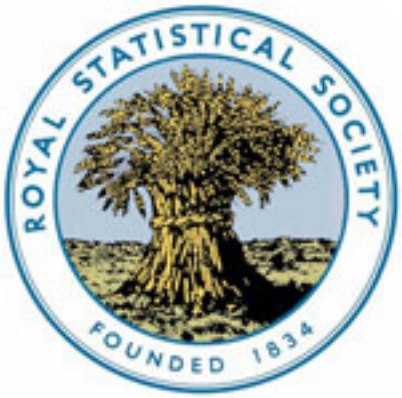

J. R. Statist. Soc. B (2002)
64, *Part* 3, *pp.* 479–498

A direct approach to false discovery rates

John D. Storey

Stanford University, USA

[Received June 2001. Revised December 2001]

Summary. Multiple-hypothesis testing involves guarding against much more complicated errors than single-hypothesis testing. Whereas we typically control the type I error rate for a single-hypothesis test, a compound error rate is controlled for multiple-hypothesis tests. For example, controlling the false discovery rate FDR traditionally involves intricate sequential *p*-value rejection methods based on the observed data. Whereas a sequential *p*-value method fixes the error rate and *estimates* its corresponding rejection region, we propose the opposite approach—we *fix* the rejection region and then estimate its corresponding error rate. This new approach offers increased applicability, accuracy and power. We apply the methodology to both the positive false discovery rate pFDR and FDR, and provide evidence for its benefits. It is shown that pFDR is probably the quantity of interest over FDR. Also discussed is the calculation of the *q*-value, the pFDR analogue of the *p*-value, which eliminates the need to set the error rate beforehand as is traditionally done. Some simple numerical examples are presented that show that this new approach can yield an increase of over eight times in power compared with the Benjamini–Hochberg FDR method.

Keywords: False discovery rate; Multiple comparisons; Positive false discovery rate; *p*-values; *q*-values; Sequential *p*-value methods; Simultaneous inference

1. Introduction

The basic paradigm for single-hypothesis testing works as follows. We wish to test a null hypothesis H_0 *versus* an alternative H_1 based on a statistic X. For a given rejection region Γ, we reject H_0 when $X \in \Gamma$ and we accept H_0 when $X \notin \Gamma$. A type I error occurs when $X \in \Gamma$ but H_0 is really true; a type II error occurs when $X \notin \Gamma$ but H_1 is really true. To choose Γ, the acceptable type I error is set at some level α; then all rejection regions are considered that have a type I error that is less than or equal to α. The one that has the lowest type II error is chosen. Therefore, the rejection region is sought with respect to controlling the *type I error*. This approach has been fairly successful, and often we can find a rejection region with nearly optimal power (power $= 1 -$ type II error) while maintaining the desired α-level type I error.

When testing multiple hypotheses, the situation becomes much more complicated. Now each test has type I and type II errors, and it becomes unclear how we should measure the overall error rate. The first measure to be suggested was the familywise error rate FWER, which is the probability of making one or more type I errors among all the hypotheses. Instead of controlling the probability of a type I error at level α for each test, the overall FWER is controlled at level α. None-the-less, α is chosen so that FWER $\leqslant \alpha$, and then a rejection region Γ is found that maintains level α FWER but also yields good power. We assume for simplicity that each test has the same rejection region, such as would be the case when using the *p*-values as the statistic.

Address for correspondence: John D. Storey, Department of Statistics, Stanford University, Stanford, CA 94305, USA.
E-mail: jstorey@stat.stanford.edu

 1369–7412/02/64479

In pioneering work, Benjamini and Hochberg (1995) introduced a multiple-hypothesis testing error measure called the false discovery rate FDR. This quantity is the expected proportion of false positive findings among all the rejected hypotheses. In many situations, FWER is much too strict, especially when the number of tests is large. Therefore, FDR is a more liberal, yet more powerful, quantity to control. In Storey (2001), we introduced the positive false discovery rate pFDR. This is a modified, but arguably more appropriate, error measure to use.

Benjamini and Hochberg (1995) provided a sequential p-value method to control FDR. This is really what an FDR controlling p-value method accomplishes: using the observed data, it estimates the rejection region so that on average FDR $\leqslant \alpha$ for some prechosen α. The product of a sequential p-value method is an estimate $\hat{k}$ that tells us to reject $p_{(1)}, p_{(2)}, \ldots, p_{(\hat{k})}$, where $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(m)}$ are the ordered observed p-values.

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 $\hat{k}$ is, the worse the procedure will work in practice. Therefore, the expected value may be that FDR $\leqslant \alpha$, but we do not know how reliable the methods are case by case. If point estimation only involved finding unbiased estimators, then the field would not be so successful. Therefore, the reliability of $\hat{k}$ case by case does matter even though it has not been explored.

Another weakness of the current approach to false discovery rates is that the error rate is controlled for all values of m_0 (the number of true null hypotheses) simultaneously without using any information in the data about m_0. Surely there is information about m_0 in the observed p-values. In our proposed method, we use this information, which yields a less stringent procedure and more power, while maintaining strong control. Often, the power of the multiple-hypothesis testing method decreases with increasing m. This should not be so, especially when the tests are independent. The larger m, the more information we have about m_0, and this should be used.

In this paper, we propose a new approach to false discovery rates. We attempt to use more traditional and straightforward statistical ideas to control pFDR and FDR. Instead of fixing α and then estimating k (i.e. estimating the rejection region), we fix the rejection region and then estimate α. Fixing the rejection region may seem counter-intuitive in the context of traditional multiple-hypothesis testing. We argue in the next section that it can make sense in the context of false discovery rates.

A natural objection to our proposed approach is that it does not offer 'control' of FDR. Actually, control is offered in the same sense as the former approach—our methodology provides a conservative bias in expectation. Moreover, since in taking this new approach we are in the more familiar point estimation situation, we can use the data to estimate m_0, obtain confidence intervals on pFDR and FDR, and gain flexibility in the definition of the error measure.

We show that our proposed approach is more effective, flexible and powerful. The multiple-hypothesis testing methods that we shall describe take advantage of more information in the data, and they are conceptually simpler. In Section 2, we discuss pFDR and its relationship to FDR, as well as using fixed rejection regions in multiple-hypothesis testing. In Section 3 we formulate our approach, and in Section 4 we make a heuristic comparison between the method proposed and that of Benjamini and Hochberg (1995). Section 5 provides numerical results, comparing our approach with the current one. Section 6 describes several theoretical results pertaining to the proposed approach, including a maximum likelihood estimate interpretation. Section 7 describes a quantity called the q-value, which is the pFDR analogue of the p-value, and Section 8 argues that the pFDR and the q-value are the most appropriate false discovery rate

quantities to use. Section 9 shows how to pick a tuning parameter in the estimates automatically. Section 10 is the discussion, and Appendix A provides technical comments and proofs of the theorems.

2. The positive false discovery rate and fixed rejection regions

As mentioned in Section 1, two error measures are commonly used in multiple-hypothesis testing: FWER and FDR. FWER is the traditional measure used; Benjamini and Hochberg (1995) recently introduced FDR. Table 1 summarizes the various outcomes that occur when testing m hypotheses.

V is the number of type I errors (or false positive results). Therefore, FWER is defined to be $\Pr(V \geqslant 1)$. Controlling this quantity offers a very strict error measure. In general, as the number of tests increases, the power decreases when controlling FWER. FDR is defined to be

$$\text{FDR} = E\left(\frac{V}{R}\middle| R > 0\right)\Pr(R > 0), \tag{1}$$

i.e. the expected proportion of false positive findings among all rejected hypotheses times the probability of making at least one rejection. Benjamini and Hochberg (1995) and Benjamini and Liu (1999) provided sequential p-value methods to control this quantity. FDR offers a much less strict multiple-testing criterion over FWER and therefore leads to an increase in power.

