Correcting false discovery rates for their bias toward false positives

David R. Bickel

February 12, 2016

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

Abstract

Conventional methods of adjusting p values for multiple comparisons seek to control a family-wise error rate (FWER) such as a genome-wise error rate. The recognition that they lead to excessive false negative rates in genomics applications has led to widespread use of false discovery rates (FDRs) in place of the conventional adjustments. While this is an improvement, the way FDRs are used in the analysis of genomics data leads to the opposite problem, excessive false positive rates. In this sense, the FDR overcorrects for the excessive conservatism (bias toward false negatives) of the FWER-controlling methods of adjusting p values.

Estimators of the local FDR (LFDR) are much less biased but have not been widely adopted because they have high variance compared to estimated FDRs. To reduce that variance, we propose estimating the LFDR by correcting an estimated FDR or the level at which an FDR is controlled.

Keywords: Bayesian false discovery rate; empirical Bayes; local false discovery rate; principle of maximum entropy; multiple comparison procedure; multiple testing

1 Introduction

1.1 Excessive conservatism of multiple comparison procedures

The two definitions of a false-positive rate most commonly seen in genomics applications are the family-wise error rate (FWER) and the false discovery rate (FDR). The former is designed to determine whether any of the hypotheses tested in the family or reference set has a false positive in the study, whereas the latter tolerates a few false positives. The traditional approach of controlling the FWER to be below some significance level would be appropriate were the goal to prevent any false positives. On the other hand, controlling the FDR permits some false positives in order to greatly increase the number of true positives, provided that the probability that a discovery is false is kept below some significance level.

As a result, the control of the FDR (Benjamini and Hochberg, 1995) or a related quantity (e.g., Korn et al., 2004; Van der Laan et al., 2004; Genovese and Wasserman, 2006; Pawitan et al., 2006; Farcomeni, 2008) is increasingly recommended as an alternative to FWER control (e.g., Van Den Oord, 2008; Dudoit and van der Laan, 2008; Glickman et al., 2014). While the FDR indeed overcomes the FWER's bias toward false negatives, it does so only by incurring a bias toward false positives.

1.2 Insufficient conservatism of false discovery rates

Suppose all null hypotheses are rejected that have p values below α, a significance level such as 0.01. The simplest FWER is $\mathrm{FWER}(\alpha)$, the frequentist probability that at least one null hypothesis is rejected at level α assuming that all null hypotheses are true (Dudoit and van der Laan, 2008). The simplest FDR, called the nonlocal FDR $\mathrm{NFDR}(\alpha)$ (Bickel, 2013), is the empirical Bayes probability that a null hypothesis rejected at level α is true given the

data (Efron, 2010). Let p_i denote the p value of the ith hypothesis test. The adjusted p value or *achieved FWER* for the ith test is $\mathrm{FWER}(p_i)$, the FWER at the significance level at which the ith null hypothesis is barely rejected ($\alpha = p_i$). Likewise, the *achieved NFDR* for the ith test is $\mathrm{NFDR}(p_i)$, the probability that a randomly selected p value less than p_i corresponds to a true null hypothesis. That is misleading as a measure of the significance of the ith hypothesis test since it can be substantially lower than the posterior probability that the null hypothesis is true. That posterior probability is the local false discovery rate (LFDR). In fact, the achieved NFDR of the ith null hypothesis is the expectation value of all LFDRs corresponding to p values less than p_i (Efron, 2010). That is why estimates of NFDRs are often much less than estimates of LFDRs (e.g., Hong et al., 2009; Bickel, 2012, Fig. 3), leading to unwarranted reports of significance and potentially many more false positives.

