LOCAL FALSE DISCOVERY RATES

by

Bradley Efron
Division of Biostatistics
Stanford University

Technical Report No. 2005-20B/234
March 2005

This research was supported in part by National Institute of Health grant 8R01 EB002784 and National Science Foundation grant DMS-0072360

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu

Local False Discovery Rates

Bradley Efron

Abstract

Modern scientific technology is providing a new class of large-scale simultaneous inference problems, with hundreds or thousands of hypothesis tests to consider at the same time. Microarrays epitomize this type of technology but similar problems arise in proteomics, time of flight spectroscopy, flow cytometry, FMRI imaging, and massive social science surveys. This paper uses local false discovery rate methods to carry out size and power calculations on large-scale data sets. An empirical Bayes approach allows the fdr analysis to proceed from a minimum of frequentist or Bayesian modeling assumptions. Microarray and simulated data sets are used to illustrate a convenient estimation methodology whose accuracy can be calculated in closed form. A crucial part of the methodology is an fdr assessment of "thinned counts", what the histogram of test statistics would look like for just the non-null cases.

1. Introduction

Large-scale simultaneous hypothesis testing problems, with hundreds or thousands of cases considered together, have become a fact of current-day statistical practice. Microarray methodology spearheaded the production of large-scale data sets, but other "high throughput" technologies are emerging, including time of flight spectroscopy, proteomic devices, flow cytometry, and functional Magnetic Resonance Imaging.

Benjamini and Hochberg's seminal (1995) paper introduced False Discovery Rates (Fdr), a particularly useful new approach to simultaneous testing. Fdr theory relies on p-values, that is on null hypothesis tail areas, and as such operates as an extension of traditional frequentist hypothesis testing to simultaneous inference, whether involving just a few cases or several thousand. Large-scale situations, however, permit another approach: empirical Bayes methods can bring Bayesian ideas to bear without the need for strong Bayesian or frequentist assumptions. Local false discovery rates (fdr), the subject of this paper, use empirical Bayes techniques to provide both size and power calculations for large-scale studies.

The data for one such study is summarized in Figure 1. Eight microarrays, four from cells of HIV infected subjects and four from uninfected subjects, have each measured expression levels for the same $N = 7680$ genes. Each gene yields a two-sample t-statistic t_i comparing the infected versus the uninfected subjects, which is then transformed to a z-value,

$$z_i = \Phi^{-1}(F_6(t_i)), \tag{1.1}$$

where F_6 is the cumulative distribution function (cdf) of a standard t variable with 6 degrees of freedom, and Φ is the standard normal cdf. Theoretically z_i should have a $N(0,1)$ distribution if gene i produces identically distributed normal expressions for infected and uninfected cells.

The histogram of z-values shown in Figure 1 looks promising: the normal-shaped central peak presumably charts the large majority of "null" genes, those behaving similarly for infected and uninfected cells, while the long tails reveal some interesting "non-null" genes, the kind the study, was intended to detect; fdr methodology, described in Section 5, has been used to provide *thinned counts*, an estimate of what a histogram of only the non-null z-values would look like.

Figure 2 shows the estimated local false discovery rate curve $\text{fdr}(z)$ based on empirical Bayes methodology discussed in Sections 3 and 4; $\text{fdr}(z)$, the conditional probability of a case being null given z, declines from one near $z = 0$ to zero at the extremes. There are 186 genes having $\text{fdr}(z) \leq 0.2$, a reasonable cutoff point discussed in Section 2, and we might report these 186 to the investigators as interesting candidates for further study. Other methods, such as Benjamini and Hochberg's Fdr procedure with cutoff $q = 0.1$, yield similar results.

Figure 2 also displays the thinned counts from Figure 1, estimating the histogram of non-null genes. Strikingly, a majority of the non-null cases lie well within the 0.2 fdr cutoff limits. However if we try to report more of the non-null cases then false discovery rates

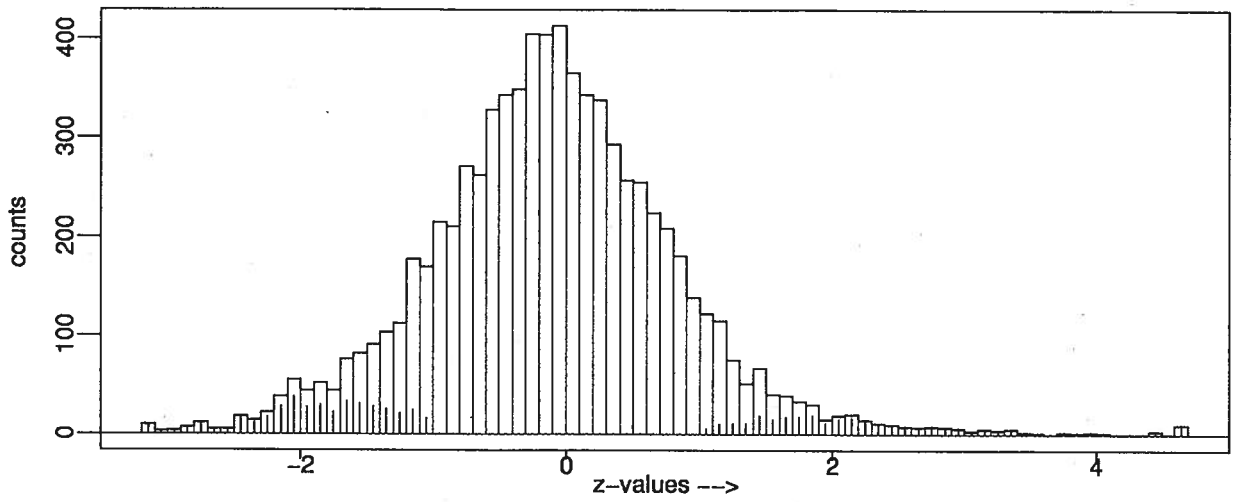

Figure 1: *Histogram of 7680 z-values from an HIV microarray experiment. Short vertical bars are estimated "thinned counts" of non-null genes, as explained in Section 5. (Extreme values have been truncated, giving small bars at each end.) Data from van't Wout et al. (2003), discussed in Gottardo et al. (2004).*

can grow unacceptably large, say to $\text{fdr}(z) = 0.5$, where the investigator would have a 50% chance of pursuing false leads.

In other words the HIV study is underpowered. Section 5 describes power diagnostics for large-scale testing situations, based on fdr calculations of the type shown Figure 2.

Section 6 discusses the non-null distribution of z-values such as (1.1). It suggests that the underlying densities for histograms like Figure 1's should be smooth normal mixtures, smoothness being an important assumption of our fdr methodology.

Ideally, a big data set like that of the HIV study should require very little parametric modeling, the data itself providing the framework for its own analysis. This ideal is approached by the fdr calculations for Figures 1 and 2, which depend on a simple model, presented in Section 2, requiring few assumptions. Section 7 examines this model in terms of a more structured formulation, clarifying its limitations in regard to bias and the choice of null hypothesis.

Focusing on z-values, rather than working within the full 7680×8 data matrix for the HIV study, greatly reduces the need for modeling assumptions. There will certainly be situations where working inside the matrix, as in Newton et al. (2004), Gottardo et al. (2004), and Kerr, Martin, and Churchill (2000), yields more information. Using such methods requires more careful attention to the details of the individual data set than our relatively crude z-value approach. A key assumption *not* made here is independence across the columns of the data matrix (e.g. independence across microarrays) which underlies the use of permutation or bootstrap methods for null hypothesis testing distributions. In fact there turns out to be curious dependences across the HIV matrix, as mentioned in Section 3.2, similar to the correlation effects in the microarray example of Efron (2004); column-wise independence seems to be a dangerous assumption for microarray studies.

A substantial microarray statistics literature has developed in the past few years, much

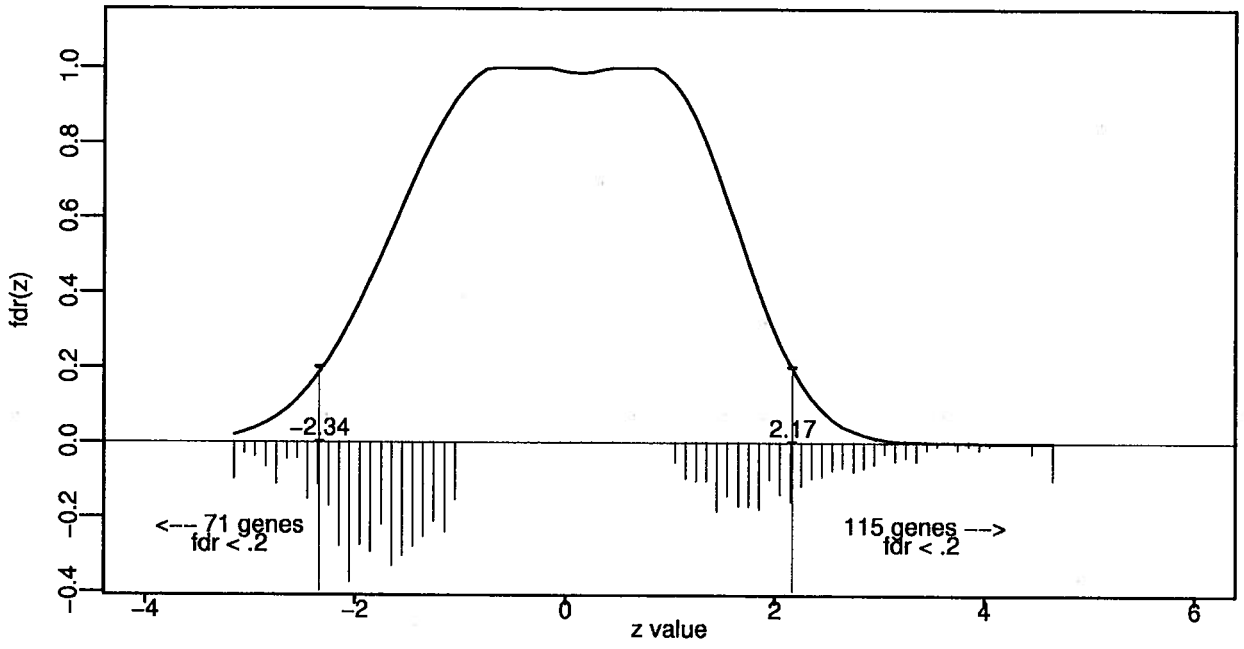

Figure 2: *Heavy curve is* $\text{fdr}(z)$*, local false discovery rate as estimated by locfdr algorithm described in Section 3;* $\text{fdr}(z) = 0.2$ *at* $z = -2.34$ *and 2.17. Vertical bars are thinned counts from Figure 1, now multiplied by 0.01 and plotted negatively.*

of it focused on the control of frequentist Type I errors, see for example Dudoit, van der Laan and Pollard (2004), and the review article by Dudoit, Shaffer, and Boldruck (2003). Bayes and empirical Bayes methods have also been advocated, as in Kendziorski et al. (2003), Johnstone and Silverman (2004), and Newton et al. (2004), while Benjamini and Hochberg's Fdr theory is increasingly influential, see Storey et al. (2004), and Genovese and Wasserman (2004). Local fdr methods, which this article argues can play a useful role, were introduced in Efron et al. (2001); several references are listed at the end of Section 3.1.

2. False Discovery Rates

Local false discovery rates, Efron et al. (2001), Efron and Tibshirani (2002), are a variant of Benjamini and Hochberg's (1995) "tail area" false discovery rates. This section relates the two ideas, reviews a few basic properties, and presents some general guidelines for interpreting fdr's. The development here is theoretical, with practical estimation procedures deferred to Section 3.

Suppose we have N null hypotheses to consider simultaneously, each with its own test statistic,

$$\begin{aligned} \text{Null hypothesis}: &\quad H_1, H_2, \ldots, H_i, \ldots, H_N \\ \text{Test statistic}: &\quad z_1,\ z_2, \ldots,\ z_i, \ldots,\ z_N \end{aligned} \tag{2.1}$$

N must be large for local fdr calculations, at least in the hundreds, but the z_i need not be independent. A simple Bayesian model, Lee et al. (2000), Newton et al. (2001), Efron et al. (2001), underlies the theory: we assume that the N cases are divided into two classes, null or non-null, occurring with prior probabilities p_0 or $p_1 = 1 - p_0$, and with the density of test

statistic z depending upon its class,

$$\begin{aligned} p_0 = Pr\{\text{null}\} \quad & f_0(z) \text{ density if null} \\ p_1 = Pr\{\text{non-null}\} \quad & f_1(z) \text{ density if non-null.} \end{aligned} \tag{2.2}$$

In context (1.1) it is natural to take $f_0(z)$ to be the standard $N(0,1)$ density – but see Section 3.2 – and $f_1(z)$ some longer-tailed density, perhaps representing a mixture of alternative possibilities; the empirical estimation theory of Section 3 does not require specification of $f_1(z)$. Practical applications of large-scale testing usually assume a large p_0 value, say

$$p_0 \geq 0.9, \tag{2.3}$$

the goal being to identify a relatively small set of interesting non-null cases.