In Storey (2001), we define a new false discovery rate, pFDR.

Definition 1.

$$\text{pFDR} = E\left(\frac{V}{R}\middle| R > 0\right). \tag{2}$$

The term 'positive' has been added to reflect the fact that we are conditioning on the event that positive findings have occurred. When controlling FDR at level α, and positive findings have occurred, then FDR has really only been controlled at level $\alpha/\Pr(R > 0)$. This can be quite dangerous, and it is not the case for pFDR. See Storey (2001) for a thorough motivation of pFDR over FDR.

Benjamini and Hochberg (1995) precisely define FDR to be expression (1) because this quantity can be controlled by a sequential p-value method. (Note, however, that weak control of FWER is implicitly embedded in this definition, i.e. FWER is controlled when all null hypotheses are true.) pFDR is identically 1 when all null hypotheses are true ($m = m_0$), so the usual sequential p-value approach cannot be applied to this quantity. Therefore, to control pFDR, it must be estimated for a particular rejection region.

A sequential p-value method allows us to fix the error rate beforehand and to estimate the rejection region, which is what has traditionally been done in multiple-hypothesis testing. In the context of FWER this makes sense. Because FWER measures the probability of making one

Table 1. Outcomes when testing m hypotheses

Hypothesis	*Accept*	*Reject*	*Total*
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

or more type I error, which is essentially a '0–1' event, we can set the rate *a priori* at which this should occur. False discovery rates, in contrast, are more of an exploratory tool. For example, suppose that we are testing 1000 hypotheses and decide beforehand to control FDR at level 5%. Whether this was an appropriate choice largely depends on the number of hypotheses that are rejected. If 100 hypotheses are rejected, then clearly this was a good choice. If only two hypotheses are rejected, then clearly this was a less useful choice.

Therefore fixing the rejection region beforehand can be more appropriate when using pFDR or FDR. For example, when performing many hypothesis tests, it can make sense to reject all p-values that are less than 0.05 or 0.01. Also, expert knowledge in a particular field may allow us to decide which rejection regions should be used.

It will also be seen that this approach allows us to control pFDR, which we find to be a more appropriate error measure. Probably the most important reason for fixing the rejecting region is that it allows us to take a conceptually simpler approach to complicated compound error measures such as pFDR and FDR.

The q-value (Section 7) is the pFDR analogue of the p-value and allows the rejection regions to be determined by the observed p-values. This is more appropriate over either fixing the rejection region or fixing the error rate. But, by first fixing the rejection region in our approach, we can formulate the q-values quite easily.

3. Estimation and inference for the positive false discovery rate and false discovery rate

In this section, we apply the proposed approach to both pFDR and FDR. We first present a few simple facts about pFDR under independence to motivate our estimates. Suppose that we are testing m identical hypothesis tests $H_1, H_2, \ldots, H_m$ with independent statistics $T_1, T_2, \ldots, T_m$. We let $H_i = 0$ when null hypothesis i is true, and $H_i = 1$ otherwise. We assume that the null $T_i|H_i = 0$ and the alternative $T_i|H_i = 1$ are identically distributed. We assume that the same rejection region is used for each test, which make the tests 'identical'. Finally, we assume that the H_i are independent Bernoulli random variables with $\Pr(H_i = 0) = \pi_0$ and $\Pr(H_i = 1) = \pi_1$. Let Γ be the common rejection region for each hypothesis test.

The following is theorem 1 from Storey (2001). It allows us to write pFDR in a very simple form that does not depend on m. For this theorem to hold we must assume that the H_i are random; however, for large m this assumption makes little difference.

Theorem 1. Suppose that m identical hypothesis tests are performed with the independent statistics $T_1, \ldots, T_m$ and rejection region Γ. Also suppose that a null hypothesis is true with *a priori* probability π_0. Then

$$\mathrm{pFDR}(\Gamma) = \frac{\pi_0 \Pr(T \in \Gamma | H = 0)}{\Pr(T \in \Gamma)} \tag{3}$$

$$= \Pr(H = 0 | T \in \Gamma), \tag{4}$$

where $\Pr(T \in \Gamma) = \pi_0 \Pr(T \in \Gamma | H = 0) + \pi_1 \Pr(T \in \Gamma | H = 1)$.

This paper will be limited to the case where we reject on the basis of independent p-values. See Storey and Tibshirani (2001) for a treatment of more general situations. It follows that for rejections based on p-values all rejection regions are of the form $[0, \gamma]$ for some $\gamma \geqslant 0$. (See remark 1 in Appendix A for a justification of this.) For the remainder of the paper, instead of denoting rejection regions by the more abstract Γ, we denote them by γ, which refers to the interval $[0, \gamma]$.

In terms of p-values we can write the result of theorem 1 as

$$\text{pFDR}(\gamma) = \frac{\pi_0 \Pr(P \leqslant \gamma | H = 0)}{\Pr(P \leqslant \gamma)} = \frac{\pi_0 \gamma}{\Pr(P \leqslant \gamma)}, \tag{5}$$

where P is the random p-value resulting from any test. Under independence the p-values are exchangeable in that each comes from the null distribution (i.e. uniform[0,1]) with probability π_0 and from the alternative distribution with probability π_1. It is easiest to think about this in terms of simple *versus* simple hypothesis tests, but the theory also works for composite alternative hypotheses with a random effect (Storey, 2001).

Since $\pi_0 m$ 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. Hence, a conservative estimate of π_0 is

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} = \frac{W(\lambda)}{(1-\lambda)m} \tag{6}$$

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.) For now we assume that λ is fixed; however, we show how to pick the optimal λ in Section 9. (Efron *et al.* (2001) used a different estimate of π_0 in an empirical Bayes method that is related to pFDR.) A natural estimate of $\Pr(P \leqslant \gamma)$ is

$$\widehat{\Pr}(P \leqslant \gamma) = \frac{\#\{p_i \leqslant \gamma\}}{m} = \frac{R(\gamma)}{m}, \tag{7}$$

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

$$\widehat{Q}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(P \leqslant \gamma)} = \frac{W(\lambda)\gamma}{(1-\lambda)\, R(\gamma)}. \tag{8}$$

pFDR and FDR are asymptotically equivalent for a fixed rejection region. We see in Section 6 that $\widehat{Q}_\lambda(\gamma)$ shows good asymptotic properties for pFDR. In fact, we show that it is a maximum likelihood estimate. However, because of finite sample considerations, we must make two slight adjustments to estimate pFDR. When $R(\gamma) = 0$, the estimate would be undefined, which is undesirable for finite samples. Therefore, we replace $R(\gamma)$ with $R(\gamma) \vee 1$. This is equivalent to making a linear interpolation between the estimate at $[0, p_{(1)}]$ and the origin. Also, $1-(1-\gamma)^m$ is clearly a lower bound for $\Pr\{R(\gamma) > 0\}$. Since pFDR is conditioned on $R(\gamma) > 0$, we divide by $1-(1-\gamma)^m$. In other words $\gamma/\{1-(1-\gamma)^m\}$ is a conservative estimate of the type I error, conditional that $R(\gamma) > 0$. (See Section 8 for more on why we do this.) Therefore, we estimate pFDR as

$$\widehat{\text{pFDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(P \leqslant \gamma)\{1-(1-\gamma)^m\}} = \frac{W(\lambda)\gamma}{(1-\lambda)\{R(\gamma) \vee 1\}\{1-(1-\gamma)^m\}}. \tag{9}$$

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

$$\widehat{\text{FDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(P \leqslant \gamma)} = \frac{W(\lambda)\gamma}{(1-\lambda)\{R(\gamma) \vee 1\}}. \tag{10}$$

For large m these two estimates are equivalent, but we find the pFDR quantity to be more appropriate and think that it should be used. When $\gamma = 1/m$, $\Pr\{R(\gamma) > 0\}$ can be as small as 0.632, so FDR can be quite misleading as mentioned in the previous section. For fixed m and $\gamma \to 0$, $\text{FDR}(\gamma)$ and $\widehat{\text{FDR}}_\lambda(\gamma)$ show unsatisfying properties, and we show this in Section 8.

