Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

THE CONTROL OF THE FALSE DISCOVERY RATE UNDER STRUCTURED HYPOTHESES

by
Gavin Lynch

The hypotheses in many multiple testing problems often have some inherent structure based on prior information such as Gene Ontology in gene expression data. However, few false discovery rate (FDR) controlling procedures take advantage of this inherent structure. In this dissertation, we develop FDR controlling methods which exploit the structural information of the hypotheses.

First, we study the fixed sequence structure where the testing order of the hypotheses has been pre-specified. We are motivated to study this structure since it is the most basic of structures, yet, it has been largely ignored in the literature on large scale multiple testing. We first develop procedures using the conventional fixed sequence method, where the procedures stop testing after the first hypothesis is accepted. Then, we extend the method and develop procedures which stop after a pre-specified number of acceptances. A simulation study and real data analysis show that these procedures can be a powerful alternative to the standard Benjamini-Hochberg and Benjamini-Yekutieli procedures.

Next, we consider the testing of hierarchically ordered hypotheses where hypotheses are arranged in a tree-like structure. First, we introduce a new multiple testing method called the generalized stepwise procedure and use it to create a general approach for testing hierarchically order hypotheses. Then, we develop several hierarchical testing procedures which control the FDR under various forms of dependence. Our simulation studies and real data analysis show that these proposed methods can be more powerful than alternative hierarchical testing methods, such as the method by Yekutieli (2008b).

Finally, we focus on testing hypotheses along a directed acyclic graph (DAG). First, we introduce a novel approach to develop procedures for controlling error rates appropriate for large scale multiple testing. Then, we use this approach to develop an FDR controlling procedure which tests hypotheses along the DAG. To our knowledge, no other FDR controlling procedure exists to test hypotheses with this structure. The procedure is illustrated through a real microarray data analysis where Gene Ontology terms forming a DAG are tested for significance.

In summary, this dissertation offers new FDR controlling methods which utilize the inherent structural information among the tested hypotheses.

THE CONTROL OF THE FALSE DISCOVERY RATE UNDER STRUCTURED HYPOTHESES

by
Gavin Lynch

Advisor
Wenge Guo

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
and Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences, NJIT
Department of Mathematics and Computer Science, Rutgers-Newark

May 2014

APPROVAL PAGE

THE CONTROL OF THE FALSE DISCOVERY RATE UNDER STRUCTURED HYPOTHESES

Gavin Lynch

Dr. Wenge Guo, Dissertation Advisor — Date
Assistant Professor of Statistics, New Jersey Institute of Technology

---

Dr. Sunil Dhar, Committee Member — Date
Professor of Statistics, New Jersey Institute of Technology

Dr. Ji Meng Loh, Committee Member — Date
Associate Professor of Statistics, New Jersey Institute of Technology

---

Dr. Sundarraman Subramanian, Committee Member — Date
Associate Professor of Statistics, New Jersey Institute of Technology

---

Dr. Xiaodong Lin, Committee Member — Date
Associate Professor of Statistics, Rutgers

BIOGRAPHICAL SKETCH

Author: Gavin Lynch

Degree: Doctor of Philosophy

Date: May 2014

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2014
- Master of Science in Applied Statistics,
  New Jersey Institute of Technology, Newark, NJ, 2010
- Bachelor of Science in Computer Science,
  The Pennsylvania State University, University Park, PA, 2004

Major: Mathematical Sciences

Presentations and Publications:

Lynch, G. (2013). On procedures controlling the FDR for testing hierarchically ordered hypotheses. New Jersey Institute of Technology, Frontiers in Applied and Computational Mathematics.

Lynch, G. (2013). On procedures controlling the FDR for testing hierarchically ordered hypotheses. Villanova, Non-clinical Biostatistics.

Lynch, G., Dhar, S., and Alvarez, T. (2013). The hidden component analysis algorithm: Recovering individual components of a sum. Working paper.

Lynch, G. and Guo, W. (2013). A novel FDR controlling procedure for testing hypotheses along a directed acyclic graph. Working paper.

Lynch, G. and Guo, W. (2013). On procedures controlling the FDR for testing hierarchically ordered hypotheses. Working paper.

Lynch, G., Guo, W., and Sarkar, S. (2013). The control of the false discovery rate in fixed sequence multiple testing. Working paper.

Lynch, G., Shih, F., and Liao, H. (2013). An efficient expanding block algorithm for image copy-move forgery detection. *Information Sciences*, 239:253–265.

Qui, Z., Guo, W., and Lynch, G. (2013). On generalized fixed sequence procedures for controlling the FWER. Working paper.

I would like to dedicate this dissertation to my mother and father.

ACKNOWLEDGMENT

First and foremost, I am extremely grateful for the support of my advisor, Dr. Wenge Guo. He provided indispensable guidance for the writing of this dissertation while also allowing me the room to work on my own - a style I particularly enjoyed. It was his patience, support, and enthusiasm that helped push the projects in this dissertation to a much better form. Additionally, I would like to thank the other members on my committee, Dr. Sundarraman Subramanian, Dr. Sunil Dhar, Dr. Ji Meng Loh, and Dr. Xiaodong Lin.

I would also like to thank my fellow classmates for keeping me company during tea time, listening to my complaints, organizing festivities, and being good friends.

Last, but certainly not least, I would like to thank my wife, Stephanie, for her support, understanding, and patience throughout my studies.

TABLE OF CONTENTS

Chapter **Page**

1 INTRODUCTION 1

1.1 Introduction 1

1.2 Basic Concepts 2

1.3 Testing Hypotheses with No Structure 4

1.3.1 PFER Controlling Procedures 5

1.3.2 FWER Controlling Procedures 6

1.3.3 FDR Controlling Procedures 6

1.4 Testing Hypotheses with Structure 8

1.4.1 FWER Controlling Procedures 8

1.4.2 FDR Controlling Procedures 9

1.5 Research Motivation and Dissertation Outline 10

2 THE TESTING OF FIXED SEQUENCES OF HYPOTHESES 13

2.1 Introduction 13

2.2 Preliminaries 15

2.3 Conventional Fixed Sequence Procedures 17

2.3.1 Procedure Under Arbitrary Dependence 17

2.3.2 Procedure Under Negative Dependence 19

2.4 Fixed Sequence Procedures that Allow More Acceptances 22

2.5 Data Driven Ordering 27

2.6 Simulation Study 31

2.7 Discussion 37

3 THE TESTING OF HIERARCHICALLY ORDERED HYPOTHESES . . 39

3.1 Introduction 40

3.2 Preliminaries 43

3.3 Generalized Stepwise Procedure 45

TABLE OF CONTENTS
(Continued)

Chapter **Page**

3.4 Hierarchical FDR Control 49

3.4.1 Procedure under Positive Dependence 51

3.4.2 Procedure under Arbitrary Dependence 54

3.4.3 Procedures under Block Dependence 56

3.5 Simulation Study 59

3.6 Real Data Analysis 64

3.7 Conclusion 66

4 THE TESTING OF HYPOTHESES ALONG A DIRECTED ACYCLIC GRAPH 68

4.1 Introduction 68

4.2 Preliminaries 70

4.3 A General Method for Developing New Procedures 72

4.4 Testing on a Directed Acyclic Graph 79

4.5 Simulation Study 87

4.6 Real Data Analysis 88

4.7 Conclusion 91

5 CONCLUSION AND FUTURE WORK 92

APPENDIX A PROOFS OF THE THEOREMS IN CHAPTER 2 94

A.1 Proof of Theorem 2.1 (ii) 94

A.2 Proof of Lemma 2.1 95

A.3 Proof of Lemma 2.2 95

A.4 Proof of Lemma 2.3 95

APPENDIX B PROOFS OF THE THEOREMS IN CHAPTER 3 100

B.1 Proof of Proposition 3.1 101

B.2 Proof of Theorem 3.2 102

B.3 Proof of Theorem 3.3 104

TABLE OF CONTENTS
(Continued)

Chapter **Page**

B.4 Proof of Theorem 3.4 104

B.5 Proof of Theorem 3.5 107

APPENDIX C PROOFS OF THE THEOREMS IN CHAPTER 4 110

C.1 Proof of Lemma 4.1 110

C.2 Proof of Proposition 4.3 110

BIBLIOGRAPHY 115

LIST OF TABLES

Table		Page
1.1	Summary of the Number and the Types of Rejected Hypotheses Regarding the Testing of m Hypotheses	2
2.1	The Number of Discoveries out of 7680 Genes in the HIV Data from van't Wout et al. (2003) by Procedure 2.4 and the BH Procedure	30
2.2	The Number of Discoveries out of 7680 Genes in the HIV Data from van't Wout et al. (2003) by Procedure 2.3 and the BY Procedure	30
3.1	A Comparison of Rejection Thresholds for the Procedure in Theorem 3.2 and Meinshausen's Procedure when Testing the Hypotheses in Figure 3.1(b)	54
3.2	The Procedure from Theorem 3.2 at Level $\alpha = 0.05$ to Hierarchically Test the Hypotheses Presented in Figure 3.1(b) with P-Values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$	60
3.3	Yekutieli's Procedure at Level $\alpha = 0.05$ to Hierarchically Test the Hypotheses Presented in Figure 3.1(b) with P-Values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$	60
3.4	Meinshausen's Procedure at Level $\alpha = 0.05$ to Hierarchically Test the Hypotheses Presented in Figure 3.1(b) with P-Values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$	61
3.5	The Number of Rejections out of 3261 Hypotheses by Procedures 3.2, 3.3, 3.4, 3.5, and Yekutieli's Procedure at Various Significance Levels for the Microbe Abundance Data Set Restricted to the Actinobacteria Phylum	66
4.1	The Number of Rejections Out of 10,362 Hypotheses by the SCR DAG Procedure, the DAG BH, and the BH procedure at Various Significance Levels for the Golub Leukemia Data Set	91

LIST OF FIGURES

Figure **Page**

2.1 A plot of the critical constants of the procedures in Theorems 2.1 (solid line), 2.2 (dashed line), and the FWER fixed sequence procedure (dotted line) for m =20, 50, and 100. 23

2.2 The FDR of Procedure 2.1 (solid line), Procedure 2.3 with k = 4 (dotted line) and k = 8 (dashed line), the BH procedure (dotted dash), and the BY procedure (long dash short dash) for 100 hypotheses under common correlation (top row) and two-class structure (bottom row) for $m_0 = 65$ (left) and $m_0 = 90$ (right). 32

2.3 Average power (%) of Procedure 2.1 (solid line), Procedure 2.3 with k = 4 (dotted line)and k = 8 (dashed line), the BH procedure (dotted dash), and the BY procedure (long dash short dash) for 100 hypotheses under common correlation (top row) and two-class structure (bottom row) for $m_0 = 65$ (left) and $m_0 = 90$ (right). 33

2.4 The FDR of Procedure 2.2 (solid line), Procedure 2.4 with k = 4 (dotted line) and k = 8 (dashed line), and the BH procedure (dotted dash) for 100 hypotheses under common correlation (top row) and two-class structure (bottom row) for $m_0 = 65$ (left) and $m_0 = 90$ (right). 34

2.5 Average power (%) of Procedure 2.2 (solid line), Procedure 2.4 with k = 4 (dotted line) and k = 8 (dashed line), and the BH procedure (dotted dash) for 100 hypotheses under independence for various values of m_0. 35

2.6 Average power (%) of Procedure 2.3 (solid line), Procedure 2.4 (dotted line), the BH procedure (dashed line), and the BY procedure (dotted dash) for 100 hypotheses under independence for $m_0 = 65$ (left) and $m_0 = 90$ (right). 36

3.1 (a) An example of a hierarchical structure with 3 hypotheses. H_2 and H_3 are only tested if H_1 is rejected. (b) An example of a hierarchical structure with 7 hypotheses.H_2 and H_3 are only tested if H_1 is rejected, H_4 and H_5 are only tested if H_2 is rejected, and H_6 and H_7 are only tested if H_3 is rejected. 42

3.2 FDR (top row) and average power (bottom row) of Procedures 3.2 (solid line), 3.3 (dashed), 3.4 (dotted), 3.5 (dot dash), and Yekutieli's procedure (long dash) under independence for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1. 63

LIST OF FIGURES
(Continued)

Figure **Page**

3.3 FDR (top row) and average power (bottom row) of Procedures 3.2 (solid line), 3.3 (dashed), 3.4 (dotted), 3.5 (dot dash), and Yekutieli's procedure (long dash) under common correlation with $\rho = 0.25$ for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1. 64

3.4 FDR (top row) and average power (bottom row) of Procedures 3.2 (solid line), 3.3 (dashed), 3.4 (dotted), 3.5 (dot dash), and Yekutieli's procedure (long dash) under common correlation with $\rho = 0.75$ for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1. 65

4.1 An example of a directed acyclic graph structure with 9 hypotheses. H_4 is only tested if H_1 and H_2 are rejected, H_5 is only tested if H_2 and H_3 are rejected, H_6 and H_7 are only tested if H_4 is rejected, and H_8 and H_9 are only tested if H_5 is rejected. 79

4.2 FDR (left column) and average power (right column) of the SCR DAG procedure (solid line), the DAG BH procedure (dashed), and the BH procedure (dotted) under independence (top row), common correlation with $\rho = 0.3$ (middle row), and common correlation with $\rho = 0.7$ (bottom row). 89

CHAPTER 1

INTRODUCTION

1.1 Introduction

Multiple testing as a statistical tool is very useful to analyze data from large scale scientific experiments such as microarray experiments, genome-wide association studies, fMRI experiments, and others. Unlike the testing of a single hypothesis, there are several possible measures for the overall type I error rate in multiple testing. Among these measures, the family wise error rate (FWER), which is the probability of making at least one false rejection, has traditionally been the most popular in small scale multiple testing. However, large scale experiments typically involve simultaneously testing a very large number of null hypotheses and the FWER may not be an appropriate choice of type I error control under such large scale multiple testing scenarios since it can be too strict, leaving many significant cases undiscovered. One popular type I error rate for large scale multiple testing is the false discovery rate (FDR), introduced by Benjamini and Hochberg (1995), which is much less strict than the FWER and hence, more appropriate for large scale multiple testing. The false discovery rate is defined as the expected proportion of falsely rejected null hypotheses among all rejected hypotheses.

In this dissertation, we consider large scale multiple testing problems where the hypotheses have some inherent structure. In many applications of multiple testing, such as genomic research, clinical trials, and quality control, the hypotheses have such an inherent structure. This structure may arise from prior knowledge, as in Goeman and Mansmann (2008) where Gene Ontology imposes a directed acyclic graph structure onto the tested hypotheses, or can be formed by re-formulating the

Table 1.1 Summary of the Number and the Types of Rejected Hypotheses Regarding the Testing of m Hypotheses

	Number of Hypotheses Not Rejected	**Number of Hypotheses Rejected**	**Total**
True Null Hypotheses	$m_0 - V$	V	m_0
False Null Hypotheses	$m - m_0 - R + V$	$R - V$	$m - m_0$
Total	$m - R$	R	m

underlying problem, as in Kropf and Läuter (2002), Kropf et al. (2004), Westfall et al. (2004), Hommel and Kropf (2005), and Finos and Farcomeni (2011), where a fixed sequence structure can be formed among the hypotheses by specifying the testing order of the hypotheses using a data-driven approach.

1.2 Basic Concepts

Throughout this dissertation, we will consider the problem of testing m null hypotheses, $H_1, \ldots, H_m$, based on the corresponding p-values $P_1, \ldots, P_m$, where m_0 of the hypotheses are true. For a multiple testing procedure, let V be the number of falsely rejected hypotheses and let R be the total number of rejected hypotheses. Table 1.1 summarizes this notation. It should be noted that m and m_0 are fixed but m_0 is usually unknown, and R and V are random but only R is observed.

The FDR is the expected proportion of falsely rejected hypotheses among all rejected hypotheses and is formally defined as

$$\text{FDR} = E\left(\frac{V}{R \vee 1}\right),$$

where $R \vee 1 = \max(R, 1)$. Our focus will be on the FDR, but we will also discuss several other type I error rates described as follows. The FWER is the probability of

making at least one false rejection and is

$$\text{FWER} = \Pr(V > 0).$$

It is easy to see that when $m_0 = m$, the FDR reduces to the FWER. The per-family error rate (PFER) is the expected number of falsely rejected hypotheses, which is expressed

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

Finally, the generalized FDR (k-FDR) is the expected ratio of k or more falsely rejected hypotheses to the total number of rejected hypotheses for some integer $1 \leq k \leq m$ (Sarkar 2007). Hence,

$$k\text{-FDR} = E\left(\frac{VI\{V \geq k\}}{R \vee 1}\right).$$

It is easy to see that when $k = 1$, the k-FDR reduces to the FDR. These four error rates are related by the following inequality

$$k\text{-FDR} \leq \text{FDR} \leq \text{FWER} \leq \text{PFER}.$$

Other type I errors include the generalized FWER, which is $\Pr(V \geq k)$ for $1 \leq k \leq m$ (Lehmann and Romano 2005b; Romano and Shaikh 2006; Sarkar 2007), the false discovery proportion, which is $\Pr(V/(R \vee 1) \geq \gamma)$ for $\gamma \in [0, 1)$ (Korn et al. 2004; Lehmann and Romano 2005a; Romano and Shaikh 2006; Guo et al. 2013), the positive false discovery rate, which is $E(V/R \mid R > 0)$ (Storey 2002, 2003), etc.

We say a method weakly controls a type I error rate, such as the FWER, if the type I error rate is controlled at a pre-specified level when all hypotheses are true. A method offers strong control of a type I error rate if the type I error rate is controlled for any configuration of true and false null hypotheses. Strong control is often desired

since we typically do not have knowledge of the exact configuration of true and false null hypotheses.

The following basic assumption is made throughout this dissertation regarding true null p-values

$$\Pr\left(P_i \leq p\right) \leq p, \text{ for any } p \in (0,1) \text{ for each true } H_i. \tag{1.1}$$

The dependence structure of the p-values plays an important role in determining whether a method can control an error rate. The dependence structures we will consider in this dissertation include independence, arbitrary dependence, positive dependence, negative dependence, and block dependence. Positive dependence is characterized by the following assumption

Assumption 1.1. *Positive Dependence Assumption*
For any coordinatewise non-decreasing function of the p-values ψ,

$$E\left(\psi(P_1, \ldots, P_m) \mid P_i \leq p\right) \text{ is non-decreasing in } p \text{ for each true } H_i. \tag{1.2}$$

This assumption is slightly more relaxed than positive regression dependence on a subset (PRDS) introduced in Benjamini and Yekutieli (2001). Independence simply means the p-values are mutually independent and arbitrary dependence means that the p-values are not known to have any specific dependence structure. The assumption of negative dependence and block dependence will be introduced in Chapter 2 and Chapter 3, respectively.

1.3 Testing Hypotheses with No Structure

Commonly used multiple testing methods are p-value based stepwise methods, including single-step, stepup, and stepdown methods. Given a single critical constant,

c, a single-step method rejects hypothesis H_i if $P_i \leq c$ for $i = 1, \ldots, m$. Stepup and stepdown methods test the hypotheses one by one in the order of the p-values and use a non-decreasing set of critical constants, $\alpha_1, \ldots, \alpha_m$. Let $P_{(1)} \leq \cdots \leq P_{(m)}$ be the ordered p-values and let $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypotheses. A stepup method rejects $H_{(1)}, \ldots, H_{(r)}$ and accepts $H_{(r+1)}, \ldots, H_{(m)}$ where r is the largest index satisfying

$$P_{(r)} \leq \alpha_r.$$

If no such r exists, the method accepts all the hypotheses. A stepdown method rejects $H_{(1)}, \ldots, H_{(r)}$ and accepts $H_{(r+1)}, \ldots, H_{(m)}$ where r is the largest index satisfying

$$P_{(1)} \leq \alpha_1, \ldots, P_{(r)} \leq \alpha_r.$$

Again, if no such r exists, then the method accepts all the hypotheses. When the critical constants are all the same, both stepup and stepdown methods reduce to a single-step method.

1.3.1 PFER Controlling Procedures

A well-known and common PFER controlling procedure is the Bonferroni procedure. The Bonferroni procedure is the single step procedure with critical constant α/m, and it strongly controls the PFER at level $m_0\alpha/m$ under arbitrary dependence. A variant of the Bonferroni procedure is the weighted Bonferroni procedure. Given weights $w_1, \ldots, w_m$ such that $\sum_{i=1}^{m} w_i = 1$, the weighted Bonferroni procedure rejects H_i if $P_i \leq w_i\alpha$. Another variant is the oracle Bonferroni procedure, which rejects H_i if $P_i \leq \alpha/m_0$. The oracle Bonferroni procedure strongly controls the PFER at level α under arbitrary dependence but m_0 must be known (Hochberg and Tamhane 1987; Gordon et al. 2007).

1.3.2 FWER Controlling Procedures

The Holm procedure, which is the stepdown procedure with critical constants $\alpha_i = \alpha/(m-i+1)$, strongly controls the FWER at level α under arbitrary dependence (Holm 1979). If the p-values are positively dependent as described by Assumption 1.1, then the stepup procedure with these same critical constants strongly controls the FWER (Sarkar and Chang 1997; Sarkar 1998), which is known as the Hochberg procedure (Hochberg 1988). A number of other procedures have also been developed, including the Simes global procedure (Simes 1986), the Sidák procedure (Holland and Copenhaver 1987), the Hommel procedure (Hommel 1988), etc.

1.3.3 FDR Controlling Procedures

Benjamini and Hochberg (1995) proposed the stepup procedure with critical constants $\alpha_i = i\alpha/m$ for $i = 1, \ldots, m$, which we will refer to as the Benjamini-Hochberg (BH) procedure. Formally, it is as follows.

1. Let $R = \max\{1 \leq r \leq m : P_{(r)} \leq r\alpha/m\}$.

2. If no such R exists, accept all the hypotheses. Otherwise, reject $H_{(1)}, \ldots, H_{(R)}$ and accept $H_{(R+1)}, \ldots, H_{(m)}$.

Benjamini and Hochberg (1995) proved that the BH procedure strongly controls the FDR at level $m_0\alpha/m$ under independence. Later, Benjamini and Yekutieli (2001) and Sarkar (2002) proved that the BH procedure strongly controls the FDR at level $m_0\alpha/m$ under positive dependence. It is easy to see that when m is replaced with m_0 in the critical constants of the BH procedure, the corresponding stepup procedure still controls the FDR at level α. Since m_0 is typically unknown, several procedures, known as adaptive procedures, replace m with a conservative estimate of m_0 (Storey et al. 2004; Benjamini et al. 2006; Blanchard and Roquain 2008; Sarkar 2008). Benjamini

and Yekutieli (2001) showed that under arbitrary dependence the BH procedure controls the FDR at level $m_0\alpha/(m\sum_{j=1}^{m} 1/j)$. Hence, the stepup procedure with critical constants

$$\alpha_i = \frac{i\alpha}{m\sum_{j=1}^{m} 1/j}, i = 1, \ldots, m \tag{1.3}$$

controls the FDR at level α under arbitrary dependence and is known as the Benjamini-Yekutieli (BY) procedure. Guo and Rao (2008) showed that the critical constants of this stepup procedure cannot be made larger without losing control of the FDR.

Several stepdown procedures have also been developed for FDR control. Benjamini and Liu (1999) introduced the stepdown procedure with critical constants

$$\alpha_i = 1 - \left[1 - \min\left(1, \frac{m\alpha}{m-i+1}\right)\right]^{1/(m-i+1)}, i = 1, \ldots, m,$$

and showed it strongly controls the FDR under independence. Gavrilov et al. (2009) proposed the stepdown procedure, which also control the FDR under independence, with critical constants

$$\alpha_i = \frac{i\alpha}{m+1-i(1-\alpha)}, i = 1, \ldots, m.$$

Storey (2002) introduced an estimation based approach to the FDR which takes the opposite approach of stepwise methods. Instead of determining the rejection region (i.e., critical constants) based on a fixed FDR level, the rejection region is fixed and the FDR of the rejection region is estimated. Several works have followed up on this approach, including Storey (2003), Storey et al. (2004), and Fan et al. (2012).

1.4 Testing Hypotheses with Structure

When the hypotheses have a known structure, this structural information can often be used by multiple testing procedures to increase the power to detect false null hypotheses and enhance interpretation of the rejected hypotheses. Some progress has been made in testing structured hypotheses; however, it has been primarily focused on controlling the FWER. There has been little research towards developing methods which take hypothesis structure into account in controlling the FDR. In the following, we briefly review these FWER and FDR controlling procedures for testing structured hypotheses.

1.4.1 FWER Controlling Procedures

There are several FWER controlling procedures which account for the structure of the hypotheses. Maurer et al. (1995) proposed the fixed sequence procedure for testing hypotheses in a fixed, pre-defined order, which we will refer to as a fixed sequence structure. This method, which strongly controls the FWER at level α under arbitrary dependence, tests each hypothesis sequentially and rejects H_i if $P_i \leq \alpha$ and $H_1, \ldots, H_{i-1}$ are all rejected for each $i = 1, \ldots, m$. The fallback procedure is also used to test hypotheses with a fixed sequence structure, but unlike the fixed sequence procedure, which does not test the remaining hypotheses once a hypothesis is accepted, the fallback procedure tests every hypothesis. The fallback procedure uses a series of weights $w_1, \ldots, w_m$ such that $\sum_{i=1}^{m} w_i = 1$. It rejects H_i if $P_i \leq \alpha_i$ where $\alpha_i = w_i\alpha + \alpha_{i-1}$ if H_{i-1} is rejected and $\alpha_i = w_i\alpha$ otherwise (Wiens 2003; Wiens and Dmitrienko 2005). Other procedures for testing hypotheses in a fixed sequence include Kropf and Läuter (2002), Kropf et al. (2004), Hommel and Kropf (2005), Qui et al. (2013).

Westfall and Kirshen (2001) and Dmitrienko et al. (2003) considered testing a fixed sequence of families of hypotheses and developed a general testing strategy known as gatekeeping procedures. A serial gatekeeping procedure tests a family only if every hypothesis in the previous families was rejected and a parallel gatekeeping procedure tests a family only if at least one hypothesis in each of the previous families was rejected. Recently, tree gatekeeping procedures, which merge the ideas of serial and parallel gatekeeping, have been introduced in the literature (Dmitrienko et al. 2007, 2008, 2013).

Meinshausen (2008) introduced a FWER controlling procedure for testing hypotheses with a hierarchical structure, where each hypothesis in the hierarchical structure can have at most one parent hypothesis and several child hypotheses. Meinshausen's procedure imposes the restriction that a hypothesis is not tested unless its parent hypothesis is rejected. Goeman and Mansmann (2008) developed a FWER controlling procedure for testing hypotheses along a directed acyclic graph and applied this procedure to the testing of Gene Ontology terms. The procedure is in part based on the Holm procedure and requires the user to select a pre-specified subset of the Gene Ontology terms the user is most interested in. Other procedures for testing hypotheses with complex structure include Huque and Alosh (2008) and Goeman and Finos (2012).

1.4.2 FDR Controlling Procedures

Among the few procedures that control the FDR while taking into account the underlying structure of the hypotheses, the method by Yekutieli (2008b) is perhaps the most general. Yekutieli's procedure tests hypotheses with a hierarchical structure and it was shown to strongly control the FDR at level α under independence. The procedure groups the hypotheses into families, where a family contains all the

hypotheses that share the same parent. The families are tested by the BH procedure at level $\alpha/2.88$ only if the family's parent hypothesis is rejected.

Other procedures for testing hierarchically ordered hypotheses have also been developed, such as Mehrotra and Heyse (2004), Benjamini and Heller (2007), Heller et al. (2009), and Guo et al. (2010), but these methods can only be applied to the special case when the hierarchy has exactly two levels. Recently, Farcomeni and Finos (2013) developed an FDR controlling fixed sequence procedure. Their testing procedure tests each hypothesis in order at level α until it reaches a stopping condition which is based on the number of rejections.

1.5 Research Motivation and Dissertation Outline

In this dissertation, we focus on developing FDR controlling methods which account for the inherent structure of the tested hypotheses. Hypotheses with inherent structure often arise in genomics research, clinical trials, fMRI studies, and quality control. While there is a large body of work for testing hypotheses with structure in the framework of FWER control, this problem has been largely ignored in the framework of FDR control, which is a more appropriate error rate for large scale multiple hypothesis testing. Most of the existing FDR controlling procedures are stepup or stepdown methods which order the hypotheses based on the p-values and thus, ignore any structural information that the hypotheses might have. Moreover, in some applications, it is not even possible to use the conventional p-value based multiple testing methods, because of the inherent structure among the tested hypotheses. For example, the hypotheses associated with stream data in sequential change detection problems (Ross et al. 2011), have a natural temporal structure. The decision concerning each hypothesis is to be made before the data associated with the

remaining hypotheses are observed. However, for any of the conventional methods, such as stepwise procedures, an implicit assumption is that all the data is available before any testing decision is made. So, these conventional methods are not applicable in such scenarios of real-time decision involving hypotheses with temporal structure.

In Chapter 2, we consider the testing of hypotheses with a fixed sequence structure where the hypotheses are tested one-by-one in a pre-specified order. Such fixed sequence structures naturally arise in clinical trials (Maurer et al. 1995; Bauer et al. 1998; Westfall and Kirshen 2001; Wiens and Dmitrienko 2005) and multiple testing problems associated with stream data (Ross et al. 2011). In this chapter, we develop new procedures controlling the FDR for testing hypotheses in a fixed sequence under both arbitrary dependence and independence. Although we only consider a basic structure, the techniques developed are used to help create FDR controlling procedures with more complex structures such as the hierarchical structure considered in Chapter 3. Also, we discuss a data driven ordering technique so that our fixed sequence testing procedures can even be used in more conventional multiple testing problems where the hypotheses do not have a fixed sequence structure.

Next, in Chapter 3, we consider the problem of testing hierarchies of hypotheses where a hypothesis is tested only if its parent hypothesis has been rejected. Recently, hierarchical testing approaches have been used for several multiple testing problems including genomics (Yekutieli 2008b; Heller et al. 2009; Guo et al. 2010; Goeman and Finos 2012), clinical safety data (Mehrotra and Heyse 2004), clinical trials (Dmitrienko et al. 2007; Huque and Alosh 2008), and variable selection (Meinshausen 2008). In this chapter, we introduce a novel testing method called the generalized stepwise method that is useful for hierarchical testing. We consider several different types of joint distributions of the p-values and develop hierarchical testing procedures controlling the FDR under these joint distributions.

In Chapter 4, we are primarily concerned with developing an FDR controlling procedure for testing hypotheses along a directed acyclic graph structure, which is a very general type of structure. In applications, multiple testing along a directed acyclic graph structure can often be found genomics studies (Goeman and Mansmann 2008) and in clinical trials (Dmitrienko et al. 2007; Dmitrienko and Tamhane 2013). In order to develop a procedure for testing such structured hypotheses, we propose a novel approach for creating new multiple testing procedures that control error rates appropriate for large scale multiple testing. This approach is used to help develop and prove FDR control of a new directed acyclic graph testing procedure.

Finally, our conclusion and future work is in Chapter 5, where we summarize our findings and discuss future research.

CHAPTER 2

THE TESTING OF FIXED SEQUENCES OF HYPOTHESES

2.1 Introduction

In this chapter, we consider the case when the hypotheses are to be tested in a fixed, pre-specified order. We will refer to testing hypotheses in this manner as fixed sequence multiple testing. The ordering of the hypotheses must be specified beforehand and is typically based on the inherent structure among the tested hypotheses. For example, the hypotheses associated with stream data in sequential change detection problems (Ross et al. 2011), have a natural temporal structure and are thus ordered by time. When the ordering of the hypotheses is unknown, a fixed sequence structure can be formed by re-formulating the underlying problem, as in Kropf and Läuter (2002), Kropf et al. (2004), Westfall et al. (2004), Hommel and Kropf (2005), and Finos and Farcomeni (2011), where the testing order of the hypotheses is specified using a data-driven approach.

