MULTIPLE HYPOTHESIS TESTING PROCEDURES WITH APPLICATIONS TO EPIDEMIOLOGIC STUDIES

by

Conghui Qu
B.Sc., Peking University, 2007

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Fall 2009

APPROVAL

Name: Conghui Qu

Degree: Master of Science

Title of project: Multiple Hypothesis Testing Procedures with Applications to Epidemiologic Studies

Examining Committee: Dr.Derek Bingham
Chair

____

Dr.Jinko Graham
Senior Supervisor
Simon Fraser University

____

Dr.John J. Spinelli
Supervisor
BC Cancer Agency and Simon Fraser University

____

Dr.Brad McNeney
External Examiner
Simon Fraser University

Date Approved: NOV 11 2009

Declaration of Partial Copyright Licence

Abstract

Epidemiologic and genetic studies often involve the testing of a large number of hypotheses with test statistics that are potentially dependent. In this project, we investigate multiple testing procedures to control the family-wise error rate and false discovery rate. We consider several classic and novel multiple hypothesis testing procedures. Furthermore, we compare the results of the procedures which take advantage of the dependent structure among test statistics to those of the procedures which do not. The data we used is from a case-control study of non-Hodgkin Lymphoma.

Keywords: multiple testing procedures; dependence; family-wise error rate; false discovery rate; adaptive procedures; adjusted p-values

Dedication

To my parents

Acknowledgments

First of all, I would like to express my deep gratitude to my supervisor Dr. Jinko Graham and co-supervisor Dr. John Spinelli for their continual guidance, suggestions and encouragement during my graduate study in Simon Fraser University. I really appreciate their dedication and patience. Their great help have made this project possible.

I am very grateful to Dr. Brad McNeney, the external examiner on my committee, for giving me valuable suggestions for both my thesis and research work. I also want to thank all the faculty members in the department for teaching and helping me; and thank all the staffs for offering me help.

Special thanks to Ji-Hyung Shin for her dedicated help from the first day I came to Vancouver. It is not possible to thank her enough. Thanks to my fellow graduate students in the department for their friendship and help, especially Kelly Burkett, Jingyu Chen, Joslin Goh, Carolyn Huston, Qifeng Jiang, Luyao Lin, Suli Ma, Zhong Wan, Vivien Wong, Donghong Wu, Huanhuan Wu and Ting Zhang. Thanks to all my friends in China and the US for their encouragement and being my listeners.

Last but not least, I want to thank my beloved parents for their everlasting love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The development of technology enables modern epidemiologic studies to measure large numbers of exposures (e.g. genetic data of the subjects, environmental risk factors). In order to obtain the associations between the exposures and disease, researchers need to test many hypotheses simultaneously. Therefore, a multiple testing procedure should be applied to control the Type I error rate over all the hypotheses tested.

During the last few decades, many multiple testing procedures have been developed and improved, and this issue has become more important recently because more and more scientific areas need multiple hypothesis testing control. Moreover, powerful procedures taking advantage of the dependence structure among the test statistics are desirable because exposure variables are frequently dependent and, therefore, so are their test statistics. Initially, procedures were developed for independent test statistics. Later, it was proven that several of these procedures also control the Type I error rate under positive dependence structure. More recently, multiple testing procedures were developed that take into account general dependence among the test statistics. This project aims to investigate selected multiple testing procedures, compare their powers, and then suggest procedures which can be applied to future studies.

1.2 Outline

This project is organized as follows. In Chapter 2, we introduce the background information for the case-control study of Spinelli *et al.* (2007) on which the project is based. We also list some basic results of the sub-study data from which our analysis data were extracted. Chapter 3 begins with the fitted model used to analyze the data. In Section 3.2, we give the basic definitions which are involved in the multiple testing procedures, and then we describe the multiple testing procedures controlling different Type I error rates (family-wise error rate and false discovery rate) that will be applied to the data. Some other procedures which we do not apply are also reviewed. An example is given in Section 3.3 to illustrate the definitions and how the procedures work. In Chapter 4, the multiple testing procedures described in Chapter 3 are applied, and results are provided in order to compare these procedures. Finally, in Chapter 5, we summarize our results, and discuss possible directions for future work.

Chapter 2

Data

The data used in the project is from an organochlorine analysis for a case-control study of non-Hodgkin lymphoma (NHL) undertaken by Spinelli *et al.* (2007).

2.1 Background

In order to have a better idea of the data we are using, we will introduce some background knowledge about NHL and organochlorines.

2.1.1 Non-Hodgkin lymphomas

Lymphoma is a cancer that originates in the lymphocytes, a type of white blood cell of the immune system. There are two classifications of lymphomas: Hodgkin lymphoma and non-Hodgkin lymphoma. The indicator of Hodgkin lymphoma, which comprises 10% of lymphoma, is the presence of Reed-Sternberg cells (Fisher and Fisher, 2004). NHL has many subclassifications depending on morphology, immunophenotype, and somatic genetics. The incidence rate of NHL has doubled over the last two decades, and it is the fifth most common cancer in Canada according to Canadian Cancer Statistics (2009).

The risk of NHL increases exponentially with age (Fisher and Fisher, 2004), and it is more common in men than women. The incidence rates of NHL are low in Asian and African countries, and high in North America and Australia. In the United States, Caucasians have higher incidence rate than African-Americans. Immune suppression is the major known risk factor for NHL.

2.1.2 Organochlorines

An organochlorine is an organic compound containing at least one covalently bonded chlorine atom. Organochlorines are lipophilic (dissolve in fat), stable and degradation resistant. Since organochlorines have these characteristics, they will accumulate in the fatty tissues of humans and therefore may affect human health. Organochlorines are often used in chemical processes in agriculture and industry. Although most of them were banned in the 1970s due to harm to the environment and health, they still exist in the environment and human bodies.

Generally, organochlorines can be classified into two groups: pesticides and non-pesticides. Pesticide organochlorines were widely used in agriculture. Polychlorinated biphenyls (PCBs), non-pesticide organochlorines with similar chemical structures and properties, were widely used as coolants and insulating fluids. There are 209 possible congeners which are numbered from 1 to 209 (Ng, 2007). Dioxins (polychlorinated dibenzo-para-dioxins (PCDDs)) have similar properties to PCBs, so the PCB congeners which have structures similar to dioxins are grouped together as dioxin-like PCBs. Since the organochlorines' descriptions are highly technical, we will not introduce them in detail. Readers are referred to Ng (2007) for the full description. Spinelli *et al.* (2007) considered eight pesticide or pesticide metabolites: β-HCCH; *cis*-Nonachlor; p,p'-DDE; p,p'-DDT; HCB; mirex; oxychlordane and *trans*-Nonachlor; and eleven PCB congeners with numbers 28, 99, 105, 118, 138, 153, 156, 170, 180, 183, 187.

2.2 Study

In the study, cases included subjects with newly diagnosed NHL, aged 20-79, diagnosed between March 2000 and February 2004, from greater Vancouver or Victoria (Canada) without evidence of HIV infection. Population controls were randomly selected from the Client Registry of the BC Ministry of Health and frequency matched to cases by sex, age and residential location in an approximate 1:1 ratio. There were 828 cases and 848 controls who participated in the study. All subjects were requested to answer a questionnaire and to provide a blood or saliva sample. Demographic characteristics, sunlight exposure, medical history and other information were included in the questionnaire.

In this project, we analyzed the organochlorine data from the study. Organochlorines were only measured in cases who provided blood before chemotherapy and who had weight loss information and under 10% weight loss in the year prior to blood collection. This is because chemotherapy treatment and weight loss have been found to significantly affect plasma levels of organochlorines (Baris *et al.*, 2000; Chevrier *et al.*, 2000). Frequency matched controls were chosen for the organochlorine analysis from the parent study in an approximate 1:1 ratio. Therefore, the total number of subjects used in the organochlorine analysis is 881, where 422 of them are cases and 459 are controls. The nineteen specific organochlorine analytes (measured PCB congeners and pesticides) and three variables representing sums of PCB congeners (total summed PCBs, total dioxin-like summed PCBs and total non dioxin-like summed PCBs) were examined. All the sums of PCB congeners, two dioxin-like PCBs – PCB 118 and PCB 156; five non dioxin-like PCBs – PCB 138, PCB 153, PCB 170, PCB 180 and PCB 187; and six pesticides – β-HCCH; p, p'-DDE; HCB; mirex; oxychlordane and *trans*-Nonachlor were significantly associated with NHL (Spinelli *et al.*, 2007).

This project further examines the associations between all twenty-two organochlorine analytes and NHL. Since we are testing multiple hypotheses simultaneously, we investigate multiple testing procedures to control the overall Type I error rate. The reason for choosing this study dataset is as follows. From the paper of Spinelli *et al.* (2007), we know that the organochlorine analytes are highly correlated. Thus the test statistics will be dependent. The power of multiple testing procedures for dependent tests has been an important issue in recent years. We can use these data to investigate these procedures. In the next chapters, we will take a closer view of the various multiple testing procedures, and compare their results on this dataset.

Chapter 3

Methods

3.1 Analysis Methods

Unconditional logistic regression is used to model the association between organochlorine exposure and the likelihood of NHL. The organochlorine variables are categorized into four groups based on the quartiles of their distribution, and then recoded as ordinal variables with values set as the medians of the quartiles in controls. (Some organochlorines with more than 25% subjects below the detection limit have three or two categories instead.) Each recoded organochlorine variable is analyzed as a continuous variable in a separate logistic regression model with case/control status as the outcome. Significance of individual regression estimates is tested by $Wald$ statistics. The same confounding variables are used in all the logistic regression models. They are age, sex, region, ethnicity, education level, family history of NHL, BMI one year before study participation and farming. Interaction terms between organochlorine exposure and confounders were not found to be statistically significant, so they are not included in the model.

Since there are 22 organochlorine exposures examined in this study, we are testing 22 hypotheses simultaneously. Therefore, we need to use a multiple testing procedure to control the overall Type I error rate.

3.2 Multiple testing procedures

Hypothesis testing is an approach using observed data to decide properties of the unknown data generating distribution. Often, one or more null hypotheses are stated that restrict or simplify the form of this data generating distribution. People decide which of these null hypotheses should be rejected by calculating test statistics (e.g., t-statistics, likelihood ratio statistics, etc.) from the observed data and applying a procedure based on the test statistics. If one simultaneously tests $m > 1$ null hypotheses, we call this is a multiple testing situation. According to Dudoit and van der Laan (2008, page 9), “a multiple testing procedure (MTP) is a data-dependent set of rejected hypotheses that estimates the set of false null hypotheses”. Our development of ideas and notation is based on their treatment. Procedures we consider are primarily from their book and from the review of Farcomeni (2008).

3.2.1 Basic Definitions

Type I Error Rates

We will focus on the two main types of errors in testing problems: Type I error and Type II error. Dudoit and van der Laan (2008) also refer to Type III error which occurs for two-sided tests when a false null hypothesis is correctly rejected but an incorrect conclusion is reached about the direction of the departure from the null hypothesis. We do not consider Type III errors in this project. A Type I error, also known as a false positive, occurs when a hypothesis test incorrectly rejects a true null hypothesis. A Type II error, or false negative, occurs when a hypothesis test incorrectly accepts a false null hypothesis. Unfortunately, one cannot minimize the two types of errors at the same time. Therefore, we aim to control the Type I error rate. By control, we mean that the Type I error rate will be less than or equal to some user-defined upper bound, which we refer to as the level α.

There are many possible definitions for the Type I error rate when testing a family of multiple hypotheses. We will focus on two of them: the family-wise error rate (FWER) and the false discovery rate (FDR). To help illustrate the ideas, we will rely on the following summary table of multiple testing decisions for m null hypotheses (Table 3.1). For the table, we use test statistics $T_1, \ldots, T_m$ to decide whether or not to reject null hypotheses $H_{01}, \ldots, H_{0m}$. The rules for rejecting depend on the MTP, the Type I error rate being controlled and the level at which we are aiming to control the Type I error rate. The m_0 null hypotheses that are true are in the set $\mathcal{H}_0$ and the m_1 null hypotheses that are false are in $\mathcal{H}_1$. The numbers of true and false hypotheses, m_0 and m_1, are non-random but unknown. The counts V, S, and R are data-dependent and therefore random.

Table 3.1: Summary of multiple testing decisions.

	Null hypotheses		
	not rejected	rejected	Total
True null hypothesis $\mathcal{H}_0$	$m_0 - V$	V	m_0
False null hypothesis $\mathcal{H}_1$	$m_1 - S$	S	m_1
Total	$m - R$	R	m

The FWER is the probability of making one or more Type I errors in a MTP; i.e., referring to the notation defined in Table 3.1,

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

The FDR, introduced by Benjamini and Hochberg (1995), is the expected proportion of false discoveries in all the rejected hypotheses; i.e.,

$$\text{FDR} = E(\frac{V}{R}),$$

where $\frac{0}{0}$ is defined as 0. Therefore,

$$\begin{aligned}
\text{FDR} &= E(\frac{V}{R}) \\
&= E(\frac{V}{R}|R > 0)Pr(R > 0) + E(\frac{V}{R}|R = 0)Pr(R = 0) \\
&= E(\frac{V}{R}|R > 0)Pr(R > 0) + E(\frac{0}{0}|R = 0)Pr(R = 0)
\end{aligned}$$

$$= \quad E(\frac{V}{R}|R > 0)Pr(R > 0) \tag{3.1}$$

In the second term on the third line, we use the fact that, when $R = 0$, $V \leq R$ is identically 0. We will often use the alternate definition of FDR given by (3.1).

