FALSE DISCOVERY RATES
THEORY AND APPLICATIONS TO DNA MICROARRAYS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

John David Storey
June 2002

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Robert Tibshirani
(Principal Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Iain Johnstone

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

David Siegmund

Approved for the University Committee on Graduate Studies:

Abstract

Multiple hypothesis testing is concerned with appropriately controlling the rate of false positives when testing several hypotheses simultaneously, while maintaining the power of each test as much as possible. One multiple hypothesis testing error measure is the False Discovery Rate (FDR), which is loosely defined to be the expected proportion of false positives among all significant hypotheses. The FDR is especially appropriate for exploratory analyses in which one is interested in finding many significant results among many tests. In this work, we introduce a modified version of the FDR called the "positive False Discovery Rate" (pFDR). We argue the pFDR is a more appropriate and useful error measure, and we investigate its statistical properties. When assuming the test statistics come from a mixture distribution, we show the pFDR can be written as a posterior probability and can be connected to classification theory. These properties remain asymptotically true under fairly general conditions, even under certain forms of dependence. Also, a new quantity called the "q-value" is introduced and investigated, which is a natural "Bayesian p-value", or rather the pFDR analogue of the p-value. This idea is also generalized to any multiple hypothesis testing error measure. Using these results, we introduce point estimates of the FDR and pFDR for fixed rejection regions. The point estimates provide proper conservative behavior in the three scenarios of (1) estimating false discovery rates for fixed rejection regions, (2) estimating rejection regions for fixed false discovery rates, and (3) simultaneously estimating false discovery rates over all possible rejection regions – even under certain forms of dependence. It is shown that this new set of methodology extends the current methodology and also provides increases in power. We apply the methodology to the problem of detecting differential gene expression between two or more biological samples based on DNA microarray data. This application is well suited because the dependence between the tests (genes) is weak and the number of tests is quite large.

Acknowledgments

Chapter 5 and part of Chapter 6 are joint work with Jonathan Taylor and David Siegmund. Chapter 8 is joint work with Rob Tibshirani. Any mistakes in these chapters or anywhere else is of course my fault.

This research was supported in part by an NSF Graduate Research Fellowship, a Stanford University Graduate Fellowship, and a Program in Mathematics and Molecular Biology National Fellowship.

Thanks to Professors Pat Brown, Iain Johnstone, David Siegmund, Jonathan Taylor, and Rob Tibshirani for being on my Ph.D. committee. Their ideas and advice have been a valuable contribution to this work.

Thanks to David Siegmund and Jonathan Taylor for being my unofficial co-advisors. Their generosity in both time and ideas have made this dissertation infinitely better. Thanks to Professor Rob Tibshirani for being a nearly perfect advisor. His friendship, guidance, and encouragement have given me the greatest experience.

I am grateful and indebted to Deborah, Dylan, Eric, Ji, Jimmy, Matt, and Steven for their friendship and support.

This dissertation is dedicated to my mother, father, and sister. Their love and support have carried me to where I am today.

Contents

Notation and Definitions

Outcomes from m hypothesis tests

	Accept	Reject	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
Total	W	R	m

Per Comparison Error Rate ..... $PCER = \mathbf{E}[V]/m$

Family Wise Error Rate ..... $FWER = \mathbf{Pr}(V \geq 1)$

False Discovery Rate ..... $FDR = \mathbf{E}[V/R|R > 0] \cdot \mathbf{Pr}(R > 0)$

positive False Discovery Rate ..... $pFDR = \mathbf{E}[V/R|R > 0]$

positive False Non-discovery Rate ..... $pFNR = \mathbf{E}[T/W|W > 0]$

Rejection regions ..... $[0, t],\ \Gamma,\ \Gamma_\alpha,\ \Gamma_t$

Random statistics ..... $X_1, X_2, \ldots, X_m$

Observed statistics ..... $x_1, x_2, \ldots, x_m$

Random p-values ..... $P_1, P_2, \ldots, P_m$

Observed p-values ..... $p_1, p_2, \ldots, p_m$

q-values ..... $\text{q-value}(t) = \inf_{t \in \Gamma_\alpha} pFDR(\Gamma_\alpha)$

Simultaneous FDR controlling curve ..... $\alpha_{FDR}(t) = \inf_{t \in \Gamma_\alpha} FDR(\Gamma_\alpha)$

Estimates for fixed rejection regions ..... $\widehat{FDR}_\lambda(\cdot),\ \widehat{pFDR}_\lambda(\cdot)$

Estimates for fixed FDR's ..... $t_\alpha\left[\widehat{FDR}_\lambda\right] = \sup\{t : \widehat{FDR}_\lambda(t) \leq \alpha\}$

Estimates for q-values and α_{FDR} ..... $\widehat{\text{q}}(\cdot),\ \widehat{\alpha}_{FDR}(\cdot)$

List of Tables

List of Figures

Chapter 1

Introduction

A major goal of statistics is to make decisions in the presence of uncertainty. Hypothesis testing, Bayesian decision theory, and classification theory all deal with this problem. This dissertation is concerned with making many related decisions simultaneously. We will specifically work within the hypothesis testing framework, but connections to Bayesian statistics and classification theory will emerge. A good starting point is to consider the basic ideas behind testing a single hypothesis, i.e., making a single decision between two choices.

1.1 Single Hypothesis Testing

Suppose we are given a set of data and we know that the data follow some distribution F_θ, where F_θ comes from a family of distributions indexed by $\theta \in \Omega$. Some subset of Ω, say, Ω_0 represents the *null hypothesis*, which is usually the state of θ that one hopes to find evidence against. Some other subset Ω_1 represents the *alternative hypothesis*. For example, if one were interested in the effect of a treatment, the null hypothesis Ω_0 would tend to be the set of θ that indicate the treatment had no effect, or the wrong effect. The alternative hypothesis Ω_1 would likely contain the set of θ that represent the treatment having the objective effect. It is always the case that $\Omega_0 \cap \Omega_1 = \emptyset$, and sometimes $\Omega_0 \cup \Omega_1 = \Omega$. If Ω_i consists of a single value, then this hypothesis is said to be *simple*. Otherwise, Ω_i is *composite*.

The *statistic* X is a function of the data that is the quantity used to decide whether $\theta \in \Omega_0$ or $\theta \in \Omega_1$. X can be anything from the unchanged set of data to a univariate quantity. If a statistic can be found such that the distribution of the data given the statistic

does not depend on θ, then the statistic is said to be *sufficient*. Therefore, a sufficient statistic can be used to test a hypothesis without any loss of information about θ. The decision is based on a rejection region, which we will denote by Γ. If $X \in \Gamma$, then we decide that the evidence favors $\theta \in \Omega_1$; if $X \notin \Gamma$, then we decide $\theta \in \Omega_0$.

There are two kinds of errors that can be committed when testing a hypothesis. The first is a *Type I error* (false positive), which occurs when $X \in \Gamma$ yet $\theta \in \Omega_0$. Therefore, the Type I error rate for a specific $\theta_0 \in \Omega_0$ is $\int_{\{X \in \Gamma\}} dF_{\theta_0}$, which we will denote by $\mathbf{Pr}(X \in \Gamma | \theta_0)$. The second class is a *Type II error* (false negative), and this occurs when $X \notin \Gamma$ yet $\theta \in \Omega_1$. The Type II error rate for a specific $\theta_1 \in \Omega_1$ is $\mathbf{Pr}(X \notin \Gamma | \theta_1)$.

The optimality criterion defined for hypothesis testing is to find the *most powerful test*. For a given $\theta_1 \in \Omega_1$, (X, Γ) represents the most powerful test of size $\sup_{\theta \in \Omega_0} \mathbf{Pr}(X \in \Gamma | \theta)$ if for all (Y, Δ) such that $\sup_{\theta \in \Omega_0} \mathbf{Pr}(X \in \Gamma | \theta) \leq \sup_{\theta \in \Omega_0} \mathbf{Pr}(Y \in \Delta | \theta)$, we have $\mathbf{Pr}(X \in \Gamma | \theta_1) \geq \mathbf{Pr}(Y \in \Delta | \theta_1)$. $\mathbf{Pr}(X \in \Gamma | \theta_1)$ is what we call the *power* of (X, Γ) at θ_1, and this is equal to 1 - Type II error rate; in words, the power is the probability of rejecting, given the alternative parameter is true. If (X, Γ) is most powerful for all $\theta_1 \in \Omega_1$, then it is said to be *uniformly most powerful*.

Given the optimality criterion for testing a single hypothesis, it makes sense to only consider nested sets of rejection regions. Therefore, if the null hypothesis is simple we can denote our set of nested rejection regions by $\{\Gamma_\alpha\}_{\alpha \in [0,1]}$, where $\alpha = \mathbf{Pr}(X \in \Gamma_\alpha | \theta_0)$. The nested property means that $\alpha' \leq \alpha$ implies $\Gamma_{\alpha'} \subseteq \Gamma_\alpha$. Lehmann (1986) covers hypothesis testing in detail, so the reader is referred there for a more thorough discussion of single hypothesis testing. For the remainder of this work, we will use the variable H to denote the state of the hypothesis: we let $H = 0$ when $\theta \in \Omega_0$ and $H = 1$ when $\theta \in \Omega_1$. Whether the hypothesis is simple or composite will be explicitly stated. We now consider testing multiple hypotheses simultaneously.

1.2 Multiple Hypothesis Testing

When testing multiple hypotheses, the situation becomes much more complicated. Now each test has possible Type I and Type II errors, and it becomes unclear how one should measure the overall error rate. Specifically, consider Table 1.1 that lists the possible outcomes when testing m hypotheses simultaneously.

For example, V is the number of Type I errors (false positives), T is the number of Type

Table 1.1: Outcomes from m Hypothesis Tests

	Accept	Reject	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
Total	W	R	m

II errors (false negatives), and $R = V + S$ is the total number of significant hypotheses. In order to measure the errors incurred in multiple hypothesis testing, it is convenient to define a compound error measure. Three that have been considered in the multiple hypothesis testing literature are defined as follows:

$$\begin{aligned} PCER &= \mathbf{E}[V]/m && \text{Per Comparison Error Rate,} \\ FWER &= \mathbf{Pr}(V \geq 1) && \text{Family Wise Error Rate,} \\ FDR &= \mathbf{E}[V/R|R > 0] \cdot \mathbf{Pr}(R > 0) && \text{False Discovery Rate.} \end{aligned}$$

The PCER and FWER have been used for many years, but the FDR was recently proposed by Benjamini & Hochberg (1995). The multiple hypothesis testing literature has mostly been concerned with deriving algorithms based on the order statistics in order to "control" the error rate of interest. We say the algorithm *strongly controls* the error rate if the inequality holds for all values of m_0 simultaneously. (In other words it is not necessary to include m_0 as an argument in the algorithm.) The algorithm *weakly controls* the error rate when it only holds for $m_0 = m$; that is, when all null hypotheses are true. Given the two types of control, Shaffer (1995) says that the multiple hypothesis testing literature indicates strong control is usually preferred over weak control, simply because it implies weak control and is adaptive over all possible values of m_0.

Without loss of generality we can consider multiple hypothesis testing algorithms to be carried out on the ordered p-values. The p-value of an observed statistic $X = x$ is defined to be (Lehmann 1986):

$$\text{p-value}(x) = \inf_{\{\Gamma_\alpha:\ x \in \Gamma_\alpha\}} \mathbf{Pr}(X \in \Gamma_\alpha | H = 0).$$

Therefore, since the set of rejected regions is nested, it can be seen that $x \in \Gamma_\alpha$ if and only if p-value$(x) \leq \alpha$. This implies the p-value is sufficient to base our decision on, given the

nested set of rejection regions and the statistic.

We will refer to the algorithms based on the order statistics that control these error measures as *sequential p-value methods*. This is how a sequential p-value method works: using the observed data, it estimates the rejection region so that on average $Err \leq \alpha$ for some pre-chosen α, where Err can be the FDR, FWER, or PCER. The product of a sequential p-value method is an estimate $\widehat{k}$ that tells us to reject the null hypotheses corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{(\widehat{k})}$, where $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ are the ordered, observed p-values.

For example taking $\widehat{k} = \max\{k : p_{(k)} \leq \alpha\}$ strongly controls the PCER at level α. (To be rigorous, we will let $\max \emptyset = 0$ and say that no hypothesis is rejected if $\widehat{k} = 0$.) The well known Bonferroni method $\widehat{k} = \max\{k : p_{(k)} \leq \alpha/m\}$ strongly controls the FWER at level α. Two other algorithms that strongly control the FWER are $\widehat{k} = \max\{k : p_{(j)} \leq \alpha/(m-j+1) \text{ for } j \leq k\}$ by Holm (1979), and $\widehat{k} = \max\{k : p_{(k)} \leq \alpha/(m-k+1)\}$ by Hochberg (1988). Shaffer (1995) provides an in-depth review of many of these methods.

Simes (1986) develops a method to weakly control the FWER (i.e., when $m_0 = m$). The algorithm is important for this work, so we summarize it in Algorithm 1.1. In recent work, Benjamini & Hochberg (1995) defined the FDR and proved that Simes's (1986) algorithm strongly controls the FDR. The FDR is roughly the expected proportion of false positives among all the rejected hypotheses, although it can be seen from its definition that V/R is set to zero when $R = 0$. In many situations, the FWER is much too strict, especially when the number of tests is large. Therefore, the FDR is a more liberal, yet more powerful quantity to control. In the next section, we consider a specific application in which the FDR is more appropriate than the FWER.

Algorithm 1.1: Simes (1986); Benjamini and Hochberg (1995)

1. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the observed, ordered p-values.
2. Calculate $\widehat{k} = \max\{k : p_{(k)} \leq \alpha \cdot k/m\}$.
3. Reject the null hypotheses corresponding to $p_{(1)}, \ldots, p_{(\widehat{k})}$.

Throughout this work, we will refer to Simes's (1986) algorithm as the BH procedure, mainly because we are interested in the fact that it strongly controls the FDR. It can be

seen in Figure 1.1 that the algorithm is easy to apply: plot each p-value versus its rank and reject all p-values less than or equal to the largest that falls under the line drawn through the origin with slope α/m. On the other hand, why does this algorithm work? This is not clear, as Benjamini & Hochberg (1995) prove strong control via an induction argument. Also, it is not clear why the sequential p-value approach is always the most appropriate for multiple hypothesis testing.

For example, what can we say about $\hat{k}$? Is there any natural way to provide an error measure on this random variable? It is a false sense of security in multiple hypothesis testing to think that we have a 100% guaranteed upper bound on the error. The reality is that this process involves estimation. The more variable the estimate of $\widehat{k}$ is, the worse the procedure is going to work in practice. Therefore, the expected value may be that FDR $\leq \alpha$, but we do not know how reliable the methods are on a case by case basis. If point estimation only involved finding unbiased estimators, then the methods wouldn't be incredibly reliable. Therefore, the reliability of $\widehat{k}$ on a case by case basis does matter even though it has not been explored.

Interestingly, the paradigm put forth in the multiple hypothesis testing literature is quite different than what is often done in practice. In the multiple hypothesis testing literature (1) the sequential p-value methods are non-parametric, and almost always rely on some sort of independence assumption, (2) strong control is heavily stressed, and (3) the error rate is chosen beforehand and the rejection region is estimated to attain conservative bias of it. On the other hand, in specific scientific settings, it is often the case that (1) methods are developed that take into account the structure of the problem at hand, including dependence, (2) considering the complete null case (weak control) suffices, and (3) the rejection region is fixed and the error rate is estimated. For example, the three seminal papers Karlin & Altschul (1990), Feingold, Brown & Siegmund (1993), and Worsley, Marrett, Neelin, Vandal, Friston & Evans (1996) all follow this latter paradigm in estimating the FWER for a given threshold.

This phenomenon is not too surprising in that large scale multiple testing problems are usually not comprised of thousands of independent experiments, but rather of thousands of potentially significant features coming from some overall process. Sequential p-value methods are therefore more appropriate for cases in which a few multiple comparisons are taken into account. False discovery rates are clearly not that applicable to only a few tests. Indeed, the strength of the FDR lies in finding many significant results among many

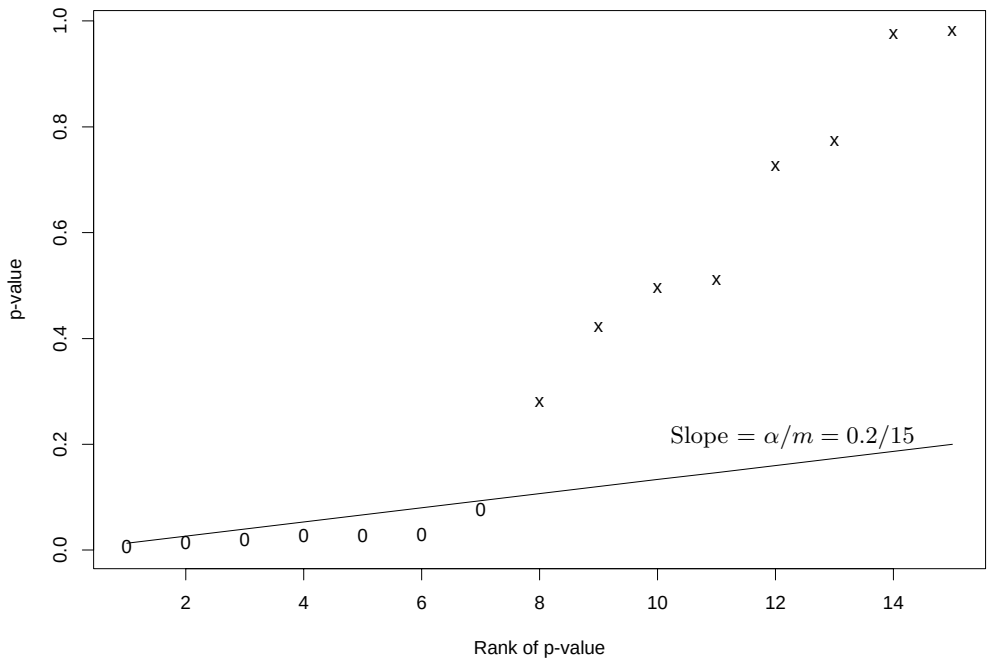


Figure 1.1: Simes's (1986) algorithm applied to 15 p-values with $\alpha = 0.20$. This algorithm is equivalent to plotting each p-value versus its rank and rejecting all p-values less than or equal to the largest that falls under the line drawn through the origin with slope α/m.

hypothesis tests, while limiting the proportion of false positives among these.

In this work, we take a new approach to false discovery rates. We attempt to use more traditional and straightforward statistical ideas to deal with the FDR. Instead of fixing α and considering ways to estimate k, we begin by considering the FDR in the context of fixed rejection regions. In the following section, we motivate our interest in the FDR in the context of a new scientific problem: detecting which of thousands of genes show differential expression between two or more biological samples. In Section 1.4, we explicitly describe the false discovery rate paradigm we develop.

1.3 DNA Microarrays

The focus of genetics has recently turned from the characterization of genes and pathways on a case by case basis to that at the genomic level. The sequencing of various genomes,

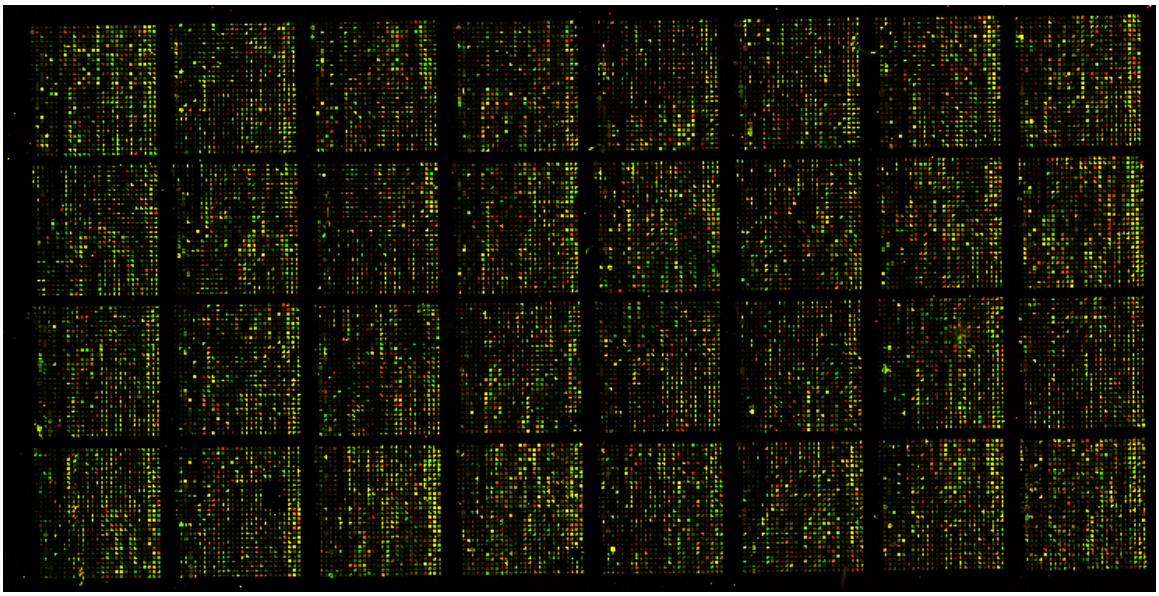


Figure 1.2: A human cDNA array. Each spot represents a different gene, and the colors represent the relative abundance of the mRNA from the red channel versus the mRNA from the green channel.

along with a surge in high-throughput technologies, has made it possible to obtain data on thousands of genes from a single biological sample. Besides high-throughput genome-wide typing of genetic markers, which can be used to map genes associated with complex diseases and understand evolution, several technologies have recently been developed that can simultaneously measure the expression levels of thousands of genes. Schena, Shalon, Davis & Brown (1995), Velculescu, Zhang, Vogelstein & Kinzler (1995), and Lockhart, Dong, Byrne, Follettie, Gallo, Chee, Mittmann, Wang, Kobayashi, Horton & Brown (1996) introduce three popular technologies that efficiently and rapidly perform this task.

Each of these technologies presents a novel method for measuring the expression levels of thousands of genes from a single biological sample. Rather than explaining the unique properties of each one, we will instead give the basic idea as to what they are accomplishing. Recall that more or less every cell of an organism contains a copy of its entire genome. The genes are encoded in the DNA of the genome. The "central dogma" of genetics can be represented by the following (oversimplified) schematic:

$$\text{DNA} \overset{transcription}{\longrightarrow} \text{RNA} \overset{translation}{\longrightarrow} \text{Protein.}$$

That is, the DNA encoding the gene is transcribed into RNA; the RNA is then translated into a protein that performs some structural or catalytic function. Based on the role and environment of the cell, this process is carried out at different levels of frequency for each gene. Moreover, genes interact in pathways (or groups with some structured ordering) to perform the complicated tasks that keep a cell functioning properly. The essential feature of the technologies put forth by Schena et al. (1995), Velculescu et al. (1995), and Lockhart et al. (1996) is that they each measure the abundance of mRNA in the sampled cells. In Schena et al. (1995) and Velculescu et al. (1995) this measurement is on a continuous scale, whereas in Lockhart et al. (1996) it is on a discrete scale.

Figure 1.2 shows a cDNA array (Schena et al. 1995) of about 15,000 human genes. It can be seen that there are red, green, yellow, and black spots present on the array. In this type of array, two biological samples are hybridized to the same slide. One of them is usually a reference (control) sample, and the other is the sample of interest. mRNA is extracted from each sample and is reverse transcribed to the gene's coding DNA. A green fluorescent dye is attached to the DNA from one sample, and a red fluorescent dye to the DNA from the other sample. The cDNA for a particular gene is attached to a unique spot on the array. The two samples are then applied to the array, and the cDNA (with dye attached) hybridizes to its complementary strand, at its particular spot. The signal of green and red dyes are measured for each spot; this gives a noisy measurement of the abundance of mRNA present in each sample for each gene. Usually the $\log_2$ ratio of the red to green signals is the number reported for each spot. Therefore, in Figure 1.2 green indicates the mRNA (corresponding to the gene at that specific spot) from the green sample is much stronger than the red, yellow indicates they are about equal, red indicates the red channel is more abundant than the green, and black indicates there is a negligible amount of mRNA present in either sample.

If one were to obtain n of these arrays, each consisting of the same m genes, then the data organized into an $m \times n$ matrix would look something like this:

	array 1	array 2	array 3	array 4	$\cdots$	array n
gene 1	1.23	-2.61	-3.57	4.22	$\cdots$	5.12
gene 2	3.98	-0.294	1.73	2.97	$\cdots$	-2.43
$\vdots$			$\vdots$			$\vdots$
gene m	0.846	3.72	1.83	-1.10	$\cdots$	-2.94

Usually the n arrays would consist of samples from two or more conditions, hopefully with replicates for each type. For example, one could gather arrays from a culture of yeast over time (Spellman et al. 1998), from tumor tissues of different cancers (Alizadeh et al. 2000), or from irradiated and untreated human cells (Tusher, Tibshirani & Chu 2001).

Some immediate statistical issues one has to deal with are the quantification of the fluorescence signals from each spot (Yang, Buckley, Dudoit & Speed 2002) and the normalization of the data across arrays (Yang, Dudoit, Luu, Lin, Peng, Ngai & Speed 2002). Also, one may apply clustering techniques to organize and visualize the data (Eisen, Spellman, Brown & Botstein 1998). Even though these are all important problems to consider, none involves making statistical inference about the behavior of the genes' expression. Consider the experiment analyzed in Tusher et al. (2001), where human cells were hybridized to eight oligonucleotide arrays. Four arrays measured the expression levels in untreated cells, and the other four measured the expression levels in irradiated cells. Each array was composed of the same 6800 genes, so that $m = 6800$ and $n = 8$. These data are very typical of that currently obtained in microarray experiments: usually we have $3000 \leq m \leq 30,000$ and $4 \leq n \leq 100$. This obviously presents new statistical challenges in that the number of covariates is large and the sample size is relatively small.

A basic yet important question we can ask is *which genes show a statistically significant change in gene expression* between the untreated and irradiated cells. Answering this question helps the biologist narrow down the search and analysis of these genes from thousands to potentially several dozen. The genetic response to radiation is obviously complex, and we would expect the expression protocol of many genes to change. This would also be the case when a tissue develops cancer, or even when observing the differences between two individuals. For example, the rapid identification of genes whose expression changes in cancer is a potentially very powerful tool. Therefore, it is of great interest to develop statistical methodology to detect differential gene expression from microarray data .

The problem of detecting differential gene expression can be broken down into four steps:

(i) Forming a statistic for each gene,
(ii) Calculating the null distribution(s) for the statistics,
(iii) Choosing the significance regions,
(iv) Assessing the false positives.

One can treat each gene as a separate experiment and perform steps (i)-(iii) within each gene as in Dudoit, Yang, Callow & Speed (2002), or one can try to "borrow strength"

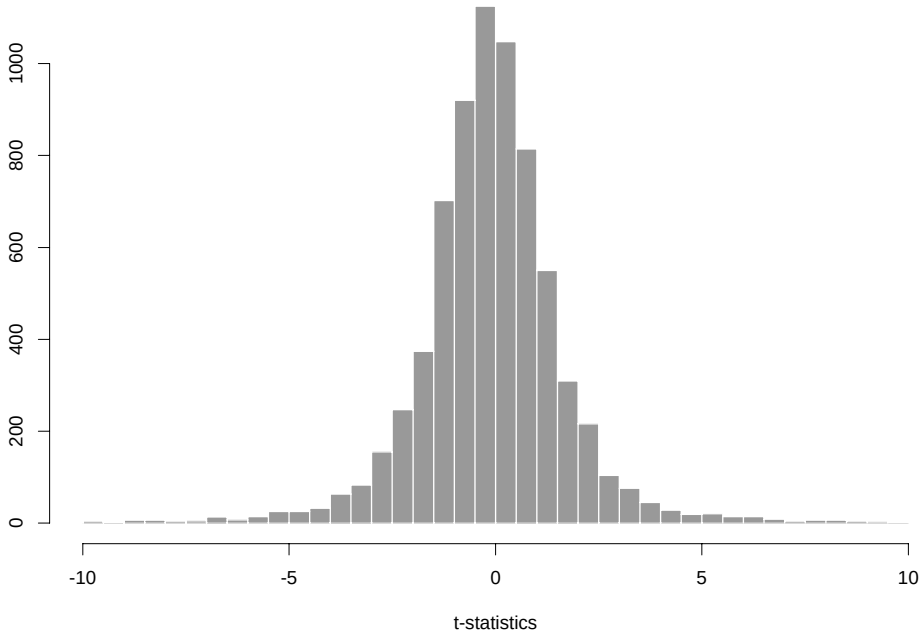


Figure 1.3: 6800 t-statistics calculated from the Tusher et al. (2001) data.

across the genes by applying a more global model as in Tusher et al. (2001) and Efron, Tibshirani, Storey & Tusher (2001). These issues will be more carefully discussed in Chapter 8. Regardless of the approach taken in steps (i)-(iii), step (iv) must be performed carefully, keeping in mind the goal of the experiment.

Using the data from Tusher et al. (2001), we formed a paired t-statistic for each gene. (The pairing was done based on the cell line and aliquot memberships.) The histogram of these 6800 t-statistics can be seen in Figure 1.3. Suppose we now simply calculate the null distribution by permuting the irradiated and untreated labels, and re-calculating the t-statistics at each permutation. Also, suppose we use symmetric significance regions and calculate a p-value for each gene. If we were to call all genes significant whose p-value was less than 0.01, then we would expect 68 false positives under the full null model – a number much too large given the scientist cannot possibly spend time further investigating 68 false positive genes. If instead, we control the FWER at level 0.01 with the Bonferroni correction, then we find only two genes significant – a number much too small given that

many genes should have changed.

Uncorrected testing prevents us from losing any power, yet the false positives are numerous; the FWER strongly prevents many false positives from occurring, yet the power of each test is greatly diminished. As opposed to many multiple testing scenarios, the goal of microarray experiments is to identify as many significant findings as possible so that further experiments and analyses on the genes can be performed. However, among these significant findings, one does not want the rate of false positives to become overly burdensome. Therefore, a delicate balance between power and Type I error must be attained, and false discovery rates are ideal for doing this. Indeed, false discovery rates measure the proportion of false positives among the significant findings, a quantity the biologist can easily put into use. Therefore, we use this important scientific problem as our motivation for studying false discovery rates. What is of particular importance is furthering the interpretation and understanding of false discovery rates, developing more powerful and flexible techniques, and considering the special type of dependence that occurs between the genes in microarray experiments.

1.4 A New Approach to False Discovery Rates

In this work, we present a new approach to false discovery rates. We directly study false discovery rates as being calculated over fixed rejection regions. This approach leads to a develop of the interpretation and definition of false discovery rates, as well as a point estimation approach that allows us to derive more powerful methodologies in a variety of settings. Recall that Benjamini & Hochberg (1995) defined the FDR to be

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

This definition is used rather than $\mathbf{E}[V/R]$ because it is almost always the case that $R = 0$ with positive probability. The additional term $\mathbf{Pr}(R > 0)$ can be problematic and obfuscate the interpretation of the FDR. In Chapter 2, we define a new error measure called the positive False Discovery Rate (pFDR):

$$pFDR = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right],$$

where we have conditioned on the event that at least one positive finding has occurred. The pFDR has a Bayesian interpretation in that it can be written as a simple posterior probability. Connections between the pFDR and classification theory will also be made. In Chapter 3, we define a new quantity called the q-value, which is the natural Bayesian (or pFDR) counterpart to the p-value. This also allows the more general concept of simultaneous controlling curves to be developed.

As mentioned in Section 1.2, the Benjamini & Hochberg (1995) procedure (Algorithm 1.1) is an example of the traditional approach to multiple hypothesis testing. We choose the level α at which to control the FDR. We then observe m p-values, and an algorithm is applied to $p_1, \ldots, p_m$ and α to yield an estimate $\widehat{k}$. The hypotheses corresponding to $p_{(1)} \leq \cdots \leq p_{(\widehat{k})}$ are called significant. The Benjamini & Hochberg (1995) procedure controls the FDR at level α when taking the long run frequency of $V/R \cdot 1(R > 0)$ over repeated applications.

The FWER represents the probability of a 0-1 outcome (i.e., either $V \geq 1$ or $V = 0$). Therefore, when using the FWER, it is easy to decide beforehand what one wants the probability of this event to be. The FDR on the other hand is more of an exploratory data analysis tool. Therefore, choosing the level α at which to control the FDR can be as arbitrary as fixing the rejection region beforehand. For example, suppose we decided to control the FDR at level 5% for 1000 hypothesis tests. The appropriateness of this choice largely depends on how many tests are significant. A 5% false discovery rate resulting in 100 significant tests is much different than one resulting in 3 significant tests. If one has a lot of experience with a certain type of data, then it may make more sense to fix the rejection region.

Benjamini & Hochberg (1995), Benjamini & Liu (1999), and Benjamini & Hochberg (2000) all treat the FDR in the typical FWER context of fixing the error rate beforehand. However, there are cases where it makes more sense to fix the rejection region beforehand and estimate the FDR. (This also allows one to estimate the potentially more appropriate error measure, the pFDR.) Due to the exploratory nature of false discovery rates, it often makes the most sense to let the rejection regions be chosen by the statistics themselves. In terms of p-values, this simply means that all rejection regions of the form $[0, p_i]$ are considered. That way, neither the error rate nor the rejection region must be chosen beforehand.

When fixing the error rate beforehand, the goal is to derive a procedure so that the

true error rate in expectation is less than or equal to the pre-chosen one. When fixing the rejection region, the goal is to provide an estimate whose expectation is greater than or equal to the true error rate over that rejection region. When letting the rejection regions be random, the goal is to provide a conservative estimate of the error rate over all rejection regions *simultaneously*.

It is unclear from Benjamini & Hochberg (1995) why Simes's (1986) algorithm was used, and why it works. We hope to explain why the procedure works, introduce more powerful methodologies, and motivate a general approach to false discovery rates. A major point of this work is to show that by taking a point estimation approach, the three scenarios of *(1) fixing the rejection region beforehand*, *(2) fixing the error rate beforehand*, and *(3) estimating the error rate over the naturally (and randomly) chosen rejection regions* can be studied within a single, unified framework. We now describe these three scenarios in more detail.