Define the *null subdensity*

$$f_0^+(z) = p_0 f_0(z) \tag{2.4}$$

and the *mixture density*

$$f(z) = p_0 f_0(z) + p_1 f_1(z). \tag{2.5}$$

The Bayes posterior probability that a case is null given z, by definition the local false discovery rate, is

$$\begin{aligned} \text{fdr}(z) &\equiv Pr\{\text{null}|z\} = p_0 f_0(z)/f(z) \\ &= f_0^+(z)/f(z). \end{aligned} \tag{2.6}$$

The Benjamini-Hochberg false discovery rate theory relies on tail areas rather than densities. Letting $F_0(z)$ and $F_1(z)$ be the cdf's corresponding to $f_0(z)$ and $f_1(z)$ in (2.2), define $F_0^+(z) = p_0 F_0(z)$ and $F(z) = p_0 F_0(z) + p_1 F_1(z)$. Then the posterior probability of a case being null given that its z-value "Z" is less than some value z is

$$\text{Fdr}(z) \equiv Pr\{\text{null}|Z \leq z\} = F_0^+(z)/F(z). \tag{2.7}$$

(It is notationally convenient to consider events $Z \leq z$ but we could just as well consider tail areas to the right, two-tailed events, etc.) Figure 3 illustrates the geometrical relationship between Fdr and fdr.

Benjamini and Hochberg's FDR control rule depends on an estimated version of (2.6) where F is replaced by the empirical cdf. Storey (2002) and Efron and Tibshirani (2002) discuss the connection of the frequentist FDR procedure with Bayesian form (2.7). $\text{Fdr}(z)$ corresponds to Storey's "q-value", the value of the tail area false discovery rate attained at a given observed value $Z = z$.

Fdr and fdr are analytically related by

$$\begin{aligned} \text{Fdr}(z) &= \int_{-\infty}^{z} \text{fdr}(Z) f(Z) dZ \Big/ \int_{-\infty}^{z} f(Z) dZ \\ &= E_f\{\text{fdr}(Z)|Z \leq z\}, \end{aligned} \tag{2.8}$$

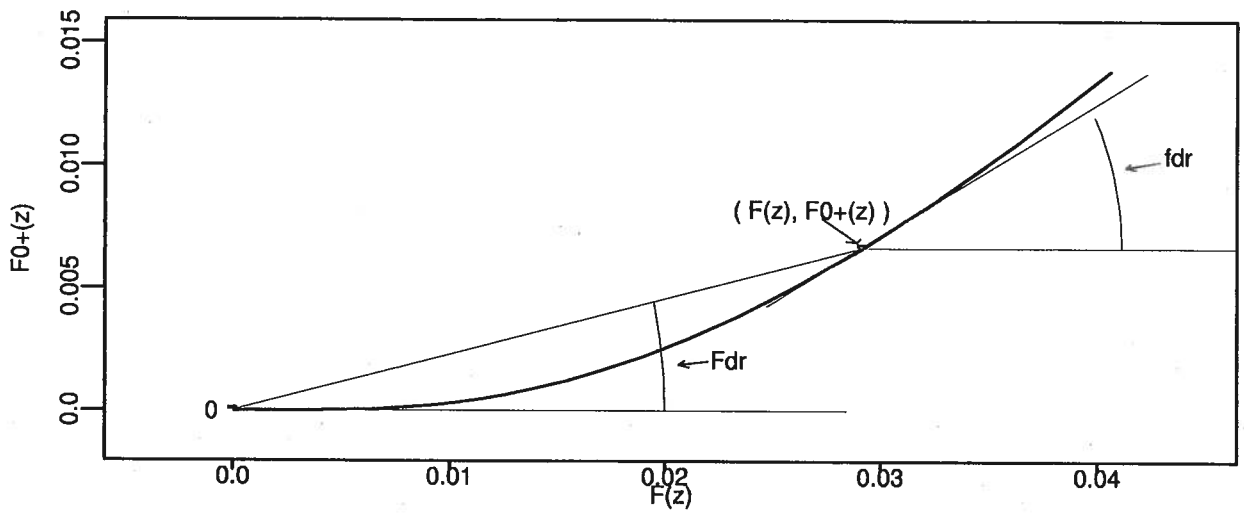

Figure 3: *Geometrical relationship of Fdr to fdr; heavy curve plots* $F_0^+(z)$ *versus* $F(z)$; $\mathrm{fdr}(z)$ *is slope of tangent,* $\mathrm{Fdr}(z)$ *slope of secant.*

"E_f" indicating expectations with respect to $f(z)$, Efron and Tibshirani (2002). That is, $\mathrm{Fdr}(z)$ is the average of $\mathrm{fdr}(Z)$ for $Z \le z$; $\mathrm{Fdr}(z)$ will be less than $\mathrm{fdr}(z)$ in the usual situation where $\mathrm{fdr}(z)$ decreases as $|z|$ gets large. For example $\mathrm{fdr}(-2.34) = 0.20$ in Figure 2 while $\mathrm{Fdr}(-2.34) = 0.12$. If the cdf's $F_0(z)$ and $F_1(z)$ are Lehmann alternatives

$$F_1(z) = F_0(z)^\alpha, \quad [\alpha < 1], \tag{2.9}$$

it is straightforward to show that

$$\log\left\{\frac{\mathrm{fdr}(z)}{1-\mathrm{fdr}(z)}\right\} = \log\left\{\frac{\mathrm{Fdr}(z)}{1-\mathrm{Fdr}(z)}\right\} + \log\left(\frac{1}{\alpha}\right), \tag{2.10}$$

giving

$$\mathrm{fdr}(z) \doteq \mathrm{Fdr}(z)/\alpha \tag{2.11}$$

for small values of Fdr. The HIV data of Figure 1 has α roughly 1/2 in the left tail and 1/3 in the right.

The local nature of $\mathrm{fdr}(z)$ is an advantage in interpreting results for individual cases. For example, a gene with $z = 2.0$ in the HIV study has an estimated fdr of 0.30 while the corresponding (right-sided) tail-area Fdr, the q-value, is 0.12. Quoting just this last number gives an overoptimistic impression of the gene's significance. In practice the methods can be combined, using the Benjamini-Hochberg algorithm to identify non-null cases, say with $q = 0.10$, but also providing individual fdr values for those cases.

The literature has not reached consensus on a standard choice of q for Benjamini-Hochberg testing, the equivalent of .05 for single tests, but Bayesian calculations offer some insight. The cutoff threshold $\mathrm{fdr} \le 0.20$ used in Figure 2 yields posterior odds ratio

$$\begin{aligned} &Pr\{\text{non-null}|z\}/Pr\{\text{null}|z\} = (1-\mathrm{fdr}(z))/\mathrm{fdr}(z) \\ &= p_1 f_1(z)/p_0 f_0(z) \ge 0.8/0.2 = 4. \end{aligned} \tag{2.12}$$

If we assume prior odds ratio $p_1/p_0 \leq 0.1/0.9$ as in (2.3), then (2.12) corresponds to Bayes factor

$$f_1(z)/f_0(z) \geq 36 \tag{2.13}$$

in favor of non-null.

This threshold requires a much stronger level of evidence against the null hypothesis then in standard one-at-a-time testing. For instance suppose we observe $x \sim N(\mu, 1)$ and wish to test $H_0 : \mu = 0$ vs $\mu = 2.80$, a familiar scenario for power calculations since rejecting H_0 for $x \geq 1.96$ yields two-sided size 0.05 and power 0.80. Here the critical Bayes factor is only $f_{2.80}(1.96)/f_0(1.96) = 4.80$. (A value closer to 3 is suggested by the more careful considerations in Efron and Gous (2001).) We might justify (2.13) as being conservative in guarding against multiple testing fallacies. More pragmatically, increasing the fdr threshold much above 0.20 can deliver unacceptably high proportions of false discoveries to the investigators. The 0.20 threshold, used in the remainder of the paper, corresponds to q-values between 0.05 and 0.15 for reasonable choices of α in (2.11); such q-value thresholds can be interpreted as reflecting a conservative Bayes factor for Fdr interpretation.

Any choice of threshold is liable to leave investigators complaining that the statisticians' list of non-null cases omits some of their *a priori* favorites. Conveying the full list of values $\text{fdr}(z_i)$, not just those for cases judged non-null, allows investigators to employ their own prior opinions on interpreting significance. This is particularly important for low-powered situations like the HIV study, where luck plays a big role in any one case's results, but it is the counsel of perfection, and most investigators will require some sort of reduced list.

False discovery rates, both fdr and Fdr, depend on only the marginal distribution of the z values, $f(z)$ or $F(z)$. This has both good and bad consequences: On the good side, independence is *not* required of the z_i's in (2.1), since all that is needed is a reasonable estimate of their marginal distribution. Less happily, results like (2.6) or (2.7) are really "one-at-a-time" Bayes inferences, that may be quite different than the (usually unknowable) posterior probability of H_i given the entire N-vector $\mathbf{z}$.

3. Estimating fdr

The heavy curve in Figure 2 is an estimate of the local false discovery rate $\text{fdr}(z)$ for the HIV study. This section concerns the estimate's empirical Bayes methodology, including the question of choosing an appropriate null hypothesis. Accuracy of the estimation procedure is taken up in Section 4. (This methodology is available through algorithm *locfdr*, Comprehensive R Archive Network, http://cran.r-project.org.) Estimating the numerator and denominator of $\text{fdr}(z) = f_0^+(z)/f(z)$ will be discussed separately.

3.1 *Estimating the Mixture Density* $f(z)$

Nonparametric density estimation has a reputation for difficulty, well-deserved in general situations. However there are good theoretical reasons for believing that z-value distributions are quite smooth, see Section 6. Our tactic here is to estimate the mixture density $f(z)$, the denominator of $\text{fdr}(z)$ in (2.6), with smooth but flexible parametric models. Section 4 discusses the accuracy of this approach.

Lindsey's method, as discussed in Section 2 of Efron and Tibshirani (1996), permits efficient and flexible parametric density estimation using standard Poisson regression software.

Suppose the N z-values have been binned, giving bin counts $y_1, y_2, \ldots y_K$ summing to N. The histogram in Figure 2, used $K = 79$ bins, each of width $\Delta = 0.1$. Lindsey's method takes the y_k to be independent Poisson counts,

$$y_k \overset{\text{ind}}{\sim} Po(\nu_k) \quad k = 1, 2, \ldots, K, \tag{3.1}$$

with ν_k proportioned to density $f(z)$ at midpoint "$z_{(k)}$" of the kth bin, approximately

$$\nu_k = N\Delta f(z_{(k)}). \tag{3.2}$$

Modeling $\log(\nu_k)$ as a pth degree polynomial function of $z_{(k)}$ makes (3.1), (3.2) a standard Poisson general linear model (GLM). The choice $p = 7$, for example, effectively amounts to estimating $f(z)$ by maximum likelihood within the seven-parameter exponential family

$$f(z) = \exp\left\{\sum_{j=0}^{7} \beta_j z^j\right\} \tag{3.3}$$

(with $(\beta_1, \beta_2, \ldots, \beta_7)$ determining β_0 from the requirement that $f(z)$ integrate to one.) The denominator of $\widehat{\text{fdr}}(z)$ in Figure 2 actually took $\log\{f(z)\}$ to be a natural spline function with seven degrees of freedom, but (3.3) gives nearly the same answers; standard Poisson deviance analysis showed a reasonably good fit, while doubling or halving the bin width Δ had little effect.

Dependence among the z_i's causes overdispersion and dependence for the y_k's in (3.1), but has little effect on (3.2). Lindsey's method remains nearly unbiased, but, as discussed in Section 4, the usual GLM accuracy estimates are liable to be overoptimistic.