When $m_0 = m$, FWER = FDR. Because all the m null hypotheses are true, all the rejected hypotheses are true null hypotheses. Thus, $V = R$ and $\frac{V}{R} = 1$ so that

$$\begin{aligned} \text{FDR} &= E(\frac{V}{R}|R > 0)Pr(R > 0) \\ &= E(1|R > 0)Pr(R > 0) \\ &= Pr(R > 0) \\ &= Pr(V > 0) = \text{FWER}. \end{aligned}$$

Procedures controlling FWER also control FDR. To see why, note that

$$\frac{V}{R} \leq I(V > 0)$$

where I is the indicator function. When $V = 0$, $\frac{V}{R} = I(V > 0) = 0$; when $V > 0$, $\frac{V}{R} \leq I(V > 0) = 1$ because $V \leq R$. Then, taking the expectation on both sides of the inequality, we obtain

$$E(\frac{V}{R}) \leq E(I(V > 0)) = Pr(V > 0).$$

Thus, FDR $\leq$ FWER and so procedures controlling FWER also control FDR.

P-values

Dudoit and van der Laan (2008, page 27) define the unadjusted p-value for the null hypothesis H_{0j} as:

$$p_j = \inf\{\alpha \in [0, 1] : H_{0j} \text{ is rejected at single test level } \alpha\}, j = 1, \ldots, m,$$

where the level α is an upper bound for any Type I error rate. The unadjusted p-value is the smallest value of the level of the Type I error rate for the single hypothesis testing procedure at which H_{0j} would be rejected given the observed value of its test statistic. For a single test ($m = 1$), the unadjusted p-values for FWER and FDR are the same. To see this, note that, since $V \in \{0, 1\}$,

$$\begin{aligned} \text{FWER} &= Pr(V > 0) \\ &= Pr(V = 1) \\ &= Pr(V = 1) * 1 + Pr(V = 0) * 0 \\ &= E(V), \end{aligned}$$

and that, by equation (3.1),

$$\begin{aligned} \text{FDR} &= E(\frac{V}{R}|R > 0)Pr(R > 0) \\ &= E(V|R = 1)Pr(R = 1) \\ &= E(V|R = 1)Pr(R = 1) + E(V|R = 0)Pr(R = 0) \\ &= E(V). \end{aligned}$$

On the third line of the equation for the FDR, $E(V|R = 0)Pr(R = 0) = 0$, because $0 \leq V \leq R$ by definition so that $E(V|R = 0) = 0$. Therefore, FWER = FDR.

In MTPs, people often use the adjusted p-values to present the results. For any Type I error rate, Dudoit and van der Laan (2008, page 32) define the adjusted p-value for the null hypothesis H_{0j} as:

$$p_j^* = \inf\{\alpha \in [0, 1] : H_{0j} \text{ is rejected at MTP level } \alpha\}, j = 1, \ldots, m,$$

where the MTP level is an upper bound for any MTP Type I error rate. Thus, for an MTP, the adjusted p-value for a hypothesis H_{0j}, $j = 1, \ldots, m$ is the smallest level of the Type I error rate for the multiple test of all m hypotheses at which the hypothesis H_{0j} would be rejected given the observed values of the test statistics; if no such level exists, the adjusted p-value is 1. When $m = 1$, $p_j^* = p_j$, the unadjusted p-value.

FWER-controlling MTPs which reduce to a single test with exact control of the type I error rate have adjusted p-values that are greater than or equal to the unadjusted p-values.

To see why, let α_u be the level of the test of the single hypothesis H_{0i}, A_i be the event that H_{0i} is rejected, and $H_{0i} \in \mathcal{H}_0$. Under exact control of the type I error rate for the single test of H_{0i}, we have $\alpha_u = Pr(A_i)$ for any i. Thus,

$$Pr(V > 0) = Pr(\bigcup_{H_{0j}} A_j) \geq Pr(A_i) = \alpha_u,$$

since $\bigcup_{H_{0j}} A_j \supset A_i$ for any i. Let α_M be the level of the FWER-controlling multiple testing procedure, then we have $Pr(V > 0) \leq \alpha_M$. Therefore, $\alpha_M \geq Pr(V > 0) \geq Pr(A_i) = \alpha_u$. It follows that the FWER adjusted p-value for a hypothesis will be at least as big as the unadjusted p-value. In general, however, the adjusted p-value need not be bigger than the unadjusted p-value. This will be illustrated in Section 3.3.

The advantages of using adjusted p-values are:

- The Type I error level does not to be chosen in advance.
- The adjusted p-values reflect the strength of evidence against a hypothesis.
- Different MTPs controlling the same Type I error rate can be conveniently compared by comparing their adjusted p-values.

Types of Multiple Testing Procedures

One way to categorize a MTP is as a single-step or stepwise procedure. In single-step procedures, each null hypothesis is tested independently, and the outcome is independent of those of other hypothesis tests. By contrast, in stepwise procedures, the decision of rejecting (or not) one null hypothesis depends on the results of other hypothesis tests.

Stepwise MTPs are of two types: step-down and step-up, depending on the order in which null hypotheses are tested. In step-down procedures, the hypothesis test with the smallest unadjusted p-value is examined first, and then the hypothesis tests with larger unadjusted p-values are examined successively, depending on the outcome of the previous tests. Once one test accepts a null hypothesis, no null hypotheses whose unadjusted p-values are greater than that of this hypothesis are rejected. In contrast, for step-up procedures, the hypothesis test with largest unadjusted p-value is examined first. Once one null hypothesis is rejected, all other null hypotheses whose unadjusted p-values are less than that of this hypothesis are rejected.

A second way to group a MTP is as a marginal or a joint procedure. Marginal multiple testing procedures are based on the marginal distribution of the test statistics, while joint procedures take into account the dependence structure among the test statistics. Joint multiple testing procedures tend to be more powerful than the marginal ones, because the joint distribution of the test statistics contains more information, i.e. the dependence structure, than the marginal distribution.

Finally, a third way to group a MTP is as an adaptive or a non-adaptive procedure. Non-adaptive MTPs conservatively take $m_0 = m$, while adaptive MTPs use an estimate of m_0 to revise a non-adaptive MTP so that it is less conservative. Adaptive MTPs can take advantage of small m_0.

3.2.2 Procedures Controlling FWER

To control the FWER, we apply the single-step Bonferroni Agresti and Franklin (2007) and the step-down Holm (1979) procedures. Both of these MTPs are marginal procedures. We also apply several resampling-based procedures that take advantage of the positive dependence among test statistics. They include the Dudoit and van der Laan (2008) bootstrap-based maxT/minP procedures, in both a single-step and step-down version, and also the permutation-based step-down minP procedure of Westfall and Young (1993).

Table 3.2 summarizes the FWER-controlling procedures applied.

Bonferroni

The Bonferroni single-step procedure is the best-known classical procedure for FWER control. It controls the FWER for arbitrary test statistics joint null distribution; i.e. under

Table 3.2: Summary of FWER-controlling procedures.*

	marginal	joint
single-step	Bonferroni	Dudoit and van der Laan maxT/minP single-step
step-down	Holm	Dudoit and van der Laan maxT/minP step-down Westfall and Young minP

* None of the procedures are adaptive.

arbitrary dependence (Dudoit and van der Laan, 2008). If one wants to control the FWER at level α, the common p-value "cut-off" is $\frac{\alpha}{m}$. The value of $\frac{\alpha}{m}$ is a cut-off in the sense that if p_i, the unadjusted p-value for H_{0i}, satisfies $p_i < \frac{\alpha}{m}$, reject H_{0i}. The rationale for the cut-off is as follows. Using the notation above,

$$\begin{aligned} Pr(V > 0) &= Pr(\bigcup_{H_{0j} \in \mathcal{H}_0} A_j) \leq \sum_{H_{0j} \in \mathcal{H}_0} Pr(A_j) \\ &\leq m_0 \times \max_{H_{0j} \in \mathcal{H}_0} Pr(A_j) \\ &\leq m \times \max_{H_{0j} \in \mathcal{H}_0} Pr(A_j) \\ &\leq m \times \max_{H_{0j} \in \mathcal{H}_0} (\alpha_j), \end{aligned}$$

where the first line of the equation uses Boole's Inequality and α_j is the single-test level for H_{0j}; and m_0 and m were defined in Table 3.1. Thus, choosing $\alpha = m \times \max_{\{H_{0j} \in \mathcal{H}_0\}}(\alpha_j)$ ensures that $Pr(V > 0) \leq \alpha$. For example, given α, we may set $\alpha_i = \frac{\alpha}{m}$ for $i = 1, \ldots, m$. The Bonferroni procedure's cut-off for any unadjusted p-value is therefore $\frac{\alpha}{m}$; i.e.

- If $p_i \leq \frac{\alpha}{m}$, reject H_{0i}.
- Else accept H_{0i}.

The adjusted p-value for a hypothesis is the smallest level α for the whole testing procedure at which the hypothesis would be rejected given the observed values of the test statistics. We would reject H_{0j} if $p_j \leq \frac{\alpha}{m}$ or if $\alpha \geq mp_j$. Hence, the adjusted p-value for H_{0j} is mp_j, provided that $mp_j \leq 1$; otherwise, the adjusted p-value is 1. Thus, the equation for the adjusted p-value for H_{0j} is

$$_{\text{Bonf}}P_j^* = \min\{mp_j, 1\}.$$

Holm

The Holm step-down procedure attempts to improve the power of the Bonferroni procedure based on the following reasoning. For m tests of hypotheses, let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ be the unadjusted p-values, ordered from smallest to largest, and let $H_{0(i)}, i = 1, \ldots, m$, be the corresponding hypotheses. If we reject $H_{0(1)}$ using the Bonferroni critical value $\frac{\alpha}{m}$, we have only $m - 1$ further hypotheses to test, so that the critical value for $H_{0(2)}$ should be $\frac{\alpha}{m-1}$ and so forth. Thus, the Holm procedure uses different cut-offs for the ordered unadjusted p-values: $\alpha_j = \frac{\alpha}{m-j+1}$, for $j = 1, \ldots, m$. For $m > 1$, the Holm critical values for the unadjusted p-values are greater than the Bonferroni critical value. Like the Bonferroni procedure, the Holm procedure controls the FWER for arbitrary test statistics joint null distribution (Dudoit and van der Laan, 2008). Adjusted p-values can be calculated as follows.

Because this is a step-down procedure, to control FWER at level α, the steps are: For $i = 1, \ldots, m$,

- If $p_{(i)} \leq \alpha_i = \frac{\alpha}{m-i+1}$, reject $H_{0(i)}$ and continue.
- Else accept $H_{0(i)}, \ldots, H_{0(m)}$ and stop.

In order to get the adjusted p-value for $H_{0(j)}$, $j = 1, \ldots, m$, we reason as follows. The adjusted p-value for $H_{0(j)}$ is the smallest FWER level at which the hypothesis would be rejected given the observed values of the test statistics. But, before we reject $H_{0(j)}$, we must reject $H_{0(1)}, \ldots, H_{0(j-1)}$ in this step-down procedure. Hence, we have the inequalities:

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

The MTP level α satisfies all the following j inequalities provided that the numbers on the right-hand side are ≤ 1:

$$\alpha \geq (m - i + 1)p_{(i)}, \quad i = 1, \ldots, j.$$

That is,

$$\alpha \geq \max_{i \in \{1,\ldots,j\}} \{(m - i + 1)p_{(i)}\},$$

provided the maximum is ≤ 1; otherwise it is 1. Hence, the adjusted p-values for $H_{0(j)}$ can be written as:

$$_{\text{Holm}}P^*_{(j)} = \min\{\max_{i \in \{1,\ldots,j\}} \{p_{(i)} \times (m - i + 1)\}, 1\}, \quad j = 1, \ldots, m.$$

Dudoit and van der Laan single-step maxT/minP

The Dudoit and van der Laan maxT and minP procedures in both single-step and step-down versions are all bootstrap-based procedures. These procedures take into account the dependence structure of the test statistics, and control FWER for arbitrary test statistic joint-null distributions (Dudoit and van der Laan, 2008, page 118). Therefore, they are joint multiple testing procedures.

To obtain the appropriate null distribution, we take the following steps. First, bootstrap B samples with replacement from cases and controls separately, such that each sample has the same number of cases and controls as in the observed data. Controls are sampled from the control covariate vectors and cases are sampled from the case covariate vectors. Second, for each of the B samples, test m hypotheses simultaneously to get m test statistics. Then we have a $B \times m$ matrix of test statistics $\mathbf{T}$ which can be used to obtain a bootstrap estimate of the joint distribution of the test statistics. The bootstrap method preserves the dependency structure among test statistics, but does not remove the systematic effects (Farcomeni, 2008). Here systematic effects are the departures from means of zero in the distributions of the test statistics that are brought about by the alternative hypotheses.