This anti-conservative bias of $\widehat{\mathrm{NFDR}}(p_i)$, an estimator of $\mathrm{NFDR}(p_i)$ defined mathematically in Section 2, undermines the most common use of the FDR in genomics research. A researcher seeking to apply a multiple-comparison procedure with a list of p values typically selects either a method of FWER control or a method FDR control (Benjamini and Hochberg, 1995). With the former now known for its excessive conservatism in genomics-scale applications (Dudoit and van der Laan, 2008, p. 145), the latter is now seen as a viable alternative, as indicated both by its availability in data analysis software and by the literature (e.g., Van Den Oord, 2008; Glickman et al., 2014). The original method of FDR control (Benjamini and Hochberg, 1995) remains the most commonly used. At significance level α, $\mathrm{FDR}(\alpha)$, the false discovery rate as defined by Benjamini and Hochberg (1995), tends to be similar in value to $\mathrm{NFDR}(\alpha)$, which is conceptually more relevant to Bayesian decision theory. In fact, with $\widehat{\mathrm{NFDR}}(\alpha)$ as the natural estimator of $\mathrm{NFDR}(\alpha)$, the practice of setting

α to the smallest value such that $\widehat{\text{NFDR}}(\alpha) \leq q$ for some specified value q guarantees that $\text{FDR}(\alpha) \leq q$ under the Benjamini and Hochberg (1995) independence assumption (Efron, 2010, pp. 53-54). Unfortunately, that means the rejected hypothesis with $p_{\max,q}$, the highest p value, is rejected merely because $\widehat{\text{NFDR}}(p_{\max,q}) \leq q$ even though $\text{NFDR}(p_{\max,q})$ is the probability of the truth of a random null hypothesis with a p value *less than* $p_{\max,q}$ rather than *equal to* $p_{\max,q}$.

As the flaw in the procedure is practical rather than mathematical, it may be most clearly seen in terms of particular applications to genomics data. Suppose the null hypothesis is that a gene is not differentially expressed. To the extent that $\widehat{\text{NFDR}}(p_{\max,q})$ accurately estimates $\text{NFDR}(p_{\max,q})$, selecting a gene for further consideration on the basis of FDR control in effect selects a gene on the basis of the probability that a more significant gene is differentially expressed. That is equivalent to inviting a job candidate to an interview on the basis of an average of his or her test score with the test scores of all the better applicants.

Later methods of FDR control do not necessarily fare any better. The opposite is often the case since many such methods are designed to be even less conservative than the Benjamini and Hochberg (1995) procedure (e.g., Benjamini and Liu, 1999), resulting in even more unwarranted discoveries and the accompanying bias toward false positives.

1.3 Insufficient stability of local false discovery rates

It would then appear that LFDR estimation provides a practical balance between unnecessarily stringent adjustments for multiple comparisons to control the FWER and overly permissive methods that only control the FDR. However, a reason that estimators of the LFDR have not been adopted for standard reporting in genetics applications in spite of other advantages (Bickel, 2013, Table 1) is that they have excessive instability as quantified

by their variance (Dudoit and van der Laan, 2008, p. 316). This instability results from the dependence of LFDR estimates on estimates of a ratio of probability densities of test statistics (Genovese and Wasserman, 2003).

This instability of LFDR methods is a valid concern, for without a way to make LFDR methods as stable as FDR methods, they are unlikely to be widely adopted by genomics researchers. Fortunately, the stability of LFDR estimation can be increased by correcting an estimated FDR or NFDR, the level q at which an FDR is controlled, or the closely related estimated q value (Storey, 2002). The strategy in Section 3 is not to replace FDR methods but rather to take advantage of their stability while correcting their bias toward false positives. With the principle of maximum entropy, the strategy results in two new estimators of the LFDR, one of which is related to the rank-doubling estimator by Bickel (2013).

For example, the achieved NFDR estimate corresponding to $p\left(x_{(i)}\right)$, the ith smallest p value, is multiplied by a simple sum to become the corrected FDR,

$$\mathrm{CFDR}\left(x_{(i)}\right) = \left(\sum_{k=1}^{i} \frac{1}{i-k+1}\right) \widehat{\mathrm{NFDR}}\left(p\left(x_{(i)}\right)\right), \tag{1}$$

a generalization of which is derived in Section 3. Wide discrepancies between the new estimators and $\widehat{\mathrm{NFDR}}\left(p\left(x_{(i)}\right)\right)$ are revealed in applications to a biomedical data set and a gene expression data set, illustrating the extent to which FDR methods require bias correction.