Most FDR controlling procedures cannot test hypotheses that have a fixed pre-define testing order and to the best of our knowledge, only Farcomeni and Finos (2013) have developed such a fixed sequence FDR controlling procedure. Their testing approach, which is different than ours, tests each hypothesis in order at level α until it reaches a stopping condition which is based on the number of rejections. Yekutieli's hierarchical testing method (Yekutieli 2008b), which controls the FDR under independence, can also be applied to test hypotheses in the fixed sequence structure. However, when testing hypotheses in a fixed sequence, this method tests each hypothesis at $\alpha/2.88$, which is too conservative.

In this chapter, we will introduce several FDR controlling methods that exploit the fixed sequence structural information. We first consider a conventional fixed sequence multiple testing method that keeps rejecting until an acceptance occurs and develop such a method controlling the FDR under arbitrary dependence. It is shown to be optimal in the sense that it cannot be improved by increasing even one of its critical values without losing control over the FDR, or even by imposing a positive dependence condition on the p-values, such as Assumption 1.1. This is different from what happens in the case of non-fixed sequence multiple testing. For instance, the so-called BY method of Benjamini and Yekutieli (2001) that controls the FDR under arbitrary dependence can be improved significantly by the BH method of Benjamini and Hochberg (1995) by imposing this positive dependence condition. Since our procedure cannot be improved under positive dependence, we consider the case of negative dependence and develop a more powerful conventional fixed sequence multiple testing method controlling the FDR under negative dependence which includes independence as a special case.

There is a potential for loss of power in a conventional fixed sequence multiple testing method if the ordering of the hypotheses, particularly for the earlier ones, does not match with that of their true effect sizes potentially leading to some earlier hypothesis being accepted and the follow-up hypotheses having no chance to be tested. To mitigate that, we consider generalizing the conventional fixed sequence multiple testing to one that allows more than one, but a pre-specified number of acceptances, and develop such generalized fixed sequence multiple testing methods controlling the FDR under both arbitrary dependence and independence.

It is not always the case in real data applications that the hypotheses will have a natural fixed sequence structure or information about how to order them will be available a priori. Nevertheless, the data itself can often provide information on how to order the hypotheses. In this chapter, we discuss such a data-driven ordering

strategy which can be applied to a broad spectrum of multiple testing problems, such as one-sample and two-sample t-tests, and one-sample and two-sample nonparametric tests. Through simulation studies and a real microarray data analysis, this strategy coupled with our proposed fixed sequence methods is seen to perform favorably against the corresponding non-fixed sequence methods under certain settings.

This chapter is organized as follows. With some concepts and background information given in Section 2.2, we present the developments of our conventional and generalized fixed sequence procedures controlling the FDR under various dependencies in Sections 2.3 and 2.4, respectively. A data-driven ordering strategy is discussed and coupled with our fixed sequence procedures, applied to a real microarray data analysis in Section 2.5. The findings from some simulation studies on the performances of our procedures are given in Section 2.6. Some concluding remarks are made in Section 2.7 and proofs of some results are given in Appendix A.

2.2 Preliminaries

Suppose that $H_i, i = 1, \ldots, m$, are the m null hypotheses to be tested and are ordered a priori. Let m_1 of these null hypotheses be false and recall that m_0 is the number of true null hypotheses. For notational convenience, we denote the index of the ith true null hypothesis by u_i and the ith true null p-value by $\widehat{P}_i$, for $i = 1, \ldots, m_0$ so that $\widehat{P}_i = P_{u_i}$. The set of indices of the true null hypotheses is I_0. Let S be the number of false null hypotheses rejected and recall that V and R are the number of true null hypotheses rejected and total number of hypotheses rejected, respectively, for a multiple testing procedure.

Typically, the hypotheses are ordered based on their p-values and multiple testing is carried out using a stepwise or single-step procedure. However, when these

hypotheses are ordered a prior and not according to their p-values, multiple testing is performed using a fixed sequence method. Given a non-decreasing sequence of critical constants $0 < \alpha_1 \leq \ldots \leq \alpha_m$, a conventional fixed sequence method is defined as follows:

Definition 2.1. (Conventional fixed sequence method)

1. *If $P_1 \leq \alpha_1$, then reject H_1 and continue to test H_2; otherwise, stop.*

2. *If $P_i \leq \alpha_i$ then reject H_i and continue to test H_{i+1}; otherwise, stop.*

Thus, a conventional fixed sequence method continues testing in the pre-determined order as long as rejections occur. Once an acceptance occurs, it stops testing the remaining hypotheses. In Section 2.4, we will generalize a conventional fixed sequence method to allow a given number of acceptances. It should be noted that a conventional fixed sequence method with common critical constant α, which is often called the fixed sequence procedure in the literature, strongly controls the FWER at level α (Maurer et al. 1995). We will refer to it as the FWER fixed sequence procedure in this dissertation in order to distinguish it from other fixed sequence methods designed to control the FDR.

Regarding the dependence of the p-values, in this chapter, we will refer to arbitrary dependence, positive dependence, independence, and negative dependence. Positive dependence is characterized by Assumption 1.1 of Chapter 1. Negative dependence is characterized by the following property:

Definition 2.2. *(Negative Association) The vector of p-values $\vec{P}$ is negatively associated with null p-values if for each $i = 1, \ldots, m_0$, the following inequality holds:*

$$\begin{aligned} & \Pr\left(\widehat{P}_i \leq p_{u_i}, P_j \leq p_j, j = 1, \ldots, m, \text{ with } j \neq u_i\right) \\ \leq \; & \Pr\left(\widehat{P}_i \leq p_{u_i}\right) \Pr\left(P_j \leq p_j, j = 1, \ldots, m, \text{ with } j \neq u_i\right), \end{aligned} \tag{2.1}$$

for all fixed p_j's.

Several multivariate distributions posses the conventional negative association property, including multivariate normal with non-positive correlation, multinomial, dirichlet, and multivariate hypergeometric (Joag Dev and Proschan 1983). It is easily seen that independence is a special case of negative dependence.

2.3 Conventional Fixed Sequence Procedures

In this section, we present the developments of two simple conventional fixed sequence procedures controlling the FDR under both arbitrary dependence and negative dependence conditions on the p-values.

2.3.1 Procedure Under Arbitrary Dependence

Since the FDR is more liberal than the FWER, a conventional fixed sequence method controlling the FDR under arbitrary dependence is expected to have critical values that are at least as large as α, the common critical constant for the FWER fixed sequence method. In the following, we present such a simple conventional fixed sequence FDR controlling procedure.

Theorem 2.1. *Consider a conventional fixed sequence procedure with critical constants*

$$\alpha_i^{(1)} = \min\left(\frac{m\alpha}{m-i+1}, 1\right),\ i = 1, \ldots, m.$$

(i) This procedure strongly controls the FDR at level α under arbitrary dependence of the p-values.

(ii) One cannot increase even one of the critical constants $\alpha_i^{(1)}, i = 1, \ldots, m$, *while keeping the remaining fixed without losing control of the FDR. This is true even under Assumption 1.1.*

Proof of 2.1(i). Since u_1 is the index of the first true null hypothesis, the first $u_1 - 1$ null hypotheses are all false. Note that the event $\{V > 0\}$ implies that $S \geq u_1 - 1$ and $\hat{P}_1 \leq \alpha_{u_1}^{(1)}$, and therefore we have

$$\begin{aligned} \text{FDR} &= E\left(\frac{V}{V+S} I\{V>0\}\right) \leq E\left(\frac{m_0}{m_0+u_1-1} I\{V>0\}\right) \\ &= \frac{m_0}{m_0+u_1-1} Pr(V>0) \leq \frac{m_0}{m_0+u_1-1} Pr(\hat{P}_1 \leq \alpha_{u_1}^{(1)}) \\ &\leq \frac{m-u_1+1}{m} \alpha_{u_1}^{(1)} \leq \alpha. \end{aligned}$$

The first inequality follows from the fact that $V/(V+S)$ is an increasing function of V and a decreasing function of S. The third inequality follows from the fact that $m_0/(m_0+u_1-1)$ is an increasing function of m_0 and $m_0 \leq m - u_1 + 1$ since there are at least $u_1 - 1$ false nulls. This proves part (i).

For a proof of part (ii), see Appendix A. ■

Remark 2.1. Theorem 2.1 shows that when controlling the FDR under arbitrary dependence, the operating characteristic of the proposed fixed sequence method is much different from that of the usual stepwise procedure of Benjamini and Yekutieli (Benjamini and Yekutieli 2001) that relies on p-value based ordering of the hypotheses. It cannot be further improved, even by imposing Assumption 1.1 on the p-values, unlike the BY procedure that is known to be greatly improved by the BH procedure under such positive dependence. Also, as shown in our proof of Theorem 2.1(ii) under arbitrary dependence (see Appendix), the least favorable configuration (the configuration which leads to the largest error rate, see Finner and Roters, 2001) of the newly suggested fixed sequence FDR controlling procedure is when the ordering of the hypotheses is perfect (i.e, when all the false null hypotheses are tested before

the true ones), the false null p-values are all 0 with probability 1, and the true null p-values are the same with each following $U(0,1)$ distribution. This least favorable configuration is much different from that given in Guo and Rao (2008) for the BY procedure under arbitrary dependence.

Although the procedure in Theorem 2.1 cannot be improved under Assumption 1.1, we consider in the next subsection the condition of negative dependence which includes independence as a special case, and under such condition, develop a more powerful conventional fixed sequence method that controls the FDR.

2.3.2 Procedure Under Negative Dependence

When the p-values are negatively associated as defined in Section 2.2, the critical constants of the conventional fixed sequence procedure in Theorem 2.1 can be further improved as in the following:

Theorem 2.2. *The conventional fixed sequence method with critical constants*

$$\alpha_i^{(2)} = \frac{i\alpha}{1+(i-1)\alpha}, i = 1, \ldots, m$$

strongly controls the FDR at level α when the p-values are negatively associated on the true null p-values.

To prove Theorem 2.2, we use the following lemma, with proof given in Appendix:

Lemma 2.1. *Let $m_{0,i}$ and $m_{1,i}$ respectively denote the numbers of true and false null hypotheses among the first i null hypotheses, and*

$$d_i = \begin{cases} I\{i \in I_0\} & \text{if } i = 1 \\ (m_{1,i-1} I\{i \in I_0\} - m_{0,i-1} I\{i \notin I_0\})/(i(i-1)) & \text{if } i > 1. \end{cases}$$

Then, the FDR of any fixed sequence procedure can be expressed as

$$\text{FDR} = \sum_{i=1}^{m} d_i \Pr\left(R \geq i\right).$$

Proof of Theorem 2.2. If $u_1 = 1$, then

$$\text{FDR} \leq \text{FWER} = \Pr\left(V > 0\right) = \Pr\left(\widehat{P}_1 \leq \alpha_1^{(2)}\right) \leq \alpha.$$

If $u_1 > 1$, then by Lemma 2.1,

$$\begin{aligned}\text{FDR} &= \sum_{i=1}^{m} d_i \Pr\left(R \geq i\right) = \sum_{i=u_1}^{m} d_i \Pr\left(R \geq i\right) \\ &= \sum_{i=u_1}^{m} \left(d_i + \frac{m_{1,i}\alpha}{i}\right) \Pr\left(R \geq i\right) - \sum_{i=u_1}^{m} \frac{m_{1,i}\alpha}{i} \Pr\left(R \geq i\right). \qquad (2.2)\end{aligned}$$

The second equality follows from the fact that $d_i = 0$ for $i = 1, \ldots, u_1 - 1$.

For each $i = u_1, \ldots, m$, the following inequality holds.

$$\left(d_i + \frac{m_{1,i}\alpha}{i}\right) \Pr\left(R \geq i\right) \leq \frac{m_{1,i-1}\alpha}{i-1} Pr(R \geq i-1). \qquad (2.3)$$

To see this, we consider, separately, the case when $i \in I_0$ and when $i \notin I_0$. Suppose $i \in I_0$, then $m_{1,i-1} = m_{1,i}$ and

$$\begin{aligned}&\left(d_i + \frac{m_{1,i}\alpha}{i}\right) \Pr\left(R \geq i\right) \\ = \;& \left(\frac{m_{1,i-1}}{i(i-1)} + \frac{m_{1,i-1}\alpha}{i}\right) \Pr\left(P_1 \leq \alpha_1^{(2)}, \ldots, P_{i-1} \leq \alpha_{i-1}^{(2)}, P_i \leq \alpha_i^{(2)}\right) \\ \leq \;& \frac{m_{1,i-1}\left(1 + (i-1)\alpha\right)}{i(i-1)} \Pr\left(P_1 \leq \alpha_1^{(2)}, \ldots, P_{i-1} \leq \alpha_{i-1}^{(2)}\right) \Pr\left(P_i \leq \alpha_i^{(2)}\right) \\ \leq \;& \frac{m_{1,i-1}\alpha}{i-1} Pr(R \geq i-1).\end{aligned}$$

The first and second inequalities follow from (2.1) and (1.1), respectively.

Now suppose $i \notin I_0$, then $m_{1,i} = m_{1,i-1} + 1$ and

$$\left(d_i + \frac{m_{1,i}\alpha}{i}\right) \Pr\left(R \geq i\right) = \left(-\frac{m_{0,i-1}}{i(i-1)} + \frac{(m_{1,i-1}+1)\alpha}{i}\right) \Pr\left(R \geq i\right)$$

$$\begin{aligned} &\leq \left(-\frac{m_{0,i-1}\alpha}{i(i-1)} + \frac{(m_{1,i-1}+1)\alpha}{i}\right)\Pr\left(R \geq i\right) \\ &= \frac{m_{1,i-1}\alpha}{i-1}\Pr\left(R \geq i\right) \leq \frac{m_{1,i-1}\alpha}{i-1}\Pr\left(R \geq i-1\right). \end{aligned}$$

In the second equality, we use the fact that $m_{0,i-1} + m_{1,i-1} = i - 1$.

Applying (2.3) to (2.2), we have

$$\begin{aligned} \text{FDR} &= \sum_{i=u_1}^{m}\left(d_i + \frac{m_{1,i}\alpha}{i}\right)\Pr\left(R \geq i\right) - \sum_{i=u_1}^{m}\frac{m_{1,i}\alpha}{i}\Pr\left(R \geq i\right) \\ &\leq \sum_{i=u_1}^{m}\frac{m_{1,i-1}\alpha}{i-1}\Pr\left(R \geq i-1\right) - \sum_{i=u_1}^{m}\frac{m_{1,i}\alpha}{i}\Pr\left(R \geq i\right) \\ &= \alpha\Pr\left(R \geq u_1 - 1\right) - \frac{m_1\alpha}{m}\Pr\left(R = m\right) \\ &\leq \alpha. \end{aligned}$$

The equality follows from that fact that $m_{1,u_1-1} = u_1 - 1$, since the first $u_1 - 1$ hypotheses are false. ■

Remark 2.2. The conventional fixed sequence procedure in Theorem 2.2 is nearly optimal in the sense that the upper bound of the FDR of this procedure is very close to α under certain configurations. Consider the following configuration: All the false null hypotheses are tested before the true null hypotheses, the false null p-values are all 0 with probability 1, and the true null p-values are independently distributed as $U(0,1)$. Under this configuration, (2.3) becomes an equality. Following the proof of Theorem 2.2, the FDR of this procedure is exactly $\alpha - \frac{m_1\alpha}{m}\Pr\left(R = m\right)$. When m_1/m approaches to π_1 as $m \to \infty$ with $0 \leq \pi_1 < 1$, an approximate lower bound of the FDR is $\alpha - \pi_1\alpha e^{-(1-\pi_1)\frac{1-\alpha}{\alpha}}$.

To see why, we first note that

$$\frac{m_1\alpha}{m}\Pr\left(R = m\right) = \frac{m_1\alpha}{m}\prod_{i=m_1+1}^{m}\alpha_i^{(2)} < \frac{m_1\alpha}{m}(\alpha_m^{(2)})^{m-m_1}.$$

Then, by simple calculation, we have

$$\begin{aligned}\lim_{m\to\infty} \text{FDR} &\geq \alpha - \lim_{m\to\infty} \frac{m_1\alpha}{m}(\alpha_m^{(2)})^{m-m_1} \\ &= \alpha - \lim_{m\to\infty} \frac{m_1\alpha}{m}\left(1+\frac{1-\alpha}{\alpha}\frac{1}{m}\right)^{-m(1-\frac{m_1}{m})} \\ &= \alpha - \pi_1\alpha e^{-(1-\pi_1)\frac{1-\alpha}{\alpha}}.\end{aligned}$$

This lower bound of the FDR is very close to the pre-specified level α. For example, for large m, if $m_1/m = 0.2$, then with $\alpha = 0.05$, the lower bound under the above configuration is about 0.04999975.

Remark 2.3. Figure 2.1 compares the critical constants for the proposed procedures along with the FWER fixed sequence procedure at level α. It should be noted that the first few critical constants are the most important ones. If the first few values are too small, then the procedure might stop too early and the remaining hypotheses will not have a chance to be tested. With this in mind, it can be seen that the critical constants of the procedure introduced in Theorem 2.2 are by far the best, and the critical constants of the procedure in Theorem 2.1 are slightly better than those of the conventional fixed sequence procedure.

2.4 Fixed Sequence Procedures that Allow More Acceptances

A conventional fixed sequence method might potentially lose power if an early null hypothesis fails to be rejected, with the remaining hypotheses having no chance of being tested. To remedy this, we generalize a conventional fixed sequence method to one that allows a certain number of acceptances. The procedure will keep testing hypotheses until a pre-specified number of acceptance has been reached. The same

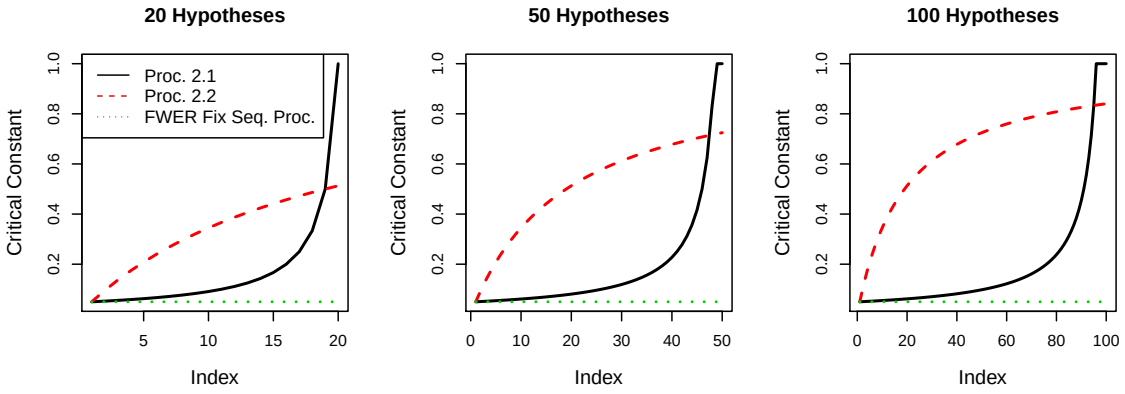


Figure 2.1 A plot of the critical constants of the procedures in Theorems 2.1 (solid line), 2.2 (dashed line), and the FWER fixed sequence procedure (dotted line) for m =20, 50, and 100.

idea has also been used by Hommel and Kropf (2005) to develop FWER controlling procedures in fixed-sequence multiple testing.

Suppose k is a pre-specified positive integer and $\alpha_1 \leq \cdots \leq \alpha_m$ is a non-decreasing sequence of critical constants. A fixed sequence method that allows more acceptances is defined below.

Definition 2.3. (Fixed sequence method stopping on the k^{th} acceptance)

1. *If $P_1 \leq \alpha_1$, then reject H_1; otherwise, accept H_1. If $k > 1$ or H_1 is rejected, then continue to test H_2; otherwise stop.*

2. *If $P_i \leq \alpha_i$, then reject H_i; otherwise, accept H_i. If the number of accepted hypotheses so far is less than k, then continue to test H_{i+1}; otherwise stop.*

It is easy to see that when $k = 1$, the fixed sequence method stopping on the k^{th} acceptance reduces to the conventional one.

Theorem 2.3. *The fixed sequence method stopping on the k^{th} acceptance with critical constants*

$$\alpha_i^{(3)} = \begin{cases} \frac{\alpha}{k} & \text{if} \quad i = 1, \ldots, k \\ \frac{(m-k+1)\alpha}{(m-i+1)k} & \text{if} \quad i = k+1, \ldots, m \end{cases}$$

strongly controls the FDR at level α under arbitrary dependence of the p-values.

Proof of Theorem 2.3. Let U be the index of the first rejected true null hypothesis. If no true null hypotheses are rejected, then set $U = 0$. We will show that for $i = 1, \ldots, m_0$,

$$E\left(\frac{V}{V+S} I\{U = u_i\}\right) \leq \frac{\alpha}{k}. \tag{2.4}$$

If $i \leq k$, then

$$E\left(\frac{V}{V+S} I\{U = u_i\}\right) \leq \Pr\left(U = u_i\right) \leq \Pr\left(\hat{P}_i \leq \frac{\alpha}{k}\right) \leq \frac{\alpha}{k}.$$

Now, assume $i > k$. Note that the event $\{U = u_i\}$ implies $V \leq m - u_i + 1$ and $S \geq u_i - k$, because the first $u_i - 1$ hypotheses were either false nulls or accepted true nulls and among the first $u_i - 1$ hypotheses tested, there can be at most $k - 1$ acceptances. Thus,

$$\begin{aligned} & E\left(\frac{V}{V+S} I\{U = u_i\}\right) \leq E\left(\frac{m - u_i + 1}{(m - u_i + 1) + (u_i - k)} I\{U = u_i\}\right) \\ \leq \quad & \frac{m - u_i + 1}{m - k + 1} \Pr\left(\hat{P}_i \leq \alpha_{u_i}^{(3)}\right) \leq \frac{\alpha}{k}. \end{aligned}$$

The first inequality follows from the fact that $V/(V+S)$ is an increasing function of V and a decreasing function of S.

From (2.4), we have

$$\text{FDR} = \sum_{i=1}^{\min(m_0,k)} E\left(\frac{V}{S+V} I\{U = u_i\}\right) \leq \sum_{i=1}^{k} \frac{\alpha}{k} = \alpha,$$

where the first equality follows from the fact that if none of the first k true null hypotheses are rejected, then $V = 0$. ■

We should point out that the result in Theorem 2.3 is weaker than that in Theorem 2.1, although the method in Theorem 2.3 reduces to that in Theorem 2.1 when $k = 1$. More specifically, we cannot prove that the procedure in Theorem 2.3 is optimal in the sense that its critical constants cannot be further improved without losing control of the FDR under arbitrary dependence of the p-values. However, under certain distributional assumptions on the p-values, the critical constants of this procedure can potentially be improved. In particular, we have the following result.

Theorem 2.4. *Consider a fixed sequence method stopping on the k^{th} acceptance with critical constants*

$$\alpha_i^{(4)} = \frac{(r_{i-1} + 1)\alpha}{k + (i - k)\alpha}, i = 1, \ldots, m,$$

where r_i is the number of rejections among the first i tested hypotheses, with $r_0 = 0$, for $i = 1, \ldots, m$. This procedure strongly controls the FDR at level α if the true null p-values are mutually independent and are independent of the false null p-values.

Before presenting a proof of the above theorem, let us introduce a few more notations. For $i = 1, \ldots, m$, let V_i and S_i be the numbers of false rejections and true rejections among the first i rejections and J_i be the index of the i^{th} rejected hypothesis. If there are less than i rejections, we define $V_i = V_{i-1}$, $S_i = S_{i-1}$ and $J_i = m + 1$. In addition, for notational convenience, define $V_0 = S_0 = J_0 = 0, V_0/0 = 0$, and $S_0/0 = 1$.

We use the following two lemmas, with proofs given in Appendix, to prove the theorem.

Lemma 2.2. *The FDR of any fixed sequence method stopping on the k^{th} acceptance can be expressed as*

$$\text{FDR} = E\left(\sum_{i=1}^{m}\left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1}\right) I\{J_i < m+1\}\right).$$

Lemma 2.3. *Consider the procedure defined in Theorem 2.4, the following inequality holds for $i = 1, \ldots, m$,*

$$\begin{aligned} & E\left(\left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1}\right) I\{J_i < m+1\}\right) \\ \leq\ & E\left(\frac{(k - J_{i-1} + i - 1)\alpha}{k}\frac{S_{i-1}}{i-1} I\{J_{i-1} < m+1\} - \frac{(k - J_i + i)\alpha}{k}\frac{S_i}{i} I\{J_i < m+1\}\right). \end{aligned}$$

Proof of Theorem 2.4. By Lemmas 2.2 and 2.3, we have

$$\begin{aligned} \text{FDR} &= E\left(\sum_{i=1}^{m}\left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1}\right) I\{J_i < m+1\}\right) \\ &\leq E\left(\alpha - \frac{(k - J_1 + 1)\alpha}{k} S_1 I\{J_i < m+1\}\right) \\ &\quad + E\left(\sum_{i=2}^{m}\left(\frac{(k - J_{i-1} + i - 1)\alpha}{k}\frac{S_{i-1}}{i-1} I\{J_{i-1} < m+1\}\right.\right. \\ &\quad \left.\left. - \frac{(k - J_i + i)\alpha}{k}\frac{S_i}{i} I\{J_i < m+1\}\right)\right) \\ &= E\left(\alpha - \frac{(k - J_m + m)\alpha}{k}\frac{S_m}{m} I\{J_m < m+1\}\right) \leq \alpha. \ \blacksquare \end{aligned}$$

Remark 2.4. When $k = 1$, the generalized fixed sequence method in the above theorem reduces to the conventional fixed sequence method in Theorem 2.2, since in this case $r_{i-1} = i - 1$, and thus continues to control the FDR when the p-values are negatively associated. However, when $k > 1$, it can only control the FDR under the independence assumption made in the theorem. It can be shown that this method, when $k > 1$, may no longer control the FDR when the p-values are negatively associated. Consider, for example, the problem of simultaneously testing two hypotheses H_1 and H_2 for which both of them are true, the associated p-values $\widehat{P}_1$ and $\widehat{P}_2$ are both $U(0, 1)$, and

$\widehat{P}_2 = 1 - \widehat{P}_1$. It is easy to see that under such configuration, when $k = 2$, the FDR of this procedure is equal to $\frac{\alpha}{2-\alpha} + \frac{\alpha}{2} > \alpha$.

2.5 Data Driven Ordering

The performances of the aforementioned fixed sequence methods depend on how well the hypotheses are ordered according to their true effect sizes. In some cases, one can use pilot data available to establish a good ordering among the hypotheses. For example, in replicated studies, the hypotheses for the follow-up study can be ordered using the data from the primary study. However, when prior information is unavailable ordering information can usually be assessed from the data itself. Such data-driven ordering methods have been used by several authors coupled with the fixed sequence methods controlling the FWER and generalized FWER (Kropf and Läuter 2002; Kropf et al. 2004; Westfall et al. 2004; Hommel and Kropf 2005; Finos and Farcomeni 2011; Goeman and Finos 2012). Assume that the variables of interest are X_i, $i = 1, \ldots, m$, with n independent observations $X_{i1}, \ldots, X_{in}$ on each X_i. An ordering statistic, Y_i, and a test statistic, T_i, are determined for each $i = 1, \ldots, m$. The Y_i's are used to order all of the hypotheses $H_1, \ldots, H_m$, T_i is used to test the hypothesis $H_i, i = 1, \ldots, m$, and P_i is the corresponding p-value. In addition, Y_i is chosen such that it is independent of the T_i's under H_i and tends to be larger as the effect size increases. The idea is outlined below.

Definition 2.4. *Data-Driven Ordering Procedure*

1. *The hypotheses are ordered based on $Y_1, \ldots, Y_m$ where the hypothesis corresponding to the largest Y_i is placed first, the hypothesis corresponding to the second largest is placed second, and so on.*

2. *The hypotheses are tested using a fixed sequence procedure based on the the p-values $P_1, \ldots, P_m$ and the testing order established in Step 1.*

We give a few examples to further illustrate the idea.

Example 2.1: One sample T-test. Consider testing $H_i : \mu_i = 0$ against $H_i' : \mu_i \neq 0$ simultaneously where X_i follows a $N(\mu_i, \sigma^2)$ distribution. Let $\bar{X}_i = \sum_{j=1}^{n} X_{ij}/n$ and $s_i^2 = \sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2/(n-1)$ be the sample mean and variance, respectively, based on the observations $X_{i1}, \ldots, X_{in}$. Let $Y_i = \sum_{j=1}^{n} X_{ij}^2$ be the ordering statistics so that the hypotheses are ordered through their sum of squares, and $T_i = \sqrt{n}\bar{X}_i/s_i$ is the usual t-test statistic for testing H_i. Then, $P_i = 2\,(1 - F(|T_i|))$, $i = 1, \ldots, m$, where $F(\cdot)$ is the CDF of the t-distribution with $n-1$ degrees of freedom, are the p-values. When $\mu_i = 0$, T_i and Y_i are independent (see, for instance, Lehmann and Romano (2005a); p. 156). Furthermore,

$$E(Y_i) = n(\mu_i^2 + \sigma^2),$$

which suggests that $|\mu_i|$ tends to increase as Y_i increases.

Example 2.2: Two sample T-test. Consider testing $H_i : \mu_i^{(1)} = \mu_i^{(2)}$ against $H_i' : \mu_i^{(1)} \neq \mu_i^{(2)}$ simultaneously using $n = n_1 + n_2$ data vectors. Suppose $X_{ij}^{(l)}$, $j = 1, \ldots, n_l$, follows a $N(\mu_i^{(l)}, \sigma^2)$ distribution, for $l = 1, 2$. Then, the hypotheses can be tested using the two-sample t-test statistics T_i and are ordered through the values of the 'total sum of squares,' which is $Y_i = \sum_{l=1}^{2}\sum_{j=1}^{n_l}(X_{ij}^{(l)} - \bar{X}_i)^2$, where $\bar{X}_i = \sum_{l=1}^{2}\sum_{j=1}^{n_l} X_{ij}^{(l)}/n$, for $i = 1, \ldots, m$. The rationale behind this is the independence between the Y_i's and T_i under H_i (see, for instance, Westfall et al. (2004)), and the following result: $E[Y_i] = (n-1)\sigma^2 + n_1 n_2(\mu_i^{(1)} - \mu_i^{(2)})^2/n$.

Example 2.3: Nonparametric test. Kropf et al. (2004) describe a data-driven ordering strategy for nonparametric tests. In the one sample case, we are interested in testing $H_i : \mu_i = 0$ against $H_i' : \mu_i \neq 0$ where $X_{ij}, j = 1, \ldots, n$ are assumed to

be symmetric about μ_i. The hypotheses are tested using the one-sample Wilcoxon test and ordered based on $Y_i = \text{med}(|X_{i1}|, \ldots, |X_{in}|)$. In the two sample case, we are interested in testing $H_i : \mu_i^{(1)} = \mu_i^{(2)}$ against $H_i' : \mu_i^{(1)} \neq \mu_i^{(2)}$ using $n = n_1 + n_2$ data vectors, where $X_{ij}^{(l)}, j = 1, \ldots, n_l$ are assumed to be symmetric about $\mu_i^{(l)}$ for $l = 1, 2$. The hypotheses are tested using the two-sample Wilcoxon test and ordered based on the interquartile range $Y_i = q_{3i} - q_{1i}$, where q_{1i} and q_{3i} are respectively the 1^{st} and 3^{rd} quartile of the mixture of $X_{ij}^{(1)}$'s and $X_{ij}^{(2)}$'s.

When our proposed fixed sequence procedures are used in applications coupled with the aforementioned data-driven ordering strategy, the FDR controls are still maintained under the independence assumption, if the ordering statistics are chosen to be independent of the test statistics in the data-driven ordering strategy, even though the same data is repeatedly used for ordering and testing the hypotheses. We have the following result.