A variety of other local fdr estimation methods have been suggested: using more specific parametric models such as normal mixtures, see Pan et al. (2003), Pounds and Morris (2003), Allison et al. (2002), or Heller and Qin (2003); isotonic regression, Broberg (2005); local smoothing, Aubert et al. (2004); and hierarchical Bayes analyses, Liao et al. (2004), Do et al. (2004). All seem to perform reasonably well. The Poisson GLM methodology of this paper has the advantage of easy implementation with familiar software, and permits a closed-form error analysis as shown in Section 4. Perhaps most usefully, it transfers density estimation to the more familiar realm of regression theory.

3.2 Estimating $f_0^+(z)$

The fdr numerator $f_0^+(z) = p_0 f_0(z)$, (2.4), is more challenging to estimate. We consider two situations: where the *theoretical null* $f_0(z)$ that would ordinarily be used for the individual hypothesis testing problems, e.g. $f_0(z) \sim N(0, 1)$ in (1.1), is deemed satisfactory for the simultaneous problem (2.1); and where it is unsatisfactory, and instead we must fit an *empirical null*, as in Efron (2004). The HIV study falls into the "unsatisfactory" category, and we begin by using it to illustrate empirical estimation of $f_0^+(z)$.

The heavy curve in Figure 4 is $\log \widehat{f}(z)$, the log of the estimated mixture density fit to the HIV counts by Poisson regression. A quadratic curve, dashed, has been fit to $\log \widehat{f}(z)$ around $z = 0$, and this is $\log \widehat{f}_0^+(z)$, the empirical estimate of $\log f_0^+$. The three coefficients of

the fitted quadratic determine $\widehat{f}_0^+(z)$ as a scaled normal density, in this case with $\widehat{p}_0 = 0.917$ and $\widehat{f}_0 \sim N(-0.10, 0.74^2)$,

$$\widehat{f}_0(z) = .917 \cdot \varphi_{-.10,.74}(z) \quad \left[\varphi_{\delta,\sigma}(z) = \exp\left\{-\frac{1}{2}\left(\frac{z-\delta}{\sigma}\right)^2\right\}\Big/\sqrt{2\pi\sigma^2}\right]. \tag{3.4}$$

The default option in *locfdr* fits a quadratic to $\log \widehat{f}(z)$ by ordinary least squares applied over the central one-third range of the z-values.

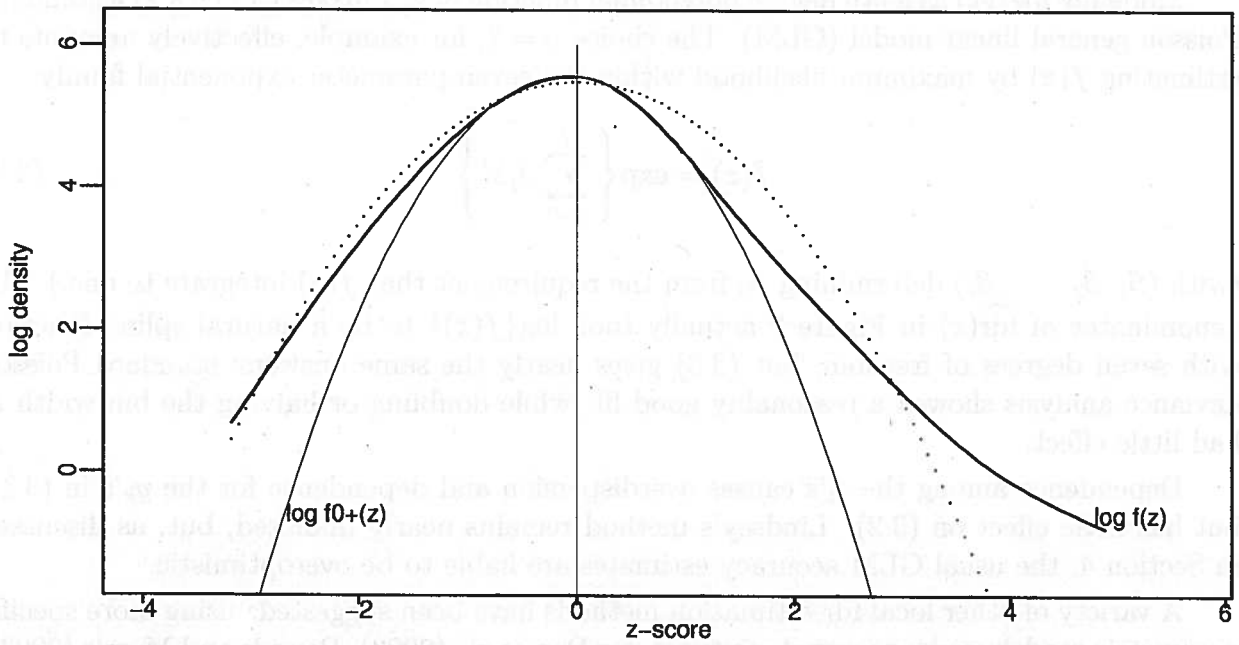

Figure 4: *Empirical estimation of the fdr numerator $f_0^+(z) = p_0 f_0(z)$, HIV study. Heavy curve is log of Poisson regression estimate $\widehat{f}(z)$ for mixture density; dashed curve is $\log \widehat{f}_0^+(z)$, best-fitting quadratic to $\log \widehat{f}(z)$ near $z = 0$; estimates $\widehat{p}_0 = 0.917$, $\widehat{f}_0 \sim N(-0.10, 0.74^2)$. Dotted curve is $\log \widehat{f}_0^+$ for theoretical null.*

The logic here is quite simple: we make the "zero assumption" that the central peak of Figure 1's histogram consists mainly of null cases, and choose p_0, δ and σ in (3.4) to quadratically approximate the histogram counts near $z = 0$. This same argument can be applied with the theoretical null, giving the dotted curve in Figure 4. Now $f_0(z)$ is assumed to be $\varphi_{0,1}(z)$, the standard normal, so only p_0 in $f_0^+ = p_0 f_0$ remains to be estimated from the central histogram counts.

The two-class model (2.2) is unidentifiable without restrictions on the form of f_0 and f_1. Some version of the zero assumption is necessary in the absence of strong parametric assumptions, see for example Section 3 of Storey (2002). (Most of the FDR literature works with p-values rather than z-values, $p_i = F_6(t_i)$ in (1.1), in which case the "zero region" occurs near $p = 1$.) The zero assumption is more believable when p_0, the proportion of null cases, is

near 1. Efron (2004), Section 5, shows that if p_0 exceeds 0.90, the fitting method of Figure 4 will have negligible bias: although the 10% or less of non-null cases might in fact contribute some counts near $z = 0$, these cannot substantially affect the estimates of δ and σ in (3.4). The estimate of p_0 *is* affected, being upwardly biased as seen in Section 4 and discussed in Section 7.

The theoretical null hypothesis $f_0 \sim N(0,1)$ is untenable for the HIV data. If it were valid then $f(z)$ should be at least as wide as f_0 near $z = 0$, assuming that non-null z's are more dispersed than nulls. Instead f is substantially narrower, forcing $f_0^+ = p_0 f_0$ to take the impossible value $\widehat{p}_0 = 1.15$ in order to match the histogram heights near $z = 0$.

The examples in Efron (2004) go the other way: in both of them the empirical null is substantially *wider* than $N(0,1)$. Various causes of overdispersion are suggested, including hidden correlations and unobserved covariates. The underdispersion here is harder to explain, but can be traced to a correlation of expression levels across microarrays: levels on the odd-numbered arrays were positively correlated, as were levels among the even-numbered arrays, the effect cutting across the Treatment-Control classification, a pattern that swelled the denominators of the t-statistics (1.1).

Misspecification of the null hypothesis, which becomes visible in large-scale testing situations, *undermines all forms of simultaneous inference*, fdr, Fdr, Bonferroni, Family-Wise Error Rate, or the sophisticated resampling based algorithms of Westfall and Young (1993). Using an empirical null avoids the problem, but at a substantial cost in estimation efficiency as shown in Section 4. Other methods are sometimes available for empirical null estimation, involving "housekeeper genes" (cases known *a priori* to be null) and designed replications, as in Lee et al. (2000).

More ambitiously, one may try to model the full error structure of the original data set, a 7680×8 matrix in the HIV study, using frequentist or Bayesian modeling as in Kerr and Churchill (2001), or Newton et al. (2004). When feasible this is the ideal approach but it can be an heroic undertaking in the complicated venue of microarray analysis. The approach here, relying only on the observed distribution of the z-values, trades some loss of efficiency for fewer assumptions and simple application.

Permutation and bootstrap null density estimates play a major role in the microarray literature, as in Tusher et al. (2001) and Pollard and van der Laan (2003). These should be considered as improved versions of the theoretical null rather than empirical nulls. The permutation null for the HIV data, permuting the eight microarrays, is about $N(0, .99^2)$.

Figure 4's quadratic construction assumes that f_0 is normal, but uses the data to estimate its mean and variance instead of accepting the theoretical choice $N(0,1)$. Under some circumstances we might wish to go further, perhaps adding a cubic term to $\log f_0$; *locfdr* includes such an option, described in Section 7.

The basic false discovery rate idea is appealingly simple: 19 of the HIV z-values fell into bin $[2.0, 2.1]$ in Figure 1, the smoothed estimate from $\widehat{f}$ being 19.95; this compares with expected number 4.70 under $\widehat{f}_0^+$, yielding estimated local false discovery rate $4.70/19.95 = 0.24$. If we report this bin as containing interesting cases, then about one-fourth of them will turn out to be false discoveries. The question of the accuracy of this estimate is taken up next.

4. Estimation Accuracy

The algorithm described in 3.2 produces an estimate $\widehat{\text{fdr}}(z)$ for the local fdr from z-values $z_1, z_2, \ldots, z_N$. How accurate is the estimate? This section derives a delta-method formula for the standard error of $\log(\widehat{\text{fdr}}(z))$, under the assumption that the z's are independent. The formula is useful for understanding the relative efficiency of local fdr compared to tail area Fdr, for assessing components of variation caused by the three $\widehat{\text{fdr}}$ components, $\widehat{p}_0$, $\widehat{f}_0$, and $\widehat{f}$, and as a lower bound and rough guide to estimation accuracy even if independence is doubtful.

Before deriving the formula we report on a small simulation study where

$$z_i \overset{\text{ind}}{\sim} N(\mu_i, 1) \quad \text{with} \quad \begin{cases} \mu_i = 0 & \text{probability } 0.90 \\ \mu_i \sim N(3,1) & \text{probability } 0.10, \end{cases} \tag{4.1}$$

for $i = 1, 2, \ldots, N$. Three choices of N were used, $N = 500, 1500, 4500$, with 250 simulations each; for example the $N = 1500$ choice had 1350 μ_i's equaling 0 and 150 exactly following $N(3,1)$. The table reports standard deviations for $\log\{\widehat{\text{fdr}}(z)\}$, and for the tail area quantity $\log\{\widehat{\text{Fdr}}(z)\}$, obtained by integrating the parametric estimates $\widehat{f}$ and $\widehat{f}_0^+$ to get $\widehat{F}$ and $\widehat{F}_0^+$ for insertion into (2.7).

The most striking fact in Table 1 is the high cost of using an empirical null, a factor of 3 increase in standard deviation in the critical interval $[2.5, 3.5]$ for z where $\text{fdr}(z)$ ranges from 0.45 down to 0.05. The local-tail area comparison is much less dramatic: $\widehat{\text{fdr}}$ is about 50% more variable than $\widehat{\text{Fdr}}$ when using the theoretical null, but correspondingly less variable with the empirical null.

In practical terms a log standard deviation less than 0.25 will usually be tolerable, corresponding to estimates between 0.15 and 0.25 for a true 0.20 false discovery rate. All the entries based on the theoretical null are less than 0.25, and this would hold for smaller sample sizes as well since the standard deviations are approximately proportional to $1/\sqrt{N}$.

The empirical null standard deviations are too big for comfort at $N = 500$ and only borderline acceptable at $N = 1500$. Of course we would prefer to use the theoretical null but, unfortunately, it does not fit the data in situations like the HIV study or the examples of Efron (2004), where inferences based on the theoretical null are dangerously misleading. One tactic is to reduce empirical variability, at the risk of bias, by using less flexible parametric models. Decreasing the degrees of freedom for the natural spline estimate of $f(z)$ from 7 to 5 reduced the standard deviations for $\log(\widehat{\text{fdr}})$ by about one-third.