Scenario # 1: Estimating the error rate for a fixed rejection region

In this setting, we would like to estimate the FDR (or pFDR) when rejecting all p-values in $[0, t]$, which we denote by $FDR(t)$. This scenario is the "inverse" of the original motivation of FDR as in Benjamini & Hochberg (1995); instead of estimating the rejection region to use for a given α (i.e., forming $\widehat{k}$), we fix the rejection region and estimate the FDR. Details of the estimates $\widehat{FDR}$ and $\widehat{pFDR}$ are presented in Chapter 4. In many cases, this approach offers increased flexibility and power over the BH procedure. Further, for any given rejection region $[0, t]$, $\widehat{FDR}(t)$ provides its own version of strong control in that $\mathbf{E}[\widehat{FDR}(t)] \geq FDR(t)$ for all values of m_0 – that is, the estimate tends to be conservative on average. The analogous result also holds for the pFDR.

Scenario # 2: Estimating the rejection region for a fixed the error rate beforehand

On the other hand, one may want to fix the error rate beforehand in certain situations, as when controlling FWER rates or in Benjamini & Hochberg's (1995) original introduction of the FDR. Therefore, it is tempting to find the largest rejection so that $\widehat{FDR}(t) \leq \alpha$ and reject all p-values falling in that region. In other words we set

$$t_\alpha = \sup\{t : \widehat{FDR}(t) \leq \alpha\}, \tag{1.1}$$

and reject all p-values in $[0, t_\alpha]$. We show in Chapter 5 that this approach (with some trivial modifications) provides strong control of the FDR in the traditional sense. We call the approach given by (1.1) an "estimation based FDR controlling procedure." That is, we show that strong control of the FDR can be provided by using the point estimate $\widehat{FDR}(t)$. Moreover, the BH procedure can be re-expressed as an estimation based FDR controlling procedure using the most conservative version of $\widehat{FDR}(t)$.

Scenario # 3: Estimating the error rate over the naturally (and randomly) chosen rejection regions

In this situation, we need to show that the estimates used in Scenario #1 work well over all rejection regions simultaneously. In other words, in some sense we have $\widehat{FDR}(\cdot) \geq FDR(\cdot)$ or $\widehat{pFDR}(\cdot) \geq pFDR(\cdot)$ regardless of the rejection region. This is the ideal approach in an exploratory setting because it allows one to choose the rejection region with respect to the values of the estimates without the risk of introducing bias. We show when the $\widehat{FDR}(\cdot)$ and $\widehat{pFDR}(\cdot)$ estimates are conservatively consistent simultaneously over all rejection regions, which is the exact property we desire. This property even holds under certain types of dependence seen in applications, such as DNA microarray data. In Chapter 3, we define a new function, called a simultaneous controlling curve, that gives the minimum error rate that can be achieved over each rejection region. By using $\widehat{FDR}(\cdot)$ and $\widehat{pFDR}(\cdot)$, we are able to conservatively estimate their simultaneous controlling curves.

In investigating the interpretation and definition of false discovery rates and their estimation in the three scenarios, we make some progress in dealing with the case of dependence between the tests. The explicit application of the new approach to false discovery rates to DNA microarrays is described in Chapter 8.

1.5 An Outline

This work can briefly be outlined as follows. Chapter 1 introduces some of the basic ideas we present in this work. Chapter 2 introduces the positive False Discovery Rate and connects this quantity to a Bayesian posterior probability and classification theory. Chapter 3 introduces the q-value and simultaneous controlling curves. Chapters 4, 5, and 6 introduce methodologies to estimate the quantities described in the aforementioned three scenarios. Chapter 7 describes methods for automatically choosing a tuning parameter present in the

estimates. Chapter 8 describes the application of the methodologies to the DNA microarray problem introduced in Section 1.3. Chapter 9 makes some brief concluding remarks.

Much of this work will appear in the forthcoming papers Storey (2001a), Storey (2001b), Storey, Taylor & Siegmund (2002), and Storey & Tibshirani (2001).

Chapter 2

The Positive False Discovery Rate

In this chapter we attempt to elucidate false discovery rates in a broad statistical sense by defining and investigating the pFDR and by making connections to other ideas in statistics. For example, hypothesis testing is traditionally known a frequentist procedure. However, classical classification theory seems to be a bridge between Bayesian modeling and hypothesis testing. In the context of the pFDR, this bridge becomes clearer. The pFDR offers the potential to be a tool for simultaneous decision making useful to both frequentists and Bayesians.

2.1 A New Multiple Hypothesis Testing Error Measure

The most natural definition of a "false discovery rate" is $\mathbf{E}[V/R]$. However, in most cases $R = 0$ with positive probability, so this definition is useless. Three natural choices for quantities that remedy the $R = 0$ problem are:

$$\begin{array}{ll} \text{(A)} & \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right] \cdot \mathbf{Pr}\left(R > 0\right) \\ \text{(B)} & \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right] \\ \text{(C)} & \mathbf{E}[V]/\mathbf{E}[R] \end{array}$$

Benjamini & Hochberg (1995) point out that if all null hypotheses are true ($m_0 = m$) then $\mathbf{E}[V/R|R > 0] = 1$ and $\mathbf{E}[V]/\mathbf{E}[R] = 1$, so neither quantity can be strongly controlled in the traditional p-value based framework. (That is, since when $m_0 = m$, (B) = (C) = 1, one can never choose a value α beforehand, and guarantee that for all m_0 (B) and (C) are less than or equal to α.) Therefore, they choose to work with definition (A) even

though much of their discussion of the FDR seems to point to quantity (B). Definition (C) is appealing because of it's simplicity, but the other two quantities measure the behavior of V and R simultaneously. One could argue that the false discovery rate is one when all null hypotheses are true, especially when we are only concerned with cases where discoveries are made. Therefore, definition (B) is more appealing under this consideration.

Definition (A) of the FDR can be written in words as "the rate that false discoveries occur", whereas definition (B) can be written as "the rate that discoveries are false." One has to ask which definition is more useful in practice. When a scientist is testing multiple hypotheses, s/he is probably not interested in cases where nothing is significant, much less in controlling a quantity that involves cases where nothing is significant.

An example where confusion between definitions (A) and (B) is dangerous is the following. One can use the Benjamini & Hochberg (1995) procedure to yield on average that $\mathbf{E}[V/R|R > 0]\mathbf{Pr}(R > 0) \leq 0.1$. But what if $\mathbf{Pr}(R > 0) = 0.5$? Then we have actually only controlled $\mathbf{E}[V/R|R > 0] \leq 0.2$, a quantity twice as large! One may suppose this example is hypothetical, but this exact confusion arises in Weller, Song, Heyen, Lewin & Ron (1998). Zaykin, Young & Westfall (1998) show that the results of Weller et al. (1998) can be very misleading if definitions (A) and (B) are confused. Also, Shaffer (1995) states that the inclusion of $\mathbf{Pr}(R > 0)$ into the definition of the FDR is unsatisfying.

Therefore, it is our opinion that definition (A) covers a situation that is much broader than what the researcher needs. Given a small number of tests or when dependence exists, $\mathbf{Pr}(R > 0)$ may be somewhat less than one. In this case definition (A) is not very useful to the researcher who has a significant finding. Definition (B) covers the exact situation about which the researcher is concerned.

Given this discussion, it seems appropriate to consider a different definition of a "false discovery rate" multiple hypothesis testing error measure than definition (A). We propose (B) as a new definition as was mentioned in the introduction. Hopefully, it will be shown that the modified definition of the FDR is intuitively pleasing as well as mathematically tractable. To this end, we propose the following alternative quantity to the FDR, which we call the *positive False Discovery Rate* because it is conditioned on the fact that at least one positive finding has occurred.

Definition 2.1 *We define the* **positive False Discovery Rate** – the rate that discoveries

are false – *to be:*

$$pFDR = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right].$$

Using the FDR over the pFDR is necessary when one chooses the error rate beforehand and estimates a rejection region to fall below that error rate. That is, the pFDR quantity must be calculated over fixed rejection regions. By considering false discovery rates for fixed rejection regions, one can gain insight into the operating characteristics of the quantities, and make improvements over the Benjamini & Hochberg (1995) methodology. Using the results of this work, Chapter 4 develops conservative point estimates for both the FDR and pFDR that show improvements over the power of the Benjamini & Hochberg (1995) methodology. Taking this a step further, Chapter 5 uses these point estimates to develop a new FDR controlling method. Chapter 6 uses the point estimates to conservatively estimate the FDR and pFDR over all rejection regions *simultaneously*. Clearly this last scenario is the most useful, as it allows the researcher to perform a truly exploratory analysis. Hence, it is clear that the pFDR is a useful quantity to study, and one can overcome Benjamini & Hochberg's (1995) concerns.

2.2 A Bayesian Interpretation

In this section we present a Bayesian interpretation of the pFDR. As it turns out, the pFDR can be written as a simple posterior probability under certain assumptions. Instead of basing the pFDR on p-values, we will use general statistics and a nested set of rejection regions. We will also treat the pFDR as being calculated on fixed rejection regions.

Suppose we wish to perform m identical tests of a null hypothesis versus an alternative hypothesis based on the statistics $X_1, X_2, \ldots, X_m$. For a given rejection region Γ, define the positive False Discovery Rate as we defined it in Section 2:

$$pFDR(\Gamma) = \mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right],$$

where $V(\Gamma) = \#\{\text{ null } X_i : X_i \in \Gamma\}$ and $R(\Gamma) = \#\{X_i : X_i \in \Gamma\}$. Let $H_i = 0$ when the i^{th} null hypothesis is true and $H_i = 1$ when the alternative is true, $i = 1, \ldots, m$. Let π_0 be the *a priori* probability that a hypothesis is true. That is, we assume that the H_i are i.i.d. Bernoulli random variables with $\mathbf{Pr}(H_i = 0) = \pi_0$ and $\mathbf{Pr}(H_i = 1) = 1 - \pi_0 = \pi_1$.

Before we present the Bayesian form of the pFDR, consider the pFDR when $m = 1$.

In this case, it is also easily seen that $pFDR(\Gamma) = \mathbf{Pr}(H = 0|X \in \Gamma)$. Fortunately, when $m > 1$, this result does not change.

Theorem 2.1 *Suppose m identical hypothesis tests are performed with the statistics $X_1, \ldots, X_m$ and rejection region Γ. Assume that $X_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot F_0 + H_i \cdot F_1$ for null distribution F_0 and alternative distribution F_1, and assume $H_i \overset{i.i.d.}{\sim}$ Bernoulli (π_1) for $i = 1, \ldots m$. Then*

$$pFDR(\Gamma) = \mathbf{Pr}(H = 0|X \in \Gamma), \tag{2.1}$$

where $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.

It is surprising that the pFDR, a compound error measure, can be written in such a simple way. Moreover, the posterior probability (2.1) does not depend on m. (Also note that $\mathbf{Pr}(H_i = 0|X_i \in \Gamma)$ is the same for each $i = 1, \ldots, m$, which is why we left out the index in the statement of the theorem.) We can explicitly write

$$\begin{aligned} pFDR(\Gamma) &= \mathbf{Pr}(H = 0|X \in \Gamma) \\ &= \frac{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma|H = 0)}{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma|H = 0) + \pi_1 \cdot \mathbf{Pr}(X \in \Gamma|H = 1)} \\ &= \frac{\pi_0 \cdot \{\text{Type I error of } \Gamma\}}{\pi_0 \cdot \{\text{Type I error of } \Gamma\} + \pi_1 \cdot \{\text{Power of } \Gamma\}}. \end{aligned}$$

This shows that the pFDR increases with increasing Type I error and decreases with increasing power. We now prove Theorem 2.1.

Proof: First note that

$$\begin{aligned} pFDR(\Gamma) &= \mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right] \\ &= \sum_{k=1}^{m} \mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) = k\right] \mathbf{Pr}(R(\Gamma) = k|R(\Gamma) > 0) \\ &= \sum_{k=1}^{m} \mathbf{E}\left[\frac{V(\Gamma)}{k}\middle| R(\Gamma) = k\right] \mathbf{Pr}(R(\Gamma) = k|R(\Gamma) > 0). \end{aligned}$$

Since the statistics are independent, it intuitively follows that $V(\Gamma)|R(\Gamma) = k$ is a binomial random variable with probability of success $\mathbf{Pr}(H = 0|X \in \Gamma)$, in which case the proof easily follows. However, we can be more precise. Because of the i.i.d. assumption, it follows

that

$$\begin{aligned}
\mathbf{E}\left[V(\Gamma)|\, R(\Gamma)=k\right] &= \mathbf{E}\left[\sum_{i=1}^{m} 1(X_i\in\Gamma)1(H_i=0)\,\middle|\, \begin{array}{c} X_1,\ldots,X_k\in\Gamma \\ X_{k+1},\ldots,X_m\notin\Gamma \end{array}\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{k} 1(H_i=0)\,\middle|\, \begin{array}{c} X_1,\ldots,X_k\in\Gamma \\ X_{k+1},\ldots,X_m\notin\Gamma \end{array}\right] \\
&= \sum_{i=1}^{k}\mathbf{E}\left[1(H_i=0)\,|X_i\in\Gamma\right] \\
&= k\cdot\mathbf{Pr}(H=0|X\in\Gamma).
\end{aligned}$$

Therefore,

$$\begin{aligned}
pFDR(\Gamma) &= \sum_{k=1}^{m}\frac{k\cdot\mathbf{Pr}(H=0|X\in\Gamma)}{k}\,\mathbf{Pr}(R(\Gamma)=k|R(\Gamma)>0) \\
&= \mathbf{Pr}(H=0|X\in\Gamma).
\end{aligned}$$

□

Note that when the H_i are not random, then this theorem no longer holds since there is the deterministic constraint that $\sum_{i=1}^{m} H_i = m_1$. However, for large m, the mixture distribution assumption can hold; this is formally dealt with in Section 2.3. Also, this theorem holds for a simple versus simple test, but composite hypotheses can also be considered as long as one models the alternative parameter as a random variable. Then F_1 is simply the mixture of the alternative distributions.

Two corollaries easily follow from Theorem 2.1. The first shows that the Bayesian interpretation holds when we condition on $R(\Gamma)=k$ for any $k>0$. This result was shown as part of the previous proof.

Corollary 2.1 *Under the assumptions of Theorem 2.1, for $k>0$ we have*

$$\mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\,\middle|\, R(\Gamma)=k\right]=\mathbf{Pr}(H=0|X\in\Gamma).$$

Recall that in Section 2.1, we considered three definitions, the third of which was $\mathbf{E}[V]/\mathbf{E}[R]$. This definition is equivalent to the pFDR under the assumptions of Theorem 2.1.

Corollary 2.2 *Under the assumptions of Theorem 2.1,*

$$\mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right] = \frac{\mathbf{E}[V(\Gamma)]}{\mathbf{E}[R(\Gamma)]}.$$

Writing the pFDR as $P(H = 0|X \in \Gamma)$ is very similar to the Type I error. One could call it a "posterior Bayesian Type I error." (See Morton (1955) for a very similar development of this concept in the context of genetic linkage analysis.) Whereas the FWER is very much frequentist, we have shown that the pFDR is quite flexible in its interpretation. This is especially appealing in that it is a multiple testing error measure that can be used by both Bayesians and frequentists. We will see in later examples that this is easily accomplished.

$pFDR(\Gamma) = \mathbf{Pr}(H = 0|X \in \Gamma)$ gives a global measure in that it doesn't provide specific information about the value of each statistic – only whether it fell in Γ or not. In Section 3.1, we use the pFDR to give each statistic a measure of its significance in terms of the pFDR, which we call the q-value. This continues to have a Bayesian interpretation, yet allows one to make simultaneous inferences. Reporting marginal posterior probabilities $\mathbf{Pr}(H = 0|X = x)$, as is the case in typical Bayesian modeling, also gives a specific measure for each statistic, but it does not take into the multiple inferences.

2.3 Dependence and Asymptotic Properties

In this section, we will consider the pFDR when the test statistics are dependent, along with some asymptotic properties that have direct applications to certain cases of dependence. Here we assume we are testing m identical hypotheses based on statistics $X_1, \ldots, X_m$. We first present the following simple result.

Theorem 2.2 *Under any distributional assumptions about $X_1, \ldots, X_m$ and $H_1, \ldots, H_m$, it follows*

$$pFDR(\Gamma) = \sum_{k=1}^{m} \sum_{i_1,\cdots,i_k} \frac{1}{k} \sum_{j=1}^{k} \mathbf{Pr}\left(H_{i_j} = 0\ ,\quad \begin{array}{c} X_{i_1}, \ldots, X_{i_k} \in \Gamma \\ X_{i_{k+1}}, \ldots, X_{i_m} \notin \Gamma \end{array} \middle| R(\Gamma) > 0\right),$$

where the middle sum is taken over all distinct subsets of size k of $\{1, 2, \ldots, m\}$.

Proof:

$$pFDR(\Gamma) = \sum_{k=1}^{m} \sum_{i_1,\cdots,i_k} \mathbf{E}\left(\frac{V(\Gamma)}{R(\Gamma)} \,\middle|\, \begin{array}{c} X_{i_1},\ldots,X_{i_k} \in \Gamma \\ X_{i_{k+1}},\ldots,X_{i_m} \notin \Gamma \end{array} \right) \cdot \mathbf{Pr}\left(\begin{array}{c} X_{i_1},\ldots,X_{i_k} \in \Gamma \\ X_{i_{k+1}},\ldots,X_{i_m} \notin \Gamma \end{array} \,\middle|\, R(\Gamma) > 0 \right)$$

$$= \sum_{k=1}^{m} \sum_{i_1,\cdots,i_k} \mathbf{E}\left(\frac{\sum_{j=1}^{k}(1-H_{i_j})}{k} \,\middle|\, \begin{array}{c} X_{i_1},\ldots,X_{i_k} \in \Gamma \\ X_{i_{k+1}},\ldots,X_{i_m} \notin \Gamma \end{array} \right) \cdot \mathbf{Pr}\left(\begin{array}{c} X_{i_1},\ldots,X_{i_k} \in \Gamma \\ X_{i_{k+1}},\ldots,X_{i_m} \notin \Gamma \end{array} \,\middle|\, R(\Gamma) > 0 \right)$$

$$= \sum_{k=1}^{m} \sum_{i_1,\cdots,i_k} \frac{1}{k} \sum_{j=1}^{k} \mathbf{Pr}\left(H_{i_j} = 0 \,\middle|\, \begin{array}{c} X_{i_1},\ldots,X_{i_k} \in \Gamma \\ X_{i_{k+1}},\ldots,X_{i_m} \notin \Gamma \end{array} \right) \cdot \mathbf{Pr}\left(\begin{array}{c} X_{i_1},\ldots,X_{i_k} \in \Gamma \\ X_{i_{k+1}},\ldots,X_{i_m} \notin \Gamma \end{array} \,\middle|\, R(\Gamma) > 0 \right).$$

□

The representation of $pFDR(\Gamma)$ in Theorem 2.2 appears intractable at first glance, but under a fully parametric model it is feasible to calculate this quantity or a numerical approximation to it. When the statistics have the same marginal distribution, but are dependent in some arbitrary way, we may simplify this result.

Corollary 2.3 *Suppose that $X_1,\ldots,X_m$ are identically distributed and $H_1,\ldots,H_m$ are identically distributed. Then*

$$pFDR(\Gamma_\alpha) = \sum_{k=1}^{m} \mathbf{Pr}\left(H_1 = 0 \,\middle|\, \begin{array}{c} X_1,\ldots,X_k \in \Gamma_\alpha \\ X_{k+1},\ldots,X_m \notin \Gamma_\alpha \end{array} \right) \cdot \mathbf{Pr}(R = k | R > 0).$$

From these results it can be seen that Theorem 2.1 does not hold under general dependence. Therefore, we now determine when Theorem 2.1 holds approximately, or rather asymptotically. Recall that we can represent the nested set of rejection regions by $\{\Gamma_\alpha\}_{\alpha>0}$, where α is the Type I error of Γ_α. In our notation

$$\begin{aligned} \frac{V_m(\Gamma_\alpha)}{\sum_{i=1}^{m}(1-H_i)} &= \frac{\sum_{i=1}^{m}(1-H_i)\cdot 1(X_i \in \Gamma_\alpha)}{\sum_{i=1}^{m}(1-H_i)}, \\ \frac{S_m(\Gamma_\alpha)}{\sum_{i=1}^{m} H_i} &= \frac{\sum_{i=1}^{m} H_i \cdot 1(X_i \in \Gamma_\alpha)}{\sum_{i=1}^{m} H_i} \end{aligned}$$

represent the empirical distribution functions of the null and alternative p-values, respectively, as a function of α. If these quantities converge in the pointwise sense, then the realized proportion of false discoveries, the FDR, and the pFDR converge to a posterior probability, simultaneously for all Γ_α. This is explicitly stated in the following theorem.

Theorem 2.3 *Suppose that with probability 1 we have* $\sum_{i=1}^{m}(1-H_i)/m \to \pi_0$, *and*

$$\frac{V_m(\Gamma_\alpha)}{\sum_{i=1}^{m}(1-H_i)} \to G_0(\alpha),$$
$$\frac{S_m(\Gamma_\alpha)}{\sum_{i=1}^{m} H_i} \to G_1(\alpha),$$

for each $\alpha > 0$ *for some functions* G_0 *and* G_1, *as* $m \to \infty$. *Then for any* $\delta > 0$ *where* $\pi_0 \cdot G_0(\delta) + (1-\pi_0) \cdot G_1(\delta) > 0$,

$$(1) \lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \frac{V_m(\Gamma_\alpha)}{R_m(\Gamma_\alpha) \vee 1} - \mathbf{Pr}_\infty(H=0|X \in \Gamma_\alpha) \right| \stackrel{a.s.}{=} 0,$$
$$(2) \lim_{m\to\infty} \sup_{\alpha\geq\delta} |FDR_m(\Gamma_\alpha) - \mathbf{Pr}_\infty(H=0|X \in \Gamma_\alpha)| = 0,$$
$$(3) \lim_{m\to\infty} \sup_{\alpha\geq\delta} |pFDR_m(\Gamma_\alpha) - \mathbf{Pr}_\infty(H=0|X \in \Gamma_\alpha)| = 0,$$

where we define

$$\mathbf{Pr}_\infty(H=0|X \in \Gamma_\alpha) = \frac{\pi_0 \cdot G_0(\alpha)}{\pi_0 \cdot G_0(\alpha) + (1-\pi_0) \cdot G_1(\alpha)}.$$

The functions G_0 and G_1 are the asymptotic Type I error and power of the p-values as a function of α. In general, Theorem 2.3 says that if the statistics are "weakly dependent" then the realized proportion of false discoveries, the FDR, and the pFDR converge simultaneously over all rejection regions to the Bayesian posterior probability defined above. If one is able to calculate or estimate G_0, G_1 and π_0, then for large m these provide good approximations for the realized proportion of false discoveries, the FDR, and the pFDR for all rejection regions simultaneously.

Proof: Let

$$Q_m(\alpha) = \frac{V_m(\Gamma_\alpha)}{[V_m(\Gamma_\alpha) + S(\Gamma_\alpha)] \vee 1}.$$

By an easy modification of the Glivenko-Cantelli Theorem (see for example Billingsley 1968), it follows:

$$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \frac{V_m(\Gamma)}{m} - \pi_0 \cdot G_0(\alpha) \right| \stackrel{a.s.}{=} 0,$$

$$\lim_{m\to\infty}\sup_{\alpha\geq\delta}\left|\frac{[V_m(\Gamma_\alpha)+S_m(\Gamma_\alpha)]\vee 1}{m}-[\pi_0\cdot G_0(\alpha)+\pi_1\cdot G_1(\alpha)]\right|\stackrel{a.s.}{=}0.$$

Since $\pi_0\cdot G_0(\delta)+(1-\pi_0)\cdot G_1(\delta)>0$ and these are both non-decreasing functions, it is easy to show that

$$\lim_{m\to\infty}\sup_{\alpha\geq\delta}\left|\frac{m}{[V_m(\Gamma_\alpha)+S_m(\Gamma_\alpha)]\vee 1}-\frac{1}{\pi_0\cdot G_0(\alpha)+\pi_1\cdot G_1(\alpha)}\right|\stackrel{a.s.}{=}0.$$

Finally noticing that

$$\left|\frac{V_m(\Gamma_\alpha)-m\pi_0\cdot G_0(\alpha)}{[V_m(\Gamma_\alpha)+S_m(\Gamma_\alpha)]\vee 1}\right|+\left|\frac{m\pi_0\cdot G_0(\alpha)}{[V_m(\Gamma_\alpha)+S_m(\Gamma_\alpha)]\vee 1}-\frac{\pi_0\cdot G_0(\alpha)}{\pi_0\cdot G_0(\alpha)+\pi_1\cdot G_1(\alpha)}\right|$$

$$\geq|Q_m(\alpha)-\mathbf{Pr}_\infty(H=0|X\in\Gamma_\alpha)|$$

completes the proof of the first convergence.

Now $|Q_m(\alpha)-\mathbf{Pr}_\infty(H=0|X\in\Gamma_\alpha)|\leq 1$ almost surely, so that it easily follows

$$\begin{aligned}0&=\mathbf{E}\left[\lim_{m\to\infty}\sup_{\alpha\geq\delta}|Q_m(\alpha)-\mathbf{Pr}_\infty(H=0|X\in\Gamma_\alpha)|\right]\\&=\lim_{m\to\infty}\mathbf{E}\left[\sup_{\alpha\geq\delta}|Q_m(\alpha)-\mathbf{Pr}_\infty(H=0|X\in\Gamma_\alpha)|\right]\\&\geq\sup_{\alpha\geq\delta}|\mathbf{E}[Q_m(\alpha)]-\mathbf{Pr}_\infty(H=0|X\in\Gamma_\alpha)|\geq 0,\end{aligned}$$

where $\mathbf{E}[Q_m(\alpha)]=FDR_m(\Gamma_\alpha)$. Finally,

$$\begin{aligned}\lim_{m\to\infty}\sup_{\alpha\geq\delta}|pFDR_m(\Gamma_\alpha)-FDR_m(\Gamma_\alpha)|&\leq\\\lim_{m\to\infty}\left|\frac{1}{\mathbf{Pr}(R_m(\delta)>0)}-1\right|&=0.\end{aligned}$$

□

Two useful case where the theorem holds are when the statistics $X_1, X_2, \ldots$ are such that there exists a k where $|i-j|\geq k$ implies X_i and X_j are independent (i.e., the dependence is in finite clumps), or when the statistics are a stationary ergodic sequence. There are other forms of dependence where this result holds, for example, for certain Markov chains and certain mixing distributions. The following is a numerical example involving locally dependent statistics.

Example: Locally Dependent Statistics. As a numerical example to illustrate the result of Theorem 2.3, consider the following situation. Suppose $X_i|H_i = 0 \sim N(0,1)$ and $X_i|H_i = 1 \sim N(2,1)$. We have $\mathbf{Cov}(X_i, X_{i+k}) = \rho$ where $0 \leq \rho \leq 1$ for $k = 1, 2, \ldots, 9$ and $i = 1, 11, 21, \ldots$, and zero covariance otherwise. In other words the statistics have correlation ρ in groups of 10. Suppose we take $\Gamma_\alpha = [\Phi^{-1}(1-\alpha), \infty)$ where Φ is the cdf of a $N(0,1)$, and $H_i \overset{i.i.d.}{\sim} Bernoulli(1-\pi_0)$. Then by Theorem 2.3, we have for example that

$$\lim_{m\to\infty} \sup_{\alpha>0} |pFDR_m(\Gamma_\alpha) - \mathbf{Pr}_\infty(H = 0|X \in \Gamma_\alpha)| = 0,$$

where $\mathbf{Pr}_\infty(H = 0|X \in \Gamma_\alpha) = \pi_0 \cdot \alpha / [\pi_0 \cdot \alpha + (1-\pi_0) \cdot \mathbf{Pr}(N(2,1) \geq \Phi^{-1}(1-\alpha))]$. Table 2.1 shows $\mathbf{Pr}_\infty(H = 0|X \in \Gamma_\alpha)$ compared to the $pFDR_m(\Gamma_\alpha)$ at several m for $\alpha = 0.005$ and $\alpha = 0.001$. It can be seen that there is quite good agreement between the limiting case and the finite cases, especially for large m. Most of the differences at $m = 5000$ are within the Monte Carlo standard error, which is listed parenthetically.

Table 2.1: Simulation results: $pFDR_m(\Gamma_\alpha)$ converging to $\mathbf{Pr}_\infty(H = 0|X \in \Gamma_\alpha)$

$\alpha = 0.005$, $\mathbf{Pr}_\infty(H = 0\|X \in \Gamma_\alpha) = 0.137$						
m	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
100	0.142 (0.004)	0.126 (0.004)	0.120 (0.004)	0.102 (0.004)	0.094 (0.004)	0.041 (0.003)
500	0.136 (0.003)	0.136 (0.003)	0.133 (0.003)	0.127 (0.003)	0.117 (0.003)	0.091 (0.003)
1000	0.138 (0.003)	0.136 (0.003)	0.134 (0.003)	0.132 (0.003)	0.128 (0.003)	0.113 (0.003)
3000	0.138 (0.003)	0.137 (0.003)	0.137 (0.003)	0.137 (0.003)	0.134 (0.003)	0.129 (0.003)
5000	0.138 (0.003)	0.138 (0.003)	0.137 (0.003)	0.137 (0.003)	0.135 (0.003)	0.132 (0.003)

$\alpha = 0.001$, $\mathbf{Pr}_\infty(H = 0\|X \in \Gamma_\alpha) = 0.061$						
m	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
100	0.061 (0.003)	0.063 (0.004)	0.053 (0.003)	0.047 (0.003)	0.036 (0.003)	0.010 (0.002)
500	0.061 (0.002)	0.063 (0.002)	0.060 (0.002)	0.052 (0.002)	0.047 (0.002)	0.028 (0.002)
1000	0.063 (0.002)	0.062 (0.002)	0.060 (0.002)	0.060 (0.002)	0.055 (0.002)	0.038 (0.002)
3000	0.061 (0.001)	0.063 (0.001)	0.061 (0.001)	0.061 (0.001)	0.058 (0.001)	0.051 (0.002)
5000	0.061 (0.001)	0.063 (0.001)	0.062 (0.001)	0.062 (0.001)	0.060 (0.001)	0.054 (0.002)

2.4 A Connection to Classification Theory

When assuming the statistics follow a mixture distribution, as we have assumed throughout this chapter, it is possible to view multiple hypothesis testing as a classification problem. For each test, we observe X_i and we have to decide whether to classify H_i as 0 or H_i as 1 based on X_i. There are four possible outcomes for each test with two of them being misclassifications. Consider Table 2.2 listing these outcomes, with the penalties for each type of misclassification parameterized by λ.

Table 2.2: Outcomes of "Classifying" H_i with Misclassification Penalties

	Classify H_i as 0	Classify H_i as 1
$H_i = 0$	0	$1 - \lambda$
$H_i = 1$	λ	0

We use several of the basic facts about classification theory found in Cherkassky & Mulier (1998), for example. A rejection region Γ can be thought of as a classification rule in the following way: if $X_i \in \Gamma$ then we classify H_i as 1, and if $X_i \notin \Gamma$, then we classify H_i as 0. The "Bayes Error" of a classification rule (in terms of the rejection region representation) is

$$\mathrm{BE}(\Gamma) = (1 - \lambda)\mathbf{Pr}(X_i \in \Gamma, H_i = 0) + \lambda\mathbf{Pr}(X_i \notin \Gamma, H_i = 1). \tag{2.2}$$

That is, $\mathrm{BE}(\Gamma)$ is the expected loss under Table 2.2.

Genovese & Wasserman (2001) notice that one can define a dual quantity to the FDR, which they call the False Non-discovery Rate (FNR). It is defined to be the expected proportion of false negatives among all hypotheses that *are not* rejected, with the ratio being set to zero if all hypotheses are rejected:

$$FNR = \mathbf{E}\left[\frac{T}{W}\middle| W > 0\right]\mathbf{Pr}(W > 0), \tag{2.3}$$

where W is the total number of non-significant hypotheses, and T is the number of non-significant alternative statistics (false negatives). We make the following modified definition of the FNR, in the spirit of the pFDR.

Definition 2.2 *We define the* **positive False Non-discovery Rate** – the rate that non-discoveries are false – *to be:*

$$pFNR = \mathbf{E}\left[\frac{T}{W}\middle| W > 0\right].$$

Using an analogous argument to Theorem 2.1, we can show the following result.

Theorem 2.4 *Under the assumptions of Theorem 2.1, it follows*

$$pFNR(\Gamma) = \mathbf{Pr}(H = 1 | T \notin \Gamma),$$

where $\pi_1 = 1 - \pi_0$ *is the implicit prior probability in the above posterior probability.*

Now the Bayes Error can be written as a weighted sum of $pFDR(\Gamma)$ and $pFNR(\Gamma)$.

Corollary 2.4 *Under the assumptions of Theorem 2.1,*

$$\text{BE}(\Gamma) = (1-\lambda)\mathbf{Pr}(X \in \Gamma) \cdot pFDR(\Gamma) + \lambda \mathbf{Pr}(X \notin \Gamma) \cdot pFNR(\Gamma). \tag{2.4}$$

In the Benjamini & Hochberg (1995) framework, one decides beforehand at what level to control the FDR and then applies their procedure to control it at that level. We described in Section 1.4 how this can be a difficult choice to make, since one has to make the choice of α before any data are seen. Using the classification theory connection, we suggest two other ways to use the pFDR. One can choose the rejection region based on the relative cost of a false positive to a false negative and then minimize the Bayes Error; or one can decide the relative importance of the pFDR to the pFNR, and then minimize their weighted average.

In Section 1.3 and Chapter 8, we consider a problem in which one is concerned with deciding which of several thousand genes show a statistically significant change in gene expression between two types of cells (e.g., normal versus diseased cells). Here, it is feasible that the scientist can decide on the relative cost of a false positive gene to a false negative gene. In that case, one can derive the Bayes rule to minimize the Bayes Error. By Corollary 2.4, one can interpret the Bayes Error in terms of the multiple hypothesis testing quantities pFDR and pFNR. In fact, the manner in which the Bayes Error weights the pFDR and pFNR makes a lot of sense. Another equally useful quantity to minimize is the weighted average of the pFDR and pFNR:

$$(1-w) \cdot pFDR(\Gamma) + w \cdot pFNR(\Gamma).$$

In words, one can decide how important the rate of false discoveries is to the rate of false non-discoveries. We now show how to minimize this weighted average.

