Selection and Estimation for Large-Scale Simultaneous Inference

Bradley Efron
Stanford University

Abstract

Modern scientific technology is providing a new class of simultaneous inference problems for the applied statistician, where there are hundreds or thousands or even more 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, and functional Magnetic Resonance Imaging. The paper considers two related questions: given a large number of simultaneous hypothesis testing problems, how can we Select the Non-Null cases? And how can we Estimate effect sizes for the Non-Nulls? Selection and Estimation combine to assess power, a large-scale experiment's ability to correctly identify individual cases of interest. Microarray data is used to illustrate a simple methodology that requires a minimum of mathematical modeling.

This research was supported by grants NSF DMS-0072360 and NIH 8R01 EB002784.

1. Introduction

Massive data sets have become a fact of current-day life for applied statisticians. Often these sets have parallel structure comprising a large number "N" of similar problems. For example a microarray experiment comparing control subjects with treatment subjects might produce N two-sample t-tests, one for each of N genes. "Large-scale" problems of the kind considered here typically have N at least several hundred and perhaps several thousand or even more.

The microarray literature has focussed on simultaneous hypothesis testing, where each of the N component problems is represented by its own null hypothesis and test statistic, say

$$\begin{array}{ll} Null\ Hypothesis: & H_1, H_2, H_3, \ldots, H_i, \ldots, H_N \\ Test\ Statistic: & z_1,\ z_2,\ \ z_3, \ldots,\ z_i, \ldots,\ z_N. \end{array} \tag{1.1}$$

This is the framework used in Efron (2004), which considered the choice of an appropriate density for the null hypotheses, the point there being that large-scale situations can provide their own "empirical null", which may differ in important ways from the traditional theoretical null appropriate for any individual problem. (These are different considerations than adjustments for multiple comparisons, another crucial aspect of large-scale inference.)

Frequentist, Bayesian, and empirical Bayes methods have been developed for the *Selection Problem*, i.e. how to select the Non-Null cases from among the N possibilities. Recent references include Newton et al. (2004), Dudoit et al. (2003), Pollard and van der Laan (2003), Gottardo et al. (2004), and Efron and Tibshirani (2002). However Selection by itself may not deliver the data's full message. *Estimation*, the assignment of effect sizes to the Non-Null cases, can add important information: how powerful was the test procedure? How reproducible are the results? How useful would it be to increase the experiment's size?

This paper concerns both the Selection and Estimation problems. Selection is approached via the simple Bayesian model used in Efron (2004), going back to Newton et al. (2001) and Efron et al. (2001); we assume that the N cases are divided into two classes, Null ("Uninteresting") or Non-Null ("Interesting"), occurring with prior probability p_0 or $p_1 = 1 - p_0$, and with the density of test statistic z depending on its class,

$$\begin{array}{ll} p_0 = Pr\{\text{Null}\}, & f_0(z) \text{ density if Null} \\ p_1 = Pr\{\text{Non-Null}\}, & f_1(z) \text{ density if Non-Null.} \end{array} \tag{1.2}$$

Defining the sub-densities

$$f_0^+(z) = p_0 f_0(z) \quad , \quad f_1^+(z) = p_1 f_1(z) \tag{1.3}$$

and the mixture density

$$f(z) = p_0 f_0(z) + p_1 f_1(z) = f_0^+(z) + f_1^+(z), \tag{1.4}$$

Bayes theorem gives posterior probability

$$\begin{aligned} fdr(z) &\equiv Pr\{\text{Null}|z\} = p_0 f_0(z)/f(z) \\ &= f_0^+(z)/f(z). \end{aligned} \tag{1.5}$$

Here $fdr(z)$ is the *local false discovery rate*, a localized version of Benjamini and Hochberg's 1995 tail-area False Discovery Rate, reviewed in Section 2. Selection proceeds by means of an empirical Bayes argument: $f_0^+(z)$ and $f(z)$ are estimated from the empirical distribution of the N z-values, giving an estimate of $fdr(z)$; cases with small fdr values are reported back to the scientific investigators as Non-Null, or perhaps Significant, or Interesting, the last terminology being appropriate for the situation where large-scale inference is preliminary to further, more detailed, investigation.

Figure 1 concerns a typical large-scale inference problem: eight microarrays, four from cells of HIV infected subjects and four from uninfected subjects' cells, 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-score

$$z_i = \Phi^{-1}(F_6(t_i)), \tag{1.6}$$

where F_6 is the cumulative distribution function (cdf) of a standard t-variable with 6 degrees of freedom, and Φ is the cdf of a standard normal variable. Theoretically, z_i should have a $N(0,1)$ distribution if gene$_i$ produces identically distributed normal expressions for infected and uninfected cells. Estimates of p_0, $f_0(z)$, (1.2), and $f(z)$, (1.4), were obtained from the histogram of the 7680 z-scores, using methodology discussed in Section 3. The empirical null is noticeably narrower than $N(0,1)$.

The estimated fdr curve (1.5) for the HIV data appears in Figure 2; 71 genes on the left, those with $z_i \le -2.34$, have $\text{fdr}(z_i) \le 0.2$, and these might be reported back as being significantly *underexpressed* in the HIV cells; likewise 115 genes with $z_i \ge 2.17$ are *overexpressed* according to a 0.2 fdr cutoff. (Remark B of Section 7 discusses the somewhat arbitrary 0.2 cutoff.)

The dashed curve in Figure 2 concerns the Estimation Problem. Here the Selection Model (1.2) has been augmented by a *structural model* for the z-scores,

$$z_i \sim N(\mu_i, \sigma_0^2), \tag{1.7}$$

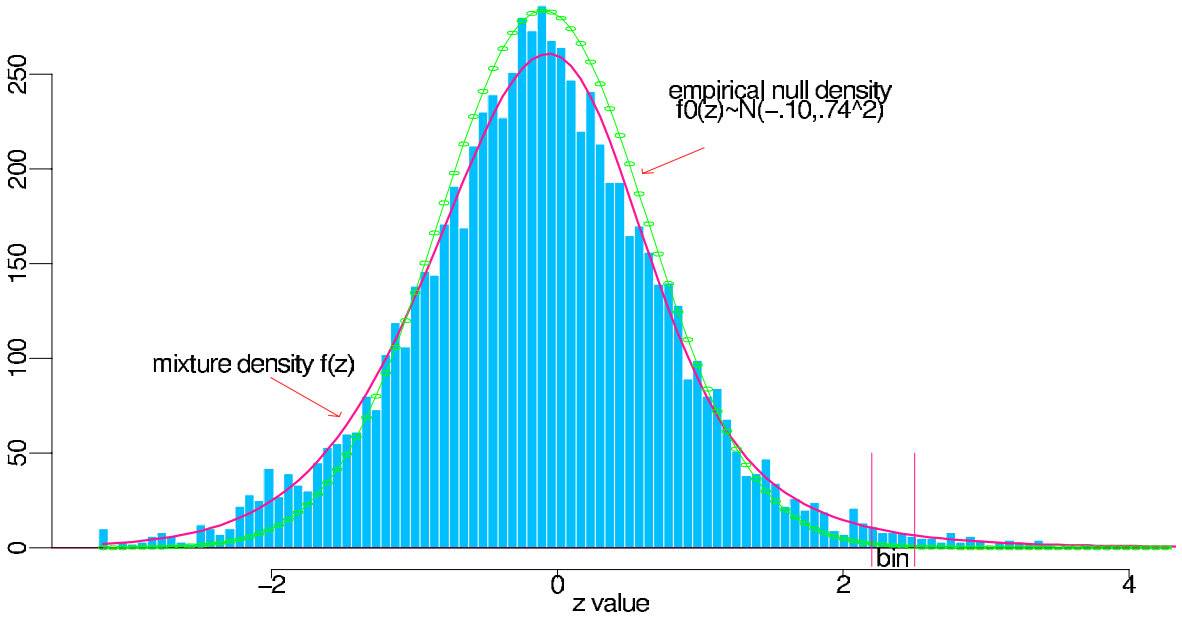

Figure 1: *Histogram of 7680 z-values from an HIV microarray experiment. Solid Curve estimates mixture density $f(z)$, (1.4); beaded curve estimates null density $f_0(z)$, (1.2). Here the empirical null $N(-.10, .74^2)$ is much narrower than the theoretical $N(0, 1)$ null. Proportion p_0 of null genes estimated as 0.917. Data from van't Wout et al. (2003), discussed in Gottardo et al. (2004). Note that f and f_0 are rescaled to each have total sum 7680.*

with σ_0 equaling the empirical null standard deviation 0.74. The unobservable "true scores" μ_i are assumed to be randomly generated in two varieties, Null (uninterestingly small in magnitude) or Non-Null, and, as in (1.2),

$$\begin{aligned} p_0 &= Pr\{\text{Null}\}, \qquad & g_0(\mu) \text{ density if Null} \\ p_1 &= Pr\{\text{Non-Null}\}, \qquad & g_1(\mu) \text{ density if Non-Null.} \end{aligned} \tag{1.8}$$

Model (1.8) implies (1.2) by convolution, with $f_0 = g_0 * N(0, \sigma_0^2)$ and $f_1 = g_1 * N(0, \sigma_0^2)$.