Theorem 2.5. *Suppose $X_1, \ldots, X_m$ are mutually independent. If the hypotheses $H_1, \ldots, H_m$ are ordered based on the ordering statistics $Y_i, i = 1, \ldots, m$, tested using the test statistics $T_i, i = 1, \ldots, m$, and Y_i is independent of T_i under H_i, then the fixed sequence multiple testing procedures introduced in Theorems 2.1-2.4 can still strongly control the FDR at level α.*

Proof of Theorem 2.5. Assume without any loss of generality that $Y_1 \geq \cdots \geq Y_m$, so that conditional on the Y_i's, H_i is the ith hypotheses to be tested in our fixed sequence multiple testing methods. When H_i is true, P_i is independent of both Y_i and $X_j, j = 1, \ldots, m$ with $j \neq i$. This follows from independence of the X_i's and that of Y_i and T_i under H_i. Thus, conditional on the Y_i's, each true null p-value P_i still satisfies 1.1 and is independent of all other p-values P_j with $j \neq i$. Therefore, we have for each of the procedures in Theorems 2.1, 2.2, 2.3, and 2.4,

$$E\left(\frac{V}{\max(R,1)} \,\middle|\, Y_1, \ldots, Y_m\right) \leq \alpha. \tag{2.5}$$

This proves the desired result. ■

Table 2.1 The Number of Discoveries out of 7680 Genes in the HIV Data from van't Wout et al. (2003) by Procedure 2.4 and the BH Procedure

	Procedure 2.4				
	k = 1	k = 384	k = 768	k = 1152	BH Procedure
$\alpha = 0.001$	11	13	9	8	8
$\alpha = 0.01$	11	18	17	16	13
$\alpha = 0.025$	11	18	18	18	13
$\alpha = 0.05$	11	20	19	19	18
$\alpha = 0.1$	20	21	24	20	22

Table 2.2 The Number of Discoveries out of 7680 Genes in the HIV Data from van't Wout et al. (2003) by Procedure 2.3 and the BY Procedure

	Procedure 2.3				
	k = 1	k = 384	k = 768	k = 1152	BY Procedure
$\alpha = 0.001$	11	10	8	8	0
$\alpha = 0.01$	11	13	13	11	8
$\alpha = 0.025$	11	15	13	13	10
$\alpha = 0.05$	11	16	15	13	10
$\alpha = 0.1$	11	18	16	16	13

We applied our proposed methods to the HIV microarray data (van't Wout et al. 2003) used by Efron (2008). These data consist of 7680 gene expression levels across eight subjects, four HIV infected and four uninfected. The data were log-transformed and normalized. Our goal is to determine which genes are differentially expressed by testing $H_i : \mu_i^{(1)} = \mu_i^{(2)}$ versus $H_i' : \mu_i^{(1)} \neq \mu_i^{(2)}$ simultaneously for $i = 1, \ldots, 7680$,

where $\mu_i^{(1)}$ and $\mu_i^{(2)}$ are the gene specific mean expressions for HIV infected and uninfected subjects, respectively.

We applied our proposed procedures with the p-values generated from two sample t-tests for the genes. Since there is no natural ordering among the genes, we used the ordering statistics for two sample t-tests in Example 2.2 to order these tested hypotheses. We compared the procedure in Theorem 2.4 (labeled Procedure 2.4) with the BH procedure. The results are summarized in Table 2.1 for different values of k where $k = 384$, $k = 768$, and $k = 1152$, representing 5%, 10%, and 15% of the total number of tested hypotheses. As seen from Table 2.1, for all values of k except $k = 1$, the procedure in Theorem 2.4 usually has more rejections than the BH procedure. When α is small, $k = 384$ tends to have the most rejections, but for large α, $k = 768$ has the most rejections. Also, we compared the procedure in Theorem 2.3 (labeled Procedure 2.3) with the BY procedure. The results are displayed in Table 2.2. As seen from Table 2.2, for most values of k, our procedure outperforms the BY procedure in terms of the number of rejections. When $\alpha = 0.001$, the BY procedure cannot reject any hypotheses, but the procedure in Theorem 2.3 has at least 8 rejections for all the values of k considered.

2.6 Simulation Study

A simulation study was conducted to address the performances of the proposed procedures. We will refer to the procedures in Theorems 2.1-2.4 as Procedures 2.1-2.4, respectively. Specifically, we addressed the following two questions:

1. How do Procedures 2.1 and 2.3 compare against the BH and BY procedures in terms of FDR and power?

2. How do Procedures 2.2 and 2.4 compare against the BH procedure in terms of FDR and power?

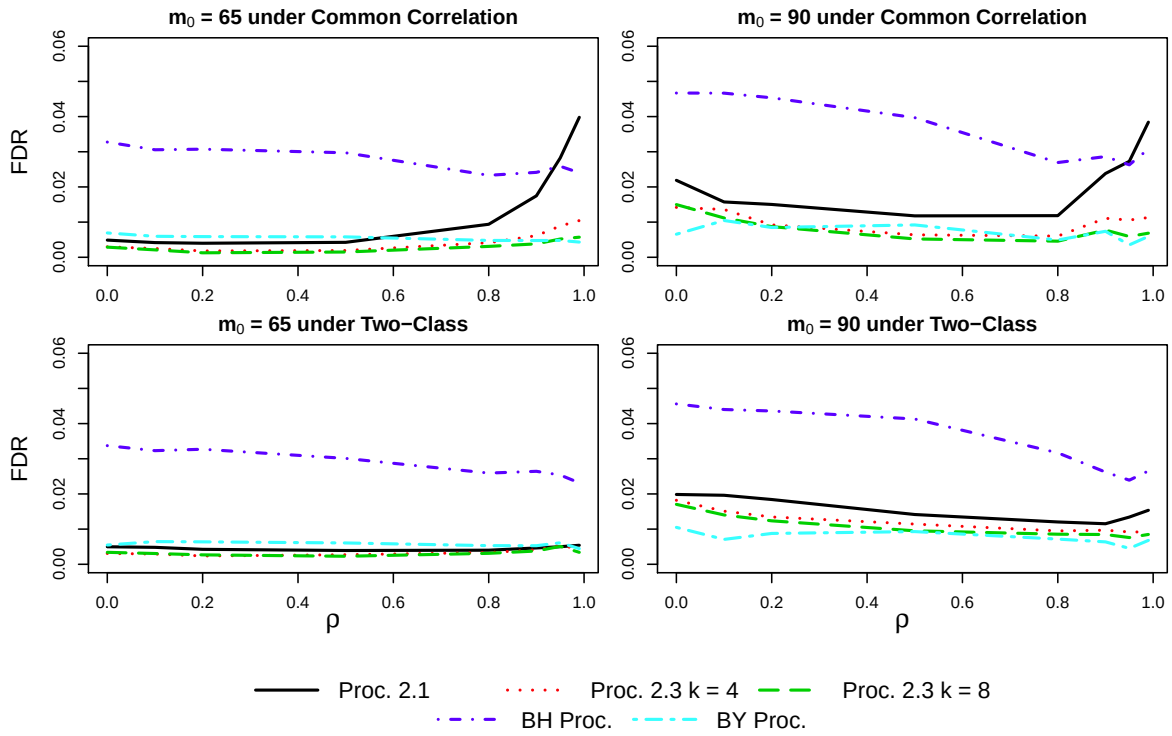


Figure 2.2 The FDR of Procedure 2.1 (solid line), Procedure 2.3 with k = 4 (dotted line) and k = 8 (dashed line), the BH procedure (dotted dash), and the BY procedure (long dash short dash) for 100 hypotheses under common correlation (top row) and two-class structure (bottom row) for $m_0 = 65$ (left) and $m_0 = 90$ (right).

In each simulation, n independent m dimensional random normal vectors with covariance matrix Σ and components $Z_i \sim N(\mu_i, 1), i = 1, \ldots, m$, were generated. The p-value for testing $H_i : \mu_i = 0$ vs. $H_i' : \mu_i > 0$ was calculated using a one-sample t-test for each i. The μ_i corresponding to each false null hypothesis is set to the value at which the power of one-sample t-test at level 0.05 is 0.75. We considered three different structures for the covariance matrix $\Sigma = ((\sigma_{ij}))$:

- Independence Structure: $\sigma_{ij} = 0$ for $i \neq j$.
- Common Correlation Structure: $\sigma_{ij} = \rho$ for $i \neq j$.

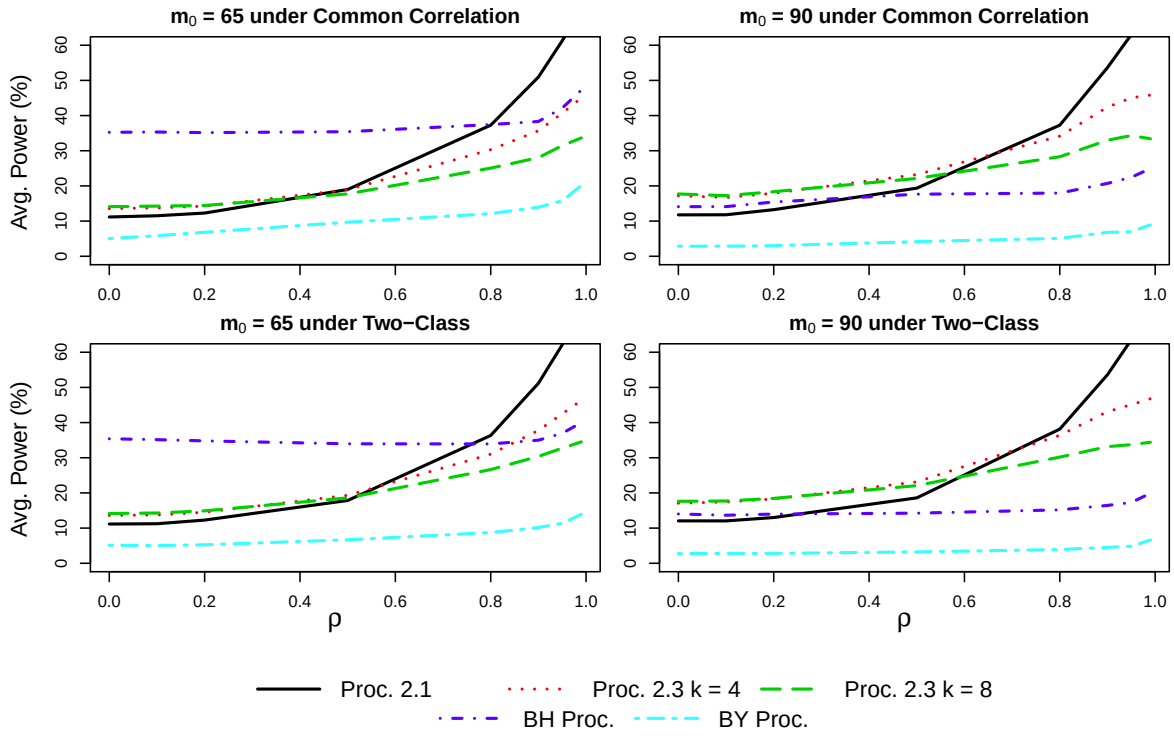


Figure 2.3 Average power (%) of Procedure 2.1 (solid line), Procedure 2.3 with k = 4 (dotted line)and k = 8 (dashed line), the BH procedure (dotted dash), and the BY procedure (long dash short dash) for 100 hypotheses under common correlation (top row) and two-class structure (bottom row) for $m_0 = 65$ (left) and $m_0 = 90$ (right).

- Two-Class Structure: We used the two-class structure as described in Romano et al. (2008). The variables are placed into two classes of size m/2. Within each class, there is common correlation of ρ and across classes there is correlation -ρ. For $i \neq j$,

$$\sigma_{ij} = \begin{cases} \rho & \text{if both } i, j \in \{1, \ldots, m/2\} \text{ or both i, j} \in \{m/2+1, \ldots, m\} \\ -\rho & \text{otherwise.} \end{cases}$$

We set $\alpha = 0.05$ and $m = 100$. The sample size n was set to 10, and the number of true null hypotheses m_0 was set to 65 and 90. The hypotheses were ordered using the 'sum of squares ordering' used in Example 2.1 from Section 2.5. For the two-class structure, the true and false null hypotheses were randomly distributed between the

two classes (specifically, each hypothesis was assigned an independent random number and the hypotheses corresponding to the smallest $m/2$ random numbers were assigned to the first class with the remaining assigned to the second one). We had 5,000 runs of simulation for each of the procedures considered. We noted the false discovery proportion and the proportion of correctly rejected false null hypotheses for each procedure in each of these runs. The simulated FDR and average power (the expected proportion of correctly rejected false null hypotheses) were obtained by averaging out the corresponding 5,000 values.

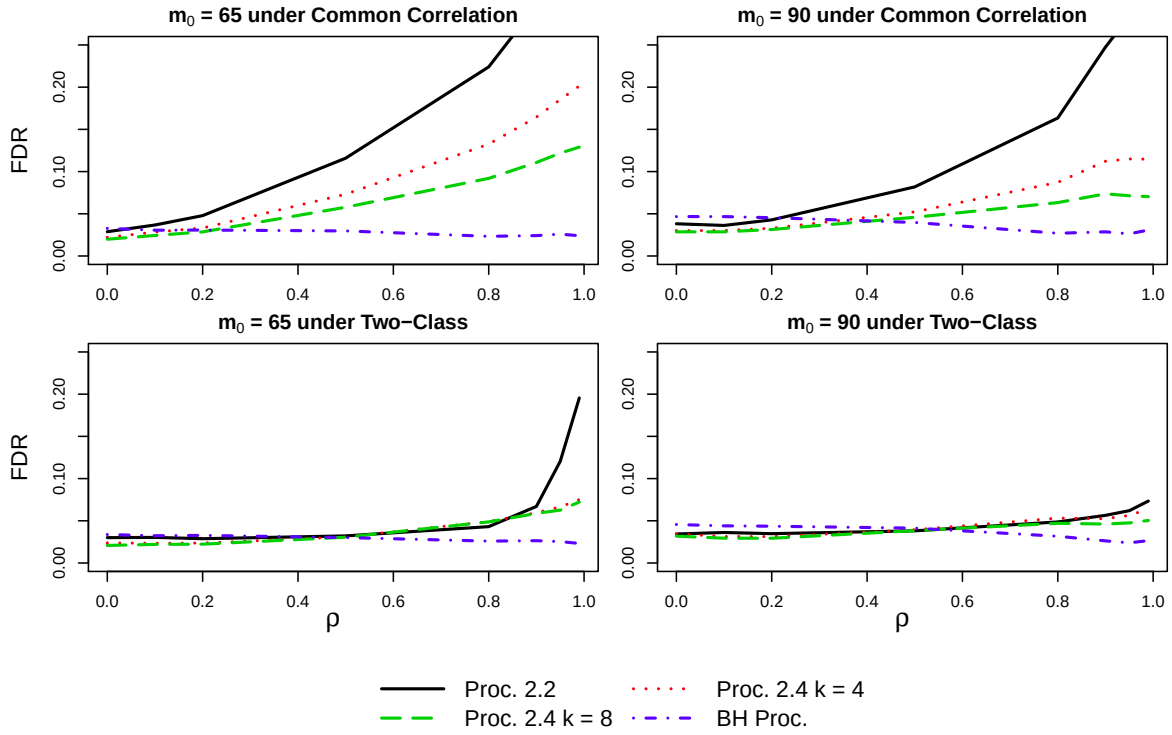


Figure 2.4 The FDR of Procedure 2.2 (solid line), Procedure 2.4 with k = 4 (dotted line) and k = 8 (dashed line), and the BH procedure (dotted dash) for 100 hypotheses under common correlation (top row) and two-class structure (bottom row) for $m_0 = 65$ (left) and $m_0 = 90$ (right).

We first looked at Procedures 2.1 and 2.3 with $k = 4$ and 8, and compared them with the BH and BY procedures. Figures 2.2 and 2.3 display these comparisons in terms of the simulated FDR and average power, respectively, under common

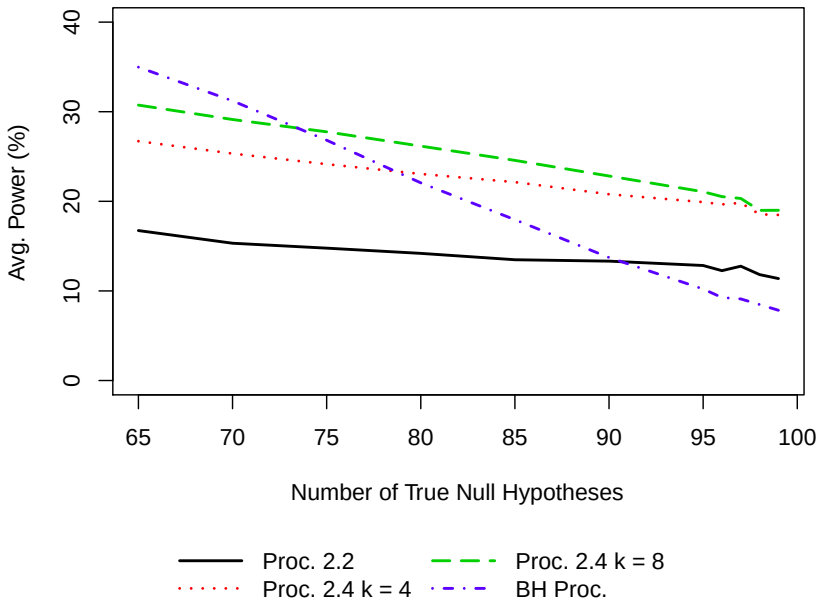


Figure 2.5 Average power (%) of Procedure 2.2 (solid line), Procedure 2.4 with k = 4 (dotted line) and k = 8 (dashed line), and the BH procedure (dotted dash) for 100 hypotheses under independence for various values of m_0.

correlation and two-class structures across different values of ρ from 0 to 1. All of the procedures are seen to control the FDR at the desired 5% level. In terms of power, both Procedures 2.1 and 2.3 perform much better than the BY procedure. When the proportion of true null hypotheses is large, Procedure 2.3 tends to outperform the BH procedure in terms of average power. The powers of all the procedures considered tend to increase as ρ increases, with Procedure 2.1 showing the steepest increase.

Next, we looked at Procedures 2.2 and 2.4 with $k = 4$ and 8, and compared them with the BH procedure. Figure 2.4 compares these procedures in terms of the simulated FDR under common correlation and two-class structures with ρ varying from 0 to 1. Under common correlation structure, Procedures 2.2 and 2.4 are seen to control the FDR with mild correlation, but as the correlation increases these procedures seem to lose control of the FDR, although Procedure 2.4 seems more robust

against departures from independence as k is increased. Under two-class structure, Procedures 2.2 and 2.4 both control the simulated FDR for nearly all values of ρ.

Since Procedures 2.2 and 2.4 do not maintain control of the FDR for all values of ρ, we only considered the independence case when assessing the powers of these procedures. Figure 2.5 shows that the average powers of Procedure 2.2 and Procedure 2.4 with $k = 4$ and 8 along with that of the BH procedure under independence as m_0 varies from 65 to 99. The average powers of all these procedures tend to decrease as the number of true null hypotheses increases; however, the powers of Procedures 2.2 and 2.4 decrease much more slowly than that of the BH procedure. When the number of true null hypotheses is large, Procedure 2.4 tends to outperform the BH procedure in terms of average power.

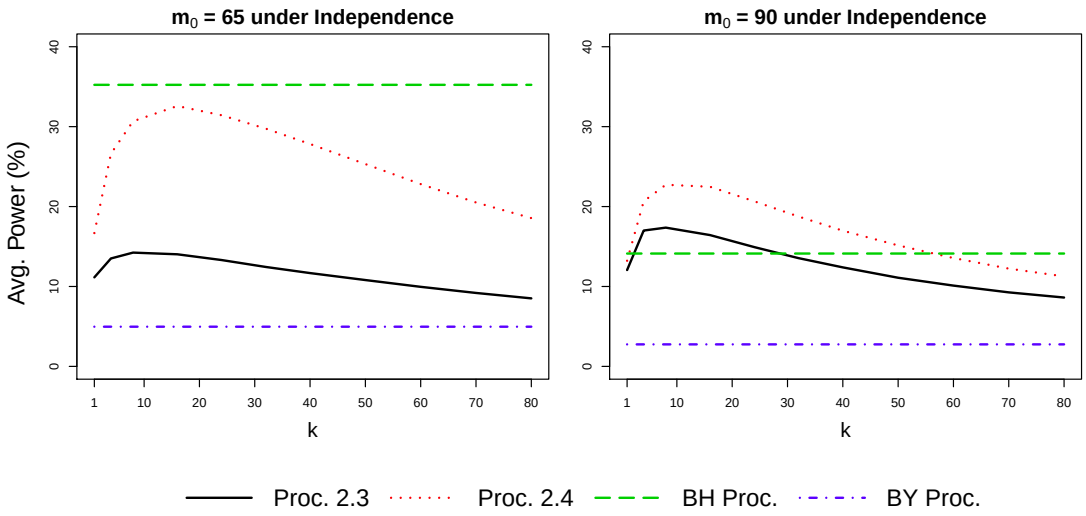


Figure 2.6 Average power (%) of Procedure 2.3 (solid line), Procedure 2.4 (dotted line), the BH procedure (dashed line), and the BY procedure (dotted dash) for 100 hypotheses under independence for $m_0 = 65$ (left) and $m_0 = 90$ (right).

Finally, in order to better understand how the value of k affects the powers of our proposed procedures, additional simulation studies were performed with $\rho = 0$ and varying k from 1 to 80. As seen from Figure 2.6, the powers of our proposed

procedures first increase with k, then slowly decrease with k and a value of k between about 5 and 15 seems to be a good choice for maximizing the power. When $m_0 = 65$, the BH procedure is more powerful, but when $m_0 = 90$, procedures 2.3 and 2.4 are more powerful for a well-chosen k. For all values of k, procedures 2.3 and 2.4 are more powerful than the BY procedure.

2.7 Discussion

In this chapter, we have developed methods which control the FDR and exploit the structure of pre-ordered hypotheses. We have been able to produce the desired methods in the most simple as well as a general setting covering different dependence scenarios. Our simulation study shows that in some cases, such as when the proportion of true null hypotheses is large, our procedures can outperform the BH and BY procedures.

Using some of the techniques developed in this chapter, it is possible to develop other types of fixed sequence procedures controlling the FDR, such as a fallback-type procedure. Unlike the conventional and generalized fixed sequence procedures developed in this chapter, the fallback-type procedure tests the remaining hypotheses no matter how many earlier hypotheses are accepted. The FWER controlling fallback procedure uses a series of weights $w_1, \ldots, w_m$ such that $\sum_{i=1}^{m} w_i = 1$. The procedure rejects H_i if $P_i \leq \alpha_i$ where $\alpha_i = w_i\alpha + \alpha_{i-1}$ if H_{i-1} is rejected and $\alpha_i = w_i\alpha$ otherwise (Wiens 2003; Wiens and Dmitrienko 2005). Using similar techniques that we used to derive the results in this chapter, we were able to derive a fallback-type FDR controlling procedure. The procedure, which controls the FDR at level α under Assumption 1.1, is to reject H_i if $P_i \leq (1 + R_{i-1}/(m - i + 1))\alpha_i$, where R_i is the number of rejected hypotheses among the first i hypotheses tested. Such a procedure

has interesting applications towards network traffic analysis using stream data. The proof of the FDR control and some theoretical discussion of this procedure are left for another communication.

CHAPTER 3

THE TESTING OF HIERARCHICALLY ORDERED HYPOTHESES

Due to the amount of notation used in this chapter, for convenience, we have summarizes the commonly used notation here. Each commonly used notation is listed along with its definition and where each symbol can be found.

Symbol	Description	Section	Page
$\mathcal{M}, m$	The set of tested hypotheses $\{H_1, \ldots, H_m\}$ and its cardinality.	3.2	43
$\mathcal{M}_i, m_i$	The set of descendant hypotheses of H_i and its cardinality.	3.2	43
$\mathcal{D}_i, d_i$	The set of ancestor hypotheses of H_i and its cardinality, also referred to as depth.	3.2	43
$T(\cdot)$	A function that takes an index of a hypothesis and returns the index of its parent hypothesis.	3.2	43
$\mathcal{F}_d$	The set of hypotheses with depth d, $\mathcal{F}_d = \{H_i : d_i = d\}$.	3.2	43
$\mathcal{G}_d$	The union of $\mathcal{F}_1, \ldots, \mathcal{F}_d$.	B.2	102
D	The maximum depth of the hypotheses so that $\mathcal{G}_D = \mathcal{M}$.	3.2	43
ℓ	The total number of leaf hypotheses.	3.2	43
ℓ_i	The number of leaf hypotheses in set $\mathcal{M}_i$.	3.2	43
$R(\mathcal{A}), R$	The number of rejected hypotheses belonging to set $\mathcal{A}$ and $R = R(\mathcal{M})$.	3.2	43
$V(\mathcal{A}), V$	The number of falsely rejected hypotheses belonging to set $\mathcal{A}$ and $V = V(\mathcal{M})$.	3.2	43
$\alpha_i(\cdot)$	The critical function for testing the i^{th} hypothesis.	3.2	44

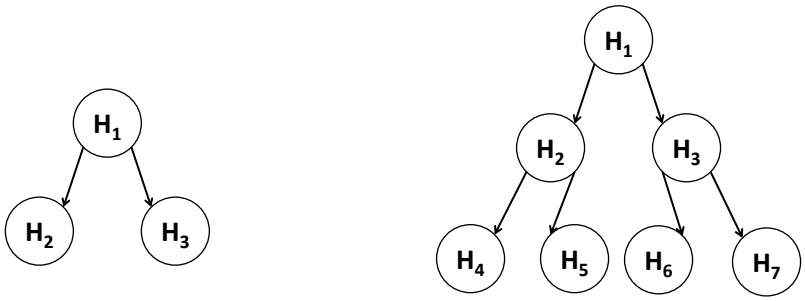
3.1 Introduction

In many problems involving the testing of multiple hypotheses, the hypotheses have an intrinsic, hierarchical structure such as a tree-like or graphical structure. Hierarchical structures often arise in multiple testing problems involving clinical trials (Mehrotra and Heyse 2004; Dmitrienko et al. 2007; Huque and Alosh 2008), genomics research (Yekutieli et al. 2006; Goeman and Mansmann 2008; Heller et al. 2009; Guo et al. 2010) and fMRI studies (Benjamini and Heller 2007). In general, hierarchical testing typically occurs in multiple testing problems where, upon the rejection of one hypothesis, followup hypotheses are to be tested. For instance, Heller et al. (2009) introduced a hierarchical testing approach for analyzing microarray data where individual genes were grouped into gene sets. The gene sets were tested and upon successfully rejecting a gene set, the associated individual genes were tested. Guo et al. (2010) and Mehrotra and Heyse (2004) used a similar hierarchical testing approach for time-course microarray data and clinical safety data, respectively. Meinshausen (2008) introduced a hierarchical familywise error rate (FWER) controlling method for addressing the problem of variable selection in a multiple regression model. Yekutieli et al. (2006) applied a hierarchical testing method controlling the FDR, introduced in Yekutieli (2008b), to a microarray experiment involving different strains of inbred mice among various regions of the brain. Benjamini and Heller (2007) used a hierarchical testing approach to study fMRI data, where the brain was divided into brain regions and each brain region was tested for significance. If a brain region was significant, the voxels within the brain region were tested.

In the field of multiple testing, the problem of controlling the FWER for testing hierarchically ordered hypotheses has received considerable attention (Dmitrienko et al. 2006, 2007; Goeman and Mansmann 2008; Huque and Alosh 2008; Meinshausen 2008; Brechenmacher et al. 2011; Goeman and Finos 2012); however, FWER control

can be too conservative for large-scale multiple testing. There has been very little work towards developing general methods for testing hierarchically ordered hypotheses that control the FDR, even though FDR is a more appropriate error measure for large scale multiple testing. To our knowledge, only Yekutieli (2008b) has provided a general method for testing hierarchically ordered hypotheses that is specifically intended for FDR control. Yekutieli's procedure, which is based on the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg 1995), is only shown to control the FDR under independence. Some of the aforementioned procedures (Mehrotra and Heyse 2004; Benjamini and Heller 2007; Heller et al. 2009; Guo et al. 2010) can only be applied to special hierarchies consisting of only 2 levels.

In this chapter, we propose new FDR controlling methods for testing hierarchically ordered hypotheses under various dependencies. Our approach towards controlling the FDR for testing hierarchically ordered hypotheses is different from that of Yekutieli's. First, to assist in the development of our hierarchical testing procedures, we introduce a generalized stepwise procedure which generalizes stepup, stepdown, and stepup-down procedures to the case where each hypothesis is tested with a different set of critical constants. Then, we describe a general hierarchical testing approach where the hypotheses are organized into different families according to their depth in the hierarchical structure and each family is tested by using the generalized stepwise procedure. Based on this approach, we were able to develop several new hierarchical testing procedures which control the FDR under various dependence structures including positive and arbitrary dependence. To the best of our knowledge, our procedures are the first procedures developed for testing hierarchically ordered hypotheses with control of the FDR under dependence structures other than independence. Furthermore, our simulation study shows that these procedure are quite powerful. Our most powerful procedure, which we prove can control the FDR under positive block dependence, significantly outperforms Yekutieli's procedure in



(a) (b)

Figure 3.1 (a) An example of a hierarchical structure with 3 hypotheses. H_2 and H_3 are only tested if H_1 is rejected. (b) An example of a hierarchical structure with 7 hypotheses.H_2 and H_3 are only tested if H_1 is rejected, H_4 and H_5 are only tested if H_2 is rejected, and H_6 and H_7 are only tested if H_3 is rejected.

terms of power, even though Yekutieli's procedure is only shown to control the FDR under independence, which is a special case of positive block dependence.

Another interesting finding of this research is that when the hierarchy takes on some special configurations, our procedures reduce to existing procedures. For example, when there is no hierarchical structure among the tested hypotheses, our proposed procedures reduce to the BH procedure and the BY procedure. When the hierarchy takes on a fixed sequence structure, our procedures are equivalent to the fixed sequence procedures of Chapter 2. This shows that our procedures are the combination of stepwise and fixed sequence methods.

The rest of this chapter is outlined as follows. In Section 3.2, we provide relevant notation and definitions that will be used throughout this chapter. Section 3.3 presents our proposed generalized stepwise procedure. Section 3.4 presents our new procedures which control the FDR for testing hierarchically order hypotheses. Section 3.5 and 3.6 presents a simulation study and real data analysis where we compare our procedures with Yekutieli's procedure. Finally, Section 3.7 provides some brief discussion.

3.2 Preliminaries

Suppose the m test hypotheses are organized hierarchically in a tree-like structure where each hypothesis can have several child hypotheses but at most one parent hypothesis. Let $\mathcal{M} = \{H_1, \ldots, H_m\}$ be the set of the m tested hypotheses. Let $T : \{0, \ldots, m\} \to \{0, \ldots, m\}$ be a function that takes an index of a hypothesis and returns the index of the parent hypothesis where $T(0) = 0$. That is, if H_i has a parent hypothesis, its parent hypothesis is $H_{T(i)}$; otherwise, H_i does not have a parent hypothesis and $T(i) = 0$. Define $T^0(i) = i$ and $T^k(i) = T(T^{k-1}(i))$. Let $\mathcal{D}_i = \{H_j : T^k(i) = j, j \neq 0 \text{ for } k = 0, \ldots, m\}$ so that $\mathcal{D}_i$ is the set of ancestor hypotheses of H_i, which includes H_i. Let d_i be the cardinality of $\mathcal{D}_i$, $d_i = |\mathcal{D}_i|$. The depth of H_i in the hierarchy is defined as d_i. If $\mathcal{D}_i = \{H_i\}$, then H_i does not have a parent hypothesis. Let $\mathcal{M}_i = \{H_j : T^k(j) = i \text{ for } k = 0, \ldots, m\}$ so that $\mathcal{M}_i$ is the set of descendant hypotheses of H_i, which includes H_i. We will refer to the hypotheses in set $\mathcal{M}_i$ as the subtree under H_i. Let m_i be the cardinality of $\mathcal{M}_i$, $m_i = |\mathcal{M}_i|$. If $\mathcal{M}_i = \{H_i\}$, then H_i has no children and it is referred to as a leaf hypothesis. We denote the number of leaf hypotheses by ℓ and the number of leaf hypotheses in the subtree under H_i by ℓ_i. Formally, $\ell = \sum_{H_j \in \mathcal{M}} I\{m_j = 1\}$ and $\ell_i = \sum_{H_j \in \mathcal{M}_i} I\{m_j = 1\}$. Our procedures described in Section 3.4 group the hypotheses into D families, $\mathcal{F}_1, \ldots, \mathcal{F}_D$, by depth, where $\mathcal{F}_d$ contains the hypotheses with depth d for $d = 1, \ldots, D$. Let $\mathcal{F}_d = \{H_i \in \mathcal{M} : d_i = d\}$. For example, in figure 3.1(a), $T(2) = T(3) = 1$ and H_2 and H_3 are leaf hypotheses. In figure 3.1(b), $T(6) = T(7) = 3, \mathcal{D}_6 = \{1, 3, 6\}, \mathcal{M}_2 = \{2, 4, 5\}$, and $\mathcal{F}_3 = \{4, 5, 6, 7\}$.