Recall that we assume $X_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot F_0 + H_i \cdot F_1$ for $i = 1, \ldots, m$, and $H_i \overset{i.i.d.}{\sim} Bernoulli(1-\pi_0)$. Also assume that F_0 and F_1 are continuous distributions with common support, with respective densities f_0 and f_1. Define the set of rejection regions $\{\mathcal{B}_\lambda\}$ for $0 \leq \lambda \leq 1$ by

$$\mathcal{B}_\lambda = \left\{ x : \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \leq \lambda \right\}.$$

The sets $\{\mathcal{B}_\lambda\}$ define the Bayes rule for the cost matrix given by Table 2.2. That is, for each λ, $\mathcal{B}_\lambda$ minimizes $\mathrm{BE}(\mathcal{B}_\lambda)$. Note that by Corollary 2.4, $\mathcal{B}_\lambda$ also minimizes (2.4) for each λ.

As it turns out, the nested set of rejection regions $\{\mathcal{B}_\lambda\}$ can also be used to minimize $(1-w)\cdot pFDR(\Gamma) + w \cdot pFNR(\Gamma)$. We state this formally in the following theorem.

Theorem 2.5 *Let* $\lambda(w) = \arg\min_\lambda [(1-w)\cdot pFDR(\mathcal{B}_\lambda) + w \cdot pFNR(\mathcal{B}_\lambda)]$. *Then* $(1-w)\cdot pFDR(\mathcal{B}_{\lambda(w)}) + w \cdot pFNR(\mathcal{B}_{\lambda(w)})$ *minimizes* $(1-w)\cdot pFDR(\Gamma) + w \cdot pFNR(\Gamma)$ *among all measurable* Γ.

Proof: Recall that by the Neyman-Pearson lemma, the $\{\mathcal{B}_\lambda\}$ form a set of uniformly most powerful rejection regions. Without loss of generality, we can assume that for each $\alpha \in [0,1]$, there exists a $\mathcal{B}_\lambda$ such that $\mathbf{Pr}(X \in \mathcal{B}_\lambda | H = 0) = \alpha$. Otherwise, $\{\mathcal{B}_\lambda\}$ can be extended in the natural way to accomplish this and still remain uniformly most powerful (Lehmann 1986).

Consider any measurable Γ. Then there exists a $\mathcal{B}_\lambda$ such that $\mathbf{Pr}(X \in \Gamma | H = 0) = \mathbf{Pr}(X \in \mathcal{B}_\lambda | H = 0)$. Since the $\{\mathcal{B}_\lambda\}$ are uniformly most powerful, it follows that $\mathbf{Pr}(X \in \Gamma | H = 1) \leq \mathbf{Pr}(X \in \mathcal{B}_\lambda | H = 1)$. Therefore,

$$\begin{aligned}
pFDR(\Gamma) &= \tfrac{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H=0)}{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H=0) + \pi_1 \cdot \mathbf{Pr}(X \in \Gamma | H=1)} \\
&\geq \tfrac{\pi_0 \cdot \mathbf{Pr}(X \in \mathcal{B}_\lambda | H=0)}{\pi_0 \cdot \mathbf{Pr}(X \in \mathcal{B}_\lambda | H=0) + \pi_1 \cdot \mathbf{Pr}(X \in \mathcal{B}_\lambda | H=1)} &= pFDR(\mathcal{B}_\lambda), \\
pFNR(\Gamma) &= \tfrac{\pi_1 \cdot \mathbf{Pr}(X \notin \Gamma | H=1)}{\pi_1 \cdot \mathbf{Pr}(X \notin \Gamma | H=1) + \pi_0 \cdot \mathbf{Pr}(X \notin \Gamma | H=0)} \\
&\geq \tfrac{\pi_1 \cdot \mathbf{Pr}(X \notin \mathcal{B}_\lambda | H=1)}{\pi_1 \cdot \mathbf{Pr}(X \notin \mathcal{B}_\lambda | H=1) + \pi_0 \cdot \mathbf{Pr}(X \notin \mathcal{B}_\lambda | H=0)} &= pFNR(\mathcal{B}_\lambda).
\end{aligned}$$

Hence for any w, $(1-w)\cdot pFDR(\mathcal{B}_\lambda) + w \cdot pFNR(\mathcal{B}_\lambda) \leq (1-w)\cdot pFDR(\Gamma) + w \cdot pFNR(\Gamma)$, and the overall minimizing $\mathcal{B}_{\lambda(w)}$ can be found among the $\{\mathcal{B}_\lambda\}$ as stated in the theorem. □

Example: Normal Distributions. Suppose $X_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot N(0,1) + H_i \cdot N(2,1)$, and $H_i \overset{i.i.d}{\sim} Bernoulli(0.2)$. Also suppose we want to minimize

$$\frac{1}{3}\cdot pFDR(\Gamma) + \frac{2}{3}\cdot pFNR(\Gamma)$$

over all measurable Γ. Therefore, we have made the rate that non-discoveries are false 2 times more important than the rate that discoveries are false. By Theorem 2.5, we only have to consider rejection regions of the form:

$$\mathcal{B}_\lambda = \left\{ x : \frac{0.8\cdot\phi_{0,1}(x)}{0.8\cdot\phi_{0,1}(x) + 0.2\cdot\phi_{2,1}(x)} \leq \lambda \right\},$$

where ϕ_{μ,σ^2} is the density of a $N(\mu,\sigma^2)$. By calculating $\lambda(2/3) = \arg\min_\lambda[1/3\cdot pFDR(\mathcal{B}_\lambda) + 2/3\cdot pFNR(\mathcal{B}_\lambda)]$, we get $\lambda(2/3) = 0.193$, which implies $\mathcal{B}_{0.193} = \{X \geq 2.41\}$. Therefore $\inf_\Gamma 1/3\cdot pFDR(\Gamma) + 2/3\cdot pFNR(\Gamma) = 0.123$ and this occurs at $\Gamma = \mathcal{B}_{0.193} = \{X \geq 2.41\}$. Figure 2.1 shows $1/3\cdot pFDR(\mathcal{B}_\lambda) + 2/3\cdot pFNR(\mathcal{B}_\lambda)$ as a function of λ.

Since it will tend to be the case that $\pi_0 \gg \pi_1$, one may also wish to find Γ to minimize

$$(1-w)\cdot\frac{pFDR(\Gamma)}{\pi_0} + w\cdot\frac{pFNR(\Gamma)}{\pi_1}.$$

The minimizing set can also be found among the $\{\mathcal{B}_\lambda\}$ using some $\lambda'(w)$ defined similarly to above.

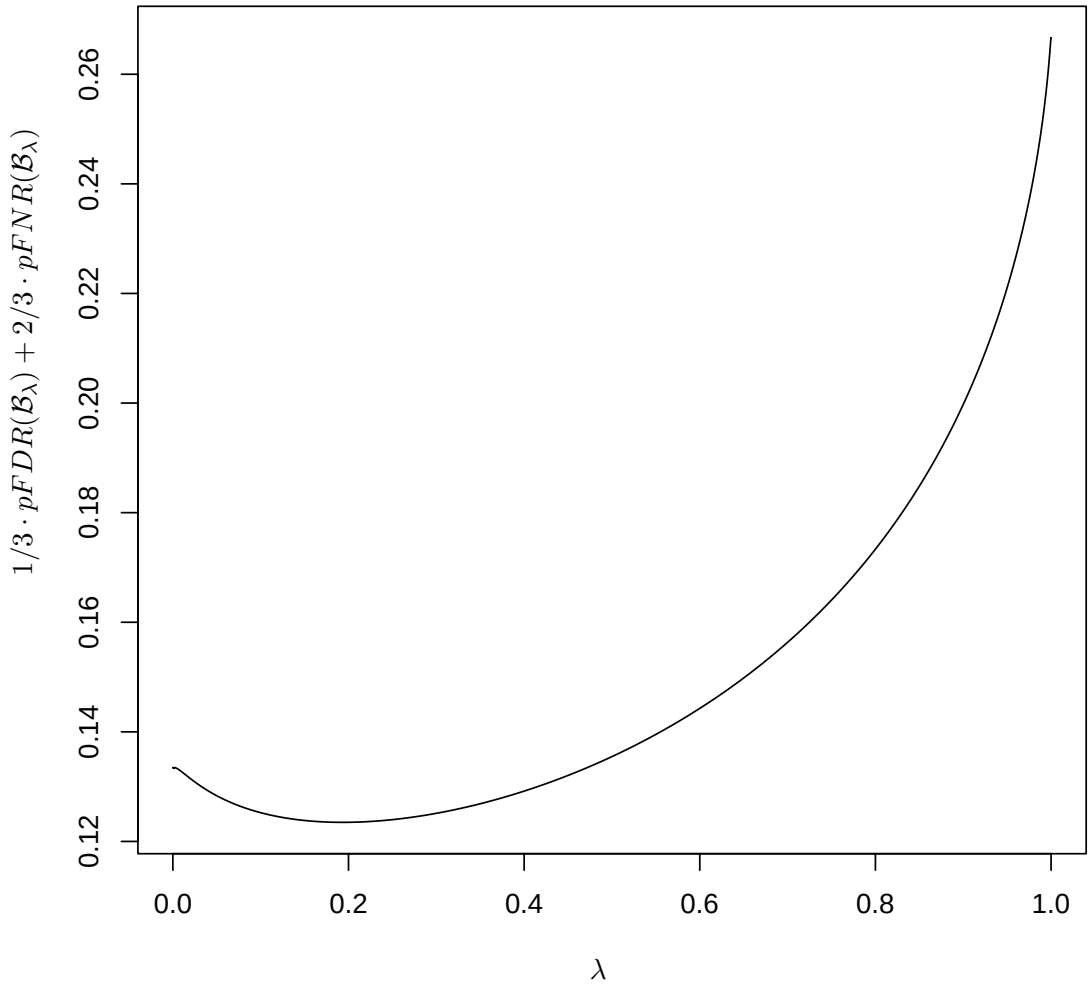


Figure 2.1: A plot of $1/3 \cdot pFDR(\mathcal{B}_\lambda) + 2/3 \cdot pFNR(\mathcal{B}_\lambda)$ as a function of λ.

Chapter 3

The q-value and Simultaneous Controlling Curves

As discussed in Chapter 1, it is potentially most useful to consider the FDR or pFDR over the naturally occurring rejection regions – in terms of p-values, this is $[0, p_i]$ for $i = 1, \ldots, m$. Moreover, one might want to calculate the minimum error rate that can be attained when rejecting a particular statistic. We first consider this quantity in terms of the pFDR, then we generalize it to general multiple hypothesis testing error rates, paying particular attention to the FDR.

3.1 The q-value

We now introduce the pFDR analogue of the p-value, which we call the *q-value*. Because of the connection made in Section 2.2, the q-value is useful in both Bayesian and frequentist settings. It gives the scientist a hypothesis testing error measure for each observed statistic with respect to the pFDR. Again, assume that $X_i|H_i \overset{i.i.d.}{\sim} (1 - H_i) \cdot F_0 + H_i \cdot F_1$ and $H_i \overset{i.i.d.}{\sim} Bernoulli(1 - \pi_0)$ for $i = 1, \ldots, m$. We introduce the q-value by first showing an example.

Example: Testing the Mean of a $N(\theta, 1)$ Random Variable. Suppose we perform m hypothesis tests of $\theta = 0$ versus $\theta = 2$ for m independent $N(\theta, 1)$ random variables $X_1, \ldots, X_m$, where $X_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot N(0,1)+H_i\cdot N(2,1)$. Given we observe the random

variables to be $X_1 = x_1, \ldots, X_m = x_m$, the p-value of $X_i = x_i$ can be calculated as

$$\text{p-value}(x_i) = \mathbf{Pr}(X \geq x_i | H = 0) = \mathbf{Pr}(N(0,1) \geq x_i).$$

In words, p-value(x_i) gives the Type I error rate if we reject any statistic as extreme or more extreme than x_i.

What is the pFDR if we reject any statistic as extreme or more extreme than x_i among all m hypotheses? By Theorem 2.1,

$$\begin{aligned} pFDR(\{T \geq x_i\}) &= \frac{\pi_0 \mathbf{Pr}(X \geq x_i | H = 0)}{\pi_0 \mathbf{Pr}(X \geq x_i | H = 0) + \pi_1 \mathbf{Pr}(X \geq x_i | H = 1)} \\ &= \frac{\pi_0 \mathbf{Pr}(N(0,1) \geq x_i)}{\pi_0 \mathbf{Pr}(N(0,1) \geq x_i) + \pi_1 \mathbf{Pr}(N(2,1) \geq x_i)}. \end{aligned} \tag{3.1}$$

It can be seen that $pFDR(\{T \geq x_i\})$ is a natural pFDR analogue to p-value(x_i). The relationship between these two quantities can also be understood graphically. Figure 3.1 shows a graph of the $N(0,1)$ and $N(2,1)$ distributions with the point $X_i = x_i$ marked by a vertical line. The area under the $N(0,1)$ density to the right of x_i is p-value(x_i). In order to calculate $pFDR(\{T \geq x_i\})$, we use formula (3.1), which involves the areas to the right of x_i under the $N(0,1)$ and the $N(2,1)$ densities, and their respective weights π_0 and π_1.

As we show in this section, (3.1) is what we call q-value(x_i). In many situations, it is the pFDR obtained when rejecting a statistic as extreme or more extreme than x_i among all m hypotheses; but the q-value can be defined more generally, as can the p-value.

Until now, we have only considered a single rejection region. Hypothesis tests are usually derived according to a nested set of rejection regions. As long as F_0 and F_1 have a common support, we can denote this nested set of rejection regions without loss of generality by $\{\Gamma_\alpha\}_{\alpha=0}^1$, where α is such that $\mathbf{Pr}(X \in \Gamma_\alpha | H = 0) = \alpha$. Note that $\alpha' \leq \alpha$ implies that $\Gamma_{\alpha'} \subseteq \Gamma_\alpha$, giving the nested property. Using this notation, the p-value(x) of an observed statistic $X = x$ is defined to be (Lehmann 1986)

$$\text{p-value}(x) = \inf_{\{\Gamma_\alpha:\ x \in \Gamma_\alpha\}} \mathbf{Pr}(X \in \Gamma_\alpha | H = 0).$$

This quantity gives a measure of the strength of the observed statistic with respect to making a Type I error – it is the minimum Type I error rate that can occur when rejecting

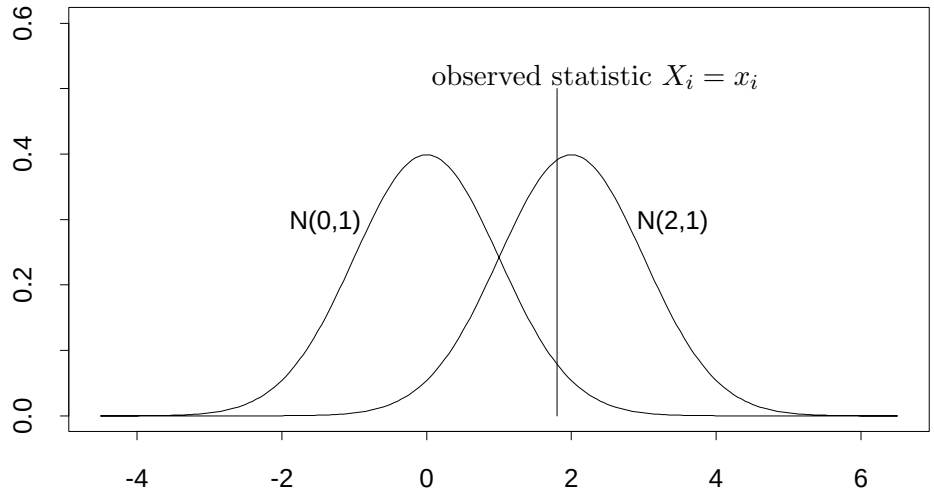


Figure 3.1: A plot of the N(0,1) and N(2,1) densities. The vertical line denotes the observed statistic $X_i = x_i$. The p-value can be calculated from the area under the N(0,1) density to the right of $X_i = x_i$. The q-value is calculated using the area under both densities to the right of $X_i = x_i$ and their prior probabilities π_0 and π_1.

a statistic with value t, given the set of nested rejection regions.

In light of Theorem 2.1, we define an analogous quantity in terms of the pFDR that has both frequentist multiple hypothesis testing and Bayesian interpretations.

Definition 3.1 *For an observed statistic $X = x$ define the **q-value** of x to be:*

$$\begin{aligned} \textit{q-value}(x) &= \inf_{\{\Gamma_\alpha:\ x \in \Gamma_\alpha\}} \mathbf{Pr}(H = 0 | X \in \Gamma_\alpha) && (3.2) \\ &= \inf_{\{\Gamma_\alpha:\ x \in \Gamma_\alpha\}} pFDR(\Gamma_\alpha) && (3.3) \end{aligned}$$

Therefore, according to (3.2), the q-value is a Bayesian version of the p-value – say a "posterior Bayesian p-value" – the minimum posterior probability $H = 0$ over all rejection regions containing the statistic. We call (3.2) the q-value because it is equivalent to the p-value with the events $\{X \in \Gamma_\alpha\}$ and $\{H = 0\}$ reversed. In words, (3.3) says the q-value is a measure of the strength of an observed statistic with respect to the pFDR – it is the minimum pFDR that can occur when rejecting a statistic with value x for the set of nested rejection regions.

Remark: Is the q-value a "pFDR adjusted p-value"? *The answer is no*, and this is immediately clear when recalling the definition of an adjusted p-value. Shaffer (1995) says, "Given any test procedure, the adjusted p-value corresponding to a test of a single hypothesis H_i can be defined as the level of the entire test procedure at which H_i would be rejected, given the values of all test statistics involved." Therefore, since the pFDR cannot be controlled by a test procedure (i.e., a sequential p-value method), then it cannot be used to define adjusted p-values. But more importantly, notice that the adjusted p-values are defined *in terms of a particular procedure*. The q-value in no way depends on a sequential p-value procedure. It is a deterministic function evaluated at a random quantity.

3.2 A Connection Between q-values and p-values

Notice that

$$\begin{aligned}\underset{\{\Gamma_\alpha:\, x\in\Gamma_\alpha\}}{\arg\min}\ \mathbf{Pr}(H=0|X\in\Gamma_\alpha) &= \underset{\{\Gamma_\alpha:\, x\in\Gamma_\alpha\}}{\arg\min}\ \frac{\pi_0\mathbf{Pr}(X\in\Gamma_\alpha|H=0)}{\pi_0\mathbf{Pr}(X\in\Gamma_\alpha|H=0)+\pi_1\mathbf{Pr}(X\in\Gamma_\alpha|H=1)} \\ &= \underset{\{\Gamma_\alpha:\, x\in\Gamma_\alpha\}}{\arg\min}\ \frac{\mathbf{Pr}(X\in\Gamma_\alpha|H=0)}{\mathbf{Pr}(X\in\Gamma_\alpha|H=1)}.\end{aligned}$$

Therefore, the rejection region that determines the q-value minimizes the ratio of the Type I error to the power over all rejection regions that contain the statistic. This makes sense because the pFDR is concerned with measuring how frequent the false positives occur in relation to true positives.

One can understand this observation in terms of a plot of power versus Type I error for a given set of rejection regions. Recall that $X_i \overset{i.i.d.}{\sim} \pi_0 \cdot F_0 + \pi_1 \cdot F_1$ for $i = 1, \ldots, m$. We will write

$$\begin{aligned}G_1(\alpha) = \int_{\Gamma_\alpha} dF_1 &= \mathbf{Pr}(X\in\Gamma_\alpha|H=1), \\ G_0(\alpha) = \int_{\Gamma_\alpha} dF_0 &= \mathbf{Pr}(X\in\Gamma_\alpha|H=0) = \alpha.\end{aligned} \tag{3.4}$$

It is easily shown that G_0 and G_1 are the cdf's of the null and alternative p-values, respectively. Suppose that G_1 is continuous and differentiable. Then it can be shown through simple calculus that $\alpha/G_1(\alpha)$ is minimized at $\alpha = G_1(\alpha)/G_1'(\alpha)$. Therefore we can minimize $\alpha/G_1(\alpha)$ graphically by drawing all lines from the origin that are tangent to a concave

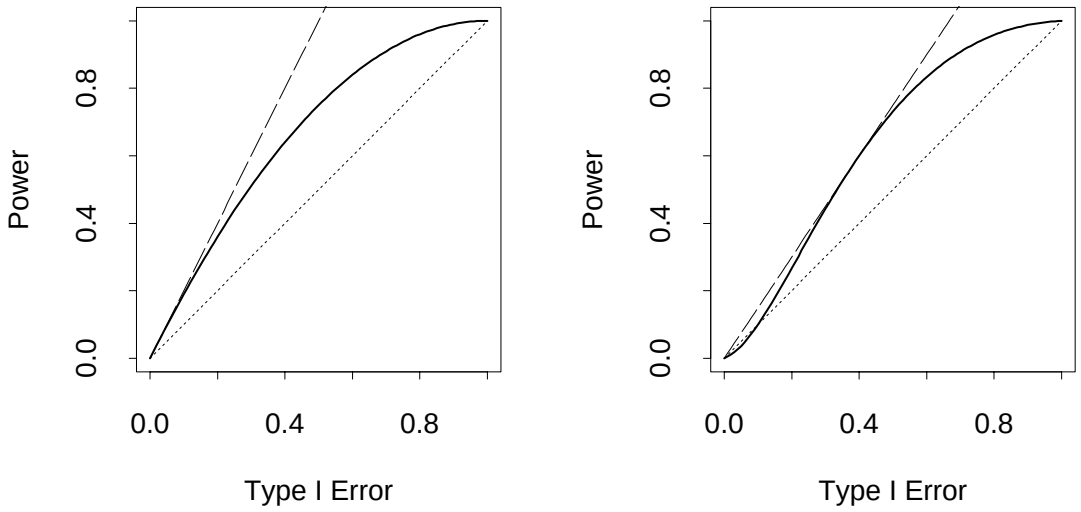


Figure 3.2: A plot of power versus Type I error rate for two hypothetical sets of rejection regions. The solid line is power as a function of Type I error, $G_1(\alpha)$; the dotted line is the identity function; the dashed line is the line from the origin tangent to $G_1(\alpha)$.

portion of the function. The line with the largest slope is tangent to the point on the curve where $\alpha/G_1(\alpha)$ is minimized.

See Figure 3.2 for a picture of this maximization. The left panel has a strictly concave $G_1(\alpha)$. In this case, the ratio of power to Type I error increases as $\alpha \to 0$. In other words, as the rejection regions get smaller, the ratio of power to Type I error gets larger. Therefore, for a concave G_1, we can conclude that the Γ_α that contains x and minimizes $pFDR(\Gamma_\alpha)$ also minimizes $\mathbf{Pr}(X \in \Gamma_\alpha | H = 0)$. This follows since we would take the rejection region with the smallest α, where $x \in \Gamma_\alpha$, in order to minimize $\alpha/G_1(\alpha)$.

Therefore, when the cdf of the alternative p-values G_1 is concave, the same rejection region is used to define the q-value and the p-value. More generally, we only need $G_1(\alpha)/\alpha$ to be a decreasing function of α if we do not assume that G_1 is differentiable. (Note that if G_1 is concave then this holds). We state this formally in the following proposition.

Proposition 3.1 *The q-value of a statistic is based on the same rejection region as the*

p-value, as long as $G_1(\alpha)/\alpha$ is decreasing in α. That is,

$$\underset{\{\Gamma_\alpha:\ x\in\Gamma_\alpha\}}{\arg\min}\ \mathbf{Pr}(H=0|X\in\Gamma_\alpha) = \underset{\{\Gamma_\alpha:\ x\in\Gamma_\alpha\}}{\arg\min}\ \mathbf{Pr}(X\in\Gamma_\alpha|H=0)$$

when $G_1(\alpha)/\alpha$ is decreasing in α.

The right panel of Figure 3.2 shows an example where G_1 is not concave, nor is $G_1(\alpha)/\alpha$ decreasing in α. The rejection region that minimizes the ratio of the Type I error to the power is the one that corresponds to the point that the shown line from the origin intersects. No similar connection can be made with the p-value under this kind of G_1.

Example: Likelihood Ratio Based Rejection Regions. We have assumed that $X_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot F_0 + H_i \cdot F_1$. Suppose that f_0 and f_1 are the densities corresponding to F_0 and F_1, respectively. Also, suppose that we consider rejection regions of the form

$$\left\{x : \frac{f_0(x)}{f_1(x)} \leq \lambda\right\}. \tag{3.5}$$

Then it follows that the power to Type I error curve is concave, and therefore Proposition 3.1 holds. Moreover, the result of Theorem 3.1 that we present below also holds.

It is natural to consider whether defining the q-value in terms of the original statistics is equivalent to defining the q-value in terms of the statistics' p-values. We will denote the pFDR based on the original statistics as $pFDR^X(\Gamma_\alpha)$, and the analogous pFDR based on the p-values by $pFDR^P(\{p \leq \alpha\})$. Is $pFDR^X(\Gamma_\alpha) = pFDR^P(\{p \leq \alpha\})$, and when is it the case that q-value$(x_i) = pFDR^P(\{p \leq \text{p-value}(x_i)\})$? We answer these questions in the following theorem.

Theorem 3.1 *For m identical hypothesis tests, $pFDR^X(\Gamma_\alpha) = pFDR^P(\{p : p \leq \alpha\})$, which implies that the q-value can be calculated from either the original statistics or their p-values. Also, when the statistics are independent then*

$$\textit{q-value}(x) = pFDR^P(\{p : p \leq \textit{p-value}(x)\}) \tag{3.6}$$

if and only if $G_1(\alpha)/\alpha$ is decreasing in α.

Proof: Because the set of rejection regions is nested, it is trivial to show that p-value$(x) \leq \alpha$ if and only if $x \in \Gamma_\alpha$. This implies $pFDR^X(\Gamma_\alpha) = pFDR^P(\{p : p \leq \alpha\})$. For the second statement, first suppose that $G_1(\alpha)/\alpha$ is decreasing in α. For any $X = x$, let $\Gamma_{\alpha'} = \underset{\{\Gamma_\alpha:\ x \in \Gamma_\alpha\}}{\arg\min}\ pFDR^X(\Gamma_\alpha)$, so that q-value$(x) = pFDR^X(\Gamma_{\alpha'}) = pFDR^P(\{p : p \leq \alpha'\})$. Since $G_0(\alpha) = \alpha$ also decreases in α, it is also the case that $\Gamma_{\alpha'} = \underset{\{\Gamma_\alpha:\ x \in \Gamma_\alpha\}}{\arg\min}\ \mathbf{Pr}(X \in \Gamma_\alpha | H = 0)$, i.e., p-value$(x) = \alpha'$. Now suppose that q-value$(x) = pFDR(\{p : p \leq \text{p-value}(x)\})$ for each x. By the definition of the q-value, this implies q-value(x) is an increasing function of p-value(x). Therefore $G_1(\alpha)/G_0(\alpha)$ is a decreasing function of α. □

Suppose that we perform m different hypothesis tests, so that each one has its own nested set of rejection regions, possibly on different spaces. One can transform these tests into the same space by calculating their p-values. By the results presented in this section, it follows that the p-value based q-values are a natural way to transform these tests onto the same space with respect to the pFDR. In Chapter 6 we provide a method for estimating q-values and show that these estimates are simultaneously conservatively consistent.

3.3 Simultaneous Controlling Curves

We find the concept of an adjusted p-value to be incomplete. Firstly, its definition and values are dependent on the controlling procedure. Therefore, two sets of FDR adjusted p-values derived from the same set of realized p-values, but using different FDR controlling procedures, can be completely different. For example, one should call them "FDR BH adjusted p-values" when using the BH procedure to adjust them. Secondly, for a given rejection region $[0, p_i]$ there is a true minimum error rate that can be attained, regardless of the procedure used or the p-values observed. This quantity in itself hasn't been considered in general. The adjusted p-value is one way to estimate the minimum attainable error rate, but the strength of this estimate is never evaluated in a rigorous statistical sense. Adjusted p-values are taken as is, with the assumption that they are "good." They should be compared to the minimum attainable error rate as the quantity they are estimating.

With respect to our criticism of the ambiguity of the "adjusted p-value" definition and its dependence upon a specific controlling procedure, we offer the following new definition. This new definition provides a generalization that extends beyond quantities that can be controlled by sequential p-value methods, and it provides an opportunity for one to define

rigorous criteria in placing a multiple hypothesis testing error measure on each of several p-values simultaneously.

Definition 3.2 *Let Err be some hypothesis testing error measure, where $Err(t)$ is the value of Err when rejecting all p-values in $[0,t]$. Define the* **simultaneous Err controlling curve** $\alpha_{Err}(\cdot)$ *to be*

$$\alpha_{Err}(t) = \inf_{s \geq t} Err(s) \text{ for each } t \in [0,1].$$

Specifically, consider the definition of the simultaneous FDR controlling curve of the FDR.

Definition 3.3 *The* **simultaneous FDR controlling curve** $\alpha_{FDR}(\cdot)$ *is defined to be*

$$\alpha_{FDR}(t) = \inf_{s \geq t} FDR(s), \text{ for each } t \in [0,1].$$

We conclude this chapter with the following observations about simultaneous controlling curves:

- The simultaneous controlling curve evaluated at t, $\alpha_{Err}(t)$, gives the minimum Err that can be attained when rejecting all p-values in $[0,t]$.
- The simultaneous controlling curve is deterministic, and independent of any particular estimate or sequential p-value method.
- The simultaneous controlling curve can be defined for an error measure, even those that cannot be controlled by a sequential p-value method in the traditional sense, such as the pFDR.
- The q-values are equivalent to the simultaneous pFDR controlling curve evaluated at the (random) p-values.
- Traditional "FWER adjusted p-values" are estimates of the simultaneous FWER controlling curve evaluated at the observed p-values. (The analogous statement holds true for "FDR adjusted p-values.") The performance of these "estimates" is typically not evaluated in a rigorous sense.
- The simultaneous controlling curve (or a conservative estimate of it) can be reported at each of the observed p-values, circumventing the need to choose a level beforehand at which to control the error measure.

Chapter 4

Estimating False Discovery Rates for Fixed Rejection Regions

From our own experience in applying false discovery rates to DNA microarrays, the biologists are rarely satisfied or interested in setting the FDR beforehand because they are more concerned about the balance between the FDR and the total number of hypotheses rejected. This makes sense from a practical viewpoint because the real use of the FDR for large numbers of hypothesis tests is to gauge how reliable it is to take the smallest 100 p-values as significant, for example. Scientists are very comfortable with p-values, and they often want to reject all p-values less than 0.001 or 0.005, while taking into account the multiple comparisons. The methodology we present in this chapter is geared towards this kind of situation.

Experts in a particular field (for example, DNA microarrays) run similar experiments over and over. Often they are able to judge from their experience which statistics are likely to be significant. Fixing the rejection region also makes sense for this reason. Of course, we are not stating that it never makes sense to fix the FDR beforehand and estimate the rejection region, but there are many situations, particularly when a thousand or more hypotheses are being tested, when it is more reasonable to fix the rejection region. It will also be seen that this approach allows us to control the positive False Discovery Rate (pFDR), which we find to be a more appealing error measure. Probably the most important reasons for fixing the rejection region are that it allows us to take a conceptually simpler approach to complicated compound error measures such as the FDR and pFDR, and it provides a unified framework for studying false discovery rates as mentioned in Section 1.4.

4.1 Estimation and Inference of the pFDR and FDR

In this section, we derive point estimates for the pFDR and FDR for a fixed rejection regions. This work will be limited without loss of generality to the case where we reject based on p-values. It follows that for p-value based rejections, all rejection regions are of the form $[0, t]$ for some $t \geq 0$. Instead of denoting rejection regions by the more abstract Γ, we denote them by t, which refers to the interval $[0, t]$. In terms of p-values we can write the result of Theorem 2.1 as

$$pFDR(t) = \frac{\pi_0 \cdot \mathbf{Pr}(P \leq t | H = 0)}{\mathbf{Pr}(P \leq t)} = \frac{\pi_0 \cdot t}{\mathbf{Pr}(P \leq t)},$$

where P is the random p-value resulting from any test.

Since $\pi_0 \cdot m$ of the p-values are expected to be null, then the largest p-values are most likely to come from the null, uniformly distributed p-values. Moreover, $\pi_0 \cdot (1 - \lambda)$ of the null p-values are expected to fall in $(\lambda, 1]$, and a small proportion of alternatives p-values will fall in $(\lambda, 1]$. Hence, a good estimate of π_0 is

$$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1 - \lambda)m} = \frac{W(\lambda)}{(1 - \lambda)m}$$

for some well chosen λ, where $p_1, \ldots, p_m$ are the observed p-values and $W(\lambda) = \#\{p_i > \lambda\}$. (Recall the definitions of W and R from Table 1.1.) For now we assume that λ is fixed, however, we show how to pick the optimal λ in Section 7.1. A natural estimate of $\mathbf{Pr}(P \leq t)$ is

$$\widehat{\mathbf{Pr}}(P \leq t) = \frac{\#\{p_i \leq t\}}{m} = \frac{R(t)}{m},$$

where $R(t) = \#\{p_i \leq t\}$. Therefore, a good estimate of $pFDR(t)$ for fixed λ is

$$\widehat{Q}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\widehat{\mathbf{Pr}}(P \leq t)} = \frac{W(\lambda) \cdot t}{(1 - \lambda) \cdot R(t)}. \tag{4.1}$$

The pFDR and the FDR are asymptotically equivalent for a fixed rejection region. We see in Section 4.4 that $\widehat{Q}_\lambda(t)$ shows good asymptotic properties for the pFDR. In fact, in Section 4.6 we show it is a maximum likelihood estimate. However, due to finite sample considerations, we have to make two slight adjustments in order to estimate the pFDR.

When $R(t) = 0$, the estimate would be undefined, which is undesirable for finite samples.