2 False discovery rates

2.1 Basic definitions and notation

Let A_i stand for the random quantity that is equal to 1 if the ith feature is affected by a treatment, associated with a disease, etc. If the feature is unaffected (or unassociated), then A_i is equal to 0. In general, $A_i = 1$ if the ith alternative hypothesis is true, but $A_i = 0$ if the ith null hypothesis is true.

The *local false discovery rate* (LFDR) is the posterior probability that null hypothesis i is true given the p value:

$$\text{LFDR}\left(p\left(x_i\right)\right) \quad = \quad P\left(A_i = 0 | p\left(X_i\right) = p\left(x_i\right)\right). \tag{2}$$

In the gene expression case, the LFDR is a conditional probability of the hypothesis that gene i is not differentially expressed given the p value. Conversely, $1 - \text{LFDR}\left(p\left(x_i\right)\right)$ is $P\left(A_i = 1 | p\left(X_i\right) = p\left(x_i\right)\right)$, the posterior probability of the hypothesis that feature i is differentially expressed.

Another "Bayesian false discovery rate" (Efron and Tibshirani, 2002) is the *nonlocal false discovery rate* (NFDR) (Bickel, 2013) at a significance level of α:

$$\text{NFDR}\left(\alpha\right) = P\left(A_i = 0 | p\left(X_i\right) \leq \alpha\right).$$

The NFDR evaluated at significance level α is equal to the average LFDR over all genes with p values less than α. That is to say,

$$\text{NFDR}\left(\alpha\right) = P\left(A_i = 0 | p\left(X_i\right) \leq \alpha\right)$$

$$= E\left(P\left(A_i = 0|p\left(X_i\right)\right)|p\left(X_i\right) \leq \alpha\right) = E\left(\text{LFDR}\left(p\left(X_i\right)\right)|p\left(X_i\right) \leq \alpha\right), \tag{3}$$

by the law of total probability (Efron, 2010). The NFDR may be estimated by this ratio of the estimated average number of false discoveries to the number of null hypotheses rejected at significance level α:

$$\widehat{\text{NFDR}}\left(\alpha\right) \quad = \quad \begin{cases} \frac{\alpha d}{\#(p(x_i) \leq \alpha)} & \text{if } \frac{\alpha d}{\#(p(x_i) \leq \alpha)} \leq 1 \\ 1 & \text{if } \frac{\alpha d}{\#(p(x_i) \leq \alpha)} > 1, \end{cases} \tag{4}$$

where the dimension d is number of tests performed.

2.2 Misleading usage of false discovery rates

Many software packages substitute each p value for α in equations (3) and (4), yielding the achieved FDR and its estimate for test j:

$$\text{NFDR}\left(p\left(x_j\right)\right) = P\left(A_i = 0|p\left(X_i\right) \leq p\left(x_j\right)\right)$$

$$= E\left(P\left(A_i = 0|p\left(X_i\right)\right)|p\left(X_i\right) \leq p\left(x_j\right)\right) = E\left(\text{LFDR}\left(p\left(X_i\right)\right)|p\left(X_i\right) \leq p\left(x_j\right)\right); \tag{5}$$

$$\widehat{\text{NFDR}}\left(p\left(x_j\right)\right) = \begin{cases} \frac{p(x_j)d}{\#(p(x_i) \leq p(x_j))} & \text{if } \frac{p(x_j)d}{\#(p(x_i) \leq p(x_j))} < 1 \\ 1 & \text{if } \frac{p(x_j)d}{\#(p(x_i) \leq p(x_j))} > 1. \end{cases} \tag{6}$$

However, that can be misleadingly low, for equation (5) says the achieved FDR of the test corresponding to the p value $p\left(x_j\right)$ is the average posterior probability that the null hypothesis is true given that the p value is less than $p\left(x_j\right)$. That probability is often much less than $\text{LFDR}\left(p\left(x_j\right)\right)$, the posterior probability that the null hypothesis is true given that

the p value is equal to $p(x_j)$; in short, $\text{NFDR}(p(x_j)) \ll \text{LFDR}(p(x_j))$.