The hypotheses in the hierarchical structure are tested hierarchically by some testing procedure based on their corresponding p-values $P_1, \ldots, P_m$. By hierarchical testing, we mean a hypothesis is only tested if its parent hypothesis has been rejected or it does not have a parent hypothesis. For any set $\mathcal{A} \subseteq \mathcal{M}$ define $R(\mathcal{A})$ and $V(\mathcal{A})$ to be the number of rejected hypotheses and falsely rejected hypotheses, respectively,

for the testing procedure among hypotheses in set $\mathcal{A}$. For example, $R(\mathcal{M})$ and $V(\mathcal{M})$ are the number of rejected hypotheses and falsely rejected hypotheses among all the m tested hypotheses, respectively, and $R(\mathcal{M}_i)$ and $V(\mathcal{M}_i)$ are number of rejected hypotheses and falsely rejected hypotheses among the hypotheses in the subtree $\mathcal{M}_i$, respectively. In keeping consistent with the notation used in the rest of this dissertation, we will use R and V which is taken to mean $R(\mathcal{M})$ and $V(\mathcal{M})$, respectively.

Many testing procedures are stepwise methods which are based on the ordered p-values $P_{(1)}, \ldots, P_{(m)}$ with corresponding hypotheses $H_{(1)}, \ldots, H_{(m)}$. Typically, the rejection thresholds of a stepwise procedure are described by a set of non-decreasing critical constants, but in this chapter, for convenience, we will instead test the hypotheses using a non-decreasing, non-negative function $\alpha_0 : \{0, \ldots, m+1\} \to [0, \infty)$ called a critical function, where $\alpha_0(0) = 0$. For example, the critical function of the BH procedure is $\alpha_0(r) = r\alpha/m$. Recall that a stepwise procedure first determines the number of rejections R based on the critical function, then for each $i = 1, \ldots, m$, it rejects H_i if $P_i \leq \alpha_0(R)$ and accepts H_i if $P_i > \alpha_0(R)$. With $P_{(0)} \equiv 0$ and $P_{(m+1)} \equiv \infty$, a stepup procedure sets $R = \max\{0 \leq r \leq m : P_{(r)} \leq \alpha_0(r)\}$. A stepdown procedure sets $R = \min\{1 \leq r \leq m+1 : P_{(r)} > \alpha_0(r)\} - 1$. Finally, a stepup-down procedure of order k, which generalizes stepup and stepdown procedures, sets $R = \max\{0 \leq r \leq k-1 : P_{(r)} \leq \alpha_0(r)\}$ if $P_{(k)} > \alpha_0(k)$ and $R = \min\{k+1 \leq r \leq m+1 : P_{(r)} > \alpha_0(r)\} - 1$ if $P_{(k)} \leq \alpha_0(k)$. When $k = m$, the stepup-down procedure reduces to the stepup procedure and when $k = 1$, it reduces to the stepdown procedure. It should be noted that the event $\{P_{(r)} \leq \alpha_0(r)\}$ is equivalent to the event $\{r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_0(r)\}\}$. Thus, the number of rejections can also be expressed by

$$R = \max\left\{0 \leq r \leq m : r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_0(r)\}\right\} \tag{3.1}$$

for the stepup procedure,

$$R = \min\left\{1 \le r \le m+1 : r > \sum_{i=1}^{m} I\{P_i \le \alpha_0(r)\}\right\} - 1 \tag{3.2}$$

for the stepdown procedure, and

$$R = \begin{cases} \max\left\{0 \le r \le k-1 : r \le \sum_{i=1}^{m} I\{P_i \le \alpha_0(r)\}\right\} & \text{if } k > \sum_{i=1}^{m} I\{P_i \le \alpha_0(k)\} \\ \min\left\{k+1 \le r \le m+1 : r > \sum_{i=1}^{m} I\{P_i \le \alpha_0(r)\}\right\} - 1 & \text{if } k \le \sum_{i=1}^{m} I\{P_i \le \alpha_0(k)\} \end{cases} \tag{3.3}$$

for the stepup-down procedure of order k. Refer to Tamhane et al. (1998) and Sarkar (2002) for further discussion on stepwise procedures.

We consider several types of joint dependence throughout this chapter: arbitrary dependence, positive dependence, and block dependence. Positive dependence is characterized by Assumption 1.1 of Chapter 1. Block dependence is characterized by the following assumption.

Assumption 3.1. Block Dependence Assumption

For each $d = 1, \ldots, D$, the p-values corresponding to the hypotheses in $\mathcal{F}_d$ are independent of the p-values corresponding to the hypotheses not in $\mathcal{F}_d$.

Assumption 3.1 only characterizes the joint dependence of the p-values across families but does not describe the joint dependence within families.

3.3 Generalized Stepwise Procedure

In order to present our hierarchical testing procedures in the next section, in this section, we present a new type of procedure called the generalized stepwise procedure, which generalizes the usual stepup, stepdown, and stepup-down procedures. In a non-hierarchical multiple testing problem where a stepwise procedure is used to test

the hypotheses, the tested hypotheses often have the same importance and thus, it is natural to test those hypotheses with the same critical function, as shown in (3.1), (3.2), and (3.3). However, when the hypotheses have a hierarchical structure, the importance of a hypothesis depends on where it is located in the hierarchy. Hence, each hypothesis should be tested with a critical function that reflects its importance, and so for this reason, we generalize the usual stepwise procedure.

Given m non-decreasing critical functions $\alpha_1(r), \ldots, \alpha_m(r)$ our proposed generalized stepwise procedure rejects H_i if $P_i \leq \alpha_i(R)$ for each $i = 1, \ldots, m$, where R is determined as follows. For the generalized stepup procedure,

$$R = \max \left\{ 0 \leq r \leq m : r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\} \right\}, \tag{3.4}$$

for the generalized stepdown procedure,

$$R = \min \left\{ 1 \leq r \leq m+1 : r > \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\} \right\} - 1, \tag{3.5}$$

and for the generalized stepup-down procedure of order k,

$$R = \begin{cases} \max \left\{ 0 \leq r \leq k-1 : r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\} \right\} & \text{if } k > \sum_{i=1}^{m} I\{P_i \leq \alpha_i(k)\} \\ \min \left\{ k+1 \leq r \leq m+1 : r > \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\} \right\} - 1 & \text{if } k \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_i(k)\}. \end{cases} \tag{3.6}$$

A simple algorithm for determining R is presented at the end of this section.

It is easy to see that when $\alpha_i(r) = \alpha_0(r)$ for each $i = 1, \ldots, m$, (3.4), (3.5), and (3.6) reduce to (3.1), (3.2), and (3.3), respectively, so that the generalized stepwise procedure reduces to the usual stepwise procedure. It should be noted that when $k = m$, (3.6) reduces to (3.4), and when $k = 1$, (3.6) reduces to (3.5).

The generalized stepwise procedure is fairly general and we present two examples to show its broad applicability.

Example 3.1. Consider a weighted multiple testing problem where H_i has corresponding weight $w_i, i = 1, \ldots, m$. A weighted stepwise procedure with critical function $\alpha_0(r)$ tests H_i based on weight-adjusted p-values P_i/w_i instead of P_i. This is equivalent to the generalized stepwise procedure with critical function $\alpha_i(r) = w_i\alpha_0(r)$ so that weighted stepwise procedures can be regarded as a special case of the generalized stepwise procedure. The generalized stepwise procedure can handle more complex scenarios where the weights are a function of the number of rejections, say $w_i(r)$. For example, suppose there are two families of hypotheses where the second family is tested only if there is at least one rejection in the first family (such a testing strategy is often called a parallel gatekeeping strategy in clinical trials, see Dmitrienko et al. (2003)). This configuration of hypotheses can be tested using the generalized stepdown procedure with critical functions of the form $\alpha_i(r) = w_i(r)\alpha_0(r)$ where $\alpha_0(r)$ is a common critical function, $w_i(r) = 1$ if H_i belongs to the family 1, and $w_i(r) = I\{r \geq 2\}$ if H_i belongs to family 2. Thus, parallel gatekeeping procedures can also be regarded as a special case of the generalized stepwise procedure.

Example 3.2. Fixed sequence procedures assume the testing order of the hypotheses has been specified a priori and that H_i is not tested unless $H_1, \ldots, H_{i-1}$ have all been rejected. Chapter 2 showed that the fixed sequence procedure that rejects H_i when $P_i \leq m\alpha/(m - i + 1)$ controls the FDR at level α under arbitrary dependence. This procedure is a special case of the generalized stepwise procedure with critical functions $\alpha_i(r) = I\{r \geq i\}m\alpha/(m - i + 1), i = 1, \ldots, m$. Other fixed sequence procedures can be defined similarly.

From equations (3.4)-(3.6), it can be seen that many of the familiar properties of stepwise procedures also hold for the generalized stepwise procedure. For example, R is a coordinatewise non-increasing function of the p-values, and R is a non-decreasing function of k (i.e., a stepup-down procedure of order k rejects more hypotheses than

a stepup-down procedure of order $k-1$). Another important property is a self-consistency property, which allows us to express R in (3.4)-(3.6) with the following equality:

$$R = \sum_{i=1}^{m} I\{P_i \leq \alpha_i(R)\}. \tag{3.7}$$

(Blanchard and Roquain (2008) discussed a weaker self-consistency condition, which is the inequality $R \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_0(R)\}$). In words, (3.7) shows that R as determined by (3.4)-(3.6) is the number of rejections by the generalized stepwise procedure. Hence, the event $\{H_i \text{ is rejected}\}$ can be expressed as $\{P_i \leq \alpha_i(R)\}$ with R being the number of rejections. We will show the property holds for (3.6), noting that (3.4) and (3.5) are special cases of (3.6). Define $\psi(r) = \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}$. When $k > \psi(k)$, then $R = \max\{0 \leq r \leq k-1 : r \leq \psi(r)\}$ and if $k \leq \psi(k)$, then $R+1 = \min\{k+1 \leq r \leq m+1 : r > \psi(r)\}$. In either case, it is easy to see that $R \leq \psi(R)$ and $R+1 > \psi(R+1)$. The fact that $\psi(R+1) < R+1$ implies $\psi(R+1) \leq R$. Thus, $R = \psi(R)$ since $R \leq \psi(R) \leq \psi(R+1) \leq R$.

To conclude this section, we present an efficient algorithm for finding the number of rejections by the generalized stepwise procedure. The algorithm is particularly useful when the number of hypotheses is very large.

Algorithm 3.1. Given a positive integer $1 \leq k \leq m$ and critical functions $\alpha_i(\cdot), i = 1, \ldots, m$, define $\psi(r) = \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}$.

1. Let $t = 1$ and $r_t = k$.

2. If $k > \psi(k)$, then

   (a) Increment t by 1, and set $r_t = \psi(r_{t-1})$.

   (b) If $r_t \leq \psi(r_t)$, then let $R = r_t$ and stop; otherwise, if $r_t > \psi(r_t)$, repeat step 2(a).

3. Otherwise, if $k \leq \psi(k)$, then

   (a) Increment t by 1, and set $r_t = \psi(r_{t-1}) + 1$.

   (b) If $r_t > \psi(r_t)$, then let $R = r_t - 1$ and stop; otherwise, if $r_t \leq \psi(r_t)$, repeat step 3(a).

Proposition 3.1. *The value of R in (3.6) can be solved by Algorithm 3.1.*

Proof. The proof is in Appendix B. ■

3.4 Hierarchical FDR Control

In this section, we describe our general approach to test hierarchically ordered hypotheses. The hypotheses are arranged into D families, $\mathcal{F}_1, \ldots, \mathcal{F}_D$, where $\mathcal{F}_d$ is the family of hypotheses with depth $d, d = 1, \ldots, D$. Given non-decreasing critical functions $\alpha_1(r), \ldots, \alpha_m(r)$, the hypotheses are tested as follows.

Definition 3.1. Hierarchical Testing Procedure

1. *Test $\mathcal{F}_1$ by using the generalized stepup procedure with critical functions $\alpha_i(r), H_i \in \mathcal{F}_1$. Let $\mathcal{S}_1$ be the set of rejected hypotheses and $R(\mathcal{F}_1)$ be the number of rejected hypotheses in $\mathcal{F}_1$.*

2. *Generally, sequentially test $\mathcal{F}_d, d = 2, \ldots, D$, using the generalized stepup procedure with critical functions $\alpha_i^*(r) = I\{H_{T(i)} \text{ is rejected}\}\alpha_i(r + \sum_{j=1}^{d-1} R(\mathcal{F}_j))$, $H_i \in \mathcal{F}_d$. Let $\mathcal{S}_d$ be the set of rejected hypotheses and $R(\mathcal{F}_d)$ be the number of rejected hypotheses in $\mathcal{F}_d$.*

3. *The set of rejected hypotheses is $\bigcup_{d=1}^{D} \mathcal{S}_d$ and the total number of rejections is $R = \sum_{d=1}^{D} R(\mathcal{F}_d)$.*

We describe the above procedure as a hierarchical testing procedure since the procedure will accept any hypothesis whose parent hypothesis has been accepted. This can be seen in the construction of the critical functions in step 2 where $\alpha_i^*(r) = 0$ if H_i's parent has not been rejected so that H_i cannot be rejected. It should be noted that the parents of the hypotheses in $\mathcal{F}_d$ are $\mathcal{F}_{d-1}$, which is tested before testing $\mathcal{F}_d$. Hence, for each $H_i \in \mathcal{F}_d$, the event $\{H_{T(i)}$ is rejected$\}$ is observed by the time $\mathcal{F}_d$ is tested.

Remark 3.1. In Definition 1, when all the hypotheses in $\mathcal{F}_d$ have the same critical functions and every $H_i \in \mathcal{F}_d$ can be tested (i.e., $H_{T(i)}$ is rejected), the generalized stepup procedure used for testing $\mathcal{F}_d$ reduces to the usual stepup procedure. However, our critical functions for testing hierarchically ordered hypotheses, which are presented in the next subsections, are not the same and depend on where the hypothesis is located in the hierarchy. Furthermore, since the hypotheses in $\mathcal{F}_d$ may not have the same parent, $H_{T(i)}$ could be rejected for some, but not all, of $H_i \in \mathcal{F}_d$. Hence, only in an uncommon case does the generalized stepup procedure reduce to the stepup procedure for testing $\mathcal{F}_d$.

Following (3.7), the hierarchical testing procedure has the following self-consistency property in each family

$$R(\mathcal{F}_d) = \sum_{H_i \in \mathcal{F}_d} I\{P_i \leq \alpha_i^*(R(\mathcal{F}_d))\}, d = 1, \ldots, D,$$

where $\alpha_i^*(r) = I\{H_{T(i)}$ is rejected$\}\alpha_i(r + \sum_{j=1}^{d-1} R(\mathcal{F}_j))$. Hence, $\{H_i$ is rejected$\}$ is equivalent to the event $\{H_{T(i)}$ is rejected$, P_i \leq \alpha_i \left(\sum_{j=1}^{d_i} R(\mathcal{F}_j)\right)\}$ where $\sum_{j=1}^{d_i} R(\mathcal{F}_j)$ is the number of rejections in the first d_i families. This property will be useful to prove FDR control of our procedures.

We will often compare our proposed procedures with two existing hierarchical testing procedures: Yekutieli's procedure (Yekutieli 2008b) and Meinshausen's

procedure (Meinshausen 2008). Yekutieli's hierarchical testing procedure controls the FDR at level α under independence. Yekutieli's procedure groups the hypotheses into families, where a family contains all the hypotheses that share the same parent and the first family contains the hypotheses with no parent. Hence, the grouping is different than our depth-based grouping. Testing begins with the first family, which is tested by the BH procedure at level $\alpha/2.88$. The remaining families are tested by the BH procedure at level $\alpha/2.88$ only if the family's parent hypothesis is rejected. Although both procedures test families of hypotheses using a stepup based procedure, the generalized stepup procedure provides the flexibility needed to control the FDR under dependence.

Meinshausen's hierarchical testing procedure controls the FWER at level α under arbitrary dependence. It rejects H_i if $P_i \leq \ell_i \alpha/\ell$ and $H_{T(i)}$ is rejected or H_i does not have a parent hypothesis. Meinshausen's procedure is equivalent to the hierarchical testing procedure with critical constants $\alpha_i(r) = \ell_i \alpha/\ell, i = 1, \ldots, m$.

Now that we have defined our general hierarchical testing approach, we will consider various dependence structures, such as positive dependence, arbitrary dependence, and block dependence, and develop hierarchical testing procedures which control the FDR under these dependence structures. The proofs of all the theorems in this section are in Appendix B.

3.4.1 Procedure under Positive Dependence

We first consider positive dependence. Positive dependence has received a fair amount of attention in multiple testing due to the fact that several multiple testing procedures can control the type I error rate under this type of dependence (see Sarkar (1998), Benjamini and Yekutieli (2001), Sarkar (2002), Yekutieli (2008a), Guo and Sarkar (2013)). Our procedure under positive dependence is as follows.

Theorem 3.2. *FDR Control under Positive Dependence*

Under Assumption 1.1, the hierarchical testing procedure with critical functions

$$\alpha_i(r) = \frac{\ell_i \alpha}{\ell} \frac{m_i + r - 1}{m_i}$$

strongly controls the FDR at level α.

Consider the special case where there is no hierarchical ordering (i.e., $\ell = m$ and $\ell_i = m_i = 1$) so that the hypotheses do not have any pre-defined structure and the problem resembles a non-hierarchical multiple testing problem. We will refer to this configuration as the non-hierarchical configuration. Under this setting, all the hypotheses belong to the same family, $\mathcal{F}_1$, so that the hierarchical testing procedure reduces to the generalized stepup procedure with the same critical functions $r\alpha/m$, which is the BH procedure. Thus, Theorem 3.2 can be viewed as a generalization of the BH procedure in the context of testing hierarchically ordered hypotheses.

Now, we consider another special case where each family has exactly one hypothesis so that $\mathcal{F}_i = \{H_i\}, i = 1, \ldots, m$, and the hypotheses are all sequentially ordered where $\ell = 1$ and $m_i = m - i + 1$. We will refer to this configuration as the fixed sequence configuration. Under this configuration, the hierarchical testing procedure is equivalent to the fixed sequence method where hypothesis H_i is rejected if, and only if, hypotheses $H_1, \ldots, H_{i-1}$ have all been rejected and $P_i \leq m\alpha/(m - i + 1)$. This fixed sequence procedure matches the fixed sequence procedure from Theorem 2.1 in Chapter 2, which was shown to control the FDR at level α under arbitrary dependence. Thus, our result also generalizes this fixed sequence procedure to the testing of hierarchically ordered hypotheses.

Remarkably, Theorem 3.2 has connected two opposing testing methods: the testing of p-value ordered hypotheses (through the BH procedure) and the testing of pre-ordered hypotheses (through the fixed sequence procedure).

Finally, we consider a third configuration which we call the binary tree configuration. This configuration is helpful for evaluating the critical functions of the proposed procedures in the hierarchical setting and it is defined as follows. There is one hypothesis in $\mathcal{F}_1$ and each hypothesis has two child hypotheses except for the leaf hypotheses in $\mathcal{F}_D$. Hence, $\ell = 2^{D-1}$ and $m = 2^D - 1$. For each $d = 1, \ldots, D$, there are 2^{d-1} hypotheses in $\mathcal{F}_d$ and for each $H_i \in \mathcal{F}_d$, $\ell_i = 2^{D-d}$ and $m_i = 2^{D-d+1} - 1$. Under this configuration, the critical functions of the procedure introduced in Theorem 3.2 are, after simplification,

$$\alpha_i(r) = \frac{\alpha}{2^{d-1}} \left(1 + \frac{r-1}{2^{D-d+1}-1} \right), H_i \in \mathcal{F}_d, d = 1, \ldots, D, r = d, \ldots, 2^d - 1. \quad (3.8)$$

Compared to Meinshausen's FWER controlling hierarchical testing procedure, which is equivalent to the hierarchical testing procedure with critical functions $\ell_i \alpha / \ell, i = 1, \ldots, m$, the critical functions of Theorem 3.2 are $(m_i + r - 1)/m_i$ times larger for H_i. Table 3.1 lists the critical functions of Theorem 3.2 and Meinshausen's procedure for testing the hypotheses in Figure 3.1(b), which has the binary tree configuration. For family d, only the values of r between d and $\sum_{j=1}^{d} |\mathcal{F}_j|$ are listed in Table 3.1 (where $|\mathcal{F}_j|$ is the cardinality of $\mathcal{F}_j$) due to the fact that if a hypothesis in family d is rejected, then all d of its ancestor hypotheses are rejected so that $d \leq \sum_{j=1}^{d} R(\mathcal{F}_j) \leq \sum_{j=1}^{d} |\mathcal{F}_j|$.

Remark 3.2. It should be noted that the hierarchical testing procedure relies on the generalized stepup procedure to test each family; however, our proof of FDR control for Theorem 3.2 (and our proof of FDR control for the remaining procedures in this section) still holds if the generalized stepup-down procedure of any arbitrary order is used to test each family. Nevertheless, in practice we are generally trying to maximize the number of rejections subject to FDR control. Since the generalized stepup procedure is more powerful than the generalized stepup-down and generalized

Table 3.1 A Comparison of Rejection Thresholds for the Procedure in Theorem 3.2 and Meinshausen's Procedure when Testing the Hypotheses in Figure 3.1(b)

		Theorem 3.2							Meinshausen
		$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	
Family 1	$\alpha_i(r)$	α	-	-	-	-	-	-	α
Family 2	$\alpha_i(r)$	-	$2\alpha/3$	$5\alpha/6$	-	-	-	-	$\alpha/2$
Family 3	$\alpha_i(r)$	-	-	$3\alpha/4$	α	$5\alpha/4$	$3\alpha/2$	$7\alpha/4$	$\alpha/4$

stepdown procedures, we opted to use the generalized stepup procedure to test each family.

3.4.2 Procedure under Arbitrary Dependence

In this subsection we develop a FDR controlling hierarchical testing procedure under arbitrary dependence. Since arbitrary dependence is a more general type of joint dependence than positive dependence, it follows that the procedure under arbitrary dependence will not be quite as powerful as the procedure from Theorem 3.2.

Theorem 3.3. *FDR Control under Arbitrary Dependence*

The hierarchical testing procedure with critical functions

$$\alpha_i(r) = \frac{\ell_i \alpha}{\ell} \frac{m_i + r - 1}{m_i} \frac{1}{c_i} \quad \textit{where} \quad c_i = 1 + \sum_{j=d_i}^{|\mathcal{G}_{d_i}|-1} 1/(m_i + j)$$

and $|\mathcal{G}_{d_i}|$ *is the cardinality of* $\bigcup_{j=1}^{d_i} \mathcal{F}_j$*, strongly controls the FDR at level* α*.*

Just like Theorem 3.2, we consider the non-hierarchical configuration of hypotheses. In this special case, all of the critical functions of the procedure in

Theorem 3.3 are $r\alpha/(mc)$, where $c = \sum_{j=1}^{m} 1/j$, so that this procedure reduces to the stepup procedure with critical function $r\alpha/(mc)$, which is the BY procedure. Thus, this result extends the BY procedure to the context of testing hierarchically ordered hypotheses.

We also consider the fixed sequence configuration. Here, the rejection threshold for H_i is $(m - i + 1)/m\alpha$, which is the same as the procedure from Theorem 3.2 under this configuration. It should be noted that under the arbitrary dependence assumption, the procedure did not have smaller critical values than under the positive dependence assumption for this configuration. This is due to the fact that this fixed sequence procedure under arbitrary dependence cannot be improved with a positive dependence assumption (see Chapter 2, Theorem 2.1(b)).

It is easy to see that the critical functions of this procedure are scaled down compared with the procedure from Theorem 3.2, similar to the way the critical function of the BY procedure is scaled down compared with the BH procedure. Consider the example in Figure 3.1(b) which consists of 7 hypotheses. Here, $c_1 = 1, c_2 = c_3 = 1.2$, and $c_4 = c_5 = c_6 = c_7 = 1.76$, which means the critical functions of the procedure in Theorem 3.2 are as large, 1.2 times larger, and 1.76 times larger than the critical functions of the procedure in Theorem 3.3 for testing $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$, respectively. The critical function of the BH procedure, on the other hand, is $\sum_{i=1}^{7} 1/i = 2.59$ times larger than the critical function of the BY procedure for testing 7 hypotheses in the non-hierarchical setting. This holds in general, that the adjusting factors for the critical functions of the procedure in Theorem 3.3 are smaller in the hierarchical setting than in the non-hierarchical setting (i.e., the BY procedure). Thus, the procedure from Theorem 3.3 tends to be less affected by not having a positive dependence assumption in the hierarchical setting than in the non-hierarchical setting.

3.4.3 Procedures under Block Dependence

In this subsection, we consider block dependence and develop more powerful versions of the procedures in Theorems 3.2 and 3.3 by taking this dependence into account. Since block dependence only describes the dependence of the p-values across families, we consider both positive dependence and arbitrary dependence to describe the dependence of the p-values within the families, which we will refer to as block positive dependence and block arbitrary dependence, respectively.

In the fixed sequence configuration, block dependence reduces to independence. Under this configuration of hypotheses, both of our procedures presented in this subsection reduce to the more powerful FDR controlling fixed sequence procedure under independence, which is the procedure in Theorem 2.2 of Chapter 2, whereas the procedures in the last two subsections reduce to the less powerful FDR controlling fixed sequence procedure under arbitrary dependence, which is the procedure in Theorem 2.1 of Chapter 2.

First, we consider block positive dependence.

Theorem 3.4. *FDR Control under Block Positive Dependence*
Under Assumptions 1.1 and 3.1, the hierarchical testing procedure with critical functions

$$\alpha_i(r) = \begin{cases} \dfrac{\ell_i r\alpha}{\ell + \ell_i(r-1)\alpha} & \text{if } H_i \text{ is not a leaf hypothesis} \\ \dfrac{r\alpha}{\ell} & \text{if } H_i \text{ is a leaf hypothesis} \end{cases}$$

strongly controls the FDR at level α.

In the non-hierarchical configuration, this procedure reduces to the BH procedure since all the critical functions are $r\alpha/m$. It should be noted that under this configuration there is only one family so that block dependence is irrelevant and we are left with just the positive dependence assumption. Thus, both this procedure and

the procedure from Theorem 3.2, which both assume positive dependence, reduce to the BH procedure in the non-hierarchical configuration.

In the hierarchical setting, this procedure offers a large improvement over the critical functions of Theorem 3.2. To see this, consider the binary tree configuration. In this case, the critical functions are

$$\alpha_i(r) = \frac{r\alpha}{2^{d-1} + (r-1)\alpha}, H_i \in \mathcal{F}_d, d = 1, \ldots, D-1 \text{ and}$$
$$\alpha_i(r) = \frac{r\alpha}{2^{D-1}}, H_i \in \mathcal{F}_D. \tag{3.9}$$

Comparing (3.8) to (3.9), one can see that (3.9) is, in general, much larger than (3.8). For example, when $D = 5, d = 3$, and $r = 4$ the increase is by a factor of almost 3. Also, compared to Meinshausen's FWER controlling hierarchical testing procedure, which uses critical function $\ell_i\alpha/\ell, i = 1, \ldots, m$, the critical functions of Theorem 3.4 are approximately r times larger for small α. Hence, the procedure from Theorem 3.4, which requires the strongest dependence assumption to control the FDR, is our most powerful hierarchical testing procedure.

Finally, we consider block arbitrary dependence.

Theorem 3.5. *FDR Control under Block Arbitrary Dependence*

Under Assumption 3.1, the hierarchical testing procedure with critical functions

$$\alpha_i(r) = \begin{cases} \dfrac{\ell_i r\alpha}{\ell + \ell_i(r-1)\alpha}\dfrac{1}{c_i} & \text{if } H_i \text{ is not a leaf hypothesis} \\ \dfrac{r\alpha}{\ell}\dfrac{1}{c_i} & \text{if } H_i \text{ is a leaf hypothesis} \end{cases}$$

where

$$c_i = \begin{cases} 1 + \sum_{j=1}^{|\mathcal{F}_{d_i}|-1} \dfrac{\ell - \ell_i\alpha}{(j+d_i)(\ell + \ell_i(j+d_i-2)\alpha)} & \text{if } H_i \text{ is not a leaf hypothesis} \\ 1 + \sum_{j=1}^{|\mathcal{F}_{d_i}|-1} \dfrac{1}{j+d_i} & \text{if } H_i \text{ is a leaf hypothesis} \end{cases}$$

and $|\mathcal{F}_{d_i}|$ is the cardinality of $\mathcal{F}_{d_i}$, strongly controls the FDR at level α.

This procedure reduces to the BY procedure in the non-hierarchical configuration. Similar to the procedures from Theorems 3.2 and 3.3, the critical functions of this procedure are those of Theorem 3.4 adjusted by a factor $c_i, i = 1, \ldots, m$. Again, we consider the hypotheses in Figure 3.1(b). Here, $c_1 = 1, c_2 = c_3 = 1.317$, and $c_4 = c_5 = c_6 = c_7 = 1.760$ at $\alpha = 0.05$. In this example, the c_i's are significantly smaller than the adjusting factor for the BY procedure for testing 7 hypotheses, which is 2.59. However, c_2 and c_3 for Theorem 3.5 are larger than c_2 and c_3 for Theorem 3.3, which are both 1.2, but this is not true in general for the c_i's. The portion of the critical function without the adjusting factor, is generally much larger for Theorem 3.5 than for Theorem 3.3 so that the procedure from Theorem 3.5 is typically more powerful than the procedure from Theorem 3.3.

Remark 3.3. Our proofs of the theorems in this section heavily rely on the approach of mathematical induction. The hierarchical structure of the hypotheses implies a recursive property where the hypotheses in the subtree under any hypothesis also form a hierarchical structure. Hence, mathematical induction is a natural choice for proving results for hierarchical structures.

Below, we present an example of our hierarchical testing procedure along with Yekutieli's and Meinshausen's hierarchical testing procedures.

Example 3.3. Let us demonstrate how the hierarchical testing procedure works through an example using the critical functions of Theorem 3.2 as well as Yekutieli's and Meinshausen's procedure. Consider the example presented in Figure 1(b). The total depth of the tree is 3 and the 3 families are $\{H_1\}$, $\{H_2, H_3\}$, and $\{H_4, H_5, H_6, H_7\}$. Suppose the p-values are $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$ and the hypotheses are tested using the procedure from Theorem 3.2, Yekutieli's procedure, and Meinshausen's procedure at level $\alpha = 0.05$.

Table 3.2 shows the value of the variables step-by-step for the procedure from Theorem 3.2. The first family is tested using the generalized stepup procedure, and H_1, the only hypothesis in this family, is rejected. Now, $R(\mathcal{F}_1) = 1$. The second family is tested using the generalized stepup procedure with critical functions $\alpha_2^*(r) = \alpha_2(r+1)$ and $\alpha_3^*(r) = \alpha_3(r+1)$. H_3 can be rejected but H_2 cannot. Thus, $R(\mathcal{F}_2) = 1$. Finally, the third family is tested. Since H_2 was accepted and H_3 was rejected, we have $\alpha_4^*(r) = \alpha_5^*(r) = 0$, $\alpha_6^*(r) = \alpha_6(r+2)$, and $\alpha_7^*(r) = \alpha_7(r+2)$. Hypotheses H_6 and H_7 are rejected by the generalized stepup procedure.