Table 1's standard errors for the tail area false discovery rates $\widehat{\text{Fdr}}(z)$, i.e. for q-values, are based on the same parametric models as $\widehat{\text{fdr}}(z)$. Replacing the parametric cdf estimate $\widehat{F}$ with the nonparametric empirical cdf $\bar{F}$ increased the standard errors in Table 1 by several percent, worsening an already bad situation for the empirical null. Benjamini and Hochberg's (1995) Fdr-controlling algorithm depends on $\bar{F}$ (as well as independence); there the high variability of $\widehat{\text{Fdr}}$ does not affect the claimed control rates, but does reduce the power of the procedure to identify non-null cases. Power is considered here in Section 5.

Poisson GLM calculations provide convenient approximation formulas for $\text{stdev}(\log \widehat{\text{fdr}})$ and $\text{stdev}(\log \widehat{\text{Fdr}})$. Let X be the $K \times m$ structure matrix used for estimating $\log(f)$ in

Section 3.1; X has $m = 8$, kth row $(1, z_{(k)}, z_{(k)}^2, \ldots, z_{(k)}^7)$ in model (3.3). Also let X_0 be the $K \times m_0$ matrix used to describe $\log(f_0^+)$ in Section 3.2: X_0 has kth row $(1, z_{(k)}, z_{(k)}^2)$ for the empirical null, $m_0 = 3$, while X_0 is the $K \times 1$ matrix $(1, 1, \ldots, 1)'$ for the theoretical null.

Locfdr fits $\log \widehat{f}_0^+(z)$ to $\log \widehat{f}(z)$ over a central subset of K_0 bins, with index set say "$\mathbf{i}_0$", defining submatrices with rows in $\mathbf{i}_0$,

$$\widetilde{X} = X[\mathbf{i}_0,] \quad \text{and} \quad \widetilde{X}_0 = X_0[\mathbf{i}_0,] \tag{4.2}$$

of dimensions $K_0 \times m$ and $K_0 \times m_0$. Also define inner product matrices

$$\widehat{G} = X' \text{ diag } (\widehat{\nu}) X \quad \text{and} \quad \widetilde{G}_0 = \widetilde{X}_0' \widetilde{X}_0, \tag{4.3}$$

where diag $(\widehat{\nu})$ is the $K \times K$ diagonal matrix having diagonal elements $\widehat{\nu}_k = N \Delta \widehat{f}(z_{(k)})$ as in (3.2).

Finally, let $\widehat{\ell}$ indicate the K-vector with elements $\widehat{\ell}_k = \log \widehat{f}(z_{(k)})$, likewise $\widehat{\ell}_0^+$ for vector $(\log \widehat{f}_0^+(z_{(k)}))$ and $\widehat{\ell f dr}_k$ for $\log \widehat{\text{fdr}}(z_{(k)})$.

Lemma 1 The $K \times K$ derivative matrix of $\log \widehat{\text{fdr}}$ with respect to the bin counts is

$$\left(\frac{d \, \ell f dr_k}{dy_\ell} \right) = A \widehat{G}^{-1} X', \tag{4.4}$$

where

$$A = X_0 \widetilde{G}_0^{-1} \widetilde{X}_0' \widetilde{X} - X. \tag{4.5}$$

Proof A small change dy in the count vector (considered as continuous) produces change $d\widehat{\ell}$ in $\widehat{\ell}$,

$$d\widehat{\ell} = X \widehat{G}^{-1} X' dy. \tag{4.6}$$

Similarly if $\widehat{\ell}_0^+ = X_0 \widehat{\gamma}$, $\widehat{\gamma}$ a m_0-vector, is fit by least squares to $\widetilde{\ell} = \widehat{\ell}[\mathbf{i}_0]$, we have

$$d\widehat{\gamma} = \widetilde{G}_0^{-1} \widetilde{X}_0' d\widetilde{\ell} \quad \text{and} \quad d\widehat{\ell}_0^+ = X_0 \widetilde{G}_0^{-1} \widetilde{X}_0' d\widetilde{\ell}. \tag{4.7}$$

Both (4.6) and (4.7) are standard regression results. Then (4.6) gives $d\widetilde{\ell} = d\widehat{\ell}[\mathbf{i}_0] = \widetilde{X} \widehat{G}^{-1} X' dy$, yielding

$$d\widehat{\ell}_0^+ = X_0 \widetilde{G}_0^{-1} \widetilde{X}_0' \widetilde{X} \widehat{G}^{-1} X' dy \tag{4.8}$$

from (4.7). Finally,

$$d(\widehat{\ell f dr}) = d(\widehat{\ell}_0^+ - \widehat{\ell}) = (X_0 \widehat{G}_0^{-1} \widetilde{X}_0' \widetilde{X} - X) \widehat{G}^{-1} X' dy, \tag{4.9}$$

verifying (4.4). ■

The delta-method estimate of covariance for the K-vector $\widehat{\ell \text{fdr}}$ is derived from the lemma as

$$\begin{aligned} \widehat{\text{cov}}(\widehat{\ell f dr}) &= (A \widehat{G}^{-1} X') \widehat{\text{cov}}(y) (A \widehat{G}^{-1} X')' \\ &= (A \widehat{G}^{-1} X') \text{diag}(\widehat{\nu}) (A \widehat{G}^{-1} X')', \end{aligned} \tag{4.10}$$

under Poisson assumptions (3.1), (3.2). Since $\widehat{G} = X'$ diag $(\widehat{\nu})X$ this reduces to a relatively simple formula:

Theorem The delta-method estimate of covariance for the vector of $\log \widehat{\text{fdr}}(z_{(k)})$ values is

$$\widehat{\text{cov}}(\widehat{\ell fdr}) = A\widehat{G}^{-1}A \tag{4.11}$$

with A as in (4.5).

The entries "form" in Table 1 are square roots of diagonal elements of $\widehat{\text{cov}}$ in (4.11), averaged over the 250 simulations. They produced reasonable estimates of the actual standard deviations of log $(\widehat{\text{fdr}})$, especially for the empirical null.

A formula similar to (4.11) exists for the tail area false discovery rates $\widehat{\ell Fdr}_k = \log \widehat{\text{Fdr}}(z_{(k)})$,

$$\widehat{\text{cov}}(\widehat{\ell Fdr}) = B\widehat{G}^{-1}B', \tag{4.12}$$

$$B = \widehat{S}_0 X_0 \widetilde{G}_0^{-1} \widetilde{X}_0' \widetilde{X} - \widehat{S}X, \tag{4.13}$$

where, for the case of left-tail $\widehat{\text{Fdr}}$'s, $\widehat{S}_0$ and $\widehat{S}$ are lower triangular matrices,

$$\widehat{S}_{k\ell} = \frac{\widehat{f}_\ell}{\widehat{F}_k} \quad \text{and} \quad \widehat{S}_{0k\ell} = \frac{\widehat{f}_{0\ell}}{\widehat{F}_{0k}} \quad \text{for} \quad \ell \le k. \tag{4.14}$$

Comparisons of (4.11) with (4.12) in various situations confirm the general story of Table 1: $\widehat{\text{fdr}}$ is somewhat more variable than $\widehat{\text{Fdr}}$ when using theoretical nulls, the opposite being true for empirical nulls; however both methods are much more variable in the empirical case, this effect dwarfing their comparative differences. (Empirical nulls fare better in the power calculations of Section 5.)

Table 2 displays means and standard deviations in simulation (4.1) for the three estimated parameters of the empirical null, p_0, δ, and σ, (3.4). Notice that $\widehat{p}_0$ is biased upward from the simulation value $p_0 = 0.90$. This makes little difference to $\widehat{\text{fdr}}(z) = \widehat{p}_0 \widehat{f}_0(z)/\widehat{f}(z)$, only increasing it by factor $.0924/0.90 = 1.03$. (The power calculations of Section 5 are more sensitive to bias.) Upward bias arises from the zero assumption: the $\mu \sim N(3,1)$ component of (4.1) gives $z \sim N(3,2)$, resulting in a small proportion of non-null z-values near 0. However the "bias" here reflects, at least partly, an ambiguity in what p_0 actually means, as discussed in Section 7.

The variability in the estimated mean and standard deviation of the empirical null, $\widehat{\delta}$ and $\widehat{\sigma}$, has an order of magnitude bigger effect than $\widehat{p}_0$ on $\widehat{\text{fdr}}$ and $\widehat{\text{Fdr}}$. The theoretical null "knows" that $(\delta, \sigma) = (0,1)$, eliminating this variability and accounting for its much smaller standard deviations.

For fixed z, $\log \widehat{\text{fdr}}(z)$ is a sum of three terms,

$$\log \widehat{\text{fdr}}(z) = \log \widehat{p}_0 + \log \widehat{f}_0(z) - \log \widehat{f}(z), \tag{4.15}$$

allowing an exact apportionment of variability of $\log \widehat{\text{fdr}}(z)$ to the three components. For the empirical null with $N = 1500$ and $z = 2.9$ (the point where $\text{fdr}(z) = 0.20$ in model (4.1)) the

z	ave($\widehat{\text{fdr}}$)	THEORETICAL NULL local	(form)	tail	EMPIRICAL NULL local	(form)	tail
$N = 500$							
1.5	.95	**.08**	(.09)	**.09**	**.06**	(.07)	**.17**
2.0	.77	**.15**	(.15)	**.08**	**.17**	(.17)	**.27**
2.5	.45	**.17**	(.18)	**.08**	**.28**	(.28)	**.40**
3.0	.17	**.15**	(.18)	**.10**	**.45**	(.45)	**.56**
3.5	.05	**.18**	(.24)	**.12**	**.68**	(.67)	**.72**
4.0	.01	**.20**	(.27)	**.16**	**.89**	(.90)	**.90**
$N = 1500$							
1.5	.96	**.05**	(.05)	**.05**	**.04**	(.04)	**.10**
2.0	.76	**.08**	(.09)	**.05**	**.09**	(.10)	**.15**
2.5	.44	**.09**	(.10)	**.05**	**.16**	(.16)	**.23**
3.0	.16	**.08**	(.10)	**.06**	**.25**	(.25)	**.32**
3.5	.04	**.10**	(.13)	**.07**	**.38**	(.38)	**.42**
4.0	.01	**.11**	(.15)	**.10**	**.50**	(.51)	**.52**
$N = 4500$							
1.5	.96	**.03**	(.03)	**.03**	**.02**	(.02)	**.05**
2.0	.77	**.05**	(.05)	**.03**	**.05**	(.06)	**.08**
2.5	.43	**.06**	(.06)	**.03**	**.09**	(.09)	**.12**
3.0	.16	**.05**	(.06)	**.03**	**.14**	(.14)	**.18**
3.5	.04	**.06**	(.08)	**.04**	**.21**	(.22)	**.23**
4.0	.01	**.06**	(.09)	**.05**	**.28**	(.29)	**.29**

Table 1: Accuracy comparison for local and tail area false discovery rates, simulation study (4.1); *boldface* stdev(log $\widehat{\text{fdr}}$), "local", and stdev (log $\widehat{\text{Fdr}}$), "tail"; "form" from formula (4.11), delta-method approximation for stdev(log $\widehat{\text{fdr}}$). Second column shows average $\widehat{\text{fdr}}(z)$ over the 250 simulations. Simulations used natural spline bases, seven degrees of freedom, for estimating $f(z)$.

	p_0	δ	σ	$p_0^{(\text{theo})}$
$N = 500$:	.924 (.020)	.021 (.078)	1.018 (.056)	.917 (.023)
$N = 1500$:	.924 (.011)	.023 (.046)	1.020 (.031)	.915 (.015)
[form]:	[.013]	[.049]	[.032]	[.015]
$N = 4500$:	.922 (.006)	.024 (.026)	1.017 (.018)	.915 (.007)

Table 2: Means and standard deviations (parentheses) for estimated empirical null parameters p_0, δ, σ, (3.4); simulation study (4.1). Last column for theoretical null p_0 estimates. Bracketed numbers from formulas (4.17)-(4.18), $N = 1500$.

term $\log \widehat{f}_0(z)$ is completely dominant: even knowing the true value of p_0 and $f(z)$ would reduce the standard deviation of $\log \widehat{\text{fdr}}(z)$ by less than 1%. Employing a theoretical null assumes away variability in $\log \widehat{f}_0(z)$. Now the $\log \widehat{f}(z)$ term dominates variance: knowing p_0 exactly reduces $sd(\log \widehat{\text{fdr}}(z))$ by only 9%.

