ESTIMATING THE POSITIVE FALSE DISCOVERY RATE UNDER DEPENDENCE, WITH APPLICATIONS TO DNA MICROARRAYS

by

JOHN D. STOREY
ROBERT TIBSHIRANI

Technical Report No. 2001-28
October 2001

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305

ESTIMATING THE POSITIVE FALSE DISCOVERY RATE UNDER DEPENDENCE, WITH APPLICATIONS TO DNA MICROARRAYS

by

John D. Storey
Department of Statistics
Stanford University

Robert Tibshirani
Department of Statistics
Stanford University

Technical Report No. 2001-28
October 2001

This research was supported in part by National Institute of Health grant 2R01 CA72028 and National Science Foundation grants DMS-9971405 and Graduate Research Fellowship

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305

http://www-stat.stanford.edu

Estimating the positive False Discovery Rate Under Dependence, with Applications to DNA Microarrays

John D. Storey *
Robert Tibshirani †

October 22, 2001

Abstract

When conducting multiple hypothesis tests, it is important to assess the number of false positives in some fashion. One useful error measure is the positive False Discovery Rate (pFDR). We show how to estimate the pFDR when general dependence between the hypotheses exists. This can be done using general statistics, not necessarily p-values, where the Type I error rate for a given rejection region may not even be known. We apply the proposed methodology to the problem of detecting differential gene expression in replicated DNA microarray experiments, where unknown dependence is likely to occur.

Keywords: multiple hypothesis testing, DNA microarrays, multiple comparisons, simultaneous inference.

*Department of Statistics, Stanford University, Stanford, CA 94305. Email: jstorey@stat.stanford.edu.
†Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305. Email: tibs@stat.stanford.edu.

1 Introduction

When testing multiple hypotheses, one must guard against Type I errors. Traditionally, the Family Wise Error Rate (FWER) has been controlled using sequential p-value methods. The FWER is defined to be the probability of making one or more Type I error among all hypotheses tested. (See Shaffer (1995) for a review of these pre-FDR multiple hypothesis testing methods.) In a seminal paper, Benjamini and Hochberg (1995) suggested controlling a new quantity called the False Discovery Rate (FDR), which is defined to be the expected proportion of Type I errors among rejected hypotheses.

More specifically, consider Table 1 displaying the various outcomes when testing m hypotheses:

Table 1: *Possible Outcomes from m Hypothesis Tests*

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

For example, V is the number of false positives (Type I errors), while R is the total number of hypotheses rejected. The FWER is defined to be $\mathbf{Pr}(V \geq 1)$, and the FDR is loosely defined to be

$$\mathbf{E}\left[\frac{V}{V+S}\right] = \mathbf{E}\left[\frac{V}{R}\right]. \tag{1}$$

Benjamini and Hochberg (1995) and Benjamini and Liu (1999) provide sequential p-value methods to control this quantity. The FDR offers a much less strict multiple testing criterion than the FWER, and therefore leads to an increase in power.

One must deal with the case where $R = 0$ in which V/R is undefined. Therefore, the precise definition of the FDR that Benjamini and Hochberg (1995) use is

$$\mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]\mathbf{Pr}(R > 0). \tag{2}$$

Two other possible definitions that circumvent the $R = 0$ problem are

$$\mathbf{E}\left[\frac{V}{R}\middle| R > 0\right] \text{ and } \frac{\mathbf{E}[V]}{\mathbf{E}[R]}. \tag{3}$$

Neither of these definitions is used by Benjamini and Hochberg (1995) because the two quantities are identically 1 when all null hypotheses are true, and therefore they cannot be controlled by a sequential p-value method.

In Storey (2001a), we define the *positive False Discovery Rate* (pFDR) to be

$$pFDR = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]. \tag{4}$$

The additional term *positive* refers to the fact that we are only interested in estimating an error rate where positive findings have occurred (as is the case in single hypothesis testing). For example, suppose a researcher controls the BH-FDR at level α, and rejects at least one hypothesis. Conditional that positive findings occur, the FDR has really only been controlled at level $\alpha/\mathbf{Pr}(R > 0)$. When m is small or dependence exists, $\mathbf{Pr}(R > 0)$ can be less than one, resulting in control of a misleading error measure much larger than α. See Weller et al. (1998) for one such case.

Therefore, we find the BH-FDR inconvenient in that when a researcher has made a discovery, he has controlled an error measure covering a much wider range of cases (i.e., when no discoveries are made) than what he is interested in. The error measure he wants is the pFDR because it is conditioned on and is an error measure of his exact situation.

We have argued that traditional sequential p-value methods are not appropriate for false discovery rates. In Storey (2001a,b) and in this paper, we take a different, more direct approach: we directly estimate $pFDR$ for a fixed rejection region. A sequential p-value method does the opposite: it estimates the rejection region needed to give a certain error measure on average. Therefore, the fact that $pFDR = 1$ when all null hypotheses are true is not a problem using our approach. See Storey (2001b) for a thorough treatment of the advantages and flexibility of this approach.

As it turns out, estimating the pFDR when the tests are independent is a fairly straightforward task. A review of this case will be given in Section 4. The objective of this paper is to estimate the pFDR when there is arbitrary dependence between the tests. This situation has growing importance, especially since large data sets with many dependent variables are becoming more abundant. For example, in the burgeoning field of statistical genomics, one is forced to test hypotheses on thousands of dependent genes. In this paper, we show how to apply our methodology to DNA microarrays, a new biotechnology that allows the simultaneous measurement of the expression levels of thousands of genes.

We propose a non-parametric method to cover all levels of dependence, with modifications being possible if some parametric assumptions are made.

Example 1 *DNA Microarrays*

Here is an example and a capsule summary of our main proposal. Rieger (2001, unpublished) analyzes DNA expression data on 3000 genes from a study of the effects of ionizing radiation, comparing normal patients to radiation sensitive patients. (For background on DNA expression data see section 9.) There are 15 samples in group 1 (normal) and 13 in group 2 (radiation sensitive). Figure 1 is a histogram of the 3000 two-sample t-statistics from the genes. They range from -4.54

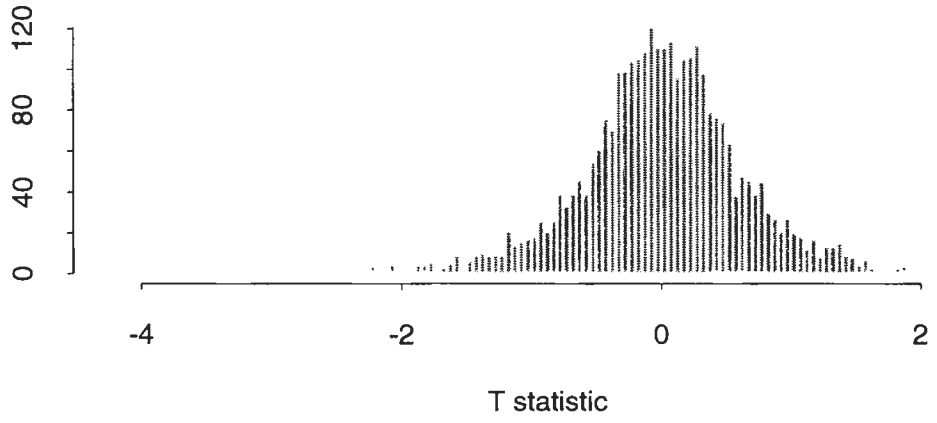

Figure 1: *Histogram of 3000 t-statistics, from DNA microarray example.*

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 pFDR among these 146 genes? To assess this, we do a random permutation of the sample labels

$$(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2),$$

recompute the t-statistics and count how many exceed ± 2. Doing this for 100 permutations, we find the average number is 12.3. Thus a simple estimate of the pFDR is $12.3/146 = 8.44\%$.