To obtain a bootstrap estimate of the joint-null distribution, one may center the test statistics to get a $B \times m$ matrix $\mathbf{Z}$ such that

$$\mathbf{Z}[i,j] = |\mathbf{T}[i,j] - E(\mathbf{T}[j])|, \quad i = 1, \ldots, B, j = 1, \ldots, m,$$

where $\mathbf{Z}[i,j]$ and $\mathbf{T}[i,j]$ are the $(i,j)^{\text{th}}$ entry of $\mathbf{Z}$ and $\mathbf{T}$, respectively, and $E(\mathbf{T}[j])$ is the bootstrap mean of the j^{th} column of $\mathbf{T}$. Dudoit and van der Laan (2008) suggest dividing $\mathbf{Z}[i,j]$ by $sd(\mathbf{T}[j])$, the bootstrap standard deviation of the j^{th} column of $\mathbf{T}$. Under a correctly specified model, dividing the bootstrapped test statistics by their bootstrap standard deviation should not impact their asymptotic variance of 1. However, when we have missing covariates (confounders), the asymptotic variance of test statistics can be greater than 1 (Efron, 2007). Thus, dividing by the standard deviation can lead to a null reference distribution that is too narrow and to false positive results.

In the single-step maxT/minP procedure, let q_α be the $(1-\alpha)^{\text{th}}$ quantile of the null distribution of $\max(|T_1|, |T_2|, \ldots, |T_m|)$ and q'_α be the α^{th} quantile of the null distribution of $\min(p_1, p_2, \ldots, p_m)$, where T_1, T_2, $\ldots$, T_m are test statistics with the same marginal null distribution and p_1, p_2, $\ldots$, p_m are their unadjusted p-values. We reject H_{0i} if $|T_i| \geq q_\alpha$ or equivalently, if $p_i \leq q'_\alpha$.

To obtain the adjusted p-values p_i^*, $i = 1, \ldots, m$, we compare each of the observed $|t_i|$ to the bootstrap null distribution of $\max(|T_1|, |T_2|, \ldots, |T_m|)$. The tail area of this bootstrap null distribution to the right of $|t_i|$ is a bootstrap estimate of p_i^*. Specifically, the proportion of the B row maxima of $\mathbf{Z}$ that are greater than or equal to $|t_i|$ estimates p_i^*.

The observed p_i, $i = 1, \ldots, m$, are monotonically decreasing functions of the $|t_i|$. Hence p_i^*, $i = 1, \ldots, m$, can also be obtained by comparing each p_i to the bootstrap null distribution of $\min(p_1, p_2, \ldots, p_m)$. Specifically, the proportion of the B row minima of the unadjusted p-values for $\mathbf{Z}$ less than or equal to p_i will also estimate p_i^*.

Dudoit and van der Laan step-down maxT/minP

To obtain the bootstrap estimate of the joint null distribution of the test statistics, the step-down procedure (Dudoit and van der Laan, 2008, page 126) uses the same bootstrap method as the single-step procedure.

In the step-down maxT procedure, let the absolute values of the observed test statistics be ordered from largest to smallest and denote them by $|t|_{(1)} \geq \ldots \geq |t|_{(m)}$. Let $i_1, \ldots, i_m$ index the corresponding test statistics such that $|t_{i_1}| = |t|_{(1)}$, $\ldots$, $|t_{i_m}| = |t|_{(m)}$. The procedure can be described as follows.

Step1 Let $q_{\alpha 1}$ be the $(1-\alpha)$ quantile of the null distribution for

$\max_{\{j=1,\ldots,m\}}(|T_j|) = \max_{\{j=1,\ldots,m\}}(|T_{i_j}|)$. Reject $H_{0(1)}$ if $|t|_{(1)} \geq q_{\alpha 1}$ and continue to next step. Else stop and accept $H_{0(1)}, \ldots, H_{0(m)}$.

Step2 Let $q_{\alpha 2}$ be the $(1-\alpha)$ quantile of the null distribution for $\max_{\{j \neq i_1; j=1,\ldots,m\}}(|T_j|)$. Reject $H_{0(2)}$ if $|t|_{(2)} \geq q_{\alpha 2}$ and continue to next step. Else stop and accept $H_{0(2)}$, $\ldots$, $H_{0(m)}$.

Step3 Let $q_{\alpha 3}$ be the $(1-\alpha)$ quantile of the null distribution for $\max_{\{j \notin \{i_1, i_2\}; j=1,\ldots,m\}}(|T_j|)$. Reject $H_{0(3)}$ if $|t|_{(3)} \geq q_{\alpha 3}$ and continue to next step. Else stop and accept $H_{0(3)}$, $\ldots$, $H_{0(m)}$.

etc.

If we use the unadjusted p-values instead of the observed test statistics:

Step1 Let $q'_{\alpha 1}$ be the α quantile of the null distribution for

$\min_{\{j=1,\ldots,m\}}(p_j) = \min_{\{j=1,\ldots,m\}}(p_{i_j})$. Reject $H_{0(1)}$ if $p_{(1)} \leq q'_{\alpha 1}$ and continue to next step. Else stop and accept $H_{0(1)}, \ldots, H_{0(m)}$.

Step2 Let $q'_{\alpha 2}$ be the α quantile of the null distribution for $\min_{\{j \neq i_1; j=1,\ldots,m\}}(p_j)$. Reject $H_{0(2)}$ if $p_{(2)} \leq q'_{\alpha 2}$ and continue to next step. Else stop and accept $H_{0(2)}$, $\ldots$, $H_{0(m)}$.

Step3 Let $q'_{\alpha 3}$ be the α quantile of the null distribution for $\min_{\{j \notin \{i_1, i_2\}; j=1,\ldots,m\}}(p_j)$. Reject $H_{0(3)}$ if $p_{(3)} \leq q'_{\alpha 3}$ and continue to next step. Else stop and accept $H_{0(3)}$, $\ldots$, $H_{0(m)}$.

etc.

The adjusted p-values are based on the distributions of the maxima of test statistics over successive nested and decreasing subsets of ordered null hypotheses. In the step-down maxT procedure, we start with $H_{0(1)}$ which has the smallest unadjusted p-value and also the largest test statistic $|t|_{(1)}$. The B row maxima of $\mathbf{Z}$ comprise the bootstrap null distribution of $\max(|T_1|, \ldots, |T_m|)$. The tail area of this bootstrap null distribution to the right of $|t|_{(1)}$ is a bootstrap estimate of $p^*_{(1)}$. Specifically, the proportion of the B row maxima of $\mathbf{Z}$ that are greater than or equal to $|t|_{(1)}$ estimates $p^*_{(1)}$.

Now delete the $(i_1)^{\text{th}}$ column of $\mathbf{Z}$, so we have a new matrix $\mathbf{Z}_1$ with dimension $B \times (m-1)$. The B row maxima of $\mathbf{Z}_1$ comprise the bootstrap null distribution of $\max_{\{j \neq i_1; j=1,\ldots,m\}}(|T_j|)$. The tail area of this bootstrap null distribution to the right of $|t|_{(2)}$ is a candidate for estimating $p^*_{(2)}$. Specifically, let $prop_2$ be the proportion of the B row maxima of $\mathbf{Z}_1$ greater than or equal to $|t|_{(2)}$. Then the tail area $prop_2$ is a candidate for estimating $p^*_{(2)}$. However, the adjusted p-value for $H_{0(2)}$ depends on the outcome of $H_{0(1)}$ and is the smallest MTP level α such that both hypotheses are rejected. Therefore, $p^*_{(2)} = \min\{p^*_{(1)}, prop_2\}$.

The steps in the paragraph above are repeated for other hypothesis tests to obtain the adjusted p-values $p^*_{(h)}$, for $h = 3, \ldots, m$. Specifically, for $h = 3, \ldots, m$, delete the $(i_{h-1})^{\text{th}}$ column of $\mathbf{Z}_{h-2}$ to obtain a new matrix $\mathbf{Z}_{h-1}$ with dimension $B \times (m-h+1)$. The B row maxima of $\mathbf{Z}_{h-1}$ comprise the bootstrap null distribution of $\max_{\{j \notin \{i_1,\ldots,i_h\}; j=1,\ldots,m\}}(|T_j|)$. The tail area of this bootstrap null distribution is a candidate for estimating $p^*_{(h)}$. In particular, the proportion, $prop_h$, of the B row maxima of $\mathbf{Z}_{h-1}$ greater than or equal to $|t|_{(h)}$ is a candidate for estimating $p^*_{(h)}$. However, $p^*_{(h)}$ must satisfy $H_{0(1)}, \ldots, H_{0(h)}$ being rejected, and so $p^*_{(h)} = \min\{p^*_{(h-1)}, prop_h\}$. It follows that the adjusted p-values can be written as

$$_{\mathrm{mT}}P^*_{(i)} = \min_{h \in \{1,\ldots,i\}} \{\frac{\sum_{b=1}^{B}(I(maxZ_{h-1}(b) \geq |t|_{(h)}))}{B}\},$$

where $maxZ_{h-1}(b)$ is the maximum of the b^{th} row in the matrix $\mathbf{Z}_{h-1}$ and I is the indicator function.

In the step-down minP procedure, the adjusted p-values are the same as for the maxT procedure, but the algorithm we use to get them is more computationally efficient (Pesarin, 2001; Ge *et al.*, 2003) at the expense of being less understandable. First, re-arrange the columns of $\mathbf{Z}$ in the order of $(i_1, \ldots, i_m)$, and then define a matrix $\mathbf{P}$ of dimension $B \times m$ with the corresponding unadjusted p-values as entries. Therefore, the i^{th} column of $\mathbf{P}$ corresponds to $H_{0(i)}$. Create a new matrix $\mathbf{P'}$ from $\mathbf{P}$ such that

$$\mathbf{P'}[i,m] = \mathbf{P}[i,m] \text{ and } \mathbf{P'}[i,j] = \min\{\mathbf{P'}[i,j+1], \mathbf{P}[i,j]\}, \quad j = 1, \ldots, m-1,$$

where $\mathbf{P}[i,j]$ is the $(i,j)^{\text{th}}$ entry of $\mathbf{P}$. Then we can use $\mathbf{P'}$ to obtain the estimates of the adjusted p-values as follows. The proportion, $prop'_j$, of elements in the j^{th} column of $\mathbf{P'}$ less than or equal to $p_{(j)}$ is a candidate for estimating $p^*_{(j)}$. Again, however, since this is a step-down procedure, we should compare the proportion to the previous adjusted p-values in order to obtain the smallest level α for $H_{0(j)}$ at which all the hypotheses $H_{0(i)}, i = 1, \ldots, j$ are rejected. Hence, $p^*_{(j)} = \max\{p^*_{(j-1)}, prop'_j\}$. The adjusted p-values can therefore be expressed as

$$_{\text{mP}}P^*_{(i)} = \max\{\frac{\sum_{b=1}^{B}(I(\mathbf{P'}[b,i] \le p_{(i)}))}{B}, _{\text{mP}} P^*_{(i-1)}\}, \quad i = 1, \ldots, m,$$

where we define $_{\text{mP}}P^*_{(0)} = 0$.

For both the single-step and step-down versions, the maxT and minP multiple testing procedures of Dudoit and van der Laan are equivalent because the unadjusted p-value is a monotone decreasing transformation of the test statistics; i.e, there is a one-to-one correspondence between $|t|_{(i)}$ and $p_{(i)}, i = 1, \ldots, m$. Hence,

$$_{\text{mT}}P^*_{(i)} =_{\text{mP}} P^*_{(i)}, \quad i = 1, \ldots, m.$$

Westfall and Young minP

The MTP is the same as for the Dudoit and van der Laan step-down minP procedure, except that the Westfall and Young minP procedure uses the permutation distribution instead of the bootstrap distribution to obtain the critical values for q'_α. This procedure is a step-down procedure that controls FWER under a general dependence structure for the test statistics.

The B permuted datasets are generated by shuffling the case/control status of the observed data. The algorithm for calculating the adjusted p-values is the same as that of the Dudoit and van der Laan minP procedure. The adjusted p-values can therefore be expressed as

$$_{\mathrm{WY}}P^*_{(i)} = \max\{\frac{\sum_{b=1}^{B}(I(\mathbf{P'}[b,i] \le p_{(i)}))}{B}, _{\mathrm{WY}}P^*_{(i-1)}\}, \quad i = 1, \ldots, m,$$

where we define $_{\mathrm{WY}}P^*_{(0)} = 0$.

Other FWER-controlling procedures

Following the Holm step-down procedure, Hochberg (1988) proved that using the same cut-offs, a step-up procedure is more powerful, although this procedure is only valid under certain dependence structures for the test statistics (Farcomeni, 2008). Sidak (1967) developed a single-step procedure with common cut-off $1 - \sqrt[m]{1-\alpha}$, and after that an improved step-down procedure with common cut-offs $1 - \sqrt[m-j+1]{1-\alpha}$ for $j = 1, \ldots, m$ (Sidak, 1971). Both the Sidak procedures are valid under positive orthant dependence (Farcomeni, 2008). These other FWER-controlling procedures were not applied in this project.