Yekutieli's hierarchical testing procedure groups the hypotheses into families that share the same parent hypothesis so that the 4 families are $\{H_1\}$, $\{H_2, H_3\}$, $\{H_4, H_5\}$, and $\{H_6, H_7\}$. This procedure rejects hypotheses H_1 and H_3 (Table 3.3). Meinshausen's hierarchical testing procedure uses a fixed rejection threshold $\ell_i\alpha/\ell$ for testing H_i. This procedure rejects H_1 and H_3 (Table 3.4).

3.5 Simulation Study

We conducted a simulation study to evaluate the performance of the proposed procedures. Specifically, the simulation study compared the performance of our proposed procedures, which are labeled Procedures 3.2-3.5 corresponding to the procedures introduced in Theorems 3.2-3.5, against Yekutieli's FDR controlling procedure in terms of both FDR and average power. Several dependence configurations were considered as well as different hierarchical structures.

We generated m normal random variables with covariance matrix Σ and mean vector $\vec{\mu} = (\mu_1, \ldots, \mu_m)$ to test the m hypotheses $H_i : \mu_i \leq 0$ versus $H_i' : \mu_i > 0, i = 1, \ldots, m$. The p-value for testing the i^{th} leaf hypothesis was calculated using a one sided, one-sample Z-test. When H_i was true, we set $\mu_i = 0$. When H_i was

Table 3.2 The Procedure from Theorem 3.2 at Level $\alpha = 0.05$ to Hierarchically Test the Hypotheses Presented in Figure 3.1(b) with P-Values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$

Procedure 3.4		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	Outcome
Family 1									
generalized stepup	$\alpha_i^*(R)$	0.05	-	-	-	-	-	-	$R=1$
Reject H_1 and set $R(\mathcal{F}_1)=1$									
Family 2									
generalized stepup	$\alpha_i^*(R)$	-	0.033	0.033	-	-	-	-	$R=1$
Accept H_2, reject H_3 and set $R(\mathcal{F}_2)=1$									
Family 3									
generalized stepup	$\alpha_i^*(R)$	-	-	-	0	0	0.05	0.05	$R=2$
Accept H_4 and H_5 and reject H_6 and H_7. Set $R(\mathcal{F}_3)=2$									

Table 3.3 Yekutieli's Procedure at Level $\alpha = 0.05$ to Hierarchically Test the Hypotheses Presented in Figure 3.1(b) with P-Values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$

Yekutieli's Procedure		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	Outcome
Family 1									
BH procedure	$\alpha_i(R)$	0.0174	-	-	-	-	-	-	$R=1$
Reject H_1									
Family 2									
BH procedure	$\alpha_i(R)$	-	0.009	0.009	-	-	-	-	$R=1$
Accept H_2 and reject H_3									
Family 3									
Not Tested		-	-	-	-	-	-	-	
Accept H_4 and H_5									
Family 4									
BH procedure	$\alpha_i(R)$	-	-	-	-	-	0	0	$R=0$
Accept H_6 and H_7									

Table 3.4 Meinshausen's Procedure at Level $\alpha = 0.05$ to Hierarchically Test the Hypotheses Presented in Figure 3.1(b) with P-Values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$

Meinshausen's Procedure		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	Outcome
Family 1									
single-step	α_i	0.05	-	-	-	-	-	-	$R=1$
Reject H_1									
Family 2									
single-step	α_i	-	0.025	.025	-	-	-	-	$R=1$
Accept H_2 and reject H_3									
Family 3									
single-step	α_i	-	-	-	-	-	0.0125	0.0125	$R=0$
Accept H_4, H_5, H_6, and H_7									

false, we set μ_i to a positive value which was non-increasing in d_i. Our intention was to simulate the setting where hypotheses that are near the top of the hierarchy are easier to reject than hypotheses near the bottom. As for the joint dependence, we considered a common correlation structure where Σ had off-diagonal components equal to ρ and diagonal components equal to 1.

We constructed two types of hierarchies: a shallow hierarchy and a deep hierarchy. Both hierarchies had 1000 leaf hypotheses. A proportion π_0 of the leaf hypotheses were randomly chosen and set to be true with the remaining set to false. Each non-leaf hypothesis was set to true only if all of its child hypotheses were true; otherwise, it was set to false. For both hierarchies, the tree was balanced so that each parent hypothesis had the same number of child hypotheses. The two hierarchies are described in detail below.

Shallow Hierarchy: The total depth of this tree is 2 so that a hypothesis is either a leaf or a top-level hypothesis. There are 10 top-level hypotheses each of which have

100 child hypotheses giving a total of 1010 hypotheses. For each false hypothesis H_i, $\mu_i = 3$ if $d_i = 1$ and $\mu_i = 2$ if $d_i = 2$.

Deep Hierarchy: The total depth of this tree is 4 and there are 8 top-level parents. Each parent hypothesis has 5 child hypotheses giving a total of 1248 hypotheses. For each false hypothesis H_i, $\mu_i = 3.5$ if $d_i = 1$, $\mu_i = 3$ if $d_i = 2$ or 3, and $\mu_i = 2$ if $d_i = 4$.

We set $\alpha = 0.05$ and for each procedure we noted the false discovery proportion, which is the proportion of falsely rejected hypotheses among all rejected hypotheses, and the the proportion of rejected false null hypotheses among all false null hypotheses. Each tree was generated and tested 5000 times and the simulated values of the FDR and average power were obtained by averaging out the 5000 values of these two proportions, respectively.

Figure 3.2 displays the FDR and average power under independence as π_0 varies from 0.2 to 1. As seen from Figure 3.2, all the procedures control the FDR at level 0.05. In terms of power, Procedure 3.4 outperforms Yekutieli's procedure quite substantially and in some cases even doubles the power of Yekutieli's procedure. Procedure 3.2, which controls the FDR under positive dependence, outperforms Yekutieli's procedure under the shallow hierarchy but is outperformed by Yekutieli's procedure in the deep hierarchy. In the deep hierarchy, Procedure 3.5 and Yekutieli's procedure are comparable in terms of power. Not surprisingly, Procedure 3.3, which controls the FDR under arbitrary dependence, performs the worst.

Figures 3.3 and 3.4 display the FDR and average power under common correlation with $\rho = 0.25$ and $\rho = 0.75$, respectively, as π_0 varies from 0.2 to 1. As seen from these figures, the FDRs of all the procedures are controlled at level 0.05 under both weak and strong correlation. It should be noted that Assumption 3.1 (block dependence) does not hold under this dependence configuration, but Procedures 3.4 and 3.5 still control the FDR suggesting that both procedures are fairly robust to

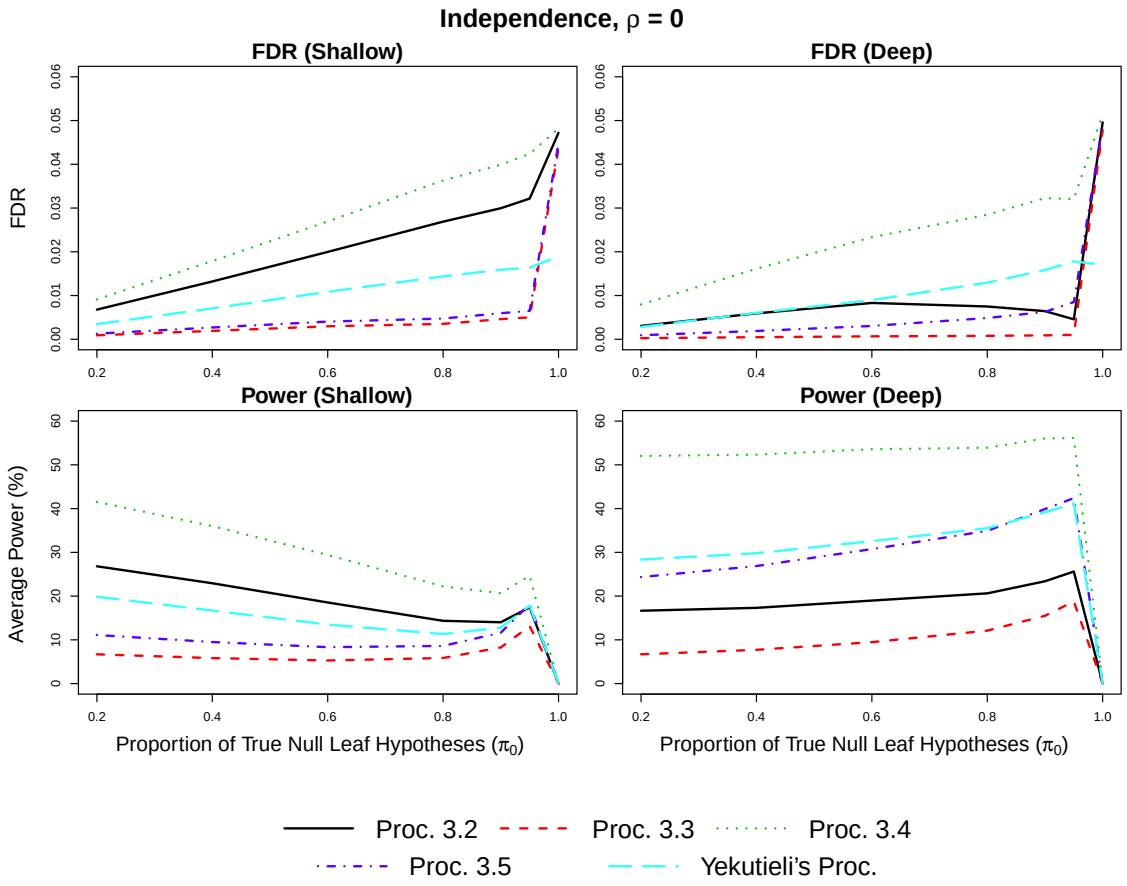


Figure 3.2 FDR (top row) and average power (bottom row) of Procedures 3.2 (solid line), 3.3 (dashed), 3.4 (dotted), 3.5 (dot dash), and Yekutieli's procedure (long dash) under independence for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1.

departures from this assumption. In terms of power, these figures show a similar pattern to Figure 3.2, where Procedure 3.4 is the most powerful and Procedure 3.3 is the least powerful. The remaining three procedures fall somewhere in the middle depending on the setting.

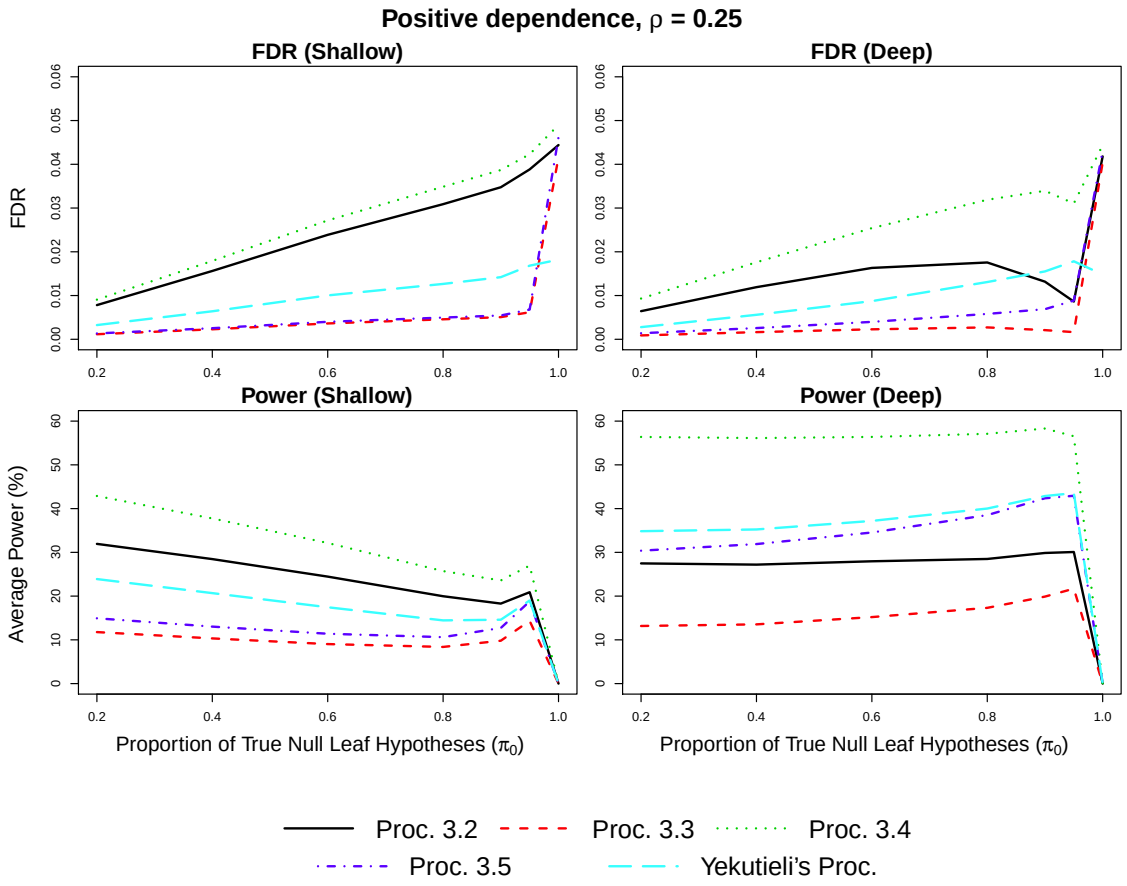


Figure 3.3 FDR (top row) and average power (bottom row) of Procedures 3.2 (solid line), 3.3 (dashed), 3.4 (dotted), 3.5 (dot dash), and Yekutieli's procedure (long dash) under common correlation with $\rho = 0.25$ for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1.

3.6 Real Data Analysis

We applied our proposed procedures as well as Yekutieli's procedure to a real data set. We used the data set of Caporaso et al. (2011), available in the phyloseq Bioconductor package at www.bioconductor.org, which provides the abundances of individual microbes in different ecological environments as well as their phylogenetic relationships. The data can be naturally organized into a hierarchy consisting of taxonomic units according to their phylogenetic relationships. The question of interest is whether there is an association between a taxonomic unit and ecological environment. Specifically, we tested the null hypothesis that the mean abundance for

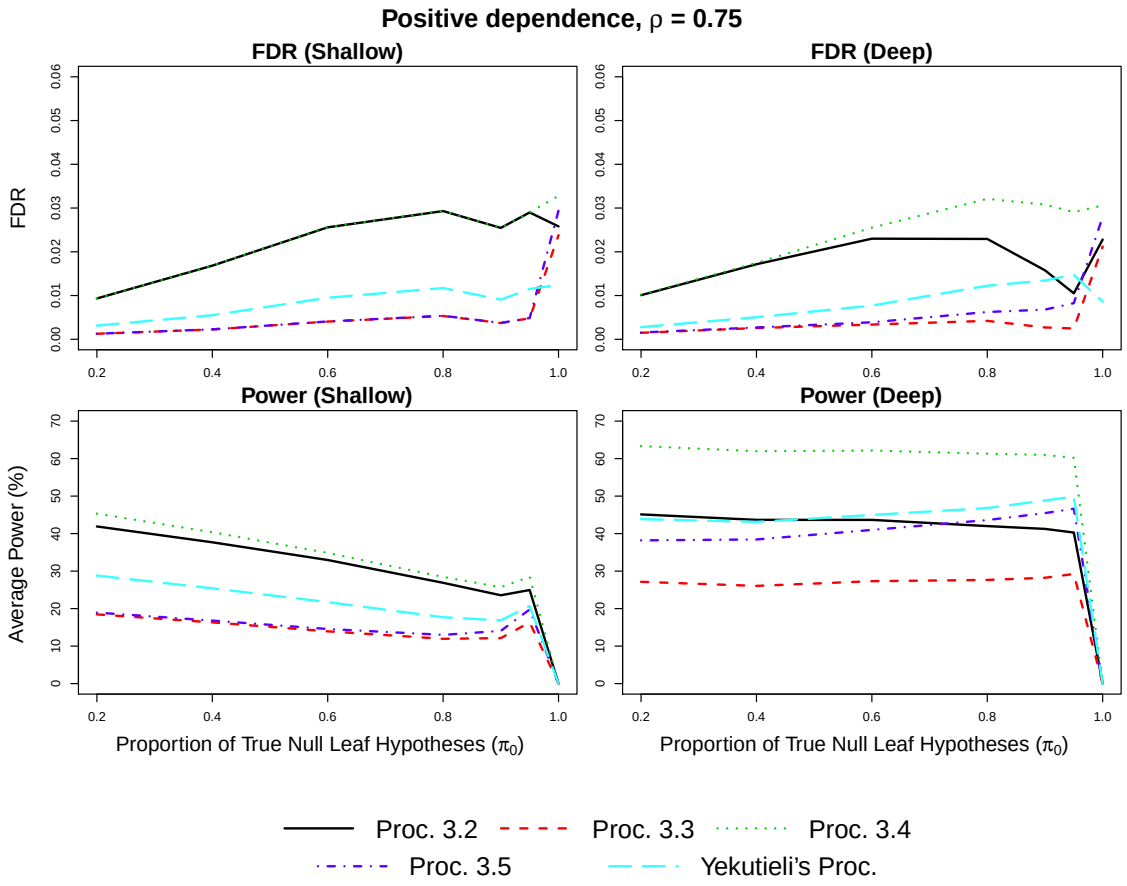


Figure 3.4 FDR (top row) and average power (bottom row) of Procedures 3.2 (solid line), 3.3 (dashed), 3.4 (dotted), 3.5 (dot dash), and Yekutieli's procedure (long dash) under common correlation with $\rho = 0.75$ for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1.

the taxonomic unit is the same across environments versus the alternative hypotheses that the mean abundance for the taxonomic unit is different across environments. The p-value for each hypothesis was determined by using an F-test (for more information see Sankaran and Holmes (2013)).

We restricted our analysis to the microbes in the Actinobacteria phylum which had 1631 individual microbes. The taxonomic hierarchy in the Actinobacteria phylum consisted of 3261 taxonomic units so that the total number of hypotheses is 3261 across 39 families. We tested the hypotheses at various significance levels and the number of rejections for each procedure are displayed in Table 3.5. All of

the procedures are seen to make a substantial number of discoveries, even when $\alpha = 0.01$. In terms of the number of rejections, one can easily see that Procedure 3.4 is by far the best, significantly outperforming the other procedures. Procedure 3.5 outperforms Yekutieli's procedure when α is moderate to large, but Yekutieli's procedure outperforms Procedure 3.5 when α is small. Procedure 3.3 is, not surprisingly, the worst since it is the only procedure that controls the FDR despite not requiring any assumption on the dependence structure.

Table 3.5 The Number of Rejections out of 3261 Hypotheses by Procedures 3.2, 3.3, 3.4, 3.5, and Yekutieli's Procedure at Various Significance Levels for the Microbe Abundance Data Set Restricted to the Actinobacteria Phylum

α	**Procedure 3.2**	**Procedure 3.3**	**Procedure 3.4**	**Procedure 3.5**	**Yekutieli's Procedure**
0.01	75	68	144	107	123
0.025	88	75	574	148	165
0.05	118	92	1156	353	230
0.1	138	108	1497	813	253

3.7 Conclusion

In this chapter, we have developed several FDR controlling procedures for testing hierarchically ordered hypotheses. We have, for the first time, presented hierarchical testing methods with proven FDR control under various dependencies. Specifically, we have developed a method which controls the FDR under block positive dependence, and in our simulation study, it was shown to be more powerful than Yekutieli's FDR controlling procedure. A particularly interesting aspect of this work is that we have connected two contrasting testing methods: fixed sequence procedures, which assume

the hypotheses have a fixed pre-defined testing order, and stepwise procedures, in which the hypotheses are ordered based on the corresponding p-values.

CHAPTER 4

THE TESTING OF HYPOTHESES ALONG A DIRECTED ACYCLIC GRAPH

4.1 Introduction

In this chapter, we propose a novel approach for developing procedures which control type I error rates appropriate for large scale multiple testing, such as the FDR, which we will refer to as large scale error rates. We then use this approach to develop a procedure which controls the FDR for testing hypotheses with a directed acyclic graph (DAG) structure. Testing on a complex structure such as a DAG has applications to testing Gene Ontology terms for phenotype association (Goeman and Mansmann 2008) and clinical trials (Dmitrienko et al. 2007; Dmitrienko and Tamhane 2013). Moreover, a DAG structure is a very generic structure and can take the form of a hierarchy, as discussed in Chapter 3, or a fixed sequence, as discussed in Chapter 2.

With regard to procedures that test hypotheses along a DAG, there has been some work in the framework of FWER control. Goeman and Mansmann (2008) proposed a FWER controlling procedure for testing Gene Ontology terms called the focus level method which preserves the graph structure of Gene Ontology. The procedure is based on the closed testing procedure and requires the user to select a pre-specified subset of the Gene Ontology terms the user is most interested in. Dmitrienko et al. (2007) and Dmitrienko and Tamhane (2013) proposed a general FWER controlling method for testing hypotheses with applications to clinical trials. The hypotheses are organized into families and the families are tested sequentially. A decision regarding whether a hypothesis can be tested is made based on which hypotheses are rejected in the previously tested families. We are motivated to study

the DAG structure since it is a very general structure and to our knowledge, no procedures exist which control a large scale error rate, such as the FDR, which is more appropriate for large scale multiple testing.

To assist in the development of our new DAG testing procedure, we propose a new approach for developing multiple testing methods which control large scale error rates. This approach splits a large scale error rate of the form $E(WL)$ into two components: a weight, W, and another error rate, L. For example, the FDR $= E(V/(R \vee 1))$ can be split into a weight $W = 1/(R \vee 1)$ and an error rate $L = V$. Our main idea is to handle these two components separately. We do this by creating or using an existing procedure which controls $E(L)$, then determining the weight W, and finally by applying the procedure at level α/W, thereby controlling $E(WL)$ at level α. This approach greatly facilitates the development of new multiple testing procedures and we will show that by using this method we can obtain new procedures as well as several existing procedures, such as the BH and adaptive BH procedures.

By using our proposed approach, we create a new FDR controlling procedure for testing hypotheses which have a DAG structure. Specifically, our procedure only rejects a hypothesis if all of its parent hypotheses are rejected. The advantages of accounting for the underlying structure of the hypotheses are two-fold. First, there is a potential gain in power in the testing of these hypotheses. Indeed, our simulation study and real data analysis show that our proposed procedure compares favorably against the BH procedure in terms of power by accounting for the underlying DAG structure. Second, the rejected hypotheses maintain their hierarchical integrity which can enhance interpretation. That is, the rejected hypotheses form a DAG that is a subset of the tested DAG. With regard to testing hypotheses along a DAG, we also show that the BH procedure can be slightly modified so that the rejected hypotheses maintain their hierarchical integrity. This modification reduces the power of the

usual BH procedure but benefits from the fact that the rejection set preserves the DAG structure.

This rest of this chapter is organized as follows. Section 4.2 introduces notation and concepts that will be used in this chapter. Section 4.3 describes our new approach for developing multiple testing procedures. In Section 4.4, we propose a new DAG testing procedure controlling the FDR. Our simulation study and real data analysis are in Section 4.5 and 4.6, respectively. Finally, Section 4.7 offers concluding remarks.

4.2 Preliminaries

Recall that there are m_0 true null hypotheses among the m tested hypotheses and for a testing procedure, V and R are the number of false and total rejections, respectively. For a testing procedure applied at level α, let $L(\alpha)$ be some measure of the overall type I error after testing all the m hypotheses, and we will assume throughout this chapter that R is a non-decreasing function of α. With regard to notation, we will often express $L(\alpha)$ simply as L when the significance level has not been specified. Typically, we will set $L = V$ so that L is the total number of false rejections; however, we will also consider, with less emphasis, other error measures such as $L = VI\{V \geq k\}$. We assume we have available to us some procedure which controls $E(L)$ at some fixed, pre-specified level. We will call this procedure the base procedure. For example, if $L = V$, then any PFER controlling procedure could be a base procedure.

Since we are primarily interested in the case when $L = V$, let us review some PFER controlling procedures. We assume each of the procedures are applied at the fixed, pre-specified level α.

Example 4.1. *The Bonferroni Procedure.* For each $i = 1, \ldots, m$, the Bonferroni procedure rejects H_i if $P_i \leq \alpha/m$. The Bonferroni procedure controls the PFER under arbitrary dependence. A variant of the Bonferroni procedure is the weighted Bonferroni procedure. Given weights $w_1, \ldots, w_m$ such that $\sum_{i=1}^{m} w_i = 1$, the weighted Bonferroni procedure rejects H_i if $P_i \leq w_i\alpha$. Another variant is the oracle Bonferroni procedure, which rejects H_i if $P_i \leq \alpha/m_0$. The oracle Bonferroni procedure controls the PFER at level α under arbitrary dependence but m_0 must be known.

Example 4.2. *The Adaptive Bonferroni Procedure.* For each $i = 1, \ldots, m$, the adaptive Bonferroni procedure rejects H_i if $P_i \leq \alpha/\widehat{m}_0$ where $\widehat{m}_0$ is a conservative estimate of m_0 such that $E(\widehat{m}_0) \leq m_0$. One estimate for m_0 is $\widehat{m}_0 = (\sum_{i=1}^{m} I\{P_i > \lambda\}+1)/(1-\lambda)$ where $\lambda \in (0, 1)$ is a tuning parameter. The adaptive Bonferroni procedure controls the PFER under independence.

In this chapter, we will describe a general approach to control a two-factor error rate where one factor is a weight and the other is an error rate. Given an error rate L and a non-increasing function of the number of rejections $W : \{0, \ldots, m\} \to [0, \infty)$, the overall error rate that we are interested in is

$$E(W(R)L). \tag{4.1}$$

Let us consider some examples. Let $L = V$. If $W(R) = 1/(R \vee 1)$, then (4.1) is the FDR. We will give particular attention to this case of $L = V$ throughout this chapter. If $W(R) = 1$, then (4.1) is simply the PFER across all the hypotheses.

We are interested in controlling (4.1) at a fixed, pre-specified level α. We do this by determining some significance level that may be a function of the p-values, say β, such that when the base procedure is applied at level β, (4.1) is controlled at level α. Before proceeding, let us make our terminology clear. When we say that a procedure controls $E(L)$ at level β, we mean that $E(L) \leq \beta$. When we say a

procedure is applied at level β, we mean that the procedure rejects hypotheses based on parameter β, where β could be a function of the p-values. If, however, β is a fixed, pre-specified value, then applying the base procedure at level β also has the effect of controlling $E\left(L(\beta)\right)$ at level β. In general, we will use the convention that α is a fixed, pre-specified significance level and β is a significance level that may be random (i.e., it may depend on the p-values).

4.3 A General Method for Developing New Procedures

In this section, we present a testing procedure, which we call the self-consistent rejection (SCR) procedure, and develop the theoretical conditions under which it controls overall error rate (4.1). We motivate the development of this procedure by studying the relationship between the Bonferroni procedure and the BH procedure in the example below.

Example 4.3. Suppose $W = 1/(R \vee 1)$ and $L = V$ so that (4.1) becomes the FDR. We consider the Bonferroni procedure as our base procedure since this procedure controls $E\left(L\right)$. Recall that the BH procedure is the stepup procedure with critical constants $i\alpha/m, i = 1, \ldots, m$. The number of rejections by the BH procedure is $R = \max\{0 \leq r \leq m : P_{(r)} \leq r\alpha/m\}$ where $P_{(0)} \equiv 0$. Let $R_{base}(\beta) = \sum_{i=1}^{m} I\{P_i \leq \beta/m\}$ be the number of rejections by the Bonferroni procedure at level β. Since the event $\{P_{(r)} \leq r\alpha/m\}$ is equivalent to the event $\{r \leq R_{base}(r\alpha)\}$, we have that the number of rejections by the BH procedure can equivalently be expressed as

$$R = \max\{0 \leq r \leq m : r \leq R_{base}(r\alpha)\}. \tag{4.2}$$

Thus, the Bonferroni applied at level $R\alpha$ is equivalent to the BH procedure applied at level α since both procedures reject each H_i where $P_i \leq R\alpha/m, i = 1, \ldots, m$. In this example, we have shown that the FDR controlling procedure, the BH procedure, can be characterized by the PFER controlling procedure, the Bonferroni procedure.

Notice that expression (4.2) does not determine the number of rejections based on the p-values directly, but instead determines the number of rejections through the base procedure, which is the Bonferroni procedure. Hence, we can readily generalize this expression to any other base procedure. Given a non-increasing function of the number of rejections $W : \{0, \ldots, m\} \to [0, \infty)$ and a base procedure, the SCR procedure is as follows.

Definition 4.1. The Self-Consistent Rejection Procedure
Given a base procedure, let $R_{base}(\beta)$ be the number of rejections by the base procedure at level β.

1. *Let*

$$R = \max \{r \in \{0, \ldots, m\} : r \leq R_{base}(\alpha/W(r))\}. \tag{4.3}$$

2. *Test the hypotheses $H_1, \ldots, H_m$ by applying the base procedure at level $\alpha/W(R)$.*

Remark 4.1. It should be noted that R in (4.3) is equal to the number of rejections by the base procedure at level $\alpha/W(R)$. That is, $R = R_{base}(\alpha/W(R))$. In the literature, this property has been referred to as self-consistency, see Blanchard and Roquain (2008). To see why this property holds, we note that R is the largest integer r in $\{0, \ldots, m\}$ that is less than or equal to $R_{base}(\alpha/W(r))$. Hence, $R + 1 > R_{base}(\alpha/W(R + 1))$, which implies $R \geq R_{base}(\alpha/W(R + 1)) \geq R_{base}(\alpha/W(R))$ where the latter inequality follows from the fact that $R_{base}(\alpha/W(r))$ is a non-decreasing function of r. Thus, $R \leq R_{base}(\alpha/W(R)) \leq R$ so that $R = R_{base}(\alpha/W(R))$.

Based on the discussion in example 4.3, it is easy to see that if the $W = 1/(R \vee 1)$ and the base procedure is the Bonferroni procedure, then (4.3) reduces to (4.2). In fact, the SCR procedure reduces to a stepup procedure when the base procedure is any single-step procedure, not just the Bonferroni procedure.

Proposition 4.1. *If the base procedure is a single-step procedure whose critical constant is the non-decreasing function $t : \mathbb{R}^+ \to \mathbb{R}^+$ where H_i is rejected at level α if $P_i \leq t(\alpha)$ for each $i = 1, \ldots, m$, then the SCR procedure using this base procedure is equivalent to the stepup procedure with critical constants $t(\alpha/W(1)), \ldots, t(\alpha/W(m))$.*

Proof. For $\alpha > 0$, the event $\{r \leq R_{base}(\alpha/W(r))\}$ is equivalent to the event $\{P_{(r)} \leq t(\alpha)\}$. Hence, (4.3) reduces to $R = \max\left\{r \in \{0, \ldots, m\} : P_{(r)} \leq t(\alpha/W(r))\right\}$ so that the SCR procedure rejects each H_i where $P_i \leq t(\alpha/W(R)), i = 1, \ldots, m$, which is the same as the stepup procedure with critical constants $t(\alpha/W(1)), \ldots, t(\alpha/W(m))$. ■

Before discussing how to control the overall error rate (4.1), we will discuss two cases for the error rate L: the binary case and the non-binary case. In the binary case, L can only take on 0 and a positive value. In the non-binary case, L can take on any number of values; however, we will focus on a specific case where L is a finite sum of binary error rates. For example, if $L = I\{V > 0\}$, so that $E(L)$ is the FWER, then L is binary, but if $L = V$, so that $E(L)$ is the PFER, then L is non-binary.

Given a binary error rate L, positive dependence among the p-values is characterized by the following assumption.

Assumption 4.1. Positive Dependence Assumption

Given a binary error rate L and a base procedure, for any coordinatewise non-decreasing function of the p-values ψ,

$$E(\psi(P_1, \ldots, P_m) \mid L(\alpha) > 0) \text{ is non-decreasing in } \alpha. \tag{4.4}$$

It should be noted that if $L(\alpha) = I\{P_i \leq \alpha\}$ where H_i is true, then assumption 4.1 reduces to Assumption 1.1, which is a commonly used assumption in multiple testing problems (Benjamini and Yekutieli 2001). If the p-values are independent, then we have the following result.