Now it turns out that this simple estimate tends to be biased upward. The reason is clear from Table 1: the permutations make all m genes null, but in our data only a proportion $\pi_0 = m_0/m$ are null. Hence to improve our estimate of the pFDR, we multiply it by an estimate of π_0. To obtain the latter, we look at small values of the t-statistic (in absolute value), where null statistics are much more abundant than alternatives. Looking for example at t-statistics below 0.15 in absolute value, we find that 668 of the observed t-statistics fall in that range, while on the average 750 of the t-statistics from the permutations fall in that range. Hence our estimate of π_0 is $\widehat{\pi}_0 = 668/750 \approx 0.890$. Finally, our revised estimate of the pFDR is $0.890 \cdot 0.0844 = 7.52\%$ The above computation is an example of our main proposal given on page 6.

What if we try a simpler estimate, assuming independence and a marginal t-distribution for the statistics? If T_{26} has a t distribution on 26 degrees of freedom, then $\Pr(|T_{26}| > 2) = 0.056$. Hence we would expect $3000 \cdot 0.056 = 168$ false positive genes, giving a $pFDR$ of $168/146 \gg 1$. It turns out that the marginal t-distribution, not independence, is the problem. A total of 1232 out of the $3000 \cdot 100$ permutation statistics exceeded 2 in absolute value, a proportion of 0.0046. This gives an estimate of $0.0041 \cdot 3000 = 12.3$ false positive genes, the same value as above.

Tusher et al. (2001) propose the "SAM" (Significance Analysis of Microarrays) procedure. In the case of unpaired samples, as above, SAM is essentially a method for choosing cutpoints like the values ± 2. These cutpoints can be asymmetric around zero. Having chosen the cutpoints, they estimate the pFDR as above, except that they use $\widehat{\pi}_0 = 1$.

This paper is organized as follows. Section 2 presents our proposal for the dependent case, and Section 3 makes some comparisons between our method and existing methods based on the Rieger DNA microarray data set. As part of the motivation and theoretical justification for our method, Section 4 reviews the independence case. Theoretical properties of the method under dependence are presented in Section 5, and we discuss a simulation study in section 6. Section 8 proposes and investigates a method for choosing a tuning parameter λ. Section 9 discusses how and why this methodology can be applied to DNA microarrays in general.

2 The Proposed Method for Estimation of pFDR

We assume we are testing m hypotheses using the statistics $T_1, \ldots, T_m$. We also assume that the null hypothesis is simple, and it is the same for all tests. The alternative hypothesis can be simple, or it can be composite in the sense that the alternative is different for each test, but comes from some common family of alternatives. Since the same null and alternative hypotheses exist for each test, each one is based on the same rejection region. The dependence between the T_i can be arbitrary, regardless of whether they follow the null or alternative distributions.

For generality, we denote the rejection regions by the set $\{\Gamma\}$. (Note that the set of possible rejection regions is nested.) We provide an estimate of the pFDR over the fixed rejection region Γ rather than estimating the rejection region for a fixed pFDR (as in Storey (2001b)). We make the important assumption that null versions of the statistics can be simulated; denote these simulated null statistics by $T_1^0, \ldots, T_m^0$. An example where these are available is Example 1.

The task of estimating the pFDR when strong dependence exists is difficult because we can only observe $R(\Gamma) = \#\{T_i \in \Gamma\}$ along with $R^0(\Gamma) = \#\{T_i^0 \in \Gamma\}$ (Also, note we can observe $W(\Gamma) = m - R(\Gamma)$. See Table 1.) In the work of Westfall and Young (1993), using the simulated null $T_1^0, \ldots, T_m^0$ turns out to be very important in preserving the dependence structure in calculating adjusted p-values. However, for the pFDR, the dependence structure obtained from $T_1^0, \ldots, T_m^0$ is not so useful, especially given that the pFDR involves sums of indicator variables.

Yekutieli and Benjamini (1999) attempt to use the $T_1^0, \ldots, T_m^0$ to capture the dependence structure of V and S in estimating the BH-FDR. However, upon close examination of their method, the $T_1^0, \ldots, T_m^0$ are more or less used to estimate the expected number of false positives when all hypotheses are null. Since $R^0(\Gamma) = V^0(\Gamma) + S^0(\Gamma)$ and the dependence structure of V and S can radically differ, we find it futile to capture the dependence through $R^0(\Gamma)$. We directly use $R^0(\Gamma)$

to calculate $\mathbf{E}[R^0(\Gamma)]$. In doing so, we are able to estimate the expected number of null hypotheses $\mathbf{E}[V(\Gamma)]$. This leads to a greater use of the data and a more powerful method, as will be seen.

We propose the same estimate of the pFDR regardless of the form of dependence, since under our assumptions the dependence structure is unknown. It will be shown that

$$\begin{aligned} \mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\right] &\approx \frac{\mathbf{E}[V(\Gamma)]}{R(\Gamma)} \\ &= \pi_0 \cdot \frac{\mathbf{E}[R^0(\Gamma)]}{R(\Gamma)} \end{aligned} \tag{5}$$

We calculate $\mathbf{E}[R^0(\Gamma)]$ from simulations of null statistics; in order to estimate π_0, we form the ratio

$$\widehat{\pi}_0 = \frac{W(\Gamma')}{\mathbf{E}[W^0(\Gamma')]} \tag{6}$$

for some well chosen rejection region Γ'. (We will show how this region can be optimally chosen. In Section 4 and thereafter, Γ' will be treated more formally.) Recalling that $W(\Gamma') = m - R(\Gamma')$, we can see that $W(\Gamma')$ provides a count of statistics that do not fall into the rejection region Γ'. If Γ' is chosen well, then $W(\Gamma')$ will mostly consist of null statistics, and $\mathbf{E}[W(\Gamma')] \approx m_0/m \cdot \mathbf{E}[W^0(\Gamma')]$, so $\widehat{\pi}_0$ provides a good estimate of π_0. Finally, the estimate of $pFDR(\Gamma)$ is

$$\widehat{pFDR}_\lambda(\Gamma) = \frac{W(\Gamma')}{\mathbf{E}[W^0(\Gamma')]} \cdot \frac{\mathbf{E}[R^0(\Gamma)]}{R(\Gamma)}. \tag{7}$$

The method is detailed in Algorithm 1 below.