3.2.3 Procedures Controlling FDR

As m increases, so does m_0 and the chance of rejecting any true null hypothesis. Thus, procedures controlling FWER can become too strict and lose power when some of the null hypotheses are false (i.e. $m_1 = m - m_0 > 0$). In these situations, procedures controlling FDR are a useful alternative. When m and m_1 are large FDR-controlling procedures tend to reject more false null hypotheses than FWER-controlling procedures (Benjamini and Hochberg, 1995). In this section, we consider some FDR-controlling procedures.

To control the FDR, we applied the general step-up Benjamini and Hochberg (1995) procedure, the step-down Gavrilov *et al.* (2009) procedure and the step-down Benjamini and Liu (1999) procedure. These procedures are marginal procedures and the latter requires independent test statistics to be valid. We also applied the joint bootstrap-based procedure that takes advantage of the positive dependence among test statistics: the single-step Dudoit and van der Laan empirical Bayes procedure (Dudoit and van der Laan, 2008, page 319).

Table 3.3 summarizes the FDR-controlling procedures applied.

Table 3.3: Summary of FDR-controlling procedures.

	marginal	joint
single-step	—	Dudoit and van der Laan empirical Bayes
step-up	Benjamini and Hochberg	—
step-down	Gavrilov* Benjamini and Liu	—

* the only adaptive procedure.

Benjamini and Hochberg

The Benjamini and Hochberg step-up procedure is the first developed and most commonly used method to control FDR. This procedure is a marginal multiple testing procedure. Benjamini and Hochberg (1995) showed that, if m_0 were known, this procedure would control FDR at level $\alpha m_0/m$ for independent test statistics. Later Benjamini and Yekutieli (2001) extended the result to test statistics that have positive regression dependency with test statistics from $\mathcal{H}_0$.

For this procedure the cut-offs are $\frac{j}{m}\alpha$, for $j = 1, \ldots, m$. Since it is a step-up procedure, we test the least significant hypothesis first, and once we reject one hypothesis, we reject all other more significant hypotheses. The steps are:

For $j = m, \ldots, 1$:

- If $p_{(j)} > \frac{j}{m}\alpha$, accept $H_{0(j)}$ and continue.

- Else reject $H_{0(j)}, \ldots, H_{0(1)}$.

Another way to understand the procedure is this: we reject $H_{0(j)}$, $j = 1, 2, \ldots, m$, if there exists $i \in \{j, \ldots, m\}$ such that $p_{(i)} \leq \frac{i}{m}\alpha$. In order to get the adjusted p-value for $H_{0(j)}$, we use this understanding. To obtain the smallest level α at which $H_{0(j)}$, $j = 1, 2, \ldots, m$, would be rejected, we need

$$p_{(i)} \leq \frac{i}{m}\alpha, \quad \text{for some } i \in \{j, \ldots, m\}.$$

Thus,

$$\alpha \geq \frac{m}{i}p_{(i)}, \quad \text{for some } i \in \{j, \ldots, m\}.$$

provided the right-hand sides are ≤ 1. Since only one i is needed to satisfy the inequality, the inequality becomes

$$\alpha \geq \min_{i \in \{j, \ldots, m\}} \{\frac{m}{i}p_{(i)}\}$$

provided the minimum ≤ 1. Thus, the adjusted p-values for the Benjamini and Hochberg step-up procedure are

$$_{\text{BH}}P^*_{(i)} = \min\{\min_{j \in \{i, \ldots, m\}} \{p_{(j)} \times \frac{m}{j}\}, 1\}, \quad i = 1, \ldots, m.$$

From the equation above, we can find that the adjusted p-values for the Benjamini and Hochberg step-up procedure have to be greater than or equal to their unadjusted p-values because,

$$\begin{aligned} \min_{j \in \{i, \ldots, m\}} \{p_{(j)} \times \frac{m}{j}\} &= \min\{p_{(i)} \times \frac{m}{i}, p_{(i+1)} \times \frac{m}{i+1}, \ldots, p_{(m)} \times \frac{m}{m}\} \\ &\geq \min\{p_{(i)} \times \frac{m}{m}, p_{(i+1)} \times \frac{m}{m}, \ldots, p_{(m)} \times \frac{m}{m}\} \\ &= \min\{p_{(i)}, p_{(i+1)}, \ldots, p_{(m)}\} \\ &= p_{(i)}, \end{aligned}$$

and $1 \geq p_{(i)}$. Hence, $\min\{\min_{j \in \{i, \ldots, m\}} \{p_{(j)} \times \frac{m}{j}\}, 1\} \geq p_{(i)}$.

In general, however, the adjusted p-values for FDR-controlling procedures need not be greater than or equal to their unadjusted p-values.

Gavrilov

The Gavrilov procedure is an adaptive step-down FDR-controlling procedure which controls the FDR when the test statistics are independent. However, through simulations, the authors find that the FDR is controlled at or slightly above the desired level under positive dependence of the test statistics. Conceptually, the procedure is *adaptive* in that it uses an estimated value of m_0 to revise the Benjamini and Hochberg procedure so that it is less conservative. The revised Benjamini and Hochberg procedure is based on the following reasoning. The usual level α Benjamini and Hochberg procedure has FDR $\leq \alpha m_0/m$, which is too conservative when m_0/m is small. Gavrilov *et al.* (2009) note that, if m_0 were known, we could increase power, while continuing to ensure FDR $\leq \alpha$, by applying the Benjamini and Hochberg procedure with level $\alpha' = \frac{m_0}{m}\alpha$.

The Gavrilov procedure is a simpler case of the multiple-stage linear step-up procedure of Benjamini *et al.* (2006), which outlines how to estimate m_0. From Table 3.1, we know

$$\begin{aligned} & S \leq m_1 \\ \Rightarrow\quad & R - V = S \leq m_1 = m - m_0 \\ \Rightarrow\quad & R - V \leq m - m_0 \\ \Rightarrow\quad & m_0 \leq m - (R - V). \end{aligned}$$

Because $E(\frac{V}{R}) \leq \alpha$, we have the approximation $V \overset{\sim}{\leq} \alpha R$. Thus,

$$\begin{aligned} m_0 &\leq m - (R - V) \\ &\overset{\sim}{\leq} m - (R - \alpha R) \\ &= m - (1 - \alpha)R, \end{aligned}$$

and we may set

$$\hat{m}_0 = m - (1 - \alpha)R.$$

If $\hat{m}_0$ is used instead of m in the Benjamini and Hochberg procedure, the resulting adaptive procedure rejects the same hypotheses as the Gavrilov procedure.

For the Gavrilov procedure, the cut-offs are $\frac{i\alpha}{m+1-i(1-\alpha)}$ for $i = 1, \ldots, m$:

- If $p_{(i)} \leq \frac{i\alpha}{m+1-i(1-\alpha)}$, reject $H_{0(i)}$ and continue.
- Else accept $H_{0(i)}, \ldots, H_{0(m)}$ and stop.

The adjusted p-value for $H_{0(j)}$ is the smallest level α at which $H_{0(j)}$ would be rejected. But, in a step-down procedure, $H_{0(j)}$ cannot be rejected unless $H_{0(1)}, \ldots, H_{0(j)}$ are. Therefore,

$$p_{(i)} \leq \frac{i\alpha}{m+1-i(1-\alpha)}, \quad i = 1, \ldots, j.$$

Thus,

$$\alpha \geq \frac{(m+1-i)p_{(i)}}{(1-p_{(i)})i}, \quad i = 1, \ldots, j,$$

given the numbers on the right-hand side of the equations are ≤ 1. Thus, α should satisfy

$$\alpha \geq \max_{i \in \{1,\ldots,j\}} \{\frac{(m+1-i)p_{(i)}}{(1-p_{(i)})i}\},$$

provided the maximum is ≤ 1. It follows that the adjusted p-values are:

$$_{\text{Gav}}P^*_{(i)} = \min\{1, \max_{j \in \{1,\ldots,i\}} \{\frac{(m+1-j)p_{(j)}}{(1-p_{(j)})j}\}\} \quad i = 1, \ldots, m. \tag{3.2}$$

Benjamini and Liu

The Benjamini and Liu procedure, which can be either step-up or step-down, also controls the FDR when the test statistics are independent. Based on a large simulation study, the step-down procedure turned out to be more powerful when the number of tested hypotheses was small and many of the hypotheses were far from being true (Benjamini and Liu, 1999). We will thus only consider the Benjamini and Liu step-down procedure. For the procedure, the cut-offs are $1 - [1 - \min(1, \frac{m}{m-j+1}\alpha)]^{\frac{1}{m-j+1}}, j = 1, \ldots, m$:

- If $p_{(i)} \leq 1 - [1 - \min(1, \frac{m}{m-i+1}\alpha)]^{\frac{1}{m-i+1}}$, reject $H_{0(i)}$ and continue.

- Else accept $H_{0(i)}, \dots, H_{0(m)}$ and stop.

The adjusted p-value for $H_{0(j)}$ is the smallest level α at which $H_{0(j)}$ would be rejected. In a step-down procedure, however, $H_{0(j)}$ cannot be rejected unless $H_{0(1)}, \dots, H_{0(j)}$ are. Therefore,

$$p_{(i)} \leq 1 - [1 - \min(1, \frac{m}{m-i+1}\alpha)]^{\frac{1}{m-i+1}}, \quad i = 1, \dots, j,$$

or

$$1 - (1 - p_{(i)})^{m-i+1} \leq \min\{1, \frac{m}{m-i+1}\alpha\}, \quad i = 1, \dots, j.$$

The latter set of inequalities can be rewritten as

$$1 - (1 - p_{(i)})^{m-i+1} \leq 1 \text{ and } 1 - (1 - p_{(i)})^{m-i+1} \leq \frac{m}{m-i+1}\alpha, \quad i = 1, \dots, j.$$

Since $1 - (1 - p_{(i)})^{m-i+1} \leq 1$ for any i, α satisfies all j inequalities when

$$\alpha \geq \frac{m-i+1}{m} \times [1 - (1 - p_{(i)})^{m-i+1}], \quad i = 1, \dots, j.$$

Moreover, the right-hand side is never greater than 1 because $\frac{m-i+1}{m} \leq 1$ and $1 - (1 - p_{(i)})^{m-i+1} \leq 1$. Hence,

$$\alpha \geq \max_{i \in \{1,\dots,j\}} \{\frac{m-i+1}{m} \times [1 - (1 - p_{(i)})^{m-i+1}]\}.$$

Therefore, the adjusted p-values equation is:

$${}_{\text{BL}}P^*_{(i)} = \max_{j \in \{1,\dots,i\}} \{\frac{m-j+1}{m} \times [1 - (1 - p_{(j)})^{m-j+1}]\}, \quad i = 1, \dots, m.$$

Dudoit and van der Laan empirical Bayes

The Dudoit and van der Laan empirical Bayes procedure is a bootstrap-based joint MTP (Dudoit and van der Laan, 2008, page 298). For a given level α at which to control the FDR, this single-step procedure uses a common cut-off $c(\alpha)$ for rejecting the test statistics:

Reject H_{0i} if $|T_i| \geq c(\alpha)$. If we knew the true distribution of the test statistics, we could let $c(\alpha)$ be the smallest c such that

$$\text{FDR} = E\left[\frac{V(c)}{V(c)+S(c)}\right] \leq \alpha,$$

where

$$V(c) = \sum_{i \in \mathcal{S}_0} I(|T_i| \geq c), \quad S(c) = \sum_{i \in \mathcal{S}_0^c} I(|T_i| \geq c),$$

$\mathcal{S}_0$ is the set of indices of the true null hypotheses and $\mathcal{S}_0^c$ is the complement of $\mathcal{S}_0$. However, the distribution of $\frac{V(c)}{V(c)+S(c)}$ is unknown and so is the set $\mathcal{S}_0$.

If $\mathcal{S}_0$ were known, the distribution of the test statistics corresponding to null hypotheses in $\mathcal{S}_0$ could be approximated by the bootstrap null distribution. Let the number of rejected hypotheses in $\mathcal{S}_0$ under this bootstrap null distribution be

$$V' = \sum_{i \in \mathcal{S}_0} I(|T_i'| \geq c),$$

where the T_i' come from the bootstrap null distribution. Dudoit *et al.* (2004) prove that, asymptotically, V' tends to be larger than V. As a result, asymptotically, $\frac{V'}{V'+S}$ tends to be larger than $\frac{V}{V+S}$. Though we don't know $\mathcal{S}_0$, we could imagine using another set of indices $\tilde{s}_0$ instead, guessed so that $\tilde{s}_0 \supset \mathcal{S}_0$. Let

$$\tilde{V} = \sum_{i \in \tilde{s}_0} I(|T_i'| \geq c),$$

be the number of rejected hypotheses in $\tilde{s}_0$ under the bootstrap null distribution and

$$\tilde{S} = \sum_{i \in \tilde{s}_0^c} I(|T_i| \geq c)$$

be the number of rejected hypotheses in $\tilde{s}_0^c$ under the true data generating distribution. Then $\tilde{V} \geq V'$ because $\tilde{V}$ is a sum of rejection decisions over a set $\tilde{s}_0$ as large or larger than $\mathcal{S}_0$ and $\tilde{S} \leq S$ because $\tilde{S}$ is a sum of rejection decisions over a set $\tilde{s}_0^c$ as small or smaller than $\mathcal{S}_0^c$. Therefore,

$$\frac{\tilde{V}}{\tilde{V}+\tilde{S}} \geq \frac{V'}{V'+S}.$$

But, asymptotically, $\frac{V'}{V'+S}$ tends to be larger than $\frac{V}{V+S}$ and so we can control $E(V/(V+S))$ by controlling $E\left(\frac{\tilde{V}}{\tilde{V}+\tilde{S}}\right)$. However, a guessed set $\tilde{s}_0$ that is too big can lead to type 1 error rates that are smaller than the nominal level. To protect against this possibility, van der Laan *et al.* (2005) propose hedging by repeatedly drawing $\tilde{\mathcal{S}}_0$ from a certain distribution (discussed below) constructed so that these random guessed sets tend to be larger than $\mathcal{S}_0$ for finite samples and, for large samples, approach $\mathcal{S}_0$ with probability 1.