We show in Section 6 that $\widehat{\text{pFDR}}_\lambda(\gamma)$ and $\widehat{\text{FDR}}_\lambda(\gamma)$ offer an analogous property to strong control in that they are conservatively biased for all π_0. However, as we argued in Section 1, the expected value of a multiple-hypothesis testing procedure is not a sufficiently broad 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 pFDR and FDR. This allows us 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 $\text{pFDR}(\gamma)$ are given in algorithm 1. The same algorithm holds for the estimation and inference of $\text{FDR}(\gamma)$, except that we obviously use $\widehat{\text{FDR}}_\lambda(\gamma)$ instead. In Section 9, we extend our methodology to include an automatic method for choosing the optimal λ.

If $\widehat{\text{pFDR}}_\lambda(\gamma) > 1$, we recommend setting $\widehat{\text{pFDR}}_\lambda(\gamma) = 1$ since obviously $\text{pFDR}(\gamma) \leqslant 1$. We could smooth the estimate so that it is always less than or equal to 1, but we have taken a simpler approach here. The same comment holds for $\widehat{\text{FDR}}_\lambda(\gamma)$.

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

3.1. Algorithm 1: estimation and inference for pFDR(γ) and FDR(γ)

(a) For the m hypothesis tests, calculate their respective p-values $p_1, \ldots, p_m$.

(b) Estimate π_0 and $\Pr(P \leqslant \gamma)$ by

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

and

$$\widehat{\Pr}(P \leqslant \gamma) = \frac{R(\gamma) \vee 1}{m},$$

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

(c) For any rejection region of interest $[0, \gamma]$, estimate $\text{pFDR}(\gamma)$ by

$$\widehat{\text{pFDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(P \leqslant \gamma)\{1-(1-\gamma)^m\}},$$

for some well-chosen λ. (See Section 9 for how to choose the optimal λ.)

(d) For B bootstrap samples of $p_1, \ldots, p_m$, calculate the bootstrap estimates $\widehat{\text{pFDR}}_\lambda^{*b}(\gamma)$ $(b = 1, \ldots, B)$ similarly to the method above.

(e) Form a $1 - \alpha$ upper confidence interval for $\text{pFDR}(\gamma)$ by taking the $1 - \alpha$ quantile of the $\widehat{\text{pFDR}}_\lambda^{*b}(\gamma)$ as the upper confidence bound.

(f) For $\text{FDR}(\gamma)$, perform this same procedure except using

$$\widehat{\text{FDR}}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(P \leqslant \gamma)}. \tag{11}$$

4. A connection between the two approaches

In this section we present a heuristic connection between the sequential p-value method of Benjamini and 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 basic point that we make is that using the Benjamini and Hochberg (1995) method to control FDR at level α/π_0 is equivalent to (i.e. rejects the same p-values as) using the proposed method to control FDR at level α. The gain in power from our approach is clear—we control a smaller error rate ($\alpha \leqslant \alpha/\pi_0$), yet reject the same number of tests.

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

$$\hat{k} = \max\{k : p_{(k)} \leqslant (k/m)\alpha\}. \tag{12}$$

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

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

$$\hat{l} = \max\{l : \widehat{\mathrm{FDR}}(p_{(l)}) \leqslant \alpha\}. \tag{13}$$

But since

$$\widehat{\mathrm{FDR}}(p_{(l)}) = \frac{\hat{\pi}_0\, p_{(l)}}{l/m}$$

this is equivalent to (with $\hat{\pi}_0 = 1$)

$$\hat{l} = \max\{l : p_{(l)} \leqslant (l/m)\alpha\}. \tag{14}$$

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

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} \tag{15}$$

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

Therefore, we have shown that $\hat{l} \geqslant \hat{k}$. In other words, using our approach, we reject a greater number of hypotheses while controlling the same error rate, which leads to greater power. The operational difference between $\widehat{\mathrm{FDR}}_\lambda(\gamma)$ and the Benjamini and Hochberg (1995) procedure is the inclusion of $\hat{\pi}_0(\lambda)$. It is important to note, however, that we did not simply reverse their method and stick in $\hat{\pi}_0(\lambda)$. Rather, we took a very different approach, starting from simple facts about pFDR under independence with fixed rejection regions. Benjamini and Hochberg (1995) did not give us much insight into why they chose their particular sequential p-value method. This comparison sheds some light on why it works.

5. A numerical study

In this section we present some numerical results to compare the power of the Benjamini and Hochberg (1995) procedure with our proposed method. As mentioned in Section 4, it is not straightforward to compare these two methods since Benjamini and Hochberg (1995) estimated the rejection region whereas our method estimates FDR. We circumvent this problem by using the Benjamini–Hochberg procedure to control FDR at level $\widehat{\mathrm{FDR}}_\lambda(\gamma)$ for each iteration.

We looked at the two rejection regions $\gamma = 0.01$ and $\gamma = 0.001$ over several values of π_0. The values of γ and π_0 were chosen 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 $p_i = \Pr\{N(0, 1) \geqslant z_i\}$, where z_i is the observed value of Z_i.

To calculate the power of our method, test i was rejected if $p_i \leqslant \gamma$, and the power was calculated accordingly. Also, $\widehat{\text{FDR}}(\gamma)$ was calculated as we outlined in Section 3. The Benjamini–Hochberg method was performed at level $\widehat{\text{FDR}}(\gamma)$, and the power was calculated. This approach should put the two methods on equal ground for comparison; reporting $\widehat{\text{FDR}}(\gamma)$ is equivalent in practice to using the Benjamini–Hochberg method to control FDR at level $\widehat{\text{FDR}}(\gamma)$.

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 FDR as if the alternative distribution was unknown. Therefore, we had to choose a value of λ to estimate π_0; we used $\lambda = \frac{1}{2}$ in all calculations for simplicity.

Table 2 shows the results of the simulation study. The first half of the table corresponds to $\gamma = 0.01$, and the second half corresponds to $\gamma = 0.001$. It can be seen that there is a substantial increase in power by using the proposed method. One case even gives an increase of over 800% in power. The power is constant over each case of our method because the same rejection region is used. The power of the Benjamini–Hochberg method increases as π_0 grows larger because

Table 2. Numerical comparison between the Benjamini–Hochberg and proposed methods

π_0	*FDR*	*Power*		$E(\widehat{FDR})$, *proposed method*	$E(\hat{\pi}_0)$, *proposed method*	$E(\hat{\gamma})$, *Benjamini–Hochberg method*
		Proposed method	*Benjamini–Hochberg method*			
$\gamma = 0.01$						
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
$\gamma = 0.001$						
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

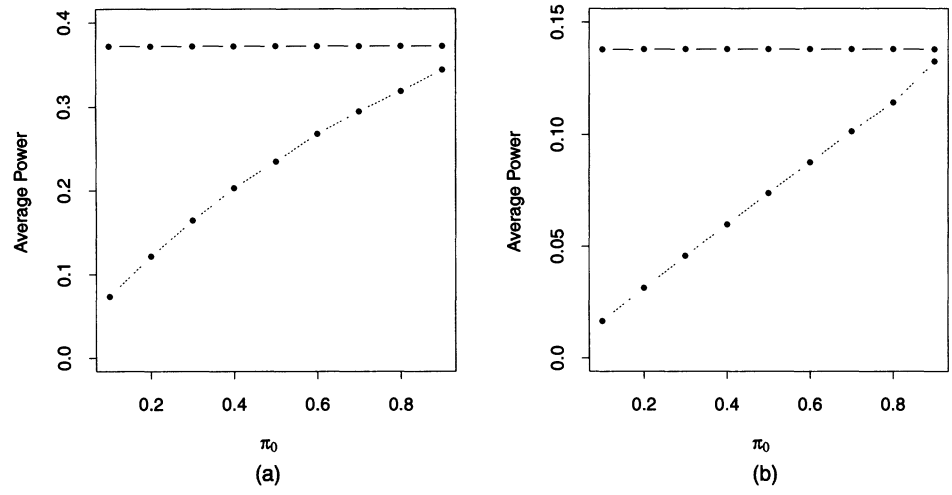

Fig. 1. Average power *versus* π_0 for the Benjamini–Hochberg method (· · · · · ·) and the proposed method (———): (a) rejection region defined by $\gamma = 0.01$; (b) rejection region defined by $\gamma = 0.001$ (it can be seen that there is a substantial increase in power under the proposed method in both cases)

the procedure becomes less conservative. In fact, it follows from Section 4 that, as $\pi_0 \to 1$, the Benjamini–Hochberg method becomes the proposed method.

