Applying multiple testing procedures to detect changes in east african vegetation. *Annals of Applied Statistics 8*, 286–308.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association 102*, 93–103.

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

Efron, B. and R. Tibshirani (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology 23*, 70–86.

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

Ephraim, Y. and N. Merhav (2002). Hidden markov processes. *IEEE Transactions on Information Theory 48*, 1518–1569.

Gavrilov, Y., Y. Benjamini, and S. K. Sarkar (2009). An adaptive step-down procedure with proven fdr control. *Annals of Statistics 37*, 619–629.

Genovese, C. and L. Wasserman (2001). Operating characteristics and extensions of the false discovery rate procedure. *Journal o f the Royal Statistical Society, Series B 64*, 499–517.

Guo, W., L. He, and S. Sarkar (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics 42*, 1070–1101.

Guo, W. and S. Sarkar (2012). Adaptive controls of fwer and fdr under block dependence. *Biometrika x*, 1–23.

Guo, W., S. Sarkar, and S. Peddada (2009). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics 66*, 485–492.

He, L., S. Sarkar, and Z. Zhao (2015). Capturing the severity of type ii errors in high-dimensional multiple testing. *Journal of Multivariate Analysis 142*, 106–116.

Heller, R., E. Manduchi, G. G.R., and W. Ewens (2009). A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics 25*, 1019–1205.

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

Korn, E., J. Troendle, L. Mcshane, and R. Simon (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*, 379–398.

Krogh, A., M. Brown, I. Mian, K. Sjolander, and D. Hausder (1994). Hidden markov models in computational biology applications to protein modeling. *Journal of Molecular Biology 235*, 1501–1531.

Lewis, C. and D. T. Thayer (2004). A loss function related to the fdr for random effects multiple comparison. *Journal of Statistical Planning and Inference 125*, 49–58.

Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *Journal of the American Statistical Association 99*, 1002–1014.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 257–285.

Romano, J. P.and Shaikh, A. M. and M. Wolf (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST 17*, 417–442.

Sarkar, S. and W. Guo (2010). Procedures controlling generalized false discovery rate using bivariate distributions of the null p-values. *Statistica Sinica 20*, 1227–1238.

Sarkar, S. and T. Zhou (2008). Controlling bayes directional false discovery rate in random effects model. *Journal of Statistical Planning and Inference 138*, 682–693.

Sarkar, S., T. Zhou, and D. Ghosh (2008). A general decision theoretic formulation of procedures controlling fdr and fnr from a bayesian perspective. *Statist. Sinica 18*, 925–946.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals o f Statistics 30*, 239–257.

Sarkar, S. K. (2004). Fdr-controlling procedures and their false neagtives rates. *Journal of Statistical Planning and Inference 125*, 119 –139.

Sarkar, S. K. (2007). Stepup procedures controlling generalized fwer and generalized fdr. *The Annals of Statistics 35*(6), 2405ÍC2420.

Sarkar, S. K. (2008). On methods controlling the false discovery rate(with discussion). *Sankhya, Ser. A 70*, 135–168.

Sarkar, S. K. and W. Guo (2009). On a generalized false discovery rate. *Annals of Statistics 37*, 1545 –1565.

Stein, J., X. Hua, S. Lee, A. Ho, A.J. Leow, A. Toga, A. Saykin, L. Shen, T. Foround, N. Pankratz, M. Huentelman, D. Craig, J. Gerber, A. Allen, J. Corneveaux, B. DeChairo, S. Potkin, M. Weiner, and P. Thompson (2010). Voxelwise genome-wide association study (vg-was). *NeuroImage 53*, 1160–1174.

Storey, J. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B 64*, 479–498.

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

Storey, J. D., J. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Soc. B 66*, 187–205.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*, 15545–15550.

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

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 Z. Wei (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association 106*(493), 73–88.

Yekutieli, D. and Y. Benjamini (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference 82*, 171–196.