Lemma 4.1. *If the p-values are independent and the binary error rate L is of the form $L(\alpha) = I\{\bigcap_{i=1}^{m} P_i \leq t_i(\alpha)\}$ for non-decreasing $t_i : \mathbb{R}^+ \to \mathbb{R}^+, i = 1, \ldots, m$, then assumption 4.1 is satisfied.*

Proof. The proof is in Appendix C. ■

For a non-binary error rate L and the associated base procedure, we will assume the following conditions are satisfied:

Condition A1: L is no larger than the finite sum of $n > 0$ binary error rates, $L_1, \ldots, L_n$.

Condition A2: For each $i = 1, \ldots, n$, $E(L_i(\alpha)) \leq c_i\alpha$ for every fixed $\alpha > 0$, where $c_1, \ldots, c_n$ are constants such that $\sum_{i=1}^{n} c_i \leq 1$.

Condition A3: Assumption 4.1 is satisfied for $L_i, i = 1, \ldots, n$.

Remark 4.2. We can regard condition A1 as a property of the error rate, condition A2 as a property of the base procedure, and condition A3 as a property of the joint dependence of the p-values. Secondly, if conditions A1 and A2 are satisfied, then it must be the case that $E(L(\alpha)) \leq \alpha$. Indeed, $E(L(\alpha)) = \sum_{i=1}^{n} E(L_i(\alpha)) \leq \sum_{i=1}^{n} c_i\alpha \leq \alpha$.

For example, consider the non-binary error rate $L = V$ with the Bonferroni procedure, which controls $E(V)$, as the base procedure. It is easy to see that $L(\alpha)$ can be broken into m binary error rates, $L_i(\alpha) = I\{P_i \leq \alpha/m, H_i \text{ is true}\}, i = 1, \ldots, m$.

Here, $c_i = 1/m$ and $\sum_{i=1}^{m} c_i = 1$. Hence, the error rate V and the Bonferroni procedure satisfy conditions A1 and A2. If the p-values have positive dependence as described by Assumption 1.1 in Chapter 1, then condition A3 also holds.

Theorem 4.2. *Given L and a base procedure if conditions A1-A3 are satisfied, then the SCR procedure along with the base procedure controls the overall error rate, $E\left(W(R)L\right)$, at level α.*

Proof of Theorem 4.2. By condition A1, $L_1, \ldots, L_n$ are the binary error rates whose sum dominates L and without loss of generality, we assume L_i only takes on 0 or 1 for each $i = 1, \ldots, n$. Based on remark 4.1, we have that the base procedure is applied at level $\alpha/W(R)$, where R is the number of rejections as determined by (4.3). Hence, for each $i = 1, \ldots, n$,

$$
\begin{aligned}
&E\left(W(R)L_i(\alpha/W(R))\right) = \sum_{r=1}^{m} W(r)\Pr\left(R = r, L_i(\alpha/W(r)) > 0\right) \\
\leq &\sum_{r=1}^{m} c_i\alpha\Pr\left(R = r \mid L_i(\alpha/W(r)) > 0\right) \\
= &\sum_{r=1}^{m} c_i\alpha\Pr\left(R \geq r \mid L_i(\alpha/W(r)) > 0\right) - \sum_{r=1}^{m-1} c_i\alpha\Pr\left(R \geq r+1 \mid L_i(\alpha/W(r)) > 0\right) \\
\leq &c_i\alpha + \sum_{r=1}^{m-1} c_i\alpha\left[\Pr\left(R \geq r+1 \mid L_i(\alpha/W(r+1)) > 0\right) - \right. \\
&\qquad\qquad \left. \Pr\left(R \geq r+1 \mid L_i(\alpha/W(r)) > 0\right)\right] \\
\leq &c_i\alpha.
\end{aligned}
$$

The first equality follows by the fact that L_i only takes on 0 or 1 so that $E\left(L_i\right) = \Pr\left(L_i > 0\right)$. The first inequality follows by condition A2 where $\Pr\left(L_i(\alpha/W(r)) > 0\right) = E\left(L_i(\alpha/W(r)\right) \leq c_i\alpha/W(r)$. The last inequality follows by condition A3 along with the fact that $\alpha/W(r+1) \geq \alpha/W(r)$.

Finally, we have

$$E\left[W(R)L(\alpha/W(R))\right] \leq \sum_{i=1}^{n} E\left[W(R)L_i(\alpha/W(R))\right] \leq \sum_{i=1}^{n} c_i\alpha \leq \alpha. \blacksquare$$

Let us consider two examples to show why Theorem 4.2 is useful.

Example 4.4. As an example, let us show how Theorem 4.2 can be used to prove an adaptive BH procedure controls the FDR under independence. Let $L = V$ and $W = 1/(R \vee 1)$. For the base procedure, consider the procedure described in example 4.2, the adaptive Bonferroni procedure. Recall that this procedure rejects H_i if $P_i \leq \alpha/\widehat{m}_0$ for $\widehat{m}_0 = (\sum_{i=1}^{m} I\{P_i > \lambda\} + 1)/(1 - \lambda)$. Let I_0 and m_0 be the index set and number of true null hypotheses, respectively. Clearly, V can be split into m_0 binary error rates $L_i = I\{P_i \leq \alpha/\widehat{m}_0\}, i \in I_0$, so that A1 is satisfied. Secondly, it can be shown under independence that $E\left(L_i(\alpha)\right) \leq \alpha/m_0$. Setting $c_i = 1/m_0$, we see that condition A2 is satisfied. Finally, for each $i \in I_0$, fix P_j for $j \neq i$ and define the function $t_i : \mathbb{R}^+ \to \mathbb{R}^+$ such that $t_i(\alpha) = \sup\{P_i : P_i \leq \alpha/\widehat{m}_0\}$. Since $\widehat{m}_0 P_i$ is increasing in P_i, t_i is an increasing function of α. Now, the event $\{L_i(\alpha) > 0\} = \{P_i \leq t_i(\alpha)\}$ so that conditional on every fixed P_j for $j \neq i$, Assumption 4.1 holds. Hence, Assumption 4.1 holds unconditionally and condition A3 is satisfied. Thus, according to Theorem 4.2, the SCR procedure using the adaptive Bonferroni procedure controls the FDR at level α under independence. By Proposition 4.1, with $\widehat{m}_0$ as the estimate of m_0, the SCR procedure is actually an adaptive BH procedure (see Storey et al. (2004), Benjamini et al. (2006), Sarkar (2008)). If other base procedures, such as the Bonferroni procedure, the oracle Bonferroni procedure, and the weighted Bonferroni, are used, the corresponding SCR procedures reduce to the BH procedure, the oracle BH procedure, and the weighted BH procedure (Genovese et al. 2006), respectively.

Example 4.5. The SCR procedure along with Theorem 4.2 can also be used to help create new procedures for controlling complex error measures other than the

FDR. For example, consider the generalized false discovery rate, k-FDR, which is $E\left(VI\{V \geq k\}/(R \vee k)\right)$ for some pre-specified $1 \leq k \leq m$ (see Sarkar (2007)). Here, let $L = VI\{V \geq k\}$ and $W = 1/(R \vee k)$. First, we will describe a base procedure for controlling $E(L)$, then we will show that conditions A1-A3 are satisfied under independence. The base procedure is the single-step procedure with critical constant $t(\alpha) = [(k-1)!\alpha/\binom{m}{k}]^{1/k}$. To show conditions A1-A3 are satisfied for such L and base procedure, let $\mathcal{C}_i^0 = \{\mathcal{J} : \mathcal{J} \subseteq \{1, \ldots, m\}$ such that $|\mathcal{J}| = i$ and H_j is true for each $j \in \mathcal{J}\}$. Let $L_{\mathcal{J}}(\alpha) = I\{\max_{j\in\mathcal{J}} P_j \leq t(\alpha)\}/(k-1)!$ and define $c_{\mathcal{J}} = 1/\binom{m}{k}, \mathcal{J} \in \mathcal{C}_k^0$. Then,

$$
\begin{aligned}
L = VI\{V \geq k\} &= \sum_{H_i \text{ is true}} I\{P_i \leq t(\alpha), V \geq k\} \\
&= \sum_{\mathcal{J}\in\mathcal{C}_1^0} I\{\max_{j\in\mathcal{J}} P_j \leq t(\alpha), V^{(-\mathcal{J})} \geq k-1\} \\
&\leq \sum_{\mathcal{J}\in\mathcal{C}_1^0} \frac{V^{(-\mathcal{J})}}{k-1} I\{\max_{j\in\mathcal{J}} P_j \leq t(\alpha), V^{(-\mathcal{J})} \geq k-1\} \\
&= \sum_{\mathcal{J}\in\mathcal{C}_1^0} \sum_{\substack{H_i \text{ is true} \\ i\notin\mathcal{J}}} \frac{1}{k-1} I\{\max_{j\in\mathcal{J}} P_j \leq t(\alpha), P_i \leq t(\alpha), V^{(-\mathcal{J})} \geq k-1\} \\
&= \sum_{\mathcal{J}\in\mathcal{C}_2^0} \frac{1}{k-1} I\{\max_{j\in\mathcal{J}} P_j \leq t(\alpha), V^{(-\mathcal{J})} \geq k-2\} \ldots \\
&\leq \sum_{\mathcal{J}\in\mathcal{C}_k^0} \frac{1}{(k-1)!} I\{\max_{j\in\mathcal{J}} P_j \leq t(\alpha)\} \leq \sum_{\mathcal{J}\in\mathcal{C}_k^0} L_{\mathcal{J}}(\alpha)
\end{aligned}
$$

where $V^{(-\mathcal{J})}$ is the number of false rejections excluding H_j for $j \in \mathcal{J}$. The first inequality follows by the fact that the event $\{V^{(-\mathcal{J})} \geq k-1\}$ implies $V^{(-\mathcal{J})}/(k-1) \geq 1$. The last equality follows from the fact that $\mathcal{C}_2^0 = \{\{\mathcal{J} \cup H_i\} : \mathcal{J} \in \mathcal{C}_1^0, H_i \text{ is true, and } i \notin \mathcal{J}\}$. Thus, $VI\{V \geq k\}$ is dominated by the sum of binary error rates $L_{\mathcal{J}}, \mathcal{J} \in C_k^0$ and condition A1 is satisfied. We have

$$
E\left(L_{\mathcal{J}}(\alpha)\right) = \frac{1}{(k-1)!} \Pr\left(\max_{j\in\mathcal{J}} P_j \leq t(\alpha)\right) = \frac{1}{(k-1)!} \prod_{j\in\mathcal{J}} \Pr\left(P_j \leq t(\alpha)\right)
$$

$$\leq \frac{1}{(k-1)!} t^k(\alpha) = c_{\mathcal{J}}\alpha,$$

and $|\mathcal{C}_k^0| = \binom{m_0}{k}$ so that $\sum_{\mathcal{J}\in\mathcal{C}_k^0} c_{\mathcal{J}} = \binom{m_0}{k}/\binom{m}{k} \leq 1$. Condition A2 is satisfied. Finally, by Lemma 4.1, condition A3 is satisfied under independence.

Thus, this single-step procedure satisfies conditions A1-A3 under independence. By Theorem 4.2, the corresponding SCR procedure controls the k-FDR using this base procedure. Since the base procedure is a single-step procedure, by Proposition 4.1 the SCR procedure is equivalent to the stepup procedure with critical constants $[(k-1)!\max(k,i)\alpha/\binom{m}{k}]^{1/k}, i = 1, \ldots, m$.

4.4 Testing on a Directed Acyclic Graph

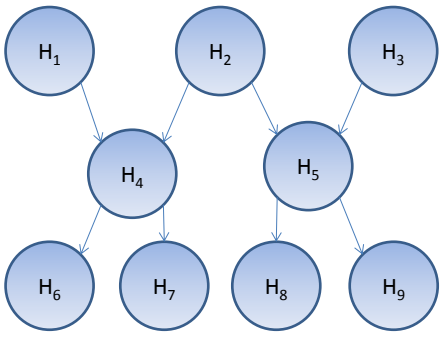


Figure 4.1 An example of a directed acyclic graph structure with 9 hypotheses. H_4 is only tested if H_1 and H_2 are rejected, H_5 is only tested if H_2 and H_3 are rejected, H_6 and H_7 are only tested if H_4 is rejected, and H_8 and H_9 are only tested if H_5 is rejected.

In this section, we consider testing hypotheses organized in a directed acyclic graph (DAG) structure (see Figure 4.1). Such graphical modeling of multiple testing has broad applications, such as the testing of terms in a Gene Ontology. A graph is a set of vertices connected by links called edges, and a directed graph implies the

edges are directed, meaning they have an associated direction. A DAG is a special type of directed graph which does not contain any cycles. That is, starting from a vertex there is no way to follow a sequence of directed edges to get back to this vertex. In the context of multiple testing along a DAG, the vertices of the graph represent the hypotheses to be tested and the edges represent the relationships between these hypotheses. We will refer to the hypotheses associated with the incoming edges of H_i as the parents of H_i, and similarly, we will refer to the hypotheses associated with the outgoing edges of H_i as the children of H_i.

For our testing approach, we restrict the testing of hypotheses to those hypotheses whose parents have all been rejected. We will create a base procedure controlling the PFER to test hypotheses in this manner, and with the help of Theorem 4.2, we will prove the SCR procedure with this base procedure controls the FDR under independence.

Let $\mathcal{T}_i$ be the set of parent hypotheses of H_i so that H_i is tested if each $H_j \in \mathcal{T}_i$ is rejected. Now, define $\mathcal{T}_i^{(0)} = \{H_i\}$ and $\mathcal{T}_i^{(k)} = \bigcup_{H_j \in \mathcal{T}_i^{(k-1)}} \mathcal{T}_j, k = 1, \ldots, m$. Then, the set of ancestor hypotheses of H_i is $\mathcal{D}_i = \bigcup_{k=0}^{m} \mathcal{T}_i^{(k)}$. It follows that H_i is rejected if, and only if, every hypothesis in $\mathcal{D}_i$ is rejected. Conversely, define $\mathcal{M}_i = \{H_j : H_i \in \mathcal{D}_j, j = 1, \ldots, m\}$ so that $\mathcal{M}_i$ is the set of descendants hypotheses of H_i. We have the following relationship: $H_j \in \mathcal{D}_i$ if, and only if $H_i \in \mathcal{M}_j$. Consider the example, in Figure 4.1. Here, $\mathcal{T}_4 = \{H_1, H_2\}$, $\mathcal{D}_6 = \{H_1, H_2, H_4, H_6\}$, and $\mathcal{M}_3 = \{H_3, H_5, H_8, H_9\}$. We will refer to a leaf hypothesis as a hypothesis that is not a parent of any other hypothesis. Mathematically, H_i is a leaf hypothesis if $H_i \notin \mathcal{T}_j$ for every $j = 1, \ldots, m$. Let ℓ be the total number of leaf hypotheses.

The testing of hypotheses with a DAG structure can be done in several possible ways. Our approach, similar to Chapter 3, is to divide the hypotheses into disjoint families such that a hypothesis and its parent hypotheses are not in the same family.

$$\mathcal{F}_1 = \{H_i : \mathcal{T}_i = \emptyset\},$$

$$\mathcal{F}_2 = \{H_i : H_i \notin \mathcal{F}_1, \mathcal{T}_i \subseteq \mathcal{F}_1\},$$

$$\ldots$$

$$\mathcal{F}_m = \left\{ H_i : H_i \notin \bigcup_{j=1}^{m-1} \mathcal{F}_j, \mathcal{T}_i \subseteq \bigcup_{j=1}^{m-1} \mathcal{F}_j \right\}.$$

Remark 4.3. $\mathcal{F}_1$ is the set of hypotheses at the top of the DAG and these hypotheses are always tested since they do not have parents. Also, there can never be more than m families as there are only m hypotheses, and in most cases the number of families will be much less than m. It should be noted that a given hypothesis H_j belongs to $\mathcal{F}_i$ if the longest path starting from a hypothesis in $\mathcal{F}_1$ to H_j has exactly i elements. Hence, one algorithm to determine which family H_i belongs to is to compute the longest path from a hypothesis in $\mathcal{F}_1$ to H_i. This can be done by the Bellman-Ford algorithm (Cormen et al. 2009).

We use the following procedure as our base procedure to test hypotheses along the DAG.

Definition 4.2. DAG Testing Procedure

Given a sequence of critical constants at level β, $\alpha_i(\beta), 1, \ldots, m$.

1. *For each hypothesis $H_i \in \mathcal{F}_1$, if $P_i \leq \alpha_i(\beta)$, then reject H_i. If $P_i > \alpha_i(\beta)$, then accept each $H_j \in \mathcal{M}_i$, which includes H_i.*

2. *Generally, to test $\mathcal{F}_k$, $k = 2, \ldots, m$, for each hypothesis $H_i \in \mathcal{F}_k$ that has not been accepted, reject H_i if $P_i \leq \alpha_i(\beta)$. If $P_i > \alpha_i(\beta)$, then accept each $H_j \in \mathcal{M}_i$, which includes H_i.*

Remark 4.4. The DAG Testing Procedure ensures a hypothesis is tested only if all of its parents are rejected. To see this, consider $H_i \in \mathcal{F}_k$. By construction of $\mathcal{F}_k$, we

have that $\mathcal{T}_i \subseteq \bigcup_{j=1}^{k-1} \mathcal{F}_j$ so that each of H_i's parents must be in $\mathcal{F}_1, \ldots, \mathcal{F}_{k-1}$ which are tested before testing $\mathcal{F}_k$. If any of H_i's parents are accepted, then H_i will be accepted before testing $\mathcal{F}_k$. Thus, H_i has a chance to be tested only if all of its parents are rejected.

Before presenting our main result, we introduce one more notation. For each pair of hypotheses, H_i and H_j, define $s_{i,j}$ as follows.

$$s_{i,j} = \begin{cases} 0 & \text{if } H_i \notin \mathcal{D}_j, \\ 1 & \text{if } i = j, \\ \sum_{H_k \in \mathcal{T}_j} \dfrac{s_{i,k}}{|\mathcal{T}_j|} & \text{if } i \neq j, H_i \in \mathcal{D}_j. \end{cases} \tag{4.5}$$

Now, define

$$\ell_i = \sum_{H_j \text{ is a leaf}} s_{i,j}.$$

In terms of interpretation, we can interpret $s_{i,j}$ by considering the following analogy. Suppose water flows from each vertex in the opposite direction of the edges. At vertex H_j, the amount of water flowing through H_j is divided evenly among its parents so that a proportion of $1/|\mathcal{T}_j|$ of the water flowing through H_j flows through each of H_j's parents. In this analogy, $s_{i,j}$ represents the proportion of water starting from H_j and flowing through H_i, and ℓ_i represents the amount of water starting from leaf hypotheses and flowing through H_i. If $H_i \notin \mathcal{D}_j$, then no water starting from H_j can reach H_i and thus, $s_{i,j} = 0$. Also, all the water starting from H_i flows through H_i so that $s_{i,i} = 1$. By means of this analogy, it is easy to see that the following property holds: $\sum_{i \in \mathcal{F}_1} s_{i,j} = 1$ (i.e., all water starting from H_j must eventually reach the top of the DAG). Similarly, we have that $s_{i,j} = \sum_{k: H_i \in \mathcal{T}_k} s_{k,j}/|\mathcal{T}_k|$. This follows from the fact that the proportion of water from H_j, flowing through one of H_i's children,

say H_k, is $s_{k,j}$, and then through H_i is $s_{k,j}/|\mathcal{T}_k|$. We will verify these two properties mathematically in the appendix (see (C.1) and (C.3)).

Below, we introduce a base procedure controlling the PFER for testing hypotheses along a directed acyclic graph.

Proposition 4.3. *Given a fixed $\lambda > 0$, the DAG testing procedure with critical constants*

$$\alpha_i(\beta) = \min\left(\lambda, \frac{\beta}{\ell}\right)\frac{\ell_i}{1+\ell_i\lambda}, i = 1, \ldots, m$$

satisfies conditions A1-A3 for $L = V$ under independence.

Proof. The proof is in Appendix C. ■

Remark 4.5. Based on the proof of Proposition 4.3, in Appendix C, it can be seen that the critical constants for leaf hypotheses can be made slightly larger (see (C.5) in the Appendix C). Specifically, if H_i is a leaf hypothesis, then the DAG testing procedure still satisfies conditions A1-A3 if $\alpha_i(\beta) = \min\left(\lambda, \frac{\beta(1+\lambda)}{\ell}\right)\frac{\ell_i}{1+\ell_i\lambda}$. Note that for small λ, this offers little improvement. Hence, we opted for a simpler form for the critical constants which does not distinguish between a leaf and non-leaf hypothesis.

Remark 4.6. It should be noted that the FWER controlling procedure for testing hierarchically ordered hypotheses introduced in Meinshausen (2008) at level α is equivalent to the DAG testing procedure at level α with critical constants $\alpha_i(\alpha) = \ell_i\alpha/\ell$. By letting $\lambda = \alpha/\ell$, the procedure introduced in Proposition 4.3 at level α reduces to the DAG testing procedure with critical constants $\alpha_i(\alpha) = \alpha\ell_i/(\ell + \ell_i\alpha/\ell)$. Following remark 4.2, our proposed base procedure controls the PFER, which is more conservative than the FWER, and when $\lambda = \alpha/\ell$, it has similar but slightly smaller critical constants than Meinshausen's FWER controlling procedure.

By using this procedure as the base procedure along with the SCR procedure, we create the new FDR controlling procedure below.

Theorem 4.4. *Given a fixed $\lambda > 0$, the SCR procedure using the the base procedure from Proposition 4.3 and weight $W(R) = 1/(R \vee 1)$ controls the FDR at level α under independence.*

Proof. The proof follows directly from Theorem 4.2 and the fact that the base procedure from Proposition 4.3 satisfies conditions A1-A3 under independence. ■

We will refer to the procedure in Theorem 4.4 as the SCR DAG procedure. The procedure first determines R based on (4.3) using the base procedure from Proposition 4.3. Then, the hypotheses are tested by applying the procedure from Proposition 4.3 at level $R\alpha$, which is summarized as follows.

1. For each hypothesis $H_i \in \mathcal{F}_1$, if $P_i \leq \min\left(\lambda, \frac{R\alpha}{\ell}\right) \frac{\ell_i}{1+\ell_i\lambda}$, then reject H_i. Otherwise, accept each $H_j \in \mathcal{M}_i$.

2. Generally, to test $\mathcal{F}_k$, $k = 2, \ldots, m$, for each hypothesis $H_i \in \mathcal{F}_k$ that has not been accepted, reject H_i if $P_i \leq \min\left(\lambda, \frac{R\alpha}{\ell}\right) \frac{\ell_i}{1+\ell_i\lambda}$. Otherwise, accept each $H_j \in \mathcal{M}_i$.

Remark 4.7. It should be noted from the above description of this procedure that the critical constants in Theorem 4.4 rely on the tuning parameter λ and that the procedure will not reject H_i if P_i is larger than $\ell_i\lambda/(1+\ell_i\lambda)$. This is true no matter how large R is. As λ increases, $\ell_i\lambda/(1+\ell_i\lambda)$ increases, but $\ell_i R\alpha/(\ell(1+\ell_i\lambda)$ decreases, so λ must be chosen sensibly to allow the procedure to be powerful. We have found that when λ is chosen to be a value close to α, the SCR DAG procedure tends to give good results in terms of the number of rejections.

In the special configuration where each hypothesis has 0 or 1 parent hypotheses, the DAG structure reduces to the hierarchical structure presented in Chapter 3. In this case, $s_{i,j}$ is 0 or 1 depending on whether $H_i \in \mathcal{D}_j$ and ℓ_i is simply the total number of leaf hypotheses in $\mathcal{M}_i$. Under the hierarchical configuration, the SCR DAG procedure is as follows.

Corollary 4.5. *Given $\lambda > 0$, if the hypotheses are arranged into a hierarchy, where each hypothesis has 0 or 1 parent hypothesis, then the SCR procedure using the DAG testing procedure with critical constants*

$$\alpha_i(\beta) = \min\left(\lambda, \frac{\beta}{\ell}\right) \frac{\ell_i}{1 + \ell_i \lambda}, i = 1, \ldots, m$$

as the base procedure, where ℓ_i is the number of leaf hypotheses in $\mathcal{M}_i$, controls the FDR at level α.

The procedure described in Corollary 4.5 is different from the hierarchical testing procedures described in Chapter 3. Hence, we have developed a new FDR controlling procedure to test hierarchically ordered hypotheses.

Now consider the fixed sequence configuration, where each hypothesis has 1 parent hypothesis (except H_1 which has 0), 1 child hypothesis (except H_m which has 0), and H_i cannot be tested unless $H_1, \ldots, H_{i-1}$ have been rejected. Under this configuration, $\ell = 1$ and $\ell_i = 1$.

Corollary 4.6. *Given $\lambda > 0$, if the hypotheses are arranged into a fixed sequence, then the SCR procedure using the DAG testing procedure with critical constants*

$$\alpha_i(\beta) = \min(\lambda, \beta) \frac{1}{1 + \lambda}, i = 1, \ldots, m$$

as the base procedure controls the FDR at level α.

Hence, we have also created a new FDR controlling fixed sequence procedure. In this case, the rejection thresholds are all $\min(\lambda, R\alpha)/(1 + \lambda)$.

In addition to our proposed FDR controlling procedure for DAG testing, we can also develop a procedure similar to the BH procedure. If all the hypotheses are tested simultaneously by the BH procedure, the set of rejected hypotheses do not preserve the underlying structure; however, the BH procedure can easily be modified so the hypotheses are tested along a DAG so that DAG structure is preserved. As the base procedure, we consider the DAG testing procedure using Bonferroni critical constants. Then, we have the following result.

Proposition 4.7. *The DAG testing procedure with critical constants*

$$\alpha_i(\beta) = \beta/m, i = 1, \ldots, m$$

satisfies conditions A1-A3 for $L = V$ under Assumption 1.1 of Chapter 1.

Proof. Let $L_i(\alpha) = I\{P_i \leq \alpha/m, H_i \text{ is true}\}$. Then,

$$V = \sum_{i=1}^{m} I\{H_i \text{ is falsely rejected}\} \leq \sum_{i=1}^{m} I\{P_i \leq \alpha/m, H_i \text{ is true}\} = \sum_{i=1}^{m} L_i(\alpha).$$

The inequality follows from the fact that the event $\{H_i \text{ is falsely rejected}\}$ implies $\{P_i \leq \alpha/m\}$. Condition A1 is satisfied. Let $c_i = 1/m$ and note that $E(L_i(\alpha)) \leq \alpha/m = c_i\alpha$. Condition A2 is satisfied. Finally, condition A3 follows directly from Assumption 1.1. ■

By using the SCR procedure with the base procedure in Proposition 4.7 and weight $W(R) = 1/(R \vee 1)$, we can create a new FDR controlling procedure. We will refer to this new FDR controlling procedure as the DAG BH procedure. After determining R as in (4.3), the testing is done as follows

1. For each hypothesis $H_i \in \mathcal{F}_1$, if $P_i \leq R\alpha/m$, then reject H_i. Otherwise, accept each $H_j \in \mathcal{M}_i$.

2. Generally, to test $\mathcal{F}_k$, $k = 2, \ldots, m$, for each hypothesis $H_i \in \mathcal{F}_k$ that has not been accepted, reject H_i if $P_i \leq R\alpha/m$. Otherwise, accept each $H_j \in \mathcal{M}_i$.

The DAG BH procedure is very similar to the BH procedure since both procedures will reject H_i if $P_i \leq R\alpha/m$ so long as H_i can be tested. The DAG BH procedure is less powerful than the usual BH procedure since not every hypothesis will be tested by using the DAG BH procedure, but the rejections by the DAG BH procedure may be easier to interpret since the DAG structure is preserved.

4.5 Simulation Study

Through simulation studies, we evaluated the performance of the SCR DAG procedure and the DAG BH procedure in terms of FDR and average power by comparing it with the BH procedure. It should be noted that the BH procedure does not preserve the DAG structure, but we still included it in our simulation study because it is a very well-known and powerful FDR controlling procedure.

We constructed a graph with 3003 hypotheses with 3 families such that $\mathcal{F}_1 = \{H_1, \ldots, H_{1000}\}, \mathcal{F}_2 = \{H_{1001}, \ldots, H_{2001}\}$, and $\mathcal{F}_3 = \{H_{2002}, \ldots, H_{3003}\}$. Each $H_i \in \mathcal{F}_1$ had two child hypotheses in $\mathcal{F}_2$: H_{i+1000} and H_{i+1001}. Similarly, each $H_i \in \mathcal{F}_2$ had two child hypotheses in $\mathcal{F}_3$: H_{i+1001} and H_{i+1002}. A randomly selected proportion of the leaf hypotheses were set to be true with the remaining set to false. Each non-leaf hypothesis was set to true only if all of its child hypotheses were true; otherwise, it was set to false.

We generated m normal random variables with covariance matrix Σ and mean vector $\vec{\mu} = (\mu_1, \ldots, \mu_m)$ to test the m hypotheses $H_i : \mu_i \leq 0$ versus $H_i' : \mu_i > 0, i = 1, \ldots, m$. The p-value for testing H_i was calculated using a one sided, one-sample Z-test. When H_i was true, we set $\mu_i = 0$. When H_i was false, we set μ_i to a value

that depended on its location in the graph. Specifically, $\mu_i = 3$ if $H_i \in \mathcal{F}_1$, $\mu_i = 2$ if $H_i \in \mathcal{F}_2$, and $\mu_i = 1$ if $H_i \in \mathcal{F}_3$. As for the joint dependence, we considered a common correlation structure where Σ had off-diagonal components equal to ρ and diagonal components equal to 1.

We set $\alpha = 0.05$ and $\lambda = 0.1$. For each procedure we noted the false discovery proportion, which is the proportion of falsely rejected hypotheses among all rejected hypotheses, and the the proportion of rejected false null hypotheses among all false null hypotheses. The graph was generated and tested 5000 times and the simulated values of the FDR and average power were obtained by averaging out the 5000 values of these two proportions, respectively.

Figure 4.2 shows the FDR and average power of the SCR DAG procedure, the DAG BH procedure, and the BH procedure under independence, common correlation with $\rho = 0.3$, and common correlation with $\rho = 0.7$. It can be seen in Figure 4.2 that all three procedures control the FDR at level 0.05 under the three dependence settings considered. It should be noted that we were only able to show theoretically that the SCR DAG procedure controls the FDR under independence, yet Figure 4.2 shows that the SCR DAG procedure controls the FDR under both mild and strong correlation. In terms of power, the SCR DAG procedure outperforms the BH procedure and the DAG BH procedure fairly significantly. It is interesting to note that under independence and weak dependence, the SCR DAG procedure has a smaller FDR than the BH procedure but larger power.