The fifth column of Table 2 shows $E(\widehat{\text{FDR}})$ for our method. 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 proposed method is nearly optimal in that it estimates $\text{FDR}(\gamma)$ 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 becomes better as the number of tests increases; the opposite has been true in the past. The sixth column shows $E(\hat{\pi}_0)$ for our method. It can be seen that this estimate is always conservative and very close to the actual value. Recall that the Benjamini–Hochberg method essentially estimates the rejection region $[0, \gamma]$. The eighth column shows $E(\hat{\gamma})$ over the 1000 realizations of $\hat{\gamma}$. It can be seen that these estimates are quite conservative. The power comparisons are also shown graphically in Fig. 1.

The success of our method also largely depends on how well we can estimate $\text{pFDR}(\gamma)$ and $\text{FDR}(\gamma)$. 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 6. If we choose λ more adaptively, the estimation is even better. This is the topic of Section 7.

6. Theoretical results

In this section, we provide finite sample and large sample results for $\widehat{\text{pFDR}}_\lambda(\gamma)$ and $\widehat{\text{FDR}}_\lambda(\gamma)$. Our goal of course is to provide conservative estimates of $\text{pFDR}(\gamma)$ and $\text{FDR}(\gamma)$. For example, we want $\widehat{\text{pFDR}}_\lambda(\gamma) \geqslant \text{pFDR}(\gamma)$ as much as possible without being too conservative. The following is our main finite sample result.

Theorem 2. $E\{\widehat{\text{pFDR}}_\lambda(\gamma)\} \geqslant \text{pFDR}(\gamma)$ and $E\{\widehat{\text{FDR}}_\lambda(\gamma)\} \geqslant \text{FDR}(\gamma)$ for all γ and π_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 1. Of course in practice we would

truncate the estimates at 1 since $\text{FDR} \leqslant \text{pFDR} \leqslant 1$, but the expected value of the estimates nevertheless behaves as we would want it to. The following result shows that truncating the estimates is a good idea when taking into account both bias and variance.

Theorem 3. $E[\{\widehat{\text{pFDR}}_\lambda(\gamma) - \text{pFDR}(\gamma)\}^2] > E[\{\widehat{\text{pFDR}}_\lambda(\gamma) \wedge 1 - \text{pFDR}(\gamma)\}^2]$ and $E[\{\widehat{\text{FDR}}_\lambda(\gamma) - \text{FDR}(\gamma)\}^2] > E[\{\widehat{\text{FDR}}_\lambda(\gamma) \wedge 1 - \text{FDR}(\gamma)\}^2]$.

We now present large sample results for $\widehat{\text{pFDR}}(\gamma)$. (These results also hold for $\widehat{\text{FDR}}(\gamma)$ since $\widehat{\text{FDR}}(\gamma) \sim \widehat{\text{pFDR}}(\gamma)$.) The tightness to which $\widehat{\text{pFDR}}(\gamma)$ converges to an upper bound of $\text{pFDR}(\gamma)$ largely depends on how the power changes with the type I error. For this, let $g(\lambda) = \Pr(P \leqslant \lambda | H = 1)$ be the power as a function of type I error λ. Note that $g(\cdot)$ is just the cumulative density function of the alternative p-values. For m identical simple tests, $g(\lambda)$ is the same for each test. If the alternative hypothesis is composite, then $g(\lambda)$ must be defined as the appropriate mixture. We assume that $g(0) = 0$, $g(1) = 1$ and $g(\lambda) \geqslant \lambda$ for $0 < \lambda < 1$.

Theorem 4. With probability 1, we have

$$\lim_{m \to \infty} \{\widehat{\text{pFDR}}_\lambda(\gamma)\} = \frac{\pi_0 + \pi_1\{1 - g(\lambda)\}/(1 - \lambda)}{\pi_0} \text{pFDR}(\gamma) \geqslant \text{pFDR}(\gamma). \tag{16}$$

This theorem can be understood graphically in terms of the plot of power against type I error for each rejection region $[0, \lambda]$. The function $g(\lambda)$ gives the power over the rejection region $[0, \lambda]$, 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 $1 - g(\lambda)$ is the probability that a p-value from the alternative distribution falls into $(\lambda, 1]$. Likewise, $1 - \lambda$ is the probability that the null p-value falls into $(\lambda, 1]$. The estimate of π_0 is better the more $g(\lambda) > \lambda$. This is the case since the interval $(\lambda, 1]$ will contain fewer alternative p-values, and hence the estimate will be less conservative. Fig. 2 shows a plot of $g(\lambda)$ *versus* λ for a concave g. For concave g, the estimate of π_0 becomes less conservative as $\lambda \to 1$. This is formally stated in the following corollary.

Corollary 1. For concave g

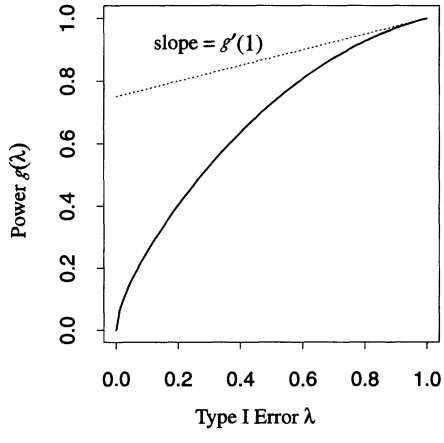

Fig. 2. Power $g(\lambda)$ *versus* type 1 error λ: it can be seen that since g is concave $\{1 - g(\lambda)\}/(1 - \lambda)$ grows smaller as $\lambda \to 1$; the line has slope $\lim_{\lambda \to 1}[\{1 - g(\lambda)\}/(1 - \lambda)]$, which is the smallest value of $\{1 - g(\lambda)\}/(1 - \lambda)$ that can be attained for concave g

$$\inf_{\lambda} \lim_{m\to\infty} \{\widehat{\text{pFDR}}_\lambda(\gamma)\} = \lim_{\lambda\to 1} \lim_{m\to\infty} \{\widehat{\text{pFDR}}_\lambda(\gamma)\}$$
$$= \frac{\pi_0 + g'(1)\pi_1}{\pi_0}\text{pFDR}(\gamma) \qquad \text{almost surely,} \qquad (17)$$

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

In other words, the right-hand side of equation (17) is the tightest upper bound that $\widehat{\text{pFDR}}(\gamma)$ can attain on pFDR as $m \to \infty$ for concave g. The corollary can be seen graphically in Fig. 3. A plot of $\{1 - g(\lambda)\}/(1 - \lambda)$ *versus* λ is shown for a concave g. It can be seen that the minimum is obtained at $\lambda = 1$. The minimum value is $g'(1)$, which happens to be $\frac{1}{4}$ in this graph. Whenever the rejection regions are based on a monotone function of the likelihood ratio between the null and alternative hypotheses, g is concave. If g is not concave, then the optimal λ used in the estimate of π_0 may not be $\lambda = 1$.

A nice property of this last result is that $g'(1) = 0$ whenever doing a one-sided test of a single parameter of an exponential family. Therefore, in many of the common cases, we can achieve exact convergence as $\lambda \to 1$.

Recall the estimate $\hat{Q}_\lambda(\gamma) = \hat{\pi}_0(\lambda)\gamma/R(\gamma)$ from equation (8) in Section 3. $\widehat{\text{pFDR}}_\lambda(\gamma)$ and $\widehat{\text{FDR}}_\lambda(\gamma)$ are modified versions of this that show good finite sample properties, as we have seen. It follows, however, that $\hat{Q}_\lambda(\gamma)$ is a maximum likelihood estimate of the limiting quantity in theorem 4.

Theorem 5. Under the assumptions of theorem 1, $\hat{Q}_\lambda(\gamma)$ is the maximum likelihood estimate of

$$\frac{\pi_0 + \pi_1\{1 - g(\lambda)\}/(1 - \lambda)}{\pi_0}\text{pFDR}(\gamma). \qquad (18)$$

This quantity is slightly greater than $\text{pFDR}(\gamma)$ for powerful tests. In situations where g is unknown, this estimate is, loosely speaking, 'optimal' in that the bias can usually be made arbitrarily small (see corollary 1), while obtaining the smallest asymptotic variance for an estimator of that bias. $\widehat{\text{pFDR}}_\lambda(\gamma)$ has good finite sample properties (avoiding the inconveniences of the pure maximum likelihood estimate), but it is asymptotically equivalent to $\hat{Q}_\lambda(\gamma)$, so it has the same large sample properties.

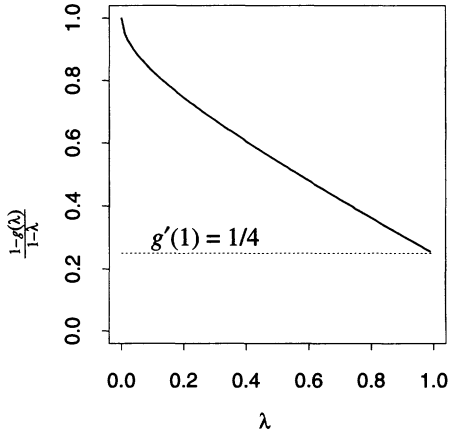

Fig. 3. $\{1 - g(\lambda)\}/(1 - \lambda)$ *versus* λ for a concave g: it can be seen that the minimum is obtained at $\lambda = 1$ with value $g'(1) = \frac{1}{4}$

7. The q-value

We now discuss a natural pFDR analogue of the p-value, which we call the *q-value*. This quantity was first developed and investigated in Storey (2001). The q-value gives the scientist a hypothesis testing error measure for each observed statistic with respect to pFDR. The p-value accomplishes the same goal with respect to the type I error, and the adjusted p-value with respect to FWER.

Even though we are only considering hypothesis testing with independent p-values, it helps to introduce the q-value formally in a general setting to motivate its definition better. We shall also define the q-value in terms of p-values. For a nested set of rejection regions $\{\Gamma\}$ (for example, $\{\Gamma\}$ could be all sets of the form $[c, \infty)$ for $-\infty \leqslant c \leqslant \infty$), the p-value of an observed statistic $T = t$ is defined to be

$$p\text{-value}(t) = \min_{\{\Gamma : t \in \Gamma\}} \{\Pr(T \in \Gamma | H = 0)\}. \tag{19}$$

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 a statistic with value t for the set of nested rejection regions. In a multiple-testing situation, we can adjust the p-values of several statistics to control FWER. The adjusted p-values give a measure of the strength of an observed statistic with respect to making one or more type I error. As a natural extension to pFDR, the q-value has the following definition.

Definition 2. For an observed statistic $T = t$, the *q-value* of t is defined to be

$$q(t) = \inf_{\{\Gamma : t \in \Gamma\}} \{\text{pFDR}(\Gamma)\}. \tag{20}$$

In words, the q-value is a measure of the strength of an observed statistic with respect to pFDR—it is the minimum pFDR that can occur when rejecting a statistic with value t for the set of nested rejection regions.

The definition is simpler when the statistics are independent p-values. The nested set of rejection regions take the form $[0, \gamma]$ and pFDR can be written in a simple form. Therefore, in terms of independent p-values, the following is the definition of the q-value of an observed p-value p.

Definition 3. For a set of hypothesis tests conducted with independent p-values, the q-value of the observed p-value p is

$$q(p) = \inf_{\gamma \geqslant p} \{\text{pFDR}(\gamma)\} = \inf_{\gamma \geqslant p} \left\{ \frac{\pi_0 \gamma}{\Pr(P \leqslant \gamma)} \right\}. \tag{21}$$

The right-hand side of the definition only holds when the H_i are random as in theorem 1. See Storey (2001) for more theoretical details about the q-value. Here, we propose the following algorithm for calculating $q(p_i)$ in practice.

This procedure ensures that $\hat{q}(p_{(1)}) \leqslant \ldots \leqslant \hat{q}(p_{(m)})$, which is necessary according to our definition. The q-values can be used in practice in the following way: $\hat{q}(p_{(i)})$ gives us the minimum pFDR that 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 $q(p_{(i)})$ so that at least $p_{(1)}, \ldots, p_{(i)}$ are rejected. Note that we write the calculated q-values as $\hat{q}(p_{(i)})$. This is because $\hat{q}(p_{(i)})$ is an *estimate* of $q(p_{(i)})$. The exact operating characteristics of $\hat{q}(p_{(i)})$ are left as an open problem, but simulations show that it behaves conservatively, as we would want.

7.1. Algorithm 2: calculating the q-value

(a) For the m hypothesis tests, calculate the p-values $p_1, \ldots, p_m$.
(b) Let $p_{(1)} \leqslant \ldots \leqslant p_{(m)}$ be the ordered p-values.
(c) Set $\hat{q}(p_{(m)}) = \widehat{\text{pFDR}}(p_{(m)})$.
(d) Set $\hat{q}(p_{(i)}) = \min\{\widehat{\text{pFDR}}(p_{(i)}), \hat{q}(p_{(i+1)})\}$ for $i = m-1, m-2, \ldots, 1$.

8. The advantages of $\widehat{\text{pFDR}}_\lambda(\gamma)$ and $\hat{q}$ over $\widehat{\text{FDR}}_\lambda(\gamma)$

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

$$\lim_{\gamma \to 0}\{\widehat{\text{pFDR}}_\lambda(\gamma)\} = \hat{\pi}_0(\lambda). \tag{22}$$

In other words, as we make the rejection region increasingly smaller, we eventually estimate pFDR as $\hat{\pi}_0(\lambda)$. This is the conservative thing to do since all that we can conclude is that

$$\lim_{\gamma \to 0}\{\text{pFDR}(\gamma)\} \leqslant \pi_0.$$

Also, under no parametric assumptions, this is exactly what we would want. For example, suppose that we take a very small rejection region. Then it is most likely that only one p-value falls into that region. Without information from other p-values, and without parametric information about the alternative distribution, there is little that we can say about whether this one p-value is null or alternative. Therefore, it makes sense to estimate pFDR by the prior probability $\hat{\pi}_0(\lambda)$ in extremely small rejection regions.

Note in contrast that

$$\lim_{\gamma \to 0}\{\widehat{\text{FDR}}_\lambda(\gamma)\} = 0. \tag{23}$$

Does this makes sense? It does in that $\lim_{\gamma \to 0}\{\text{FDR}(\gamma)\} = 0$. But the only reason why we always achieve this convergence is because of the extra term $\Pr\{R(\gamma) > 0\}$ in FDR. Therefore, as γ becomes small, 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 define the q-value in terms of FDR, then for small p-values it would be driven to 0 just because the p-value is small, even though we know little about how likely it came from the alternative hypothesis without serious assumptions. Consider Fig. 4. 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. Fig. 4(a) shows $\widehat{\text{pFDR}}_\lambda(\gamma)$ and $\widehat{\text{FDR}}_\lambda(\gamma)$ as a function of γ, 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 close to the origin.

Fig. 4(b) zooms in near zero, where we see that $\widehat{\text{pFDR}}_\lambda(\gamma)$ shoots up to $\hat{\pi}_0(\lambda)$, and $\widehat{\text{FDR}}_\lambda(\gamma)$ shoots down to 0. The q-value, however, sits steady where $\widehat{\text{pFDR}}_\lambda(\gamma)$ reaches its minimum (at about $p_{(10)}$). In other words, the q-value calibrates where we start to receive enough information to make good statements about pFDR. FDR says nothing about 'the rate that discoveries are false' near the origin and merely measures the fact that we are near the origin. Moreover, the area near zero is arguably the most important region since this is where the most significant p-values lie. Therefore, by using $\widehat{\text{pFDR}}_\lambda(\gamma)$ and the q-value, we obtain *robust* estimates of pFDR,

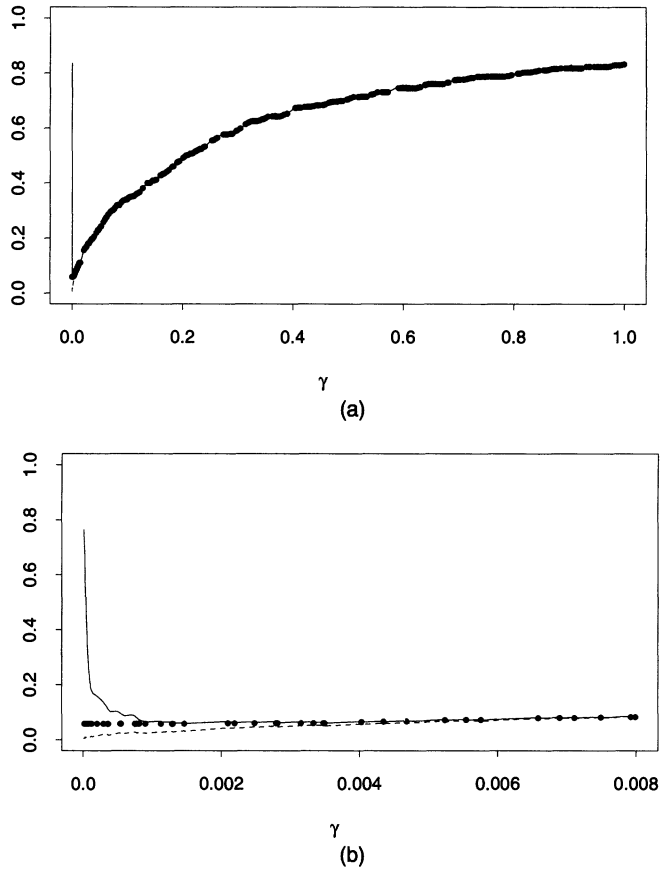

Fig. 4. Plot of $\widehat{\text{pFDR}}(\gamma)$ (———), $\widehat{\text{FDR}}(\gamma)$ (-------) and $\hat{q}$ (•) for the $N(0,1)$ *versus* $N(2,1)$ example: it can be seen that $\widehat{\text{pFDR}}(\gamma)$ and $\hat{q}$ behave more reasonably than $\widehat{\text{FDR}}(\gamma)$ near the origin

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 must be done in the traditional framework.

9. Calculating the optimal λ

In Section 3 we showed how to estimate $\text{pFDR}(\gamma)$ and $\text{FDR}(\gamma)$, using the fixed parameter λ for the estimate of π_0. In this section, we show how to pick λ optimally to minimize the mean-squared error of our estimates. We present the methodology for $\widehat{\text{pFDR}}_\lambda(\gamma)$, although the same procedure works for $\widehat{\text{FDR}}_\lambda(\gamma)$. We provide an automatic way to estimate

$$\lambda_{\text{best}} = \arg\min_{\lambda \in [0,1]} (E[\{\widehat{\text{pFDR}}_\lambda(\gamma) - \text{pFDR}(\gamma)\}^2]). \tag{24}$$

We use the bootstrap method to estimate λ_{best} and calculate an estimate of $\text{MSE}(\lambda) = E[\{\widehat{\text{pFDR}}_\lambda(\gamma) - \text{pFDR}(\gamma)\}^2]$ over a range of λ. (Call this range R; for example, we may take $R = \{0, 0.05, 0.10, \ldots, 0.95\}$.) As mentioned in Section 3, we can produce bootstrap versions

$\widehat{\text{pFDR}}_\lambda^{*b}(\gamma)$ (for $b = 1, \ldots, B$) of the estimate $\widehat{\text{pFDR}}_\lambda(\gamma)$ for any fixed λ. Ideally we would like to know $\text{pFDR}(\gamma)$, and then the bootstrap estimate of the $\text{MSE}(\lambda)$ would be

$$\frac{1}{B}\sum_{b=1}^{B}\{\widehat{\text{pFDR}}_\lambda^{*b}(\gamma) - \text{pFDR}(\gamma)\}^2. \tag{25}$$

However, we do not know $\text{pFDR}(\gamma)$, so we must form a plug-in estimate of this quantity (Efron and Tibshirani, 1993). For any λ we have

$$E\{\widehat{\text{pFDR}}_\lambda(\gamma)\} \geqslant \min_{\lambda'}[E\{\widehat{\text{pFDR}}_{\lambda'}(\gamma)\}] \geqslant \text{pFDR}(\gamma), \tag{26}$$

as was shown in Section 6. Therefore, our plug-in estimate of $\text{pFDR}(\gamma)$ is $\min_{\lambda' \in R}\{\widehat{\text{pFDR}}_{\lambda'}(\gamma)\}$. The estimate of $\text{MSE}(\lambda)$ is then

$$\widehat{\text{MSE}}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}[\widehat{\text{pFDR}}_\lambda^{*b}(\gamma) - \min_{\lambda' \in R}\{\widehat{\text{pFDR}}_{\lambda'}(\gamma)\}]^2. \tag{27}$$

This method can easily be incorporated in the main method described in Section 3 in a computationally efficient way. Our proposed method for choosing λ is formally detailed in algorithm 3. Finally note that, in choosing λ over the q-values, we can minimize the averaged $\text{MSE}(\lambda)$ over all the q-values and adjust algorithm 3 accordingly.

We provide some numerical results under the following set-up. We tested m hypotheses of $N(0, 1)$ *versus* $N(1, 1)$ with the rejection region $\Gamma = [c, \infty)$. Each statistic is independent—therefore, when we form bootstrap statistics, we simply sample from the m statistics. We calculated λ_{best} from the true mean-squared error 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 mean-squared error curves. $\hat{\lambda}$ was chosen by taking the minimum of the *averaged* mean-squared error curves. Taking the median of the 100 $\hat{\lambda}$ produces nearly identical results.

Fig. 5 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. It is not important to predict the mean-squared error 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 mean-squared error for any λ. Table 3 shows simulation results for several other sets of parameters. It can be seen that even when $\hat{\lambda} \neq \lambda_{\text{best}}$ the difference in their true mean-squared errors is not very drastic, so the minimum mean-squared error is nearly attained in almost all the situations that we simulated.