We cannot observe individual μ_i's, but it is possible to estimate, at least roughly, their overall distribution. The dashed curve in Figure 2 estimates the Non-Null component $g_1(\mu)$, using the methodology of Section 5. Proportion $p_1 = .083$ of the genes, 637 of them, are estimated to come from $g_1(\mu)$ in (1.8), twice as many to the left of zero as to the right. However the two modes of g_1 correspond to fdr values exceeding 0.4, uncomfortably large for identifying significant genes. The situation is somewhat frustrating: we know that there are many more "Interesting" genes than just the 186 identified in Figure 2, but they cannot be reported without including an unacceptably large proportion of false discoveries. Moreover

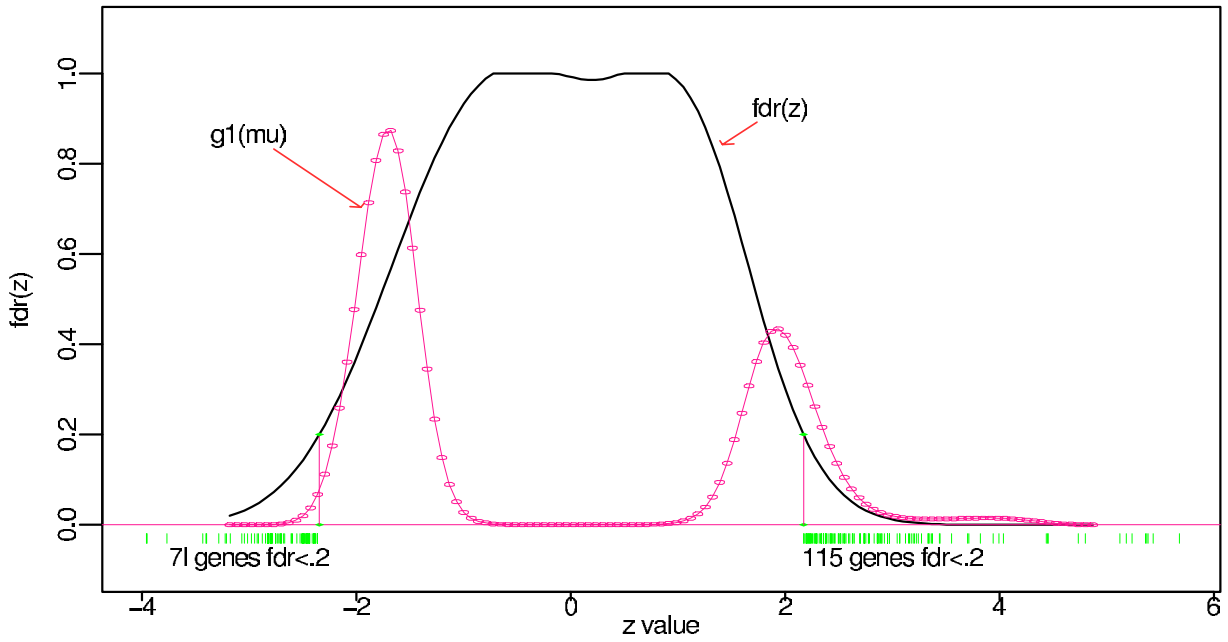

Figure 2: *Solid curve is fdr(z) (1.5), estimated from p_0, f_0 and f in Figure 1; 71 genes on the left ($z_i \leq -2.34$) and 115 genes on the right ($z_i \geq 2.17$) have fdr(z_i) ≤ 0.20. beaded curve is the estimated effect size density $g_1(\mu)$ for the non-null cases, discussed in Section 5.*

if the whole experiment were run again, roughly the same small number, 186, would be identified as having fdr < 0.2, but not the same set of genes identified here. These kinds of "power calculations" are taken up in Section 6.

Ideally, a big data set like that for 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 Selection calculations of Figures 1 and 2. By focusing attention on the gene-wise summary statistics z_i we avoid modeling *inside* the full 7680×8 data matrix. In fact there are some curious effects inside this matrix, discussed in Remark C of Section 7, causing the strangely narrow Null density, but the empirical estimation of $f(z)$ and $f_0(z)$ avoids having to model these effects.

The Estimation Problem is inherently more model dependent than Selection. Nevertheless the development here, again focusing on the z-scores, aims to minimize off-the-shelf mathematical modeling. There will certainly be situations where modeling inside the data matrix, as in Newton et al. (2004) and Gottardo et al. (2004), yields more information for both Selection and Estimation. Using such methods requires more careful attention to the details of the individual data set than the relatively crude z-score approach favored here. A reasonable compromise uses within-matrix modeling to improve the z-scores, for example by

improved normalization of the individual expression levels as in Dudoit et al. (2002), and then proceeds as in this paper.

Dudoit et al. (2003) emphasize a strict approach to the Selection problem in which error probabilities, in particular Family-Wise Error Rates (FWER), are carefully controlled. The fdr methodology featured in this paper is more relaxed, emphasizing a graded evaluation for each case, from very likely Null to very likely Interesting, rather than a strict yes or no. However our major inferential points could be made in the FWER framework just as well as with false discovery rates.

The two-group models (1.2) and (1.8) strongly suggest comparison of Treatment with Control, as in classical hypothesis testing. In practice, though, the test statistics z_i in (1.1) can arise from more general data structures. For example, each case's data might be a linear regression, with z_i the z-value corresponding to the slope parameter. Section 4 discusses false discovery rates in terms of a one-group structural model. This helps clarify the role of the empirical null density seen in Figure 1.

The paper concludes with remarks and a brief summary in Sections 7 and 8.

2. False Discovery Rates Local false discovery rates, Efron and Tibshirani (2002), used for much of the development here, are a version of Benjamini and Hochberg's (1995) "tail area" false discovery rates. This section relates the two ideas, while Section 3 discusses the practicalities of their estimation in situations like that of Figure 1.

Returning to Bayesian model (1.2), let $F_0(z)$ and $F_1(z)$ be the cdf's corresponding to $f_0(z)$ and $f_1(z)$, and likewise define

$$F_0^+(z) = p_0 F_0(z),\ F_1^+(z) = p_1 F_1(z) \text{ and } F(z) = F_0^+ + F_1^+(z). \tag{2.1}$$

Suppose we observe that z_i lies to the left of some prechosen point "z". Then Bayes theorem yields the posterior probability that null hypothesis H_i is true given $z_i \le z$,

$$\text{Fdr}(z) \equiv Pr\{\text{Null}|z_i \le z\} = F_0^+(z)/F(z). \tag{2.2}$$

The notation "Fdr" indicates the tail-area false discovery rate, as distinct from the local rate "fdr" in (1.5). (It is notationally convenient to consider events $z_{(i)} \le z$ but we could just as well consider tail areas to the right, or more general definitions). Benjamini and Hochberg's original FDR rule depends on an estimated version of (2.2), $F_0^+(z)/\bar{F}(z)$ where F_0^+ is based on the theoretical null and $\bar{F}$ is the usual empirical cdf. Storey (2002) and

Efron and Tibshirani (2002) discuss the connection between the Bayesian form (2.2) and the original frequentist definition.

Fdr and fdr are related according to (1.5) and (2.2) by

$$\begin{aligned} \mathrm{Fdr}(z) &= \textstyle\int_{-\infty}^{z} \mathrm{fdr}(z')f(z')dz' \big/ \int_{-\infty}^{z} f(z')dz' \\ &= E_f\{\mathrm{fdr}(z')|z' \le z\}, \end{aligned} \tag{2.3}$$

"E_f" indicating expectations with respect to $f(z)$. 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 $\mathrm{fdr}(-2.34) = 0.20$ while $\mathrm{Fdr}(-2.34) = 0.12$ in Figure 2. This says the obvious, that false discoveries are more likely near the boundary of a rejection region. An advantage of fdr is its local nature, which yields more relevant inferences for the individual cases.

Figure 3 illustrates the Fdr/fdr relationship in terms of the geometry of the curve $(F(z), F_0^+(z))$; $\mathrm{fdr}(z)$ is the tangent slope to the curve, while $\mathrm{Fdr}(z)$ is its secant slope. In the "Lehmann Alternative" situation where $F_0^+(z) = F(z)^\gamma$ for some power $\gamma > 1$, we have

$$\mathrm{fdr}(z) = \gamma \cdot \mathrm{Fdr}(z). \tag{2.4}$$

Remark A of Section 7 gives a version of (2.4) relating $\mathrm{fdr}(z)$ to Storey's (2002) "q-values".

The basic false discovery rate idea is appealingly simple. Consider the bin $z \in [2.20, 2.35]$ indicated in Figure 1; 41 of the 7680 HIV z-scores fell into this bin. This compares with expected number 6.12 under the null density $N(-.10, .74^2)$, which reduces to 5.62 assuming that p_0 in (1.2) equals its estimated value 0.917. Then the estimated local false discovery rate is

$$5.62/41 = 0.14. \tag{2.5}$$

If we report this bin as containing Interesting cases then about one seventh of them will turn out to be false discoveries. This is a frequentist one seventh; from a Bayesian point of view, *each* of the 41 cases will have one-seventh posterior probability of being Null.

Estimated in this way, local false discovery rates tend to be conservative, i.e. biased upwards, Efron and Tibshirani (2002). Benjamini and Hochberg's important 1995 algorithm for controlling Fdr's depends on this conservative bias, see also Storey, Taylor and Siegmund (2004). False discovery rates are expectations, and as such are fundamentally easier to control than probabilities. Lehman and Romano (2004) present generalizations of FWER that are more appropriate for large-scale inference.

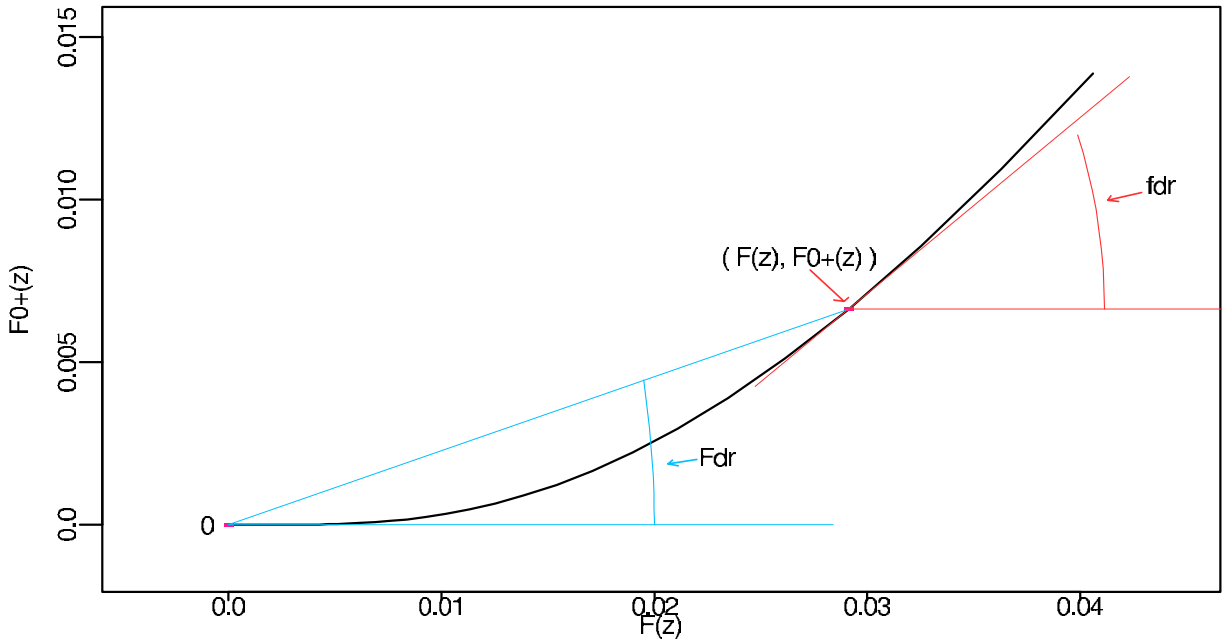

Figure 3: *Geometrical relationship of Fdr to fdr; heavy curve plots $F_0^+(z)$ versus $F(z)$; $fdr(z)$ is slope of tangent, $Fdr(z)$ slope of secant.*

Strict control is not emphasized in this paper, which is more concerned with making reasonably accurate inferences despite the range of difficulties that seem to afflict large-scale studies such as the HIV experiment. Nevertheless the expectation nature of fdr's plays an important role; in particular it means that calculations like (2.4) do not require independence across cases. We expect measurements to be correlated across genes in a microarray study for example, and would not want our methods to depend on knowing the correlation structure.