As a result, to the extent that $\widehat{\text{NFDR}}(p(X_j))$ is an unbiased estimate of $\text{NFDR}(p(X_j))$, its negative bias tends to be excessive as an estimator of $\text{LFDR}(p(X_j))$, the quantity most relevant to deciding whether to reject the jth null hypothesis given the data. Under the assumption that $\widehat{\text{NFDR}}(p(X_j))$ and $\widehat{\text{LFDR}}(p(X_j))$ are unbiased estimators of $\text{NFDR}(p(X_j))$ and $\text{LFDR}(p(X_j))$, respectively, the bias in $\widehat{\text{NFDR}}(p(X_j))$ as an estimate of $\text{LFDR}(p(X_j))$ is

$$E\left(\widehat{\text{NFDR}}(p(X_j)) - \text{LFDR}(p(X_j))\right)$$

$$\begin{aligned} &= E\left(\widehat{\text{NFDR}}(p(X_j)) - \text{NFDR}(p(X_j))\right) + E(\text{NFDR}(p(X_j)) - \text{LFDR}(p(X_j))) \\ &= 0 + E(\text{NFDR}(p(X_j)) - \text{LFDR}(p(X_j))) \ll 0. \end{aligned}$$

The practical consequence of this negative bias in the case of gene expression data is that more genes will be considered differentially expressed on the basis of a low $\text{NFDR}(p(x_j))$ than is warranted since the probability that the null hypothesis is true tends to be higher than $\text{NFDR}(p(x_j))$.

That bias toward false positives occurs not only for estimates of achieved NFDRs but also for various procedures that control the FDR (§1.2). Fortunately, the bias can be corrected.

3 Salvaging false discovery rates

3.1 Inferring local false discovery rates from a false discovery rate

Equation (3) constrains the conditional expectation value of the LFDR to be the NFDR: $E\left(\mathrm{LFDR}\left(p\left(X_i\right)\right)|p\left(X_i\right)\leq\alpha\right)=\mathrm{NFDR}\left(\alpha\right)$. With the additional constraint that $\mathrm{LFDR}\left(p\left(X_i\right)\right)\geq 0$, the maximum entropy principle yields $\mathrm{Expon}\left(\left(\mathrm{NFDR}\left(\alpha\right)\right)^{-1}\right)$, the $\left(\mathrm{NFDR}\left(\alpha\right)\right)^{-1}$-rate exponential distribution of the LFDR given $p\left(X_i\right)\leq\alpha$:

$$\mathrm{LFDR}\left(p\left(X_i\right)\right)|p\left(X_i\right)\leq\alpha\sim f_{\alpha}^{\star}\left(\bullet\right)\coloneqq\underset{f(\bullet)}{\arg\max}\left(-\int_0^{\infty}f\left(L\right)\ln f\left(L\right)dL\right)=\frac{e^{-\frac{\bullet}{\mathrm{NFDR}(\alpha)}}}{\mathrm{NFDR}\left(\alpha\right)},\tag{7}$$

where L is a dummy variable for the LFDR, and each $f\left(\bullet\right)$ is a probability density function satisfying $\int_0^{\infty}f\left(L\right)dL=1$ and $\int_0^{\infty}Lf\left(L\right)dL=\mathrm{NFDR}\left(\alpha\right)$. Thus, if $p\left(x_i\right)\leq\alpha$, then the conditional probability density of $\mathrm{LFDR}\left(p\left(x_i\right)\right)$ given $p\left(X_i\right)\leq\alpha$ is $f_{\alpha}^{\star}\left(\mathrm{LFDR}\left(p\left(x_i\right)\right)\right)=e^{-\frac{\mathrm{LFDR}(p(x_i))}{\mathrm{NFDR}(\alpha)}}/\,\mathrm{NFDR}\left(\alpha\right)$.

Under the additional constraint that $\mathrm{LFDR}\left(p\left(X_i\right)\right)\leq 1$, the distribution that maximizes the entropy becomes more complicated (Conrad, 2014). Since that distribution is very closely approximated by equation (7) if $\alpha<1/2$, it does not provide a practical benefit for tests at any significance level much less than 0.5.