APPENDIX

A.1 Local FDR scores in Chapter 4.1

Let $f_{\theta_{k|g}}(x_j) = (1-\theta_{k|g})f_0(x_k) + \theta_{k|g}f_1(x_k)$, $\tilde{\theta} = (\theta_{1|g}, \ldots, \theta_{m_g|g})$, and $\Omega = \{0,1\}^m \backslash (0, \ldots, 0)$. Then,

$$
\begin{aligned}
f(\theta_{j|g} = 0, \boldsymbol{x}|\theta_g = 1) = &\sum_{\theta_{j|g}=0, \tilde{\theta}\in\Omega} \left\{ \prod_{k=1}^{m_g} f_{\theta_{k|g}}(x_k) \right\} \frac{(1-\pi_{2|1})^{m_g - \sum_k \theta_{k|g}} \pi_{2|1}^{\sum_k \theta_{k|g}}}{1-(1-\pi_{2|1})^{m_g}} \\
=&(1-\pi_{2|1})f_0(x_j) \sum_{\theta_{j|g}=0, \tilde{\theta}\in\Omega} \left\{ \prod_{k=1, k\neq j}^{m_g} f_{\theta_{k|g}}(x_k) \right\} \frac{(1-\pi_{2|1})^{m_g - \sum_k \theta_{k|g}} \pi_{2|1}^{\sum_k \theta_{k|g}}}{1-(1-\pi_{2|1})^{m_g}} \\
=&\frac{(1-\pi_{2|1})f_0(x_j) \left[\prod_{k=1,\neq j}^{m_g} \{(1-\pi_{2|1})f_0(x_k) + \pi_{2|1}f_1(x_k)\} - (1-\pi_{2|1})^{m_g-1} \prod_{k=1,\neq j}^{m_g} f_0(x_k) \right]}{1-(1-\pi_{2|1})^{m_g}}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&f(\boldsymbol{x}|\theta_g = 1) \\
&= \sum_{\tilde{\theta}\in\Omega} \left\{ \prod_{k=1}^{m_g} f_{\theta_{k|g}}(x_j) \right\} \frac{(1-\pi_{2|1})^{m_g - \sum_k \theta_{k|g}} \pi_{2|1}^{\sum_k \theta_{k|g}}}{1-(1-\pi_{2|1})^{m_g}} \\
&= \frac{\prod_{k=1}^{m_g} \{(1-\pi_{2|1}) f_0(x_k) + \pi_{2|1} f_1(x_k)\} - (1-\pi_{2|1})^{m_g} \prod_{k=1}^{m_g} f_0(x_k)}{1-(1-\pi_{2|1})^{m_g}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
fdr_{j|g} &= \frac{f(\theta_{j|g} = 0, \boldsymbol{x}|\theta_g = 1)}{f(\boldsymbol{x}|\theta_g = 1)} \\
&= \frac{(1-\pi_{2|1}) f_0(x_j) \prod_{k=1,k\neq j}^{m_g} \Big[(1-\pi_{2|1}) f_0(x_k) + \pi_{2|1} f_1(x_k)\Big] - (1-\pi_{2|1})^{m_g} \prod_{k=1}^{m_g} f_0(x_k)}{\prod_{k=1}^{m_g} \Big[(1-\pi_{2|1}) f_0(x_k) + \pi_{2|1} f_1(x_k)\Big] - (1-\pi_{2|1})^{m_g} \prod_{k=1}^{m_g} f_0(x_k)}
\end{aligned}
$$

For between-group local FDR,

$$
fdr_g = \frac{\pi_0 f(\boldsymbol{x}_g|\theta_g = 0)}{\pi_0 f(\boldsymbol{x}_g|\theta_g = 0) + \pi_1 f(\boldsymbol{x}_g|\theta_g = 1)}
$$

where

$$
f(\boldsymbol{x}_g|\theta_g = 0) = \prod_{j=1}^{m_g} f_0(x_{gj}|\theta_{gj} = 0)
$$

and

$$
f(\boldsymbol{x}_g|\theta_g = 1) = \frac{\prod_{j=1}^{m_g}((1-\pi_{2|1}) f_0(x_j) + \pi_{2|1} f_1(x_j)) - (1-\pi_{2|1})^{m_g} \prod_{j=1}^{m_g} f_0(x_j)}{1-(1-\pi_{2|1})^{m_g}}
$$

A.2 Technical Proofs

Proof of Theorem 3.4.

For $\theta_g = 1$, from the definition of $PFDR_{W|g}$ given in 3.2, we let $PFDR_{W|g}(\boldsymbol{\delta}) = \eta_g \leq \eta$ for our proposed procedure, and $PFDR_{W|g}(\boldsymbol{\delta}') \leq \eta_g$ for any other procedure. Then we get

$$\sum_j (\delta_{j|g} - \delta'_{j|g}) \left[fdr_{j|g} - \frac{\eta_g}{1-\eta_g}(1 - fdr_{j|g}) \right] \geq 0, \tag{A.1}$$