4.6 Real Data Analysis

We applied our proposed SCR DAG procedure and the DAG BH procedure to the real microarray data set of Golub et al. (1999) involving 27 patients with acute

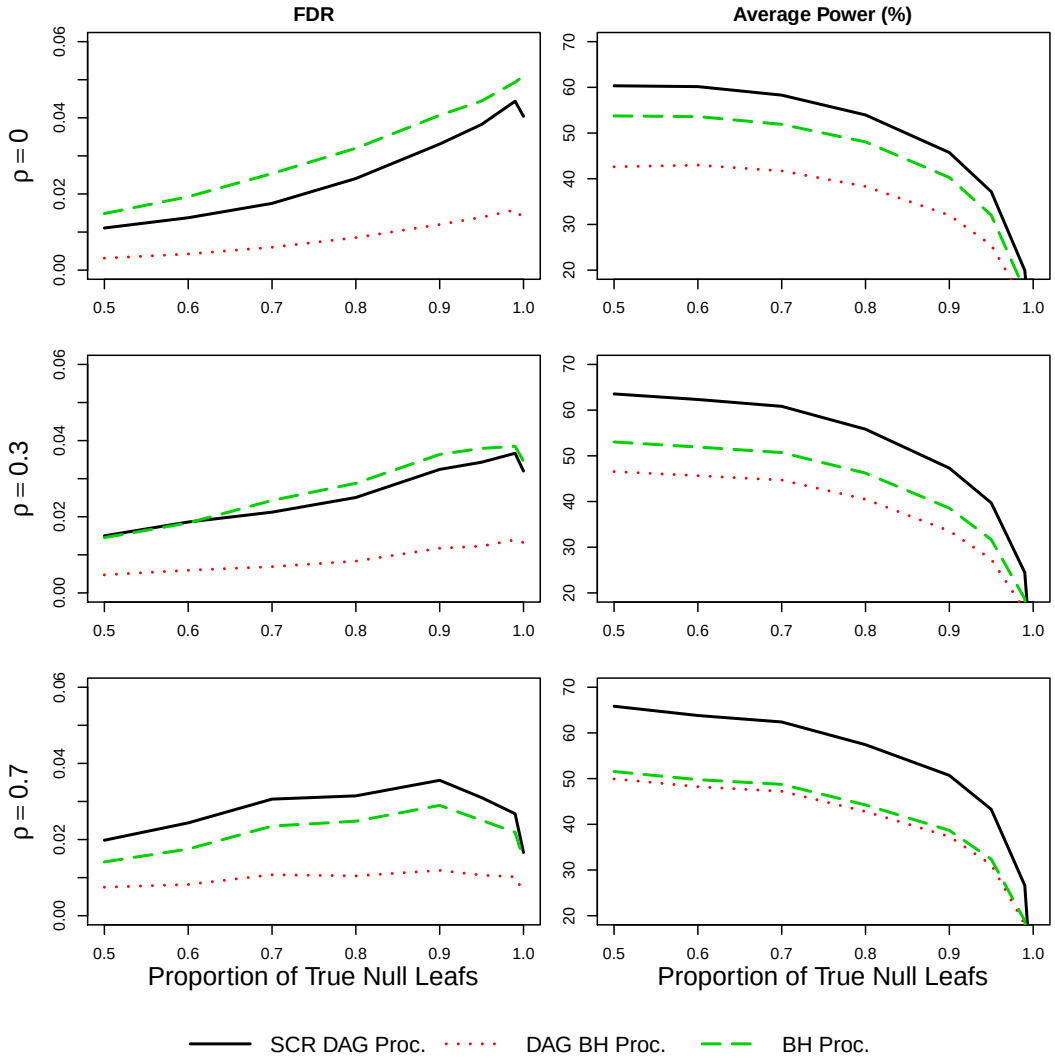


Figure 4.2 FDR (left column) and average power (right column) of the SCR DAG procedure (solid line), the DAG BH procedure (dashed), and the BH procedure (dotted) under independence (top row), common correlation with $\rho = 0.3$ (middle row), and common correlation with $\rho = 0.7$ (bottom row).

lymphoblastic leukemia and 11 patients with acute myeloid leukemia. The data set is available from the Bioconductor golubEsets package. The data consist of 7,192 expression levels across all 38 patients assayed using Affymetrix Hgu6800 chips. The probes were mapped to Gene Ontology (GO) biological process terms and of the 7,192 probes, 5,819 were mapped successfully resulting in a total of 10,362 GO terms.

The terms in the GO structure form a directed acyclic graph. If a probe is mapped to a GO term, then it is also mapped to each ancestor of this GO term. Hence, terms lower on the graph refer to more specific gene functions, but terms higher on the graph refer to more general gene functions.

Our aim is to determine which gene terms have different gene expression levels between ALL patients and AML patients. We created a null and alternative hypothesis for each GO term that had at least one probe mapped to it. The null hypothesis states that there is no difference in the expression levels between ALL and AML patients for any of the probes mapped to this term, and the alternative states that there is a difference in the expression level for at least one probe mapped to this term. Hence, if the hypothesis corresponding to a GO term is false, then all of the hypotheses corresponding to its ancestors are also false because the probe with an expression level difference is also mapped to the term's ancestors.

We tested all the 10,362 hypotheses corresponding to the GO terms which were divided into 19 families. We calculated the p-values for each GO term using the approach by Goeman et al. (2004) and tested the hypotheses using the SCR DAG procedure with $\lambda = 2\alpha$, the DAG BH, and the BH procedure. Table 4.1 lists the number of rejections for the three procedures at various significance levels. For all the significance levels, the SCR DAG procedure has the most number of rejections. The DAG BH performs the worst. It is important to remember that the rejections by both the SCR DAG procedure and the DAG BH procedure preserve the underlying GO structure and perhaps have a more natural interpretation, but the rejections by the BH procedure lack such biological interpretation.

Table 4.1 The Number of Rejections Out of 10,362 Hypotheses by the SCR DAG Procedure, the DAG BH, and the BH procedure at Various Significance Levels for the Golub Leukemia Data Set

α	**SCR DAG Procedure**	**DAG BH Procedure**	**BH Procedure**
0.0001	2226	1882	2170
0.001	3226	2746	3086
0.01	4368	3947	4245
0.025	5066	4594	4898
0.05	5705	5210	5542
0.1	6384	6017	6258

4.7 Conclusion

In this chapter, we have provided a novel approach for developing procedures controlling large scale error rates, such as the FDR. This new approach can be used to prove FDR control of existing procedures, such as the adaptive BH procedure, and can also be used to create new procedures for controlling recently introduced error rates, such as the generalized FDR controlling procedure introduced in example 4.5.

By using this novel approach, we have developed a new procedure controlling the FDR under independence for testing hypotheses along a DAG and also showed that the BH procedure can be modified for DAG testing. In our simulation study, the newly proposed procedure was shown to be more powerful than the BH procedure by accounting for the underlying DAG structure. Moreover, the set of rejected hypotheses by the SCR DAG procedure preserves the graph structure, which may enhance intuitive interpretation of our discoveries. Finally, because the DAG structure is a very general structure, the SCR DAG procedure also has applications towards the testing of hypotheses that have a hierarchical or fixed sequence structure, which can be viewed as a special DAG structure.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The literature on large scale multiple hypothesis testing has typically focused on the dependence of the test statistics and has largely ignored the structure of the tested hypotheses. In this dissertation, we considered testing hypotheses with several different types of structure and developed novel methods which exploit this inherent structure.

In Chapter 2, we presented several new FDR controlling procedures for testing hypotheses with a fixed, pre-specified testing order, and discussed a data driven ordering method to be used when the testing order is not known in advance. Hypotheses associated with stream data, such as network traffic data, can have an inherent fixed sequence structure since the hypotheses are often be ordered by time. So, a future direction of research related to the work in Chapter 2 is to explore methods for testing these stream data hypotheses. We have already suggested one possible FDR controlling procedure for this application in the conclusion of Chapter 2, which tests all the hypotheses, no matter how many are accepted.

In Chapter 3, we described a generalized stepwise procedure and a general approach to hierarchical testing. Using our hierarchical testing approach, we introduced several new and powerful hierarchical testing procedures controlling the FDR under different dependence structures. Recently, Dmitrienko et al. (2007, 2013) have discussed multiple testing in the FWER framework where the hypotheses are grouped into families and the decision to test a hypothesis depends on what was rejected in the previous tested families. A possible future work is to extend our hierarchical testing approach, which tests families sequentially and carries forward

the number of rejections, and develop a novel procedure in the FDR framework for sequentially testing multiple families of hypotheses.

Finally, in Chapter 4 we presented a novel approach that makes the task of developing procedures that control large scale error rates significantly easier. By using this approach, we developed a new FDR controlling procedure for testing hypotheses along a directed acyclic graph. We showed that the general stepup-based SCR procedure can reduce to many existing stepup procedures under different contexts, such as the BH procedure, the weighted BH procedure, the oracle BH proceudre, and an adaptive BH procedure. However, there are also many existing stepdown procedures for large scale multiple testing, such as the procedures by Benjamini and Liu (1999), Guo and Rao (2008), Gavrilov et al. (2009). One possible future work is to develop a general stepdown-based SCR procedure and study the conditions under which it controls the overall error rate.

As a concluding remark, we would like to point out that in terms of the way the hypotheses are tested there is a distinct difference between the methods described in Chapters 2 and 3 and the methods described in Chapter 4. Our approach in Chapter 4 was to determine the total number of rejections out of all the m tested hypotheses before actually testing any of the hypotheses. Hence, the p-values for all m hypotheses are needed at the time of testing. This is different from Chapters 2 and 3, where testing is done in stages. Such stage-wise testing can proceed as p-values become available and is useful when testing hypotheses associated with stream data, where a decision regarding a hypothesis is to be made before all the data is available.

APPENDIX A

PROOFS OF THE THEOREMS IN CHAPTER 2

This appendix contains the proof of the lemmas and theorems stated but not proved in Chapter 2.

A.1 Proof of Theorem 2.1 (ii)

For any u_1, $1 \leq u_1 \leq m$, consider a joint distribution of the p-values such that the first $u_1 - 1$ hypotheses are false null hypotheses whose corresponding p-values are 0 with probability one. The remaining $m - u_1 + 1$ hypotheses are true null hypotheses such that $\hat{P}_1 \sim U(0,1)$ and $\hat{P}_i = \hat{P}_1$ for $i = 2, \ldots, m_0$. Under such joint distribution of the p-values, the FDR of the conventional fixed sequence procedure is exactly $\alpha_{u_1}^{(1)}(m - u_1 + 1)/m$. If $\alpha_{u_1}^{(1)} = 1$ then the critical constant is already at its largest and cannot be improved. Otherwise, if $\alpha_{u_1}^{(1)} < 1$, then FDR $= \alpha$ and thus $\alpha_{u_1}^{(1)}$ cannot made any larger.

In the above construction $\vec{P}$ satisfies Assumption 1.1. We only need to note that for $i = 1, \ldots, m_0$, $\widehat{P}_i = p$ implies $\vec{P} = (0, \ldots, 0, p, \ldots, p)$. Thus, for any increasing set D,

$$\Pr\left(\vec{P} \in D | \widehat{P}_i = p\right) = \begin{cases} 1 & \text{if } (0, \ldots, 0, p, \ldots, p) \in D \\ 0 & \text{otherwise,} \end{cases}$$

which is an increasing function in p. ■

A.2 Proof of Lemma 2.1

Lemma 2.1 can be regarded as a special case of Lemma 2.2 with $k = 1$. Note that for $i = 1, \ldots, m$, $m_{1,i} = m_{1,i-1}$ and $m_{0,i} = m_{0,i-1} + 1$ when $i \in I_0$, $m_{0,i} = m_{0,i-1}$ when $i \notin I_0$, and $m_{0,i-1} + m_{1,i-1} = i - 1$. Thus, when $k = 1$, the event $\{R \geq i\}$ implies $V_i = m_{0,i}$ and hence

$$\frac{V_i}{i} - \frac{V_{i-1}}{i-1} = \begin{cases} I\{1 \in I_0\} & \text{for } i = 1 \\ \frac{m_{1,i-1} I\{i \in I_0\} - m_{0,i-1} I\{i \notin I_0\}}{i(i-1)} & \text{for } i = 2, \ldots, m. \end{cases} \tag{A.1}$$

By (A.1) and Lemma 2.2, the desired result follows. ■

A.3 Proof of Lemma 2.2

It is easy to see that

$$\begin{aligned} \text{FDR} &= E\left(\sum_{i=1}^{m} \frac{V_i}{i} I\{R = i\}\right) = E\left(\sum_{i=1}^{m} \left(\frac{V_i}{i} I\{R \geq i\} - \frac{V_i}{i} I\{R \geq i+1\}\right)\right) \\ &= E\left(\sum_{i=1}^{m} \left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1}\right) I\{R \geq i\}\right) \\ &= E\left(\sum_{i=1}^{m} \left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1}\right) I\{J_i < m+1\}\right), \end{aligned}$$

the desired result. ■

A.4 Proof of Lemma 2.3

We note that for $i = 1, \ldots, m$, if there are at least i rejections, then $i \leq J_i \leq \min(i+k-1, m)$. For ease of notation, let $j_i^* = \min(i+k-1, m)$. For $i, j = 1, \ldots, m$,

define $f_i(j) = \frac{(k-j+i)\alpha}{k}\frac{S_i}{i}$ and $W_i(j) = I\{J_{i-1} \leq j, J_i > j\}$. Regarding the relationship between J_i and $W_i(j)$, there are the following two equalities available:

$$I\{J_i = j\} = W_i(j-1)I\{P_j \leq \alpha_j^{(4)}\} \tag{A.2}$$

and

$$W_i(j) - W_i(j-1) = I\{J_{i-1} = j\} - I\{J_i = j\}. \tag{A.3}$$

The first equality follows from the fact that for $i = 1, \ldots, m$ and $j = i, \ldots, j_i^*$, when $J_i = j$, there are $i-1$ rejections among the first $j-1$ tested hypotheses and the i^{th} rejection is exactly the j^{th} tested hypothesis, thus

$$I\{J_i = j\} = I\{J_{i-1} \leq j-1, J_i > j-1, P_j \leq \alpha_j^{(4)}\} = W_i(j-1)I\{P_j \leq \alpha_j^{(4)}\}.$$

The second equality follows from the fact that the event $\{W_i(j) = 1\}$ implies that there are exactly $i-1$ rejections among the first j tested hypotheses, thus for $j = i-1, \ldots, j_{i-1}^*$,

$$\begin{aligned}
& W_i(j) - W_i(j-1) = I\{J_{i-1} \leq j, J_i > j\} - W_i(j-1) \\
= \;& I\{J_{i-1} = j\} + I\{J_{i-1} \leq j-1, J_i > j-1, P_j > \alpha_j^{(4)}\} - W_i(j-1) \\
= \;& I\{J_{i-1} = j\} - W_i(j-1)I\{P_j \leq \alpha_j^{(4)}\} \\
= \;& I\{J_{i-1} = j\} - I\{J_i = j\},
\end{aligned}$$

where, the third equality follows from the fact that $I\{P_j > \alpha_j^{(4)}\} = 1 - I\{P_j \leq \alpha_j^{(4)}\}$ and the fourth follows from (A.2).

By using the above two equalities, we can prove two inequalities below, which are needed in the proof of this lemma. Firstly, we show by using (A.2) that the following inequality holds:

$$E\left(\left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1} + f_i(j) - f_{i-1}(j)\right) I\{J_i = j\}\right) \leq E\left(\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1)\right). \tag{A.4}$$

Proof of (A.4). To see this, we consider, separately, the case when $j \in I_0$ and when $j \notin I_0$.

Suppose $j \in I_0$, then $S_i = S_{i-1}$ and $V_i = V_{i-1} + 1$ when $J_i = j$. Using the fact that $V_{i-1} + S_{i-1} = i - 1$, the left hand side of (A.4), after some algebra, becomes

$$\begin{aligned} & E\left(\frac{k+(j-k)\alpha}{ki}\frac{S_{i-1}}{i-1}W_i(j-1)I\{P_j \le \alpha_j^{(4)}\}\right) \\ = \;& E\left(\frac{k+(j-k)\alpha}{ki}\frac{S_{i-1}}{i-1}W_i(j-1)\right)\Pr\left(P_j \le \alpha_j^{(4)}\right) \\ \le \;& E\left(\frac{k+(j-k)\alpha}{ki}\frac{S_{i-1}}{i-1}W_i(j-1)\right)\frac{i\alpha}{k+(j-k)\alpha} \\ = \;& E\left(\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1)\right). \end{aligned}$$

The first equality follows from these two facts: (i) When $J_i = j$ (i.e., the i^{th} rejection is H_j) S_{i-1} is only dependent on the first $j-1$ p-values, since S_{i-1} is the number of false null hypotheses among the first $i-1$ rejections; (ii) $W_i(j-1)$ is also only dependent on the first $j-1$ p-values, since $W_i(j-1)$ is 1 if and only if there are exactly $i-1$ rejections among the first $j-1$ hypotheses tested. The inequality follows from (1.1).

Now suppose $j \notin I_0$, then $S_i = S_{i-1} + 1$ and $V_i = V_{i-1}$. Similarly, using the fact that $V_{i-1} + S_{i-1} = i - 1$, the left hand side of (A.4), after some algebra, becomes

$$\begin{aligned} & E\left(\left(\frac{-V_{i-1}}{i(i-1)} + \frac{((j-k)S_{i-1} + (i-1)(k-j+i))\alpha}{ki(i-1)}\right)W_i(j-1)I\{P_j \le \alpha_j^{(4)}\}\right) \\ \le \;& E\left(\left(\frac{-V_{i-1}}{i(i-1)}\frac{(k-j+i)\alpha}{k} + \frac{((j-k)S_{i-1} + (i-1)(k-j+i))\alpha}{ki(i-1)}\right)W_i(j-1)\right) \\ = \;& E\left(\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1)\right). \end{aligned}$$

The inequality follows by the fact that $j \ge i$ so that $k - j + i \le k$. In addition, in the last line we use the fact that $V_{i-1} + S_{i-1} = i - 1$. ■

Next, we show by using (A.3) that the following inequality holds:

$$E\left(f_{i-1}(J_{i-1})I\{J_{i-1}<m+1\}-f_{i-1}(J_i)I\{J_i<m+1\}\right)\geq E\left(\sum_{j=i}^{j_i^*}\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1)\right). \tag{A.5}$$

Proof of (A.5). By using (A.3), we have

$$\begin{aligned}
& E\left(f_{i-1}(J_{i-1})I\{J_{i-1}<m+1\}-f_{i-1}(J_i)I\{J_i<m+1\}\right)\\
= \;& E\left(\sum_{j=i-1}^{j_{i-1}^*}f_{i-1}(j)I\{J_{i-1}=j\}-\sum_{j=i}^{j_i^*}f_{i-1}(j)I\{J_i=j\}\right)\\
= \;& E\left(\sum_{j=i-1}^{j_{i-1}^*}f_{i-1}(j)I\{J_{i-1}=j\}-\sum_{j=i}^{j_{i-1}^*}f_{i-1}(j)I\{J_i=j\}\right)\\
= \;& E\left(\sum_{j=i}^{j_{i-1}^*}f_{i-1}(j)\left(I\{J_{i-1}=j\}-I\{J_i=j\}\right)+f_{i-1}(i-1)I\{J_{i-1}=i-1\}\right)\\
= \;& E\left(\sum_{j=i}^{j_{i-1}^*}f_{i-1}(j)(W_i(j)-W_i(j-1))+f_{i-1}(i-1)W_i(i-1)\right)\\
= \;& E\left(\sum_{j=i}^{j_{i-1}^*}\left(f_{i-1}(j-1)-f_{i-1}(j)\right)W_i(j-1)+f_{i-1}(j_{i-1}^*)W_i(j_{i-1}^*)\right)\\
\geq \;& E\left(\sum_{j=i}^{j_i^*}\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1)\right),
\end{aligned}$$

the desired result. Here, the second equality follows from the fact that if $j_{i-1}^*=m$, then $j_i^*=m$; otherwise, $j_{i-1}^*=i+k-2$ and $j_i^*=i+k-1$ so that $f_{i-1}(j_i^*)=0$. The fourth equality follows from (A.3) and the fact that $W_i(i-1)=I\{J_{i-1}=i-1\}$. The last inequality follows from the definition of $f_{i-1}(j)$. ■

Proof of Lemma 2.3. Finally, by combining these two inequalities, we can get the desired result as follows.

$$E\left(\left(\frac{V_i}{i}-\frac{V_{i-1}}{i-1}\right)I\{J_i<m+1\}\right)$$

$$
\begin{aligned}
&= E\left(\left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1} + f_i(J_i) - f_{i-1}(J_i)\right) I\{J_i < m+1\}\right) \\
&\quad -E\left(f_{i-1}(J_{i-1})I\{J_{i-1} < m+1\} - f_{i-1}(J_i)I\{J_i < m+1\}\right) \\
&\quad +E\left(f_{i-1}(J_{i-1})I\{J_{i-1} < m+1\} - f_i(J_i)I\{J_i < m+1\}\right) \\
&= E\left(\sum_{j=i}^{j_i^*}\left(\frac{V_i}{i} - \frac{V_{i-1}}{i-1} + f_i(j) - f_{i-1}(j)\right) I\{J_i = j\}\right) \\
&\quad -E\left(f_{i-1}(J_{i-1})I\{J_{i-1} < m+1\} - f_{i-1}(J_i)I\{J_i < m+1\}\right) \\
&\quad +E\left(f_{i-1}(J_{i-1})I\{J_{i-1} < m+1\} - f_i(J_i)I\{J_i < m+1\}\right) \\
&\leq E\left(\sum_{j=i}^{j_i^*}\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1) - \sum_{j=i}^{j_i^*}\frac{\alpha}{k}\frac{S_{i-1}}{i-1}W_i(j-1)\right) \\
&\quad +E\left(f_{i-1}(J_{i-1})I\{J_{i-1} < m+1\} - f_i(J_i)I\{J_i < m+1\}\right) \\
&= E\left(f_{i-1}(J_{i-1})I\{J_{i-1} < m+1\} - f_i(J_i)I\{J_i < m_1\}\right),
\end{aligned}
$$

The inequality follows from (A.4) and (A.5). ■

APPENDIX B

PROOFS OF THE THEOREMS IN CHAPTER 3

This appendix contains the proof of the proposition and theorems stated but not proved in Chapter 3. Let us first state and prove the following lemmas which will be helpful in the proofs of Theorems 3.2, 3.3, 3.4, and 3.5.

Lemma B.1. *Under Assumption 1.1, if* $\Gamma(P_1, \ldots, P_m)$ *is a discrete coordinatewise non-increasing function of the p-values taking on values* $\gamma_1 < \cdots < \gamma_n$ *and* $t(\cdot)$ *is a non-decreasing function, then for each true null* H_j*,*

$$\sum_{i=1}^{n} \Pr\left(\Gamma = \gamma_i \mid P_j \leq t(\gamma_i)\right) \leq \Pr\left(\Gamma \geq \gamma_1 \mid P_j \leq t(\gamma_1)\right).$$

Proof of Lemma B.1.

$$\begin{aligned}
&\sum_{i=1}^{n} \Pr\left(\Gamma = \gamma_i \mid P_j \leq t(\gamma_i)\right) \\
=&\sum_{i=1}^{n} \Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_i)\right) - \sum_{i=1}^{n-1} \Pr\left(\Gamma \geq \gamma_{i+1} \mid P_j \leq t(\gamma_i)\right) \\
=&\Pr\left(\Gamma \geq \gamma_1 \mid P_j \leq t(\gamma_1)\right) - \sum_{i=2}^{n} \left[\Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_{i-1})\right) - \Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_i)\right)\right] \\
\leq&\Pr\left(\Gamma \geq \gamma_1 \mid P_j \leq t(\gamma_1)\right).
\end{aligned}$$

The inequality follows by Assumption 1.1. ■

Lemma B.2. *Under arbitrary dependence, if* $\Gamma(P_1, \ldots, P_m)$ *is a discrete function of the p-values taking on values* $\gamma_0 < \cdots < \gamma_n$ *and* $t(\cdot)$ *is a positive non-decreasing function with* $t(\gamma_0) = 0$*, then for each true null* H_j*,*

$$\sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr\left(\Gamma = \gamma_i, P_j \leq t(\gamma_i)\right) \leq \sum_{i=1}^{n} \frac{t(\gamma_i) - t(\gamma_{i-1})}{t(\gamma_i)}.$$

Proof of Lemma B.2. Using the convention that $0/0 = 0$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr\left(\Gamma = \gamma_i, P_j \le t(\gamma_i)\right) \\
=&\sum_{i=1}^{n} \left[\frac{1}{t(\gamma_i)} \Pr\left(\Gamma \ge \gamma_i, P_j \le t(\gamma_i)\right) - \frac{1}{t(\gamma_{i-1})} \Pr\left(\Gamma \ge \gamma_i, P_j \le t(\gamma_{i-1})\right)\right] \\
\le&\sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr\left(\Gamma \ge \gamma_i, t(\gamma_{i-1}) < P_j \le t(\gamma_i)\right) \\
\le&\sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr\left(t(\gamma_{i-1}) < P_j \le t(\gamma_i)\right) \\
=&\sum_{i=1}^{n-1} \left(\frac{1}{t(\gamma_i)} - \frac{1}{t(\gamma_{i+1})}\right) \Pr\left(P_j \le t(\gamma_i)\right) + \frac{1}{t(\gamma_n)} \Pr\left(P_j \le t(\gamma_n)\right) \\
\le&\sum_{i=1}^{n} \frac{t(\gamma_i) - t(\gamma_{i-1})}{t(\gamma_i)}. \blacksquare
\end{aligned}$$

B.1 Proof of Proposition 3.1

Assume $k > \psi(k)$. Then, step 2(a) is repeated until for some $\ell \ge 2$, $r_\ell \le \psi(r_\ell)$. For $t = 1, \ldots, \ell - 1$, we have $r_t > \psi(r_t)$ implying $r_t > r_{t+1}$. Thus, $k = r_1 > r_\ell$. For any integer r from 0 to $k - 1$ such that $r \le \psi(r)$, we will show that $r_\ell \ge r$. To prove it, we show using induction that $r_t \ge r, t = 1, \ldots, \ell$. Since $r_1 > r$, by induction assume $r_{t-1} \ge r$. Then, $r_t = \psi(r_{t-1}) \ge \psi(r) \ge r$ and thus, $r_\ell \ge r$ follows. Since $r_\ell \le \psi(r_\ell)$, $r_\ell < k$, and $r_\ell \ge r$ for any $0 \le r \le k - 1$ such that $r \le \psi(r)$, we have $R = r_\ell = \max\{0 \le r \le k - 1 : r \le \psi(r)\}$.

Conversely, assume $k \le \psi(k)$. Then, step 2(b) is repeated until for some $\ell \ge 2$, $r_\ell > \psi(r_\ell)$. For $t = 1, \ldots, \ell - 1$, we have $r_t \le \psi(r_t)$ implying $r_t < \psi(r_t) + 1 = r_{t+1}$. Thus, $k = r_1 < r_\ell$. For any integer r from $k + 1$ to $m + 1$ such that $r > \psi(r)$, we will show that $r_\ell \le r$. To prove it, we show using induction that $r_t \le r, t = 1, \ldots, \ell$. Since $r_1 < r$, by induction assume $r_{t-1} \le r$. Then, $r_t = \psi(r_{t-1}) + 1 \le \psi(r) + 1 \le r$ and

thus, $r_\ell \le r$ follows. Since $r_\ell > \psi(r_\ell)$, $r_\ell > k$, and $r_\ell \le r$ for any $k+1 \le r \le m+1$ such that $r > \psi(r)$, we have $R = r_\ell - 1 = \min\{k+1 \le r \le m+1 : r > \psi(r)\} - 1$. ■

B.2 Proof of Theorem 3.2

In this proof and the remaining proofs, we will use the convention that $0/0 = 0$. For convenience of notation, define $\mathcal{G}_d = \bigcup_{j=1}^{d} \mathcal{F}_j$ and $R(\mathcal{G}_d)$ is the number of rejections in the first d families, $\mathcal{F}_j, j = 1, \ldots, d$. Let $|\mathcal{G}_d|$ be the cardinality of $\mathcal{G}_d$.

We will show that

$$E\left(\frac{V(\mathcal{M}_i)}{R}\right) \le \frac{\ell_i \alpha}{\ell}, i = 1, \ldots, m. \tag{B.1}$$

Proof of (B.1). The event $\{H_i \text{ is rejected}\}$ implies all ancestors of H_i are rejected so there must be at least d_i rejections in the first d_i families. Therefore, the event $\{H_i \text{ is rejected}\}$ implies the following two inequalities:

$$d_i \le R(\mathcal{G}_{d_i}) \le |\mathcal{G}_{d_i}|, \tag{B.2}$$

$$R(\mathcal{G}_{d_i}) - 1 \le R - R(\mathcal{M}_i). \tag{B.3}$$

The second inequality follows from the fact that $\mathcal{G}_{d_i}/\{H_i\} \subseteq \mathcal{M}/\mathcal{M}_i$ so that $R(\mathcal{G}_{d_i}/\{H_i\}) \le R(\mathcal{M}/\mathcal{M}_i)$.

If H_i is true,

$$\begin{aligned}
& E\left(\frac{V(\mathcal{M}_i)}{R}\right) \le E\left(\frac{V(\mathcal{M}_i)}{V(\mathcal{M}_i) + R - R(\mathcal{M}_i)}\right) \\
\le\ & E\left(\frac{m_i}{m_i + R - R(\mathcal{M}_i)} I\{H_i \text{ is rejected}\}\right) \le E\left(\frac{m_i}{m_i + R(\mathcal{G}_{d_i}) - 1} I\{H_i \text{ is rejected}\}\right) \\
=\ & \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} E\left(\frac{m_i}{m_i + r - 1} I\{R(\mathcal{G}_{d_i}) = r, H_i \text{ is rejected}\}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{m_i}{m_i + r - 1} \Pr\left(R(\mathcal{G}_{d_i}) = r, P_i \leq \alpha_i(r)\right) \\
&\leq \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{m_i \alpha_i(r)}{m_i + r - 1} \Pr\left(R(\mathcal{G}_{d_i}) = r \mid P_i \leq \alpha_i(r)\right) \\
&= \frac{\ell_i \alpha}{\ell} \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \Pr\left(R(\mathcal{G}_{d_i}) = r \mid P_i \leq \alpha_i(r)\right). \qquad \text{(B.4)}
\end{aligned}
$$

The second inequality follows from the fact that $V(\mathcal{M}_i) \leq m_i$ and $V(\mathcal{M}_i)/(V(\mathcal{M}_i) + R - R(\mathcal{M}_i))$ is an increasing function of $V(\mathcal{M}_i)$. The third inequality follows from (B.3) and the first equality follows from (B.2). The fourth inequality follows by the fact that the event $\{H_i \text{ is rejected}\}$ implies $P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))$.

Since the number of rejections by the generalized stepup procedure is a coordinatewise non-increasing function of the p-values, it follows that $R(\mathcal{G}_{d_i})$ is also a coordinatewise non-increasing function of the p-values. Therefore, by Lemma B.1,

$$
\sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \Pr\left(R(\mathcal{G}_{d_i}) = r \mid P_i \leq \alpha_i(r)\right) \leq \Pr\left(R(\mathcal{G}_{d_i}) \geq d_i \mid P_i \leq \alpha_i(d_i)\right) \leq 1. \qquad \text{(B.5)}
$$

From (B.5), we have that (B.4) is less than $\ell_i \alpha / \ell$. Thus, (B.1) holds when H_i is true.

We will use induction to show (B.1) also holds when H_i is false. When H_i is a false null leaf hypothesis, then (B.1) is true trivially. Otherwise, assume (B.1) is true for every false child hypothesis of H_i. Thus, (B.1) is true for all children of H_i. We note that when H_i is false, $V(\mathcal{M}_i) = \sum_{j:T(j)=i} V(\mathcal{M}_i)$ and

$$
E\left(\frac{V(\mathcal{M}_i)}{R}\right) = \sum_{j:T(j)=i} E\left(\frac{V(\mathcal{M}_j)}{R}\right) \leq \sum_{j:T(j)=i} \frac{\ell_j \alpha}{\ell} = \frac{\ell_i \alpha}{\ell}.
$$

Thus, (B.1) holds for all hypotheses.

Proof of Theorem 3.2. By (B.1), we have

$$
\text{FDR} = \sum_{i:T(i)=0} E\left(\frac{V(\mathcal{M}_i)}{R}\right) \leq \sum_{i:T(i)=0} \frac{\ell_i \alpha}{\ell} = \alpha. \blacksquare
$$

B.3 Proof of Theorem 3.3

We will show that (B.1) holds under arbitrary dependence for the procedure introduced in Theorem 3.3.