Therefore, we replace $R(t)$ with $R(t) \vee 1$. This is equivalent to making a linear interpolation between the estimate at $[0, p_{(1)}]$ and the origin. Also, $1-(1-t)^m$ is clearly a lower bound for $\mathbf{Pr}(R(t) > 0)$. Since the pFDR is conditioned on $R(t) > 0$, we divide by $1-(1-t)^m$. (See Section 6.2 for more on why we do this.) This is made clearer by noting that

$$pFDR(t) = \frac{\pi_0 \cdot \mathbf{Pr}(P \leq t | H = 0, R(t) > 0)}{\mathbf{Pr}(P \leq t | R(t) > 0)}.$$

$[R(t) \vee 1]/m$ is an underestimate of $\mathbf{Pr}(P \leq t | R(t) > 0)$ and $t/[1-(1-t)^m]$ is an overestimate of $\mathbf{Pr}(P \leq t | H = 0, R(t) > 0)$. Therefore, we estimate the pFDR as

$$\widehat{pFDR}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\widehat{\mathbf{Pr}}(P \leq t) \cdot [1-(1-t)^m]} = \frac{W(\lambda) \cdot t}{[1-\lambda] \cdot [R(t) \vee 1] \cdot [1-(1-t)^m]}. \tag{4.2}$$

Since the FDR is not conditioned on at least one rejection occurring, we can set

$$\widehat{FDR}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\widehat{\mathbf{Pr}}(P \leq t)} = \frac{W(\lambda) \cdot t}{[1-\lambda] \cdot [R(t) \vee 1]}. \tag{4.3}$$

For large m these two estimates are equivalent, but we find the pFDR quantity to be more appropriate and think it should be used. Note that when $t = 1/m$, $\mathbf{Pr}(R(t) > 0)$ can be as small as 0.632, so the FDR can be measuring small p-values rather than a small "false discovery rate." For fixed m and $t \to 0$, $FDR(t)$ and $\widehat{FDR}_\lambda(t)$ show unsatisfying properties, and we show this in Section 6.2.

We show in Section 4.4 that $\widehat{pFDR}_\lambda(t)$ and $\widehat{FDR}_\lambda(t)$ offer an analogous property to strong control in that they are conservatively biased for all π_0. However, as we argued in the introduction, the expected value of a multiple hypothesis testing procedure is not a broad enough picture. Since the p-values are independent, we can sample them with replacement to obtain standard bootstrap samples. From these we can form bootstrap versions of our estimate and provide upper confidence limits for the pFDR and FDR. This allows one to make much more precise statements about how much multiple hypothesis testing "control" is being offered. The full details of the estimation and inference of $pFDR(t)$ are given in Algorithm 4.1. The same algorithm holds for the estimation and inference of $FDR(t)$, except we obviously use $\widehat{FDR}_\lambda(t)$ instead. In Section 7.1, we extend our methodology to include an automatic method for choosing the optimal λ.

If $\widehat{pFDR}_\lambda(t) > 1$, we recommend setting $\widehat{pFDR}_\lambda(t) = 1$ since obviously $pFDR(t) \leq 1$.

Algorithm 4.1: Estimation and Inference of $pFDR(t)$ and $FDR(t)$

1. For the m hypothesis tests, calculate their respective p-values $p_1, \ldots, p_m$.
2. Estimate π_0 and $\mathbf{Pr}(P \leq t)$ by

$$\widehat{\pi}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)m} \text{ and } \widehat{\mathbf{Pr}}(P \leq t) = \frac{R(t) \vee 1}{m},$$

   where $R(t) = \#\{p_i \leq t\}$ and $W(\lambda) = \#\{p_i > \lambda\}$.

3. For any rejection region of interest $[0, t]$, estimate $pFDR(t)$ by

$$\widehat{pFDR}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\widehat{\mathbf{Pr}}(P \leq t) \cdot [1 - (1-t)^m]},$$

   for some well chosen λ. (See Section 7.1 for how to choose the optimal λ.)

4. For B bootstrap samples of $p_1, \ldots, p_m$, calculate the bootstrap estimates $\widehat{pFDR}_\lambda^{*b}(t)$ $(b = 1, \ldots, B)$ similarly to above.
5. Form a $1 - \alpha$ upper confidence interval for $pFDR(t)$ by taking the $1 - \alpha$ quantile of the $\widehat{pFDR}_\lambda^{*b}(t)$ as the upper confidence bound.
6. For $FDR(t)$, perform this same procedure except using

$$\widehat{FDR}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\widehat{\mathbf{Pr}}(P \leq t)}.$$

We could smooth the estimate so that it is always less or equal to one, but we have taken a simpler approach here. The same comment holds for $\widehat{FDR}_\lambda(t)$.

Even though the estimates presented in this section are new, the approach has implicitly been taken before. Yekutieli & Benjamini (1999) introduced the idea of estimating the FDR under dependence within the Benjamini & Hochberg (1995) framework. Also, Benjamini & Hochberg (2000) incorporate an estimate of m_0 into their original algorithm in a *post hoc* fashion. Tusher et al. (2001) fix the rejection region and estimate the FDR.

4.2 A Connection Between the Two Approaches

In this section we present a heuristic connection between the sequential p-value method of Benjamini & Hochberg (1995) and the approach presented in the previous section. The goal is to provide insight into the increased power and effectiveness of our proposed approach. The connection is made more rigorous in Chapter 5.

> The basic point we make is that using the Benjamini & Hochberg (1995) method to control the FDR at level α/π_0 is equivalent to (i.e., rejects the same p-values as) using the proposed method to control the FDR at level α. The gain in power from our approach is clear – we control a smaller error rate ($\alpha \leq \alpha/\pi_0$), yet reject the same number of tests.

Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered, observed p-values for the m hypothesis tests. The method of Benjamini & Hochberg (1995) finds $\widehat{k}$ such that

$$\widehat{k} = \max\{k : p_{(k)} \leq k/m \cdot \alpha\}. \tag{4.4}$$

Rejecting $p_{(1)}, \ldots, p_{(\widehat{k})}$ provides $FDR \leq \alpha$.

Now suppose we use our method and take the most conservative estimate $\widehat{\pi}_0(\lambda) = 1$. Then the estimate $\widehat{FDR}(t) \leq \alpha$ if we reject $p_{(1)}, \ldots, p_{(\widehat{l})}$ such that

$$\widehat{l} = \max\{l : \widehat{FDR}(p_{(l)}) \leq \alpha\}.$$

But since $\widehat{FDR}(p_{(l)}) = \frac{\widehat{\pi}_0(\lambda) \cdot p_{(l)}}{l/m}$ this equivalent to (with $\widehat{\pi}_0(\lambda) = 1$)

$$\widehat{l} = \max\{l : p_{(l)} \leq l/m \cdot \alpha\}.$$

Therefore, $\widehat{k} = \widehat{l}$ when $\widehat{\pi}_0(\lambda) = 1$. Moreover, if we take the better estimate

$$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m}$$

then $\widehat{l} > \widehat{k}$ with high probability.

Therefore, we have shown that $\widehat{l} \geq \widehat{k}$. In other words, using our approach, we reject a greater number of hypotheses while controlling the same error measure, which leads to greater power. The operational difference between our method and the Benjamini &

Hochberg (1995) method is the inclusion of $\widehat{\pi}_0(\lambda)$. It is important to note, however, that we did not simply reverse their method and stick in $\widehat{\pi}_0(\lambda)$. Rather, we took a very different approach, starting from simple facts about the pFDR under independence with fixed rejection regions. Benjamini & Hochberg (1995) do not give us much insight into why they chose their particular sequential p-value method. This comparison sheds some light onto why it works.

An asymptotic comparison can also be made by using the results of Genovese & Wasserman (2001) and those of Section 4.4. The pFDR and FDR are asymptotically equivalent under independence since $\mathbf{Pr}(R(t) > 0) \to 1$. The p-value cut-off determined by equation (4.4) goes to a fixed value as $m \to \infty$, and it is the solution to Theorem 2.1. Genovese & Wasserman (2001) define the optimal method in terms of the FDR and the FNR (False Non-discovery Rate). We conjecture that the asymptotically optimal method (or estimate) is always in or a limit of the family $\widehat{Q}_\lambda(t)$ for $\lambda \in [0, 1)$. Genovese & Wasserman (2001), however, only define optimality in terms of expected values, and we have stated that the variance of the process should also be taken into account. The results of Section 4.6 show that our method also minimizes the asymptotic variance.

4.3 A Numerical Study

In this section we present some numerical results in order to compare the power of the Benjamini & Hochberg (1995) procedure to our proposed method. We denote the results from the proposed method as "PM", and from the Benjamini & Hochberg (1995) procedure as "BH". As mentioned in Section 4.2, it is not straightforward to compare these two methods since BH estimates the rejection region while PM estimates the FDR. We circumvent this problem by using the BH procedure to control the FDR at $\widehat{FDR}_\lambda(t)$ for each iteration.

We looked at the two rejection regions $t = 0.01$ and $t = 0.001$ over several values of π_0. The values of t and π_0 were chosen in order to cover a wide variety of situations. We performed $m = 1000$ hypothesis tests of $\mu = 0$ versus $\mu = 2$ for independent random variables $Z_i \sim N(\mu, 1)$, $i = 1, \ldots, 1000$, over 1000 iterations. The null hypothesis for each test is that $\mu = 0$, so the proportion of $Z_i \sim N(0, 1)$ was set to π_0; hence, π_1 of the statistics have the alternative distribution $N(2, 1)$. For each test the p-value is defined as $p_i = \mathbf{Pr}(N(0, 1) \geq z_i)$, where z_i is the observed value of Z_i.

In order to calculate the power of PM, test i was rejected if $p_i \leq t$, and the power was

calculated accordingly. Also, $\widehat{FDR}(t)$ was calculated as we outlined in Section 4.1. The BH method was performed at level $\widehat{FDR}(t)$, and the power was calculated. This approach should put the two methods on equal ground for comparison; reporting $\widehat{FDR}(t)$ is equivalent in practice to using the BH method to control the FDR at level $\widehat{FDR}(t)$.

The simulations were performed for $\pi_0 = 0.1, 0.2, \ldots, 0.9$. Even though here we know the alternative distribution of the p-values, we did not use this knowledge. Instead, we estimated the FDR as if the alternative distribution was unknown. Therefore, we had to choose a value of λ in order to estimate π_0; we used $\lambda = 1/2$ in all calculations for simplicity.

Table 4.1: A numerical comparison between the BH and proposed methods

π_0	FDR	Power (PM)	Power (BH)	$\mathbf{E}[\widehat{FDR}]$ (PM)	$\mathbf{E}[\widehat{\pi}_0]$ (PM)	$\mathbf{E}[\widehat{t}]$ (BH)
$t = 0.01$						
0.1	0.003	0.372	0.074	0.004	0.141	0.0003
0.2	0.007	0.372	0.122	0.008	0.236	0.0008
0.3	0.011	0.372	0.164	0.013	0.331	0.001
0.4	0.018	0.372	0.203	0.019	0.426	0.002
0.5	0.026	0.372	0.235	0.027	0.523	0.003
0.6	0.039	0.372	0.268	0.040	0.618	0.004
0.7	0.060	0.371	0.295	0.061	0.714	0.005
0.8	0.097	0.372	0.319	0.099	0.809	0.007
0.9	0.195	0.372	0.344	0.200	0.905	0.008
$t = 0.001$						
0.1	0.0008	0.138	0.016	0.001	0.141	1×10^{-5}
0.2	0.002	0.138	0.031	0.002	0.236	5×10^{-5}
0.3	0.003	0.137	0.046	0.003	0.331	0.0001
0.4	0.005	0.138	0.060	0.005	0.426	0.0002
0.5	0.007	0.138	0.074	0.008	0.523	0.0003
0.6	0.011	0.138	0.088	0.011	0.618	0.0004
0.7	0.017	0.138	0.101	0.017	0.714	0.0005
0.8	0.028	0.138	0.129	0.030	0.809	0.0006
0.9	0.061	0.137	0.133	0.066	0.905	0.0008

Table 4.1 shows the results of the simulation study. The first half of the table corresponds to $t = 0.01$, and the second half corresponds to $t = 0.001$. It can be seen that there is a substantial increase in power using the proposed method. One case even gives over an 800% increase in power. The power is constant over each case of PM because the same rejection region is used. The power of BH increases as π_0 gets larger because the procedure becomes less conservative. In fact, it follows from Section 4.2 that as $\pi_0 \to 1$, the BH method

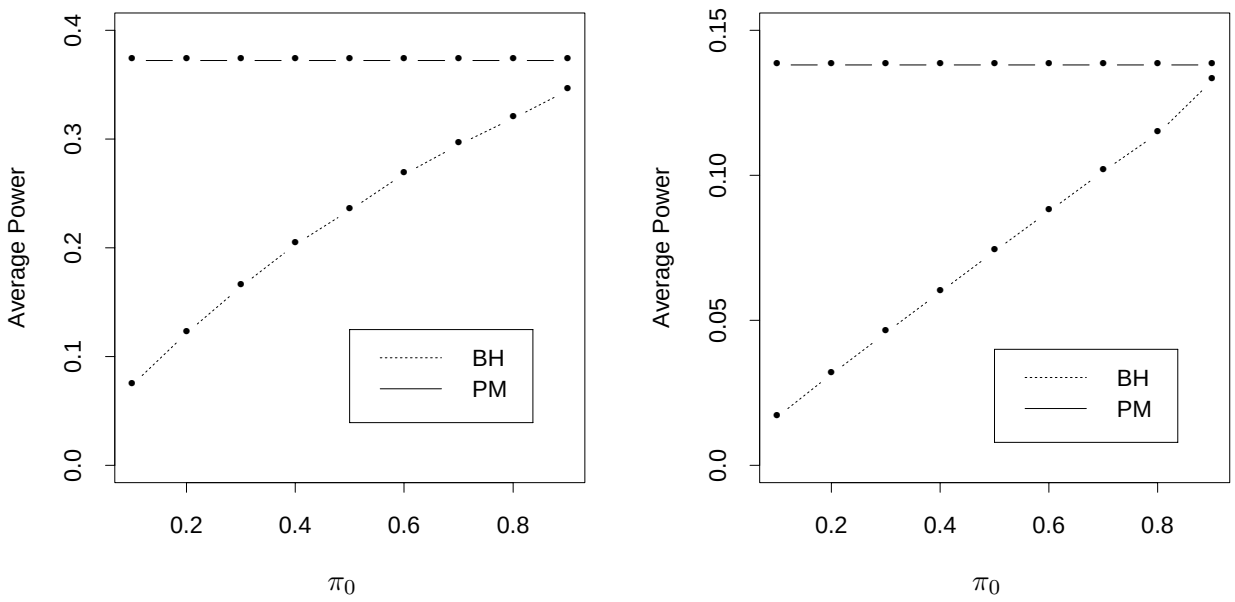


Figure 4.1: A plot of average power versus π_0 for the BH method (BH) and the proposed method (PM). The left panel is the case where the rejection region is defined by $t = 0.01$, and the right panel where $t = 0.001$. It can be seen that there is a substantial increase in power under the proposed method in both situations.

becomes the PM method.

The fifth column of Table 4.1 shows $\mathbf{E}[\widehat{FDR}]$ for PM. It can be seen that this is very close to the true FDR in the second column (usually within 0.1%), and it is always conservative. The PM method is nearly optimal in that it estimates the $FDR(t)$ basically as close as conservatively possible for each rejection region. Therefore, we essentially lose no power regardless of the value of π_0. Moreover the method gets better as the number of tests increases; the opposite has been true in the past. The sixth column shows $\mathbf{E}[\widehat{\pi}_0]$ for PM. It can be seen that this estimate is always conservative and very close to the actual value. Recall that the BH method essentially estimates the rejection region $[0, t]$. The eighth column shows $\mathbf{E}[\widehat{t}]$ over the 1000 realizations of $\widehat{t}$. It can be seen that these estimates are quite conservative. The power comparisons are also shown graphically in Figure 4.1.

The success of our method also largely depends on how well we can estimate $pFDR(t)$ and $FDR(t)$. It is seen in this simulation that the estimates are very good. This is especially

due to the fact that the power-Type I error curve is well behaved in the sense discussed in Section 4.4. If we choose λ more adaptively (Section 7.1), the estimation is even better since we take into account both bias and variance.

4.4 Finite Sample Results

In this section, we provide finite sample results for $\widehat{pFDR}_\lambda(t)$ and $\widehat{FDR}_\lambda(t)$. Our goal of course is to provide conservative estimates of $pFDR(t)$ and $FDR(t)$. For example, we want $\widehat{pFDR}_\lambda(t) \geq pFDR(t)$ as much as possible without being too conservative. Note that these results *do not depend of the mixture distribution assumptions of Chapter 2*, although the mixture model motivated the estimates. We cover several levels of dependence, and these are explicitly stated in the theorems.

First recall the notation

$$\begin{aligned} V(t) &= \#\{\text{null } p_i : p_i \leq t\} \\ S(t) &= \#\{\text{alternative } p_i : p_i \leq t\}, \\ R(t) &= \#\{p_i : p_i \leq t\}, \end{aligned}$$

which we will use throughout this section. The following is our main finite sample result, and it only depends on the null p-values being independent.

Theorem 4.1 *If the p-values corresponding to the true null hypotheses are independent then*
$\mathbf{E}[\widehat{pFDR}_\lambda(t)] \geq pFDR(t)$ *and* $\mathbf{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$ *for all* t *and* π_0.

Proof: Note that

$$\widehat{pFDR}_\lambda(t) - pFDR(t) \geq \frac{1}{\mathbf{Pr}(R(t) > 0)} \left[\widehat{FDR}_\lambda(t) - FDR(t)\right],$$

so it suffices to show $\mathbf{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$. It follows that

$$\frac{(m - R(\lambda)) \cdot t}{(1-\lambda) \cdot R(t)} \geq \frac{(m_0 - V(\lambda)) \cdot t}{(1-\lambda) \cdot R(t)}.$$

Moreover,

$$FDR(t) = \mathbf{E}\left[\frac{V(t)}{R(t) \vee 1}\right].$$

Thus,

$$\begin{aligned}
\widehat{FDR}_\lambda(t) - FDR(t) &\geq \mathbf{E}\left[\frac{\frac{m_0 - V(\lambda)}{1-\lambda}t - V(t)}{R(t) \vee 1}\right] \\
&= \mathbf{E}\left[\left.\frac{\frac{m_0 - V(\lambda)}{1-\lambda}t - V(t)}{R(t)}\right| R(t) > 0\right]\mathbf{Pr}(R(t) > 0) \\
&+ \mathbf{E}\left[\left.\frac{m_0 - V(\lambda)}{1-\lambda}t\right| R(t) = 0\right]\mathbf{Pr}(R(t) = 0).
\end{aligned}$$

Conditioning on $V(t)$, we get

$$\mathbf{E}\left[\left.\frac{\frac{m_0 - V(\lambda)}{1-\lambda}t - V(t)}{R(t)}\right| V(t)\right] = \frac{\frac{m_0 - \mathbf{E}[V(\lambda)|V(t)]}{1-\lambda}t - V(t)}{V(t) + S(t)}, \tag{4.5}$$

since $V(t)$ and $S(t)$ are independent. Also by independence, $\mathbf{E}[V(\lambda)|V(t)]$ is a non-decreasing linear function of $V(t)$. Therefore, by conditioning on $S(t)$ and Jensen's inequality on $V(t)$, and then Jensen's inequality on $S(t)$, we get

$$\begin{aligned}
&\mathbf{E}\left[\left.\frac{\frac{m_0 - V(\lambda)}{1-\lambda}t - V(t)}{R(t)}\right| R(t) > 0\right]\mathbf{Pr}(R(t) > 0) \geq \\
&\frac{\mathbf{E}[\frac{m_0 - V(\lambda)}{1-\lambda}|R(t) > 0]t - \mathbf{E}[V(t)|R(t) > 0]}{\mathbf{E}[R(t)|R(t > 0]}\mathbf{Pr}(R(t) > 0).
\end{aligned}$$

Since $\mathbf{E}[R(t)|R(t) > 0] \geq 1$, it follows

$$\mathbf{E}\left[\left.\frac{m_0 - V(\lambda)}{1-\lambda}\right| R(t) = 0\right] t \cdot \mathbf{Pr}(R(t) = 0) \geq \frac{\mathbf{E}\left[\left.\frac{m_0 - V(\lambda)}{1-\lambda}\right| R(t) = 0\right] t}{\mathbf{E}[R(t)|R(t) > 0]}\mathbf{Pr}(R(t) = 0)$$

Putting all of this together we get

$$\mathbf{E}[\widehat{FDR}_\lambda(t)] - FDR(t) \geq \frac{\mathbf{E}[\frac{m_0 - V(\lambda)}{1-\lambda}]t - \mathbf{E}[V(t)]}{\mathbf{E}[R(t)|R(t) > 0]} \geq 0.$$

□

This result is analogous to showing "strong control" of our method. The theorem is stated under the assumption that we do not truncate the estimates at one. Of course in practice we would truncate the estimates at one since $FDR \leq pFDR \leq 1$, but the expected value of the estimates nevertheless behaves as we would want it. The following result shows that truncating the estimates is a good idea when taking into account both bias and variance.

Lemma 4.1 $\mathbf{E}[(\widehat{pFDR}_\lambda(t) - pFDR(t))^2] > \mathbf{E}[(\widehat{pFDR}_\lambda(t) \wedge 1 - pFDR(t))^2]$ *and* $\mathbf{E}[(\widehat{FDR}_\lambda(t) - FDR(t))^2] > \mathbf{E}[(\widehat{FDR}_\lambda(t) \wedge 1 - FDR(t))^2]$.

Proof: This easily follows by noting

$$\mathbf{E}[(\widehat{pFDR}_\lambda(t) - pFDR(t))^2 | \widehat{pFDR}_\lambda(t) > 1] > \mathbf{E}[(\widehat{pFDR}_\lambda(t) \wedge 1 - pFDR(t))^2 | \widehat{pFDR}_\lambda(t) > 1]$$

since $pFDR(t) \leq 1$. The proof for $\widehat{FDR}_\lambda(t)$ follows similarly. □

We can also extend Theorem 4.1 to cases where stronger dependence holds. Note that the factor $1/[1-(1-t)^m]$ is in denominator of $\widehat{pFDR}_\lambda(t)$. This factor was chosen because $1-(1-t)^m \leq \mathbf{Pr}(R(t) > 0)$. When dependence exists, this factor can no longer be used. But it is feasible to simulate $\mathbf{Pr}(R^0(t) > 0)$, where $R^0(t)$ is the number of significant hypotheses under the full null model. Therefore, in the following results, we assume that $1-(1-t)^m$ has been replaced with $\mathbf{Pr}(R^0(t) > 0)$. See Chapter 8 for more on how to calculate this quantity.

For the following result, the dependence can exist between the null statistics, as well as between the alternative statistics, but V and S are independent. We can show the estimates provide a conservative bias in expectation under the curious condition that

$$\mathbf{E}\left[\frac{V(\lambda)}{\mathbf{E}[V(\lambda)]}\middle| V(t)\right] \leq \frac{V(t)}{\mathbf{E}[V(t)]}.$$

Note that since it will often be the case that $\lambda \geq t$, this condition is equivalent to showing that $V(t)/\mathbf{E}[V(t)]$ is a supermartingale in t with $\mathcal{F}_t = \sigma(V(\Gamma_{t'}), 0 \leq t' \leq t)$ being the filtration.

Theorem 4.2 $\mathbf{E}[\widehat{pFDR}_\lambda(t)] \geq pFDR(t)$ *and* $\mathbf{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$ *if* $V(t)$ *and* $S(t)$ *are*

independent and

$$\mathbf{E}\left[\frac{V(\lambda)}{\mathbf{E}[V(\lambda)]}\middle| V(t)\right] \leq \frac{V(t)}{\mathbf{E}[V(t)]}.$$

Proof: As in Theorem 4.1 it suffices to show $\mathbf{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$. Also, this proof is very similar to that of Theorem 4.1. Note that under our assumption

$$\frac{\frac{m_0-\mathbf{E}[V(\lambda)|V(t)]}{1-\lambda}t - V(t)}{V(t)+S(t)} \geq \frac{\frac{m_0-\mathbf{E}[V(\lambda)]/\mathbf{E}[V(t)]\cdot V(t)}{1-\lambda}t - V(t)}{V(t)+S(t)},$$

where the left hand side is from equation (4.5). The rest follows similarly. □

For this last result, the dependence exists between all statistics. If we have that

$$\mathbf{E}\left[\frac{R(\lambda)}{\mathbf{E}[R(\lambda)]}\middle| R(t)\right] \leq \frac{R(t)}{\mathbf{E}[R(t)]} \text{ and } \mathbf{E}\left[\frac{V(t)}{\mathbf{E}[V(t)]}\middle| R(t)\right] \leq \frac{R(t)}{\mathbf{E}[R(t)]},$$

then our estimates are conservatively biased. Note that when we restrict $\lambda \geq t$, then the first condition is equivalent to that $R(t)/\mathbf{E}[R(t)]$ is a supermartingale in t with $\mathcal{F}_t = \sigma(R(\Gamma_{t'}), 0 \leq t' \leq t)$ being the filtration.

Theorem 4.3 *For arbitrary dependence among all statistics,* $\mathbf{E}[\widehat{pFDR}_\lambda(t)] \geq pFDR(t)$ *and* $\mathbf{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$ *if*

$$\mathbf{E}\left[\frac{R(\lambda)}{\mathbf{E}[R(\lambda)]}\middle| R(t)\right] \leq \frac{R(t)}{\mathbf{E}[R(t)]} \text{ and } \mathbf{E}\left[\frac{V(t)}{\mathbf{E}[V(t)]}\middle| R(t)\right] \leq \frac{R(t)}{\mathbf{E}[R(t)]}.$$

Proof: Once again, as in Theorem 4.1 it suffices to show $\mathbf{E}[\widehat{FDR}_\lambda(t)] \geq FDR(t)$. Similarly to Theorem 4.1, we have

$$\begin{aligned}
\widehat{FDR}_\lambda(t) - FDR(t) &\geq \mathbf{E}\left[\frac{\frac{m-R(\lambda)}{1-\lambda}t - V(t)}{R(t)\vee 1}\right] \\
&\geq \mathbf{E}\left[\frac{\frac{m-R(\lambda)}{1-\lambda}t - V(t)}{R(t)}\middle| R(t)>0\right]\mathbf{Pr}(R(t)>0).
\end{aligned}$$

Conditioning on $R(t)$, we get

$$\begin{aligned}\mathbf{E}\left[\left.\frac{\frac{m-R(\lambda)}{1-\lambda}t - V(t)}{R(t)}\right| R(t)\right] &= \frac{\frac{m-\mathbf{E}[R(\lambda)|R(t)]}{1-\lambda}t - \mathbf{E}[V(t)|R(t)]}{R(t)} \\ &\geq \frac{\frac{m-\mathbf{E}[R(\lambda)]/\mathbf{E}[R(t)]\cdot R(t)}{1-\lambda}t - \mathbf{E}[V(t)]/\mathbf{E}[R(t)]\cdot R(t)}{R(t)}.\end{aligned}$$

Therefore, by Jensen's inequality on $R(t)$, we get

$$\mathbf{E}\left[\left.\frac{\frac{m-R(\lambda)}{1-\lambda}t - V(t)}{R(t)}\right| R(t) > 0\right]\mathbf{Pr}(R(t) > 0) \geq \frac{\frac{m-\mathbf{E}[R(\lambda)]}{1-\lambda}t - \mathbf{E}[V(t)]}{\mathbf{E}[R(t)|R(t) > 0]} \geq 0.$$

It follows $\mathbf{E}[\widehat{FDR}_\lambda(t)] - FDR(t) \geq 0$ and $\mathbf{E}[\widehat{pFDR}_\lambda(t)] - pFDR(t) \geq 0$. □

It is feasible that the assumptions of Theorems 4.2 and 4.3 can be checked in practice or may hold under certain models.

4.5 Large Sample Results

We now present large sample results for $\widehat{pFDR}_\lambda(t)$ and $\widehat{FDR}_\lambda(t)$. False discovery rates are most useful in cases in which many hypotheses are tested. Moreover, in the application we have in mind (gene expression data), m is typically on the order of several thousand, in which case the assumption of independence can be replaced with "weak dependence." In this section, we prove several theorems that all require the almost sure pointwise convergence of the empirical distributions of the null p-values and alternative p-values. Given these conditions, we are able to prove much stronger almost sure results for the estimates $\widehat{FDR}_\lambda$ and $\widehat{pFDR}_\lambda$ than their pointwise convergence.

The almost sure pointwise convergence of the empirical distributions of the null p-values and alternative p-values is likely for many applications. For example, the dependence we will discuss in Chapter 8 regarding detecting differential gene expression in DNA microarrays likely meets this "weak dependence" criterion.

Note that $V(t)/m_0$ and $S(t)/m_1$ are the empirical distribution functions of the null and alternative hypotheses, respectively. Almost sure convergence in the point-wise sense as

$m \to \infty$ means that with probability 1:

$$\begin{aligned} \frac{V(t)}{m_0} &\to G_0(t) \text{ for each } t \in [0,1], \\ \frac{S(t)}{m_1} &\to G_1(t) \text{ for each } t \in [0,1], \end{aligned} \tag{4.6}$$

for some functions G_0 and G_1.

Theorem 4.4 *Suppose that $V(\cdot)/m_0$ and $S(\cdot)/m_1$ converge almost surely point-wise to G_0 and G_1, respectively, where $G_0(t) = t$. Also suppose that $\lim_{m\to\infty} m_0/m = \pi_0$ exists. Then for each $\delta > 0$,*

$$\lim_{m\to\infty} \sup_{t\geq\delta} \left| \widehat{pFDR}_\lambda(t) - \frac{\pi_0 + \frac{1-G_1(\lambda)}{1-\lambda} \cdot \pi_1}{\pi_0} pFDR(t) \right| \stackrel{a.s.}{=} 0.$$

The analogous result holds for $\widehat{FDR}_\lambda(t)$ and $FDR(t)$.

Proof: It easily follows that

$$\lim_{m\to\infty} \sup_{t\geq\delta} \left| \widehat{pFDR}_\lambda(t) - \frac{\pi_0 + \frac{1-G_1(\lambda)}{1-\lambda} \cdot \pi_1}{\pi_0} \frac{mt}{[R_m(t) \vee 1] \cdot [1-(1-t)^m]} \right| \stackrel{a.s.}{=} 0.$$

Moreover,

$$\begin{aligned} &\lim_{m\to\infty} \sup_{t\geq\delta} \left| \frac{mt}{[R_m(t) \vee 1] \cdot [1-(1-t)^m]} - \frac{mt}{R_m(t) \vee 1} \right| \stackrel{a.s.}{\leq} \\ &\lim_{m\to\infty} \sup_{t\geq\delta} \left| \frac{mt}{[R_m(t) \vee 1] \cdot [1-(1-\delta)^m]} - \frac{mt}{R_m(t) \vee 1} \right| \stackrel{a.s.}{=} 0. \end{aligned}$$

Also, from arguments given in the proof of Theorem 2.3, we know

$$\lim_{m\to\infty} \sup_{t\geq\delta} \left| \frac{mt}{R_m(t) \vee 1} - pFDR(t) \right| \stackrel{a.s.}{=} 0.$$

The argument for the FDR case follows similarly. □

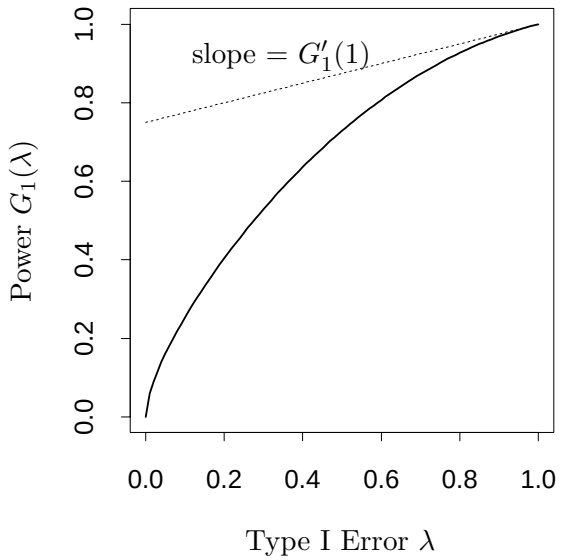


Figure 4.2: A plot of power $G_1(\lambda)$ versus Type I error λ. It can be seen that since G_1 is concave $\frac{1-G_1(\lambda)}{1-\lambda}$ gets smaller as $\lambda \to 1$. The line has slope equal to $\lim_{\lambda\to 1} \frac{1-G_1(\lambda)}{1-\lambda}$, which is the smallest value of $\frac{1-G_1(\lambda)}{1-\lambda}$ that can be attained for concave G_1.

Corollary 4.1 *Under the assumptions of Theorem 4.4,*

$$\lim_{m\to\infty} \inf_{t\geq\delta} \left[\widehat{FDR}_\lambda(t) - FDR(t)\right] \geq 0, \ \lim_{m\to\infty} \inf_{t\geq\delta} \left[\widehat{pFDR}_\lambda(t) - pFDR(t)\right] \geq 0,$$

$$\lim_{m\to\infty} \inf_{t\geq\delta} \left[\widehat{FDR}_\lambda(t) - \frac{V_m(t)}{R_m(t)\vee 1}\right] \geq 0, \textit{ and } \lim_{m\to\infty} \inf_{t\geq\delta} \left[\widehat{pFDR}_\lambda(t) - \frac{V_m(t)}{R_m(t)\vee 1}\right] \geq 0$$

with probability 1.