Standard error formulas are available for the trio of empirical null parameter estimates $\widehat{\theta} \equiv (\log \widehat{p}_0, \widehat{\delta}, \widehat{\sigma})$ obtained as in Figure 4. Defining

$$D = \begin{pmatrix} 1 & \widehat{\delta} & \widehat{\sigma}^2 + \widehat{\delta}^2 \\ 0 & \widehat{\sigma}^2 & 2\widehat{\delta}\,\widehat{\sigma}^2 \\ 0 & 0 & \widehat{\sigma}^3 \end{pmatrix} \widetilde{G}_0^{-1} \widetilde{X}_0' \widetilde{X} \tag{4.16}$$

in the notation of (4.2)-(4.9), the delta method covariance matrix is

$$\widehat{\text{cov}}(\widehat{\theta}) = D\widehat{G}^{-1}D' - \begin{pmatrix} \frac{1}{N} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{4.17}$$

this following after some calculation from $d\widehat{\gamma}$ in (4.7). Applied to the simulations for $N = 1500$, (4.16) gave average standard errors "form" in Table 2, close to the empirical values; variation was moderate across trials, coefficients of variation 13%, 6%, and 8% respectively.

Simpler calculations provide a delta-method formula for the variance of $\log(\widehat{p}_0)$ when using the theoretical null, also shown in Table 2,

$$\widehat{\text{var}}\{\log \widehat{p}_0\} = \bar{x}_0^{-1} \widehat{G}^{-1} \bar{x}_0' - \frac{1}{N}, \tag{4.18}$$

where $\bar{x}_0$ is the column-wise average of $\widetilde{X}_0$.

For the HIV study, formula (4.17) yielded standard errors $(.0087, .014, .014)$ for the empirical null estimates $(\widehat{p}_0, \widehat{\delta}, \widehat{\sigma}) = (0.917, -0.10, 0.735)$. The objection here is that z_i's are likely to be correlated in a microarray study, which would usually increase $\text{cov}(y)$ above the Poisson estimate $\text{diag}(\widehat{\nu})$ used in (4.10). ("Correlated" refers to the random errors in the expression readings, not the fact that genes have related functions; if for example $z_i \overset{\text{ind}}{\sim} N(\mu_i, 1)$ as in (4.1), then it is easy to show that $\text{cov}(y)$ will be *smaller* than $\text{diag}(\nu)$, even if the μ_i's for related genes tend toward similar values.)

Other methods of microarray error assessment, not requiring independence, may be available: resampling microarrays instead of genes (the latter giving almost the same results as (4.11) or (4.17)); blocking genes into groups suspected to be intracorrelated, and then bootstrapping or jackknifing with the groups as units; decomposing the gene-microarray data matrix into some form of random effects model that can then be resampled to give presumably more dependable standard error estimates. The HIV study, with its small number of microarrays and uncertain correlation structure sprawling across both genes and arrays, is not a promising candidate for these methods. The independence-based results of this section are useful even when not definitive, serving as lower bounds on variability for microarray analysis; *locfdr* returns standard errors from (4.11) along with $\widehat{\text{fdr}}(z)$.

5. Power Diagnostics

The microarray statistics literature has focussed on controlling Type I error, false rejection of genuinely null cases. Dudoit et al. (2003) provides a good review. Local fdr methods can also help assess power, the probability of rejecting genuinely non-null cases. This section discusses power diagnostics based on $\widehat{\text{fdr}}(z)$, showing for example why the HIV study might easily fail to identify important genes. The emphasis here is on diagnostic statistics that are dependable and simple to calculate.

The *Null subdensity*

$$f_1^+(z) = p_1 f_1(z) = (1 - \text{fdr}(z))f(z), \tag{5.1}$$

the last equality following from (2.5)-(2.6), plays a central role in power calculations. Integrating $f_1^+(z)$ yields the non-null proportion $p_1 = 1 - p_0$.

$$p_1 = \int_{-\infty}^{\infty} f_1^+(z)dz = \int_{-\infty}^{\infty} (1 - \text{fdr}(z))f(z)dz, \tag{5.2}$$

so that

$$f_1(z) = (1 - \text{fdr}(z))f(z) / \int_{-\infty}^{\infty} (1 - \text{fdr}(z'))f(z')dz'. \tag{5.3}$$

Power diagnostics are obtained by comparing $f_1(z)$ with $\text{fdr}(z)$. We hope to see $f_1(z)$ supported in regions having low values of $\text{fdr}(z)$.

The fdr methodology of Section 3 provides a useful estimate of f_1. Returning to notation (3.1), (3.2), with counts y_k in K bins of width Δ and midpoints $z_{(k)}$, let $\widehat{f}_k = \widehat{f}(z_{(k)})$ and $\widehat{\text{fdr}}_k = \widehat{\text{fdr}}(z_{(k)})$, where $\widehat{f}$ and $\widehat{\text{fdr}}$ are obtained as in Section 3. Substituting into (5.2), (5.3) gives estimates

$$\widehat{p}_1 = \sum_{k=1}^{K} (1 - \widehat{\text{fdr}}_k)\widehat{f}_k = 1 - \widehat{p}_0 \tag{5.4}$$

and

$$\widehat{f}_{1k} \equiv \widehat{f}_1(z_{(k)}) = (1 - \widehat{\text{fdr}}_k)\widehat{f}_k / \widehat{p}_1, \tag{5.5}$$

The latter is shown as the heavy curve in Figure 5. It is similar to "f_1" in Figure 2 of Efron et al. (2001) (though now based on a more stable estimation methodology), where the goal was to choose, from a class of modified student t formulas, summary statistics "z_i" that maximized the number of genes having $\widehat{\text{fdr}} \leq 0.10$. Here the form of the summary statistic is assumed given, as in (1.1), the goal being to assess the power of the resulting analysis.

Power diagnostics are obtained from the comparison of $\widehat{f}_1(z)$ with $\widehat{\text{fdr}}(z)$. The expectation of $\widehat{\text{fdr}}$ under f_1, say $\widehat{\text{Efdr}}_1$, provides a particularly simple diagnostic,

$$\begin{aligned} \widehat{\text{Efdr}} &= \sum_{k=1}^{K} \widehat{\text{fdr}}_k \widehat{f}_{1k} \\ &= \sum_{k=1}^{K} \widehat{\text{fdr}}_k (1 - \widehat{\text{fdr}}_k)\widehat{f}_k \Big/ \sum_{k=1}^{K} (1 - \widehat{\text{fdr}}_k)\widehat{f}_k, \end{aligned} \tag{5.6}$$

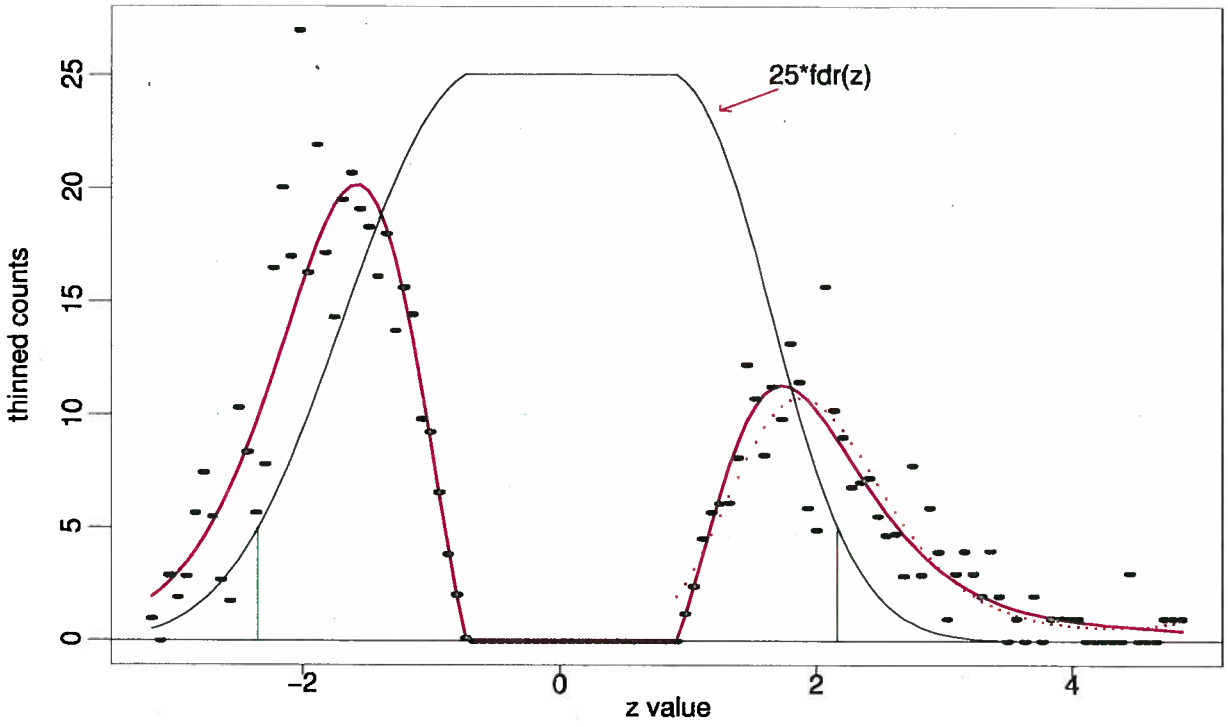

Figure 5: *Heavy curve proportional to non-null density estimate $\widehat{f}_1(z)$, (5.5), for HIV study; light curve proportional to $\widehat{\text{fdr}}(z)$. Points are thinned counts (5.9); a regression curve, dotted, has been fit directly to the thinned counts on the right.*

the expected non-null false discovery rate. A small value of $\widehat{\text{Efdr}}_1$, suggests good power, with a typical non-null case likely to show up on a list of interesting candidates for further study.

Table 3 shows $\widehat{\text{Efdr}}_1$'s behavior in simulation (4.1), $N = 1500$. The situation is seen to be favorable, with $\widehat{\text{Efdr}}_1$, averaging only 0.23 or 0.29 using empirical or theoretical nulls (Section 7 explains the disparity between the two $\widehat{\text{Efdr}}_1$ values). Moreover the individual $\widehat{\text{Efdr}}_1$ values were reasonably stable, having standard deviations only 0.04 or 0.06. The empirical null performs well here, in contrast to Table 3.

	empirical null		theoretical null	
	$\widehat{p}_1$	$\widehat{\text{Efdr}}_1$	$\widehat{p}_1$	$\widehat{\text{Efdr}}_1$
mean:	0.76	**.232**	.085	**.285**
stdev:	.011	.040	.015	.060
coeffvar:	.14	.17	.18	.21

Table 3: Means, standard deviations, and coefficients of variation of $\widehat{p}_1$ and $\widehat{\text{Efdr}}_1$ for $N = 1500$ case of Tables 1 and 2.

On the other hand, $\widehat{\text{Efdr}}_1$ equals 0.45 for the HIV study (by necessity using the empirical null) so a typical non-null gene is likely to receive a substantial fdr estimate, high enough to exclude it from the list of those having $\widehat{\text{fdr}} \leq 0.20$.

The Efdr_1, computations can be carried out separately to the left and right of $z = 0$ by appropriately restricting the range of summation in the numerator and denominator of (5.6). Doing so gives $\widehat{\text{Efdr}}_{\text{left}} = 0.51$ and $\widehat{\text{Efdr}}_{\text{right}} = 0.35$ for the HIV data. This says that it will be particularly difficult to detect genes that *underexpress* in HIV-positive subjects.

Other moments or probabilities of $\widehat{\text{fdr}}$ with respect to $\widehat{f}_1$ are as simple to calculate as $\widehat{\text{Efdr}}_1$, for example the standard deviation

$$\widehat{\text{Sd}}_1 = \left[\sum_{k=1}^{K} \widehat{\text{fdr}}_k^2 \cdot f_{1k} - \widehat{\text{Efdr}}_1^2\right]^{1/2}, \tag{5.7}$$

which equals 0.30 for the HIV data. The possibility of dependence among the z_i's does not bias estimates such as (5.6) or (5.7), though it increases their variability.