The question of dependence brings up an important limitation of statements like (1.5) and (2.2). These are really "one-at-a-time" Bayes results: if we knew the complete probability structure of the entire N-vector $\mathbf{z}$ we might find that $Pr\{H_i \text{ is true}|\mathbf{z}\}$ was much different than (1.5). Full Bayes modeling, as in Gottardo et al. (2004), provides full Bayesian answers, at the expense of substantial modeling assumptions. Notice that one-at-a-time calculations are not limited to expectations. In Section 5 we will extend results like (1.5) to include posterior means and variances given observation z_i.

3. Estimating fdr The heavy curve in Figure 2 is an estimate of $\text{fdr}(z)$ as defined by (1.5), carried out using the algorithm "locfdr" (available under that name from the Comprehensive R Archive Network: http://cran.r-project.org.) This Section concerns key points in the calculations, accuracy of estimation, comparison with Fdr estimates, and the

choice of the null density $f_0(z)$.

Mixture density $f(z)$, (1.4), the denominator of fdr(z), is estimated directly from the empirical distribution of the z-values. The solid curve in Figure 1 is a Poisson regression fit to the histogram counts of the z-scores, in 119 small bins spanning most of their range; locfdr allows two choices of the regression function, a natural spline basis or a polynomial, with seven degrees of freedom as the default option.

The polynomial option amounts to supposing that $f(z)$ has a seven-parameter exponential family of densities,

$$f(z) = \exp\left\{\sum_{j=0}^{7} \beta_j z^j\right\}, \tag{3.1}$$

with the β_j's estimated by maximum likelihood, assuming independence across cases. (The constant β_0 is determined from $(\beta_1, \beta_2, \cdots, \beta_7)$ by the requirement that $f(z)$ integrates to 1.) "Lindsey's method", discretization followed by Poisson regression, allows the use of standard glm software to solve the MLE equations, as discussed in Section 2 of Efron and Tibshirani (1996). Discretization is convenient but not necessary, see Remark G. As usual with regression methods, failure of the independence assumption does not bias the estimation of $f(z)$, but it does undercut glm assessments of variance.

The fdr numerator $p_0 f_0(z) = f_0^+(z)$ is more challenging to estimate. First of all, model (1.2) is unidentifiable without some further assumption. Full Bayesian specifications of $f_0(z)$ and $f_1(z)$, as in Kendziorski et al. (2003) are one way to go. Here, instead, it is assumed that all z_i values close to zero come from $f_0(z)$, the Null component. Various forms of the "zero assumption" have been proposed to estimate p_0, for example in Efron and Tibshirani (2002), Storey (2002), and Storey, Taylor and Siegmund (2004). Its use here relates to the estimation of the null density f_0 as well as p_0, or, more precisely, to $f_0^+ = p_0 f_0$, the numerator of (1.5).

Figure 4 illustrates null density estimation for the HIV data. The solid curve is log $\widehat{f}(z)$, using a natural spline estimate with seven degrees of freedom for $f(z)$, while the beaded curve estimates log $f_0^+(z) =$ log $(p_0 f_0(z))$ as a quadratic fit to log $\widehat{f}(z)$ near $z = 0$. The three coefficients of the quadratic fit completely determine the fdr numerator,

$$\widehat{f}_0^+(z) = .917 \cdot \varphi_{-10,.74}(z) \quad \left[\varphi_{\delta,\sigma}(z) = e^{-\frac{1}{2}\left(\frac{z-\delta}{\sigma}\right)^2} \Big/ \sqrt{2\pi\sigma^2}\right], \tag{3.2}$$

giving estimates of both p_0 and $f_0(z)$ at once.

The logic here is quite simple: we assume that the central peak of Figure 1's histogram

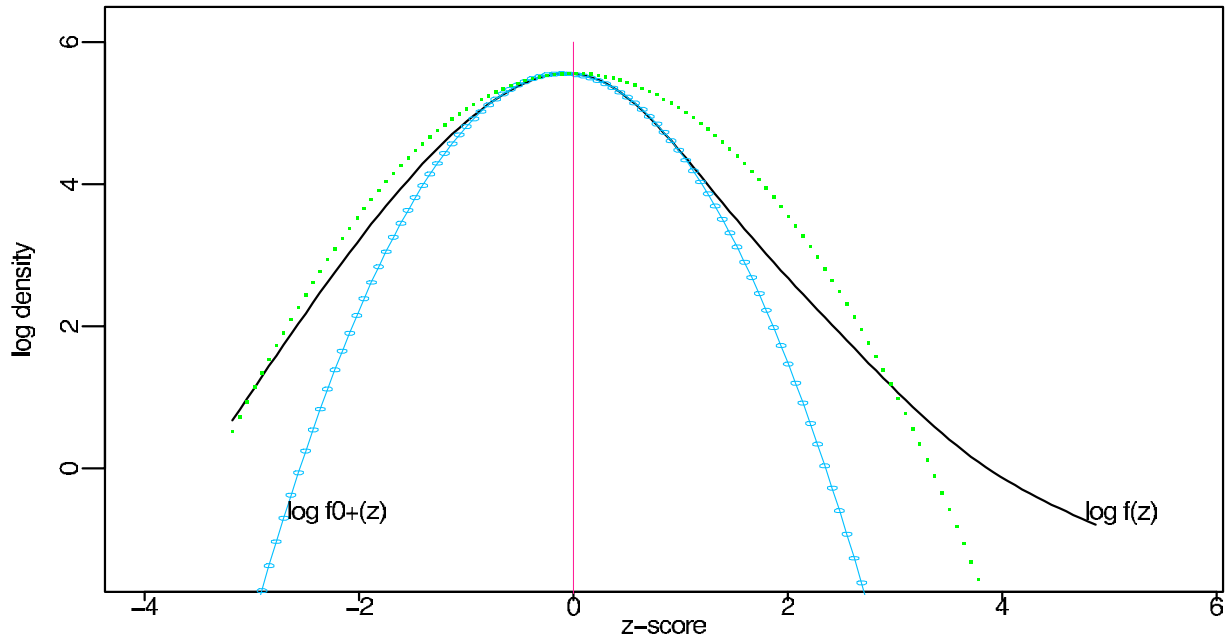

Figure 4: *Estimating the Null density for HIV data. Solid curve is log of* $\widehat{f}(z)$*, smooth fit to the histogram counts in Figure 1; beaded curve, a quadratic fit to log* $\widehat{(}z)$ *near* $z = 0$*, gives estimate of* $f_0^+(z) = p_0 f_0(z)$*. Dotted curve indicates theoretical null estimate of* $f_0^+(z)$*.*

consists mainly of Null cases, so p_0 and $f_0(z)$ are chosen to quadratically approximate the histogram counts near $z = 0$. This same argument can be used with fitting methods other than quadratic. The dotted curve in Figure 4 relates to the "theoretical null": it is assumed that $f_0(z)$ is $N(0, 1)$ so that only p_0 remains to be selected to match the histogram heights near $z = 0$. Section 4 gives a more general analysis of such procedures.

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 quadratic fitting method of Figure 4 will have negligible bias. That is, although the 10% or less of non-Null cases might in fact contribute some counts near $z = 0$, these cannot substantially affect estimates like (3.2). The development in Section 4 says more about the zero assumption, and its important relationship to the assessment of Interesting versus Null cases.

The theoretical null hypothesis $f_0 \sim N(0, 1)$ is untenable for the HIV data. If it were valid than $f(z)$ should be at least as wide as f_0 near $z = 0$, assuming that Non-Null z's tend to more dispersed than Nulls. Instead $f(z)$ is substantially narrower, forcing p_0 to have the impossible value $p_0 = 1.15$ 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 the theoretical null. Various causes of overdispersion are suggested,

including hidden correlations and unobserved covariates. The underdispersion observed here in Figure 1 is harder to explain, but see Remark C. Theoretical null densities are dangerous to use in large-scale inference problems, if for no other reason than that they can be negated by the data. Using an empirical null avoids these problems, and in situations like Figure 1 seems necessary.

There is, however, a substantial price to pay in terms of estimation accuracy. Figure 5 reports on a simulation experiment where data was generated from structural model (1.7), (1.8), with $N = 3000$ cases, independent z_i's, $\sigma_0^2 = 1$, and

$$p_0 = 0.90, \quad g_0(\mu) = \text{delta function at } \mu = 0, \ g_1(\mu) \sim N(-3, 1), \tag{3.3}$$

This is a situation where the theoretical $N(0, 1)$ null is appropriate. Standard errors of $\log \widehat{\text{fdr}}(z)$ were calculated from 100 simulations of (3.3), using the theoretical and empirical estimates of $f_0^+(z)$. In the crucial region near $z = -3$, the empirical method more than doubles standard errors.