Consider $\mathrm{LFDR}_{(1)},\ldots,\mathrm{LFDR}_{(m)}$, the order statistics of the m independent draws $\mathrm{LFDR}\left(p\left(X_1\right)\right)$, $\ldots$, $\mathrm{LFDR}\left(p\left(X_m\right)\right)$ from $\mathrm{Expon}\left(\left(\mathrm{NFDR}\left(\alpha\right)\right)^{-1}\right)$. The ith expected order statistic is known for the exponential distribution (e.g., Cox and Hinkley, 1979) to be

$$E\left(\mathrm{LFDR}_{(i)}\,|p\left(X_1\right),\ldots,p\left(X_m\right)\leq\alpha\right)=\left(\sum_{k=1}^{m}\frac{1}{m-k+1}\right)\mathrm{NFDR}\left(\alpha\right).\tag{8}$$

Let $m(\alpha)$ denote the number of hypotheses rejected at level α:

$$m(\alpha) = \left|\{p(x_i) : i = 1, \ldots, d; p(x_i) \leq \alpha\}\right|, \tag{9}$$

assuming there are no ties. With $p\left(x_{(1)}\right), \ldots, p\left(x_{(m(\alpha))}\right)$ as the $m(\alpha)$ observed order statistics of the p values corresponding to rejected null hypotheses, equation (8) suggests estimating $\text{LFDR}\left(p\left(x_{(i)}\right)\right)$, the LFDR corresponding to ith-highest such p value, by

$$\widehat{\text{LFDR}}^{\star}\left(x_{(i)}; q\right) \coloneqq \left(\sum_{k=1}^{m} \frac{1}{m-k+1}\right) q \tag{10}$$

for $i = 1, \ldots, m(\alpha)$, where, for now, $q = \widehat{\text{NFDR}}(\alpha)$ and $m = m(\alpha)$. The $\star$ indicates dependence on the maximum entropy solution (7).

As FDR and NFDR methods analyze data much more similarly to each other than to LFDR estimators, the proposed method is not limited to the NFDR. Rather than setting $q = \widehat{\text{NFDR}}(\alpha)$, the q in equation (10) may instead be a level at which the FDR is controlled according to a method that determines which null hypotheses to reject, as in Section 1.2. The number of null hypotheses rejected by the method is the m in equation (10). In some cases, the level of FDR control is identical to an achieved NFDR estimate (§1.2).

Example 1. Of the $d = 15$ hypotheses on thrombolytic-treatment outcomes tested by Neuhaus et al. (1992), 4 are rejected when controlling the FDR at the 0.05 level (Benjamini and Hochberg, 1995). However, their expected order statistics of the LFDR according to $f^{\star}_{0.05}$ are $\widehat{\text{LFDR}}^{\star}\left(x_{(1)}; 0.05\right) = 0.012$, $\widehat{\text{LFDR}}^{\star}\left(x_{(2)}; 0.05\right) = 0.029$, $\widehat{\text{LFDR}}^{\star}\left(x_{(3)}; 0.05\right) = 0.054$, and $\widehat{\text{LFDR}}^{\star}\left(x_{(4)}; 0.05\right) = 0.104$, only 2 of which are below 0.05. That the hypotheses with estimated LFDRs of 0.054 and 0.104 would not be rejected merely because the mean of

0.012, 0.029, 0.054, and 0.104 is 0.05 highlights the problem identified in above (§1.2, §2.2).

▲

3.2 Correcting and re-ranking false discovery rates

Let $q(\alpha)$ denote the smallest value of q such that all hypothesis with p values in $[0, \alpha]$ are rejected according to some procedure that guarantees that the FDR, NFDR, or an estimate of either is no higher than q. The substitutions $q = q(\alpha)$ and $m = m(\alpha)$ (9) turn equation (10) into

$$\widehat{\text{LFDR}}^{\star}\left(x_{(i)}; q(\alpha), \alpha\right) \coloneqq \left(\sum_{k=1}^{m(\alpha)} \frac{1}{m(\alpha) - k + 1}\right) q(\alpha). \tag{11}$$

As explained in Sections 1.2 and 2.2, $q\left(p\left(x_{(i)}\right)\right)$ would be overly optimistic as an estimator of $\text{LFDR}\left(x_{(i)}\right)$. It does not follow that $q\left(p\left(x_{(i)}\right)\right)$ is useless, for its bias may be corrected by equation (11):

$$\begin{aligned} \text{CFDR}\left(x_{(i)}\right) &\coloneqq \widehat{\text{LFDR}}^{\star}\left(x_{(i)}; q\left(p\left(x_{(i)}\right)\right), p\left(x_{(i)}\right)\right) \\ &= \left(\sum_{k=1}^{m\left(p\left(x_{(i)}\right)\right)} \frac{1}{m\left(p\left(x_{(i)}\right)\right) - k + 1}\right) q\left(p\left(x_{(i)}\right)\right) \\ &= \left(\sum_{k=1}^{i} \frac{1}{i - k + 1}\right) q\left(p\left(x_{(i)}\right)\right) \end{aligned} \tag{12}$$

since $m\left(p\left(x_{(i)}\right)\right) = i$ according to equation (9) as, under the assumption that there are no ties among the p values, $p\left(x_{(i)}\right)$ is the p value of rank i. The quantity $\widehat{\text{LFDR}}^{\star}\left(x_{(i)}; q\left(p\left(x_{(i)}\right)\right), p\left(x_{(i)}\right)\right)$ is called the *corrected false discovery rate* (CFDR) of the ith smallest p value and is abbreviated by $\text{CFDR}\left(x_{(i)}\right)$. The special case with $q\left(p\left(x_{(i)}\right)\right) = \widehat{\text{NFDR}}\left(p\left(x_{(i)}\right)\right)$ appears in equation (1). Like other false discovery rate estimates (Efron, 2010), CFDR values greater

than 1 are reset to 1.

A different estimator of the LFDR arises from the simplifying assumption that the p values and LFDRs are similarly ordered in the sense that they have the same ranks. Let $F^{\star}_{\alpha}$ denote the cumulative distribution function of the maximum-entropy LFDR: $F^{\star}_{\alpha}(L^{\star}) = \int_0^{L^{\star}} f^{\star}_{\alpha}(L)\, dL$ for $0 \leq L^{\star} \leq 1$. Then, for any α, the quantile rank of the NFDR is $F^{\star}_{\alpha}(\text{NFDR}(\alpha)) = 1 - e^{-1}$ since $F^{\star}_{\alpha}$ is exponential with rate $(\text{NFDR}(\alpha))^{-1}$. As $1 - e^{-1}$ is constant in α, $F^{\star}_{\alpha}(\text{NFDR}(\alpha))$ will be abbreviated by $F^{\star}(\text{NFDR})$.

Let $[i/F^{\star}(\text{NFDR})]$ denote the closest integer to $i/F^{\star}(\text{NFDR})$, with $i = 1, 2, \ldots$ small enough that $[i/F^{\star}(\text{NFDR})] \leq d$. For large i, the expected value of a single LFDR drawn from $F^{\star}_{p\left(x_{([i/F^{\star}(\text{NFDR})])}\right)}$ is $\text{NFDR}\left(p\left(x_{([i/F^{\star}(\text{NFDR})])}\right)\right)$, which, with high probability, is approximately $\text{LFDR}\left(p\left(x_{(i)}\right)\right)$ since $\text{LFDR}_{(i)} = \text{LFDR}\left(p\left(x_{(i)}\right)\right)$ by the assumed equality of ranks and since $\text{LFDR}_{(i)}$ is approximately equal to the $F^{\star}(\text{NFDR})$-quantile of $F^{\star}_{p\left(x_{([i/F^{\star}(\text{NFDR})])}\right)}$ with high probability. The *re-ranked false discovery rate* (RFDR) is accordingly defined by:

$$\text{RFDR}\left(x_{(i)}\right) := \begin{cases} q\left(p\left(x_{([i/F^{\star}(\text{NFDR})])}\right)\right) & \text{if } [i/F^{\star}(\text{NFDR})] \leq d \\ 1 & \text{if } [i/F^{\star}(\text{NFDR})] > d. \end{cases} \tag{13}$$