A random guessed set $\tilde{\mathcal{S}}_0$ is constructed by sampling m independent Bernoulli random variables indicating the truth of the corresponding null hypothesis. If the i^{th} Bernoulli random variable is a success, the i^{th} null hypothesis is included in the set. Each Bernoulli random variable is assigned a success probability equal to the posterior probability of the corresponding hypothesis given its test statistic. This posterior probability is calculated under a working model for the marginal distribution f of the test statistics such that

$$f = \pi_0 f_0 + (1 - \pi_0) f_1,$$

where $\pi_0 = m_0/m$, f_0 is the null distribution of the test statistics and f_1 is the non-null distribution. Under this working model, the required posterior probability is

$$\Pr(H_{0i}|T_i) = \pi_0 \, \frac{f_0(T_i)}{f(T_i)}. \tag{3.3}$$

For good finite-sample behaviour, the random sets should tend to cover the actual set S_0 of true null hypotheses. Thus, setting $\pi_0 = 1$ is recommended, even thought this can be too conservative when $\pi_0 = \frac{m_0}{m} < 1$. We estimate f_0 by a normal distribution with mean zero and variance equal to the sample variance of the bootstrap null distribution. We use a kernel density estimator to estimate f from the uncentered bootstrapped test statistics.

Once $\tilde{\mathcal{S}}_0$ is sampled, we may sample $\tilde{V}$ and $\tilde{S}$. To sample $\tilde{V}$, the number of rejected hypotheses in $\tilde{\mathcal{S}}_0$ under the bootstrap null distribution, test statistics are sampled from the bootstrap null distribution. However, sampling $\tilde{S}$, the number of rejected hypotheses in $\tilde{\mathcal{S}}_0^c$ under the true data generating distribution, is not possible because the true data generating distribution is unknown. van der Laan *et al.* (2005) suggest approximating the distribution of $\tilde{S}$ by the distribution of the number of rejections in $\tilde{\mathcal{S}}_0^c$ based on the observed test statistics. However, for finite samples, this ignores the variation in $\tilde{S}$ due to randomness in the test statistics. Instead we propose to use the number of rejections in $\tilde{\mathcal{S}}_0^c$ based on test statistics sampled from the uncentered bootstrap distribution.

Therefore, the distribution of $\frac{V(c)}{V(c)+S(c)}$ is estimated by the distribution of $\frac{\tilde{V}(c)}{\tilde{V}(c)+\tilde{S}(c)}$. Substituting this related proportion leads to the following procedure:

- For a given level α, let $c(\alpha)$ be that smallest c such that

$$g(c) \equiv E\left[\frac{\tilde{V}(c)}{\tilde{V}(c)+\tilde{S}(c)}\right] \leq \alpha.$$

- Reject H_{0i} if $|T_i| \geq c(\alpha)$.

To obtain the adjusted p-values from this procedure, we could naively try $p_i^* = g(|t_i|)$ in the hopes that $g(c)$ is monotone decreasing. However, for finite sample sizes, $g(c)$ can be non-monotone as shown in Figure 3.1, leading to the undesirable behaviour that $p_{(i)}^* = g(|t|_{(i)}) < g(|t|_{(j)}) = p_{(j)}^*$ for some $i > j$. To avoid this undesirable behaviour, we reason as follows. As shown in Figure 3.1, for a point c_* at which $g(c)$ starts to increase, let $c_*' > c_*$ be the next largest value such that $g(c_*') = g(c_*) = \alpha_*$. Suppose $c_* < |t_i| < c_*'$. Then, from the figure, we can see that

$$\alpha_* = \inf\{g(c) : |t_i| \geq c\}.$$

The cut-off function is defined as $c(\alpha) = \inf\{c : g(c) \leq \alpha\}$. Hence $c(\alpha_*) = c_*$ and, for any $\alpha < \alpha_*$, $c(\alpha) > c_*' > |t_i| > c_* = c(\alpha_*)$. The adjusted p-value for H_{0i} is defined as $p_i^* = \inf\{\alpha : |t_i| \geq c(\alpha)\}$ and it follows that $p_i^* = \alpha_*$. But $\alpha_* = \inf\{g(c) : |t_i| \geq c\}$ and so we may conclude that

$$p_i^* = \inf\{g(c) : |t_i| \geq c\} \;\text{ or }\; p_{(i)}^* = \inf\{g(c) : |t|_{(i)} \geq c\}.$$

To approximate this infimum, we can take the minimum over the set of m observed test statistics:

$$\begin{aligned} p_{(i)}^* &\approx \min_{j \in \{1,..,m\}} \left\{ g(|t|_{(j)}) : |t|_{(i)} \geq |t|_{(j)} \right\} \\ &= \min_{j \in \{i,..,m\}} \{ g(|t|_{(j)}) \} \end{aligned} \tag{3.4}$$

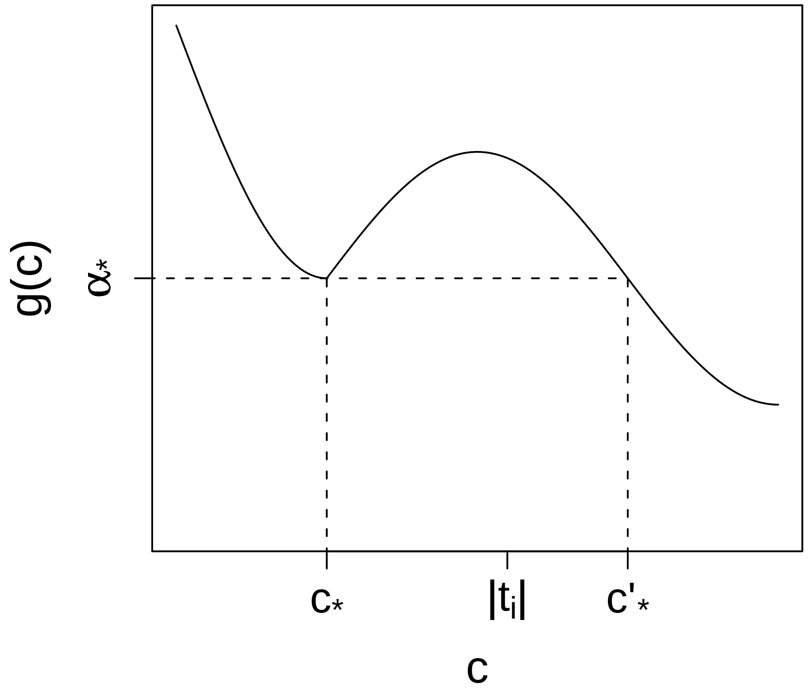

Figure 3.1: Hypothetical FDR curve from the empirical Bayes procedure.

R pseudo-code to implement the procedure is provided in the **Appendix A**.

Other FDR-controlling procedures

Although the Benjamini and Hochberg procedure controls FDR for test statistics that have positive regression dependency with test statistics from the set of true null hypotheses, the procedure is conservative when false null hypotheses exist ($m_1 > 0$). Therefore, Benjamini and Hochberg (2000) developed an adaptive step-up procedure to improve the power of their original procedure. Their adaptive procedure estimates the number of true null hypotheses $\hat{m}_0$ first, and then use this estimator instead of m in the cut-offs. Thus, the cut-offs are $\frac{j}{\hat{m}_0}\alpha$, for $j = 1, \ldots, m$. Their adaptive procedure controls FDR for independent test statistics, but is still useful in cases of dependency. Storey *et al.* (2004) developed another adaptive procedure based on their estimator of the number of true null hypotheses. The Storey procedure is valid when weak dependence exists and the number of tests is large. A resampling-based procedure of Yekutieli and Benjamini (1999) controls FDR under dependence, but this procedure is not guaranteed to yield FDR control (Farcomeni, 2008). Previous studies have shown that these other FDR-controlling procedures all have their own advantages in certain situations. However, we do not examine them in this project due to time limitations.

3.3 Example

To illustrate the definitions and procedures, we go through the following example.

Example 1: Suppose T_1 and T_2 are independent normally distributed test statistics with variance 1 and means $\mu_1 = 10$ and $\mu_2 = 0$, respectively. We know that the test statistics are independent and normally distributed with variance 1 but we don't know their means. We would like to use data to simultaneously test

$$H_{0i} : \mu_i = 0 \text{ versus } H_{1i} : \mu_i \neq 0, i = 1, 2.$$

We collect data that leads to observed test statistics $t_1 = 9.7$ and $t_2 = -0.8$.

In Example 1, the test statistic $T_1 \sim N(0, 1)$ under H_{01}. The single-test FWER for H_{01} when H_{01} is false is zero; otherwise it is the tail probability

$$Pr(V = 1|H_{01}) = Pr(|T_1| \geq q_\alpha|H_{01}) = \alpha,$$

where critical value q_α is the $(1 - \alpha)^{\text{th}}$ quantile of $|T_1|$ under $H_{01} : \mu_1 = 0$. We only worry about controlling the single-test FWER when H_{01} is true because when it is false, FWER $= 0$. To control the single-test FWER at level α, we reject H_{01} if $|t_1| \geq q_\alpha$ given $T_1 = t_1$. The unadjusted p-value is therefore the smallest α at which $|t_1| \geq q_\alpha$. Any quantile q to the left of $|t_1|$ will satisfy the inequality, but the one that has the smallest tail probability is $q = |t_1|$. Hence, the unadjusted p-value for H_{01} is $p_1 = Pr(|T_1| \geq |t_1||H_{01})$. The numeric values of the unadjusted p-values for the two tests in the example are $< 1 * 10^{-16}$ and 0.42.

Applying the step-down multiple testing procedure of Gavrilov *et al.* (2009), we use the equation (3.2) of ${}_{\text{Gav}}P^*_{(i)}, i = 1, \ldots, m$ to calculate the adjusted p-values. Substituting the numeric values of $p_{(1)} = 0$ and $p_{(2)} = 0.42$, we have ${}_{\text{Gav}}P^*_{(1)} = 0$ and ${}_{\text{Gav}}P^*_{(2)} = 0.36$. The adjusted p-value for the second hypothesis test (0.36) is less than the unadjusted p-value (0.42). Intuitively, the small $p_{(1)}$ observed from our data indicates that $H_{0(1)}$ is a sure bet for a correct rejection. Thus, a step-down procedure controlling $E(\frac{V}{R})$ could reject $H_{0(2)}$ at the second step, in spite of the large $p_{(2)}$, because there is room for a mistake given the certainty about $H_{0(1)}$ being false. This example shows that, in general, the adjusted p-values for FDR-controlling procedures need not be bigger than the unadjusted p-values, in contrast to what is shown in Section 3.2.1 for FWER-controlling procedures.

Chapter 4

Analysis

We fit the models outlined in Section 3.1, and got unadjusted p-values for all the organochlorine exposures in our study. Then, we applied all the MTPs we considered in Chapter 3. Our aim was to control the overall Type I error rate at level 5%.

In order to approximate the marginal distributions of the test statistics, we generated 5000 bootstrap samples and calculated twenty-two $Wald$ statistics for each sample. Boxplots of the twenty-two distributions are shown in Figure 4.1. The boxes are ordered by the size of the bootstrapped means of the test statistics. The red boxes with red labels represent the organochlorines with more than 20% subjects below the detection limit. Four out of twenty-two organochlorines have test statistics with bootstrapped 75^{th} percentiles below 2. Thus, we can roughly estimate that about four null hypotheses are true, $m_0 \approx 4$.