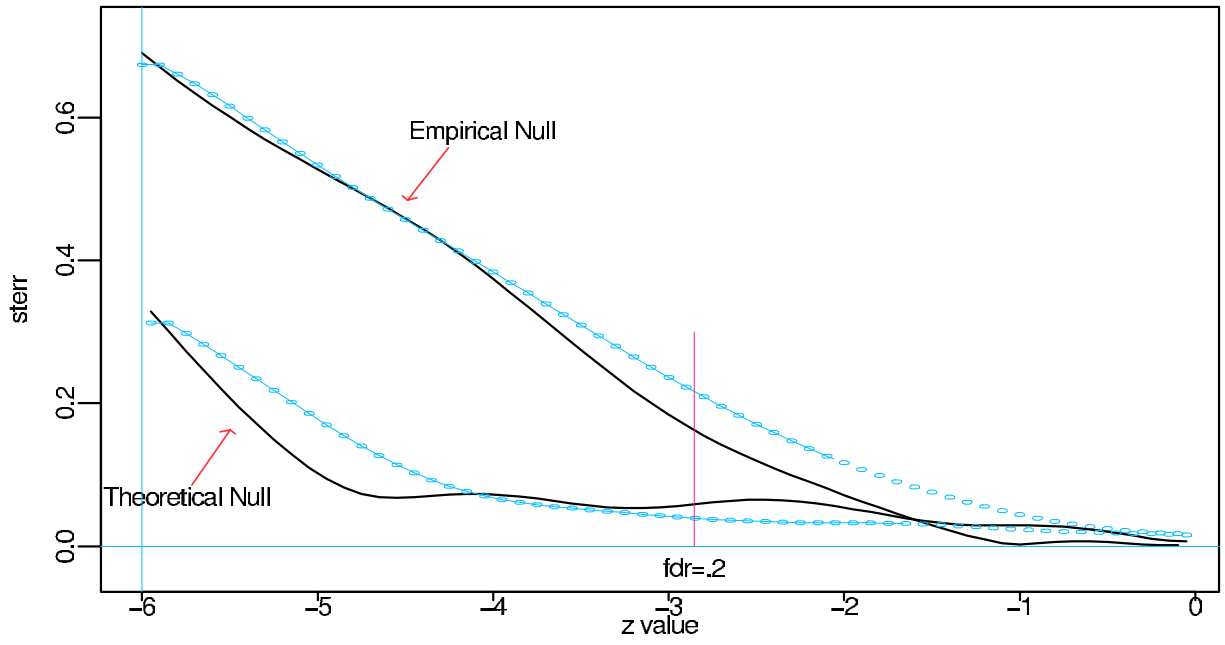

Figure 5: *Simulation comparison of standard errors for situation (3.3). Heavy curves sterrs of* $\log \widehat{\text{fdr}}(z)$*, empirical or theoretical nulls; using the empirical null more than doubles standard errors. Beaded curves sterrs of* $\log \widehat{\text{Fdr}}(z)$*, tail area false discovery rates.*

One might expect the tail-area Fdr (2.2) to be easier to estimate than fdr since it does not involve densities. This is contradicted by the dashed curves in Figure 5. Here $\widehat{\text{Fdr}}(z) = \widehat{F}_0^+/\widehat{F}(z)$ was calculated by numerically integrating the corresponding estimates

$\widehat{f}_0^+(z)$ and $\widehat{f}(z)$ used for $\widehat{\text{fdr}}(z)$. Surprisingly, log ($\widehat{\text{Fdr}}$) has about the same variability as log ($\widehat{\text{fdr}}$). Remark H confirms the simulation results with a delta-method theoretical comparison.

Other methods may be available for estimating an empirical null density, involving "housekeeper genes" (cases knowing *a priori* to be Null), designed replications, or perhaps an estimation technique different than that of Figure 4. In fact, the amount of variability of $\widehat{\text{fdr}}$ seen in Figure 5, even using the empirical null, would not be ruinous for practical applications: the standard error of the point where $\widehat{\text{fdr}}$ equaled 0.20 was only about 0.09 in the simulations. This is effectively negligible compared to the sampling variability of z_i for any one case, as discussed in Remark H.

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 $N(0, .99^2)$. Remark D suggests a method for combining permutation techniques with empirical null estimation.

4. One-Group Models The Selection Model (1.2) 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 (the μ_i in (1.7) for instance) ranging smoothly from zero or near zero to very large. This Section considers a "one-group" structural model that allows for smooth effects. We can still apply false discovery rate methods to data from one-group models, and doing so helps clarify the relationship between theoretical and empirical null hypotheses.

Consider the structural model

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim N(\mu, 1); \tag{4.1}$$

where each μ_i is drawn randomly according to a density $g(\mu)$, and then z_i is normally distributed around μ_i. (A generalization of (4.1) is suggested in Section 5.) Here we imagine that the number of cases N has gone to infinity, so that we can do "population" calculations for $f(z), p_0, \text{fdr}(z)$ etc., as in Section 2. The density $g(\mu)$ is allowed to have discrete atoms. It might, in particular, have an atom at zero, but this is not required, and in any case there is no *a priori* partition of $g(\mu)$ into Null and Non-Null components as in (1.8).

Model (4.7) gives

$$f(z) = \int_{-\infty}^{\infty} \varphi(\mu - z) g(\mu) d\mu \quad [\varphi(x) \equiv e^{-\frac{1}{2}x^2}/\sqrt{2\pi}] \tag{4.2}$$

for the marginal density of z, with value at $z = 0$

$$f(0) = \int_{-\infty}^{\infty} \varphi(\mu) g(\mu) d\mu. \tag{4.3}$$

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

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

$$\ell_J(z) = \sum_{j=0}^{J} \ell^{(j)}(0) z^j / j!, \tag{4.4}$$

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

$$\ell^{(j)}(0) = \frac{d^j \log\ f(z)}{dz^j}|_{z=0}. \tag{4.5}$$

The sub-density

$$f_0^+(z) = e^{\ell_J(z)} \tag{4.6}$$

matches $f(z)$ at $z = 0$ (a convenient use of the zero assumption) and leads to an fdr expression as in (1.5),

$$\text{fdr}(z) = e^{\ell_J(z)} / f(z). \tag{4.7}$$

Larger choices of J match $f_0^+(z)$ more accurately to $f(z)$, increasing ratio (4.7); the Interesting z-values, those with smaller 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 (4.1) leads to a helpful interpretation of the derivatives $\ell^{(j)}(0)$:

Lemma The derivative $\ell^{(j)}(0)$, (4.5), 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

$$\ell^{(1)}(0) = E_0 \quad \text{and} \quad -\ell^{(2)}(0) = \bar{V}_0, \tag{4.8}$$

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

$$\begin{aligned} \ell(z) &= \log \textstyle\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(\mu - z)^2}}{\sqrt{2\pi}} g(\mu) d\mu \\ &= -\tfrac{1}{2} z^2 + \log\ f(0) + \log \textstyle\int_{-\infty}^{\infty} e^{z\mu} [\varphi(\mu) g(\mu) / f(0)] d\mu. \end{aligned} \tag{4.9}$$

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 $\varphi(\mu)g(\mu)/f(0)$,

$$\frac{d^j m(z)}{dz^j}|_{z=0} = \int_{-\infty}^{\infty} \mu^j \frac{\varphi(\mu)g(\mu)}{f(0)} d\mu, \tag{4.10}$$

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 (4.7), (4.8) 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 1 (with p_0 obtained through definition (1.3),

$$p_0 = \Big[\int_{-\infty}^{\infty} f_0^+(z)dz \Big]^{-1}. \ ) \tag{4.11}$$

Formulas are also available for $\mathrm{Fdr}(z)$, (2.3), see Remark F. 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.

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

Table 1: Expressions for p_0, f_0 and fdr first three choices of J in (4.6), (4.7); numerator of $\mathrm{fdr}(z)$ is $f_0^+(z)$. $J = 0$ gives theoretical null, $J = 2$ empirical null.

Here is a simple example, borrowed from Johnstone and Silverman (2004), illustrating the relationships in Table 1. We assume that $g(\mu)$ in (4.1) is a mixture of a normal and a uniform density, each symmetric about 0,

$$g(\mu) = \begin{cases} N(0, A) \text{ with probability } \pi_0 \\ U(-a, a) \text{ with probability } \pi_1 = 1 - \pi_0. \end{cases} \tag{4.12}$$

Denoting

$$\Phi[u,v] = \Phi(v) - \Phi(u) \quad [\Phi(x) \text{ the standard normal cdf}], \tag{4.13}$$

and letting $A_1 = A + 1$, it is easy to calculate

$$f(z) = \frac{\pi_0}{\sqrt{2\pi A_1}}\, e^{-\frac{1}{2}\frac{z^2}{A_1}} + \pi_1 \frac{\Phi[-a-z, a-z]}{2a}, \quad f(0) = \frac{\pi_0}{\sqrt{2\pi A_1}} + \frac{\pi_1 \Phi[-a,a]}{2a}, \tag{4.14}$$

and

$$E_0 = 0, \quad \bar{V}_0 = \frac{\pi_0/\sqrt{2\pi a^3} + \pi_1 \varphi(a)}{\pi_0/\sqrt{2\pi A_1} + \pi_1 \Phi[-a,a]/2a}. \tag{4.15}$$

Table 2 shows p_0 and the standard deviation $\sigma_0 = \bar{V}_0^{-\frac{1}{2}}$ for the empirical null $N(0, \sigma_0^2)$, and also $p_{0,\text{theo}} = f(0)\sqrt{2\pi}$ for the theoretical null, for various combinations of (a, A) in (4.12), $\pi_0 = 0.90$. For the case $A = 0$, where $g(\mu)$ has atom 0.90 at zero, there are only small differences between the empirical and theoretical nulls; (The fact that σ_0 does not increase much above 1 is predicted by the theorem in Efron (2004).) Both p_0 and $p_{0,\text{theo}}$ exceed $\pi_0 = 0.90$, notably so for small "a". This happens because the uniform component in (4.12) contributes some probability at $z = 0$, affecting the zero-matching assumption. The assumption is really a definition that resolves the unidentifiability of $f_0(z)$ in (1.2).