Proof: This easily follows by noting that

$$\frac{\pi_0 + \frac{1-G_1(\lambda)}{1-\lambda}\cdot\pi_1}{\pi_0} \geq 1,$$

and combining the results of Theorems 2.3 and 4.4. □

Theorem 4.4 can be understood graphically in terms of the plot of power to Type I error for each rejection region $[0, \lambda]$. The function $G_1(\lambda)$ gives the power over the rejection region

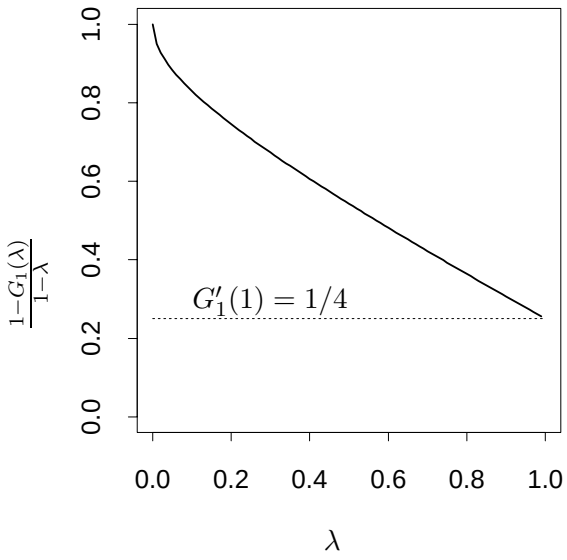


Figure 4.3: A plot of $\frac{1-G_1(\lambda)}{1-\lambda}$ versus λ is shown for a concave G_1. It can be seen that the minimum is obtained at $\lambda = 1$ with value $G_1'(1) = 1/4$.

$[0, \lambda]$, and of course the Type I error over this region is λ. The estimate of π_0 is taken over the interval $(\lambda, 1]$, so that $1 - G_1(\lambda)$ is the probability of a p-value from the alternative distribution falling into $(\lambda, 1]$. Likewise, $1 - \lambda$ is the probability of null p-value falling into $(\lambda, 1]$. The estimate of π_0 is better the more $G_1(\lambda) > \lambda$. This is the case since the interval $(\lambda, 1]$ will contain less alternative p-values, and hence the estimate will be less conservative. Figure 4.2 shows a plot of $G_1(\lambda)$ versus λ for a concave G_1. For concave G_1, the estimate of π_0 becomes less conservative as $\lambda \to 1$. This is formally stated in the following corollary.

Corollary 4.2 *Suppose the assumptions of Theorem 4.4 hold. If G_1 is concave then*

$$\inf_{\lambda} \lim_{m\to\infty} \widehat{pFDR}_\lambda(t) \overset{a.s.}{=} \lim_{\lambda\to 1} \lim_{m\to\infty} \widehat{pFDR}_\lambda(t) \overset{a.s.}{=} \frac{\pi_0 + G_1'(1)\cdot\pi_1}{\pi_0} \lim_{m\to\infty} pFDR(t), \qquad (4.7)$$

where $G_1'(1)$ is the derivative of G_1 evaluated at 1.

Proof: Since $G_1(\lambda)$ is concave in λ, $\frac{1-G_1(\lambda)}{1-\lambda}$ is non-increasing in λ. Therefore, the minimum of $\frac{1-G_1(\lambda)}{1-\lambda}$ is obtained at $\lim_{\lambda\to 1} \frac{1-G_1(\lambda)}{1-\lambda}$. By L'Hopital's rule, $\lim_{\lambda\to 1} \frac{1-G_1(\lambda)}{1-\lambda} = G_1'(1)$. □

In other words, the right hand side of equation (4.7) is the tightest upper bound $\widehat{pFDR}(t)$

can attain on the pFDR as $m \to \infty$ for concave G_1. The corollary can be seen graphically in Figure 4.3. A plot of $\frac{1-G_1(\lambda)}{1-\lambda}$ versus λ is shown for a concave G_1. It can be seen that the minimum is obtained at $\lambda = 1$. The minimum value is $G_1'(1)$, which happens to be $1/4$ in this graph. Whenever the rejection regions are based on a monotone function of the likelihood ratio between the null and alternative hypotheses, G_1 is concave. Note that if G_1 is not concave, then the optimal λ used in the estimate of π_0 may not be attained as $\lambda \to 1$. A nice property of this last result is that $G_1'(1) = 0$ whenever testing a single parameter of an exponential family. Therefore, in many of the common cases, we can get exact convergence as $\lambda \to 1$.

4.6 $\widehat{Q}(t)$ is a Maximum Likelihood Estimate

In order to estimate the pFDR, a sensible place to start is to find the maximum likelihood estimate of $pFDR(t)$. We will assume throughout this section that $P_i \overset{i.i.d.}{\sim} \pi_0 \cdot G_0 + \pi_1 \cdot G_1$ for $i = 1, \ldots, m$ with $G_0(t) = t$. Also, let $G = G_0 + G_1$. We are interested in finding maximum likelihood estimates of π_0 and $G(t)$, so that we may combine them to find a maximum likelihood estimate of $pFDR(t) = \pi_0 t / G(t)$. It easily follows that the likelihood of the data can be written as

$$G(t)^{R(t)} \cdot [1 - G(t)]^{m-R(t)} = [\pi_0 \cdot t + (1-\pi_0) \cdot G_1(t)]^{R(t)} [1 - \pi_0 \cdot t - (1-\pi_0) \cdot G_1(t)]^{m-R(t)} .$$

Regardless of our knowledge of G_1, the maximum likelihood estimate of $G(t)$ is

$$\widehat{G}(t) = \frac{R(t)}{m}.$$

If G_1 is known, the maximum likelihood estimate of π_0 is

$$\widetilde{\pi}_0 = \frac{G_1(t) - \widehat{G}(t)}{G_1(t) - t} = \frac{G_1(t) - R(t)/m}{G_1(t) - t}.$$

Therefore, when G_1 is known, the mle of $pFDR(t)$ is

$$\widetilde{Q}(t) = \frac{\widetilde{\pi}_0 \cdot t}{\widehat{G}(t)} = \frac{[G_1(t) - R(t)/m]t}{[G_1(t) - t]\frac{R(t)}{m}}.$$

The behavior of this estimate should be good for large m since it is consistent and efficient.

Recall that in Section 4.1 we introduced the estimate

$$\widehat{Q}(t) = \frac{\widehat{\pi}_0(\lambda) \cdot t}{\widehat{G}(t)} = \frac{\frac{W(\lambda)}{m} \cdot t}{(1-\lambda) \cdot \frac{R(t)}{m}},$$

where $\widehat{\pi}_0(\lambda) = \frac{W(\lambda)}{m \cdot (1-\lambda)}$ served as our estimate of π_0. This estimate was derived because we did not assume G_1 to be known. Ideally, we would like to find mle's of both π_0 and $G_1(t)$ when $G_1(t)$ is unknown. Since $\widehat{G}(t) = R(t)/m$ regardless of the knowledge of G_1, it follows that the mle of $G_1(t)$, say $\widetilde{G_1}(t)$ and $\widetilde{\pi}_0$, would have to satisfy the equation

$$\widetilde{\pi}_0 \cdot t + (1 - \widetilde{\pi}_0) \cdot \widetilde{G_1}(t) = \frac{R(t)}{m}.$$

This is one equation with two unknowns, so it is impossible to find both mle's simultaneously. Therefore, our remedy was to find a conservative estimate of π_0. Note that when we observe the p-values $p_1, \ldots, p_m$, we can form any reject region $[0, \lambda]$. Also note that

$$\frac{1 - G(\lambda)}{1 - \lambda} = \pi_0 + \frac{1 - G_1(\lambda)}{1 - \lambda} \cdot (1 - \pi_0).$$

Without knowing G_1 we can form the mle of $1 - G(\lambda)$ as $W(\lambda)/m$. Therefore, $\widehat{\pi}_0(\lambda)$ is estimating a parameter with conservative additive bias over π_0 of size

$$\frac{1 - G_1(\lambda)}{1 - \lambda} \cdot (1 - \pi_0).$$

One could choose $\lambda = t$, however this does not have to be the case. Since $\frac{1-G_1(\lambda)}{1-\lambda}$ usually gets smaller as λ gets larger, it may be better to take a larger λ than t, because t will likely be very small.

Therefore, $\widehat{Q}_\lambda(t)$ is the maximum likelihood estimate of

$$\frac{\pi_0 + \frac{1-G_1(\lambda)}{1-\lambda} \cdot \pi_1}{\pi_0} pFDR(t),$$

a quantity slightly greater than $pFDR(t)$. We made two adjustments to $\widehat{Q}_\lambda(t)$ in order to formulate $\widehat{pFDR}_\lambda(t)$. Now $\widehat{pFDR}_\lambda(t)$ has good finite sample properties (avoiding the inconveniences of the pure mle), but it is asymptotically equivalent to $\widehat{Q}(t)$, so it has the same large sample properties. Also, the variance of $\widehat{pFDR}_\lambda(t)$ is not that different than

that of $\widetilde{Q}(t)$ for large m and powerful tests.

One criticism of the current approach to multiple hypothesis testing that we have made is that the variability of the estimated rejection region resulting from a sequential p-value method is not calculated. Even if it is calculated, it is hard to interpret what $\mathbf{Var}(\widehat{k})$ means in terms of the effectiveness of the sequential p-value method. The only property of $\widehat{k}$ that is assessed is that the expected error rate is less than or equal to α under $\widehat{k}$. Using our approach, the variance of $\widehat{pFDR}_\lambda(t)$ can easily be calculated. In a parametric situation where G_1 is known, one can calculate $\mathbf{Var}(\widehat{pFDR}_\lambda(t))$ either in the finite sample or asymptotic sense. If G_1 is unknown, the bootstrap can be employed to estimate $\mathbf{Var}(\widehat{pFDR}_\lambda(t))$.

Chapter 5

Estimating Rejection Regions for Fixed False Discovery Rates

We will now use $\widehat{FDR}_\lambda(t)$ to derive a new, more powerful class of FDR controlling procedures, of which the BH procedure is the most conservative version. The motivation for the method is the following. The point estimate $\widehat{FDR}_\lambda(t)$ shows improvements over Benjamini & Hochberg's (1995) estimate of $FDR(t)$, which was shown to implicitly be $\widehat{FDR}_{\lambda=0}(t)$ in Section 4.2. Thus, especially for $\lambda \gg 0$, it is tempting to try to use $\widehat{FDR}_\lambda(t)$ in the context of Benjamini & Hochberg's (1995) original FDR controlling procedure: if one wants to control the FDR at level α, reject all p-values in the largest rejection region $[0, t]$ such that $\widehat{FDR}_\lambda(t) \leq \alpha$. In this chapter we show when this heuristic procedure controls the FDR and when it is less conservative than the BH procedure.

5.1 A New Class of FDR Controlling Procedures

Sequential p-values methods essentially estimate the region based on the p-values and the pre-chosen level α at which to control the error rate. Thus, we define the following function that chooses the cut-point based some function G defined on $[0, 1]$:

$$t_\alpha[G] = \sup\{0 \leq t \leq 1 : G(t) \leq \alpha\}.$$

In words, $t_\alpha[G]$ finds the largest t such that $G(t) \leq \alpha$. Now if $G(t)$ is an estimate of $FDR(t)$, then $t_\alpha[G]$ is a random variable, and it finds the largest rejection region so that

this estimate is less than or equal to α. We will only consider G defined on $[0,1]$ that are right continuous functions defined with left limits, i.e., the *cadlag* functions $D([0,1],\mathbb{R})$ (cf. Billingsley 1968). It follows that $t_\alpha[G]$ over all $G \in D([0,1],\mathbb{R})$ defines a general class of p-value step-up methods, including the standard step-up p-value methods.

We show that $t_\alpha[\widehat{FDR}_\lambda]$ provides asymptotic control of the FDR under fairly general assumptions, even allowing for certain forms of dependence. For the finite sample case with independent null p-values, we propose a new FDR controlling step-up method with two minor modifications to $\widehat{FDR}_\lambda$, and we summarize it in Algorithm 5.1 on page 63. As one modification, we estimate π_0 in $\widehat{FDR}_\lambda$ with

$$\widehat{\pi}_0^*(\lambda) = \frac{W(\lambda) + 1(\lambda > 0)}{m(1-\lambda)}$$

so that it is always the case that $\widehat{\pi}_0^*(\lambda) > 0$. This avoids nonsensical estimates of π_0, since an estimate of 0 would be problematic here. We also have to limit our rejections to occurring within the region $[0,\lambda]$. Therefore, the estimate of $FDR(t)$ we use for the finite sample case is

$$\widehat{FDR}_\lambda^*(t) = \begin{cases} m \cdot \widehat{\pi}_0^*(\lambda) \cdot t/R(t) & \text{if } t \le \lambda \\ 1 & \text{if } t > \lambda \end{cases}$$

These modifications allow us to prove in the finite sample case that the $t_\alpha[\widehat{FDR}_\lambda^*]$ procedure controls the FDR (see Theorem 5.2). This last modification has little impact on the procedure as explained later in this section, and both modifications are totally unnecessary to prove asymptotic control, as is shown in Theorem 5.3 in Section 5.4.

A natural question to ask is what is the motivation for the $t_\alpha[\widehat{FDR}_\lambda]$ procedure. There are two reasons, one of which is that using the $t_\alpha[\cdot]$ cut point procedure gives a general way of formulating FDR controlling methods, perhaps tailored to specific examples, in which the estimate of FDR may depend on the specific application. Indeed, the BH procedure can be written as $t_\alpha[\widehat{FDR}_{\lambda=0}]$, which, considering our results in later sections, clarifies why the BH procedure works, something the original proof by induction fails to do. It also allows us to conclude that the BH procedure is unnecessarily conservative.

The second reason for studying rules of the form $t_\alpha[\cdot]$ has to do with the form of our estimate $\widehat{FDR}_\lambda(t)$. We show, using martingale methods, in Section 5.3 that the BH procedure controls the FDR at exactly level $\alpha \cdot m_0/m$. It also follows from this result that if we were to replace m with m_0, we would control the FDR exactly at level α, eliminating

Algorithm 5.1: Proposed FDR Controlling Procedure

1. Let α be the pre-chosen level at which to control the FDR.
2. For any fixed rejection region $[0, t]$, estimate $FDR(t)$ by
$$\widehat{FDR}^*_\lambda(t) = \left[\frac{W(\lambda) + 1(\lambda > 0)}{1 - \lambda} \cdot \frac{t}{R(t) \vee 1}\right]^{1(t \le \lambda)}$$
for small m, with the null p-values all being independent.
3. Estimate $FDR(t)$ by
$$\widehat{FDR}_\lambda(t) = \frac{W(\lambda)}{1 - \lambda} \cdot \frac{t}{R(t) \vee 1}$$
for large m that meet the conditions of Theorem 5.3.
4. For small m with independent null p-values, reject all hypotheses corresponding to $p_i \le t_\alpha[\widehat{FDR}^*_\lambda]$ for $\lambda > 0$, and $p_i \le t_\alpha[\widehat{FDR}_{\lambda=0}]$ for $\lambda = 0$.
5. For large m that meet the conditions of Theorem 5.3, reject all hypotheses corresponding to $p_i \le t_\alpha[\widehat{FDR}_\lambda]$.

See Chapter 7 for how to choose λ automatically.

the conservative bias of the BH procedure, and therefore gaining power. By replacing m with $\widehat{m}_0(\lambda) = m \cdot \widehat{\pi}_0(\lambda)$, which is less conservative estimate of m_0 than m, we make the procedure less conservative, more powerful – and more accurate in the sense that the true level of FDR control is asymptotically closer to α than the BH procedure. In particular, consider the following two lemmas.

Lemma 5.1 *The p-value step-up method $t_\alpha[\widehat{FDR}_{\lambda=0}]$ is equivalent to the BH procedure.*

Proof: Noting that $\widehat{\pi}_0(\lambda = 0) = 1$, this follows by the proof of the next proposition. □

Lemma 5.2 *In general, the p-value step-up method $t_\alpha[\widehat{FDR}_\lambda]$ is equivalent to the BH procedure with m replaced by $\widehat{m}_0(\lambda) = m \cdot \widehat{\pi}_0(\lambda)$.*

Proof: What we have to show is $p_{(\widehat{k}_\lambda)} \le t_\alpha[\widehat{FDR}_\lambda] < p_{(\widehat{k}_\lambda + 1)}$ where $\widehat{k}_\lambda$ is the $\widehat{k}$ in the BH procedure (Algorithm 1.1) with m replaced by $\widehat{m}^*_0(\lambda)$. Ignoring the cases where $R(t) = 0$, in which case both procedures reject no p-values, the inequalities follow immediately once it is noted that the BH procedure for selecting $\widehat{k}_\lambda$ is simply $\widehat{k}_\lambda = \max\{k : \widehat{FDR}_\lambda(p_{(k)}) \le \alpha\}$.

□

In practice there should not be much of a difference between $t_\alpha[\widehat{FDR}^*_\lambda]$ and $t_\alpha[\widehat{FDR}_\lambda]$ since λ will tend to be large and $t_\alpha[\widehat{FDR}_\lambda]$ will tend to be small. To be specific, consider the following scenario: under the mixture model for the p-values Theorem 2.1 implies that $\alpha = pFDR(t)$ if and only if

$$\frac{\pi_0}{1-\pi_0} \cdot t\left(\frac{1}{\alpha} - 1\right) = G_1(t),$$

where G_1 is the alternative distribution. Since $G_1(t) \leq 1$, t must satisfy

$$t \leq \frac{(1-\pi_0)\cdot\alpha}{\pi_0 \cdot (1-\alpha)}.$$

Even with a small $\pi_0 = 0.75$ and a large $\alpha = 0.20$, it follows that if $pFDR(t) \leq \alpha$, then $t \ll 0.1$. Therefore, using λ over the range $\lambda = 0, 0.1, 0.2, \ldots, 0.9$ for example, implies $t_\alpha[\widehat{FDR}^*_\lambda]$ will essentially be equivalent to $t_\alpha[\widehat{FDR}_\lambda]$. The method we propose for automatically choosing λ in Chapter 7 should also avoid choosing λ that are less than $t_\alpha[\widehat{FDR}_\lambda]$ too often.

5.2 A Numerical Study: Independence

We performed $m = 1000$ one-sided hypothesis tests of $N(0,1)$ (null) versus $N(2,1)$ (alternative). We let $m_0 = 100, \ldots, 900$, and generated 1000 sets of 1000 normal random variables for each m_0 value. The BH procedure and proposed finite sample procedure $t_\alpha[\widehat{FDR}^*_\lambda]$ were performed at levels $\alpha = 0.01$ and $\alpha = 0.05$. For simplicity, we set $\lambda = 1/2$ for both cases in our procedure. Figure 5.1 shows the average power of the proposed procedure versus BH. It can be seen that the increase in power we achieve is greater the smaller m_0 is. This makes sense because the difference in our proposed procedure over BH is that it estimates m_0.

5.3 Finite Sample Results

In this section, we prove the following two theorems using martingale methods.

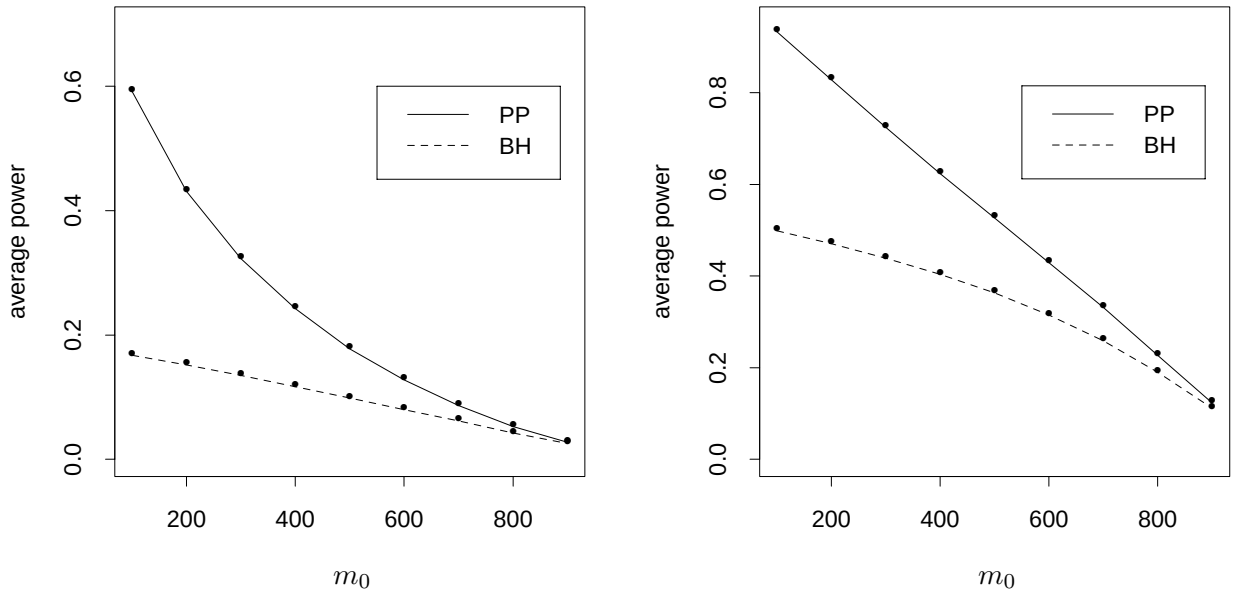


Figure 5.1: A plot of average power versus m_0 for the proposed procedure for small m (PP) and the Benjamini and Hochberg (1995) procedure (BH). The left panel is the case where the FDR is controlled at level $\alpha = 0.01$, and the right panel where $\alpha = 0.05$. It can be seen that there is an increase in power under the proposed procedure in both situations.

Theorem 5.1 *(Benjamini & Hochberg 1995) If the p-values corresponding to the true null hypotheses are independent, then*

$$FDR\left(t_\alpha[\widehat{FDR}_{\lambda=0}]\right) = \frac{m_0}{m} \cdot \alpha \leq \alpha.$$

That is, the BH procedure controls the FDR at level α. Now consider the procedure with $\lambda > 0$.

Theorem 5.2 *If the p-values corresponding to the true null hypotheses are independent, then for $\lambda > 0$,*

$$FDR\left(t_\alpha[\widehat{FDR}^*_\lambda]\right) \leq (1 - \lambda^{m_0}) \cdot \alpha \leq \alpha.$$

Remark: Note that since $t_\alpha[\widehat{FDR}^*_\lambda]$ is a random variable, we have for example that

$$FDR\left(t_\alpha[\widehat{FDR}^*_\lambda]\right) = \mathbf{E}\left[\frac{V(t_\alpha[\widehat{FDR}^*_\lambda])}{R(t_\alpha[\widehat{FDR}^*_\lambda]) \vee 1}\right].$$

We use the fact that, for our estimate $\widehat{FDR}^*_\lambda$ the random variable $t_\alpha[\widehat{FDR}^*_\lambda]$ is a stopping time with respect to a certain filtration, which allows us to use the Optional Stopping Theorem. We recall the following empirical processes, which have values in $D([0,1], \mathbb{R})$:

$$\begin{aligned} V(t) &= \#\{\text{null } p_i : p_i \leq t\} \\ S(t) &= \#\{\text{alternative } p_i : p_i \leq t\} \\ R(t) &= V(t) + S(t) = \#\{p_i : p_i \leq t\} \end{aligned}$$

We view time as running backwards in these empirical processes. The proof that $V(t)/t$ is a backwards martingale is easily shown, so we omit the proof here.

Lemma 5.3 *If the p-values of the m_0 true null hypotheses are independent, then $V(t)/t$ for $0 \leq t < 1$ is a martingale, with time running backwards, with respect to the filtration $\mathcal{F}_t = \sigma(1_{\{p_i \leq s\}}, t \leq s \leq 1, i = 1, \ldots, m)$. That is, for $s \leq t$, $\mathbf{E}[V(s)/s|\mathcal{F}_t] = V(t)/t$.*

The following elementary lemma, whose proof is also omitted, incorporates the stopping times (thresholding rules) into our martingale framework.

Lemma 5.4 *The random variable $t_\alpha[\widehat{FDR}_{\lambda=0}]$ is a stopping time with respect to $\mathcal{F}_t$ with time running backwards. Further, for $\lambda > 0$, $t_\alpha[\widehat{FDR}^*_\lambda]$ is a stopping time with respect to $\mathcal{F}^\lambda_t \triangleq \mathcal{F}_{t \wedge \lambda}$.*

We are now ready to prove Theorems 5.1 and 5.2. The proof of Theorem 5.1 is originally due to D. Siegmund and J. Taylor (personal communication).

Proof of Theorem 5.1: Noting that the process $m \cdot t/R(t)$ has only upward jumps and has a final value of 1, we see $R(t_\alpha[\widehat{FDR}_{\lambda=0}]) = t_\alpha[\widehat{FDR}_{\lambda=0}] \cdot m/\alpha$. Therefore, the ratio whose expectation we must calculate can be expressed as

$$\frac{V(t_\alpha[\widehat{FDR}_{\lambda=0}])}{R(t_\alpha[\widehat{FDR}_{\lambda=0}])} = \frac{\alpha}{m}\frac{V(t_\alpha[\widehat{FDR}_{\lambda=0}])}{t_\alpha[\widehat{FDR}_{\lambda=0}]}.$$

Noting that $V(t)/t$ stopped at $t_\alpha[\widehat{FDR}_{\lambda=0}]$ is bounded by m/α, the Optional Stopping Theorem then implies

$$FDR(t_\alpha[\widehat{FDR}_{\lambda=0}]) = \frac{\alpha}{m}\mathbf{E}\left[\frac{V(t_\alpha[\widehat{FDR}_{\lambda=0}])}{t_\alpha[\widehat{FDR}_{\lambda=0}]}\right] = \frac{\alpha}{m}\mathbf{E}\left[V(1)\right] = \frac{m_0}{m}\cdot\alpha.$$

□

Proof of Theorem 5.2: Abbreviate $t_\alpha[\widehat{FDR}^*_\lambda]$ by t^λ_α. If $\widehat{FDR}^*_\lambda(\lambda) \geq \alpha$ then it can be seen that $R(t^\lambda_\alpha) = t^\lambda_\alpha \cdot m\widehat{\pi}^*_0(\lambda)/\alpha$ similarly to above. Moreover, when $\widehat{FDR}^*_\lambda(\lambda) \geq \alpha$, then $V(t)/t$ stopped at t^λ_α is bounded by $m(m-1)/[(1-\lambda)\alpha]$. Thus,

$$FDR(t^\lambda_\alpha) = \mathbf{E}\left[\frac{V(t^\lambda_\alpha)}{R(t^\lambda_\alpha)}; \widehat{FDR}^*_\lambda(\lambda) \geq \alpha\right] + \mathbf{E}\left[\frac{V(t^\lambda_\alpha)}{R(t^\lambda_\alpha)}; \widehat{FDR}^*_\lambda(\lambda) < \alpha\right].$$

Now

$$\begin{aligned}
\mathbf{E}\left[\frac{V(t^\lambda_\alpha)}{R(t^\lambda_\alpha)}; \widehat{FDR}^*_\lambda(\lambda) \geq \alpha\right] &= \mathbf{E}\left[\alpha\frac{1-\lambda}{W(\lambda)+1}\cdot\frac{V(t^\lambda_\alpha)}{t^\lambda_\alpha}; \widehat{FDR}^*_\lambda(\lambda) \geq \alpha\right] \\
&= \mathbf{E}\left[\alpha\frac{1-\lambda}{W(\lambda)+1}\cdot\mathbf{E}\left[\frac{V(t^\lambda_\alpha)}{t^\lambda_\alpha}\middle|\mathcal{F}_\lambda\right]; \widehat{FDR}^*_\lambda(\lambda) \geq \alpha\right] \\
&= \mathbf{E}\left[\alpha\frac{1-\lambda}{W(\lambda)+1}\cdot\frac{V(\lambda)}{\lambda}; \widehat{FDR}^*_\lambda(\lambda) \geq \alpha\right],
\end{aligned}$$

where the last step follows by the Optional Stopping Theorem. It also easily follows that

$$\mathbf{E}\left[\frac{V(t^\lambda_\alpha)}{R(t^\lambda_\alpha)}; \widehat{FDR}^*_\lambda(\lambda) < \alpha\right] \leq \mathbf{E}\left[\alpha\frac{1-\lambda}{W(\lambda)+1}\cdot\frac{V(\lambda)}{\lambda}; \widehat{FDR}^*_\lambda(\lambda) < \alpha\right].$$

The upper bound follows by

$$\begin{aligned} FDR(t_\alpha^\lambda) &\leq \mathbf{E}\left[\frac{1-\lambda}{W(\lambda)+1}\cdot\frac{V(\lambda)}{\lambda}\alpha\right] \\ &\leq \mathbf{E}\left[\frac{1-\lambda}{m_0-V(\lambda)+1}\cdot\frac{V(\lambda)}{\lambda}\alpha\right] \\ &= (1-\lambda^{m_0})\cdot\alpha \leq \alpha. \end{aligned}$$

□

5.4 Large Sample Results

In this section, we use Theorem 4.4 to show when the proposed large m FDR controlling procedure asymptotically controls the FDR. Note that here we use the thresholding rule $t_\alpha[\widehat{FDR}_\lambda]$ instead of $t_\alpha[\widehat{FDR}^*_\lambda]$. Recall that the assumptions of Theorem 4.4 are that $V(\cdot)/m_0$ and $S(\cdot)/m_1$ converge almost surely point-wise to some G_0 and G_1, respectively, and that $\lim_{m\to\infty} m_0/m = \pi_0$ exists. Also, for $t \in [0,1]$ we define

$$\widehat{FDR}^\infty_\lambda(t) \triangleq \frac{\left\{\pi_0 + \frac{1-G_1(\lambda)}{1-\lambda}\pi_1\right\}G_0(t)}{\pi_0 G_0(t) + \pi_1 G_1(t)}$$

where $G_0(t) = t$.

Theorem 5.3 *Assume the conditions of Theorem 4.4 hold, and assume there exists a $t > 0$ such that $\widehat{FDR}^\infty_\lambda(t) < \alpha$. Then*

$$\limsup_{m\to\infty} FDR\left(t_\alpha[\widehat{FDR}_\lambda]\right) \leq \alpha.$$

Proof: Let t' be the $t' > 0$ such that $\alpha - \widehat{FDR}^\infty_\lambda(t') = \epsilon > 0$. Therefore, we can take m large enough so that $|\widehat{FDR}^\infty_\lambda(t') - \widehat{FDR}_\lambda(t')| < \epsilon/2$ which implies $\widehat{FDR}_\lambda(t') < \alpha$ and $t_\alpha[\widehat{FDR}_\lambda] \geq t'$. Therefore, $\liminf_{m\to\infty} t_\alpha[\widehat{FDR}_\lambda] \geq t'$ with probability 1. The "lim" in Corollary 4.1 can be replaced with "liminf" since these limits exist. Therefore, by this corollary we have that probability 1

$$\liminf_{m\to\infty}\left[\widehat{FDR}_\lambda(t_\alpha[\widehat{FDR}_\lambda]) - \frac{V(t_\alpha[\widehat{FDR}_\lambda])}{R(t_\alpha[\widehat{FDR}_\lambda])\vee 1}\right] \geq \liminf_{m\to\infty}\inf_{t\geq\delta}\left[\widehat{FDR}_\lambda(t) - \frac{V(t)}{R(t)\vee 1}\right] \geq 0$$

for $\delta = t'/2$. Since $\widehat{FDR}_\lambda(t_\alpha[\widehat{FDR}_\lambda]) = \alpha$ it follows

$$\limsup_{m\to\infty} \frac{V(t_\alpha[\widehat{FDR}_\lambda])}{R(t_\alpha[\widehat{FDR}_\lambda]) \vee 1} \leq \alpha$$

with probability 1. By Fatou's Lemma,

$$\limsup_{m\to\infty} \mathbf{E}\left[\frac{V(t_\alpha[\widehat{FDR}_\lambda])}{R(t_\alpha[\widehat{FDR}_\lambda]) \vee 1}\right] \leq \mathbf{E}\left[\limsup_{m\to\infty} \frac{V(t_\alpha[\widehat{FDR}_\lambda])}{R(t_\alpha[\widehat{FDR}_\lambda]) \vee 1}\right] \leq \alpha$$

□

We can generalize the main result (Theorem 1) of Genovese & Wasserman (2001) using a different approach than they take. We show that if $\widehat{FDR}_\lambda$ converges almost surely pointwise to some limit $\widehat{FDR}_\lambda^\infty$, then the (random) thresholding rule $t_\alpha[\widehat{FDR}_\lambda]$ converges to the deterministic rule $t_\alpha[\widehat{FDR}_\lambda^\infty]$.

Theorem 5.4 *Suppose the conditions of Theorem 4.4 hold. Then*

$$\lim_{m\to\infty} t_\alpha\left[\widehat{FDR}_\lambda\right] \stackrel{a.s.}{=} t_\alpha\left[\widehat{FDR}_\lambda^\infty\right]$$

if $\widehat{FDR}_\lambda^\infty(\cdot)$ *has a non-zero derivative at* $t_\alpha[\widehat{FDR}_\lambda^\infty] > 0$.

Proof: For $t' > t_\alpha[\widehat{FDR}_\lambda^\infty]$, we have that $\widehat{FDR}_\lambda^\infty(t') - \widehat{FDR}_\lambda^\infty(t_\alpha[\widehat{FDR}_\lambda^\infty]) = \epsilon$ for some $\epsilon > 0$. Thus, we can take m large enough so that $|\widehat{FDR}_\lambda^\infty(t') - \widehat{FDR}_\lambda(t')| < \epsilon/2$, and thus $\widehat{FDR}_\lambda(t') > \alpha$ eventually with probability 1. Hence $\limsup_{m\to\infty} t_\alpha\left[\widehat{FDR}_\lambda\right] \leq t_\alpha\left[\widehat{FDR}_\lambda^\infty\right]$ a.s. If $\widehat{FDR}_\lambda^\infty(\cdot)$ has a non-zero derivative at $t_\alpha[\widehat{FDR}_\lambda^\infty]$, then it has to be a positive derivative. Otherwise, the definition of $t_\alpha[\widehat{FDR}_\lambda^\infty]$ would be violated. Thus, there exists a neighborhood, say of size $\delta > 0$, such that for $t' \in \left[t_\alpha[\widehat{FDR}_\lambda^\infty] - \delta, t_\alpha[\widehat{FDR}_\lambda^\infty]\right)$, we have $\widehat{FDR}_\lambda^\infty(t') < \widehat{FDR}_\lambda^\infty(t_\alpha[\widehat{FDR}_\lambda^\infty])$. By a similar argument to the previous case, we have that for any t' in this neighborhood, $\widehat{FDR}_\lambda(t') < \alpha$ eventually with probability 1. Thus $\liminf_{m\to\infty} t_\alpha\left[\widehat{FDR}_\lambda\right] \geq t_\alpha\left[\widehat{FDR}_\lambda^\infty\right]$ a.s. Putting these together, we have the result. □

We now show the conservative consistency of $t_\alpha[\widehat{FDR}_\lambda]$, but we first define:

$$FDR^\infty(t) \stackrel{\Delta}{=} \frac{\pi_0 \cdot G_0(t)}{\pi_0 \cdot G_0(t) + \pi_1 \cdot G_1(t)},$$

where $G_0(t) = t$.

Corollary 5.1 *Under analogous conditions of Theorem 5.4 we have that* $\lim_{m\to\infty} t_\alpha\left[FDR\right] = t_\alpha\left[FDR^\infty\right]$ *and* $\lim_{m\to\infty}\left(t_\alpha[\widehat{FDR}_\lambda] - t_\alpha[FDR]\right) \overset{a.s.}{\leq} 0$.

This final result shows that $t_\alpha[\widehat{FDR}_\lambda]$ is asymptotically less conservative than the BH procedure and is therefore more powerful.

Corollary 5.2 *Suppose that the assumptions of Theorem 5.4 hold for* $\widehat{FDR}_0^\infty, \widehat{FDR}_\lambda^\infty$, *and* FDR^∞. *Also suppose that* $G_1(t) > G_0(t)$ *for* $0 < t < 1$. *Then for any* $0 < \lambda < 1$

$$\lim_{m\to\infty} t_\alpha[\widehat{FDR}_0] < \lim_{m\to\infty} t_\alpha[\widehat{FDR}_\lambda] < \lim_{m\to\infty} t_\alpha[FDR].$$

The proofs of these last two corollaries are straightforward, so we omit them. Note that from the discussion in Section 4.5, it is conceivable that under the assumptions of this last corollary the least conservative thresholding procedure is a limit of the family of $t_\alpha[\widehat{FDR}_\lambda]$.

5.5 A Numerical Study: Dependence

In this section we present a simple numerical study to illustrate the asymptotic control of the FDR, even under certain forms of dependence. We generated $m = 3000$ statistics where $T_1, \ldots, T_{600} \sim N(2,1)$ and $T_{601}, \ldots, T_{3000} \sim N(0,1)$. The null distribution is $N(0,1)$ with $m_0 = 2400$ and the alternative distribution is $N(2,1)$ with $m_1 = 600$. The statistics have correlation ± 0.4 in groups of 10. Specifically, $\mathbf{Cov}(T_{i+j}, T_{i+k}) = \pm 0.4$ for $1 \leq j,k \leq 10$ and $i = 0, 10, \ldots, 2990$. (Otherwise, the statistics are independent.) If $1 \leq j,k \leq 5$ or $6 \leq j,k \leq 10$, then the correlation is positive; otherwise it is negative. Note that these statistics *do not* meet the "positive regression dependence" condition of Benjamini & Yekutieli (2001).

1000 data sets were generated, and for each one $t_\alpha[\widehat{FDR}_\lambda]$ was calculated for and $\alpha = 0.005, 0.01, 0.05, 0.10, 0.20$. We used $\lambda = 0$ (BH), $\lambda = 0.5$ (PP), and the BH algorithm with m replaced by m_0 (Optimal). The true FDR and the average power were calculated for all of these procedures. It can be seen from Figure 5.2 that all three procedures control the FDR at level α as the theory predicts. Moreover, $t_\alpha[\widehat{FDR}_{\lambda=0.5}]$ attains the control and power near that of the optimal procedure, showing the improvement of our proposed methodology over Benjamini & Hochberg's (1995) methodology. The numbers from this study are listed in Table 5.1.

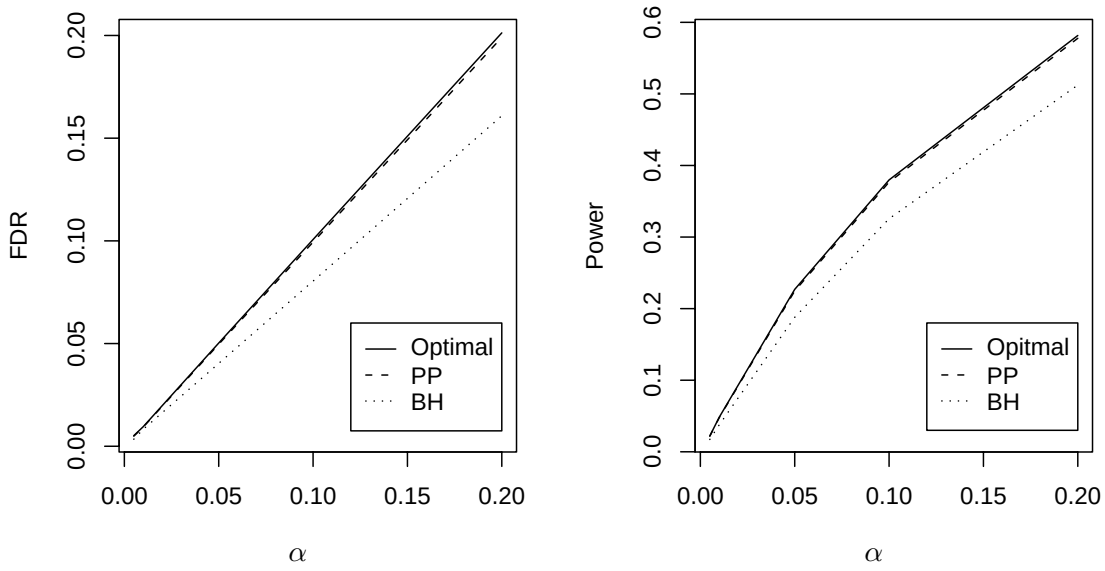


Figure 5.2: The left panel is the FDR when performing the BH, PP, and Optimal procedures at level α. The right panel is the average power attained under these three procedures. It can be seen that $t_\alpha[\widehat{FDR}_{\lambda=0.5}]$ (PP) attains the control and power near that of the optimal procedure whereas the conservative $t_\alpha[\widehat{FDR}_{\lambda=0}]$ (BH) does not.

Table 5.1: A numerical study of $t_\alpha[\widehat{FDR}_{\lambda=0}]$ (BH), $t_\alpha[\widehat{FDR}_{\lambda=0.5}]$ (PP), and the optimal procedure under dependence. The Monte Carlo standard error is listed below each number.

α	FDR			Average Power		
	BH	PP	Optimal	BH	PP	Optimal
0.005	0.00343 (5×10^{-4})	0.00492 (6×10^{-4})	0.00516 (7×10^{-4})	0.0172 (4×10^{-4})	0.0218 (4×10^{-4})	0.0221 (4×10^{-4})
0.01	0.00828 (7×10^{-4})	0.00934 (6×10^{-4})	0.00952 (6×10^{-4})	0.0376 (6×10^{-4})	0.0477 (6×10^{-4})	0.0483 (6×10^{-4})
0.05	0.0403 (6×10^{-4})	0.0497 (6×10^{-4})	0.0503 (6×10^{-4})	0.188 (9×10^{-4})	0.225 (9×10^{-4})	0.227 (9×10^{-4})
0.10	0.0804 (7×10^{-4})	0.0994 (7×10^{-4})	0.101 (7×10^{-4})	0.326 (9×10^{-4})	0.377 (9×10^{-4})	0.380 (9×10^{-4})
0.20	0.161 (8×10^{-4})	0.199 (8×10^{-4})	0.201 (8×10^{-4})	0.512 (8×10^{-4})	0.578 (9×10^{-4})	0.582 (8×10^{-4})

Chapter 6

Estimating the q-values and Simultaneous Controlling Curve

In this chapter we introduce estimates of the q-values and the simultaneous FDR controlling curve. We show that these estimates are simultaneously conservatively consistent. Recall from Chapter 4 that we estimate the pFDR and FDR over the rejection region $[0, t]$ by

$$\begin{aligned} \widehat{pFDR}_\lambda(t) &= \frac{W(\lambda) \cdot t}{[1-\lambda] \cdot [R(t) \vee 1] \cdot [1-(1-t)^m]}, \\ \widehat{FDR}_\lambda(t) &= \frac{W(\lambda) \cdot t}{[1-\lambda] \cdot [R(t) \vee 1]}, \end{aligned}$$

where $W(\lambda) = \#\{p_i : p_i > \lambda\}$ and $R(t) = \#\{p_i : p_i \leq t\}$. Also recall from Chapter 3 that for a given p-value p_i, q-value$(p_i) = \inf_{t \geq p_i} pFDR(t)$ is the q-value for p_i, and $\alpha_{FDR}(p_i) = \inf_{t \geq p_i} FDR(t)$ is the simultaneous FDR controlling curve evaluated at p_i.

6.1 The Nonparametric Estimates

Since q-value$(t) = \inf_{s \geq t} pFDR(s)$, we simply estimate it by

$$\widehat{\mathrm{q}}_\lambda(t) = \inf_{s \geq t} \widehat{pFDR}_\lambda(s).$$

We propose Algorithm 6.1 for non-parametrically estimating the q-value(p_i). (If a model is assumed, then the results from Chapter 3 can be used to make a more precise estimate.) This procedure ensures that $\widehat{\mathrm{q}}(p_{(1)}) \leq \ldots \leq \widehat{\mathrm{q}}(p_{(m)})$, which is necessary according to our

---

Algorithm 6.1: Estimating the q-values

1. For the m hypothesis tests, calculate the p-values $p_1, \ldots, p_m$.
2. Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered p-values.
3. Set $\widehat{\mathrm{q}}_\lambda(p_{(m)}) = \widehat{pFDR}_\lambda(p_{(m)})$.
4. Set $\widehat{\mathrm{q}}_\lambda(p_{(i)}) = \min\left(\widehat{pFDR}_\lambda(p_{(i)}), \widehat{\mathrm{q}}_\lambda(p_{(i+1)})\right)$ for $i = m-1, m-2, \ldots, 1$.

---

definition. The q-values can be used in practice in the following way: it gives us the minimum pFDR we can achieve for rejection regions containing $[0, p_{(i)}]$ for $i = 1, \ldots, m$. In other words, for each p-value there is a rejection region with pFDR equal to q-value($p_{(i)}$) so that at least $p_{(1)}, \ldots, p_{(i)}$ are rejected.

In order to estimate the simultaneous FDR controlling curve, we use the exact same approach as in estimating the q-value. Therefore, for each $t \in [0, 1]$, we estimate $\alpha_{FDR}(t)$ by

$$\widehat{\alpha}_{FDR,\lambda}(t) = \inf_{s \geq t} \widehat{FDR}_\lambda(s).$$

Algorithm 6.2 is proposed to estimate $\widehat{\alpha}_{FDR,\lambda}$ at the observed p-values.

Algorithm 6.2: Estimating α_{FDR} at the Observed p-values

1. For the m hypothesis tests, calculate the p-values $p_1, \ldots, p_m$.
2. Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered p-values.
3. Set $\widehat{\alpha}_{FDR,\lambda}(p_{(m)}) = \widehat{FDR}_\lambda(p_{(m)})$.
4. Set $\widehat{\alpha}_{FDR,\lambda}(p_{(i)}) = \min\left(\widehat{FDR}_\lambda(p_{(i)}), \widehat{\alpha}_{FDR,\lambda}(p_{(i+1)})\right)$ for $i = m-1, m-2, \ldots, 1$.

We show in Chapter 7 how to automatically choose λ for both of these sets of estimates.

6.2 The Advantages of $\widehat{pFDR}_\lambda$ and $\widehat{\mathbf{q}}_\lambda$ Over $\widehat{FDR}_\lambda$

We now take a closer look at the differences between $\widehat{pFDR}_\lambda(t)$ and $\widehat{FDR}_\lambda(t)$, and why it makes sense to use the pFDR and the q-value. Consider the following fact for fixed m:

$$\lim_{t\to 0} \widehat{pFDR}_\lambda(t) = \widehat{\pi}_0(\lambda).$$

In other words, as we make the rejection region smaller and smaller, we eventually estimate the pFDR as $\widehat{\pi}_0(\lambda)$. This is the conservative thing to do since all we can conclude is that $\lim_{t\to 0} pFDR(t) \leq \pi_0$. Also, under no parametric assumptions, this is exactly what we would want. For example, suppose we take a very small rejection region. Then it is most likely that only one p-value at most would fall into that region. Without information from other p-values, and without parametric information about the alternative distribution, there is little we can say about whether this one p-value is null or alternative. Therefore, it makes sense to estimate the pFDR by the prior probability $\widehat{\pi}_0(\lambda)$ in extremely small rejection regions.

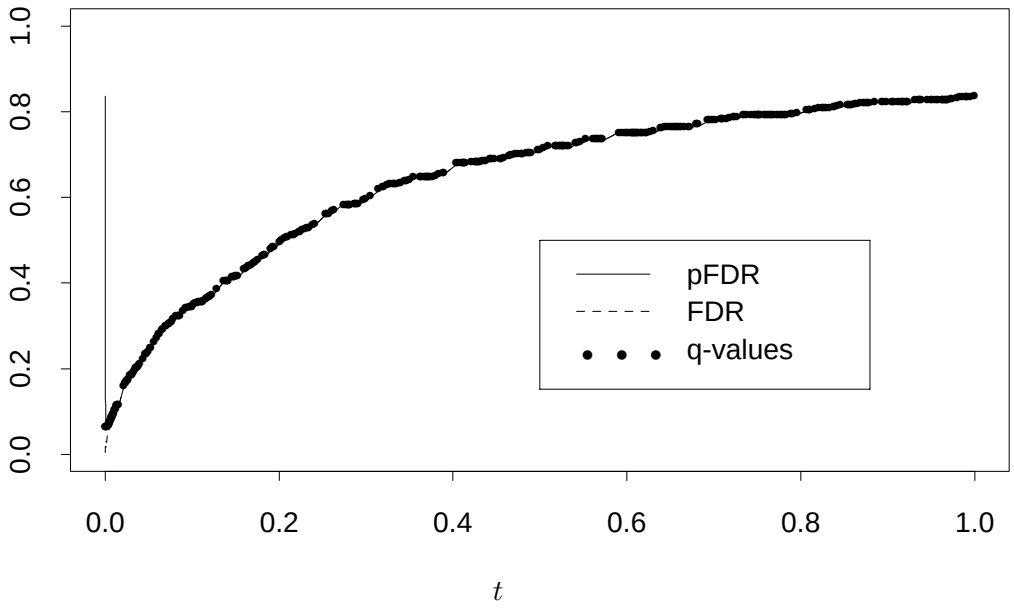
Note on the other hand that

$$\lim_{t\to 0} \widehat{FDR}_\lambda(t) = 0.$$

Does this makes sense? It does in that $\lim_{t\to 0} FDR(t) = 0$. But the only reason why we always get this convergence is because of the extra term $\mathbf{Pr}(R(t) > 0)$ in the FDR. Therefore, as t gets small, the FDR is driven by the fact that the rejection region is small rather than the fact that the "rate that discoveries are false" is small. After all, as we said above, there is not enough information about the alternative distribution in these small intervals to know how likely it would be that a p-value is null or alternative.

Therefore, if we were to report $\widehat{\alpha}_{FDR,\lambda}$, then for small p-values it would be driven to zero just because the p-value is small, even though we know little about how likely it came from the alternative without serious assumptions. Consider Figure 6.1. We performed 1000 hypothesis tests of $N(0,1)$ versus $N(2,1)$. 800 came from the null $N(0,1)$ distribution and 200 came from the alternative $N(2,1)$ distribution. The top panel shows $\widehat{pFDR}_\lambda(t)$ and $\widehat{FDR}_\lambda(t)$ as a function of t, as well as the q-value as a function of the observed p-values. It can be seen that all three functions look similar except for close to the origin.

The bottom panel zooms in near zero, where we see that $\widehat{pFDR}_\lambda(t)$ shoots up to $\widehat{\pi}_0(\lambda)$,



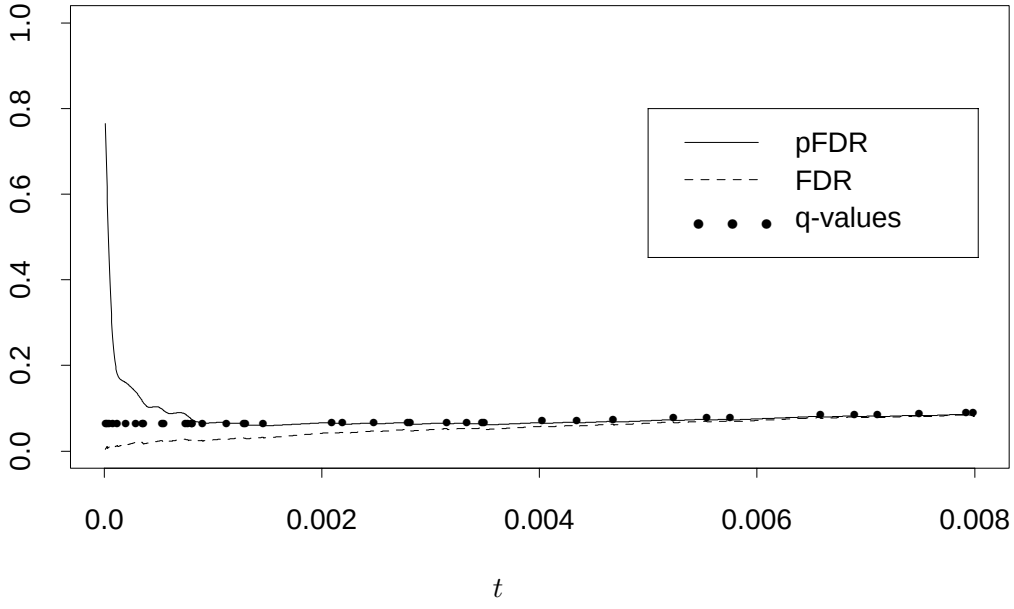


Figure 6.1: A plot of $\widehat{pFDR}(t)$, $\widehat{FDR}(t)$, and $\widehat{\mathrm{q}}$ for the $N(0,1)$ versus $N(2,1)$ example. It can be seen that $\widehat{pFDR}(t)$ and $\widehat{\mathrm{q}}$ behave more reasonably than $\widehat{FDR}(t)$ near the origin.

and $\widehat{FDR}_\lambda(t)$ shoots down to 0. The q-value, however, sits steady where $\widehat{pFDR}_\lambda(t)$ reaches its minimum (at about $p_{(10)}$). In other words, the q-value calibrates where we start getting enough information to make good statements about the pFDR. The FDR totally misses the point near the origin and merely measures the fact that we are near the origin. Moreover, the origin is arguably the most important region since this is where the most significant p-values lie. Therefore, by using $\widehat{pFDR}_\lambda(t)$ and the q-value, we obtain *robust* estimates of the pFDR, which we argue is the more appropriate error measure. The q-value bypasses our having fixed the rejection regions, and makes the rejection regions random in the appropriate way. It also bypasses any need to fix the error rate beforehand, as has to be done in the traditional framework.

6.3 Large Sample Results

It is tempting to conclude that $\mathbf{E}[\widehat{\mathrm{q}}_\lambda(p_i)] \geq \mathbf{E}[q(p_i)]$ since we have $\inf_{t \geq p_i} \mathbf{E}[\widehat{pFDR}_\lambda(t)] \geq \inf_{t \geq p_i} pFDR(t)$. However, the "inf" must be placed inside the expectation on the left hand side of the inequality, so the desired result is not trivial. Using the ideas from Section 5.4, we can prove that $\widehat{\mathrm{q}}_\lambda$ is asymptotically conservative. Moreover, we can show that it is *simultaneously* so in the following theorem.

Theorem 6.1 *Suppose that with probability 1, $V(t)/m_0$ converges to some function $G_0(t)$ and $S(t)/m_1$ converges to some function $G_1(t)$ for each $t \in [0,1]$. Also suppose that $\lim_{m\to\infty} m_0/m = \pi_0$ exists. Then for each $\delta > 0$*

$$\lim_{m\to\infty} \inf_{t \geq \delta} [\widehat{q}_\lambda(t) - q(t)] \geq 0 \ \textit{a.s.}$$

The convergence in the condition of Theorem 6.1 occurs under independence, as well as under the forms of "weak dependence" that we have discussed throughout this work. Not surprisingly, the simultaneous conservative consistency of $\widehat{\alpha}_{FDR,\lambda}$ also holds under the same conditions as Theorem 6.1.

Theorem 6.2 *Suppose that with probability 1, $V(t)/m_0$ converges to some function $G_0(t)$ and $S(t)/m_1$ converges to some function $G_1(t)$ for each $t \in [0,1]$. Also suppose that $\lim_{m\to\infty} m_0/m = \pi_0$ exists. Then for each $\delta > 0$*

$$\lim_{m\to\infty} \inf_{t \geq \delta} [\widehat{\alpha}_{FDR,\lambda}(t) - \alpha_{FDR}(t)] \geq 0 \ \textit{a.s.}$$

Proof of Theorems 6.1 and 6.2: It follows from Theorem 4.4 that with probability 1

$$\lim_{m\to\infty} \inf_{t\geq\delta} [\widehat{pFDR}_\lambda(t) - pFDR(t)] \geq 0$$

for each $\delta > 0$. Therefore, for each $t > 0$, it is straightforward to show that with probability 1

$$\lim_{m\to\infty} \left[\inf_{s\geq t} \widehat{pFDR}_\lambda(s) - \inf_{s\geq t} pFDR(s)\right] \geq 0.$$

This is just another way of writing that with probability 1

$$\lim_{m\to\infty} [\widehat{\mathrm{q}}_\lambda(t) - \mathrm{q}(t)] \geq 0$$

for each $t > 0$. Since $\widehat{\mathrm{q}}_\lambda(\cdot) \in D([0,1],\mathbb{R})$ and $\widehat{\mathrm{q}}_\lambda(t)$ is an increasing function of t, it follows that with probability 1

$$\lim_{m\to\infty} \inf_{t\geq\delta} [\widehat{\mathrm{q}}_\lambda(t) \wedge 1 - \mathrm{q}(t)] \geq 0.$$

The proof of Theorem 6.2 is completely analogous, so we omit the details. □

Remark: Due to the robustness of $\widehat{\mathrm{q}}_\lambda(t)$ near zero that we discussed, it is feasible that $\lim_{m\to\infty} [\widehat{\mathrm{q}}_\lambda(0) - \mathrm{q}(0)] \geq 0$ with probability 1. In this case Theorem 6.1 holds with $\delta = 0$. This is left as an open problem.

6.4 A Numerical Example

We performed 3000 hypothesis tests of the above distributions with $m_0 = 2400$. $\widehat{\mathrm{q}}_\lambda(p_i)$ was calculated at each p-value p_i as well as the true $\mathrm{q}(p_i)$ with $\lambda = 0.5$. These calculations are displayed in Figure 6.2. It can be seen that $\widehat{\mathrm{q}}_\lambda(\cdot) \geq \mathrm{q}(\cdot)$ over all p-values simultaneously, which is exactly the result of Theorem 6.1. The simulation results for $\widehat{\alpha}_{FDR}$ are equally good, but we omit them here.

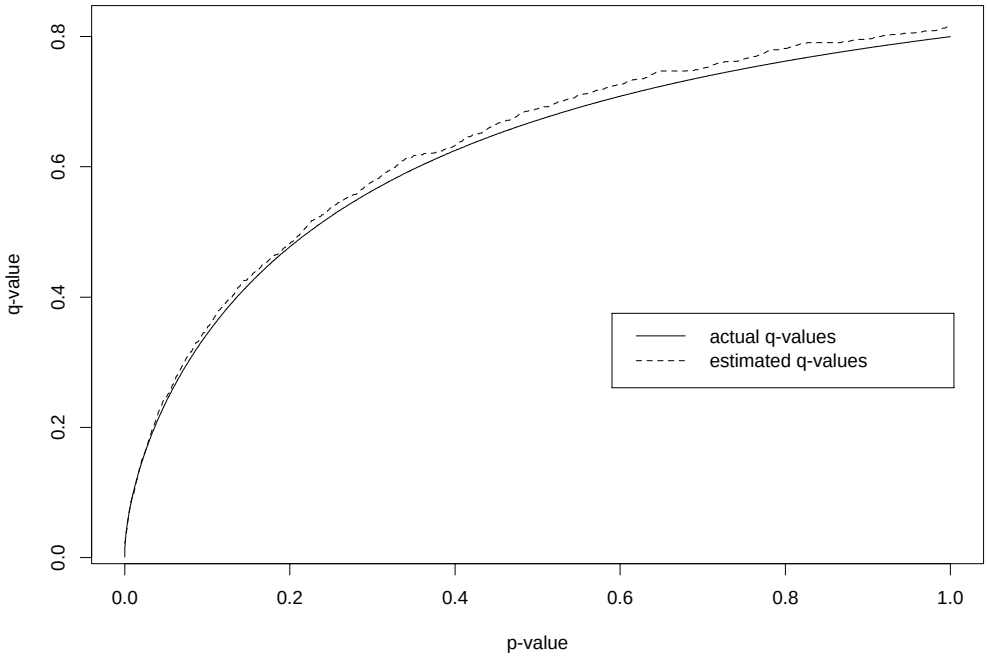


Figure 6.2: A plot of $\widehat{\mathrm{q}}(\cdot)$ and $q(\cdot)$ evaluated at each p-value for 3000 tests of $N(0,1)$ versus $N(2,1)$ with $m_0 = 2400$.

Chapter 7

Automatically Choosing λ

Throughout this work we have used the estimates $\widehat{FDR}_\lambda$ and $\widehat{pFDR}_\lambda$ to tackle the three scenarios discussed in Section 1.4. These estimates involve the tuning parameter λ used in estimating π_0, the proportion of true null hypotheses. In each of these scenarios, we showed the methodologies provide the correct conservative property for a fixed λ. Now we consider choosing λ adaptively to simultaneously minimize the bias and variance of the procedures.

Recall that

$$\widehat{\pi}_0(\lambda) = \frac{W(\lambda)}{m(1-\lambda)} = \frac{\#\{p_i : p_i > \lambda\}}{m(1-\lambda)}.$$

Suppose $G_1(\lambda)$ is the power over $[0, \lambda]$ averaged over all alternative hypotheses. Then

$$\mathbf{E}[\widehat{\pi}_0(\lambda)] = \pi_0 + \frac{1 - G_1(\lambda)}{1-\lambda}\pi_1 \geq \pi_0,$$

so clearly $\widehat{\pi}_0(\lambda)$ conservatively biased, as we would want. Moreover, if the p-values are independent, and each alternative p-value follows the distribution G_1, then

$$\mathbf{Var}[\widehat{\pi}_0(\lambda)] = \frac{\pi_0 \cdot \lambda}{m(1-\lambda)} + \frac{\pi_1 \cdot [1 - G_1(\lambda)]G_1(\lambda)}{m(1-\lambda)^2}.$$

Now it easily seen that $\lim_{\lambda \to 0} \mathbf{Var}[\widehat{\pi}_0(\lambda)] = 0$ and $\lim_{\lambda \to 1} \mathbf{Var}[\widehat{\pi}_0(\lambda)] = \infty$. The bias of $\widehat{\pi}_0(\lambda)$ is also greatest at $\lambda = 0$; if $G_1(\lambda)/\lambda$ decreases in λ (which is a desirable property when deriving rejection regions) then the bias of $\widehat{\pi}_0(\lambda)$ decreases in λ. Therefore, there is clearly a bias-variance trade-off in the choice of λ.

There are four ways in which one can choose λ to minimize a mean-squared error (which is a common measure to balance a bias-variance trade-off situation). If one is interested in

only performing the methodology described in each of Chapters 4, 5, and 6, then λ can be chosen to minimize these respective mean-squared errors. For example, we can choose λ to minimize one of:

$$\begin{aligned}
&\mathbf{E}\left[\left(\widehat{pFDR}_\lambda(t) - pFDR(t)\right)^2\right] \quad \text{(Chapter 4)} \\
&\mathbf{E}\left[\left(t_\alpha[\widehat{FDR}_\lambda] - t_\alpha[FDR]\right)^2\right] \quad \text{(Chapter 5)} \\
&\sum_{p_i} \mathbf{E}\left[\{\widehat{\mathrm{q}}_\lambda(p_i) - \mathrm{q}(p_i)\}^2\right] \quad \text{(Chapter 6),}
\end{aligned}$$

depending on the method of interest. Or if one is interested in some combination of the three methodologies, we can simply find λ to minimize

$$\mathbf{E}\left[(\widehat{\pi}_0(\lambda) - \pi_0)^2\right].$$

We present methods to approximate the λ that minimizes each of these quantities. Throughout this chapter, we assume the p-values are independent. For the type of dependence encountered in DNA microarrays (repeated experiments with dependent measurements within each), we show how to automatically pick λ in Section 8.7.

7.1 Fixed Rejection Regions

We now present methodology for the fixed rejection region scenario and limit our discussion to $\widehat{pFDR}_\lambda(t)$, although the same procedure works for $\widehat{FDR}_\lambda(t)$. We provide an automatic way to estimate:

$$\lambda_{best} = \arg\min_{\lambda \in [0,1]} \mathbf{E}\left[\left(\widehat{pFDR}_\lambda(t) - pFDR(t)\right)^2\right].$$

We use a bootstrap method in order to estimate λ_{best}, and calculate an estimate of $MSE(\lambda) = \mathbf{E}[(\widehat{pFDR}_\lambda(t) - pFDR(t))^2]$ over a range of λ. (Call this range $\mathcal{R}$; for example, we may take $\mathcal{R} = \{0, 0.05, 0.10, \ldots, 0.95\}$.) As mentioned in Section 4.1, we can produce bootstrap versions $\widehat{pFDR}^{*b}_\lambda(t)$ (for $b = 1, \ldots B$) of the estimate $\widehat{pFDR}_\lambda(t)$ for any fixed λ. Ideally we would like to know $pFDR(t)$, and then the bootstrap estimate of the $MSE(\lambda)$ would be

$$\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{pFDR}^{*b}_\lambda(t) - pFDR(t)\right)^2.$$

However, we do not know $pFDR(t)$, so we have to form a plug-in estimate of this quantity (Efron & Tibshirani 1993). Notice that for any λ we have:

$$\mathbf{E}[\widehat{pFDR}_\lambda(t)] \geq \min_{\lambda'} \mathbf{E}[\widehat{pFDR}_{\lambda'}(t)] \geq pFDR(t),$$

as was shown in Section 4.4. Therefore, our plug-in estimate of $pFDR(t)$ is $\min_{\lambda' \in \mathcal{R}} \widehat{pFDR}_{\lambda'}(t)$. The estimate of $MSE(\lambda)$ is then

$$\widehat{MSE}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{pFDR}_\lambda^{*b}(t) - \min_{\lambda' \in \mathcal{R}} \widehat{pFDR}_{\lambda'}(t) \right)^2.$$

This method can easily be incorporated in the main method described in Section 4.1 in a computationally efficient way. Our proposed method for choosing λ is formally detailed in Algorithm 7.1.

Algorithm 7.1: Estimation and Inference of $pFDR(t)$ and $FDR(t)$ with Optimal λ

1. For some range of λ, say $\mathcal{R} = \{0, 0.05, 0.10, \ldots, 0.95\}$, calculate $\widehat{pFDR}_\lambda(t)$ as in Section 4.1.
2. For each $\lambda \in \mathcal{R}$, form B bootstrap versions $\widehat{pFDR}_\lambda^{*b}(t)$ of the estimate, $b = 1, \ldots, B$.
3. For each $\lambda \in \mathcal{R}$, estimate its respective mean squared error as:

$$\widehat{MSE}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{pFDR}_\lambda^{*b}(t) - \min_{\lambda' \in \mathcal{R}} [\widehat{pFDR}_{\lambda'}(t)] \right)^2.$$

4. Set $\widehat{\lambda} = \arg\min_{\lambda \in \mathcal{R}} \widehat{MSE}(\lambda)$. Our overall estimate of $pFDR(t)$ is

$$\widehat{pFDR}(t) = \widehat{pFDR}_{\widehat{\lambda}}(t).$$

5. Form a $1 - \alpha$ upper confidence interval of $\widehat{pFDR}(t)$ by taking the $1 - \alpha$ quantile of $\{\widehat{pFDR}_{\widehat{\lambda}}^{*1}(t), \ldots, \widehat{pFDR}_{\widehat{\lambda}}^{*B}(t)\}$ as the upper endpoint (the lower endpoint being 0).
6. In estimating the FDR, perform this same procedure with $\widehat{FDR}(t)$ instead.

We provide some numerical results under the following set up. We tested m hypotheses

Table 7.1: Simulation results for the bootstrap procedure to pick the optimal λ.

π_0	m	cut-point	λ_{best}	$\widehat{\lambda}$	$MSE(\lambda_{best})$	$MSE(\widehat{\lambda})$
1	1000	2	0	0	0.0602	0.0602
0.8	1000	2	0.8	0.8	0.00444	0.00444
0.5	1000	2	0.9	0.9	0.000779	0.000779
0.2	1000	2	0.95	0.9	0.000318	0.000362
0.8	100	2	0.75	0.65	0.123	0.127
0.8	500	2	0.75	0.75	0.00953	0.00953
0.8	10000	2	0.9	0.9	0.000556	0.000556
0.8	1000	0	0.7	0.85	0.00445	0.00556
0.8	1000	1	0.7	0.8	0.00361	0.00385
0.8	1000	3	0.85	0.9	0.0323	0.0326

of $N(0,1)$ versus $N(1,1)$ with the rejection region $\Gamma = [c, \infty)$. We calculated λ_{best} from the true MSE for each case. For each set of parameters, we performed the bootstrap procedure on 100 data sets with $B = 500$, and then averaged their predicted MSE curves. $\widehat{\lambda}$ was chosen by taking the minimum of the *averaged* MSE curves. Taking the median of the 100 $\widehat{\lambda}$ produces nearly identical results.

Figure 7.1 shows the results for $m = 1000$ and $c = 2$ over the values $\pi_0 = 1, 0.8, 0.5, 0.2$. Averaging over applications of the procedure only 100 times gives us the correct λ_{best} for the first three cases. Note that it is not important to predict the MSE curve, but rather where its minimum is. It can also be seen from the plots that the bootstrap procedure produces a conservative estimate of the MSE for any λ. Table 7.1 shows simulation results for several other sets of parameters. It can be seen that even when $\widehat{\lambda} \neq \lambda_{best}$, the difference in their true MSE's is not very drastic, so the minimum MSE is nearly attained in almost all situations we simulated.