9.1. *Algorithm 3: estimation and inference of pFDR(γ) and FDR(γ) with optimal λ*

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

$$\widehat{\text{MSE}}(\lambda) = \frac{1}{B}\sum_{b=1}^{B}[\widehat{\text{pFDR}}_\lambda^{*b}(\gamma) - \min_{\lambda' \in R}\{\widehat{\text{pFDR}}_{\lambda'}(\gamma)\}]^2. \tag{28}$$

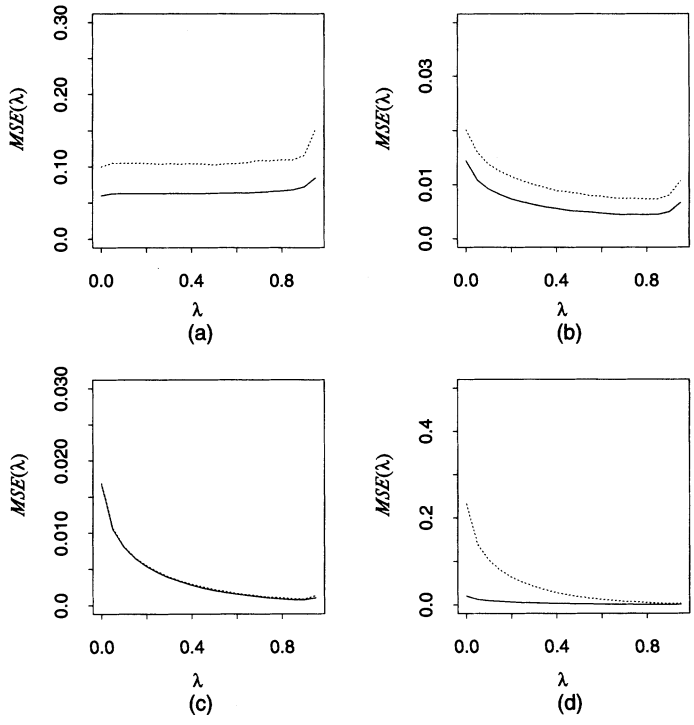

Fig. 5. Plots of MSE(λ) *versus* λ for $\Gamma = [2, \infty)$ for (a) $\pi_0 = 1$, (b) $\pi_0 = 0.8$, (c) $\pi_0 = 0.5$ and (d) $\pi_0 = 0.2$ (———, true mean-squared error; · · · · · · · , mean-squared error predicted by the bootstrap procedure averaged over 100 applications)

(d) Set $\hat{\lambda} = \arg\min_{\lambda \in R}\{\widehat{\mathrm{MSE}}(\lambda)\}$. Our overall estimate of pFDR(γ) is

$$\widehat{\mathrm{pFDR}}(\gamma) = \widehat{\mathrm{pFDR}}_{\hat{\lambda}}(\gamma). \tag{29}$$

(e) Form a $1 - \alpha$ upper confidence interval of $\widehat{\mathrm{pFDR}}(\gamma)$ by taking the $1 - \alpha$ quantile of $\{\widehat{\mathrm{pFDR}}^{*1}_{\hat{\lambda}}(\gamma), \ldots, \widehat{\mathrm{pFDR}}^{*B}_{\hat{\lambda}}(\gamma)\}$ as the upper end point (the lower end point being 0).

(f) In estimating FDR, perform this same procedure with $\widehat{\mathrm{FDR}}(\gamma)$ instead.

10. Discussion

In this paper, we have proposed a new approach to multiple-hypothesis testing. Instead of setting the error rate at a particular level and estimating the rejection region, we have proposed fixing the rejection region and estimating the error rate. This approach allows a more straightforward analysis of the problem. We have seen that the result is a more powerful and applicable methodology. For example, we estimated a new definition, the positive false discovery rate, one which we argued is usually much more appropriate. And we successfully 'controlled' it. By using theoretical results about pFDR with fixed rejection regions, we could derive well-behaved estimates of both pFDR and FDR. Interestingly, the Benjamini and Hochberg (1995) step-up method naturally falls out of these results.

Everything that we have discussed in this paper has been under the assumption that we are working with independent p-values. In more general cases, such as with dependence or in

Table 3. Simulation results for the bootstrap procedure to pick the optimal λ

π_0	m	*Cut point*	λ_{best}	$\hat{\lambda}$	$MSE(\lambda_{\text{best}})$	$MSE(\hat{\lambda})$
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

nonparametric situations, it is possible to apply very similar ideas to obtain accurate estimates of pFDR and FDR. See Storey and Tibshirani (2001) for a treatment of this. There are several other open questions that this approach brings to light. Other, better, estimates may be available. One could also possibly prove optimality theorems with respect to estimating pFDR within certain frameworks.

The q-value was discussed, which is the pFDR analogue of the p-value. Whereas it can be inconvenient to have to fix the rejection region or the error rate beforehand, the q-value requires us to do neither. By developing our methodology with fixed rejection regions, we could formulate the q-value in a conceptually simple manner. As an open problem, it is of interest to investigate the operating characteristics of the calculated q-values, which we called $\hat{q}$.

In a very interesting paper, Friedman (2001) discusses the role that statistics can play in the burgeoning field of data mining. Data mining involves investigating huge data sets in which 'interesting' features are discovered. The classical example is determining which products tend to be purchased together in a grocery store. Often the rules for determining interesting features have no simple statistical interpretation. It is understandable that hypothesis testing has not played a major role in this field because, the more hypotheses we have, the less power there is to discover effects. The methodology presented here has the opposite property—the more tests we perform, the better the estimates are. Therefore, it is an asset under this approach to have large data sets with many tests. The only requirement is that the tests must be exchangeable in the sense that the p-values have the same null distribution.

Even if the tests are dependent, our approach can be fully applied. It was shown in Storey (2001) that the effect of dependence is negligible if m is large for a large class of dependence. Also, Storey and Tibshirani (2001) treated the case where dependence cannot be ignored. Therefore, we hope that this proposed multiple-hypothesis testing methodology is useful not only in fields like genomics or wavelet analysis but also in the field of data mining where it is desired to find several interesting features out of many, while limiting the rate of false positive findings.

Acknowledgements

Thanks go to Bradley Efron and Ji Zhu for helpful comments and suggestions. I am especially grateful for the ideas and encouragement of my advisor, Robert Tibshirani. This research was supported in part by a National Science Foundation graduate research fellowship.

Appendix A: Remarks and proofs

Remark 1. Here, we explain why rejection regions for p-values should be of the form $[0, \gamma]$. Recall that, for a nested set of rejection regions $\{\Gamma\}$, the p-value of $X = x$ is defined to be

$$p\text{-value}(x) = \inf_{\{\Gamma : x \in \Gamma\}} \{\Pr(X \in \Gamma | H = 0)\}. \tag{30}$$

Therefore, for two p-values p_1 and p_2, $p_1 \leqslant p_2$ implies that the respective observed statistics x_1 and x_2 are such that $x_2 \in \Gamma$ implies $x_1 \in \Gamma$. Therefore, whenever p_2 is rejected, p_1 should also be rejected.

Proof of theorem 1. See Storey (2001) for a proof of theorem 1.