For $A = 0$:				For $a = 5$:				
a	σ_0	p_0	$p_{0,\text{theo}}$	**A**	σ_0	$(\sqrt{A+1})$	p_0	$p_{0,\text{theo}}$
3	1.02	0.96	0.94	**0.0**	1.01	(1.00)	0.94	0.93
5	1.01	0.94	0.93	**0.4**	1.20	(1.18)	0.94	0.79
10	1.01	0.92	0.91	**0.8**	1.37	(1.34)	0.95	0.70
15	1.00	0.91	0.91	**1.2**	1.51	(1.48)	0.96	0.63
20	1.00	0.91	0.91	**1.6**	1.65	(1.61)	0.96	0.58
25	1.00	0.91	0.91	**2.0**	1.77	(1.73)	0.97	0.54

Table 2: $\sigma_0 = \bar{V}_0^{-\frac{1}{2}}$, the standard deviation of the empirical null, for various combinations of a and A in model (4.12), $\pi_0 = 0.90$; also p_0 for empirical null, and $p_{0,\text{theo}}$ for theoretical null.

The right panel of Table 2 shows A increasing from 0 to 2, with $a = 5$ fixed. Now σ_0 closely follows the standard deviation of the z-scores coming from the dominant $N(0, A)$ component in (4.12), $z \sim N(0, A+1)$. Note that $p_{0,\text{theo}}$ declines precipitously, to 0.54 for

$A = 2$. If we believe Selection model (1.2), then $f_0(z)$ cannot be the theoretical null $N(0,1)$ unless there is a very large proportion of Non-Null cases. Whether or not this is reasonable depends on scientific context: it *is* in Johnstone and Silverman's "haystack" setting, but *not* for most microarray experiments.

The empirical and theoretical nulls judge "Significance" in different ways. Suppose we take $(a, A) = (20, 4)$ in model (4.1), (4.12), $\pi_0 = .90$, and choose $\text{fdr}(z) \leq 0.2$ as our significance threshold. Using the empirical null, we calculate that cases with $|z| \geq 7.54$ will be reported as Significant. This means that almost all (99.9%) of the cases from the $N(0, A)$ component in (4.12) will be declared Null. If instead we use the theoretical $N(0,1)$ null to compute fdr's, then $|z| \geq 2.00$ is the cutoff, and more than one-third of the $N(0, A)$ cases will be declared Significant. These are cases that look unusually big compared to the theoretical standard deviation 1, but not compared to the empirical standard deviation $\sigma_0 = 2.25$.

Simply put, the empirical null approach judges significance of the extremes by the spread of the center. Emphasis is on comparative rather than absolute significance, suggesting a preference for the "Interesting" terminology rather than "Significant" or "Non-Null". Scientific context determines the appropriate choice for the null density, but in situations like the HIV experiment, an observational study where the investigator hopes to identify a relatively small number of cases for further study, the empirical approach seems preferable.

Suppose $\pi_0 = 1$ in (4.12), so all of the cases draw μ from a $N(0, A)$ distribution. The empirical null gives $\text{fdr}(z) = 1$ for all z in this situation, indicating *no* Interesting cases at all. This is the correct interpretation if the cause of the nonzero μ_i's is itself uninteresting, for example unobserved covariates in an observational study or failed assumptions as in Remark C. On the other hand we might believe that the μ_i's are interesting effects connected with the individual cases, perhaps true gene effects that would persist in future studies. In this situation *all* the cases are Non-Null. The hypotheses testing formulation (1.1) is inappropriate now, and we would prefer a pure estimation approach, with $\widehat{\mu}_i = (A/(A+1))z_i$ as the Bayes estimate of μ_i. Using the $N(0,1)$ theoretical null in model (1.1) gives misleading results either way.

The choices $J = 0, 1, 2$ in Table 1 all 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 (4.7) and the Lemma,

$$\text{fdr}(z) = \frac{f(0)}{f(z)}\, e^{E_0 z - \bar{V}_0 z^2/2 + S_0 z^3/6}, \tag{4.16}$$

where S_0 is the posterior third center moment of μ given $z = 0$ in model (4.1). The program

locfdr uses a variant of (4.16), the "split normal", to model asymmetric null densities, Remark E.

5. Estimation The beaded curve in Figure 2 is an estimated density for the true effects in the HIV study. This Section concerns what the curve means and how it is computed. Various kinds of power calculations based on the effect curve are discussed in Section 6.

We begin with a slightly generalized version of the one-group model (4.1),

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim N(\mu, \sigma_0^2), \tag{5.1}$$

σ_0^2 fixed. (Later, σ^2 will be allowed to vary with μ.) This determines the mixture density by convolution, $f(z) = g \star N(0, \sigma_0^2)$, which in turn provides $f_0^+(z) = p_0 f_0(z)$ as in (4.4)-(4.6). The numerical examples below use the empirical null, $J = 2$, as in Table 1. By subtraction in (1.4) this yields the Non-Null, (or "Interesting") density

$$f_1^+(z) = p_1 f_1(z) = f(z) - f_0^+(z) \tag{5.2}$$

We estimated $p_1 = .083$ for the HIV data, indicating 637 genes in the Interesting class. Each of these genes has a true effect μ_i in model (5.1). The beaded curve in Figure 2 is an estimate of the density "$g_1(\mu)$" for the Interesting cases. It is like the density $g_1(\mu)$ in (1.8), except here we do not begin by assuming a two-group model for the μ's. Instead, $g_1(\mu)$ will be obtained by inverting $f_1^+(z)$ in (5.3).

Suppose first that we wished to recover $g(\mu)$ in (5.1) from $f(z) = g \star N(0, \sigma_0^2)$. This kind of inverse problem can be notoriously difficult, at least in the situations we will be interested in where f must be estimated from the observed z_i's. Brown (1971) and Stein (1981) provided a useful formula for the posterior mean function $E\{\mu|z\}$, which can easily be extended to higher moments:

Lemma (Brown and Stein) Under model (5.1), μ given z has posterior mean and variance

$$\mu|z \sim (z + \sigma_0^2 \ell'(z),\ \sigma_0^2(1 + \sigma_0^2 \ell''(z))), \tag{5.3}$$

where $\ell(z) = \log f(z)$, and ℓ', ℓ'' indicate first and second derivatives.

The estimation option in the *locfdr* program begins by applying (5.3) to $f_1^+(z)$ – actually to $\widehat{f}_1^+(z)$, but we will ignore this distinction for notational simplicity. Let

$$\ell_1(z) = \log f_1(z) = \log f_1^+(z) + \log p_1, \tag{5.4}$$

and define

$$a(z) = z + \sigma_0^2 \ell_1'(z) \quad \text{and} \quad b(z) = \sigma_0[1 + \sigma_0^2 \ell_1''(z)]^{\frac{1}{2}}, \tag{5.5}$$

with σ_0 the standard deviation of the empirical null. (Note that it makes no difference to the derivatives whether ℓ_1 equals $\log f_1$ or $\log f_1^+$.) Then, using notation (3.2),

$$g_1^{(0)}(\mu) = \int_{-\infty}^{\infty} f_1(z) \varphi_{a(z),b(z)}(\mu) dz \tag{5.6}$$

is a reasonable first guess for $g_1(\mu)$. Notice that (5.6) adds a normality assumption to the relationships in (5.3). A third-moment relationship, in terms of ℓ_1''', can be appended to (5.3), leading to a skewed version of (5.6), but this seemed unstable in applications.

Ideally we might hope that the convolution

$$f_1^{(0)} = g_1^{(0)} \star N(0, \sigma_0^2) \tag{5.7}$$

would equal f_1, but this is unlikely since usually neither f_1^+ nor f_1 themselves will be of convolution form (5.7). Instead we can try to update $g_1^{(0)}$ to $g_1^{(1)}$ so that $f_1^{(1)} = g_1^{(1)} \star N(0, \sigma_0^2)$ minimizes a measure of discrepancy with f_1, say

$$D(f_1, f_1^{(1)}) = \int_{-\infty}^{\infty} \frac{[f_1(z) - f_1^{(1)}(z)]^2}{f_1(z)} \, dz. \tag{5.8}$$

Here is a Gibbs-type updating scheme: beginning with $(g_1^{(0)}, f_1^{(0)})$ as in (5.6), (5.7), let $g_1^{(0)}(\mu|z) = g_1^{(0)}(\mu)\varphi_{\mu,\sigma_0}(z)/f_1^{(0)}(z)$; then update to

$$g_1^{(1)}(\mu) = \int_{-\infty}^{\infty} g_1^{(0)}(\mu|z) f_1(z) dz \quad \text{and} \quad f_1^{(1)} = g_1^{(1)} \star N(0, \sigma_0^2). \tag{5.9}$$

The *locfdr* algorithm, beginning with estimates of $f_1, g_1^{(0)}$, and $f_1^{(0)}$, iterates (5.9) several times, which is how the effect density in Figure 2 was computed. Remark J describes the connection between minimizing (5.8) and iterating (5.9),

Model (5.1) can be generalized to

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim N(\mu, \sigma_\mu^2), \tag{5.10}$$

now allowing the variance of z to depend up on its mean. We might expect $z_i = \Phi^{-1}(F_6(t_i))$ in (1.6) to follow (5.10): the diagnostic theory of Efron (1982) shows that the non-central student-t family closely approximates a "Normal Scaled Transformation Family" $z \sim N(\mu, \sigma_\mu^2)$, with σ_μ^2 an even function of μ; σ_μ^2 is well-approximated by a cubic function of μ^2, declining from $\sigma_0^2 = 1$ to $\sigma_4^2 = 0.6$ when t has 6 degrees of freedom.

The Brown-Stein Lemma does not directly apply to situation (5.10). Letting $\mu(z) = E\{\mu|z\}$, an obvious approximation to (5.3) is

$$\mu|z \doteq (z + \sigma^2_{\mu(z)}\ell'(z),\ \sigma^2_{\mu(z)}[1 + \sigma^2_{\mu(z)}\ell''(z)]), \tag{5.11}$$

which must be solved iteratively since $\mu(z)$ figures on both sides. A more careful analysis changes $z + \sigma^2_{\mu(z)}\ell'(z)$ to $z + V(z) + \sigma^2_{\mu(z)}\ell'(z)$, where $V(z) = \text{var}\{\mu|z\}$. Replacing (5.1) with (5.10) did not noticeably change the estimated effect density curve in Figure 2.