Figure 4.2 shows the proportions of subjects with organochlorine levels below the detection limit for each organochlorine. Six organochlorines, PCB 28, *cis*-Nonachlor, PCB 105, p, p'-DDT, mirex and PCB 183, have more than 20% of subjects with levels below the detection limit and they are marked by red filled squares in all figures. Going back to check Figure 4.1, we find that the organochlorines with higher proportions below the detection limit tend to have smaller test statistics. Figure 4.3 shows the relationship between the unadjusted p-value and the proportion below the detection limit. The 0.05 value of the unadjusted p-value is marked by the dashed line. As can be clearly seen, the four organochlorines with

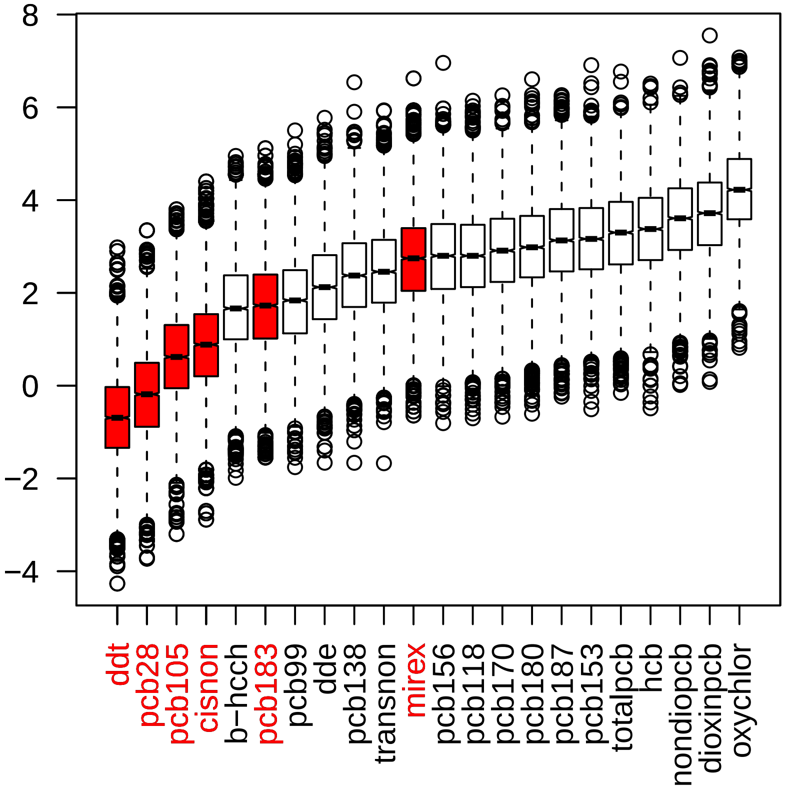

Figure 4.1: Boxplot of bootstrap test statistics.

the largest proportion below the detection limit have the largest p-values.

We also examined the bootstrap test statistic correlations to get a better idea of the dependence structure. Figure 4.4 is the histogram of the test statistic correlations. No negative correlation is observed, and around 50 out of 231 correlations are greater than 0.6. Figure 4.5 is the heatmap of the squared Pearson correlations between the bootstrap test statistics, where red represents correlation 1 and white correlation 0. The names of the organochlorines with more than 20% subjects below the detection limit are in red with a star mark follows. We can see that the test statistics for some of organochlorines are

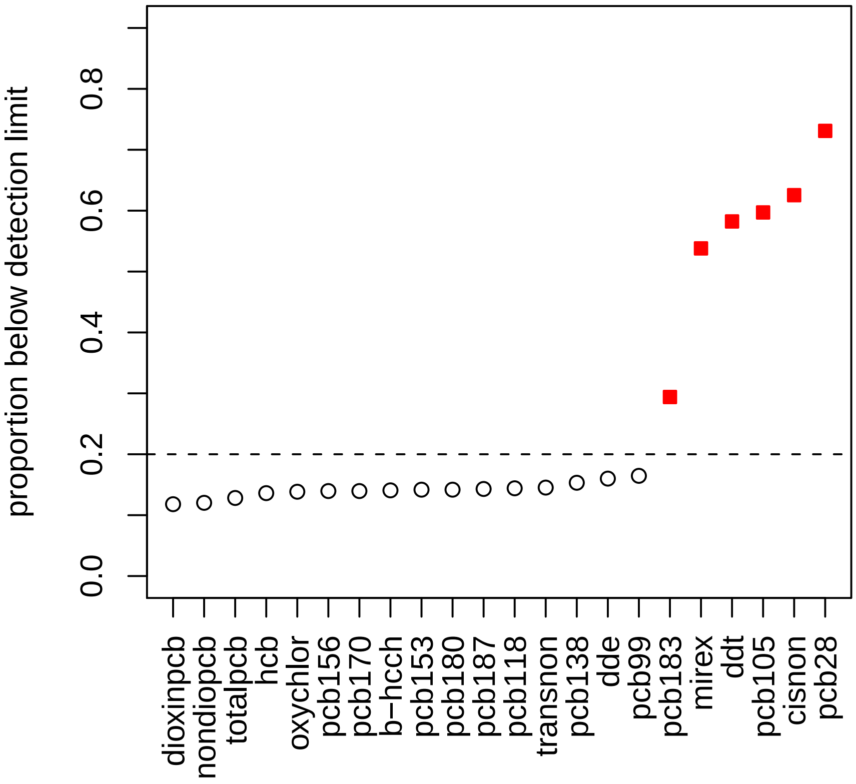

Figure 4.2: Plot of ordered proportions below detection limit.

highly correlated (e.g. PCB 156, PCB 187, PCB 170 and PCB 180), while the test statistics for some other organochlorines have very low correlations with all other test statistics for other ones (e.g. mirex, PCB 28 and β-HCCH). For each organochlorine, we calculated the average correlation between its test statistic and the test statistics for all other organochlorines. Figure 4.6 shows the relationship between the average correlations for organochlorines and the proportion of subjects below the detection limit. The four points with the lowest average correlations are β-HCCH, mirex, p,p'-DDT and PCB 28. Except for β-HCCH, these organochlorines are among the six organochlorines which have more than 20% of subjects below the detection limit. Therefore, the test statistics of organochlorines with higher proportions below the detection limit tend to have lower average correlation with the test statistics of other organochlorines.

4.1 Results for FWER-controlling procedures

Table 4.1 shows the unadjusted p-values and the adjusted p-values for all the FWER-controlling procedures we considered. The p-values are rounded to 4 decimal places and a horizontal line indicating the 0.05 threshold is shown. Since the Dudoit and van der Laan maxT and minP procedures are equivalent in this case, only one column for each of the single-step (DL-SS) and step-down procedures (DL-SD) are shown. Figure 4.7 plots of the unadjusted p-values versus the adjusted p-values for the FWER-controlling procedures, on the $\log_{10}$ scale for both axes. The lower the curve is, the smaller the adjusted p-values are.

Comparing the entries in the table and the curves in the plot, we can see that the adjusted p-values for single-step procedures are generally larger than those of step-down procedures for a specific organochlorine analyte. The Dudoit and van der Laan procedures are good examples of this. The adjusted p-values for their step-down procedures are smaller than the adjusted p-values for their single-step procedures. For a 0.05 MTP level, the Bonferroni procedure has the least number (seven) of adjusted p-values less than the threshold while all the other procedures have nine adjusted p-values less than the 0.05 threshold.

Take PCB 180 as an example. First, from Table 4.1, we see that the Bonferroni procedure does not reject the null hypothesis for it at the 0.05 level, but all other procedures do. Second, for PCB 180, all the adjusted p-values from marginal MTPs are greater than those from joint MTPs. Finally, the step-down procedures give smaller adjusted p-values for PCB 180 than the single-step procedures.

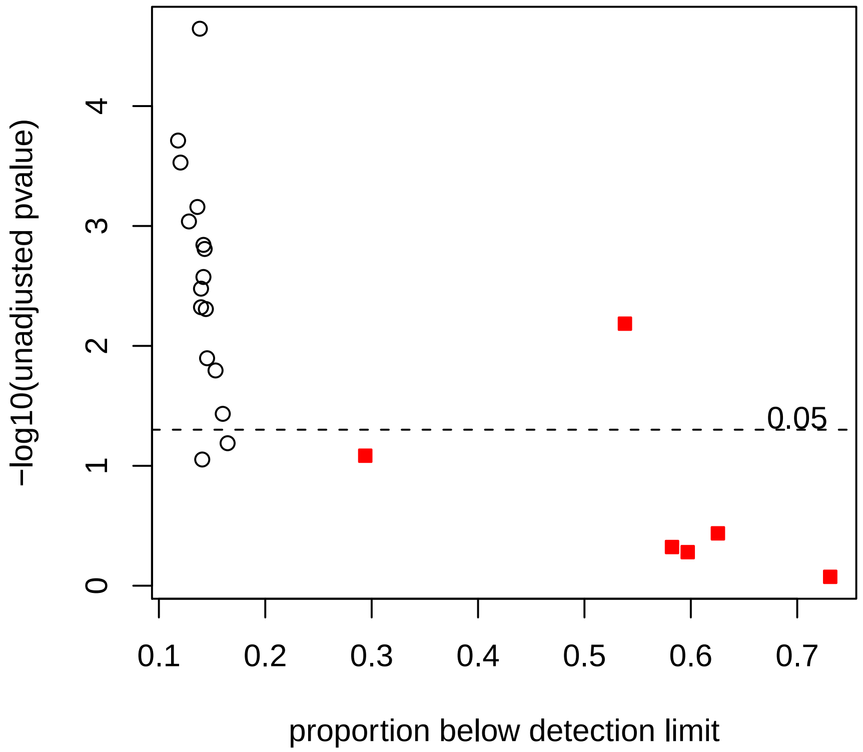

Figure 4.3: Relationship between $-\log_{10}$(unadjusted p-values) and proportions below detection limit.

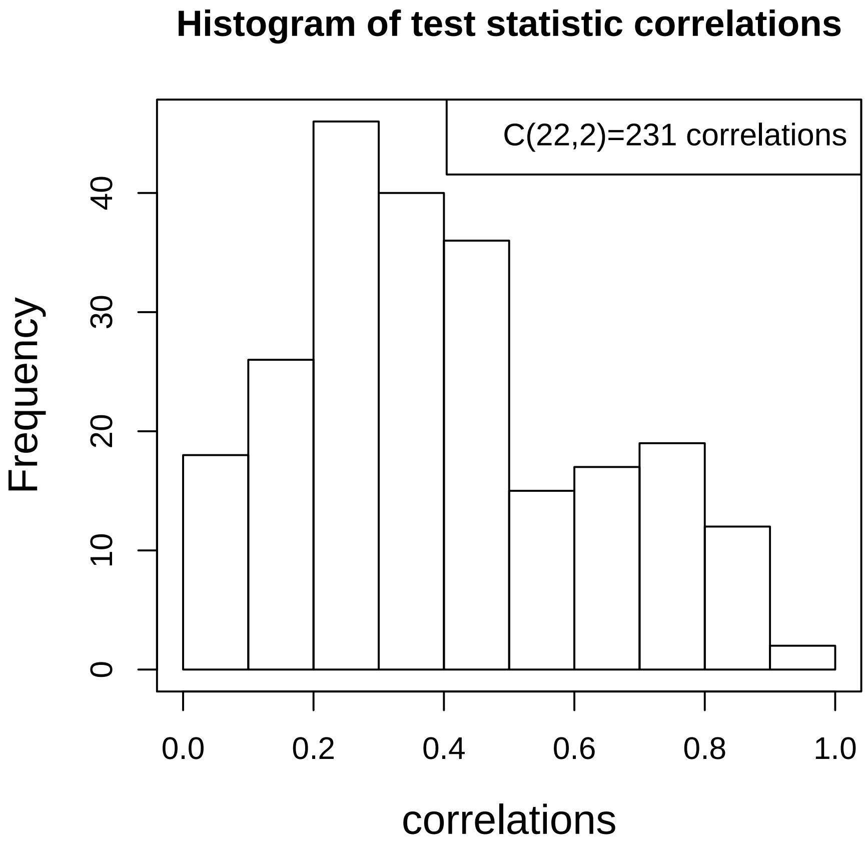

Figure 4.4: Histogram of test statistic correlations.

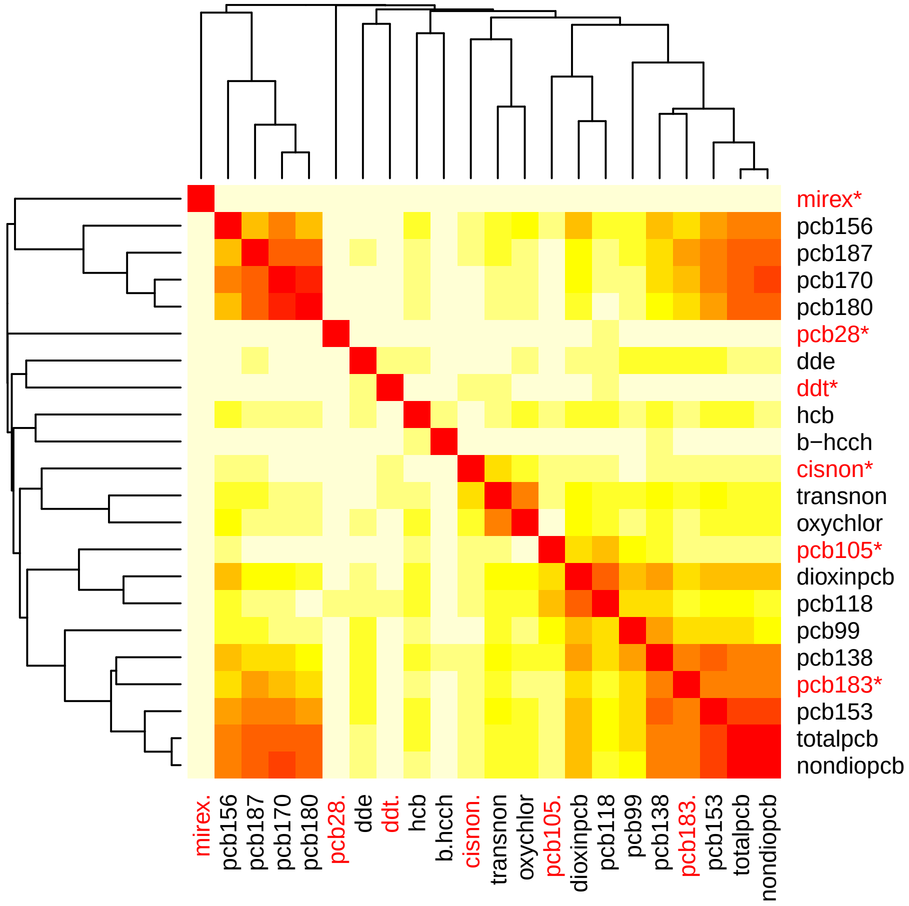

Figure 4.5: Heatmap of correlations between bootstrap test statistics.

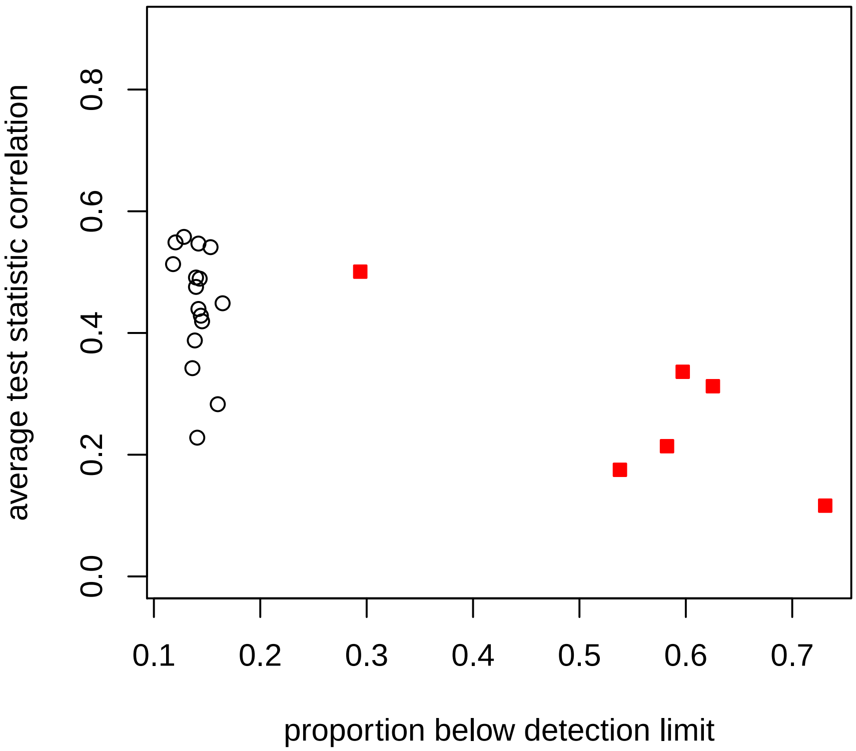

Figure 4.6: Relationship between test statistic correlations and proportions below detection limit.

Table 4.1: Summary of FWER-controlling procedure results.

		single-step		step-down		
	unadj-p[a]	Bonferroni	DL-SS[b]	Holm	DL-SD[c]	Westfall-Young
Oxychlordane	< 0.0001	0.0005	< 0.0001	0.0005	< 0.0001	0.0004
Summed Dioxin-like PCBs	0.0002	0.0043	0.0018	0.0041	0.0018	0.0042
Summed non Dioxin-like PCBs	0.0003	0.0065	0.0028	0.0059	0.0026	0.0056
HCB	0.0007	0.0153	0.0094	0.0132	0.0086	0.0128
Summed total PCBs	0.0009	0.0202	0.0126	0.0165	0.0110	0.0150
PCB 153	0.0014	0.0316	0.0198	0.0244	0.0176	0.0222
PCB 187	0.0016	$\underline{0.0342}$	0.0224	0.0249	0.0190	0.0224
PCB 180	0.0027	0.0586	0.0392	0.0399	0.0342	0.0350
PCB 170	0.0033	0.0733	$\underline{0.0474}$	$\underline{0.0466}$	$\underline{0.0386}$	$\underline{0.0420}$
PCB 156	0.0048	0.1047	0.0660	0.0619	0.0518	0.0550
PCB 118	0.0049	0.1084	0.0664	0.0619	0.0518	0.0550
Mirex	0.0065	0.1436	0.0852	0.0718	0.0610	0.0638
trans-Nonachlor	0.0127	0.2787	0.1544	0.1267	0.1052	0.1042
PCB 138	0.0160	0.3525	0.1912	0.1442	0.1226	0.1160
p, p'-DDE	$\underline{0.0369}$	0.8107	0.3730	0.2948	0.2426	0.2334
PCB 99	0.0648	1.0000	0.5642	0.4535	0.3552	0.3300
PCB 183	0.0824	1.0000	0.6460	0.4944	0.3912	0.3766
β-HCCH	0.0885	1.0000	0.6728	0.4944	0.3912	0.3766
cis-Nonachlor	0.3659	1.0000	0.9972	1.0000	0.8288	0.8232
p, p'-DDT	0.4759	1.0000	0.9998	1.0000	0.8448	0.8520
PCB 105	0.5250	1.0000	1.0000	1.0000	0.8448	0.8520
PCB 28	0.8429	1.0000	1.0000	1.0000	0.8448	0.8520

[a] unadjusted p-value; [b] the Dudoit and van der Laan single-step maxT/minP procedures; [c] the Dudoit and van der Laan step-down maxT/minP procedures. The Bonferroni and Holm procedures are marginal MTPs; others are joint MTPs.

4.2 Results for FDR-controlling procedures

Table 4.2 shows the unadjusted p-values and the adjusted p-values for all the FDR-controlling procedures we considered. As in Table 4.1, the p-values are rounded to 4 decimal places and a horizontal line indicating the 0.05 threshold is shown. For the empirical Bayes procedure, we approximated the distribution of $\tilde{S}$ by the distribution of the number of rejections in $\tilde{\mathcal{S}}_0^c$ based on both the observed test statistics and the test statistics sampled from the uncentered bootstrap distribution. The two methods gave similar results, so we only provide the results based on the test statistics sampled from the uncentered bootstrap distribution in the table. Figure 4.8 plots of the unadjusted p-values versus the adjusted p-values for the FDR-controlling procedures, on the $\log_{10}$ scale for both axes. As in Figure 4.7, the lower the curve is, the smaller the adjusted p-values are.

Comparing the entries in the table, the adjusted p-values for all the procedures do not show a pattern related to whether the procedure is single-step or stepwise procedure. For a 0.05 MTP level, the Benjamini and Liu procedure has the least number of adjusted p-values less than the threshold; twelve out of twenty-two. The numbers of rejected null hypotheses for the Benjamini and Hochberg, the Gavrilov and the empirical Bayes procedures are respectively fourteen, eighteen and fifteen. Except for the largest unadjusted p-value, the Gavrilov procedure has the smallest adjusted p-values; and except for the four largest unadjusted p-values, the Benjamini and Liu procedure has the largest adjusted p-values. The empirical Bayes procedure (the only joint procedure considered) does not always have smaller adjusted p-values than the marginal procedures.

Take the organochlorine p,p'-DDE as an example. Neither the Benjamini and Hochberg nor the Benjamini and Liu procedures reject the null hypothesis at the 0.05 threshold. However, the adjusted p-value for the Benjamini and Hochberg procedure is smaller (0.0540) than the one for the Benjamini and Liu procedure (0.0944). Both the Gavrilov and the empirical Bayes procedures reject the null hypothesis, but the adjusted p-value for the empirical Bayes procedure (0.0486) is twice Gavrilov's adjusted p-value (0.0204). In fact, for organochlorines with unadjusted p-values greater than 0.0033, the adjusted p-values for the empirical Bayes procedure are about twice the adjusted p-values for the Gavrilov procedure. However, the Gavrilov procedure is adaptive: it estimates $\hat{m}_0 = 4.9$, and uses this estimate to revise the Benjamini and Hochberg procedure in an attempt to gain power. By contrast, all the other MTPs, including the empirical Bayes procedure, are not adaptive and conservatively take $m_0 = m = 22$.

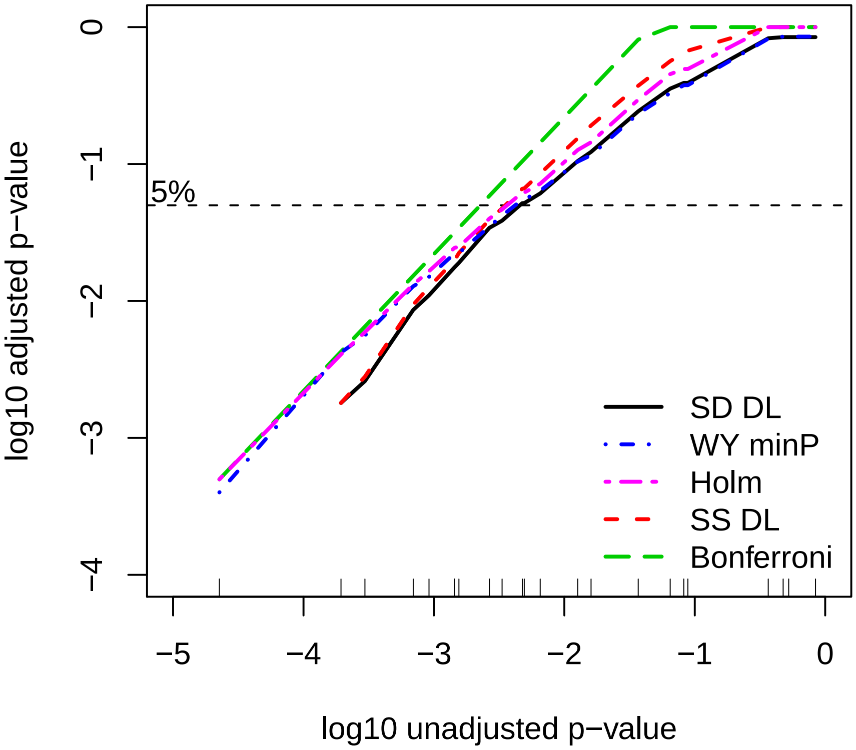

Figure 4.7: Relationship between adjusted and unadjusted p-values, on the $\log_{10}$-scale, for FWER-controlling procedures. The locations of unadjusted p-values are marked by the rug on the x-axis. The 5% cut-off for adjusted p-values is marked by a horizontal dashed line. 'SS DL', Dudoit and van der Laan's single-step maxT/minP procedure; 'SD DL', Dudoit and van der Laan's step-down maxT/minP procedure; 'WY minP', Westfall and Young minP procedure. The smallest adjusted p-values for the 'SS DL' and 'SD DL' procedures are 0 and are not shown.

Table 4.2: Summary of FDR-controlling procedure results.

		step-up	step-down		single-step
	unadj-p	BH[1]	Gavrilov[4]	BL[2]	empirical Bayes[3]
Oxychlordane	< 0.0001	0.0005	0.0005	0.0005	< 0.0001
Summed Dioxin-like PCBs	0.0002	0.0021	0.0020	0.0039	0.0004
Summed non Dioxin-like PCBs	0.0003	0.0022	0.0020	0.0054	0.0005
HCB	0.0007	0.0038	0.0033	0.0113	0.0026
Summed total PCBs	0.0009	0.0040	0.0033	0.0134	0.0033
PCB 153	0.0014	0.0049	0.0041	0.0187	0.0048
PCB 187	0.0016	0.0049	0.0041	0.0187	0.0053
PCB 180	0.0027	0.0073	0.0050	0.0267	0.0079
PCB 170	0.0033	0.0081	0.0052	0.0290	0.0097
PCB 156	0.0048	0.0099	0.0062	0.0355	0.0122
PCB 118	0.0049	0.0099	0.0062	0.0355	0.0124
Mirex	0.0065	0.0120	0.0062	$\underline{0.0355}$	0.0150
trans-Nonachlor	0.0127	0.0214	0.0099	0.0544	0.0236
PCB 138	0.0160	$\underline{0.0252}$	0.0105	0.0553	0.0281
p,p'-DDE	$\underline{0.0369}$	0.0540	0.0204	0.0944	$\underline{0.0486}$
PCB 99	0.0648	0.0891	0.0303	0.1191	0.0719
PCB 183	0.0824	0.1066	0.0317	0.1191	0.0856
β-HCCH	0.0885	0.1082	$\underline{0.0317}$	0.1191	0.0898
cis-Nonachlor	0.3659	0.4237	0.1215	0.1524	0.2517
p,p'-DDT	0.4759	0.5235	0.1362	0.1524	0.3029
PCB 105	0.5250	0.5500	0.1362	0.1524	0.3239
PCB 28	0.8429	0.8429	0.2438	0.1524	0.4339

[1] Benjamini and Hochberg procedure; [2] Benjamini and Liu procedure, the only MTP that assumes independent test statistics; [3] empirical Bayes procedure, the only joint MTP. [4] The only adaptive MTP estimates $\frac{\hat{m}_0}{m} = \frac{4.9}{22} = 0.22$, while all other MTPs take $\frac{m_0}{m} = 1$.