Proof of (B.1). When H_i is true, by the fourth inequality of (B.4), we have

$$\begin{aligned} E\left(\frac{V(\mathcal{M}_i)}{R}\right) &\leq \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{m_i}{m_i+r-1}\Pr\left(R(\mathcal{G}_{d_i})=r, P_i \leq \alpha_i(r)\right) \\ &= \frac{\ell_i \alpha}{\ell}\frac{1}{c_i}\sum_{r=d_i}^{|\mathcal{G}_{d_i}|}\frac{1}{\alpha_i(r)}\Pr\left(R(\mathcal{G}_{d_i})=r, P_i \leq \alpha_i(r)\right) \\ &\leq \frac{\ell_i \alpha}{\ell}\frac{1}{c_i}\left(1+\sum_{r=d_i+1}^{|\mathcal{G}_{d_i}|}\frac{\alpha_i(r)-\alpha_i(r-1)}{\alpha_i(r)}\right) \\ &= \frac{\ell_i \alpha}{\ell}. \end{aligned}$$

The second inequality follows by Lemma B.2. Thus, (B.1) holds when H_i is true. When H_i is false, (B.1) also holds by the same proof of induction used in the proof of Theorem 3.2. Hence, (B.1) holds for all hypotheses.

Proof of Theorem 3.3. Since (B.1) holds for each $i = 1, \ldots, m$, FDR control follows by the same argument used in the proof of Theorem 3.2. ■

B.4 Proof of Theorem 3.4

Recursively define the random variables $A_1, \ldots, A_m$ as follows:

$$A_i = \begin{cases} 1 & T(i) = 0, \\ A_{T(i)} & T(i) \neq 0 \text{ and } H_{T(i)} \text{ is false}, \\ A_{T(i)}(1-(1/R(\mathcal{G}_{d_i-1}))I\{H_{T(i)} \text{ is rejected}\} & T(i) \neq 0 \text{ and } H_{T(i)} \text{ is true}. \end{cases}$$

Notice that A_i is a function of the p-values corresponding to the hypotheses in families $\mathcal{F}_1, \ldots, \mathcal{F}_{d_i-1}$ so that P_i and A_i are independent due to Assumption 3.1.

When H_i is a true null hypothesis, we have the following useful inequality

$$E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \leq E\left(A_i\right). \tag{B.6}$$

Proof of (B.6). With $R(\mathcal{G}_0) = 0$, we have

$$\begin{aligned}
& E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) = E\left(A_i \frac{I\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \\
= \; & E\left(A_i \frac{I\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1} + R(\mathcal{F}_{d_i}))\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + R(\mathcal{F}_{d_i}))}\right) \\
\leq \; & E\left(A_i \sum_{r=1}^{|\mathcal{F}_{d_i}|} E\left(\frac{I\{P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{F}_{d_i}) = r\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)} \;\middle|\; R(\mathcal{G}_{d_i-1}), A_i\right)\right) \\
\leq \; & E\left(A_i \sum_{r=1}^{|\mathcal{F}_{d_i}|} \Pr\left(R(\mathcal{F}_{d_i}) = r \mid P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{G}_{d_i-1}), A_i\right)\right) \\
\leq \; & E\left(A_i \Pr\left(R(\mathcal{F}_{d_i}) \geq 1 \mid P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + 1), R(\mathcal{G}_{d_i-1}), A_i\right)\right) \\
\leq \; & E\left(A_i\right).
\end{aligned}$$

The first equality follows by the fact that the event $\{H_i \text{ is rejected}\}$ is the same as the event $\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))\}$. The second inequality follows by the fact that P_i is independent of $R(\mathcal{G}_{d_i-1})$ and A_i due to Assumption 3.1 so that conditional on $R(\mathcal{G}_{d_i-1})$ and A_i, P_i is still stochastically greater than or equal to $U(0,1)$. Finally, the third inequality is due to Lemma B.1.

Now, we will show that

$$E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) \leq \frac{\ell_i \alpha}{\ell} E\left(A_i\right). \tag{B.7}$$

Proof of (B.7). If H_i is a false null leaf hypothesis, then the left hand side of (B.7) is 0. If H_i is a true null leaf hypothesis, then

$$E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) \leq E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})}\right)$$

$$= \frac{\ell_i \alpha}{\ell} E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha(R(\mathcal{G}_{d_i}))}\right) \leq \frac{\ell_i \alpha}{\ell} E\left(A_i\right).$$

The first inequality follows by the fact that $R(\mathcal{G}_{d_i}) \leq R$ and $V(\mathcal{M}_i) = I\{H_i \text{ is rejected}\}$ when H_i is a true null leaf hypothesis. The equality follows by $\alpha_i(r) = \ell_i r\alpha/\ell$. The second inequality follows by (B.6). Thus, (B.7) holds when H_i is a leaf hypothesis.

Now, we will show that (B.7) holds when H_i is a non-leaf hypothesis. By induction assume (B.7) holds for all children of H_i. If H_i is false, then we note that $V(\mathcal{M}_i) = \sum_{j:T(j)=i} V(\mathcal{M}_j)$ and $A_i = A_j$ for each j such that $T(j) = i$.

$$E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) = E\left(\sum_{j:T(j)=i} A_j \frac{V(\mathcal{M}_j)}{R}\right) \leq \sum_{j:T(j)=i} \frac{\ell_j \alpha}{\ell} E\left(A_j\right) = \frac{\ell_i \alpha}{\ell} E\left(A_i\right).$$

The inequality follows by induction.

Now, assume H_i is true. We will use the following inequality.

$$\begin{aligned}\frac{1}{R} = \frac{1}{R(\mathcal{G}_{d_i})} - \frac{R - R(\mathcal{G}_{d_i})}{RR(\mathcal{G}_{d_i})} &\leq \frac{1}{R(\mathcal{G}_{d_i})} - \sum_{j:T(j)=i} \frac{R(\mathcal{M}_j)}{RR(\mathcal{G}_{d_i})} \\ &\leq \frac{1}{R(\mathcal{G}_{d_i})} - \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{RR(\mathcal{G}_{d_i})}. \end{aligned} \tag{B.8}$$

The equality follows by simple algebra and the first inequality follows by the fact that $\mathcal{M}_j \subseteq \mathcal{M}/\mathcal{G}_{d_i}$ for each j with $T(j) = i$ so that $\sum_{j:T(j)=i} R(\mathcal{M}_j) \leq R - R(\mathcal{G}_{d_i})$. It should also be noted that $V(\mathcal{M}_i) = (1 + \sum_{j:T(j)=i} V(\mathcal{M}_j)) I\{H_i \text{ is rejected}\}$.

$$\begin{aligned} & E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) \\ = \; & E\left(A_i \left(\frac{1}{R} + \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{R}\right) I\{H_i \text{ is rejected}\}\right) \\ \leq \; & E\left(A_i \left(\frac{1}{R(\mathcal{G}_{d_i})} - \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{RR(\mathcal{G}_{d_i})} + \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{R}\right) I\{H_i \text{ is rejected}\}\right) \\ = \; & E\left(A_i \left(\frac{1}{R(\mathcal{G}_{d_i})} + \left(1 - \frac{1}{R(\mathcal{G}_{d_i})}\right) \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{R}\right) I\{H_i \text{ is rejected}\}\right) \end{aligned}$$

$$
\begin{aligned}
&= E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})} + \sum_{j:T(j)=i} A_j \frac{V(\mathcal{M}_j)}{R}\right) \\
&\leq E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})} + \sum_{j:T(j)=i} \frac{\ell_j \alpha}{\ell} A_j\right) \\
&= E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})} + A_i \left(1 - \frac{1}{R(\mathcal{G}_{d_i})}\right) \sum_{j:T(j)=i} \frac{\ell_j \alpha}{\ell} I\{H_i \text{ is rejected}\}\right) \\
&= E\left(A_i \left(\frac{1}{R(\mathcal{G}_{d_i})} + \left(1 - \frac{1}{R(\mathcal{G}_{d_i})}\right) \frac{\ell_i \alpha}{\ell}\right) I\{H_i \text{ is rejected}\}\right) \\
&= \frac{\ell_i \alpha}{\ell} E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \\
&\leq \frac{\ell_i \alpha}{\ell} E\left(A_i\right).
\end{aligned}
$$

The first inequality follows by (B.8). The third and forth equality follow by the fact that $A_j = A_i(1 - (1/R(\mathcal{G}_{d_i}))I\{H_i \text{ is rejected}\}$ for j such that $T(j) = i$. The second inequality follows by induction. The last equality follows by $\alpha_i(r) = \ell_i r\alpha/(\ell + \ell_i(r-1)\alpha)$ and the last inequality follows by (B.6).

Proof of Theorem 3.4. Finally, by (B.7),

$$
\text{FDR} = E\left(\frac{V}{R}\right) = \sum_{i:T(i)=0} E\left(\frac{V(\mathcal{M}_i)}{R}\right) = \sum_{i:T(i)=0} E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) \leq \sum_{i:T(i)=0} \frac{\ell_i \alpha}{\ell} = \alpha. \blacksquare
$$

B.5 Proof of Theorem 3.5

To prove Theorem 3.5, we show that the following inequality holds when H_i is true

$$
E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))c_i}\right) \leq E\left(A_i\right). \tag{B.9}
$$

Proof of (B.9). It can be shown through simple algebra that

$$
c_i = 1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(r + d_i - 1) - \alpha_i(r + d_i - 2)}{\alpha_i(r + d_i - 1)}.
$$

It should be noted that c_i is canceled out on the right hand side of the equation. Assume H_i is true. Then,

$$
\begin{aligned}
& E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))c_i}\right) \\
= \;& \frac{1}{c_i} E\left(A_i \frac{I\{P_i \le \alpha_i(R(\mathcal{G}_{d_i})), H_{T(i)} \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \\
= \;& \frac{1}{c_i} E\left(A_i \frac{I\{P_i \le \alpha_i(R(\mathcal{G}_{d_i-1}) + R(\mathcal{F}_{d_i})), H_{T(i)} \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + R(\mathcal{F}_{d_i}))}\right) \\
= \;& \frac{1}{c_i} E\Bigg(A_i I\{H_{T(i)} \text{ is rejected}\} \times \\
& \quad E\left(\sum_{r=1}^{|\mathcal{F}_{d_i}|} \frac{I\{P_i \le \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{F}_{d_i}) = r\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)} \;\middle|\; P_j, H_j \in \mathcal{F}_1, \ldots, \mathcal{F}_{d_i-1}\right)\Bigg) \\
= \;& \frac{1}{c_i} E\Bigg(A_i I\{H_{T(i)} \text{ is rejected}\} \times \\
& \quad \sum_{r=1}^{|\mathcal{F}_{d_i}|} \frac{\Pr\left(P_i \le \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{F}_{d_i}) = r \mid P_j, H_j \in \mathcal{F}_1, \ldots, \mathcal{F}_{d_i-1}\right)}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)}\Bigg) \\
\le \;& \frac{1}{c_i} E\left(A_i I\{H_{T(i)} \text{ is rejected}\}\left(1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(R(\mathcal{G}_{d_i-1}) + r) - \alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)}\right)\right) \\
\le \;& \frac{1}{c_i} E\left(A_i I\{R(\mathcal{G}_{d_i-1}) \ge d_i - 1\}\left(1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(R(\mathcal{G}_{d_i-1}) + r) - \alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)}\right)\right) \\
\le \;& \frac{1}{c_i} E\left(A_i\left(1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(r + d_i - 1) - \alpha_i(r + d_i - 2)}{\alpha_i(r + d_i - 1)}\right)\right) \\
= \;& E\left(A_i\right).
\end{aligned}
$$

The first equality follows by the fact that the event $\{H_i \text{ is rejected}\}$ is the same as the event $\{H_{T(i)} \text{ is rejected}, P_i \le \alpha_i(R(\mathcal{G}_{d_i}))\}$. The first inequality follows by Lemma B.2 and the fact that P_i and $R(\mathcal{F}_{d_i})$ are independent of the p-values associated with the hypotheses in $\mathcal{F}_1, \ldots, \mathcal{F}_{d_i-1}$ due to Assumption 3.1. The second inequality follows by the fact that the event $\{H_{T(i)} \text{ is rejected}\}$ implies all ancestors of $H_{T(i)}$ are rejected so there must be at least $d_i - 1$ rejections in the first $d_i - 1$ families. The third inequality follows by the fact that $[\alpha_i(R(\mathcal{G}_{d_i-1}) + r) - \alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)]/\alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)$ is a decreasing function of $R(\mathcal{G}_{d_i-1})$.

Proof of Theorem 3.5. By using the same argument for the proof of (B.7), we have that $E\left(A_i V(\mathcal{M}_i)/R\right) \leq \ell_i \alpha E\left(A_i\right)/\ell$. Thus, FDR control follows by the same argument used in the proof of Theorem 3.4. ■

APPENDIX C

PROOFS OF THE THEOREMS IN CHAPTER 4

This appendix contains the proof of the Lemma 4.1 and Proposition 4.3 stated in Chapter 4.

C.1 Proof of Lemma 4.1

Suppose L takes the form $L(\alpha) = I\{\bigcap_{i=1}^{m} P_i \leq t_i(\alpha)\}$ for non-decreasing $t_i, i = 1, \ldots, m$. For $0 \leq \alpha \leq \alpha'$ and each $i = 1, \ldots, m$, define

$$P_i^{(1)} = P_i \mid P_i \leq t_i(\alpha) \text{ and } P_i^{(2)} = P_i \mid P_i \leq t_i(\alpha') \, .$$

Since t_i is a non-decreasing function, $P_i^{(1)}$ is stochastically less than or equal to $P_i^{(2)}$. By independence, $P_i^{(1)}, i = 1, \ldots, m$ are mutually independent and can be equivalently expressed as $P_i^{(1)} = P_i | \bigcap_{j=1}^{m} P_j \leq t_j(\alpha)$. Similarly, $P_i^{(2)}, i = 1, \ldots, m$ are mutually independent and can be equivalently expressed as $P_i^{(2)} = P_i | \bigcap_{j=1}^{m} P_j \leq t_j(\alpha')$. Thus, for any coordinatewise non-decreasing function of the p-values ψ, we have

$$\begin{aligned} &E\left(\psi(P_1, \ldots, P_m) \mid L(\alpha) > 0\right) = E\left(\psi(P_1^{(1)}, \ldots, P_m^{(1)})\right) \\ \leq &E\left(\psi(P_1^{(2)}, \ldots, P_m^{(2)})\right) = E\left(\psi(P_1, \ldots, P_m) \mid L(\alpha') > 0\right) . \blacksquare \end{aligned}$$

C.2 Proof of Proposition 4.3

$L(= V)$ can be split into a sum of m binary error rates $L_i = I\{H_i \text{ is falsely rejected}\}$, $i = 1, \ldots, m$ so that condition A1 is satisfied. Since H_i cannot be tested unless all of

its ancestors are rejected, we have

$$L_i(\alpha) = I\{\bigcap_{j\in\mathcal{D}_i} P_j \leq \alpha_j(\alpha)\}, i = 1, \ldots, m.$$

By Lemma 4.1, we also have that condition A3 is satisfied under independence. Finally, we will show that condition A2 is satisfied. For each $i = 1, \ldots, m$, if H_i is false, then let $c_i = 0$ and if H_i is true, then let

$$c_i = \frac{1}{\lambda\ell} \prod_{j\in\mathcal{D}_i} \left(\frac{\ell_j\lambda}{1+\ell_j\lambda}\right)^{I\{H_j \text{ is true}\}}.$$

To show that condition A2 holds, we show that the following two statements hold:

(a) $E\left(L_i(\alpha)\right) \leq c_i\alpha$ for each $i = 1, \ldots, m$,

(b) $\sum_{i=1}^{m} c_i \leq 1$.

Proof of part (a). If H_i is false, then $L_i = 0$ and clearly part (a) holds. Assume H_i is true. Then,

$$\begin{aligned} E\left(L_i(\alpha)\right) &= \Pr\left(\bigcap_{j\in\mathcal{D}_i}\{P_j \leq \alpha_j(\alpha)\}\right) \leq \Pr\left(P_i \leq \frac{\ell_i\alpha}{\ell(1+\ell_i\lambda)}, \bigcap_{\substack{j\in\mathcal{D}_i \\ j\neq i}}\{P_j \leq \frac{\ell_j\lambda}{1+\ell_j\lambda}\}\right) \\ &\leq \frac{\ell_i\alpha}{\ell(1+\ell_i\lambda)} \prod_{\substack{j\in\mathcal{D}_i \\ j\neq i}} \left(\frac{\ell_j\lambda}{1+\ell_j\lambda}\right)^{I\{H_j \text{ is true}\}} = \frac{\alpha}{\ell\lambda} \prod_{j\in\mathcal{D}_i} \left(\frac{\ell_j\lambda}{1+\ell_j\lambda}\right)^{I\{H_j \text{ is true}\}} \\ &= c_i\alpha. \end{aligned}$$

The second inequality follows by independence.

Proof of part (b). Part (b) follows by the equality and inequality below:

$$\sum_{H_i\in\mathcal{F}_1} s_{i,j} = 1, j = 1, \ldots, m, \tag{C.1}$$

$$\sum_{H_j \in \mathcal{M}_i} s_{i,j} c_j \leq \frac{\ell_i}{\ell}, i \in \mathcal{F}_1. \tag{C.2}$$

Thus, we have

$$\sum_{j=1}^{m} c_j = \sum_{j=1}^{m} \sum_{H_i \in \mathcal{F}_1} s_{i,j} c_j = \sum_{H_i \in \mathcal{F}_1} \sum_{H_j \in \mathcal{M}_i} s_{i,j} c_j \leq \sum_{H_i \in \mathcal{F}_1} \frac{\ell_i}{\ell}$$
$$= \sum_{H_i \in \mathcal{F}_1} \sum_{H_j \text{ is a leaf}} \frac{s_{i,j}}{\ell} = \sum_{H_j \text{ is a leaf}} \frac{1}{\ell} = 1.$$

The second equality follows by the fact that $s_{i,j} = 0$ if $H_j \notin \mathcal{M}_i$. The inequality follows by (C.2).

Proof of (C.1). Assume $H_j \in \mathcal{F}_1$. Then, $\sum_{i \in \mathcal{F}_1} s_{i,j} = \sum_{i \in \mathcal{F}_1} I\{i = j\} = 1$ and (C.1) holds. By induction, assume (C.1) holds for all k such that $H_k \in \mathcal{T}_j$, then

$$\sum_{H_i \in \mathcal{F}_1} s_{i,j} = \sum_{H_i \in \mathcal{F}_1} \sum_{H_k \in \mathcal{T}_j} \frac{s_{i,k}}{|\mathcal{T}_j|} = \sum_{H_k \in \mathcal{T}_j} \sum_{H_i \in \mathcal{F}_1} \frac{s_{i,k}}{|\mathcal{T}_j|} = \sum_{H_k \in \mathcal{T}_j} \frac{1}{|\mathcal{T}_j|} = 1.$$

The first equality follow by the definition of $s_{i,j}$, (4.5). The third equality follows by induction. Thus, (C.1) holds for all hypotheses.

Proof of (C.2). First, we will show that

$$s_{i,j} = \sum_{k: H_i \in \mathcal{T}_k} \frac{s_{k,j}}{|\mathcal{T}_k|}, i \neq j. \tag{C.3}$$

Suppose $H_i \in \mathcal{F}_t$ and $H_j \in \mathcal{F}_{t+1}$, then (C.3) holds since it can be seen from (4.5) that $s_{i,j} = 1/|\mathcal{T}_j|$ and from the fact that the right hand side of (C.3) is $1/|\mathcal{T}_j|$. Now, suppose $H_j \in \mathcal{F}_u, u > t+1$. By induction, assume (C.3) holds on $s_{i,k}$ for each $H_k \in \bigcup_{\ell=t+1}^{u-1} \mathcal{F}_\ell$. Hence,

$$s_{i,j} = \sum_{H_k \in \mathcal{T}_j} \frac{s_{i,k}}{|\mathcal{T}_j|} = I\{H_i \in \mathcal{T}_j\} \frac{1}{|\mathcal{T}_j|} + \sum_{\substack{H_k \in \mathcal{T}_j \\ k \neq i}} \sum_{\ell: H_i \in \mathcal{T}_\ell} \frac{s_{\ell,k}}{|\mathcal{T}_j||\mathcal{T}_\ell|}$$
$$= I\{H_i \in \mathcal{T}_j\} \frac{1}{|\mathcal{T}_j|} + \sum_{\substack{\ell: H_i \in \mathcal{T}_\ell \\ \ell \neq j}} \sum_{H_k \in \mathcal{T}_j} \frac{s_{\ell,k}}{|\mathcal{T}_j||\mathcal{T}_\ell|} = I\{H_i \in \mathcal{T}_j\} \frac{1}{|\mathcal{T}_j|} + \sum_{\substack{\ell: H_i \in \mathcal{T}_\ell \\ \ell \neq j}} \frac{s_{\ell,j}}{|\mathcal{T}_\ell|} = \sum_{\ell: H_i \in \mathcal{T}_\ell} \frac{s_{\ell,i}}{|\mathcal{T}_\ell|}.$$

The second equality follows by induction where it should be noted that $H_k \in \mathcal{T}_j$ implies H_k is in one of $\mathcal{F}_1, \ldots, \mathcal{F}_{u-1}$, but $k \neq i$ and $s_{i,k} \neq 0$ (i.e., $H_i \in \mathcal{D}_k$) implies H_k is in one of $\mathcal{F}_{t+1}, \ldots, \mathcal{F}_m$. Hence, H_k belongs to one of $\mathcal{F}_{t+1}, \ldots, \mathcal{F}_{u-1}$ so that induction is valid. The third equality follows from the fact that $s_{\ell,k} = 0$ if $k = i$ since H_i is a parent of H_ℓ and the fact that $s_{\ell,k} = 0$ if $\ell = j$ since H_k is a parent of H_j. Thus, (C.3) holds.

To show that (C.2) holds, we will show the following inequality holds

$$\sum_{H_j \in \mathcal{M}_i} s_{i,j} c_j \leq \frac{\ell_i}{\ell} \prod_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}}, i = 1, \ldots, m. \tag{C.4}$$

It should be noted that (C.4) reduces to (C.2) when $H_i \in \mathcal{F}_1$.

Suppose that H_i is a leaf hypothesis so that $\mathcal{M}_i = \{H_i\}$ and $\ell_i = 1$. If H_i is false, (C.4) holds trivially since $c_i = 0$. If H_i is true, then

$$\begin{aligned} \sum_{H_j \in \mathcal{M}_i} s_{i,j} c_j = c_i &= \frac{\ell_i}{\ell(1 + \ell_i \lambda)} \prod_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}} \\ &< \frac{\ell_i}{\ell} \prod_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}}. \end{aligned} \tag{C.5}$$

Thus, (C.4) holds for all leaf hypotheses.

Suppose H_i is not a leaf hypothesis and by induction assume (C.4) holds on the index of all child hypotheses of H_i. Then,

$$\begin{aligned} &\sum_{H_j \in \mathcal{M}_i} s_{i,j} c_j = c_i + \sum_{\substack{H_j \in \mathcal{M}_i \\ i \neq j}} s_{i,j} c_j = c_i + \sum_{H_j \in \mathcal{M}_i} \sum_{k: H_i \in \mathcal{T}_k} \frac{s_{k,j}}{|\mathcal{T}_k|} c_j \\ &= c_i + \sum_{k: H_i \in \mathcal{T}_k} \sum_{H_j \in \mathcal{M}_k} \frac{s_{k,j}}{|\mathcal{T}_k|} c_j \leq c_i + \sum_{k: H_i \in \mathcal{T}_k} \frac{1}{|\mathcal{T}_k|} \frac{\ell_k}{\ell} \prod_{\substack{j \in \mathcal{D}_k \\ j \neq k}} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}} \\ &\leq c_i + \sum_{k: H_i \in \mathcal{T}_k} \frac{1}{|\mathcal{T}_k|} \frac{\ell_k}{\ell} \prod_{j \in \mathcal{D}_i} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}}. \end{aligned} \tag{C.6}$$

The second equality follows from (C.3). The third equality follows from the fact that $s_{k,j} = 0$ if $H_j \notin \mathcal{M}_k$. The first inequality follows by induction and the second by the fact that $\mathcal{D}_i \subseteq \mathcal{D}_k / H_k$ for $H_i \in \mathcal{T}_k$.

Using the following argument

$$\sum_{k:H_i \in \mathcal{T}_k} \frac{\ell_k}{|\mathcal{T}_k|} = \sum_{k:H_i \in \mathcal{T}_k} \sum_{H_j \text{ is a leaf}} \frac{s_{k,j}}{|\mathcal{T}_k|} = \sum_{H_j \text{ is a leaf}} \sum_{k:H_i \in \mathcal{T}_k} \frac{s_{k,j}}{|\mathcal{T}_k|} = \sum_{H_j \text{ is a leaf}} s_{i,j} = \ell_i,$$

the right hand side of (C.6) can be simplified to

$$c_i + \frac{\ell_i}{\ell} \prod_{j \in \mathcal{D}_i} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}}. \tag{C.7}$$

If H_i is false, then $c_i = 0$ and by (C.7), it is easy to see that (C.2) holds. If H_i is true, then from (C.7), the right hand side of (C.6) can be simplified to

$$\begin{aligned}
&\frac{1}{\lambda \ell} \prod_{j \in \mathcal{D}_i} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}} + \frac{\ell_i}{\ell} \prod_{j \in \mathcal{D}_i} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}} \\
=& \left(\frac{1}{\lambda \ell} + \frac{\ell_i}{\ell} \right) \prod_{j \in \mathcal{D}_i} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}} \\
=& \frac{\ell_i}{\ell} \prod_{\substack{j \in \mathcal{D}_i \\ j \neq i}} \left(\frac{\ell_j \lambda}{1 + \ell_j \lambda} \right)^{I\{H_j \text{ is true}\}}.
\end{aligned}$$

(C.4) holds when H_i is a non-leaf true or false hypothesis; hence, (C.4) holds for all $i = 1, \ldots, m$. ■

BIBLIOGRAPHY

Bauer, P., Röhmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17:2133–2146.

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *J. Amer. Satist. Assoc.*, 102(480):1272–1281.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57:289–300.

Benjamini, Y., Krieger, A., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82:163–170.

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

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992.

Brechenmacher, T., Xu, J., Dmitrienko, A., and Tamhane, A. (2011). A mixture gatekeeping procedure based on the Hommel test for clinical trial applications. *Journal of Biopharmaceutical Statistics*, 21(4):748–767.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2011). Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *PNAS*, 108 Suppl 1:4516–4522.

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (2009). *Introduction to Algorithms*. The MIT Press, Cambridge, MA, 3rd edition.

Dmitrienko, A., D'Agostino, R., and Huque, M. (2013). Key multiplicity issues in clinical drug development. *Statistics in Medicine*, 32(7):1079–1111.

Dmitrienko, A., Offen, W., Wang, O., and Xiao, D. (2006). Gatekeeping procedures in dose-response clinical trials based on the Dunnet test. *Pharmaceutical Statistics*, 5:19–28.

Dmitrienko, A., Offen, W., and Westfall, P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine*, 22:2387–2400.

Dmitrienko, A. and Tamhane, A. (2013). General theory of mixture procedures for gatekeeping. *Biometrical Journal*, 55(3):402–419.

Dmitrienko, A., Tamhane, A., Liu, L., and Wiens, B. (2008). A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine*, 27:3446–3451.

Dmitrienko, A., Wiens, B., Tamhane, A., and Wang, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine*, 26:2465–2478.

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

Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Satist. Assoc.*, 107:1019–1034.

Farcomeni, A. and Finos, L. (2013). FDR control with pseudo-gatekeeping based on a possibly data driven order of the hypotheses. *Biometrics*, 69(3):606–613.

Finos, L. and Farcomeni, A. (2011). k-FWER control without multiplicity correction, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics*, 67:174–181.

Gavrilov, Y., Benjamini, Y., and Sarkar, S. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.*, 37(2):619–629.

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

Goeman, J. and Finos, L. (2012). The inheritance procedure: Multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology*, 11(1):1–18.

Goeman, J. and Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24(4):537–544.

Goeman, J., van de Geer, S., de Kort, F., and van Houwelingen, H. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, 20:93–99.

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomeld, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Gordon, A., Glazko, G., Qiu, X., and Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, 1:179–190.

Guo, W., He, L., and Sarkar, S. (2013). Further results on controlling the false discovery proportion. Working paper.

Guo, W. and Rao, M. (2008). On control of the false discovery rate under no assumption of dependency. *Journal of Statistical Planning and Inference*, 28:3176–3188.

Guo, W. and Sarkar, S. (2013). Adaptive controls of FWER and FDR under block dependence. Working paper, available from `http://web.njit.edu/~wguo/Guo\%20&\%20Sarkar\%202012.pdf` (accessed on 12/6/2013).

Guo, W., Sarkar, S., and Peddada, S. (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics*, 66(2):485–492.

Heller, R., Manduchi, E., Grant, G., and Ewens, W. (2009). A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics*, 25(8):929–942.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York, NY.

Holland, B. and Copenhaver, M. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, 43:417–423.

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

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

Hommel, G. and Kropf, S. (2005). Testing for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal*, 47:554–562.

Huque, M. and Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference*, 138(2):321–335.

Joag Dev, K. and Proschan, F. (1983). Negative association of random variables, with applications. *Ann. Statist.*, 11:286–295.

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

Kropf, S. and Läuter, J. (2002). Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometrical Journal*, 44:789–800.

Kropf, S., Läuter, J., Eszlinger, M., Krohn, K., and Paschkeb, R. (2004). Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *Journal of Statistical Planning and Inference*, 125(1-2):31–47.

Lehmann, E. and Romano, J. (2005a). *Testing Statistical Hypotheses*. Springer, New York, NY.

Lehmann, E. L. and Romano, J. (2005b). Generalizations of the familywise error rate. *Ann. Statist.*, 33:1138–1154.

Maurer, W., Hothorn, L., and Lehmacher, W. (1995). *Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses*, volume 6. Fischer-Verlag, Stuttgart, Germany.

Mehrotra, D. and Heyse, J. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, 13:227–238.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278.

Qui, Z., Guo, W., and Lynch, G. (2013). On generalized fixed sequence procedures for controlling the FWER. Working paper.

Romano, J. and Shaikh, A. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.*, 34:1850–1873.

Romano, J., Shaikh, A., and Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST*, 17(3):417–442.

Ross, G. J., Tasoulis, D., and Adams, N. (2011). Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4):379–389.

Sankaran, K. and Holmes, S. (2013). structSSI: Simultaneous and selective inference for hierarchically structured data. Available from: `http://www.stanford.edu/%7Ekriss1/` (accessed on 12/6/2013).

Sarkar, S. (1998). Some probability inequalities for ordered MTP_2 random variables: A proof of the Simes conjecture. *Ann. Statist.*, 26:494–504.

Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30:239–257.

Sarkar, S. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *Ann. Statist.*, 35(6):2405–2420.

Sarkar, S. (2008). On methods controlling the false discovery rate. *Sankyā*, 70A(2):135–168.

Sarkar, S. and Chang, C. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Satist. Assoc.*, 92(440):1601–1608.

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

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

Storey, J. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035.

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

Tamhane, A., Liu, W., and Dunnet, C. (1998). A generalized step-up-down multiple test procedure. *The Candian Journal of Statistics*, 26(2):353–363.

van't Wout, A., Lehrma, G., Mikheeva, S., OKeeffe, G., Katze, M., Bumgarner, R., Geiss, G., and Mullins, J. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of cd4(+)-t-cell lines. *Journal of Virology*, 77(2):1392–1402.

Westfall, P. and Kirshen, A. (2001). Optimally weighted, fixed sequence and gate-keeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99:25–41.

Westfall, P., Kropf, S., and Finos, L. (2004). Weighted FWE-controlling methods in high-dimensional situations. In Benjamini, Y., Bretz, F., and Sarkar, S., editors, *Recent Developments in Multiple Comparison Procedures*, volume 47, pages 143–154. Institute of Mathematical Statistics, Beachwood, OH.

Wiens, B. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2:211–215.

Wiens, B. and Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *J. Biopharm. Stat.*, 15(6):929–942.

Yekutieli, D. (2008a). False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference*, 138(2):405–415.

Yekutieli, D. (2008b). Hierarchical false discovery rate-control methodology. *J. Amer. Satist. Assoc.*, 103:309–316.

Yekutieli, D., Reiner-Benaim, A., and Benjamini, Y. (2006). Approaches to multiplicity issues in complex research in microarray analysis. *Statistica Neerlandica*, 60(4):414–437.