Proof of theorem 2. Recall $\widehat{\text{pFDR}}_\lambda(\gamma)$ from equation (9). Also note that

$$\text{pFDR}(\gamma) = \frac{1}{\Pr\{R(\gamma) > 0\}} E\left\{\frac{V(\gamma)}{R(\gamma) \vee 1}\right\}. \tag{31}$$

Therefore,

$$E\{\widehat{\text{pFDR}}_\lambda(\gamma)\} - \text{pFDR}(\gamma) \geqslant E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\} \Pr\{R(\gamma) > 0\}}\right], \tag{32}$$

since $\Pr\{R(\gamma) > 0\} \geqslant 1 - (1-\gamma)^m$ under independence. Conditioning on $R(\gamma)$, it follows that

$$E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\} \Pr\{R(\gamma) > 0\}} \middle| R(\gamma)\right] = \frac{[E\{W(\lambda)|R(\gamma)\}/(1-\lambda)]\gamma - E\{V(\gamma)|R(\gamma)\}}{\{R(\gamma) \vee 1\} \Pr\{R(\gamma) > 0\}}. \tag{33}$$

By independence, $E\{W(\lambda)|R(\gamma)\}$ is a linear non-increasing function of $R(\gamma)$, and $E\{V(\gamma)|R(\gamma)\}$ is a linear non-decreasing function of $R(\gamma)$. Thus, by Jensen's inequality on $R(\gamma)$ it follows that

$$E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{R(\gamma) \Pr\{R(\gamma) > 0\}} \middle| R(\gamma) > 0\right] \geqslant \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{E\{R(\gamma)|R(\gamma) > 0\} \Pr\{R(\gamma) > 0\}}. \tag{34}$$

Since $E\{R(\gamma)\} = E\{R(\gamma)|R(\gamma) > 0\} \Pr\{R(\gamma) > 0\}$, it follows that

$$\frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{E\{R(\gamma)|R(\gamma) > 0\} \Pr\{R(\gamma) > 0\}} = \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{E\{R(\gamma)\}}. \tag{35}$$

Also, note that

$$E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\} \Pr\{R(\gamma) > 0\}} \middle| R(\gamma) = 0\right] = E\left[\frac{W(\lambda)\gamma}{(1-\lambda) \Pr\{R(\gamma) > 0\}} \middle| R(\gamma) = 0\right] \tag{36}$$

$$\geqslant E\left[\frac{W(\lambda)\gamma}{(1-\lambda)\, E\{R(\gamma)\}} \middle| R(\gamma) = 0\right], \tag{37}$$

where the last inequality holds since $E\{R(\gamma)\} \geqslant \Pr\{R(\gamma) > 0\}$. Therefore,

$$E\{\widehat{\text{pFDR}}_\lambda(\gamma)\} - \text{pFDR}(\gamma) \geqslant E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\} \Pr\{R(\gamma) > 0\}}\right] \tag{38}$$

$$\geqslant \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{E\{R(\gamma)\}} \Pr\{R(\gamma) > 0\} \tag{39}$$

$$+ \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma|R(\gamma) = 0]}{E\{R(\gamma)\}} \Pr\{R(\gamma) = 0\} \tag{40}$$

$$= \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)]}{E\{R(\gamma)\}}. \tag{41}$$

Now

$$E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)] \geqslant \{m\pi_0(1-\lambda)/(1-\lambda)\}\gamma - m\pi_0\gamma = 0. \tag{42}$$

Thus, $E\{\widehat{\text{pFDR}}_\lambda(\gamma)\} - \text{pFDR}(\gamma) \geqslant 0$.

Since we showed that

$$E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\}\Pr\{R(\gamma) > 0\}}\right] \geqslant 0, \tag{43}$$

and

$$\frac{1}{\Pr\{R(\gamma) > 0\}}[E\{\widehat{\text{FDR}}_\lambda(\gamma)\} - \text{FDR}(\gamma)] = E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\}\Pr\{R(\gamma) > 0\}}\right], \tag{44}$$

it follows that $E\{\widehat{\text{FDR}}_\lambda(\gamma)\} \geqslant \text{FDR}(\gamma)$.

Remark 2. It was implicitly used in the previous proof that the H_i are random. However, this assumption is unnecessary, and in fact the assumption that the alternative statistics are independent is also unnecessary. See Storey and Tibshirani (2001) for a proof of theorem 2 under these weaker assumptions.

Proof of theorem 3. The proof of theorem 3 easily follows by noting that

$$E[\{\widehat{\text{pFDR}}_\lambda(\gamma) - \text{pFDR}(\gamma)\}^2 | \widehat{\text{pFDR}}_\lambda(\gamma) > 1] > E[\{\widehat{\text{pFDR}}_\lambda(\gamma) \wedge 1 - \text{pFDR}(\gamma)\}^2 | \widehat{\text{pFDR}}_\lambda(\gamma) > 1] \tag{45}$$

since $\text{pFDR}(\gamma) \leqslant 1$. The proof for $\widehat{\text{FDR}}_\lambda(\gamma)$ follows similarly.

Proof of theorem 4. Recall $\widehat{\text{pFDR}}_\lambda(\gamma)$ from equation (9). By the strong law of large numbers, $\widehat{\Pr}(P \leqslant \gamma) \to \Pr(P \leqslant \gamma)$ almost surely. Also, $\Pr(P \geqslant \lambda | H = 0) = 1 - \lambda$ and $\Pr(P \geqslant \lambda | H = 1) = 1 - g(\lambda)$, where $g(\lambda)$ is the power of rejecting over $[0, \lambda]$ as described in Section 6. Therefore, by the strong law of large numbers $W(\lambda)/m \to (1-\lambda)\pi_0 + \{1 - g(\lambda)\}\pi_1$ almost surely. Thus, it follows that

$$\begin{aligned}\lim_{m\to\infty}\{\widehat{\text{pFDR}}_\lambda(\gamma)\} &= \frac{[\pi_0 + \pi_1\{1-g(\lambda)\}/(1-\lambda)]\gamma}{\Pr(P \leqslant \gamma)} \\ &= \frac{\pi_0 + \pi_1\{1-g(\lambda)\}/(1-\lambda)}{\pi_0}\text{pFDR}(\gamma) \geqslant \text{pFDR}(\gamma).\end{aligned} \tag{46}$$

Proof of corollary 1. Since $g(\lambda)$ is concave in λ, $\{1 - g(\lambda)\}/(1-\lambda)$ is non-increasing in λ. Therefore, the minimum of $\{1 - g(\lambda)\}/(1-\lambda)$ is obtained at $\lim_{\lambda\to 1}[\{1 - g(\lambda)\}/(1-\lambda)]$. By L'Hopital's rule,

$$\lim_{\lambda\to 1}[\{1 - g(\lambda)\}/(1-\lambda)] = g'(1).$$

Proof of theorem 5. We can observe both $R(\gamma)$ and $R(\lambda)$. Under the assumptions of theorem 1, it follows that the likelihood of the data can be written as

$$\{\pi_0\gamma + (1-\pi_0)\,g(\gamma)\}^{R(\gamma)}\{1 - \pi_0\gamma - (1-\pi_0)\,g(\gamma)\}^{m-R(\gamma)}, \tag{47}$$

and also

$$\{\pi_0\lambda + (1-\pi_0)\,g(\lambda)\}^{R(\lambda)}\{1 - \pi_0\lambda - (1-\pi_0)\,g(\lambda)\}^{m-R(\lambda)}. \tag{48}$$

The result follows from standard methods.

Remark 3. If $g(\cdot)$ is known then the maximum likelihood estimate of $\text{pFDR}(\gamma)$ is

$$\widetilde{Q}(\gamma) = \frac{\widetilde{\pi}_0\gamma}{\hat{F}(\gamma)} = \frac{\{g(\gamma) - R(\gamma)/m\}\gamma}{\{g(\gamma) - \gamma\}R(\gamma)/m}. \tag{49}$$

References

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

———(2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.

Benjamini, Y. and Liu, W. (1999) A step-down multiple-hypothesis procedure that controls the false discovery rate under independence. *J. Statist. Planng Inf.*, **82**, 163–170.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

Friedman, J. H. (2001) The role of statistics in the data revolution? *Int. Statist. Rev.*, **69**, 5–10.

Storey, J. D. (2001) The positive False Discovery Rate: a Bayesian interpretation and the q-value. To be published. (Available from `http://www-stat.stanford.edu/~jstorey/`.)

Storey, J. D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. To be published. (Available from `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. *Proc. Natn. Acad. Sci. USA*, **98**, 5116–5121.

Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Planng Inf.*, **82**, 171–196.