This is identical to the LFDR estimator $\hat{\psi}\left(x_{(i)}\right)$ proposed by Bickel (2013, p. 537) except that $\hat{\psi}\left(x_{(i)}\right)$ uses $1/2$ in place of $F^{\star}(\text{NFDR})$ (Bickel, 2015). The fact that $F^{\star}(\text{NFDR}) > 1/2$ means $\hat{\psi}\left(x_{(i)}\right)$ is based on a p value of higher rank than that of $\text{RFDR}\left(x_{(i)}\right)$, suggesting that $\hat{\psi}\left(x_{(i)}\right)$ has a positive bias.

Example 2. The p values from Example 1 yield the achieved NFDR estimates (6) and the corresponding LFDR estimates (12)-(13) displayed in the left panel of Figure 1. Using a significance threshold of 0.05 results in the rejection of 4 hypotheses $\left(\widehat{\text{NFDR}}\left(p\left(x_{(i)}\right)\right) \leq 0.05\right)$,

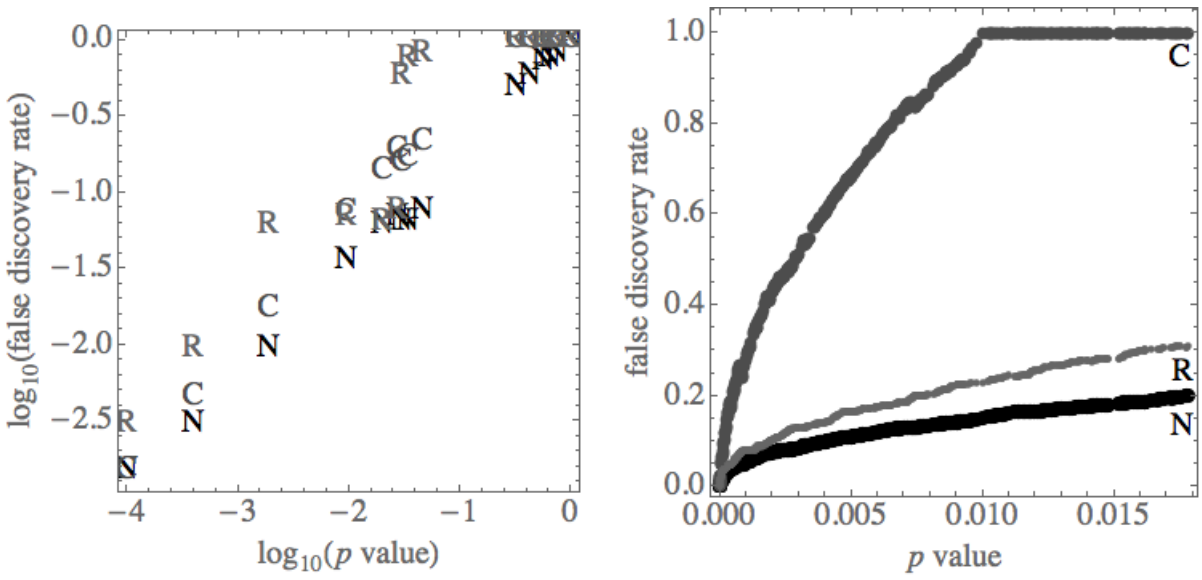

Figure 1: Local ("C"; "R") and nonlocal ("N") false discovery rate estimates as a function of $p\left(x_{(i)}\right)$ for biomedical data (left) and gene expression data (right). "N" = $\widehat{\text{NFDR}}\left(p\left(x_{(i)}\right)\right)$; "C" = $\text{CFDR}\left(x_{(i)}\right)$; "R" = $\text{RFDR}\left(x_{(i)}\right)$.