And from the definition of $\boldsymbol{\delta}$ (3.6), we can get

$$\sum_j (\delta_{j|g} - \delta'_{j|g}) \left[fdr_{j|g} - \lambda_g (1 - fdr_{j|g}) \right] \leq 0. \tag{A.2}$$

Therefore, from (A.1) and (A.2),

$$\sum_j (\delta_{j|g} - \delta'_{j|g})(1 - fdr_{j|g}) \left[\lambda_g - \frac{\eta_g}{1-\eta_g} \right] \geq 0 \tag{A.3}$$

Since for each group g,

$$\frac{\eta_g}{1-\eta_g} = \frac{\sum_j \delta_{j|g} fdr_{j|g}}{\sum_j \delta_{j|g}(1 - fdr_{j|g})} \leq \lambda_g$$

We can conclude that, $\sum_j (\delta_{j|g} - \delta'_{j|g})(1 - fdr_{j|g}) \geq 0$, which means $\sum_j (1 - \delta'_{j|g})(1 - fdr_{j|g}) \geq \sum_j (1 - \delta_{j|g})(1 - fdr_{j|g})$, or $TypeII_{W|g}(\boldsymbol{\delta}') \geq TypeII_{W|g}(\boldsymbol{\delta})$ for the groups that $\theta_g = 1$.

For $\theta_g = 0$ groups, the equality holds as $fdr_{j|g} = 1$ for these groups. Therefore, we can conclude $TypeII_{W|g}(\boldsymbol{\delta}') \geq TypeII_{W|g}(\boldsymbol{\delta})$.

Proof of Theorem 3.5.

From the definition of our proposed procedure $\boldsymbol{\delta}$, $PFDR_{T|W}(\boldsymbol{\delta}) = \frac{\sum_g \delta_g fdr^*_g R_g}{\sum_g \delta_g R_g \vee 1} = \alpha$. And for any other procedure $\boldsymbol{\delta}'$, $PFDR_{T|W}(\boldsymbol{\delta}') = \frac{\sum_g \delta'_g fdr^*_g R_g}{\sum_g \delta'_g R_g \vee 1} \leq \alpha$, we can get,

$$\sum_g (\delta_g - \delta'_g) R_g \left[fdr^*_g - \frac{\alpha}{1-\alpha}(1 - fdr^*_g) \right] \geq 0 \tag{A.4}$$

From the definition of δ_g in (3.7), we have

$$\sum_g (\delta_g - \delta_g')R_g \left[fdr_g^* - \lambda(1 - fdr_g^*)\right] \leq 0 \tag{A.5}$$

Therefore, from (A.4) and (A.5), we can conclude that

$$\sum_g (\delta_g - \delta_g')R_g(1 - fdr_g^*) \left[\lambda - \frac{\alpha}{1-\alpha}\right] \geq 0 \tag{A.6}$$

As we know, $\frac{\alpha}{1-\alpha} = \frac{\sum_g \delta_g R_g fdr_g^*}{\sum_g \delta_g R_g (1-fdr_g^*)} \leq \lambda$, we can conclude that

$$\sum_g (\delta_g - \delta_g')R_g(1 - fdr_g^*) \geq 0$$

or equivalently

$$\sum_g (1 - \delta_g')R_g(1 - fdr_g)(1 - \eta_g) \geq \sum_g (1 - \delta_g)R_g(1 - fdr_g)(1 - \eta_g)$$

which means $TypeII_{T|W}(\boldsymbol{\delta}') \geq TypeII_{T|W}(\boldsymbol{\delta})$.

A.3 EM Algorithm

A.3.1 EM Algorithm for Independent Bernoulli Model

To better present the result, define $\pi_1^1 = \pi_1, \pi_1^0 = 1 - \pi_1, \pi_{2|1}^1 = \pi_{2|1}$ and $\pi_{2|1}^0 = 1 - \pi_{2|1}$. Consider $(\boldsymbol{x}, \boldsymbol{\theta})$ as the complete data. Then the complete log-likelihood function can be written as:

$$
\begin{aligned}
& l(\boldsymbol{x}, \boldsymbol{\theta}) \\
=& \sum_{g} \sum_{k=0}^{1} I(\theta_g = k)(log\pi_1^k + logf(x_{gj}|\theta_g = k)) \\
=& \sum_{g} \left\{ I(\theta_g = 0) \left[log\pi_1^0 + \sum_{j=1}^{m_g} logf(x_{gj}|\theta_g = 0) \right] + I(\theta_g = 1) \left[log\pi_1^1 + logf(\boldsymbol{x}_g|\theta_g = 1) \right] \right\} \\
=& \sum_{g} \sum_{k=0}^{1} I(\theta_g = k) log\pi_1^k \\
& \quad + \sum_{g} \sum_{j=1}^{m_g} \sum_{l=1}^{L} \left[I(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l) log(\pi_{2|1} c_l) + I(\theta_g = 1, \theta_{j|g} = 0) log(\pi_{2|1}^0) \right] \\
& \quad + \sum_{g} \left[I(\theta_g = 0) \sum_{j=1}^{m_g} logf_0(x_{gj}) + I(\theta_g = 1) \sum_{j=1}^{m_g} I(\theta_{j|g} = 0) logf_0(x_{gj}) \right] \\
& \quad + \sum_{g} I(\theta_g = 1) \sum_{j=1}^{m_g} \sum_{l=1}^{L} I(\theta_{j|g} = 1, m_{j|g} = l) logf_l(x_{gj}|\theta_{j|g} = 1),
\end{aligned}
$$

where $m_{j|g} = l$ implies that $x_{j|g}$ is generated from $N(\mu_l, \sigma_l^2)$.

The expected value of the complete-data log-likelihood $l(\boldsymbol{x}, \boldsymbol{\theta})$ with respect to the unknown $\theta_g, \theta_{j|g}$, given the observed data $\boldsymbol{x}$ and the current value $\boldsymbol{\beta'}$ of the parameter is:

$$
\begin{aligned}
&Q(\boldsymbol{\beta},\boldsymbol{\beta}') = E\left[l(\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\beta}'\right]\\
&= \sum_{g}\sum_{k=0}^{1} log\pi_1^k P(\theta_g = k|\boldsymbol{x},\boldsymbol{\beta}')\\
&\quad + \sum_{g}\sum_{j=1}^{m_g}\sum_{k=0}^{1} \log \pi_{2|1}^k P(\theta_j = 1, \theta_{j|g} = k|\boldsymbol{x},\boldsymbol{\beta}')\\
&\quad + \sum_{g}\sum_{j=1}^{m_g}\sum_{l=1}^{L} logc_l P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x},\boldsymbol{\beta}')\\
&\quad + \sum_{g}\sum_{j=1}^{m_g} logf_0(x_{gj})P(\theta_g = 0|\boldsymbol{x},\boldsymbol{\beta}') + \sum_{g}\sum_{j=1}^{m_g} logf_0(x_{gj})P(\theta_g = 1|\boldsymbol{x},\boldsymbol{\beta}')\\
&\quad + \sum_{g}\sum_{j=1}^{m_g}\sum_{l=1}^{L} logf_l(x_{gj})P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x},\boldsymbol{\beta}').
\end{aligned}
$$

We want to maximize the Q function which can be realized by maximizing each of these parts to get the estimates of $\pi_1, \pi_{2|1}, c_l$ and μ_l, σ_l^2, since these parts are not related. To maximize the first part with the restriction that $\pi_1^0 + \pi_1^1 = 1$, using the Lagrange multipliers, we can find the maximizer for π_1^1 as

$$
\pi_1^{new} = 1 - \frac{\sum_g P(\theta_g = 0|\boldsymbol{x},\boldsymbol{\beta}')}{G} = 1 - \frac{\sum_g fdr_g(\boldsymbol{\beta}')}{G},
$$

Similarly, we can find the maximizer for $\pi_{2|1}$ and c_l as

$$
\begin{aligned}
\pi_{2|1}^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_{j|g} = 1, \theta_g = 1|\boldsymbol{x},\boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1|\boldsymbol{x},\boldsymbol{\beta}')}\\
c_l^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_{j|g} = 1, \theta_g = 1, m_{j|g} = l|\boldsymbol{x},\boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1|\boldsymbol{x},\boldsymbol{\beta}')}.
\end{aligned}
$$

For the last part of Q function, we know that $f_0(x) \sim N(0,1)$, $f_l(x) \sim N(\mu_l, \sigma_l^2)$ with probability c_l . Therefore, for each l, we need to find the MLEs for μ_l and σ_l^2 by maximizing the following log-likelihood function:

$$\begin{aligned}
&\sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} log f_l(x_{gj}) P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l | \boldsymbol{x}, \boldsymbol{\beta}') \\
&= \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} \left[-\frac{1}{2} log \sigma_l^2 - \frac{1}{2\sigma_l^2}(x_{gj} - \mu_l)^2 \right] P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l | \boldsymbol{x}, \boldsymbol{\beta}').
\end{aligned}$$

Taking derivatives with respect to μ_l and σ_1^2 and equating them to zero, we can get:

$$\begin{aligned}
\mu_l^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} x_{gj} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l | \boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l | \boldsymbol{x}, \boldsymbol{\beta}')} \\
\sigma_l^{2new} &= \frac{\sum_g \sum_{j=1}^{m_g} (x_{gj} - \mu_l)^2 P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l | \boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l | \boldsymbol{x}, \boldsymbol{\beta}')}.
\end{aligned}$$

A.3.2 EM Algorithm for Hidden Markov Model

Consider $(\boldsymbol{x}, \boldsymbol{\theta})$ as the complete data. Then the complete log-likelihood function can be written as:

$$
\begin{aligned}
& l(\boldsymbol{x},\boldsymbol{\theta}) \\
&= \sum_g \sum_{k=0}^{1} I(\theta_g = k)\left[log\pi_1^k + logf(x|\theta_g = k)\right] \\
&= \sum_g \left\{ I(\theta_g = 0)\left[log\pi_1^0 + \sum_{j=1}^{m_g} logf_0(x_{gj})\right]\right. \\
&+ \quad I(\theta_g = 1)\cdot\left[log\pi_1^1 + \sum_{h=0}^{1} I(\theta_{1|g} = h)logp_h^1 + \sum_{j=2}^{m_g}\sum_{k,h=0}^{1} loga_{kh}^j I(\theta_{j-1|g} = k, \theta_{j|q} = h)\right. \\
&\left.\left. + \sum_{j=1}^{m_g}(logf_0(x_{gj})I(\theta_{j-1|g} = 0, \theta_g = 1) + logf_1(x_{gj})I(\theta_{j-1|g} = 1, \theta_g = 1))\right]\right\} \\
&= \sum_g \sum_{k=0}^{1} I(\theta_g = k)log\pi_1^k + \sum_g \sum_{h=0}^{1} I(\theta_g = 1, \theta_{1|g} = h)logp_h^1 \\
&\quad + \sum_g \sum_{j=2}^{m_g} \sum_{k,h=0}^{1} loga_{kh}^j I(\theta_g = 1, \theta_{j-1|g} = k, \theta_{j|q} = h) \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} \left[I(\theta_g = 1, \theta_{j|g} = 0) + I(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l)logc_l\right] \\
&\quad + \sum_g \left[I(\theta_g = 0)\sum_{j=1}^{m_g} logf_0(x_{gj}) + I(\theta_g = 1)\sum_{j=1}^{m_g} I(\theta_{j|g} = 0)logf_0(x_{gj})\right] \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} I(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l)logf_l(x_{gj})
\end{aligned}
$$

where $m_{j|g} = l$ if $x_{j|g}$ is generated from $N(\mu_l, \sigma_l^2)$.

The expected value of the complete-data log-likelihood $l(x, \theta_g, \theta_{j|g})$ with respect to the unknown $\theta_g, \theta_{j|g}$ given the observed data $\boldsymbol{x}$ and the current value $\boldsymbol{\beta}'$ of the parameter is

$$\begin{aligned}
Q(\boldsymbol{\beta}, \boldsymbol{\beta}') &= E\left[l(x, \theta_g, \theta_{j|g})|\boldsymbol{x}, \boldsymbol{\beta}'\right] \\
&= \sum_g \sum_{k=0}^{1} log\pi_1^k P(\theta_g = k|\boldsymbol{x}, \boldsymbol{\beta}') + \sum_g \sum_{h=0}^{1} log p_h^1 P(\theta_g = 1, \theta_{1|g} = h|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{k,h=0}^{1} log a_{kh}^j P(\theta_g = 1, \theta_{j-1|g} = k, \theta_{j|q} = h|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{k=0}^{1} log c^k P(\theta_g = 1, \theta_{j|g} = k|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} log c_l P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} log f_l(x_{gj}) P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} log f_0(x_{gj}) P(\theta_g = 0|\boldsymbol{x}, \boldsymbol{\beta}') + \sum_g \sum_{j=1}^{m_g} log f_0(x_{gj}) P(\theta_g = 1, \theta_{j|g} = 0|\boldsymbol{x}, \boldsymbol{\beta}').
\end{aligned}$$

Similar to the independent case, we maximize the Q function to get the estimate of $\boldsymbol{\beta} = (\pi_1^k, p_h^1, a_{k,h}, c_l, \mu_l, \sigma_l^2)$ by maximizing the first six components in the above equation separately. Using the same technique as the one from the independent case, we can get the estimates of these parameters as shown in Section 4.4.2.