We make the following remarks:

- When expectations are taken, the $\approx$ in (5) turns out to be $\leq$, and so our estimate is greater than the pFDR in expectation for all π_0. This is equivalent to "strong control." The estimate becomes more conservative as the dependence is stronger. It can be seen from the simulations (Section 6) that even in the worst cases we are only conservative by 13%.

- This procedure is equivalent to what we proposed in Storey (2001b) for the independence case. This will become clear when we take a more theoretical look at the method in Sections 4 and 5.

- We only use the simulated null statistics $T_1^0, \ldots, T_m^0$ to calculate $\mathbf{E}[R^0(\Gamma)]$ and $\mathbf{E}[W^0(\Gamma')]$. In fact, we do not even consider the Monte Carlo error in this calculation because it is user specified through B.

- Depending on the level of dependence, we actually estimate a quantity that is an upper bound for $pFDR(\Gamma)$. See Section 5 for a rigorous explanation.

Algorithm 1

Estimation and Inference of pFDR

1. Let Γ be the rejection region of interest, and let Γ' be a well chosen rejection region so that its complement is likely to contain mostly null statistics. (See Sections 4 and 5 for a rigorous treatment of Γ', and Section 8 for an automatic method for choosing Γ'.)

2. Simulate the null statistics for B iterations to obtain sets $T_1^{0b}, \ldots, T_m^{0b}$ for $b = 1, \ldots, B$.

3. Calculate $\mathbf{E}[R^0(\Gamma)]$ by

$$\mathbf{E}[R^0(\Gamma)] = \frac{1}{B}\sum_{b=1}^{B} R^{0b}(\Gamma), \tag{8}$$

   where $R^{0b}(\Gamma) = \#\{T_i^{0b} \in \Gamma\}$.

4. Estimate π_0 by

$$\widehat{\pi}_0 = \frac{W(\Gamma')}{\mathbf{E}[W^0(\Gamma')]}, \tag{9}$$

   where $\mathbf{E}[W^0(\Gamma')] = m - \mathbf{E}[R^0(\Gamma')]$ is calculated similarly to the previous step but with Γ'.

5. Estimate $pFDR(\Gamma)$ by

$$\widehat{pFDR}_{\Gamma'}(\Gamma) = \frac{\widehat{\pi}_0 \cdot \mathbf{E}[R^0(\Gamma)]}{R(\Gamma)}. \tag{10}$$

- We have shown how to estimate $pFDR(\Gamma)$ using a fixed Γ'. See Section 8 on how to choose the best Γ'.

- In Example 1 we applied Algorithm 1 to the DNA microarray data and obtained $\widehat{pFDR}_{\Gamma'}(\Gamma) = 7.52\%$ for $\Gamma = \{t : |t| \geq 2\}$ and $\Gamma' = \{t : |t| \geq 0.15\}$.

3 A Comparison to Existing Methods

We now compare the results obtained in Example 1 to what is obtained by using other methods. We reported $\widehat{pFDR} = 7.52\%$ when rejecting all t-statistics beyond ± 2, for a total of **146 significant genes**. With the method described in Tusher et al. (2001) (i.e., with $\widehat{\pi}_0 = 1$), the reported pFDR would have been 8.44%. The rejection region that gives $\widehat{pFDR} = 7.52\%$ using this method rejects only 87 genes.

Controlling the FDR at level 7.52% with the Benjamini and Hochberg (1995) method results

in 87 significant genes, but this method assumes independence or positive regression dependence, neither of which we are guaranteed. Thus, if we make the correction for general dependence given in Benjamini and Yekutieli (2001), we reject only one gene, controlling the FDR at level $7.52\%/\log(3000) = 1.0\%$. Using the methodology of Yekutieli and Benjamini (1999), we get an estimate of $FDR(\{|t| \geq 2\})$ as 8.31%. For this method, the rejection region that estimates FDR at 7.52% rejects 91 genes. It theoretically follows that $FDR(\Gamma) \leq pFDR(\Gamma)$; therefore, we have given these FDR controlling procedures a slight edge, even though our method still outperforms them.

Dudoit et al. (2001) suggest controlling the FWER when detecting differential gene expression by using the methodology of Westfall and Young (1993). Controlling the FWER at level 7.52% rejects 4 genes. In order to reject 146 genes, we would have to control the FWER at level 99.1%.

4 The pFDR Under Independence

In this section we present methodology for estimating the pFDR when the hypothesis tests are independent, based on Storey (2001a) and Storey (2001b). Our approach in the dependence case is very much related to the independence case, so we present this first. For example, we formed the ratio $\widehat{\pi}_0 \mathbf{E}[R^0]/R$ as our estimate, whereas, it seems one would really need something like $\widehat{\pi}_0 \cdot \mathbf{E}[R^0/R]$. By examining properties of the pFDR under independence, it becomes clearer why our method works.

We assume that there exist m hypotheses with independent statistics $T_1, \ldots, T_m$. Let $H_i = 0$ if null hypothesis i is true, and $H_i = 1$ if it is false, $i = 1, \ldots, m$. We initially assume that each test is simple versus simple, and also that $T_i|H_i = 0 \sim f_0$ and $T_i|H_i = 1 \sim f_1$ for densities f_0 and f_1, $i = 1, \ldots, m$. Since the statistics are identically distributed under the null hypothesis, we have a nested set of rejection regions $\{\Gamma_\alpha\}$ for $\alpha \in [0, 1]$, and a single rejection region is used for each test. For generality, we have indexed $\{\Gamma_\alpha\}$ by α, where α is the Type I error rate for any single test. That is, $\alpha = \mathbf{Pr}(T \in \Gamma_\alpha | H = 0)$. Note that hypothesis tests are derived so that the set of rejection regions is nested: $\alpha < \alpha'$ implies $\Gamma_\alpha \subset \Gamma_{\alpha'}$.

In estimating $pFDR(\Gamma)$ we use the following theorem shown in Storey (2001a).

Theorem 1 *Suppose m identical hypothesis tests are performed with the i.i.d. statistics $T_1, \ldots, T_m$ and rejection region Γ. Also suppose that m_0 null hypotheses are true, and let $\pi_0 = m_0/m$. Then*

$$pFDR(\Gamma) = \frac{\pi_0 \cdot \mathbf{Pr}(T \in \Gamma | H = 0)}{\mathbf{Pr}(T \in \Gamma)}, \tag{11}$$

where $\mathbf{Pr}(T \in \Gamma) = \pi_0 \cdot \mathbf{Pr}(T \in \Gamma | H = 0) + (1 - \pi_0) \cdot \mathbf{Pr}(T \in \Gamma | H = 1)$.

This gives us a very simple form for $pFDR(\Gamma_\alpha)$ independent of m, and makes the estimation problem much more tractable. This result also holds for a simple null hypothesis versus a different

alternative hypothesis for each test. $\mathbf{Pr}(T \in \Gamma | H = 1)$ is just replaced by the average of the probabilities represented by each of the m_1 alternative hypotheses.