Going further, we can examine the entire distribution of $\widehat{\text{fdr}}$ under $\widehat{f}_1$. The heavy curve in Figure 6 shows the $\widehat{f}_1$ cdf of $\widehat{\text{fdr}}$ for the HIV study,

$$\widehat{G}(x) = \sum_{\widehat{\text{fdr}}_k \le x} \widehat{f}_{1k} \Big/ \sum_{k=1}^{K} \widehat{f}_{1k}. \tag{5.8}$$

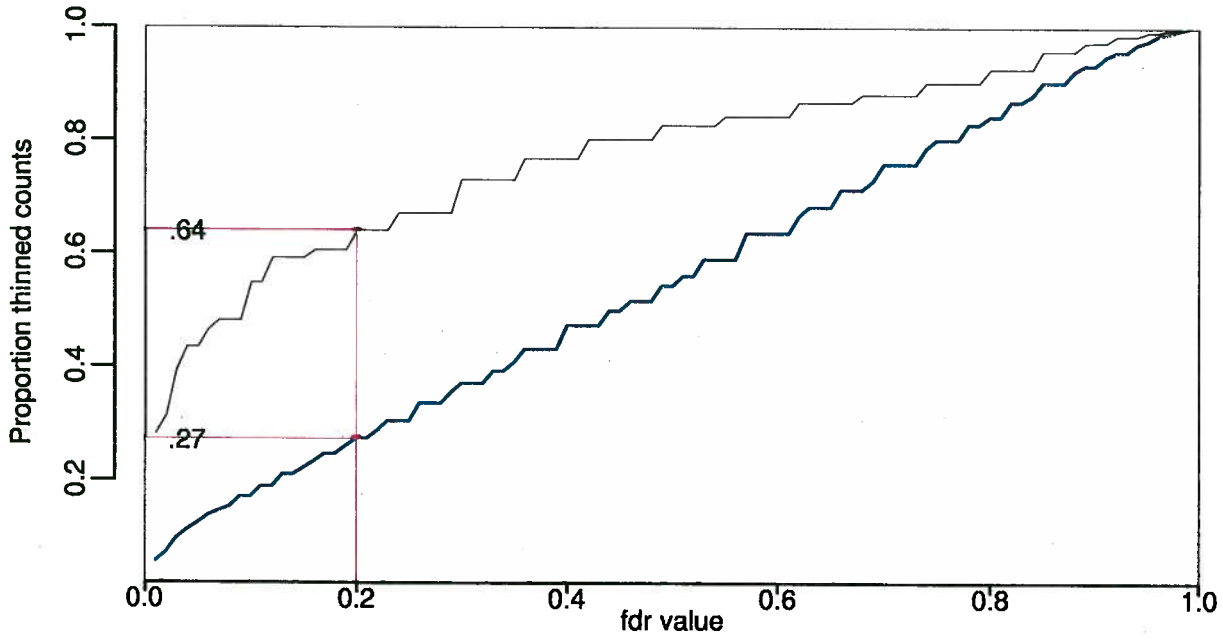

Figure 6: *Empirical cdf of* $\widehat{\text{fdr}}$ *with respect to estimated non-null density* $\widehat{f}_1$. *Heavy curve HIV study; Light curve first simulated sample (4.1),* $N = 1500$. *The simulated sample displays greater power.* $\widehat{\text{Efdr}}_1$, *equals 0.45 for HIV study, 0.23 for simulation.*

For instance $\widehat{G}(.2) = 0.27$, so only 27% of the non-null cases are estimated to have fdr values less than 0.20. By contrast, the first of the $N = 1500$ simulated data sets from Table 1 is seen to have much greater power, with $\widehat{G}(.2) = 0.64$.

The limitations of the HIV study are forcefully illustrated by Figure 6: if we wish to report 50% of the non-null cases then we must tolerate $\widehat{\text{fdr}}$ values as high as $0.45 = \widehat{G}^{-1}(.5)$, etc.

The *thinned counts* appearing in Figures 1, 2, and 5 are defined in terms of original counts y_k as

$$y_{1k} = (1 - \widehat{\text{fdr}}_k) \cdot y_k. \tag{5.9}$$

Since $1 - \text{fdr}_k$ is the probability of being non-null for a case in the kth bin, y_{1k} is, nearly, an unbiased estimate of the number of non-null cases in bin k. We can use the thinned counts to carry out sample size power calculations for large-scale studies.

Traditional sample size calculations employ preliminary data to predict how large an experiment will be required for effective power. Here we might ask, for instance, if doubling the number of subjects in the HIV study would substantially improve its detection rate. To answer the question we assume a homoskedastic model for the z-values,

$$z_i \sim (\mu_i, \sigma_0^2), \tag{5.10}$$

the notation indicating that z_i has expectation μ_i, its "true score" and variance σ_0^2, with $\mu_i = 0$ for the null cases. Sections 6 and 7 discuss the rationale for (5.10).

We imagine that c independent replicates of (5.10) are available for each case, from which a combined statistic $\widetilde{z}_i$ is formed,

$$\widetilde{z}_i = \sum_{j=1}^{c} z_{ij}/\sqrt{c} \sim (\sqrt{c}\,\mu_i, \sigma_0^2) \tag{5.11}$$

This definition maintains the distribution of the null cases, $\widetilde{z}_i \sim (0, \sigma_0^2)$, while moving the non-null true scores away from zero 1 by factor $\sqrt{c}$.

Consider a subset of the non-null cases in which the true scores have empirical mean and variance say (a, b^2). A randomly selected z statistic "Z" from this subset has marginal mean and variance

$$Z \sim (A, B^2) = (a, b^2 + \sigma_0^2) \tag{5.12}$$

according to (5.10), while the corresponding statistic "$\widetilde{Z}$" from (5.11) has

$$\widetilde{Z} \sim (\widetilde{A}, \widetilde{B}^2) = (\sqrt{c}\,a, cb^2 + \sigma_0^2) \tag{5.13}$$

Comparing (5.13) with (5.12) shows that the simple formula

$$\widetilde{Z} = \sqrt{c}\,A + d(Z - A), \quad [d^2 = c - (c-1)\sigma_0^2/B^2], \tag{5.14}$$

gives $\widetilde{Z}$ the correct mean and variance.

From the thinned counts (5.9) on the right side of Figure 5 we estimate

$$\widehat{A} = \frac{\sum z_{(k)} y_{1k}}{\sum y_{1k}} = 2.23 \quad \text{and} \quad \widehat{B} = \left[\frac{\sum z_{(k)}^2 y_{1k}}{\sum y_{1k}} - \widehat{A}^2\right]^{\frac{1}{2}} = 0.87, \tag{5.15}$$

the sums being over $z_{(k)} > 0$. Then (5.14), with $\sigma_0 = 0.735$ from the empirical null, describes how the right-side non-null z_i's might transform under increased sample sizes (5.11). A similar calculation applies on the left, while the null scores, of which there are $y_k - y_{1k}$ in the kth bin, would remain unchanged.

Table 4 reports on $\widehat{\text{Efdr}}_1$, (5.6), for hypothetical transformed data sets having $c = 1, 1.5, 2$, and 2.5. We see that doubling the number of subjects, from 4 to 8 in each group, would reduce $\widehat{\text{Efdr}}_1$ from 0.45 to 0.28, a substantial improvement. Table 3 involves a considerable amount of speculation, more so than diagnostics (5.6)-(5.8), but power computations are traditionally speculative; the calculations here, involving just means and variances, are fashioned to minimize the amount of parametric modeling.

The dotted curve on the right side of Figure 5 is a cubic Poisson GLM fit directly to the thinned counts y_{1k}; that is, we assume

$$y_{1k} \overset{\text{ind}}{\sim} Po(\nu_{1k}), \tag{5.16}$$

for $\log(\nu_{1k})$ a cubic polynomial in the bin midpoints $z_{(k)}$, say

$$(\log \nu_{1k}) = X_1 \beta, \tag{5.17}$$

with X_1 a $K_1 \times m_1$ structure matrix; K_1 is the number of bins involved and m_1 the number of parameters, $m_1 = 4$ here.

#Subjects	4-4	6-6	8-8	10-10
$\widehat{\text{Efdr}}_1$:	**0.45**	**0.33**	**0.28**	**0.22**

Table 4: Estimated values of $\widehat{\text{Efdr}}_1$, for expanded versions of HIV study; doubling the study, to 8 subjects each in the two groups reduces $\widehat{\text{Efdr}}_1$ from 0.45 to 0.28.

The usual GLM estimate of covariance for $\widehat{\beta}$ is

$$\widehat{G}_1^{-1} = (X_1' \text{ diag } (\widehat{\nu}_{1k}) X_1)^{-1}. \tag{5.18}$$

However this leads to an overestimate under model (3.1), because $y_{1k} = (1 - \widehat{\text{fdr}}_k) y_k$ has variance about $(1 - \text{fdr}_k)\nu_{1k}$, less than the Poisson value ν_{1k} assumed in (5.16). A more accurate approximation is

$$\widehat{\text{Cov}}(\widehat{\beta}) = \widehat{G}_1^{-1} [X_1' \text{ diag } ((1 - \widehat{\text{fdr}}_k)\widehat{\nu}_{1k}) X_1)^{-1} \widehat{G}_1^{-1}, \tag{5.19}$$

Estimating $\widehat{f}_1$ directly from the thinned counts is appealing since it does not involve a global fit to all N cases, as does (5.5), a fact we took advantage of in using only a cubic model for the dotted curve. It did not make much difference to the HIV analysis though, nor did simply replacing $\widehat{f}_{1k}$ with y_{1k} in (5.6)-(5.8).

6. The Non-Null Distribution of z-values

A key assumption of our fdr estimation methodology was the smooth nature of the z-value mixture density $f(z)$. This section discusses a useful approximation for the distribution of z-values, null or non-null,

$$Z \dot{\sim} N(\mu, \sigma_\mu^2), \tag{6.1}$$

where Z represents a generic z-value, μ its expectation, "$\dot{\sim}$" indicates second order accuracy with distributional errors of order $O(n^{-1})$ in the usual repeated sampling context, and $\sigma_\mu \doteq 1 + O(n^{-\frac{1}{2}})$. The smoothness assumption is justified by (6.1), which represents $f(z)$ as a well-controlled mixture of normal densities.

Figure 7 illustrates (6.1) for transformed t-statistics (1.1). We suppose that t_i has a noncentral t distribution, noncentrality θ and degrees of freedom ν,

$$t_i \sim \frac{\theta + W}{S^{1/2}} \quad [W \sim N(0, \tau) \text{ independent of } s \sim \tau x_\nu^2/\nu]. \tag{6.2}$$

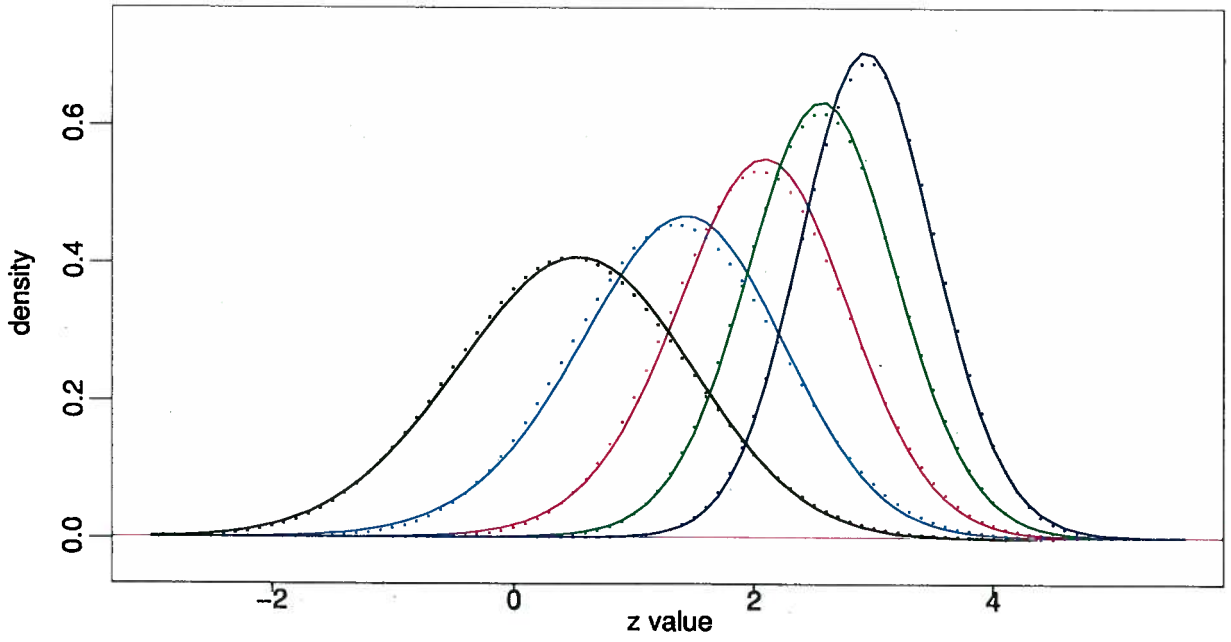

Figure 7: *Density of z-value (1.1) when t_i is non-central t variate, 6 degrees of freedom; non-centrality parameter $\theta = .5, 1.5, 2.5, 3.5, 4.5$ left to right. Means 0.42, 1.35, 2.04, 2.56, 2.96; stdevs 0.98, 0.88, 0.75, 0.65, 0.58. Dotted curves are corresponding normal densities.*

By definition $z_i \sim N(0,1)$ in the null case $\theta = 0$. (For the calculations of this section we are ignoring the failure of the theoretical null in Figure 4.) Figure 7 shows the density of $z_i = \Phi^{-1}(F_6(t_i))$ for $\theta = .5, 1.5, \cdots, 4.5$. We see σ_μ declining from 1 at $\theta = 0$ to 0.58 at $\theta = 4.5$, while the normality claimed in (6.1) is nicely maintained.

Relationship (6.1) can be verified in a wide variety of situations. Suppose Z is based on testing $H_0 : \theta = 0$ for a summary statistic $\widehat{\theta}$ having cdf F_θ,

$$Z = \Phi^{-1}(F_0(\widehat{\theta})). \tag{6.3}$$

We assume that $\widehat{\theta}$ behaves asymptotically like a maximum likelihood estimate in terms of a notional sample size "n", its bias, standard deviation, skewness, and kurtosis having the appropriate orders of magnitude,

$$\widehat{\theta} - \theta \sim (B_\theta/n, C_\theta/\sqrt{n}, D_\theta/\sqrt{n}, E_\theta/n); \tag{6.4}$$

$B_\theta, C_\theta, D_\theta$, and E_θ are smooth bounded functions of θ and n.

Following Sections 3-5 of Efron (1987), particularly Theorem 1, there exists a monotone increasing transformation $\widehat{\phi} = g(\widehat{\theta})$, $\phi = g(\theta), 0 = g(0)$, such that

$$\widehat{\phi} \dot{\sim} \phi + (1 + a\phi)(W - z_0), \tag{6.5}$$

with $W \sim N(0,1)$ and a and z_0, the "acceleration" and "bias-correction" constants, each of order $O(n^{-\frac{1}{2}})$. At $\theta = \phi = 0$ we have $\widehat{\phi} \dot{\sim} N(-z_0, 1)$, implying $Z \doteq \widehat{\phi} + z_0$. Then (6.4) gives

$$Z \dot{\sim} \phi(1 - az_0) + (1 + a\phi)W \dot{\sim} N(\phi, (1 + a\phi)^2), \tag{6.6}$$

($az_0 = O(n^{-1})$ being ignorably small) verifying (6.1) with $\mu = \phi$ and

$$\sigma_\mu = 1 + a\mu. \tag{6.7}$$

The acceleration constant "a" determines how quickly σ_μ departs from $\sigma_0 = 1$. Efron (1987) derives approximation $a = \text{skew}\ (\dot{\ell}_0)/6$, in terms of the score function $\dot{\ell}_0$ at $\theta = 0$.

As an example suppose we observe scaled one-sided exponential varieties,

$$y_1, y_2, \ldots, y_n \overset{\text{ind}}{\sim} \theta G_1 \quad [Pr\{G_1 < x\} = 1 - e^{-x}], \tag{6.8}$$

so that $\widehat{\theta} = \bar{y} \sim \theta\ \text{Gamma}_n/n$. For $n = 10$, and for any choice of the null hypothesis $H_0 : \theta = \theta_0$, the score function approximation gives $a = 1/(3\sqrt{10}) = .1054$, while direct numerical computation yielded

$$\left.\frac{d\sigma_\mu}{d\mu}\right|_{\theta_0} = .1049; \tag{6.9}$$

σ_μ varied on the range $[0.5, 1.5]$ for μ in ± 5. The normal approximation is just as impressive here as in Figure 7.

The gist of (6.1), (6.7) is that as μ departs from zero by amount $O(1)$, σ_μ changes by $O(n^{-\frac{1}{2}})$ while normality decays by only $O(n^{-1})$. The student t example of Figure 7 is not included in development (6.4)-(6.7), because of the nuisance parameter τ in (6.2), but in fact showed even greater accuracy for (6.1). This can be verified using the diagnostic function in Efron (1982).

Going further, we can consider the situation where Z is the z-value for a single parameter in a multiparameter family. A promising conjecture is that (6.1), (6.7) holds in multiparameter exponential families if z-values are obtained via the ABC method, DiCiccio and Efron (1992). Section 4 of Efron (1988) discusses a variant of (6.1) applying to sequential sampling.

Model (6.1) can be used to sharpen the sample size calculations of Section 5. Consider a subset of cases, say the non-null cases on the right side of Figure 5, and let $g(\mu)$ represent the empirical distribution of their true scores μ_i. Formulas (5.12)-(5.15) tacitly involve estimating $g(\mu) : (\widehat{A}, \widehat{B}^2)$, the mean and variance of the thinned counts, give estimates $(\widehat{a}, \widehat{b}^2)$ for the mean and variance of $g(\mu)$, (5.12), which depend on the homoskedastic model (5.10). Instead we could begin with (6.1) and directly deconvolve the thinned counts to obtain $\widehat{g}(\mu)$. Doing so made little difference to Table 4. However $\widehat{g}(\mu)$ can be useful in its own right, in particular for estimating the Bayes posterior distribution of true score μ_i given z_i.

7. Structure and Bias

Model (2.2) envisions two groups of cases, null and non-null. Realistic examples of large-scale inference are apt to be less clearcut, with true effect sizes ranging smoothly from zero or near zero to very large. Here we consider a "one-class" structural model that allows for smooth effects. We can still usefully apply fdr methods to data from one-class models; doing so helps clarify the choice between theoretical and empirical null hypothesis and explicates the biases inherent in model (2.2).

For the theoretical developments of this section we consider a Bayesian structural model where each true score μ_i is drawn randomly according to a prior density $g(\mu)$, with z_i then normally distributed around μ_i,

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim N(\mu, 1). \tag{7.1}$$

(We could use $N(\mu, \sigma_\mu^2)$ as in (6.1), but at the expense of complicating the formulas that follow.) The density $g(\mu)$ is allowed to have discrete atoms. It might have an atom at zero, as in (4.1), but this is not required, and in any case there is no *a priori* partition of $g(\mu)$ into null and non-null components.

Model (7.1) gives mixture density

$$f(z) = \int_{-\infty}^{\infty} \varphi(\mu - z) g(\mu) d\mu \qquad \left[\varphi(x) = e^{-\frac{1}{2}x^2}/\sqrt{2\pi}\right], \tag{7.2}$$

with

$$f(0) = \int_{-\infty}^{\infty} \varphi(\mu) g(\mu) d\mu. \tag{7.3}$$

The idea in what follows is to generalize the construction of Figure 4 by approximating $\ell(z) = \log f(z)$ with Taylor series other than quadratic.

The Jth Taylor approximation to $\ell(z)$ is

$$\ell_J(z) = \sum_{j=0}^{J} \ell^{(j)}(0) z^j / j!, \tag{7.4}$$

where $\ell^{(0)}(0) = \log f(0)$ and for $j \geq 1$

$$\ell^{(j)}(0) = \frac{d^j \log f(z)}{dz^j}\Big|_{z=0}. \tag{7.5}$$

The sub-density

$$f_0^+(z) = e^{\ell_J(z)} \tag{7.6}$$

matches $f(z)$ at $z = 0$ (a convenient use of the zero assumption) and leads to an fdr expression as in (2.6),

$$\text{fdr}(z) = e^{\ell_J(z)}/f(z). \tag{7.7}$$

Larger choices of J match $f_0^+(z)$ more accurately to $f(z)$, increasing ratio (7.7); the interesting z-values, those with small fdr's, are pushed farther away from zero as we allow more of the data structure to be explained by the null density.

The Bayesian model (1.1) provides a helpful interpretation of the derivatives $\ell^{(j)}(0)$:

Lemma 2 The derivative $\ell^{(j)}(0)$, (7.5), is the jth cumulant of the posterior distribution of μ given $z = 0$, except that $\ell^{(2)}(0)$ is the second cumulant minus 1. Thus

$$\ell^{(1)}(0) = E_0 \quad \text{and} \quad -\ell^{(2)}(0) = \bar{V}_0, \tag{7.8}$$

where E_0 and $V_0 \equiv 1 - \bar{V}_0$ are the posterior mean and variance of μ given $z = 0$.

Proof We have

$$\begin{aligned} \ell(z) &= \log \textstyle\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(\mu-z)^2}}{\sqrt{2\pi}} g(\mu)d\mu \\ &= -\tfrac{1}{2}z^2 + \log\ f(0) + \log\ \textstyle\int_{-\infty}^{\infty} e^{z\mu}[\varphi(\mu)g(\mu)/f(0)]d\mu. \end{aligned} \tag{7.9}$$

Notice that $m(z) \equiv \int_{-\infty}^{\infty} e^{z\mu}[\varphi(\mu)g(\mu)/f(0)]d\mu$ is the moment generating function of the probability density $\varphi(\mu)g(\mu)/f(0)$,

$$\frac{d^j m(z)}{dz^j}|_{z=0} = \int_{-\infty}^{\infty} \mu^j \frac{\varphi(\mu)g(\mu)}{f(0)} d\mu, \tag{7.10}$$

the last expression also being the posterior jth moment of μ given $z = 0$. The usual relationship between moments and cumlants, applied to the function $\ell(z) + \frac{1}{2}z^2 - \log\ f(0)$, verifies the Lemma.

For $J = 0, 1, 2$, formulas (7.7), (7.8) yield simple expressions for p_0 and $f_0(z)$ in terms of $f(0), E_0$, and $\bar{V}_0$. These are summarized in Table 5 (with p_0 obtained through definition (2.4),

$$p_0 = \Big[\int_{-\infty}^{\infty} f_0^+(z)dz\Big]^{-1}. \tag{7.11}$$

Formulas are also available for $\text{Fdr}(z)$, (2.8).

The choices $J = 0, 1, 2$ in Table 5 result in a normal null density $f_0(z)$, the only difference being the means and variances. Going to $J = 3$ allows for an asymmetric choice of $f_0(z)$; from (7.9) and the Lemma,

$$\text{fdr}(z) = \frac{f(0)}{f(z)}\, e^{E_0 z - \bar{V}_0 z^2/2 + S_0 z^3/6}, \tag{7.12}$$

J:	0	1	2
p_0:	$f(0)\sqrt{2\pi}$	$f(0)\sqrt{2\pi}\, e^{E_0^2/2}$	$f(0)\sqrt{\frac{2\pi}{\bar{V}_0}}\, e^{E_0^2/2\bar{V}_0}$
$f_0(z)$:	$N(0,1)$	$N(E_0,1)$	$N(E_0/\bar{V}_0,\ 1/\bar{V}_0)$
fdr(z):	$\frac{f(0)e^{-z^2/2}}{f(z)}$	$\frac{f(0)e^{E_0 z - z^2/2}}{f(z)}$	$\frac{f(0)e^{E_0 z - \bar{V}_0 z^2/2}}{f(z)}$

Table 5: Expressions for p_0, f_0 and fdr, first three choices of J in (7.6), (7.7); numerator of fdr(z) is $f_0^+(z)$. $J = 0$ gives theoretical null, $J = 2$ empirical null; $f(z)$ from (7.2).

	p_0	δ	σ	$p_0^{(\text{theo})}$	E_0	V_0
Model (7.13):	0.916	0.013	1.01	0.906	0.012	0.022
Model (7.14):	0.918	0.018	1.13	0.821	0.014	0.223

Table 6: p_0 and $f_0(z)$ from Table 5; δ and σ mean and standard deviation of empirical null, *Top line* Model (7.13), as used in simulation study; *Bottom line* Model (7.14).

where S_0 is the posterior third central moment of μ given $z = 0$ in model (7.1). The program *locfdr* uses a variant, the "split normal", to model asymmetric null densities with the exponent of (7.12) replaced by a quadratic spine in z.

Lemma 2 bears on the difference between empirical and theoretical nulls. Suppose that the probability mass of $g(\mu)$ occurring within a few units of the origin is concentrated in an atom at $\mu = 0$. Then the posterior mean and variance (E_0, V_0) of μ given $z = 0$ will be near 0, making $(E_0, \bar{V}_0) \doteq (0, 1)$. In this case the empirical null ($J = 2$) will approximate the theoretical null ($J = 0$). Otherwise the two nulls will differ; in particular, any mass of $g(\mu)$ around zero increases V_0, swelling the standard deviation $(1 - V_0)^{-\frac{1}{2}}$ of the empirical null.