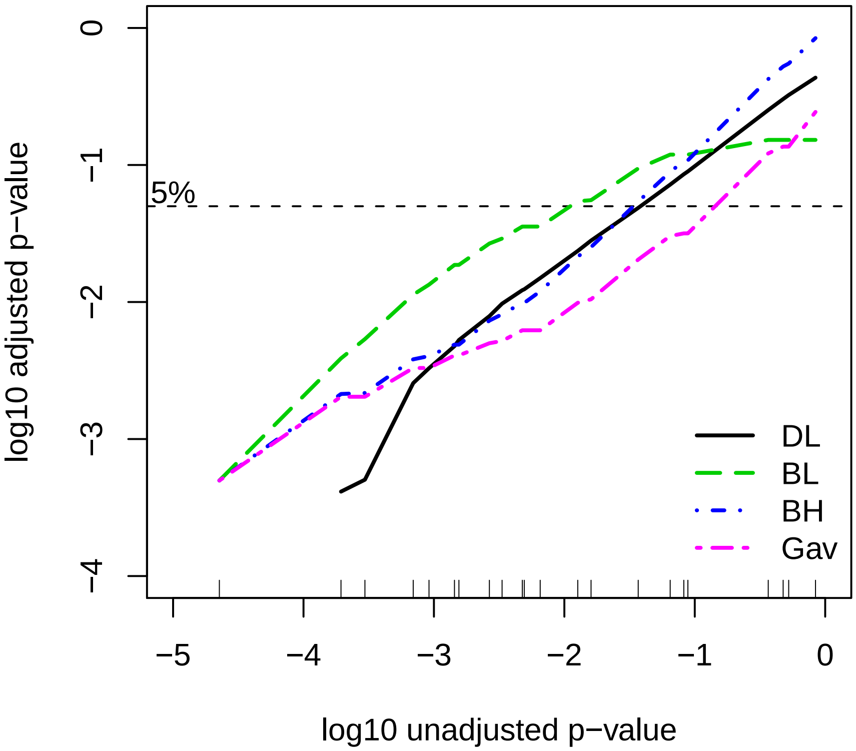

Figure 4.8: Relationship between adjusted and unadjusted p-values, on the $\log_{10}$-scale, for FDR-controlling procedures. 'DL', Dudoit and van der Laan's empirical Bayes procedure; 'BL', Benjamini and Liu procedure; 'BH', Benjamini and Hochberg procedure; 'Gav', Gavrilov procedure. The smallest adjusted p-values for the 'DL' and 'Gav' procedures are 0 and are not shown.

Chapter 5

Conclusions

5.1 Discussion

In this project, we discussed multiple testing procedures controlling the family-wise error rate and the false discovery rate using the organochlorine dataset from the NHL study.

The organochlorine analytes with large proportions (over 20%) of subjects below the detection limit tend to have large unadjusted p-values and low correlations with the test statistics. If we had assays with lower detection limit for these organochlorines, we may have found stronger associations between these organochlorines and NHL; and these organochlorines may have been highly correlated with the test statistics for other organochlorines.

For FWER-controlling procedures, the stepwise procedures have smaller adjusted p-values than the single-step procedures, as expected. This suggests that the stepwise procedures are more powerful than the single-step procedures. The joint procedures also tend to have smaller adjusted p-values than the marginal procedures, as expected. Thus, the joint procedures appear to be more powerful than the marginal procedures. Despite the high correlations between some of the organochlorine analytes (see Figure 4.4), the joint procedures which account for dependence structure among the test statistics do not lead to dramatic reductions in adjusted p-values. All the procedures controlling FWER perform similarly, and suggest more power than the Bonferroni procedure which is known to be conservative.

For FDR-controlling procedures, the Benjamini and Liu procedure, which assumes independent test statistics, stands out for rejecting two fewer hypotheses than the benchmark Benjamini and Hochberg procedure. The joint empirical Bayes procedure only rejects one more hypothesis than the marginal benchmark procedure, perhaps because the benchmark procedure can control FDR under positive regression dependency. All the test statistic correlations are positive for our dataset, and this situation is a special case of the positive regression dependency (Benjamini and Yekutieli, 2001). The most striking comparison against the benchmark procedure is for the adaptive Gavrilov procedure which rejects four more hypotheses. Interestingly, the marginal Gavrilov procedure rejects three more hypotheses than the joint empirical Bayes procedure. As test statistic correlations are quite high (see Figure 4.4), one might expect more power from the joint procedure. The reason for the apparent decreased power might be that we use $\pi_0 = m_0/m = 1$ to estimate the posterior probability in the empirical Bayes procedure, while the Gavrilov procedure uses the estimate $\hat{m}_0/m = 0.22$. In Chapter 4, from the boxplots of the marginal distributions of test statistics, we roughly estimated that only four out of twenty-two null hypotheses are true. Therefore, the empirical Bayes procedure is likely too conservative, and the Gavrilov procedure would be expected to have more power. For this specific dataset, taking advantage of small $\frac{m_0}{m}$ seems more important than taking advantage of dependence.

If an adaptive empirical Bayes procedure using the estimator of $\hat{m}_0$ in the Gavrilov procedure is applied, we expect more null hypotheses will be rejected. The test statistics are highly correlated and a joint procedure that takes advantage of the positive dependence structure among the test statistics, while simultaneously taking advantage of small m_0/m should reject at least the same number of null hypotheses as that rejected by the marginal Gavrilov procedure.

Finally, we did not consider adaptive FWER-controlling procedures in this project. However, if we consider an adaptive version of the Bonferroni procedure that also uses the estimator of $\hat{m}_0$ as in the Gavrilov procedure, this adaptive FWER-controlling procedure rejects five more null hypotheses than the non-adaptive Bonferroni procedure. Again, for this dataset, adaptive procedures appear to have better power than non-adaptive procedures.

5.2 Future work

A direction for future research would be to compare the statistical properties of the various MTPs by way of a simulation study. Knowing which null hypotheses are false, we could find which MTPs correctly reject the false null hypotheses for any simulated data set. This would allow us to understand which MTPs have optimal statistical properties in different situations controlled by simulation parameters such as:

- The number of tested null hypotheses, m.
- The proportion of tested null hypotheses that are true, m0/m.
- The dependence structure among the test statistics.

In the simulation study, we could also examine additional MTPs including those reviewed but not applied in this project.

In the future, we also plan to apply these MTPs to a new dataset from the NHL study, to analyze the associations between NHL and variants in candidate genes. This dataset has information for about 1500 genetic variants, known as single-nucleotide polymorphisms or SNPs, measured on the NHL cases and controls. In this project, we only tested 22 null hypotheses simultaneously. However, for the new dataset, we will test hundreds of null hypotheses.

Appendix A

R pseudo-code

A.1 Dudoit and van der Laan's empirical Bayes procedure

In this appendix, we will use the notation and bootstrapped data defined in the description of the Dudoit and van der Laan maxT/minP procedures in Chapter 3. We estimate the density f in equation (3.3) at a set of 13000 fixed points by

```
dens<-density(Tmat,n=13000,from=-4.5,to=8.499)
```

The limits of -4.5 and 8.499 are set so that all the observed values of the bootstrapped test statistics are included.

We use linear interpolation in `f.func()` to return the density at any value of a test statistic.

```
f.func<-function(t,dens) {
 temp.int<-floor((t--4.5)/0.001+1)
 temp.f<-dens$y[temp.int]+(t-dens$x[temp.int])/0.001
                         *(dens$y[temp.int+1]-dens$y[temp.int])
```

```
  return(temp.f)
 }
```

We program the posterior probability (q-value) function in equation (3.3) as

```
 q.func<-function(t,dens,pi0) {
  temp.q<-pi0*dnorm(t,0,var.tmat)/f.func(t,dens)
  temp.q<-pmin(1,temp.q)
  return(temp.q)
 }
```

where `var.tmat` is the variance of the bootstrapped test statistics and `pi0` is the π_0 value. Thus, the q-values can be calculated as

```
 q<-q.func(realT,dens,pi0)
```

where `realT` is the observed test statistics. Following step 1 of procedure 7.1 in Dudoit and van der Laan (2008, page 298-299), we use q to generate B binary vectors of length m representing "guessed" true and false null hypotheses for each bootstrap replicate. We put these binary vectors row-by-row into a $B \times m$ matrix `H`. The binary matrix `H` has elements coded as 1 for guessed true null hypotheses and 0 for guessed false null hypotheses. For example,

```
 H<-matrix(rbinom(n=B*m,size=1, q)),nrow=B,byrow=T)
```

Then we follow step 3 of the common cut-off version of the procedure, which is modified to control for FDR as suggested on page 319 of Dudoit and van der Laan (2008).

The matrix $\mathbf{Z}$ and a given common cut-off γ for the test statistics are used to create a matrix `R1` of rejected hypotheses under the bootstrap null distribution

```
R1<-matrix(NA, nrow=B, ncol=m)
for(i in 1:m)
 R1[,i]<-as.numeric(abs(Z[,i])> gamma)
```

The matrices `R1` and `H` are then used to create a vector `V` with `B` elements giving the number of rejected hypotheses in each bootstrapped null dataset, coming from null hypotheses guessed to be true. `V` is constructed so that, asymptotically, it tends to be larger than V, the random variable describing the number of incorrectly rejected null hypotheses under the true data-generating distribution.

```
V<-rowSums(H*R1)
```

Next we create a $B \times m$ matrix `R2` of rejected hypotheses under the bootstrap approximation to the true test statistics distribution:

```
R2<-matrix(NA,nrow=B, ncol=m)
for(i in 1:m)
 R2[,i]<-as.numeric(abs(T[,i]) > gamma)
```

The matrices `R2` and `H` are then used to create a vector `S` with `B` elements giving the guessed number of correctly rejected null hypotheses in each bootstrapped data set:

```
S<-rowSums((1-H)*R2)
```

To get the adjusted p-value for $H_{0(i)}$, $i = 1, \ldots, m$, we use $|t|_{(i)}$ as the common cut-off for the test statistics in the calculation of `R1` and `R2`. A candidate for the estimate of the adjusted p-value $p^*_{(i)}$ is the average of `V`/(`V` + `S`) over the B bootstrap replicates using the cut-off $|t|_{(i)}$. However, following equation (3.4), we take the minimum

$$p^*_{(i)} = \min_{j \in \{i+1, \ldots, m\}} \{p^*_{(j)}\}.$$

Bibliography

Agresti, A. and Franklin, C. (**2007**), *Statistics: The Art and Science of Learning from Data* (Prentice Hall), pp. 705, 730.

Baris, D., Kwak, L. W., Rothman, N., Wilson, W., Manns, A., Tarone, R. E., and Hartge, P. (**2000**), "Blood levels of organochlorines before and after chemotherapy among non-Hodgkin's lymphoma patients." Cancer Epidemiology, Biomarkers & Prevention **9**, 193–197.

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

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

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

Benjamini, Y. and Liu, W. (**1999**), "A step-down multiple hypotheses 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 of the multiple testing under dependency." The Annals of Statistics **29**, 1165–1188.

Chevrier, J., Dewailly, E., Ayotte, P., Mauriege, P., Despres, J., and Tremblay, A. (**2000**), "Body weight loss increases plasma and adipose tissue concentrations of potentially toxic pollutants in obese individuals." International Journal of Obesity **24**, 1272–1278.

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

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (**2004**), "Multiple testing. Part I. Single-step procedures for control of general Type I error rates." Statistical Applications in Genetics and Molecular Biology **3**.

Efron, B. (**2007**), "Size, power and false discovery rates." The Annals of Statistics **35**, 1351–1377.

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

Fisher, S. G. and Fisher, R. I. (**2004**), "The epidemiology of non-Hodgkin's lymphoma." Oncogene **24**, 6524–6534.

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

Ge, Y., Dudoit, S., and Speed, T. P. (**2003**), "Resampling-based multiple testing for microarray data analysis." TEST **12**, 1–77.

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

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

Ng, C. H.-M. (**2007**), *Plasma Organochlorines, Interaction between the Aryl Hydrocarbon Receptor Gene and Organochlorines, and Risk of non-Hodgkin Lymphoma.* (MSc Thesis, Department of Health Care and Epidemiology, University of British Columbia).

Pesarin, F. (**2001**), *Multivariate permutation tests : with applications in Biostatistics* (Wiley).

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

Sidak, Z. (**1971**), "On probabilities of rectangles in multivariate Student distributions: their dependence on correlations." Annals of Mathematical Statistics **42**, 169–175.

Spinelli, J. J., Ng, C. H., Weber, J.-P., Connors, J. M., Gascoyne, R. D., Lai, A. S., Brooks-Wilson, A. R., Le, N. D., Berry, B. R., and Gallagher, R. P. (**2007**), "Organochlorines and risk of non-Hodgkin lymphoma." International Journal of Cancer **121**, 2767–2775.

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

van der Laan, M. J., Birkner, M. D., and Hubbard, A. E. (**2005**), "Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the porportion of false positives." Statistical Applications in Genetics and Molecular Biology **4**.

Westfall, P. H. and Young, S. S. (**1993**), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment* (Wiley), pp. 72–74.

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