We estimate $pFDR(\Gamma_\alpha)$ by using a simple plug-in estimate that turns out to also be a maximum likelihood estimate. There are only two quantities we have to estimate because we know $\mathbf{Pr}(T \in \Gamma_\alpha | H = 0) = \alpha$. (If this is not known, it is calculated as in Algoritm 1.) We estimate $\mathbf{Pr}(T \in \Gamma_\alpha)$ by

$$\widehat{\mathbf{Pr}}(T \in \Gamma_\alpha) = \frac{\#\{T_i \in \Gamma_\alpha\}}{m} = \frac{R(\Gamma_\alpha)}{m}, \tag{12}$$

where we let $R(\Gamma_\alpha) = \#\{T_i \in \Gamma_\alpha\}$ for any α.

In order to estimate π_0, we use the following reasoning. First suppose we are rejecting based on p-values. Then most of the p-values near 1 should be null p-values. For some well chosen λ, we expect $(1-\lambda)m_0$ of the null p-values to fall in the interval $[\lambda, 1]$. Likewise for general statistics, we expect $(1-\lambda)m_0$ of the null statistics to fall into the region Γ_λ^c, where Γ_λ^c is the complement of Γ_λ. Therefore, our estimate of π_0 is

$$\widehat{\pi}_0 = \frac{\#\{T_i \in \Gamma_\lambda^c\}}{(1-\lambda)m} = \frac{W(\Gamma_\lambda)}{(1-\lambda)m}. \tag{13}$$

Note there is a trade-off between bias and variance in our choice of λ. Some alternative statistics may also fall into Γ_λ^c, so $\widehat{\pi}_0$ is conservatively biased. However, the larger we choose λ, the less conservative the bias. On the other hand, the larger we choose λ, the less statistics we expect to fall into Γ_λ^c, and hence the variance of $\widehat{\pi}_0$ becomes larger.

Therefore, our estimate of $pFDR(\Gamma_\alpha)$ is

$$\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{\pi}_0 \cdot \alpha}{\widehat{\mathbf{Pr}}(T \in \Gamma_\alpha)} = \frac{W(\Gamma_\lambda) \cdot \alpha}{(1-\lambda) \cdot R(\Gamma_\alpha)}. \tag{14}$$

In Storey (2001b), we choose λ by a bootstrap method in order to minimze the MSE. Since this requires the statistics to be independent, we show in Section 8 another way to choose λ that works under dependence. Storey (2001b) shows the following three facts about $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ under independence:

- $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ is a maximum likelihood estimate for

$$\frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha), \tag{15}$$

  where $\pi_1 = 1 - \pi_0$, and $g(\lambda)$ is the power resulting from the rejection region Γ_λ.

- $\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq pFDR(\Gamma_\alpha)$.

- $\lim_{m \to \infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \stackrel{a.s.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha) \geq pFDR(\Gamma_\alpha)$.

Our goal is to be conservative in our estimate, so these three facts are good properties of $\widehat{pFDR}_\lambda(\Gamma_\alpha)$. In fact, property 2 is analogous to offering "control" of the FDR in the Benjamini and Hochberg (1995) sense. Also note that "strong control" is provided in that the result holds for any π_0. (We think "control" is a misnomer, however, when taking our approach, since it is very different than the sequential p-value approach.) We try to show similar properties for our estimate under dependence, which we present in the following section. It can now be seen that this estimate is the same as that presented in the previous section, except here we assumed we knew the Type I error rates of the rejection regions, although these would be calculated the same if they were unknown.

5 Theoretical Properties Under Dependence

We now theoretically justify our proposed method under four cases of dependence (not necessarily mutually exclusive), showing similar properties as were shown in Storey (2001b) for the independence case. These four cases cover all levels of dependence. For Cases 1-3 we assume $V(\Gamma_\alpha)$ and $S(\Gamma_\alpha)$ are independent. In Case 4 $V(\Gamma_\alpha)$ and $S(\Gamma_\alpha)$ are dependent.

In Section 2 we estimated $pFDR(\Gamma_\alpha)$ by

$$\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda) \cdot \mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{E}[W^0(\Gamma_\lambda)] \cdot R(\Gamma_\alpha)}, \tag{16}$$

for some fixed Γ_λ. The basic idea behind our proposed method is to note the inequality:

$$\mathbf{E}\left[\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[V(\Gamma_\alpha)] + S(\Gamma_\alpha)}\right] \geq \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{V(\Gamma_\alpha) + S(\Gamma_\alpha)}\right] = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]. \tag{17}$$

which follows by Jensen's inequality. Therefore, a good conservative point estimate of $pFDR(\Gamma_\alpha)$ is

$$\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[V(\Gamma_\alpha)] + \widehat{S}(\Gamma_\alpha)}. \tag{18}$$

Since $\mathbf{E}(R(\Gamma_\alpha) - \mathbf{E}[V(\Gamma_\alpha)]) = \mathbf{E}[S(\Gamma_\alpha)]$, we take $\widehat{S}(\Gamma_\alpha) = R(\Gamma_\alpha) - \mathbf{E}[V(\Gamma_\alpha)]$. Therefore if $\mathbf{E}[V(\Gamma_\alpha)]$ were known we would estimate $pFDR(\Gamma_\alpha)$ by $\mathbf{E}[V(\Gamma_\alpha)]/R(\Gamma_\alpha)$. However, $\mathbf{E}[V(\Gamma_\alpha)]$ is unavailable. It follows that $\mathbf{E}[R^0(\Gamma_\alpha)] = m\alpha$ and $\mathbf{E}[V(\Gamma_\alpha)] = \mathbf{E}[V'(\Gamma_\alpha)] = \pi_0 \cdot m\alpha = \pi_0 \cdot \mathbf{E}[R^0(\Gamma_\alpha)]$. We can estimate π_0 exactly as in the previous section, and we arrive at our proposed estimate.

Since $\mathbf{E}[R^0(\Gamma_\alpha)] = m \cdot \alpha$ and $\mathbf{E}[W^0(\Gamma_\lambda)] = m \cdot (1 - \lambda)$ are calculated by a Monte Carlo integral, their error is not stochastic, but rather numerical. Therefore, for the remainder of the paper we will write

$$\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda) \cdot \alpha}{(1 - \lambda) \cdot R(\Gamma_\alpha)} \tag{19}$$

when we are considering $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ from a theoretical point of view.

Case 1: Large m and "loose dependence"

This is perhaps the most important case for applications, as it is the most likely situation to encounter in DNA microarrays (see Section 9). In Storey (2001a) the following theorem is shown:

Theorem 2 *Suppose the statistics $T_1, T_2, \ldots$ are such that*

$$\frac{\sum_{i=1}^{m} T_i(1-H_i)}{\sum_{i=1}^{m}(1-H_i)} \rightarrow \mathbf{E}(T|H=0),$$

$$\frac{\sum_{i=1}^{m} T_i H_i}{\sum_{i=1}^{m} H_i} \rightarrow \mathbf{E}(T|H=1)$$

in probability. Then for any Γ with $\mathbf{Pr}(T \in \Gamma) > 0$,

$$lim_{m\rightarrow\infty} pFDR_m(\Gamma) = \frac{\pi_0 \mathbf{Pr}(T \in \Gamma | H=0)}{\mathbf{Pr}(T \in \Gamma)}, \tag{20}$$

where $pFDR_m(\Gamma)$ is the pFDR resulting from the first m statistics.