6. Power Calculations The two curves appearing in Figure 2, $\widehat{\text{fdr}}(z)$ for the false discovery rate and $\widehat{g}_1(\mu)$ for the effect size density, can be combined to assess power, "power" referring to an experiment's ability to correctly identify individual Interesting cases. This Section describes a few convenient power diagnostics.

A particularly simple diagnostic is "Efdr", the estimated expectation of $\text{fdr}(z)$ for z's from the Interesting class. Defining

$$\widetilde{f}_1(z) = \int_{-\infty}^{\infty} \widehat{g}_1(\mu)\varphi_{\mu,\sigma_0}(z)d\mu, \tag{6.1}$$

σ_0 the empirical null standard deviation, gives expectation

$$\text{Efdr} \equiv \int_{-\infty}^{\infty} \widehat{\text{fdr}}(z)\widetilde{f}_1(z)dz. \tag{6.2}$$

A large value of Efdr suggests poor power, with typical Interesting cases not likely to show up on a list of low fdr cases. For example the HIV study has Efdr = 0.496, reinforcing Figure 2's impression of low power. 50 simulations of model (4.1) with $g(\mu)$ as in (3.3) gave Efdr = $0.26 \pm .04$ (mean $\pm$ standard deviation); changing $g_1(\mu)$ to $N(-4, 1)$ gave Efdr = $0.16 \pm .03$, now indicating a situation with substantial power.

For a more specific assessment of the HIV experiment we can compute

$$\int_{-\infty}^{0} \Phi\left(\frac{z_0 - \mu}{\sigma_0}\right)\widehat{g}_1(\mu)d\mu \Big/ \int_{-\infty}^{0} \widehat{g}_1(\mu)d\mu = 0.210, \tag{6.3}$$

with $z_0 = -2.34$ and $\sigma_0 = .74$, the estimated probability that a gene in the left mode of the Interesting class appears on the list of genes having $\widehat{\text{fdr}}$ less than 0.2. The corresponding number on the right is a healthier 0.43.

We can imagine running an independent duplicate of the HIV experiment and reporting the list of genes with $\widehat{\text{fdr}} \le 0.2$. What proportion of this list would overlap with the original

set indicated in Figure 2? An estimate of the overlap proportion for the left mode is

$$\int_{-\infty}^{\infty} \Phi\left(\frac{z_0-\mu}{\sigma_0}\right)^2 \widehat{g}_1(\mu)d\mu \Big/ \int_{-\infty}^{\infty} \Phi\left(\frac{z_0-\mu}{\sigma_0}\right) \widehat{g}_1(\mu)d\mu = 0.26, \tag{6.4}$$

so that 74% of the new list would be novel. The equivalent of (6.4) for the right mode is 0.53. These are rough estimates: they ignore the sampling variability in $\widehat{g}_1$, z_0, and σ_0, and assume that the structural model (5.1) holds perfectly, with σ_0^2 equal to the empirical null variance. They are useful nonetheless, and no cruder than most power calculations for standard situations.

The results of a large-scale study can be difficult to convey, particularly the uncertainty in reported outcomes for cases of special interest to the investigators. Bayesian calculations can help describe the statistical variability. The Selection and Estimation algorithms of Sections 3 and 5 provide an estimated Bayes prior $\widehat{g}(\mu)$ having two components as in (1.7), (1.8):

$$g_0 = \text{delta function at } \delta_0, \quad \text{and} \quad g_1 = \widehat{g}_1, \tag{6.5}$$

with $p_0 = \widehat{p}_0$ and δ_0 the mean of the empirical null, $\delta_0 = -0.10$ in Figure 1.

Having observed $z_i \sim N(\mu_i, \sigma_0^2)$, we can use Bayes theorem to assess the distribution of "Z_i", a hypothetical independent replication from $N(\mu_i, \sigma_0^2)$. Dropping the subscript i, the posterior density $h(Z|z)$ is estimated to be

$$h(Z|z) = \widehat{\text{fdr}}(z)\varphi_{\delta_0,\sigma_0}(Z) + [1 - \widehat{\text{fdr}}(z)]\; c(Z,z) \int_{-\infty}^{\infty} \widehat{g}_1(\mu)\varphi_{\bar{z},\sigma_0/\sqrt{2}}\;(\mu)d\mu, \tag{6.6}$$

where

$$c(Z,z) = \varphi_{z,\sqrt{2}\;\sigma_0}(Z) \Big/ \int_{-\infty}^{\infty} \widehat{g}_1(\mu)\varphi_{\mu,\sigma_0}(z)dz \quad \text{and} \quad \bar{z} = \frac{(Z+z)}{2}, \tag{6.7}$$

in notation (3.2), Formula (6.6) uses the fact that $\widehat{\text{fdr}}(z)$ approximates $\text{Prob}\{\mu = \delta_0|z\}$.

Table 3 shows $\text{Prob}\{Z \le -2.34|z\}$ and $\text{Prob}\{Z \ge 2.17|z\}$ for various values of z in the HIV experiment. For example, a gene with $z_i = 1.79$ has $\widehat{\text{fdr}}(z_i) = 0.5$, not small enough to be on the Interesting list, but there is a 19.2% chance that a hypothetical replication Z_i would have $\widehat{\text{fdr}}(Z_i)$ less than 0.2. Conversely, $z_i = -2.68$ gives $\widehat{\text{fdr}}(z_i) = 0.1$, but only 26.1% probability for $\widehat{\text{fdr}}(Z_i) \le 0.2$. A gene had to be "lucky as well as good" to show up as Interesting in the HIV experiment.

Model (5.1), (6.5) allows us to estimate the benefits of running a bigger experiment than the one at hand. In the HIV study we can imagine having "c" times as many microarrays

fdr	z	Prob	z	Prob
0.1	-2.68	0.261	2.47	0.409
0.2	-2.34	0.218	2.23	0.340
0.3	-2.15	0.183	2.06	0.286
0.4	-1.98	0.152	1.92	0.237
0.5	-1.83	0.123	1.79	0.192

Table 3: Prob$\{Z < z_0|z\}$ for $z_0 = -2.34$ (Left), and also Prob$\{Z > z_0|z\}$ for $z_0 = 2.17$ (Right); for values of z corresponding to fdr$(z) = .1, .2, \ldots, .5$ in Figure 2.

in each group. (Increasing the number of genes per microarray might improve the accuracy of the estimates in Figure 2, but would not improve power for individual genes.) This could be roughly modeled as $z_i \sim N(\mu_i, \sigma_0^2/c)$ in (5.1), or, rescaling by factor $\sqrt{c}$ to regain the original variance, $z_i \sim N(\sqrt{c}\ \mu_i, \sigma_0^2)$. Then (6.5) becomes

$$g_0 = \text{delta function at } \sqrt{c}\ \delta_0 \quad \text{and} \quad g_1 = \widehat{g}_1(\mu/\sqrt{c})/\sqrt{c}. \tag{6.8}$$

c:	1	2	3	4
Efdr	.521	.263	.145	.085

Table 4: Theoretical calculation of Efdr, (6.2) if the HIV experiment were expanded by factor c.

We could simulate from (5.1), (6.8) to assess the improvement in diagnostics like Efdr, (6.2), but that is unnecessary. Theoretical calculations as in (4.2)-(4.7) are more efficient. Table 4 shows the decline in Efdr if the HIV experiment were expanded by factor c. (The theoretical value for $c = 1$ differs slightly from the actual Efdr = 0.496 because model (4.1) does not hold exactly.) We see that doubling the experiment, from 8 to 16 microarrays, would considerably improve its power.

7. Remarks The following Remarks concern points raised in the previous Sections.

A. *fdr, Fdr, and q-values* A more familiar version of the Lehmann alternative model leading to (2.4) assumes that

$$F_0(z) = F_1(z)^\gamma \tag{7.1}$$

for some $\gamma > 1$. The population "q-value" corresponding to z, using Storey's (2002) terminology, is $\mathrm{Fdr}(z)$ in (2.2). Then (7.1) gives

$$\mathrm{logit}(fdr) = \mathrm{logit}(Fdr) + \log(\gamma) \qquad \left[\mathrm{logit}(p) \equiv \log\frac{p}{1-p}\right], \tag{7.2}$$

which reduces to (2.4) as $\mathrm{fdr}(z)$ and $\mathrm{Fdr}(z)$ go to zero.

B. *fdr 0.2 cutoff* The cutoff point $\mathrm{fdr}(z) \le 0.2$ used in Figure 2 gives posterior odds ratio

$$\begin{aligned}&\mathrm{Prob}\{\text{Non-Null}|z\}/\mathrm{Prob}\{Null|z\} = (1-\mathrm{fdr}(z))/\mathrm{fdr}(z)\\ &= p_1 f_1(z)/p_0 f_0(z) \ge 0.8/0.2 = 4.\end{aligned} \tag{7.3}$$

We have been assuming prior odds ratio $p_1/p_0 \le 0.1/0.9 = 1/9$, so (7.3) corresponds to Bayes factor

$$f_1(z)/f_0(z) \ge 36 \tag{7.4}$$

in favor of "Non-Null". This represents a much stronger level of evidence against the Null Hypothesis than in standard one-at-a-time testing. For example, suppose we observe $z \sim N(\mu, 1)$ and wish to test $H_0 : \mu = 0$ vs $\mu = 2.80$, a familiar scenario for power calculations since rejecting H_0 for $z \ge 1.96$ yields two-sided size 0.05 and power 0.80. The critical Bayes factor in this case is only $f_{2.80}(1.98)/f_0(1.96) = 4.8$. We might justify (7.4) as being conservative in guarding against multiple testing fallacies. More pragmatically, the usual purpose of large-scale testing is to winnow the number of possible cases down to a small set ripe for further investigation, while one-at-a-time testing usually hopes to reject the Null.