Model (4.1), used for the simulation study, has

$$g(\mu) = 0.9 \cdot I_0(\mu) + 0.1 \cdot \varphi_{3,1}(\mu), \tag{7.13}$$

$I_0(\mu)$ a unit atom at $\mu = 0$, which gives mixture density $f(z) = 0.9 \cdot \varphi_{0,1}(z) + 0.1 \cdot \varphi_{3,\sqrt{2}}(z)$ according to (7.1). The top line of Table 6 shows p_0 and $f_0(z)$ for (7.13), as calculated from Table 5. This amounts to having N equal infinity in Table 2 (except for matching $f_0^+(z)$ to $f(z)$ at $z = 0$ in (7.6) instead of averaging over the central third of f as in *locfdr*). We see small biases away from $f_0^+(z) = 0.9 \cdot \varphi_{0,1}(z)$: p_0 exceeds 0.9, more so for the empirical null, and (δ, σ) is slightly distorted from $(0, 1)$.

Bias is more apparent in the left panel of Figure 8, which plots $f_1(z) = (1 - \text{fdr}(z)) \cdot f(z)$ as calculated from Table 5. The left tail of $f_1(z)$ is pushed away from zero compared to the nominal f_1 density $\varphi_{3,\sqrt{2}}(z)$, again more so for the empirical null. The zero assumption is the culprit here, as mentioned before, since in fact $\varphi_{0,1}$ and $\varphi_{3,\sqrt{2}}$ overlap somewhat. The empirical's greater rightward shift, toward smaller fdr values, accounts for its smaller Efdr_1 average in Table 3: computing $\text{Efdr}_1 = \int \text{fdr}(z) \cdot f_1(z) dz$ according to Table 5 gives 0.245 for

the empirical null and 0.288 for the theoretical, close to Table 3's simulation values.

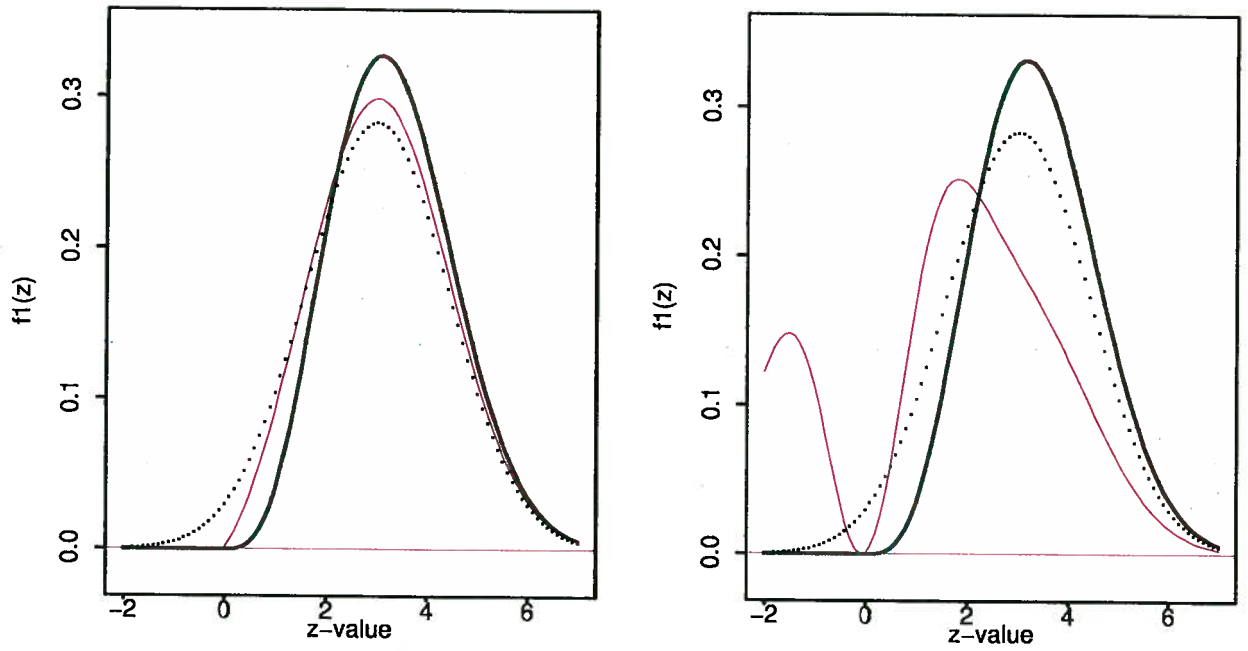

Figure 8: *Non-null density $f_1(z)$ computed from Table 5, using empirical null (heavy curve) or theoretical null (light curve); dots indicate nominal f_1 density $\varphi_{3,\sqrt{2}}(z)$. Left panel model (7.13); Right panel model (7.14).*

"Bias" can be a misleading term in model (7.1) since it tacitly assumes that each μ_i is clearly defined as either null or non-null. This seems clear enough in (7.13), where we took the 0.90 atom at $\mu = 0$ as null. Suppose though

$$g(\mu) = 0.9 \cdot \varphi_{0,.5}(\mu) + 0.1 \cdot \varphi_{3,1}(\mu), \tag{7.14}$$

which gives mixture density $f(z) = 0.9 \cdot \varphi_{0,1.12}(z) + 0.1 \cdot \varphi_{3,\sqrt{2}}(z)$. This might characterize an observational study, in which a crisp model like (7.13) has been blurred by uncontrolled covariates that cause even the "null" cases to have slightly non-zero μ_i values; see Section 4 of Efron (2004). The null/non-null distinction is less obvious in (7.14), though it still makes sense to apply model (2.2) to the search for cases that have μ_i far from 0.

The right panel of Figure 8 and the bottom line of Table 6 show the fdr analysis of Table 5 applied to model (7.14). The empirical null now estimates $f_0(z)$ as $N(0.02, 1.13^2)$, closely matching the $N(0, 1.12^2)$ first component of $f(z)$. This results in nearly the same estimates of p_0 and $f_1(z)$ as for (7.13). The fdr analysis of an actual data set $z_1, z_2, \ldots, z_N$ arising from (7.1) would identify nearly the same set of non-null cases for either (7.13) or (7.14).

Analysis based on the theoretical null changes drastically in (7.14). Twice as many cases, some for $z_i < 0$, are now identified as non-null, $p_1 = 0.179$ instead of 0.094, with the principal mode of $f_1(z)$ moved sharply towards $z = 0$. This example highlights the difference in "significance" as judged be the theoretical and empirical nulls: simply put, the empirical null judges significance in the extremes by the spread of the central z_i's, while the theoretical null uses an absolute criterion. Every inference method, Fdr, FWER,

permutation, Bonferroni, and not just fdr, yields doubtful results if model (7.14) is analyzed in terms of the theoretical null.

Summary The local false discovery rate methodology developed in Sections 3 and 5 is based on empirical Bayes analysis of the simple two-class model (2.2); fdr calculations provide both size and power estimates, while requiring a minimum of frequentist or Bayesian modeling assumptions. The methodology applies to large-scale situations, with hundreds of inference problems considered simultaneously, perhaps at least a thousand if the theoretical null hypothesis is unsatisfactory. A closed form error analysis of fdr estimation, developed in Section 4, is available when the inference problems are independent. Even when the two-class model is dubious, as discussed in Section 7, fdr methods can still be informative, though now they are more likely to require empirical estimation of the null hypothesis. All calculations are carried through using standard Poisson GLM software; program *locfdr* is available from the R library CRAN.

References

Allison, D., Gadbury, G., Heo, M., Fernandez, J., Lee, C.K., Prolla, T., and Weindruch, R. (2002), "A mixture model approach for the analysis of microarray gene expression data", *Computational Statistics and Data Analysis* **39**, 1-20.

Aubert, J., Bar-hen, A., Daudin, J., and Robin, S. (2004), "Determination of the differentially expressed genes in microarray experiments using local FDR", *BMC Bioinformatics* **5**:125.

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

Broberg, P. (2005), "A new estimate of the proportion unchanged genes in a microarray experiment", *Genome Biology* **5**:p10.

DiCiccio, T. and Efron, B. (1992), "More accurate confidence intervals in exponential families", *Biometrika* **79**, 231-45.

Do, K.A., Mueller, P., and Tang, F. (2003), "A nonparametric Bayesian mixture model for gene expression", mbi.osu.edu/2004/ws1materials/do.pdf. To appear *Jour. Roy. Stat. Soc. C.*

Dudoit, S., van der Laan, M., and Pollard, K. (2004), "Multiple testing, part I: Single step procedures for the control of general Type I error rates", *Statistical Applications in Genetics and Molecular Biology* **3**, article 13: http://www.bepress.com/sagmb/vol3/iss1/art13.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003), "Multiple hypothesis testing in microarray experiments", *Statistical Science* **18**, 71-103.

Efron, B. (2004), "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis", *JASA* **99**, 96-104.

Efron, B., and Tibshirani, R. (2002), "Empirical Bayes methods and false discovery rates for microarrays", *Genetic Epidemiology* **23**, 70-86.

Efron, B. and Gous, A. (2001), "Scales of evidence for model selection: Fisher versus Jeffreys", *Model Selection* IMS Monograph **38**, 208-256.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment", *Journal of the American Statistical Association* **96**, 1151-1160.

Efron, B. and Tibshirani, R. (1996), "Using specially designed exponential families for density estimation", *Annals Stat.* **24**, 2431-61.

Efron, B. (1988), "Three examples of computer-intensive statistical inference", *Sankhya* **50**, 338-362.

Efron, B. (1987), "Better bootstrap confidence intervals", *Jour. Amer. Stat. Assn.* **82**, 171-95.

Efron, B. (1982), "Transformation theory: how normal is a family of distributions?", *Annals Stat.* **10**, 323-339.

Genovese, C. and Wasserman, L. (2004), "A stochastic process approach to false discovery control", *Annals Stat.* **32**, 1035-1061.

Gottardo, R., Raftery, A., Yeung, K., and Bumgarner, R. (2004), "Bayesian robust inference for differential gene expression in microarrays with multiple samples", Dept. Statistics, U. Washington technical report, raph@stat.washington.edu

Heller, G. and Qing, J. (2003), "A mixture model approach for finding informative genes in microarray studies", Unpublished.

Johnstone, I. and Silverman, B. (2004), "Needles and straw in a haystacks: empirical Bayes approaches to thresholding a possibly sparse sequence", *Annals Stat.* **32**, 1594-1649.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003), "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles", *Stat. in Medicine* **22**, 3899-3914.

Kerr, M., Martin, M., and Churchill, G. (2000), "Analysis of variance in microarray data", *Jour. Computational Biology* **7**, 819-837.

Lee, M.L.T., Kuo, F., Whitmore, G., and Sklar, J. (2000), "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations", *Proc. Nat. Acad. Sci.* **97**, 9834-38.

Liao, J., Lin, Y., Selvanayagam, Z., and Weichung, J. (2004), "A mixture model for estimating the local false discovery rate in DNA microarray analysis", *Bioinformatics* **20**, 2694-2701.

Newton, M., Noveiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting differential gene expression with a semiparametric hierarchical mixture model", *Biostatistics* **5**, 155-176.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001), "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data, *Jour. Comp. Biology* **8**, 37-52.

Pan, W., Lin, J., Le, C. (2003), "A mixture model approach to detecting differentially expressed genes with microarray data", *Functional & Integrative Genomics* **3**, 117-24.

Pollard, K. and van der Laan, M. (2003), "Resampling-based multiple testing: asymptotic control of type I error and applications to gene expression data", U.C. Berkeley Biostatistics working paper 121; http://www/bepress.com/ucbiostat/paper121

Pounds, S. and Morris, S. (2003), "Estimating the occurrence of false positions and false negatives in microarray studies by approximating and partitioning the empirical distribution of the p-values", *Bioinformatics* **19**, 1236-42.

Storey, J., Taylor, J., and Siegmund, D. (2004), "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates; a unified approach", *Jour. Royal Stat. Soc. B* **66**, 187-206.

Storey, J. (2002), "A direct approach to false discovery rates", *Journal of the Royal Statistical Society*, Ser. B, **64**, 479-498.

Tusher, V., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to transcriptional responses to ionizing radiation", *Proc. Nat. Acad. Sci.* **98**, 5116-21.

Westfall, P. and Young, S. (1993), *Resampling-based multiple testing: examples and methods for p-value adjustment*, New York, Wiley.