The condition in the theorem is what we call "loose dependence." Basically it means that the average of null statistics converges in probability to its mean, as well as for the alternative statistics. This occurs when the dependence is such that it occurs in finite clumps of statistics, or if the dependence is ergodic; it can also occur in a variety of other situations. For example, the clumpy dependence example from Section 6 falls into the category of "loose dependence."

Therefore, in this case it makes sense to use the methodology prescribed for the independence case since $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ should more or less behave as if the statistics were independent (when m is large). It follows from the proof of the above theorem that we also get

$$\lim_{m\rightarrow\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \stackrel{a.s.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha) \geq pFDR(\Gamma_\alpha). \tag{21}$$

We will see from the next two cases when the finite sample result $\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq pFDR(\Gamma_\alpha)$ can also hold in this case.

Case 2: Independent null statistics

Without loss of generality, let $T_1, \ldots, T_{m_0}$ be the null statistics. When the null statistics are independent, we can write

$$\begin{aligned} pFDR(\Gamma_\alpha) &= \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\middle| R(\Gamma_\alpha) > 0\right] && (22)\\ &= \sum_{i=1}^{m_0} \mathbf{E}\left[\frac{1(T_i \in \Gamma_\alpha)}{R(\Gamma_\alpha)}\middle| R(\Gamma_\alpha) > 0\right] && (23)\\ &= m_0 \mathbf{Pr}(T_1 \in \Gamma_\alpha)\mathbf{E}\left[\frac{1}{R(\Gamma_\alpha)}\middle| R(\Gamma_\alpha) > 0, T_1 \in \Gamma_\alpha\right] && (24)\\ &= m_0\alpha\mathbf{E}\left[\frac{1}{R(\Gamma_\alpha) - 1(T_1 \in \Gamma_\alpha) + 1}\middle| R(\Gamma_\alpha) > 0, T_1 \in \Gamma_\alpha\right] && (25)\\ &\leq m_0\alpha\mathbf{E}\left[\frac{1}{R(\Gamma_\alpha)}\middle| R(\Gamma_\alpha) > 0\right] && (26) \end{aligned}$$

It is clear that $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ also provides a good estimate of (26). The quantity (26) is barely greater than $pFDR(\Gamma_\alpha)$, as was shown above. Also, $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ is only slightly more conservative when this form of dependence exists than when all statistics are independent. The following finite sample result easily follows.

Theorem 3 *When the null statistics are independent,* $\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq pFDR(\Gamma_\alpha)$.

Proof: Note that in $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ we should replace $R(\Gamma_\alpha)$ with $R(\Gamma_\alpha) \vee 1$ as in Storey (2001b), since otherwise $\widehat{pFDR}$ is undefined when $R(\Gamma_\alpha) = 0$. Moreover, we would never want to estimate the pFDR as zero, since our goal is to always be conservative in expectation. Also, in normal situations, the pFDR will never be zero. This is mainly a theoretical construct since it is doubtful that the pFDR will be reported for regions where no rejections have occurred, but we will use it here. Therefore,

$$\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq \mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)|R(\Gamma_\alpha) > 0] \cdot \mathbf{Pr}(R(\Gamma_\alpha) > 0). \tag{27}$$

Also, it follows that

$$\begin{aligned} \frac{(m - R(\Gamma_\lambda)) \cdot \alpha}{(1-\lambda) \cdot R(\Gamma_\alpha)} &\geq \frac{(m_0 - V(\Gamma_\lambda)) \cdot \alpha}{(1-\lambda) \cdot R(\Gamma_\alpha)} && (28)\\ &= \frac{(m_0 - V(\Gamma_\lambda)) \cdot \alpha}{(1-\lambda) \cdot (V(\Gamma_\alpha) + S(\Gamma_\alpha))}. && (29) \end{aligned}$$

Conditioning on $V(\Gamma_\lambda)$ and $S(\Gamma_\alpha)$, and using Jensen's inequality on $V(\Gamma_\alpha)$, we get

$$\mathbf{E}\left[\frac{(m_0 - V(\Gamma_\lambda)) \cdot \alpha}{(1-\lambda) \cdot (V(\Gamma_\alpha) + S(\Gamma_\alpha))}\middle| V(\Gamma_\lambda), S(\Gamma_\alpha)\right] \geq \frac{(m_0 - V(\Gamma_\lambda)) \cdot \alpha}{(1-\lambda) \cdot (\mathbf{E}[V(\Gamma_\alpha)|V(\Gamma_\lambda)] + S(\Gamma_\alpha))}. \tag{30}$$

By independence, $\mathbf{E}[V(\Gamma_\alpha)|V(\Gamma_\lambda)]$ is a non-decreasing function of $V(\Gamma_\lambda)$. Therefore, we get by Jensen's inequality on $V(\Gamma_\lambda)$

$$\mathbf{E}\left[\frac{(m_0 - V(\Gamma_\lambda)) \cdot \alpha}{(1-\lambda) \cdot (\mathbf{E}[V(\Gamma_\alpha)|V(\Gamma_\lambda)] + S(\Gamma_\alpha))} \middle| R(\Gamma_\alpha) > 0, S(\Gamma_\alpha)\right] \geq \frac{\mathbf{E}[V(\Gamma_\alpha)|R(\Gamma_\alpha) > 0]}{\mathbf{E}[V(\Gamma_\alpha)|R(\Gamma_\alpha) > 0] + S(\Gamma_\alpha)}. \tag{31}$$

It easily follows since $\mathbf{E}[V(\Gamma_\alpha)] = \mathbf{E}[V(\Gamma_\alpha)|R(\Gamma_\alpha) > 0] \cdot \mathbf{Pr}(R(\Gamma_\alpha) > 0)$ that

$$\frac{\mathbf{E}[V(\Gamma_\alpha)|R(\Gamma_\alpha) > 0]}{\mathbf{E}[V(\Gamma_\alpha)|R(\Gamma_\alpha) > 0] + S(\Gamma_\alpha)} \geq \frac{1}{\mathbf{Pr}(R(\Gamma_\alpha) > 0)} \cdot \frac{\mathbf{E}[V(\Gamma_\alpha)}{\mathbf{E}[V(\Gamma_\alpha)] + S(\Gamma_\alpha)} \tag{32}$$

By a third use of Jensen's inequality, it follows that

$$\mathbf{E}\left[\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[V(\Gamma_\alpha)] + S(\Gamma_\alpha)} \middle| R(\Gamma_\alpha) > 0\right] \geq \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{V(\Gamma_\alpha) + S(\Gamma_\alpha)} \middle| R(\Gamma_\alpha) > 0\right] = pFDR(\Gamma_\alpha). \tag{33}$$

Putting all of this together we get

$$\begin{aligned}
\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] &\geq \mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)|R(\Gamma_\alpha) > 0] \cdot \mathbf{Pr}(R(\Gamma_\alpha) > 0) \\
&\geq \frac{1}{\mathbf{Pr}(R(\Gamma_\alpha) > 0)} \cdot \mathbf{E}\left[\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[V(\Gamma_\alpha)] + S(\Gamma_\alpha)} \middle| R(\Gamma_\alpha) > 0\right] \cdot \mathbf{Pr}(R(\Gamma_\alpha) > 0) \\
&\geq \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{V(\Gamma_\alpha) + S(\Gamma_\alpha)} \middle| R(\Gamma_\alpha) > 0\right] = pFDR(\Gamma_\alpha). \ \square
\end{aligned} \tag{34}$$

We only get convergence of $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ as in Section 4 if "loose dependence" exists among the alternative statistics. Otherwise, $S(\Gamma_\alpha)$ does not converge in probability as $m \to \infty$, and neither does $\widehat{pFDR}_\lambda(\Gamma_\alpha)$. Also, note that if we condition on $S(\Gamma_\alpha)$, $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ is a maximum likelihood estimate (of the same quantity in Section 4).