3 hypotheses $\left(\text{CFDR}\left(x_{(i)}\right) \leq 0.05\right)$, or 2 hypotheses $\left(\text{RFDR}\left(x_{(i)}\right) \leq 0.05\right)$. ▲

Example 3. Using microarray technology and $n = 6$ biological replicates, Alba et al. (2005) recorded microarray measurements of the expression levels of 13,440 genes in tomatoes at three days after the breaker stage of ripening. Following Bickel (2012), only the $d = 6103$ genes with data available for all 6 replicates are used for multiple applications of the paired t-test, with 6 (mutant, wild type) pairs for each gene.

The right panel of Figure 1 reveals that $\widehat{\text{NFDR}}\left(p\left(x_{(i)}\right)\right)$ is much less than $\text{CFDR}\left(x_{(i)}\right)$ and $\text{RFDR}\left(x_{(i)}\right)$ for the 545 genes with $\widehat{\text{NFDR}}\left(p\left(x_{(i)}\right)\right) \leq 0.2$, only 69 or 344 of which satisfy $\text{CFDR}\left(x_{(i)}\right) \leq 0.2$ or $\text{RFDR}\left(x_{(i)}\right) \leq 0.2$, respectively. This marked discrepancy between LFDR and NFDR estimators corroborates previous findings for this data set (Bickel, 2012, Fig. 3) and for other gene expression data (Hong et al., 2009). ▲

Acknowledgments

This research was partially supported by Agriculture and Agri-Food Canada and by the Faculty of Medicine of the University of Ottawa.

References

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., Tanksley, S. D., Giovannoni, J. J., 2005. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. Plant Cell 17, 2954–2965.

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

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

Bickel, D. R., 2012. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. International Journal of Approximate Reasoning 53, 880–891.

Bickel, D. R., 2013. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Statistical Applications in Genetics and Molecular Biology 12, 529–543.

Bickel, D. R., 2015. Corrigendum to: Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Statistical Applications in Genetics and Molecular Biology 14, 225.

Conrad, K., 2014. Probability distributions and maximum entropy. Technical Report, University of Connecticut, http://www.math.uconn.edu/ kconrad/blurbs/analysis/entropypost.pdf, accessed 19 Oct. 2015.

Cox, D., Hinkley, D., 1979. Theoretical Statistics. Taylor & Francis, Boca Raton, Florida.

Dudoit, S., van der Laan, M. J., 2008. Multiple Testing Procedures with Applications to Genomics. Springer, New York.

Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, Cambridge.

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70–86.

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

Genovese, C., Wasserman, L., 2003. Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, June 2-6, 2002. Oxford University Press, Oxford, Ch. Bayesian and frequentist multiple testing, pp. 145–161.

Genovese, C., Wasserman, L., 2006. Exceedance control of the false discovery proportion. Journal of the American Statistical Association 101, 1408–1417.

Glickman, M., Rao, S., Schultz, M., 2014. False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. Journal of Clinical Epidemiology 67 (8), 850–857.

Hong, W.-J., Tibshirani, R., Chu, G., 2009. Local false discovery rate facilitates comparison of different microarray experiments. Nucleic Acids Research 37 (22), 7483–7497.

Korn, E. L., Troendle, J. F., McShane, L. M., Simon, R., 2004. Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference 124 (2), 379 – 398.

Neuhaus, K. L., von Essen, R., Tebbe, U., Vogt, A., Roth, M., Riess, M., Niederer, W., Forycki, F., Wirtzfeld, A., Maeurer, W., 1992. Improved thrombolysis in acute myocardial infarction with front-loaded administration of alteplase: results of the rt-PA-APSAC patency study (TAPS). Journal of the American College of Cardiology 19, 885–91.

Pawitan, Y., Calza, S., Ploner, A., 2006. Estimation of false discovery proportion under general dependence. Bioinformatics 22 (24), 3025–3031.

Storey, J. D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society B 64, 479–498.

Van Den Oord, E., 2008. Review article: Controlling false discoveries in genetic studies. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics 147 (5), 637–644.

Van der Laan, M. J., Dudoit, S., Pollard, K. S., 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat. Appl. in Genet. and Mol. Biol. 3, 15.