C. *HIV Null underdispersion* The empirical null standard deviation $\sigma_0 = 0.74$ is substantially smaller than the theoretical value 1.00 we would expect from (1.6), a fact confirmed by other dispersion estimates such as the interquartile range. Efron (2004) suggests reasons for *over*dispersion of the empirical null, particularly unobserved covariates in an observational study, but it is less clear how underdispersion might arise.

Let Y be the 7680 by 8 matrix of expression scores "y_{ij}" for the HIV study, with the first 4 columns representing HIV subjects and the last 4 normal controls. One possible cause of underdispersion is long tails for the y_{ij}'s, Chung (1946), but this was not evident; moreover, independently permuting entries within each column of Y gave z-values (1.6) almost perfectly $N(0, 1)$, which would not be true for long-tailed y_{ij}'s. Replacing the two-sample t-tests with Wilcoxon tests still gave underdispersion, so parametric assumptions are not the problem.

Principal components analysis indicated another possibility. The row means were subtracted from each entry of Y, to remove gene effects, and the 8 by 8 principal components matrix computed. The first principal vector turned out to be

$$(.19, -.37, .45, -.37, .25, -.29, .48, -.34), \tag{7.5}$$

showing a strong alternation pattern.

This suggests a pattern of correlation across the rows of Y, with block $(y_{i1}, y_{i3}, y_{i5}, y_{i7})$ positively intracorrelated, likewise block $(y_{i2}, y_{i4}, y_{i6}, y_{i8})$, and negative correlations across blocks. Such a pattern tends to increase the denominators of the t-statistics in (1.6), shrinking the z_i's. As a check the t-statistics were recomputed, this time comparing columns 1, 3, 5, 7 with columns 2, 4, 6, 8. The empirical standard deviation σ_0 was now much *bigger* than 1.00, $\sigma_0 = 1.53$, as predicted by the correlation theory.

Just why gene expression levels might be correlated across microarrays is a matter of speculation, but the same trouble occurred in Efron (2004) (to opposite effect there, causing overdispersion of the empirical null), a warning against casual assumptions of independence in microarray studies.

D. *Permutations and the empirical null* In our example the empirical null was tacitly justified by a two-step argument: (1) $z_i = \Phi^{-1}(F_6(t_i))$ in (1.6) has a $N(0,1)$ theoretical null distribution under the usual parametric assumptions; (2) we accept normality for the empirical null, but use data from the center of the distribution to fit the null mean and standard deviation. Suppose though that instead of $t_i = \text{num}_i/\text{den}_i$, where num and den are the usual numerator and denominator of a two-sample t statistic, we follow the suggestion in Efron et al. (2001) by adding constant "a_0" to the denominator,

$$t_i = \text{num}_i/(\text{den}_i + a_0), \tag{7.6}$$

a_0 the median of all 7680 denominators. Now it is not clear what transformation $z_i = m(t_i)$ should play the role of $\Phi^{-1}(F_6)$ in step 1.

Permutation methods can help guide the choice of $m(\cdot)$. Permuting the entries of Y independently within columns gave (7.6) a distribution approximately Student-t with 32 degrees of freedom, suggesting that (1.6) be replaced by $z_i = \Phi^{-1}(F_{32}(t_i))$. (The more familiar approach of permuting whole columns of Y produced erratic results, probably because of the correlation structure discussed above.) The empirical null is then estimated just as in Figure 4. There is a division of labor here: permutation methods suggest the form of the

null distribution, in particular its tail heaviness, and then empirical fitting yields its mean and variance.

E. *Split normal empirical null* The default option in program *locfdr* fits the estimate of $\log f_0^+(z)$, the beaded curve in Figure 4, by minimizing

$$\int_{q_1}^{q_2} \{\log \widehat{f}(x) - [b_0 + b_1 z + b_2 z^2]\}^2 dz \tag{7.7}$$

over the choice of (b_0, b_1, b_2), where q_1 and q_2 are the 1/3 and 2/3 quantiles of the z_i's. We could add a term $b_3 z^3$ to (7.7) to accommodate an asymmetric null, as in (4.16), but the cubic term can cause trouble when (z) gets large. Instead, the asymmetry option in *locfdr* uses

$$b_0 + b_1 z + b_2 z^2 + b_3 [\max(z - z_{\max}, 0)]^2, \tag{7.8}$$

where $z_{\max}$ maximizes $\widehat{\ell}(z)$. This amounts to fitting $\log f_0^+(z)$ with a quadratic spline; $\widehat{f}_0(z)$ has "split normal" form,

$$\widehat{f}_0(z) = \begin{cases} c e^{-\frac{1}{2}\left(\frac{z - z_{\max}}{\sigma_1}\right)^2} & z < z_{\max} \\ & \text{for} \\ c e^{-\frac{1}{2}\left(\frac{z - z_{\max}}{\sigma_2}\right)^2} & z \geq z_{\max}, \end{cases} \tag{7.9}$$

$c = [\sqrt{2\pi}\ (\sigma_1 + \sigma_2)/2]^{-1}$.

Suppose that the HIV study had 3 treatment groups of 4 microarray each, rather than 2 groups, so that instead of t_i the summary statistic for gene i was F_i from an F test with 2 and 9 degrees of freedom. Defining $z_i = \Phi^{-1}(F_{2,9}(F_i))$, $F_{2,9}$ the F distribution cdf, we could proceed as before, but we may suspect that now the Non-Null z_i's all lie on the positive side of zero. Using (7.9) effectively eliminates the influence of negative z_i's on assessing significance for the positive cases, which might seem appropriate, especially if the z histogram shows pronounced central asymmetry.

F. *Fdr(z) formula* The one-group model (4.1) also provides formulas for the tail-area False Discovery Rate $\mathrm{Fdr}(z)$, (2.2). For the $J = 2$ case in Table 1,

$$\mathrm{Fdr}(z) = p_0 \Phi(\bar{V}_0^{\frac{1}{2}} z - E_0/\bar{V}_0^{\frac{1}{2}}) \Big/ \int_{-\infty}^{\infty} \Phi(z - \mu) g(\mu) d\mu, \tag{7.10}$$

with $p_0 = f(0)\sqrt{2\pi/\bar{V}_0^{\frac{1}{2}}}\ \exp\{E_0^2/2\bar{V}_0\}$.

G. *Exponential family modeling* Exponential family models such as (3.1) can be elaborated to handle situations with more structure than (1.1). Suppose the N cases are partitioned into M subgroups, perhaps representing different genetic pathways in the HIV example. Of course we can apply *locfdr* separately to each subgroup, but this might invite estimation problems for the smaller groups. A more efficient model expands (3.1) to

$$\log\{f(z)\} = \sum_{j=0}^{7} \beta_j z^j + \gamma_{1m} z + \gamma_{2m} z^2, \tag{7.11}$$

with $\Sigma_m \gamma_{1m} = \Sigma_m \gamma_{2m} = 0$, effectively allowing different means and variances for each group while retaining common tail behavior. The estimates $(\widehat{\gamma}_{1m}, \widehat{\gamma}_{2m})$ can then be used to adjust $\widehat{f_0}$ for group m.

There is nothing inherently one-dimensional in the basic fdr setup (1.2)-(1.5). If the summary statistics z_i are two-dimensional, say $z_i = (u_i, v_i)$, (3.1) can be expanded to

$$\log\{f(z)\} = \sum_{j,k} \beta_{jk} w^j v^k, \tag{7.12}$$

the powers represented in (7.12) being a modeling decision. Efron and Tibshirani (2002) discuss a two-dimensional microarray example. The discretization necessary for Lindsey's method gets cumbersome in higher dimensions. Standard exponential family methods give the maximum likelihood estimates $\widehat{\beta}_{jk}$ directly, though it is no longer possible to use standard Poisson GLM software for the computations.

H. *Accuracy Estimates* Straightforward "delta method" calculations provide closed-form approximations for the standard errors obtained by simulation in Figure 5. These are based on discretization of the z-scores as in the histogram of Figure 1. Let $\mathbf{x}$ be the vector of centerpoints of the K histogram bins, $K = 119$ in Figure 1, and X the $K \times p$ structure matrix used in the Poisson glm estimation of $f(x)$. For the polynomial options (3.1), $X = (\mathbf{1}, \mathbf{x}, \mathbf{x}^2, \cdots, \mathbf{x}^7)$. Also let $X_0 = (\mathbf{1}, \mathbf{x}, \mathbf{x}^2)$, and define $\widetilde{X}$ and $\widetilde{X}_0$ as the $K_0 \times p$ and $K_0 \times 3$ submatrices confined to the central part of the histogram used in estimating the empirical null f_0, (the interval corresponding to the range of integration in (7.7)).

The delta-method estimate of covariance for $\log(\widehat{\mathbf{fdr}})$, where $\widehat{\mathbf{fdr}}$ is the K-vector of empirical null fdr estimates obtained from *locfdr*, is

$$\widehat{\mathrm{Cov}} = A\widehat{G}^{-1}A' \qquad [\widehat{G} = X^1 \widehat{V} X,\ \widehat{V} = \mathrm{diag}(N\widehat{\mathbf{f}}], \tag{7.13}$$

with

$$A = X_0(\widetilde{X}_0'\widetilde{X}_0)^{-1}\widetilde{X}_0'\widetilde{X} - X. \tag{7.14}$$

A similar expression is available for $\log(\widehat{\mathbf{Fdr}})$. Comparing the diagonal elements of the two covariance matrices confirms that $\widehat{\text{fdr}}$ and $\widehat{\text{Fdr}}$ are about equally variables.

Figure 6 illustrates the variability of $\widehat{\text{fdr}}(z)$ in 25 simulations of (3.3), as in Figure 5. Dashes indicate the points $z_{.2}$ having $\widehat{\text{fdr}}(z_{.2}) = 0.2$. These have mean and standard deviation (-2.96, 0.094), being biased downward from the true population value -2.91. Since z_i has distribution $N(\mu_i, 1)$ in (3.3), it is clear that the variability in $\widehat{\text{fdr}}(z_i)$ comes mainly from the variability of z_i, not of the curve $\widehat{\text{fdr}}(\cdot)$.

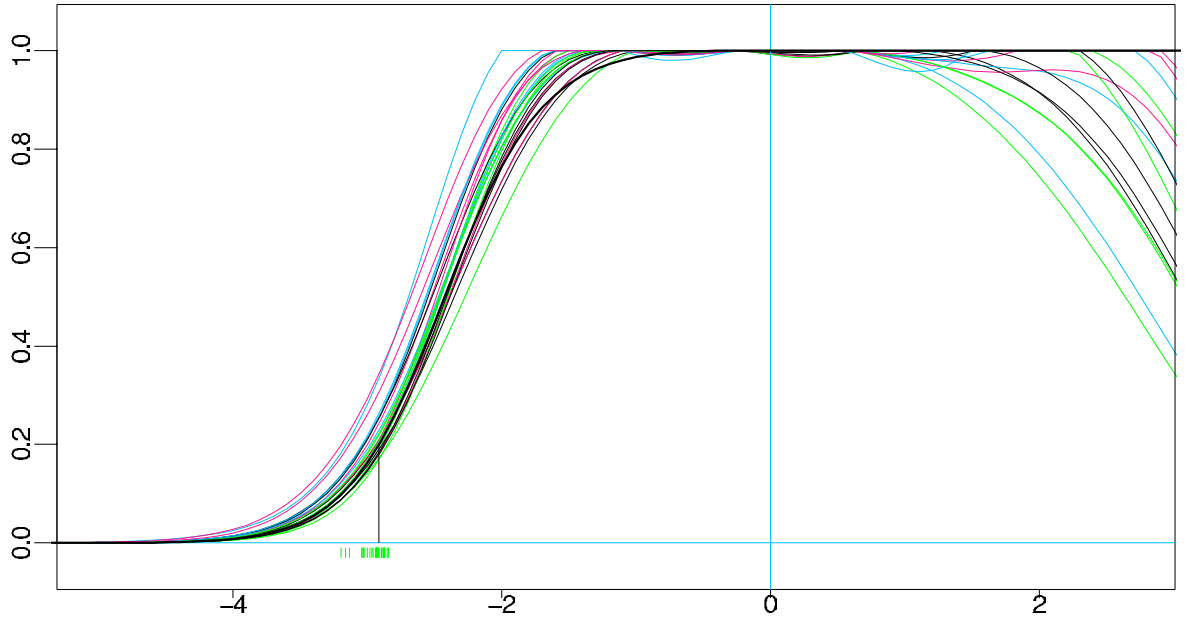

Figure 6: *Estimated curves* $\widehat{\text{fdr}}(z)$*, 25 simulations from (3.3) as in Figure 5. Dashes indicate values* $z_{.2}$ *having* $\widehat{\text{fdr}}(z_{.2}) = 0.2$*. Heavy line indicates true fdr curve (4.7),* $J = 2$*.*

I. *Independence and Accuracy* Formula (7.12) assumes independence among the z-scores. Independence manifests itself in $\widehat{V} = \text{diag}(N\widehat{\mathbf{f}})$, the diagonal matrix with entries $N\widehat{f}_k$, the smoothed estimate of expected number of z_i's in the kth histogram bin. (The estimate of $f(z)$ in Figure 1 is actually $N\widehat{\mathbf{f}}$, this being *locfdr's* convention.) $\widehat{V}$ is the approximate covariance matrix for the vector $\mathbf{y}$ of histogram counts, *assuming independence*.

Independence is sometimes believable, in the main example of Efron (2004) for instance, but more often not. In a microarray context we might intend instead try bootstrapping the *columns* of the expression matrix Y, i.e. with entire microarrays as the bootstrapping units. This gives dependable accuracy estimates for $\widehat{\text{fdr}}$, $\widehat{\text{Fdr}}$ etc. *if* the columns are independent. They are *not* independent in the HIV study, Remark C, nor the second example of Efron (2004), and independence seems a risky assumption in microarray studies.

Accuracy assessments are necessarily problematical when both the rows and columns of Y show dependence. A reasonable tactic is to try absorbing the dependencies into a resampling scheme. Table 5 shows jackknife estimates of standard error for components of the principal vector (7.5). The jackknifing was done removing genes in blocks of 384 from Y, 20 blocks in all, either with the blocks' memberships determined by the genes' ordering on the microarray, or randomly. Ordered assignment gives substantially bigger standard errors.

Ordered:	.022	.041	.033	.064	.026	.047	.027	.040
Random:	.013	.016	.036	.045	.021	.028	.025	.014

Table 5: Jackknife standard error estimates for the components of principal vector (7.5); jackknifing 20 blocks of 384 genes each. *Top* genes blocked in order of microarray listing; *Bottom* genes randomly selected.

Random assignment treats the genes as independent, which is overoptimistic here, while the original ordering captures at least some of the dependence structure. Using 10 ordered blocks of 768 genes each gene roughly the same results as the top row of Table 5, lending it credence, but of course we might still be missing important dependencies. Blocking is easier when the dependence structure is more obvious, for instance when the cases have a geometrical relationship as in fMRI brain studies.

J. *Gibbs-type Updating* The connection between (5.8) and (5.9) is easy to describe in the discrete setting of Remark H, where the vector of centerpoints $\mathbf{x}$ gives the possible values of both μ and z. Denote $g_i^{(1)} = g_1^{(1)}(x_i)$, $f_j^{(1)} = f_1^{(1)}(x_j)$, $f_j = f_1(x_j)$, and $m_{ij} = c_i\varphi_{0,\sigma_0}(x_j - x_i)$, with c_i chosen to make $\Sigma_j m_{ij} = 1$. Relationship (5.7) is $f_j^{(1)} = \Sigma_i g_i^{(1)} m_{ij}$, or $\partial f_j^{(1)}/\partial g_i^{(1)} = m_{ij}$.

We wish to choose $g_i^{(1)}$ values to

$$\text{minimize } \Sigma_j (f_j - f_j^{(1)})^2/f_j \quad \text{for} \quad f_j^{(1)} = \Sigma_i g_i^{(1)} m_{ij}, \tag{7.15}$$

which by differentiation gives the relationships

$$\Sigma_j m_{ij}[f_j^{(1)}/f_j - 1) = 0 \quad \text{for all } i. \tag{7.16}$$

On the other hand, the stable point for iteration (5.9) has $g_i^{(1)}$ values satisfying

$$g_i^{(1)} = \Sigma_j \left[\frac{g_i^{(1)} m_{ij}}{f_j^{(1)}}\right] f_j, \tag{7.17}$$

or

$$\Sigma_j m_{ij}[f_j/f_j^{(1)} - 1] = 0 \quad \text{for all } i. \tag{7.18}$$

8. Summary Recent literature on simultaneous hypothesis testing is heavily focussed on control of Type I errors, in classical terminology the *size* of a testing procedure. This paper also concerns Type II errors, the *power* of a procedure to correctly identify interesting cases amidst a large majority of uninteresting ones. The paper's main example, an HIV microarray study, turned out to be badly underpowered: a relatively small number of genes could be identified as Non-Null, but the methodology indicated a larger set of Interesting cases lost in the noise of the Null majority, most of which would have been found in an experiment two or three times as large.

"Selection and Estimation" in the paper's title refers to identifying Non-Null cases and then estimating their effect sizes, a necessary preliminary to power calculations. This is carried out using an empirical Bayes/false discovery rate algorithm, *locfdr*. Its results are not as precise as either a full Bayesian analysis or a strict frequentist control procedure for Type I errors. On the other hand, that precision is purchased at the cost of modeling assumptions that *locfdr* seeks to avoid or at least minimize.

One such assumption involves the choice of an appropriate null hypothesis. Theoretical null distributions, for instance student's t density in a two-sample comparison, are risky to use in large-scale studies, where the data itself may contradict them. The methodology here allows an empirical choice of the null, which is a different matter than using permutation or bootstrap techniques to improve upon a theoretical null distribution.

The data for the HIV study is an $N \times 8$ data matrix Y of expression scores, $N = 7680$. Row i of Y, that is the data for gene i, produces a test statistic z_i for the ith null hypothesis. *Locfdr* works directly with the N z-scores. This tactic risks losing information, but it also avoids modeling assumptions for the entries of Y. The HIV data set turned out to have some unexpected, and unpleasant, internal correlation structure, a problem that may be endemic to microarray studies.

Suppose we can assume that the z_i are normally distributed, say $z_i \sim N(\mu_i, 1)$. It is tempting to state the ith null hypothesis as $\mu_i = 0$, and take the goal of the analysis to be identification of the cases having μ_i not equal zero; as we would do in a single hypothesis test. Actual examples of large-scale inference are likely to be less clearcut, with a large majority of cases having small "Uninteresting" μ_i's, and the minority "Interesting" set having a range of bigger μ_i values. The one-group formulation of Section 4 examines Selection and Estimation

in such a framework, which helps clarify why and when the empirical null will be preferable to the theoretical null.

Large-scale testing involves more than corrections for simultaneous inference. Whether we want it or not, information accrues from the whole experiment, affecting the inference for any one case. This paper represents the accrued information with a simple empirical Bayes model, and then employs straightforward Bayesian calculations to quantify inferences for individual cases. The exact methodology used here is not crucial, but seems reasonably flexible and dependable so far in a small catalog of test problems.

References

Brown L. (1971) "Admissable estimators, recurrent diffusions, and insolvable boundary value problems" *Ann. Math. Stat.* **42** 855-903.

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.

Chung K (1946). "The approximate distribution of Student's t-statistic" *Ann. Math. Stat.* **17**, 447-465.

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

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

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

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

Efron B (2004) "Transformation theory: how normal is a family of distributions?" *Annals Stat.* **10** 323-339.

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

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.

Lehman E and Romano J (2004). "Generalizations of the familywise error rate" to appear *Annals Stat.*

Newton M, Kendzioroki 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.

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

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

Robbins H (1956) "An empirical Bayes approach to statistics" *Proc. Third Berkeley Symp* **1** 152-163, U. Cal Press.

Stein C (1981) "Estimation of the mean of a multivariate normal distribution" *Ann. Stat.* **9** 1135-1151.

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

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.

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.

van't Wout A, Lehrma G, Mikheeva S, O'Keeffe G, Katze M, Bumgarner R, Geiss G, and Mullins J (2003). "Cellular gene expression upon human immunodeficiency virus type 1 infection of CD+T-Cell lines", *Journal of Virology* **77**, 1392-1402.