Case 3: General dependence: V and S independent

Here, the dependence between the null statistics can be arbitrary, as well as between the alternative statistics. Therefore, we estimate the pFDR using the reasoning given in the beginning of this section and equation (18):

$$\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[V(\Gamma_\alpha)] + \widehat{S}(\Gamma_\alpha)}.$$

Since $\mathbf{E}[V(\Gamma_\alpha)] = m_0\alpha$ regardless of the dependence, we estimate it with $\widehat{m}_0\alpha = m\widehat{\pi}_0\alpha = \frac{m - R(\Gamma_\lambda)}{1-\lambda}\alpha$ as before. An unbiased estimate of $S(\Gamma_\alpha)$ is $R(\Gamma_\alpha) - m_0\alpha$. And since $\mathbf{E}[\widehat{m}_0\alpha] \geq m_0\alpha$, it follows that $R(\Gamma_\alpha) - \widehat{m}_0\alpha$ is an anti-conservative estimate of $S(\Gamma_\alpha)$ (which is exactly what we want since we always want our overall estimate to be conservative).

Therefore, the overall estimate is

$$\frac{\widehat{m}_0\alpha}{\widehat{m}_0\alpha + \widehat{S}(\Gamma_\alpha)} = \frac{\widehat{m}_0\alpha}{\widehat{m}_0\alpha + (R(\Gamma_\alpha) - \widehat{m}_0\alpha)} = \widehat{pFDR}_\lambda(\Gamma_\alpha). \tag{35}$$

Except for pathological cases of dependence, it follows that $\mathbf{E}[V(\Gamma_\alpha)|V(\Gamma_\lambda)]$ should be a non-decreasing function of $V(\Gamma_\lambda)$. In that case we can show the following theorem.

Theorem 4 *For arbitrary dependence among the null statistics and among the alternative statistics, if* $\mathbf{E}[V(\Gamma_\alpha)|V(\Gamma_\lambda)]$ *is a non-decreasing function of* $V(\Gamma_\lambda)$*, then* $\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq pFDR(\Gamma_\alpha)$.

Proof: The proof follows exactly as that for Theorem 3, except $\mathbf{E}[V(\Gamma_\alpha)|V(\Gamma_\lambda)]$ is a non-decreasing function of $V(\Gamma_\lambda)$ by assumption rather than by independence of the null statistics. □

Case 4: General dependence: V and S dependent

Here, the dependence between all the statistics can be arbitrary. Everything follows as in the previous case, except the conditions under which $\widehat{pFDR}$ is conservative in expectation. This is explicitly stated in the following theorem.

Theorem 5 *For arbitrary dependence among all statistics, if* $\mathbf{E}[R(\Gamma_\alpha)|R(\Gamma_\lambda)]$ *is a non-decreasing function of* $R(\Gamma_\lambda)$*, then* $\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq pFDR(\Gamma_\alpha)$.

Proof: The proof follows similarly to that for Theorem 3. □

It should finally be mentioned that in Cases 3 and 4 as well, we only get convergence of $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ if "loose dependence" between the statistics exists.

6 A Simulation Study

In this section we carry out a simulation study of the pFDR estimate in three settings: independence, clumpy dependence, and general dependence. These are discussed from a theoretical viewpoint in the previous two sections. Also note that clumpy dependence is a special case of "loose dependence" described in the previous section. We use $m = 1000$ genes and 20 samples, simulating a DNA microarray data set in the spirit of Example 1. Letting x_{ij} be the measurement for gene i and sample j, here is how the data were generated:

$$\begin{aligned} x_{ij} &\sim N(0,1) + 3 \cdot I(i \leq 50 \;\&\; j \geq 11) \;\; (\textit{independence}) \\ x_{ij} &\sim N(0,1) + 3 \cdot I(i \leq 50 \;\&\; j \geq 11) + \nu_i \;\; (\textit{clumpy}) \\ x_{ij} &\sim N(0,1) + 3 \cdot I(i \leq 50 \;\&\; j \geq 11) + \mu \;\; (\textit{general}) \end{aligned} \tag{36}$$

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

Table 2: *Simulation results. Values are the mean and standard error of the mean over 20 simulations.*

	Threshold quantile						
	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	0.0000
	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

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 2.

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 2 explore further variations of the clumpy dependence setup. In the first, the "3" in equation (36) 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.

7 Bootstrap Confidence Intervals

In Storey (2001b), we bootstrapped the statistics (or p-values) in order to obtain upper confidence intervals for $\widehat{pFDR}$. 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 Example 1, 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 and Tibshirani (1998). For example, suppose we want to find a bootstrap estimate of our confidence that $pFDR \in [0,c]$ for some c. In Example 1 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 and 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 example. Future work will adapt their methodology to obtaining confidence intervals for the pFDR for multidimensional data with independence in at least on direction (as would be expected in most multiple hypothesis testing).

In Storey (2001b), 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 Choosing the Optimal λ

In Section 2 we showed how to estimate $pFDR(\Gamma_\alpha)$, using the fixed region Γ_λ in the estimate of π_0. In this section, we will show how to approximately pick the optimal λ in order to minimize the mean-squared error between $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ and $pFDR(\Gamma_\alpha)$. That is, we provide an automatic way to estimate:

$$\lambda_{best} = \arg\min_{\lambda \in [0,1]} \mathbf{E}[(\widehat{pFDR}_\lambda(\Gamma_\alpha) - pFDR(\Gamma_\alpha))^2]. \tag{37}$$

In Storey (2001b), we use the bootstrap in order to estimate λ_{best}, and calculate an estimate of $MSE(\lambda) = \mathbf{E}[(\widehat{pFDR}_\lambda(\Gamma_\alpha) - pFDR(\Gamma_\alpha))^2]$ 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 7, we cannot yet effectively produce bootstrap versions $\widehat{pFDR}_\lambda^{*b}(\Gamma_\alpha)$ of the estimate $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ under general dependence assumptions. Therefore, we will use a different approach involving a jackknife estimate of the variance, and an estimate of the bias that is not much different from what was obtained using the bootstrap under independence.