7.2 Fixed False Discovery Rates

If we knew $FDR(\cdot)$, then in order to control the FDR at level α, we would simply calculate $t_\alpha[FDR]$. We have shown $t_\alpha[\widehat{FDR}_\lambda]$ conservatively estimates $t_\alpha[FDR]$ in the asymptotic sense, so we want to find

$$\arg\min_\lambda \mathbf{E}\left[\left(t_\alpha[\widehat{FDR}_\lambda] - t_\alpha[FDR]\right)^2\right].$$

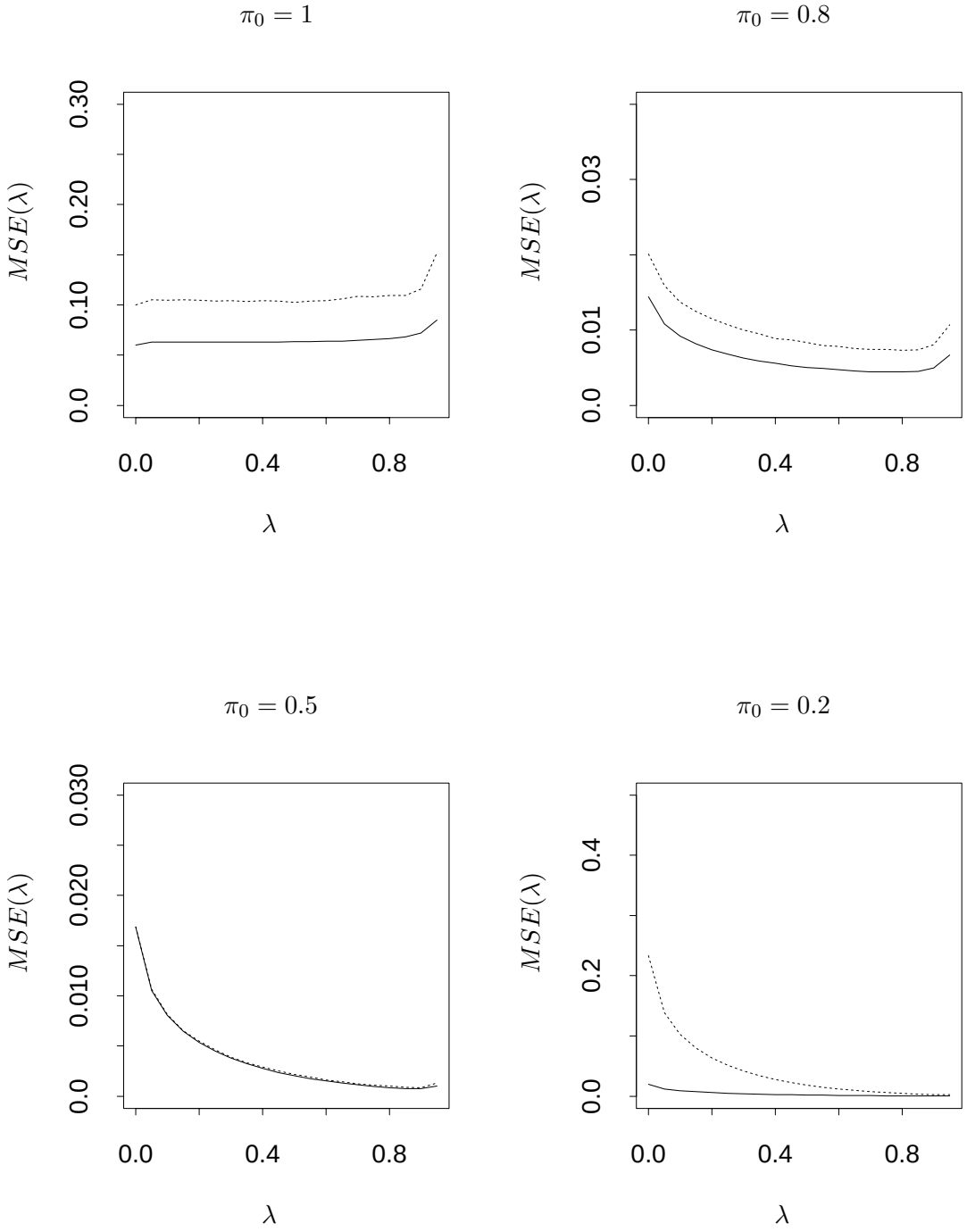


Figure 7.1: Plots of $MSE(\lambda)$ versus λ for $\Gamma = [2, \infty)$ over various values of π_0. The solid line is the true MSE. The dashed line is the MSE predicted by the bootstrap procedure averaged over 100 applications.

Asymptotically, as λ gets larger the bias of $t_\alpha[\widehat{FDR}_\lambda]$ decreases, but the variance increases. Therefore, it makes sense to use this MSE as an optimality criterion for choosing λ.

Under independence, we can re-sample the p-values with replacement, and form a bootstrap version $t_\alpha[\widehat{FDR}_\lambda^{*(b)}]$ of $t_\alpha[\widehat{FDR}_\lambda]$ for each bootstrap set $b = 1, \ldots, B$. Then a good estimate of the MSE would be

$$\frac{1}{B}\sum_{b=1}^{B}\left(t_\alpha[\widehat{FDR}_\lambda^{*(b)}] - t_\alpha[FDR]\right)^2,$$

however, similarly to the previous scenario, we do not know $t_\alpha[FDR]$. From Theorem 5.4, we know that

$$\sup_\lambda t_\alpha\left[\widehat{FDR}_\lambda\right] \approx \sup_\lambda \lim_{m\to\infty} t_\alpha\left[\widehat{FDR}_\lambda\right] \leq t_\alpha\left[FDR\right]$$

so that a good plug-in estimate of $t_\alpha[FDR]$ is $\sup_\lambda t_\alpha[\widehat{FDR}_\lambda]$.

Now to make this procedure computationally simple, we again only consider a range of λ values denoted by $\mathcal{R}$. Then for each $\lambda \in \mathcal{R}$, we calculate

$$\widehat{MSE}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\left(t_\alpha[\widehat{FDR}_\lambda^{*(b)}] - \max_{\lambda'\in\mathcal{R}} t_\alpha[\widehat{FDR}_{\lambda'}]\right)^2.$$

Then taking $\widehat{\lambda} = \arg\max_{\lambda\in\mathcal{R}} \widehat{MSE}(\lambda)$, we use $t_\alpha[\widehat{FDR}_{\widehat{\lambda}}]$ as our thresholding rule. This procedure is outlined in Algorithm 7.2 on page 88.

To test the performance of this procedure, we performed the following simulation. 1000 tests of $N(0,1)$ (null) versus $N(2,1)$ (alternative) were performed using one-sided rejection regions with either $\pi_0 = 0.5$ or $\pi_0 = 0.8$. The range of λ we considered is $\mathcal{R} = \{0, 0.10, \ldots, 0.90\}$. 100 data sets were generated for each value of π_0, and $B = 500$ bootstrap iterations were performed for each data set.

Figure 7.2 shows the results of our simulation. The true MSE curve is shown, as well as the average of the MSE curves over the 100 data sets. Notice that the *shapes* of the curves are very similar. (The shape is the most important feature because we are trying to find the minimum of the curve.) Also, the histograms show the 100 realized values of $\widehat{\lambda}$ for each case. It can be seen that the $\widehat{\lambda}$ are well concentrated around the low portions of the MSE curves.

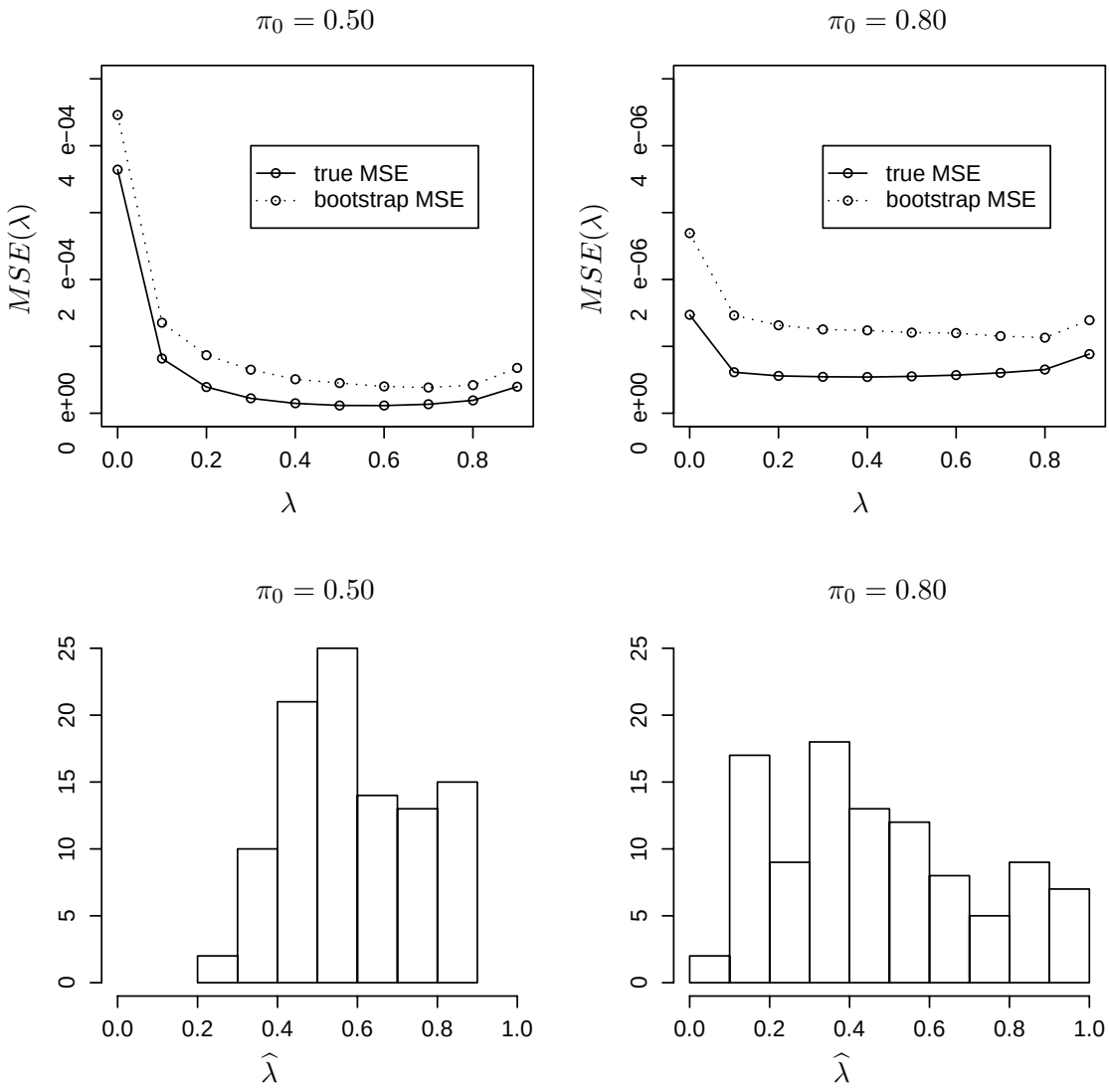


Figure 7.2: Simulation results for automatically choosing λ in the FDR controlling procedure $t_\alpha[\widehat{FDR}_\lambda]$. The plots show a comparison between the true MSE curve and the average of 100 estimated MSE curves, the similarity in shape being the most important feature. The histograms show the 100 realized $\widehat{\lambda}$ values, chosen to find the λ that minimizes the true $MSE(\lambda)$.

Algorithm 7.2: FDR Controlling Procedure with Automatically Chosen λ

1. For some range of λ, say $\mathcal{R} = \{0, 0.10, \ldots, 0.90\}$, calculate $t_\alpha[\widehat{FDR}_\lambda]$ as in Algorithm 5.1 on page 63.

2. For each $\lambda \in \mathcal{R}$, form B bootstrap versions $t_\alpha[\widehat{FDR}_\lambda^{*(b)}]$ of the estimate, $b = 1, \ldots, B$ (by sampling with replacement from the p-values, and repeating the procedure on these).

3. For each $\lambda \in \mathcal{R}$, estimate its respective mean squared error as:

$$\widehat{MSE}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\left(t_\alpha\left[\widehat{FDR}_\lambda^{*(b)}\right] - \max_{\lambda\in\mathcal{R}} t_\alpha\left[\widehat{FDR}_\lambda\right]\right)^2.$$

4. Set $\widehat{\lambda} = \arg\min_{\lambda\in\mathcal{R}} \widehat{MSE}(\lambda)$. Our overall FDR controlling thresholding rule is

$$\text{reject all p-values } \leq t_\alpha\left[\widehat{FDR}_{\widehat{\lambda}}\right].$$

5. For the final sample case, use $t_\alpha[\widehat{FDR}_\lambda^{*}]$ (as was outlined in Algorithm 5.1) in the above steps in place of $t_\alpha[\widehat{FDR}_\lambda]$.

7.3 q-values and Simultaneous Controlling Curves

By Theorem 6.1, we know that the estimated q-values given by Algorithm 6.1 are asymptotically conservative over all rejection regions simultaneously. This holds irrespective of the choice of λ in the calculation of $\widehat{\mathrm{q}}_\lambda$. Therefore, similar ideas can be used as those in the previous scenarios. The problem here, however, is how to calculate an MSE. Recall that $\widehat{\mathrm{q}}(\cdot)$ is an estimate over the entire interval $[0, 1]$, with the important arguments being the observed p-values. As a simple remedy to this problem, we propose estimating λ that minimizes

$$MSE_{\mathcal{T}}(\lambda) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \mathbf{E}\left[(\widehat{\mathrm{q}}_\lambda(t) - \mathrm{q}(t))^2\right]$$

for some user-defined set of rejection regions denoted by the set of their end-points $\mathcal{T}$. For example, if one is concerned with the q-values being most accurate over small rejection regions, one could take $\mathcal{T} = \{0.001, 0.002, \ldots, 0.01\}$. Or if one wanted the overall behavior of the estimated q-values to be good, one could use $\mathcal{T} = \{0.1, 0.2, \ldots, 0.9\}$.

Given this convention, the remainder of the method is straightforward. As in the previous cases, we use $\min_{\lambda\in\mathcal{R}} \widehat{\mathrm{q}}_\lambda(t)$ as the plug-in estimate of $\mathrm{q}(t)$ for each $t \in \mathcal{T}$. (And again $\mathcal{R}$ is a well chosen grid of λ values over which to do the calculations.) The MSE is then estimated by

$$\widehat{MSE}(\lambda) = \frac{1}{B \cdot |\mathcal{T}|} \sum_{t\in\mathcal{T}} \sum_{b=1}^{B} \left(\widehat{\mathrm{q}}_\lambda^{*b}(t) - \min_{\lambda'\in\mathcal{R}} [\widehat{\mathrm{q}}_{\lambda'}(t)] \right)^2 .$$

We then find $\widehat{\lambda} = \arg\min_{\lambda\in\mathcal{R}} \widehat{MSE}(\lambda)$ and use $\widehat{\lambda}$ in calculating all the q-values by Algorithm 6.1. This procedure is outlined in Algorithm 7.3 on page 89. Also, the methodology for choosing λ in $\widehat{\alpha}_{FDR,\lambda}$ is exactly analogous, and is included in Algorithm 7.3.

Algorithm 7.3: Estimation of q-values and α_{FDR} with Automatically Chosen λ

1. Let $\mathcal{T}$ be the end points of a set of "important" rejections regions, say, $\mathcal{T} = \{0.001, 0.002, \ldots, 0.01\}$. ($\mathcal{T}$ can be the set of observed p-values, or some subset of them.)
2. For some range of λ, say $\mathcal{R} = \{0, 0.10, 0.20 \ldots, 0.90\}$, calculate $\widehat{\mathrm{q}}_\lambda(t)$ for $t \in \mathcal{T}$.
3. For each $\lambda \in \mathcal{R}$, form B bootstrap versions $\widehat{\mathrm{q}}_\lambda^{*b}(t)$ of the estimate, $b = 1, \ldots, B$ for each $t \in \mathcal{T}$.
4. For each $\lambda \in \mathcal{R}$, estimate its respective mean squared error as:
$$\widehat{MSE}(\lambda) = \frac{1}{B \cdot |\mathcal{T}|} \sum_{t\in\mathcal{T}} \sum_{b=1}^{B} \left(\widehat{\mathrm{q}}_\lambda^{*b}(t) - \min_{\lambda'\in\mathcal{R}} [\widehat{\mathrm{q}}_{\lambda'}(t)] \right)^2 .$$
5. Set $\widehat{\lambda} = \arg\min_{\lambda\in\mathcal{R}} \widehat{MSE}(\lambda)$.
6. Calculate $\widehat{\mathrm{q}}(\cdot)$ using Algorithm 6.1 with $\widehat{\lambda}$.
7. In order to estimate $\widehat{\lambda}$ for $\widehat{\alpha}_{FDR,\lambda}$, perform the exact same procedure using the $\widehat{\alpha}_{FDR,\lambda}$.

Using simulation, the behavior of this method is very similar to that shown above, so we omit the simulation results here.

7.4 An Overall Automatically Chosen λ

If one is interested in applying all methodologies from the three scenarios with an overall optimally chosen λ, or if one is merely interested in the $\widehat{\pi}_0(\lambda)$ estimate itself, it makes sense to estimate the λ that minimizes

$$\mathbf{E}\left[\left(\widehat{\pi}_0(\lambda) - \pi_0\right)^2\right].$$

Since the reasoning and simulation results for this case are very similar to the previous three sections, we will simply list the procedure in Algorithm 7.4.

Algorithm 7.4: Estimation and Inference of π_0 with Optimal λ

1. For some range of λ, say $\mathcal{R} = \{0, 0.05, 0.10, \ldots, 0.95\}$, calculate $\widehat{\pi}_0(\lambda)$ as in Section 4.1.
2. For each $\lambda \in \mathcal{R}$, form B bootstrap versions $\widehat{\pi}_0^{*b}(\lambda)$ of the estimate, $b = 1, \ldots, B$.
3. For each $\lambda \in \mathcal{R}$, estimate its respective mean squared error as:

$$\widehat{MSE}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\pi}_0^{*b}(\lambda) - \min_{\lambda' \in \mathcal{R}}[\widehat{\pi}_0(\lambda')]\right)^2.$$

4. Set $\widehat{\lambda} = \arg\min_{\lambda \in \mathcal{R}} \widehat{MSE}(\lambda)$. Our overall estimate of π_0 is

$$\widehat{\pi}_0 = \widehat{\pi}_0(\widehat{\lambda}).$$

5. Form a $1-\alpha$ upper confidence interval of $\widehat{\pi}_0$ by taking the $1-\alpha$ quantile of $\{\widehat{\pi}_0^{*1}(\widehat{\lambda}), \ldots, \widehat{\pi}_0^{*B}(\widehat{\lambda})\}$ as the upper endpoint.

Chapter 8

Applications to DNA Microarrays

DNA microarrays are a relatively new biotechnology that allow the simultaneous measurement of the expression levels of thousands of genes from a biological sample (Brown & Botstein 1999). This exciting area of biological research has created several challenging statistical problems, including the multiple hypothesis testing problem we have considered in this work. In this chapter, we discuss the application of the methods we have introduced to DNA microarray data. Several adjustments to the methods are suggested to accommodate the dependence structure encountered in these data. We specifically consider the problem of detecting differential gene expression between two or more biological conditions, which was already introduced in Section 1.3.

8.1 An Example

Here is a simple example of estimating the FDR over a fixed rejection region. Rieger, Tusher, Hong, Tibshirani, and Chu (2001, unpublished[1]) analyze DNA expression data on 3000 genes from a study of the effects of ionizing radiation, comparing normal patients to radiation sensitive patients. There are 15 samples in group 1 (normal) and 13 in group 2 (radiation sensitive). Figure 8.1 is a histogram of the 3000 two-sample t-statistics from the genes. They range from -4.54 to 3.72. Suppose we decide to reject all genes whose t-statistic is greater than 2 in absolute value; there are 146 such genes. What is the FDR among these 146 genes? To assess this, we do a random permutation of the sample

[1]Data available on request.

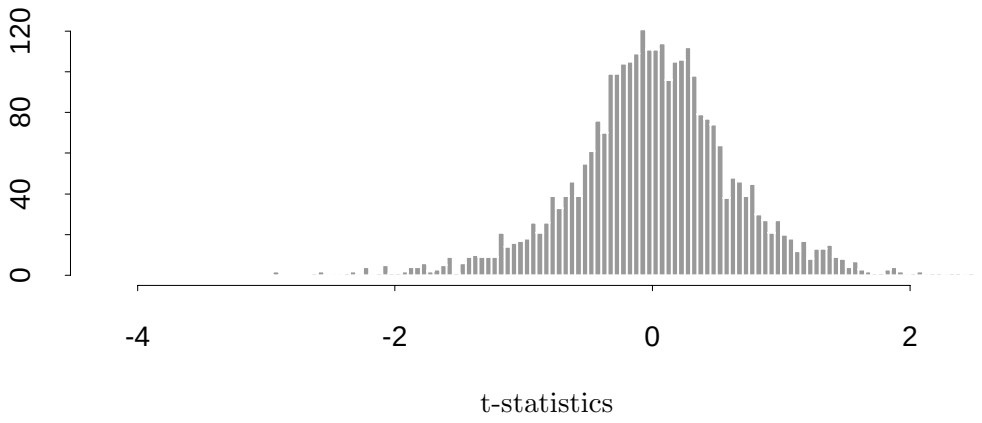


Figure 8.1: Histogram of 3000 t-statistics from the DNA microarray example.

labels $(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2)$, and recompute the t-statistics to simulate the full null case. Using the permutations, we can calculate the Type I error rate of the rejection region ± 2. A total of 1232 out of the $3000 \cdot 100$ permutation statistics exceed 2 in absolute value, a proportion of 0.0046. If we assume each null statistic has a t-distribution on 26 degrees of freedom, then the Type I error rate over this region is calculated to be 0.056, so clearly it is important to use a permutation distribution. The p-value of each statistic can also be calculated. Since we are using symmetric rejection regions, the p-value of an observed statistic x_i is the proportion of permutation statistics whose absolute value exceeds $|x_i|$. Given these p-values and the Type I error rate of ± 2, we can calculate our estimate of the FDR. For example, for two different choices of λ we have:

$$\begin{aligned} \widehat{FDR}_{\lambda=0}(0.0046) &= \tfrac{1 \cdot 0.0046}{146/3000} &&= 0.0844 \\ \widehat{FDR}_{\lambda=0.7}(0.0046) &= \tfrac{668/3000 \cdot 0.0046}{0.3 \cdot 146/3000} &&= 0.0752. \end{aligned}$$

As we have mentioned, it is equivalent to use p-values or the statistics, so there was no loss of information in compressing the t-statistics and symmetric rejection regions into p-values. Also, even though simulating the full null case is unrealistic because we expect a fraction of

the genes to be significant, these calculations were only used to calculate the Type I error over the possible rejections. As we will see, this calculation is unaffected by dependence. We formally describe the steps to perform our methodology on DNA microarray data in Section 8.3.

8.2 Dependence in DNA Microarrays

Suppose we collect data from n microarrays with the same m genes on each. Essentially, we observe the vectors $\mathbf{D}(j) = (D_{1j}, \ldots, D_{mj})$ for $j = 1, \ldots n$. This corresponds to the m expression measurements on the m genes for the j^{th} array. The components of the vectors are dependent, but the observations are independent in some way. In other words, we assume that the D_{ij} are independent across the $j = 1, \ldots, n$ observations for each i, but that they are not necessarily independent or identically distributed across the $i = 1, ..., m$ components of the vector for each j. Therefore, the data may be represented as a $m \times n$ matrix $\mathbf{D}$, with each column corresponding to an observed m-vector. The columns are independent (as in the Rieger et al. data), but the rows are dependent.

For each row of $\mathbf{D}$, we form a statistic X_i that is some function of $D_{i1}, \ldots, D_{in}$, $i = 1, \ldots, m$. We wish to test a hypothesis about a parameter of interest for each X_i. Therefore, we are testing m dependent hypotheses using the statistics $X_1, \ldots, X_m$. The null $X_1^0, \ldots, X_m^0$ are likely generated by permuting the columns in the appropriate way to simulate the null case, as was done above.

Ideally, the statistics can be formed so that they are exchangeable in the sense that the $X_i|H_i = 0$ are identically distributed. That way, all the statistics can be used in gathering information about the null distribution, and the same rejection region (in the original space) can be used for each test. If this is not possible, then a p-value can be calculated for each statistic by simulating the null distribution individually. The problem with this is that these p-values are on a much more granular scale than if the statistics are identically distributed under the null hypothesis.

Throughout this work, we have proved several results under the assumption that the statistics are "weakly dependent." We hypothesize that the most likely form of dependence encountered in DNA microarrays is "weak dependence", and more specifically, "clumpy dependence." In other words, the measurements on the genes are dependent in small groups,

each group being independent of the others. There are two reasons for this clumpy dependence. The first is that genes tend to work in pathways, that is, small groups of genes interact to produce some overall process. This can involve just a few to 50 or more genes. The second reason is that there tends to be cross-hybridization in DNA microarrays. In other words, the signals between two genes can cross because of molecular similarity at the sequence level. Cross-hybridization would only occur in small groups; genes that have a molecular similarity do so because of an evolutionary and/or functional relationship, not by random chance.

Typically microarrays measure the expression levels on 3000 to 30,000 genes, and each gene makes up a hypothesis test. Therefore, given the clumpy dependence and large number of tests, we expect the results involving the weak dependence assumption should apply; these include the asymptotic results in Sections 2.3, 4.5, 5.4, and 6.3. The specific dependence structure between the genes is unknown at the point for basically all organisms whose genome has been sequenced. Therefore, it is impossible to incorporate this knowledge *a priori* into our methodology. If the dependence structure is estimated, then it can be incorporated by using Theorem 2.2, for example. In Section 8.5, we perform a simple numerical study where we simulate the hypothesized dependence structure. It is shown there that the estimates still remain conservative, further supporting the arguments we have made.

8.3 Applying the Methodologies to DNA Microarray Data

The steps we present now can be used for any situation where one is performing a hypothesis test on each gene from microarray data. Let $X_1, \ldots, X_m$ be the m statistics, each one corresponding to a different gene. For generality, we denote the rejection regions by the set $\{\Gamma_t\}$ where t is the Type I error rate of Γ_t. Note that the set of possible rejection regions is nested. Moreover, we have that $x_i \in \Gamma_t$ if and only if p-value$(x_i) \leq t$. We make the important assumption that null versions of the statistics can be simulated; denote these simulated null statistics by $X_1^0, \ldots, X_m^0$. For example, in the Rieger et al. data $X_1^0, \ldots, X_m^0$ were generated by permuting the "normal" and "radiation sensitive" labels, and recomputing the m statistics.

The task of nonparametrically capturing the dependence structure for false discovery rates is difficult because we can only observe $R(\Gamma_t) = \#\{X_i \in \Gamma_t\}$ along with $R^0(\Gamma_t) =$

$\#\{X_i^0 \in \Gamma_t\}$ (Also, note we can observe $W(\Gamma_t) = m - R(\Gamma_t)$ and $W^0(\Gamma_t) = m - R^0(\Gamma_t)$.) In the work of Westfall & Young (1993), using the simulated null $X_1^0, \ldots, X_m^0$ turns out to be very important in preserving the dependence structure in calculating FWER adjusted p-values. However, for the pFDR and FDR, the dependence structure obtained from $X_1^0, \ldots, X_m^0$ is not so useful, especially given that false discovery rates involve sums of indicator variables.

Yekutieli & Benjamini (1999) attempt to use the $X_1^0, \ldots, X_m^0$ to capture the dependence structure of V and S in estimating the FDR. However, upon close examination of their method, the $X_1^0, \ldots, X_m^0$ are more or less used to estimate the expected number of false positives when all hypotheses are null. Since $R^0(\Gamma_t) = V^0(\Gamma_t) + S^0(\Gamma_t)$ and the dependence structures of V and S can radically differ, we find it futile to capture the dependence through $R^0(\Gamma_t)$. We directly use the simulated $R^0(\Gamma_t)$ to calculate $\mathbf{E}[R^0(\Gamma_t)]$. Since $\mathbf{E}[R^0(\Gamma_t)] = m \cdot t$, the simulated null distribution is essentially used to identify the "t" associated with each Γ_t. Also, note that the calculation for $\mathbf{E}[R^0(\Gamma_t)]$ is theoretically unaffect by dependence because it is a sum of indicator random variables.

In this general notation, however, we can re-write our estimates as:

$$\widehat{FDR}_\lambda(\Gamma_t) = \frac{W(\Gamma_\lambda)}{\mathbf{E}[W^0(\Gamma_\lambda)]} \cdot \frac{\mathbf{E}[R^0(\Gamma_t)]}{R(\Gamma_t) \vee 1},$$

$$\widehat{pFDR}_\lambda(\Gamma_t) = \frac{W(\Gamma_\lambda)}{\mathbf{E}[W^0(\Gamma_\lambda)]} \cdot \frac{\mathbf{E}[R^0(\Gamma_t)]}{\mathbf{Pr}(R^0(\Gamma_t) > 0) \cdot (R(\Gamma_t) \vee 1)}.$$

This is equivalent to calculating the p-values of each statistic using the nested set of rejection regions and then computing $\widehat{FDR}_\lambda(t)$ and $\widehat{pFDR}_\lambda(t)$ from Chapter 4, except for one small difference. Note that because of the dependence it will most likely be the case that $\mathbf{Pr}(R^0(\Gamma_t) > 0) \neq 1 - (1-t)^m$. In fact it can be much smaller. Therefore, for dependent statistics, we recommend using $\mathbf{Pr}(R^0(\Gamma_t) > 0)$ in place of $1 - (1-t)^m$ in $\widehat{pFDR}_\lambda(t)$. Once the p-values are calculated, and this small adjustment is made, then we can apply all the methodology described in Chapters 4, 5, and 6. This is summarized in Algorithm 8.1 on page 96.

Two remarks should be made about the rejection regions. The first is that we assume the rejection regions are chosen before any data are seen. If the data are used to derive asymmetric rejection regions, then these same data cannot be used in applying our methodology. This issue is further discussed in Section 8.8. Secondly, Algorithm 8.1 implies that

---

Algorithm 8.1: Applying the Methodologies to Microarray Data

1. Let $\{\Gamma\}$ be the nested set rejection regions chosen *a priori*, and let $x_1, \ldots, x_m$ be the observed statistics.

2. Simulate the null statistics for B iterations to obtain sets $X_1^{0b}, \ldots, X_m^{0b}$ for $b = 1, \ldots, B$.

3. Calculate the Type I error rate for each relevant rejection region in $\{\Gamma\}$ by

$$t = \frac{1}{m \cdot B} \sum_{b=1}^{B} \#\{X_i^{0b} \in \Gamma\},$$

   and index Γ by Γ_t. It is only be necessary to do this for at most m rejection regions, each one being the smallest containing x_i for $i = 1, \ldots, m$.

4. Calculate the p-values $p_1, \ldots, p_m$ from the statistics $x_1, \ldots, x_m$ and $\{\Gamma_t\}$.

5. For any $0 \leq t \leq 1$, form $\widehat{pFDR}_\lambda(t)$ by

$$\widehat{pFDR}_\lambda(t) = \frac{W(\lambda) \cdot t}{(1-\lambda) \cdot \{R(t) \vee 1\} \cdot \mathbf{Pr}(R^0(t) > 0)},$$

   where $\mathbf{Pr}(R^0(t) > 0) = \frac{1}{B} \sum_{b=1}^{B} 1\left(\#\{X_i^{0b} \in \Gamma_t\} > 0\right)$, $R(t) = \#\{p_i : p_i \leq t\}$, and $W(\lambda) = \#\{p_i : p_i > \lambda\}$.

6. Form the $\widehat{FDR}_\lambda(\cdot)$, $t_\alpha[\widehat{FDR}_\lambda]$, $\widehat{\alpha}_{FDR,\lambda}(\cdot)$, and $\widehat{\mathrm{q}}_\lambda(\cdot)$ exactly as before, using the new $\widehat{pFDR}_\lambda(\cdot)$ in the calculation of $\widehat{\mathrm{q}}_\lambda(\cdot)$.

---

the Type I error rate must be computed for every rejection region. However, this will usually not be the case. For example, with symmetric rejection regions, then obviously only $\pm x_i$ must be considered, where $x_1, \ldots, x_m$ are the observed statistics.

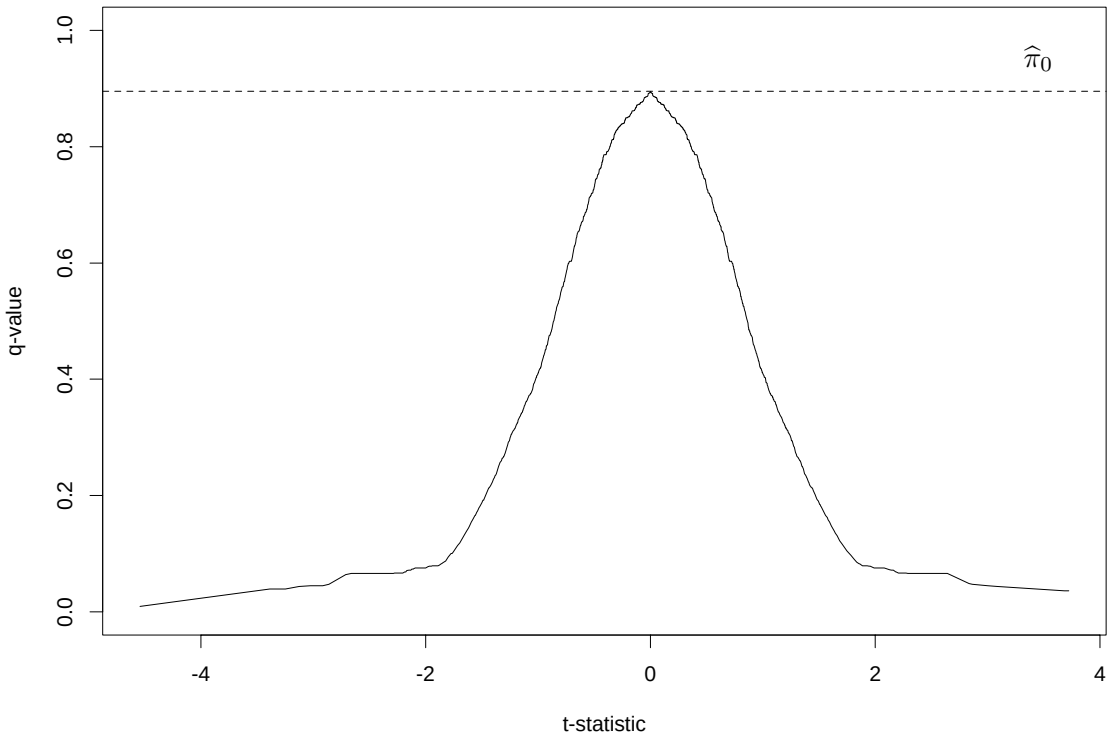


Figure 8.2: The q-values for each t-statistic from the DNA microarray example.

Using Algorithm 8.1, we calculated the q-values for the t-statistics (with $\lambda = 0.7$ again). It can be seen from Figure 8.2 that $\widehat{\mathrm{q}}_\lambda(x_i) \leq \widehat{\pi}_0 \leq 1$, for $i = 1, \ldots, 3000$, and they decrease as the t-statistics become more extreme. For each gene, its q-value estimates the minimum pFDR that can be attained when calling that gene significant. Moreover, the q-values have the asymptotic posterior probability interpretation from Chapters 3 and 6.

8.4 A Comparison to Existing Methods

We now compare the results obtained in the Rieger et al. data example to what is obtained by using other methods. We reported $\widehat{FDR}_\lambda = 7.52\%$ ($\lambda = 0.7$) when rejecting all t-statistics beyond ± 2, for a total of 146 significant genes. With the method described in Tusher et al. (2001), the reported FDR would have been 8.44%. The rejection region that estimates FDR at 7.52% using this method rejects only 87 genes. Using the methodology of Yekutieli & Benjamini (1999), we estimate the FDR as 8.31%. For this method, the rejection region that estimates FDR at 7.52% rejects 91 genes.

Now by the material presented in Chapter 5, we can use control the FDR at level 7.52% by computing $t_\alpha[\widehat{FDR}_\lambda]$ with $\alpha = 0.0752$. In doing so, we obviously find 146 genes significant again. Controlling the FDR at level 7.52% with the Benjamini & Hochberg (1995) method results in 87 significant genes. If we make the correction for general dependence given in Benjamini & Yekutieli (2001), we reject only 1 gene, controlling the FDR at level $7.52\%/\log(3000) = 1.0\%$.

8.5 A Numerical Study

In this section we carry out a simulation study of the pFDR estimate in three settings: independence, clumpy dependence, and general dependence. (Results for the FDR estimate are similar.) Note that clumpy dependence is a special case of "weak dependence" used throughout this work. We use $m = 1000$ genes and 20 samples, simulating a DNA microarray data set in the spirit of the Rieger et al. data example. Letting x_{ij} be the measurement for gene i and sample j, here is how the data were generated:

$$\begin{aligned} x_{ij} &\sim N(0,1) + 3 \cdot I(i \le 50 \;\&\; j \ge 11) \;\; (\textit{independence}) \\ x_{ij} &\sim N(0,1) + 3 \cdot I(i \le 50 \;\&\; j \ge 11) + \nu_i \;\; (\textit{clumpy}) \\ x_{ij} &\sim N(0,1) + 3 \cdot I(i \le 50 \;\&\; j \ge 11) + \mu \;\; (\textit{general}) \end{aligned} \tag{8.1}$$

Hence samples 11–20 are over-expressed by 3 units for the first 50 genes. In the general dependence setting, μ is a vector of 20 $N(0, 0.25)$ random variables, and the same μ is added to every gene. In clumpy dependence, each ν_i is vector of 20 $N(0, 0.04)$ random variables, and the same ν_i is added to each consecutive gene block of size 50. The rejection regions for the two-sample t-statistics are formed at thresholds chosen by various quantiles of the

null distribution. The calculations were done as if the null distribution were unknown. The results are shown in Table 8.1.

In the first two settings, the pFDR estimate is very accurate. In the third setting, it is biased upward, sometimes by as much as 13%. This is partly due to its overestimation of π_0, and it would not be as bad if $\widehat{\pi}_0$ were truncated at 1. (Although, this would affect the theoretical results of the previous section.) The last two sections of Table 8.1 explore further variations of the clumpy dependence setup. In the first, the "3" in equation (8.1) is replaced by zero, and hence no genes are affected. In the last setting, the "3" is replaced by a random effect from the $N(2,1)$ distribution, and is applied to the first 500 (rather than 50) genes. The estimate of the pFDR is accurate in both cases.

8.6 Bootstrap Confidence Intervals

In Chapter 4, we bootstrapped the statistics (or p-values) in order to obtain upper confidence intervals for the pFDR and FDR. So why don't we do that here? First, we cannot bootstrap the statistics because they are dependent, and we don't know the dependence structure. In most multiple hypothesis testing situations, whether the statistics are dependent or not, there will be some independent dimension to the data. In the Rieger et al. data example, the experiments are independent, so we could bootstrap these leaving the dependence structure intact.

When bootstrapping the experiments, we run into an issue that is similar to the problem of regions investigated in Efron & Tibshirani (1998). For example, suppose we want to find a bootstrap estimate of our confidence that $pFDR \in [0,c]$ for some c. In the Rieger et al. data example there are 28 independent experiments, so for each bootstrap iteration we sample with replacement from these to get a bootstrap sample of 28 experiments with 3000 measurements (genes) for each. We form 3000 new t-statistics, and apply our procedure to estimate the pFDR over the region exceeding ± 2 as before. Therefore, we get $\widehat{pFDR}^{*b}$ for $b = 1, \ldots, B$ and we count how many fall in $[0,c]$; $\widehat{\theta} = \#\{\widehat{pFDR}^{*b} \in [0,c]\}/B$ would be our estimated confidence for this region.

As it turns out $\widehat{\theta}$ would be grossly inflated because the R^{*b} tend to be too big and the W^{*b} tend to be too small, making the $\widehat{pFDR}^{*b}$ too small. Efron & Tibshirani (1998) propose methods to correct for this. In our case, we do not wish to calculate the confidence for a particular interval $[0,c]$, rather we want to calculate a 90% confidence interval, for

Table 8.1: Simulation results for $\widehat{pFDR}_\lambda$ under dependence. Values are the mean and standard error of the mean over 20 simulations.

	Threshold quantile $(1-t)$						
	0.800	0.900	0.950	0.975	0.990	0.995	0.999
	Independence						
π_0				$pFDR$			
0.9500	0.7890	0.6421	0.4734	0.3008	0.1535	0.0843	0.0183
	0.0029	0.0052	0.0073	0.0092	0.0094	0.0064	0.0043
$\hat{\pi}_0$				$\widehat{pFDR}$			
0.9623	0.8138	0.6910	0.5080	0.3377	0.1634	0.0882	0.0189
0.0221	0.0221	0.0208	0.0146	0.0104	0.0046	0.0020	0.0004
	Clumpy dependence						
π_0				$pFDR$			
0.9500	0.7889	0.6570	0.4906	0.3237	0.1678	0.1012	0.0166
	0.0045	0.0072	0.0092	0.0116	0.0099	0.0094	0.0025
$\hat{\pi}_0$				$\widehat{pFDR}$			
0.9412	0.7917	0.6426	0.4779	0.3174	0.1565	0.0845	0.0185
0.0320	0.0274	0.0210	0.0159	0.0109	0.0054	0.0029	0.0006
	General dependence						
π_0				$pFDR$			
0.9500	0.7723	0.6140	0.4415	0.2862	0.1455	0.0824	0.0265
	0.0239	0.0361	0.0437	0.0425	0.0350	0.0247	0.0093
$\hat{\pi}_0$				$\widehat{pFDR}$			
1.0297	0.8640	0.7481	0.5522	0.3593	0.1742	0.0942	0.0204
0.0527	0.0594	0.0582	0.0417	0.0246	0.0105	0.0053	0.0010
	Clumpy dependence- all genes null						
π_0				$pFDR$			
1.000	1	1	1	1	1	1	1
	0	0	0	0	0	0	0
$\hat{\pi}_0$				$\widehat{pFDR}$			
0.9913	0.9880	0.9995	0.9999	0.9730	0.9781	0.9803	1.0220
0.0348	0.0296	0.0326	0.0381	0.0415	0.0376	0.0458	0.0597
	Clumpy dependence - half of genes affected						
π_0				$pFDR$			
0.500	0.0091	3e-03	0.0015	1e-03	7e-04	0.001	0.000
	0.0010	9e-04	0.0005	7e-04	7e-04	0.001	0.000
$\hat{\pi}_0$				$\widehat{pFDR}$			
0.5192	0.0083	3e-03	0.0013	5e-04	3e-04	3e-04	1e-04
0.0149	0.0007	4e-04	0.0002	1e-04	1e-04	1e-04	1e-04

example. Future work will adapt their methodology to obtaining confidence intervals for the pFDR and FDR for multidimensional data with independence in at least on direction (as would be expected in most multiple hypothesis testing).

In Chapter 7, we also the bootstrapped the statistics to choose the optimal λ. For dependent hypotheses, we are able to choose λ nearly as effectively using an older idea.

8.7 Automatically Choosing λ

In Chapter 7, we presented numerical methods for estimated the λ that minimizes the mean-squared error for the various estimates. Since the statistics in the microarray problem are dependent, we have to introduce new methods for estimating the optimal λ. We will exploit the fact that the experiments are independent measurements of the dependent genes. We use an approach involving a jackknife estimate of the variance, and an estimate of the bias that is not much different from what was obtained in Chapter 7. For simplicity, we concentrate on the $\widehat{pFDR}_\lambda(t)$ estimate, although the other cases are adjusted analogously.

We assume that there is some independent dimension of the data of size n. In the Rieger et al. data example, the experiments are independent observations of the 3000 dependent genes, so $n = 28$ in that case. In most problems, there will be a repeated observation of some sort that will give us the required property. By removing the i^{th} experiment, we can form a new estimate of the pFDR with the remaining data. For each fixed $\lambda \in \mathcal{R}$, denote this estimate by $\widehat{pFDR}_\lambda^{(-i)}(t)$ for $i = 1, \ldots, n$. The jackknife estimate of variance is:

$$\widehat{\mathbf{Var}}_\lambda = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{pFDR}_\lambda^{(-i)}(t) - \widehat{pFDR}_\lambda(t) \right)^2 .$$

The jackknife estimate of bias works poorly here, so we use a different estimate. Ideally, if we knew $pFDR(t)$, we could estimate the squared bias by $(\widehat{pFDR}_\lambda(t) - pFDR(t))^2$, however, we obviously do not know $pFDR(t)$. As was done in Chapter 7, we use the same plug-in estimate of $pFDR(t)$: $\min_{\lambda \in \mathcal{R}} \widehat{pFDR}_\lambda(t)$ for some range of $\mathcal{R}$, say $\mathcal{R} = \{0, 0.05, \ldots, 0.95\}$. The estimate of the squared bias is

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(t) - \min_{\lambda' \in \mathcal{R}} \widehat{pFDR}_{\lambda'}(t) \right)^2 .$$

Each of these estimates gives a nice estimate of the shape of the squared bias and

Table 8.2: Simulation results for the procedure to pick the optimal λ.

m_0	u	λ_{best}	median $\widehat{\lambda}$	mean $\widehat{\lambda}$
200	0.3	0.60	0.575	0.56
200	0.5	0.75	0.55	0.54
200	0.75	0.45	0.45	0.45
500	0.3	0.75	0.60	0.59
m_0	u	$MSE(\lambda_{best})$	MSE(median $\widehat{\lambda}$)	MSE(mean $\widehat{\lambda}$)
200	0.3	0.026	0.027	0.027
200	0.5	0.0057	0.0058	0.0058
200	0.75	8.2×10^{-4}	8.2×10^{-4}	8.2×10^{-4}
500	0.3	0.035	0.037	0.037

variance curves over λ. However, each one tends to be inflated, and the jackknife estimate of variance can be unpredictably inflated. Therefore, we scale each estimate by its median over the $\lambda \in \mathcal{R}$, and make the following adjustments to our estimates:

$$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{\text{median}_{\lambda' \in \mathcal{R}}(\widehat{bias}_{\lambda'}^2)}$$
$$\widehat{\mathbf{Var}}_\lambda^* = \frac{\widehat{\mathbf{Var}}_\lambda}{\text{median}_{\lambda' \in \mathcal{R}}(\widehat{\mathbf{Var}}_{\lambda'})}$$

This puts the two estimates more or less on the same scale. Note we do not care about the overall scale because we only want to estimate the *shape* of the curve. Therefore, we estimate the shape of the mean squared error curve by

$$\widehat{MSE}(\lambda) = \widehat{bias}_\lambda^{2*} + \widehat{\mathbf{Var}}_\lambda^*,$$

and λ_{best} is estimated by $\widehat{\lambda} = \arg\min_{\lambda \in \mathcal{R}} \widehat{MSE}(\lambda)$. Our proposed method for choosing λ is formally detailed below in Algorithm 8.2. This method can easily be incorporated into Algorithm 8.1. The null distribution only has to be simulated once, giving the Type I error for each rejection region.

We provide some numerical results under the following set up. We generated normal random variables for $m = 1000$ genes and $n = 40$ samples, say x_{ij} $i = 1, \ldots, 1000$ $j =$

Algorithm 8.2: Automatically Choosing λ in Microarray Data

1. For some range of λ, say $\mathcal{R} = \{0, 0.05, 0.10, \ldots, 0.95\}$, calculate $\widehat{pFDR}_\lambda(t)$ as in Section 8.3.

2. For each $\lambda \in \mathcal{R}$, estimate the squared bias of $\widehat{pFDR}_\lambda(t)$ by

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(t) - \min_{\lambda\in\mathcal{R}} \widehat{pFDR}_\lambda(t)\right)^2,$$

$$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{\text{median}_{\lambda'\in\mathcal{R}}(\widehat{bias}_{\lambda'}^2)}.$$

3. Also, for each $\lambda \in \mathcal{R}$, estimate the variance of $\widehat{pFDR}_\lambda(t)$ by

$$\widehat{\mathbf{Var}}_\lambda = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{pFDR}_\lambda^{(-i)}(t) - \widehat{pFDR}_\lambda(t)\right)^2,$$

$$\widehat{\mathbf{Var}}_\lambda^* = \frac{\widehat{\mathbf{Var}}_\lambda}{\text{median}_{\lambda'\in\mathcal{R}}(\widehat{\mathbf{Var}}_{\lambda'})},$$

where $\widehat{pFDR}_\lambda^{(-i)}(t),\ i = 1, \ldots, n$, are jackknifed versions of $\widehat{pFDR}_\lambda(t)$ taken over the n independent aspects of the data.

4. For each $\lambda \in \mathcal{R}$, estimate its respective mean squared error curve by:

$$\widehat{MSE}(\lambda) = \widehat{bias}_\lambda^{2*} + \widehat{\mathbf{Var}}_\lambda^*.$$

5. Set $\widehat{\lambda} = \arg\min_{\lambda\in\mathcal{R}} \widehat{MSE}(\lambda)$. Our overall estimate of $pFDR(t)$ is

$$\widehat{pFDR}(t) = \widehat{pFDR}_{\widehat{\lambda}}(t).$$

6. For the estimates $\widehat{FDR}_\lambda(\cdot)$, $t_\alpha[\widehat{FDR}_\lambda]$, $\widehat{\alpha}_{FDR,\lambda}(\cdot)$, and $\widehat{\mathrm{q}}_\lambda(\cdot)$ the procedures are adjusted analogously from those presented in Chapter 7.

$1, \ldots, 40$. In the notation of Section 8.5, we have

$$x_{ij} \sim N(0,1) + u \cdot I(i \leq m_0 \ \& \ j \geq 21).$$

Each block of 50 genes has correlation 0.1, and the parameters u and m_0 varied over different simulations. The first 20 observations were designated as group 1, and the second 20 as group 2. A two-sample t-statistic was formed for each gene, and any t-statistic exceeding 2 in absolute value was rejected.

For each set of parameters u and m_0, we generated 100 data sets and performed the procedure on each. Table 8.2 displays the results. We list both λ_{best}, the mean and median $\widehat{\lambda}$, and their respective true mean squared errors. We also used $\mathcal{R} = \{0, 0.05, 0.10, \ldots, 0.95\}$. It can be seen that even in the worst cases, the optimal MSE and the MSE's corresponding to the observed median and mean $\widehat{\lambda}$ are not that different. Figure 8.3 shows the MSE curve and a histogram of the 100 $\widehat{\lambda}$'s for the case where $m_0 = 200$ and $u = 0.3$.

Applying the method for choosing λ to the Rieger et al. data, we find that $\widehat{\lambda} = 0.15$. Therefore, our overall estimate of $pFDR(\{x : |x| \geq 2\})$ is $\widehat{pFDR}(\{x : |x| \geq 2\}) = 7.56\%$. This has only slightly greater bias than in the Rieger et al. data example where we set $\lambda = 0.7$, but the variance has been reduced significantly.

8.8 Modeling Versus Hypothesis Testing

We have used symmetric rejection regions on the DNA microarray data. However, asymmetric rejection regions are more useful because the change in gene expression is not necessarily equally likely to be positive or negative. Tusher et al. (2001) provide a method for choosing asymmetric cut-points based on a rule involving a quantile-quantile plot of the original statistics versus the simulated null statistics. Therefore, their rejection regions have the form $(-\infty, c_1] \cup [c_2, \infty)$ for data dependent c_1 and c_2. Another form of rejection regions that has been used is $\Gamma = \{x : \widehat{\mathbf{Pr}}(H = 0|X = x) \leq r\}$ for some pre-chosen r. The posterior probabilities are estimated from a non-parametric empirical Bayes model in Efron et al. (2001). It can be shown that this is equivalent to a likelihood ratio based rejection region, where the likelihood ratio is estimated non-parametrically.

Besides traditional multiple hypothesis testing, we have considered the pFDR in the context of both Bayesian posterior probabilities and classification. Recall that in modeling, one uses the data to fit parameters in some optimal way. In classification theory one has to

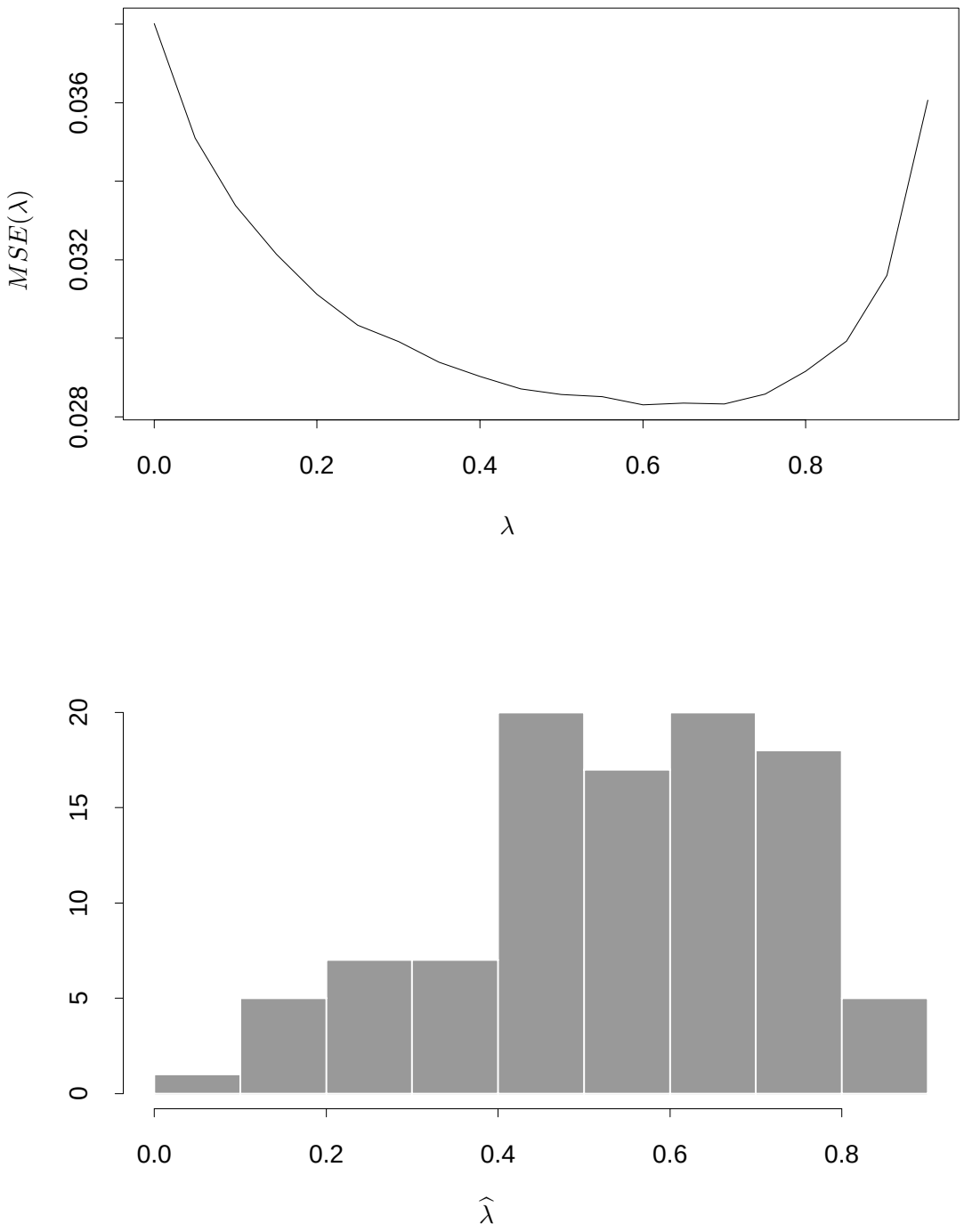


Figure 8.3: $m_0 = 200$ and $u = 0.3$. Upper panel: The mean squared error curve as a function of λ. Lower panel: Histogram of the 100 observed $\widehat{\lambda}$.

use observed data to fit a model that can predict future observations. The misclassification rate, however, is estimated from a new, independent set of data. Likewise, if one is going to use a Bayesian model to estimate $\{\mathcal{B}_r\}$, the pFDR must be calculated using a new independent set of data. In other words, estimating $\mathbf{Pr}(H = 0|X \in \Gamma)$ must be treated differently than estimating $\mathbf{Pr}(H = 0|X = x)$. Efron et al. (2001) do both with one set of data, thereby calculating the pFDR incorrectly. Also, Tusher et al. (2001) use the data to adaptively form rejection regions, and then calculate the FDR on the same data, which can lead to the same anti-conservative bias.

Efron et al. (2001) calculate modified two-sample t-statistics for the data in Tusher et al. (2001). It is assumed that $X_i|H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$, as we assumed in Chapters 2 and 3. Moreover, null versions of the statistics are calculated. Using these null versions along with the observed statistics, a non-parametric estimate of

$$\mathbf{Pr}(H_i = 0|X_i = x_i) = \frac{\pi_0 \cdot f_0(x_i)}{\pi_0 \cdot f_0(x_i) + \pi_1 \cdot f_1(x_i)}$$

is calculated, which we will denote by $\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i)$. The sets $\{\mathcal{B}_r\}$ are then estimated by $\widehat{\mathcal{B}}_r = \{\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i) \leq r\}$.

Let $f = \pi_0 \cdot f_0 + \pi_1 \cdot f_1$. Efron et al. (2001) suggest reporting the $\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i) \leq 0.10$ as being significant genes. However, this is only a marginal statistic, and it doesn't take into account the multiple comparisons. Therefore, using an earlier version of this work (Storey 2001c), they say that by integrating over $\widehat{\mathcal{B}}_{0.10}$ with the density $\widehat{f}(\cdot|X \in \widehat{\mathcal{B}}_{0.10})$, they have successfully estimated $pFDR(\widehat{\mathcal{B}}_{0.10}) = \mathbf{Pr}(H = 0|X \in \widehat{\mathcal{B}}_{0.10})$.

Is this really a fair estimate of $pFDR(\widehat{\mathcal{B}}_{0.10})$? The answer is no, because one cannot perform modeling and hypothesis testing with one set of data. The statistics are *observed* and then "rejection regions" are *estimated* to best discriminate the null statistics from the observed statistics. One cannot estimate rejection regions in this manner and then calculate a hypothesis testing error measure. As an exaggerated example, this is similar to performing a two-sided test of the mean of a Normal random variable. The statistic is observed to be, say, negative. Since one-sided rejection regions (to the left) now best discriminate the observed statistic from the null distribution, these regions are used. Clearly, this estimates the Type I error to be half of its true size. Therefore, the pFDR calculated from these rejection regions is going to be half of its true size. This same bias is likely to occur in the most extreme statistics in Efron et al. (2001). And the most extreme statistics happen to

be among the most important.

Therefore, it is not correct to report $pFDR(\widehat{\mathcal{B}}_{0.10})$. One cannot do modeling and hypothesis testing simultaneously. It is well-known that an estimate of $BE(\widehat{\mathcal{B}}_{0.10})$ (from Section 2.4) obtained in the same way could be a significant underestimate. And we have seen that the Bayes Error and the pFDR are functionally related. On the other hand, with so many statistics observed, it is tempting to try to learn something about the alternative distribution. One potential solution to this problem is to split the statistics into two groups. The rejection regions estimated from one group are applied to the other to calculate q-values, and vice versa.

Moreover, Efron et al. (2001) suggest reporting both $\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i)$ and $pFDR(\widehat{\mathcal{B}}_{0.10})$. They say $\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i)$ should be reported because it gives more specific information about the significance of the gene. In this work, we argue that the information contained in $\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i)$ is not useful because the interpretation of this quantity changes radically depending on the total number considered. We suggest instead using the q-value as the quantity to report for each statistic. In the context of hypothesis testing and decision making, one would never call one gene significant while not calling another gene significant that has a more extreme statistic. The q-value takes this into account, thereby incorporating the simultaneous inference *and* the statistic-specific significance measure into one quantity that has a clear interpretation. Moreover, it is easy to see that there are cases where the "weak dependence" assumption is met and the q-values are robust against dependence but the $\widehat{\mathbf{Pr}}(H_i = 0|X_i = x_i)$ can be greatly affected by dependence.

Chapter 9

Concluding Remarks

This work has been concerned with assessing the false discovery rates, FDR and pFDR, when testing thousands of hypotheses simultaneously. We have introduced and investigated the statistical properties of the pFDR. These measures are especially useful for the exploratory nature of testing for differential gene expression in DNA microarrays. In particular the pFDR has a sound interpretation, and the q-value is the quantity that should be reported for each gene in the microarray problem. We mainly dealt with the case of dependent statistics by appealing to asymptotic arguments. Often asymptotic results are unrealistic, but in the microarray problem the number of tests is very large and we have seen the asymptotic results work for these data.

Of obvious future interest is to further understand the role of dependence in false discovery rates, particularly in the DNA microarray problem. Also, it is worth further studying the optimality properties of our estimates $\widehat{FDR}_\lambda$ and $\widehat{pFDR}_\lambda$ among all such conservatively consistent estimates, as well as the finer behavior of these estimates, such as the rates of convergence to their deterministic limits.

In this work, we have basically made no assumptions about the statistics, except that the null p-values are uniformly distributed. If more information about the alternative distribution is known, then more precise methods will likely emerge. For example, if we know that the power to Type I error curve is concave, then consistent estimates of this curve exist that are more robust than the empirical distribution function that we used.

With the shift in the focus of biology to genomics, it is worth mentioning a few other applications of our false discovery rate methodology. For example, one can apply false discovery rates to association studies where thousands of SNP markers are typed on each

individual. There exist methods to calculate p-values for sequence alignment, but not much attention has been paid to the multiple comparisons problem. (A single sequence is usually compared to thousands of other sequences in a data base.) The goals of sequence alignment are also exploratory, and the dependence is clumpy, so this methodology is well suited to that problem as well. In general, the available genome sequences allow one to ask biologically interesting questions and narrow down the search for the answers by performing some statistical analysis of the genome data. False discovery rates are well suited for this data mining task, giving the biologist a meaningful error measure on their findings.

To conclude this work, we mention two additional useful applications of the ideas we have discussed. The first is the estimate $\widehat{\pi}_0$. In the DNA microarray problem, $\widehat{\pi}_0$ is a useful measure in itself – it estimates the proportion of genes that are not different between the biological samples. Rather than clustering the data and seeing that gene patterns look different between the samples, $1 - \widehat{\pi}_0$ gives a rigorous measurement of how molecularly different the samples are. Moreover, by using the simulated null statistics, on can easily calculate a p-value for $\widehat{\pi}_0$ in testing that it is different than 1. The second application is that power calculations in microarrays can easily be done. If $X_i \sim \pi_0 \cdot F_0 + \pi_1 \cdot F_1$, then the power over Γ is $\mathbf{Pr}(X_i \in \Gamma | H = 1) = [\mathbf{Pr}(X_i \in \Gamma) - \pi_0 \mathbf{Pr}(X_i \in \Gamma | H = 0)]/\pi_1$. Using our methodology, one can estimate each quantity on the right hand side of this equation.

Bibliography

Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. G., Sabet H., Tran T., Yu X., Powell J. I., Yang L. M., Marti G. E., Moore T., Hudson J., Lu L. S., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T. C., Weisenburger D. D., Armitage J. O., Warnke R., Levy R., Wilson W., Grever M. R., Byrd J. C., Botstein D., Brown P. O. & Staudt L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**: 503–511.

Benjamini Y. & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Stat. Soc., Ser. B.* **85**: 289–300.

Benjamini Y. & Hochberg Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics, *J. Edu. and Behav. Stat.* **25**: 60–83.

Benjamini Y. & Liu W. (1999). A step-down multiple hypothesis procedure that controls the false discovery rate under independence, *J. Stat. Plan. and Inference* **82**: 163–170.

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

Billingsley P. (1968). *Weak Convergence of Probability Measures*, John Wiley & Sons, New York.

Brown P. O. & Botstein D. (1999). Exploring the new world of the genome with DNA microarrays, *Nature Genetics* **21**: 33–37.

Cherkassky V. S. & Mulier F. M. (1998). *Learning from Data : Concepts, Theory, and Methods*, Adaptive and Learning Systems for Signal Processing, Communications and Control Series, Wiley-Interscience, New York.

Dudoit S., Yang Y., Callow M. & Speed T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* **12**: 111–139.

Efron B. & Tibshirani R. (1998). The problem of regions, *Ann. Stat.* **26**: 1687–1718.

Efron B. & Tibshirani R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Efron B., Tibshirani R., Storey J. D. & Tusher V. (2001). Empirical Bayes analysis of a microarray experiment, *J. Amer. Stat. Assoc.* **96**: 1151–1160.

Eisen M. B., Spellman P. T., Brown P. O. & Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns, *PNAS* **95**: 14863–14868.

Feingold E., Brown P. O. & Siegmund D. (1993). Gaussian models for genetic-linkage analysis using complete high-resolution maps of identity by descent, *Am. J. Hum. Gen.* **53**: 234–251.

Genovese C. & Wasserman L. (2001). Operating characteristics and extensions of the FDR procedure. Technical Report, Department of Statistics, Carnegie Mellon University.

Hochberg Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**: 800–803.

Holm S. (1979). A simple sequentially rejective multiple test procedure, *Scand. J. Stat.* **6**: 65–70.

Karlin S. & Altschul S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *PNAS* **87**: 2264–2268.

Lehmann E. L. (1986). *Testing Statistical Hypotheses*, second edn, Springer-Verlag, New York.

Lockhart D. J., Dong H., Byrne M. C., Follettie M. T., Gallo M. V., Chee M. S., Mittmann M., Wang C., Kobayashi M., Horton H. & Brown E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnol.* **14**: 1675–1680.

Morton N. E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**: 277–318.

Schena M., Shalon D., Davis R. W. & Brown P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**: 467–470.

Shaffer J. (1995). Multiple hypothesis testing, *Annual Rev. Psych.* **46**: 561–584.

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

Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. & Futcher B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Bio. of the Cell* **9**: 3273–3297.

Storey J. D. (2001a). A direct approach to false discovery rates. *J. Roy. Stat. Soc. Ser. B*, in press. Available at `http://www-stat.stanford.edu/~jstorey/`.

Storey J. D. (2001b). The positive false discovery rate: A Bayesian interpretation and the q-value. Submitted. Available at `http://www-stat.stanford.edu/~jstorey/`.

Storey J. D. (2001c). The positive false discovery rate: A Bayesian interpretation and the q-value. Technical Report 2001-12, Department of Statistics, Stanford University.

Storey J. D., Taylor J. E. & Siegmund D. (2002). A unified estimation approach to false discovery rates. Submitted. Available at `http://www-stat.stanford.edu/~jstorey/`.

Storey J. D. & Tibshirani R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Submitted. Available at `http://www-stat.stanford.edu/~jstorey/`.

Tusher V., Tibshirani R. & Chu C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proc. Natl. Acad. Sci.* **98**: 5116–5121.

Velculescu V. E., Zhang L., Vogelstein B. & Kinzler K. W. (1995). Serial analysis of gene expression, *Science* **270**: 484–487.

Weller J. I., Song J. Z., Heyen D. W., Lewin H. A. & Ron M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits, *Genetics* **150**: 1699–1706.

Westfall P. H. & Young S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley.

Worsley K. J., Marrett S., Neelin P., Vandal A. C., Friston K. J. & Evans A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation, *Human Brain Mapping* **4**: 58–73.

Yang Y. H., Buckley M. J., Dudoit S. & Speed T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *J. Comp. Graph. Stat.*, in press.

Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J. & Speed T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research* **30**: e15.

Yekutieli D. & Benjamini Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, *J. Stat. Plan. and Inference* **82**: 171–196.

Zaykin D. V., Young S. S. & Westfall P. H. (1998). Using the false discovery approach in the genetic dissection of complex traits: A response to weller et al., *Genetics* **150**: 1917–1918.