We assume that there is some independent dimension of the data of size n. In Example 1, 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} copy of the independent dimension, we can form a new estimate of the pFDR with the remaining data. For each fixed $\lambda \in \mathcal{R}$, denote this estimate by $\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha)$ for $i = 1, \ldots, n$. The jackknife estimate of variance is:

$$\widehat{var}_\lambda = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha) - \widehat{pFDR}_\lambda(\Gamma_\alpha) \right)^2. \tag{38}$$

The jackknife estimate of bias works poorly here, so we use a different estimate. Ideally, if we knew $pFDR(\Gamma_\alpha)$, we could estimate the squared bias by $(\widehat{pFDR}_\lambda(\Gamma_\alpha) - pFDR(\Gamma_\alpha))^2$, however, we obviously do not know $pFDR(\Gamma_\alpha)$. As was done in Storey (2001b), we use a plug-in estimate of $pFDR(\Gamma_\alpha)$. Notice that for any λ we have:

$$\mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geq \min_{\lambda'} \mathbf{E}[\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)] \geq pFDR(\Gamma_\alpha), \tag{39}$$

as was shown in Section 5. Therefore, our plug-in estimate of $pFDR(\Gamma_\alpha)$ is $\min_{\lambda \in \mathcal{R}} \widehat{pFDR}_\lambda(\Gamma_\alpha)$. The estimate of the squared bias is

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(\Gamma_\alpha) - \min_{\lambda' \in \mathcal{R}} \widehat{pFDR}_{\lambda'}(\Gamma_\alpha) \right)^2. \tag{40}$$

Each of these estimates gives a nice estimate of the shape of the squared bias and variance curves over λ. However, each one tends to be inflated, and the jackknife estimate of variance can

Table 3: *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

be unpredictably inflated. Therefore, we scale each estimate by its median over the $\lambda \in \mathcal{R}$, and we make the following adjustments to our estimates:

$$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{\text{median}_{\lambda' \in \mathcal{R}}(\widehat{bias}_{\lambda'}^2)} \tag{41}$$

$$\widehat{var}_\lambda^* = \frac{\widehat{var}_\lambda}{\text{median}_{\lambda' \in \mathcal{R}}(\widehat{var}_{\lambda'})} \tag{42}$$

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

$$\widehat{MSE}(\lambda) = \widehat{bias}_\lambda^{2*} + \widehat{var}_\lambda^*, \tag{43}$$

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.

This method can easily be incorporated into the main method described in Section 2. When the null distribution is simulated, it may not always be efficient to fix $\mathcal{R}$ and then find their corresponding set of rejection regions. In that case, a sensible series of rejection regions can be chosen, and then their respective λ values can be calculated via the simulated null statistics.

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} $i = 1, \ldots, 1000$ $j = 1, \ldots, 40$. In the notation of Section 6, we have

$$x_{ij} \sim N(0,1) + u \cdot I(i \leq m_0 \ \& \ j \geq 21). \tag{50}$$

Algorithm 2

Estimation and Inference of pFDR with Optimal λ

1. For some range of λ, say $\mathcal{R} = \{0, 0.05, 0.10, \ldots, 0.95\}$, calculate $\widehat{pFDR}_\lambda(\Gamma_\alpha)$ as in Section 2.

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

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(\Gamma_\alpha) - \min_{\lambda \in \mathcal{R}} \widehat{pFDR}_\lambda(\Gamma_\alpha)\right)^2, \tag{44}$$

$$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{\text{median}_{\lambda' \in \mathcal{R}}(\widehat{bias}_{\lambda'}^2)}. \tag{45}$$

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

$$\widehat{var}_\lambda = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha) - \widehat{pFDR}_\lambda(\Gamma_\alpha)\right)^2, \tag{46}$$

$$\widehat{var}_\lambda^* = \frac{\widehat{var}_\lambda}{\text{median}_{\lambda' \in \mathcal{R}}(\widehat{var}_{\lambda'})}, \tag{47}$$

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

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

$$\widehat{MSE}(\lambda) == \widehat{bias}_\lambda^{2*} + \widehat{var}_\lambda^*. \tag{48}$$

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

$$\widehat{pFDR}(\Gamma_\alpha) = \widehat{pFDR}_{\widehat{\lambda}}(\Gamma_\alpha). \tag{49}$$

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 3 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 2 shows the MSE curve and a histogram of the 100 $\widehat{\lambda}$'s for the case where $m_0 = 200$ and $u = 0.3$.

Example 2 *DNA Microarrays continued*

Applying the method for choosing λ to the Rieger data, we find that $\widehat{\lambda} = 0.15$. Therefore, our overall estimate of $pFDR(\{t : |t| \geq 2\})$ is $\widehat{pFDR}(\{t : |t| \geq 2\}) = 7.56\%$. This has only slightly greater bias than in Example 1 where we set $\lambda = 0.75$, but the variance has been reduced significantly.

9 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. This exciting area of biological research has created several challenging statistical problems, including the multiple hypothesis testing problem we have faced here. See Brown and Botstein (1999) for an overview of DNA microarrays.

In Examples 1 and 2, we presented an application of our method to DNA microarrays. The following is a description of how one would typically organize microarray data for detecting a statistically significant change in gene expression across experimental conditions. It is the kind of data we have had in mind for the methodology presented here, although many other types of data should benefit from the procedure.

Suppose we collect data from n microarrays with the same m genes on each. Essentially, we observe the vectors $\mathbf{X}(j) = (X_{1j}, \ldots, X_{mj})$ 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 have arbitrary dependence, but the observations are independent in some way. In other words, we assume that the X_{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 $\mathbf{X}$, with each column corresponding to an observed m-vector. The columns have an independence structure (such as the two sample problem presented in this paper), but the rows are dependent.

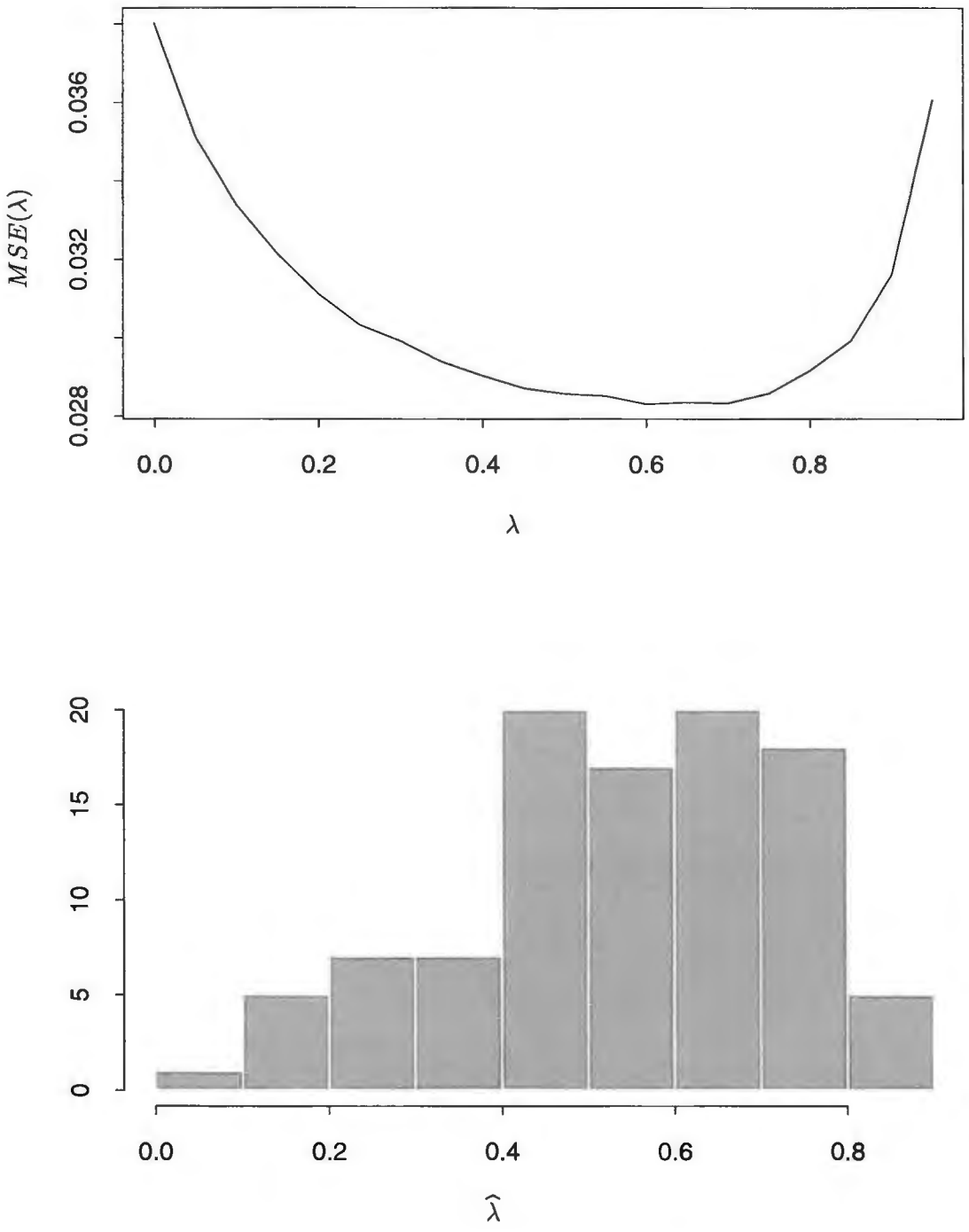

Figure 2: $m_0 = 200$ *and* $u = 0.3$. *Upper panel: The mean squared error curve as a function of* λ. *Lower panel: Histogram of the 100 observed* $\widehat{\lambda}$.

For each row of $\mathbf{X}$, we form a statistic T_i that is some function of $X_{i1}, \ldots, X_{in}$, $i = 1, \ldots, m$. We wish to test a hypothesis about a parameter of interest for each T_i. Therefore, we are testing m dependent hypotheses using the statistics $T_1, \ldots, T_m$. The null $T_1^0, \ldots, T_m^0$ are likely generated by permuting the columns in the appropriate way to simulate the null case.

Ideally, the statistics can be formed so that they are exchangeable in the sense that the $T_i|H_i = 0$ 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 exchangeable under the null hypothesis. Information can be lost, and it is a nuisance to have to include p-values as a middle step in our procedure. However, our procedure is definitely applicable to dependent p-values with rejection regions of the form $[0, \gamma]$.

In Section 5 we discussed three types of dependence: "loose dependence", dependent alternative statistics, and general dependence. We hypothesize that the most likely form of dependence encountered in DNA microarrays is "loose dependence", and more specifically, "clumpy dependence" as was used in the simulations in Section 6. 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, we expect that in most cases where one uses the pFDR to detect differential gene expression, Theorem 2 should apply. The pFDR should approximately have the independence form of the pFDR (see Theorem 1 in Section 4). This is a nice property because the bias and variance of our estimate should nearly be that obtained under independence, which has optimality properties (Storey 2001b). We can also express the pFDR as a Bayesian posterior probability as in Storey (2001a) and Efron et al. (2001), making use of the broader interpretation of the pFDR.

Suppose as in Example 1, we reject all genes with t-statistics exceeding 2 in absolute value. We then obtain a list of significant genes, with one error measure assigned to all the genes. This is unsatisfactory in that some of the genes in this list will have much bigger absolute statistics than the others, and therefore are more significant. On the other hand, we don't want to give each of

these genes a significance measure that only applies marginally because ignoring the multiplicity defeats the purpose of having the list of genes anyway. After all, the usefulness of DNA microarrays is the *simultaneous* measurement of the genes. If one or just a few genes are of interest, a Northern blot is a more precise assay anyway.

In Storey (2001a, 2001b) we introduce the q-value, which is the pFDR analogue of the p-value. The q-value of a statistic is defined to be the minimum pFDR over which that statistic can be rejected. (Recall the p-value of a statistic is the minimum Type I error over which the statistic can be rejected.) For the microarray example given in this paper, the q-value of the statistic t_0 is going to be the pFDR over the rejection region $\{t : |t| \geq |t_0|\}$. It is hoped that when using this methodology, the q-value will also be reported with each statistic. This methodology will be implemented in the SAM software that accompanies the work of Tusher et al. (2001), and the q-value will be calculated for each gene (see `http://www-stat.stanford.edu/~tibs/SAM/`).

We used a symmetric rejection region 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 cutpoints, 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 $\Gamma = \{t : \widehat{\mathbf{Pr}}(H = 0|T = t) \leq \lambda\}$ for some chosen λ. 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.

Acknowledgments

We thank Bradley Efron for continued useful discussions on this topic. RT was supported in part by NIH grant 2 R01 CA72028 and NSF grant DMS-9971405, and JDS was supported in part by a NSF graduate research fellowship and a PMMB national fellowship.

References

Benjamini Y and Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS-B* **57:** 289-300.

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

Benjamini Y and Yekutieli D. (2001) The control of the false discovery rate in multiple testing

under dependency. *Ann Stat*, in press.

Brown PO and Botstein D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21(SS):** 33-37.

Dudoit S, Yang YH, Speed TP, and Callow MJ. (2001) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, in press.

Efron B and Tibshirani RJ. (1993) *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Efron B and Tibshirani R. (1998) The problem of regions. *Ann Stat* **26:** 1687-1718.

Efron B, Tibshirani R, Storey JD, and Tusher V. (2001) Empirical Bayes analysis of a microarray experiment. *JASA*, in press.

Shaffer J. (1995) Multiple hypothesis testing. *Ann Review Psych*, **46:** 561-584.

Storey JD. (2001a) The positive false discoveray rate: A Bayesian interpretation and the q-value, submitted.
`http://www-stat.stanford.edu/~jstorey/`

Storey JD. (2001b) A direct approach to the positive false discovery rate and multiple hypothesis testing, submitted.
`http://www-stat.stanford.edu/~jstorey/`

Tusher V, Tibshirani R, and Chu G. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *PNAS* **98:** 5116-5121.

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

Weller JI, Song JZ, Heyen DW, Lewin HA, and Ron M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150:** 1699-1706.

Westfall PH and Young SS. (1993) *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley.