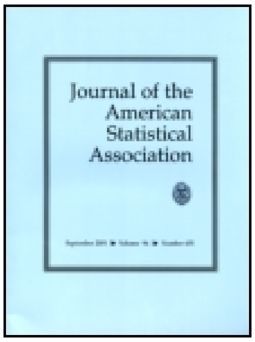

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Correlation and Large-Scale Simultaneous Significance Testing

Bradley Efron

To cite this article: Bradley Efron (2007) Correlation and Large-Scale Simultaneous Significance Testing, Journal of the American Statistical Association, 102:477, 93-103, DOI: 10.1198/016214506000001211

To link to this article: http://dx.doi.org/10.1198/016214506000001211

Correlation and Large-Scale Simultaneous Significance Testing

Bradley EFRON

Large-scale hypothesis testing problems, with hundreds or thousands of test statistics z_i to consider at once, have become familiar in current practice. Applications of popular analysis methods, such as false discovery rate techniques, do not require independence of the z_i's, but their accuracy can be compromised in high-correlation situations. This article presents computational and theoretical methods for assessing the size and effect of correlation in large-scale testing. A simple theory leads to the identification of a single omnibus measure of correlation for the z_i's order statistic. The theory relates to the correct choice of a null distribution for simultaneous significance testing and its effect on inference.

KEY WORDS: Correlated processes; Empirical null; False discovery rate; Microarray.

1. INTRODUCTION

Modern computing machinery and improved scientific equipment have combined to revolutionize experimentation in such fields as biology, medicine, genetics, and neuroscience. One effect on statistics has been to vastly magnify the scope of multiple hypothesis testing, now often involving thousands of cases considered simultaneously. The cases themselves are typically of familiar form, each perhaps a simple two-sample comparison, but with their test statistics correlated in some unknown fashion. This article concerns the effect of correlation on multiple testing procedures, particularly false discovery rate techniques (Benjamini and Hochberg 1995).

Test statistics from two microarray experiments are displayed in Figure 1. The experiments, described in Section 2, report two-sample t-statistics t_i comparing expression levels under two different conditions for N genes, $N = 3{,}226$ for the breast cancer study in Figure 1(a), and $N = 7{,}680$ for the human immunodeficiency virus (HIV) experiment in Figure 1(b); t_i tests the null hypothesis that gene i has the same expression distribution under both conditions. The t_i's have been converted to z-values for easy analysis later,

$$z_i = \Phi^{-1}(G_0(t_i)), \qquad i = 1, 2, \ldots, N, \tag{1}$$

where Φ is the standard normal cumulative distribution function (cdf) and G_0 is a putative null cdf for the t-values. G_0 was taken to be a standard Student t cdf with appropriate degrees of freedom for the HIV study, whereas a permutation method described in Section 4 provided G_0 for the breast cancer experiment (also nearly a Student t cdf). Assuming that G_0 is the correct null distribution for t_i, transformation (1) yields

$$z_i \sim \mathrm{N}(0, 1) \tag{2}$$

for the null cases, called the *theoretical null* in what follows. Form (2) for the null distribution is convenient for general discussion and can be achieved, or at least approximated, in most testing situations through transformations like (1).

Microarray experiments involving genome-wide scans usually presuppose most of the genes to be null, the goal being to identify a small subset of interesting nonnull genes for future study, so we expect N(0, 1) to fit the center of the z-value histogram. This is not the case in Figure 1, where N(0, 1) is too narrow for the breast cancer histogram and too wide for the HIV data.

This article concerns two related results:

(1) Correlation can cause effects like those shown in Figure 1, considerably widening or narrowing the distribution of the null z-values.
(2) These effects have a substantial impact on simultaneous significance testing and must be accounted for in deciding which cases should be reported as nonnull.

Sections 2 and 3 begin with a normal-theory analysis of z-value correlations. A surprisingly simple result emerges in which the main effect of all the pairwise correlations (several million of them for Fig. 1's examples) is summarized by a single dispersion variate, A: a positive value of A widens the central peak of the z-value histogram, even assuming that the theoretical null (2) is individually correct for all the cases, whereas negative A narrows it, as in Figure 1(b). The random variable A is identically 0 if the z_i's are independent, but correlation allows A to vary around 0 with positive variance α^2, as summarized by the theorem in Section 3, the more correlation, the bigger the α. Section 4 replaces normal theory with permutation methods, carried out in detail for the breast cancer data, showing nice agreement with the theory.

The effect of correlation on simultaneous inference, particularly in terms of false discovery rates (FDRs), is discussed in Section 5. Broadly speaking, a wide central histogram like that for the breast cancer data implies more null z-values in the tails, so that significance levels judged according to the theoretical N(0, 1) null are too liberal. Conversely, the theoretical null is too conservative for the HIV data. This provides some support for using an *empirical null* distribution, a normal curve fit to the central portion of the z-value histogram (Efron 2004, 2006). The light curves in Figure 1 are empirical nulls,

$$\begin{aligned} &\text{breast cancer:} \quad \mathrm{N}(-.09, 1.55^2); \\ &\text{HIV:} \quad \mathrm{N}(-.11, .75^2). \end{aligned} \tag{3}$$

Looking ahead to Section 5, Figure 2 illustrates the disturbing effects of correlation on large-scale simultaneous inference. The simulation model involved $N = 3{,}000$ genes, with 95%

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *brad@stat.stanford.edu*). The author is very grateful to Trevor Hastie and Rob Tibshirani for suggesting the applicability of permutation methods to correlation calculations. Support was provided by the National Institutes of Health (grant 8R01 EB002784) and the National Science Foundation (grant DMS-05-05673).

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000001211

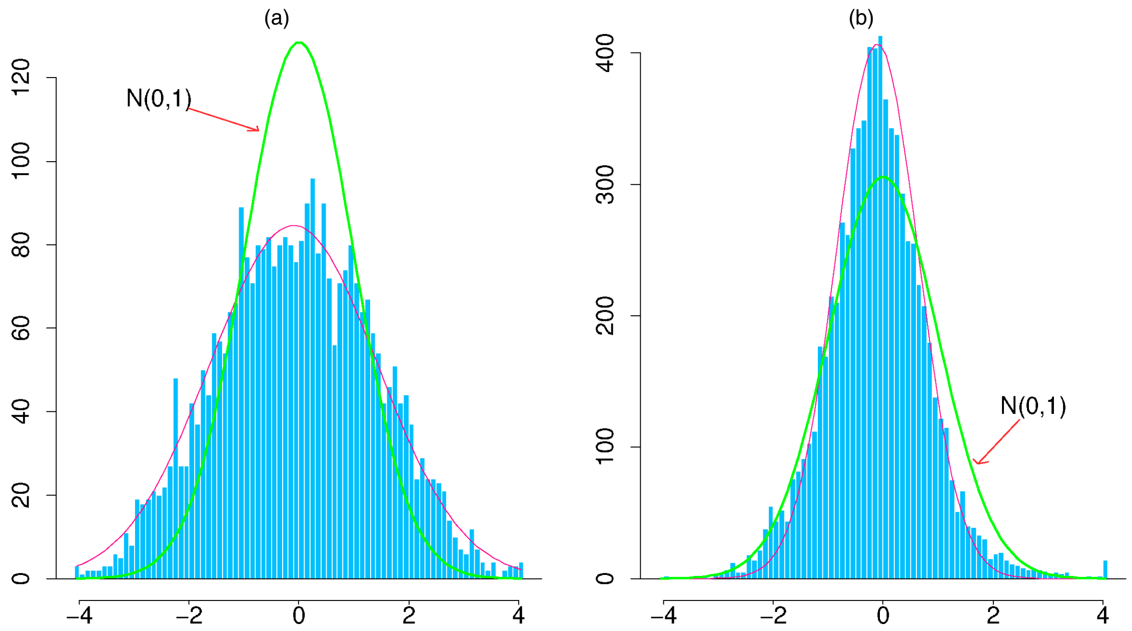

Figure 1. Histograms of z-Values From Two Microarray Experiments. (a) Breast cancer study, 3,226 genes. (b) HIV study, 7,680 genes. Heavy curves indicate N(0, 1) theoretical null densities; Light curves indicate empirical null densities fit to central z-values, as done by Efron (2004). The theoretical null distributions are too narrow in (a) and too wide in (b). Both effects can be caused by correlations among the null z-values. Data are from (a) Hedenfalk et al. (2001) and (b) van't Wout et al. (2003).

obeying the theoretical null distribution $z_i \sim \mathrm{N}(0, 1)$ and the other 5% nonnull, $z_i \sim \mathrm{N}(2.5, 1.25)$; the null z_i's were correlated about the same as in the breast cancer data, $\alpha = .15$.

For each of 1,000 replications, a standard Benjamini–Hochberg FDR analysis [i.e., based on the theoretical N(0, 1) null distribution] was run at control level $q = .10$ and used to identify a set of "significant" genes. Each of the 1,000 points in Figure 2 has coordinates

$$(A, \mathrm{FDP}), \tag{4}$$

where A is the dispersion variate and FDP is the true "false discovery proportion," the proportion of identified genes that actually came from the null class.

The overall mean FDP was .096, close to the theoretical control level .10, but we see a strong dependence on the dispersion variate A; FDP averaged .34 for the upper 5% of A realizations but only .03 for the lower 5%. The true FDP reported by the ".10" procedure easily varies by a factor of 10.

The trouble here stems from unconditional use of the theoretical null, not from the Benjamini–Hochberg procedure itself. Section 5 shows that A can be estimated from the width of the central peak of the z-value histogram; narrow peaks imply negative A's and wide peaks imply positive A's. In terms of Figure 2, this enables the statistician to *condition* the FDR estimate on its approximate location along the A axis. Replacing the theoretical null (2) with an empirical null (3) automatically carries out the conditioning process, although in a somewhat noisy fashion.

There is a lot at stake here. Table 1 gives the number of gene discoveries identified by the Benjamini–Hochberg two-sided procedure, FDR control level $q = .10$, for the two studies of Figure 1. The HIV results look much more dramatic using the empirical null distribution $\mathrm{N}(-.11, .75^2)$. In fact, a null standard deviation of $\sigma_0 = .75$ is quite believable given the amount of correlation, whereas Figure 1 argues against the theoretical

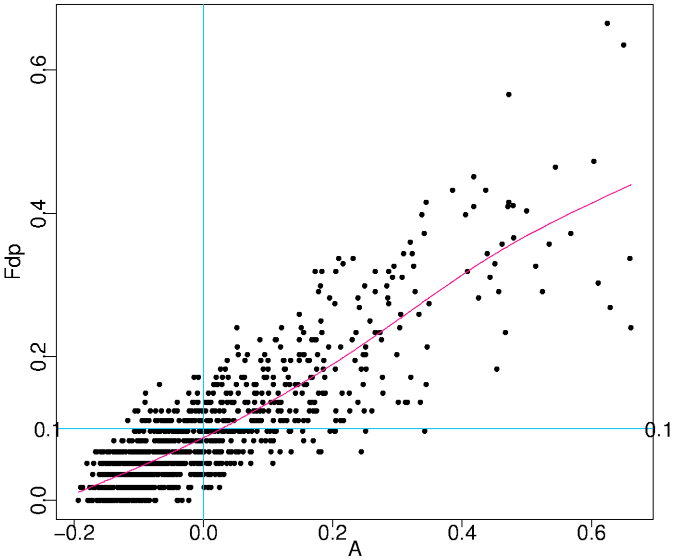

Figure 2. Benjamini–Hochberg FDR Controlling Procedure, $q = .10$, Run for 1,000 Simulation Trials; True False Discovery Proportion, FDP, Plotted versus Dispersion Variate A. Overall FDP averaged .096, close to q, but with a strong dependence on A, as shown by smooth regression curve.

Table 1. Number of Genes Identified as Significant Discoveries by Two-Sided Benjamini–Hochberg Procedure, .10 Control Level

	Breast cancer	*HIV*
Theoretical null	107	22
Empirical null	0	180

NOTE: The top row is based on theoretical N(0, 1) null distribution, and the bottom row is based on the empirical null distribution (3).

null. The breast cancer data have been used in the microarray literature to compare analysis techniques, under the presumption that better techniques will produce more discoveries (see, e.g., Pawitan, Murthy, Michiels, and Ploner 2005; Storey, Dai, and Leek 2005). Table 1 suggests caution in this interpretation, where using the empirical null negates any discoveries at all. In earlier work (Efron 2004) I gave reasons other than correlation for distrusting the theoretical null in large-scale testing problems, but only correlation effects are discussed here.

Five pertinent references are discussed in what follows: Qui, Klebanov, and Yakovlev (2005b), Qui, Brooks, Klebanov, and Yakovlev (2005a), Owen (2005), Dudoit, van der Laan, and Pollard (2004), and Hsu (1992, 1996). Permutations and correlated z-values also play a role in Westfall and Young's (1993) theory of adjusted p values, as well as in the work of Westfall (1997) and Ge, Dudoit, and Speed (2003), but with less direct bearing on the ideas here. A brief discussion and some remarks conclude the article.

2. CORRELATION EFFECTS ON THE NULL DISTRIBUTION

We begin with a normal-theory analysis for the effects of correlation on the null distribution of z-values. For these calculations, it is assumed that *all* cases are null,

$$z_i \sim \mathrm{N}(0,1) \quad \text{for } i = 1, 2, \ldots, N, \tag{5}$$

so that the theoretical N(0, 1) null distribution is individually correct. Nevertheless, it will turn out that correlation among the z_i's can make the null distribution effectively wider or narrower than N(0, 1), as in Figure 1. Section 5 shows that in real problems, where we hope to detect some nonnull cases, correlation effects can play a major role in their correct identification.

Here is a thumbnail description of the studies featured in Figure 1, along with some of the notation used in what follows. The breast cancer study compared gene activity in 15 patients observed to have one of two different genetic mutations known to increase breast cancer risk, BRCA1 or BRCA2 (Hedenfalk et al. 2001). It included seven BRCA1 and eight BRCA2 women, each providing a microarray of expression levels on the same $N = 3{,}226$ genes. The usual two-sample t-statistic, t_i, comparing BRCA2 and BRCA1 for the 15 gene i expression levels gave z_i as in (1), with G_0 nearly a Student t distribution with 13 degrees of freedom; see Section 4. Similarly, the HIV study compared four HIV-positive patients versus four HIV-negative controls, with $N = 7{,}680$ genes per microarray (van't Wout et al. 2003). In this case G_0 was taken to be Student t with 6 degrees of freedom. These datasets were discussed further in earlier work (Efron 2004, 2005).

Let X represent the full dataset, for example, an $N \times n$ matrix for the breast cancer study, with $N = 3{,}226$ rows corresponding to genes and $n = 15$ columns corresponding to microarrays. There each row of X yielded a t statistic t_i and then a z-value z_i as in (1), with $\mathbf{z}$ representing the vector of all N z_i's. Note that is not necessary that the z_i's be obtained from t-tests, only that null distribution (5) can be achieved or approximated. For example, each of the original N cases might involve a separate linear regression, with the ith case yielding p value p_i for some parameter of special interest, and $z_i = \Phi^{-1}(p_i)$.

It is helpful to work with histogram counts rather than with the vector of z-values itself. Each histogram in Figure 1 has its z-axis partitioned into $K = 82$ bins of width $\Delta = .1$, running from -4.1 to 4.1. We denote the count vector by $\mathbf{y}$,

$$y_k = \#\{z_i \text{ in } k\text{th bin}\}, \qquad k = 1, 2, \ldots, K; \tag{6}$$

$\mathbf{y}$ *is essentially the order statistic of* $\mathbf{z}$, exactly so if we let $\Delta \to 0$. FDR methods depend only on the ordered z-values.

The histogram counts y_k arise from a partition of $\mathcal{Z}$, the sample space for the z-values, into K bins,

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k, \tag{7}$$

each bin being of width Δ, with center point "$z[k]$." The following definitions lead to useful representations for the mean and covariance of $\mathbf{y}$,

$$\pi_k(i) = \Pr\{z_i \in \mathcal{Z}_k\}, \qquad \pi_{k\cdot} = \frac{\sum_{i=1}^{N} \pi_k(i)}{N} \tag{8}$$

and

$$\gamma_{k\ell}(i,j) = \Pr\{z_i \in \mathcal{Z}_k \text{ and } z_j \in \mathcal{Z}_\ell\}, \qquad \gamma_{k\ell\cdot} = \frac{\sum_{i \neq j} \gamma_{k\ell}(i,j)}{N(N-1)}. \tag{9}$$

Because of assumption (5), all of the $\pi_k(i)$ values are determined by $\varphi(z)$, the standard normal density, with Taylor approximation around centerpoint $z[k]$,

$$\pi_{k\cdot} = \pi_k(i) \doteq \Delta \cdot \varphi(z[k]) \qquad [\varphi(z) = e^{-1/2z^2}/\sqrt{2\pi}]. \tag{10}$$

The *expectation vector*, $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_K)'$, of $\mathbf{y}$ is determined by (5),

$$\boldsymbol{\nu} = N\boldsymbol{\pi}_{\cdot} \doteq \big(\ldots, N\Delta\varphi(z[k]), \ldots\big)'. \tag{11}$$

Definitions (8) and (9) lead to a convenient expression for the covariance matrix of $\mathbf{y}$.

Lemma 1.

$$\mathrm{cov}(\mathbf{y}) = C_0 + C_1, \tag{12}$$

where C_0 is the multinomial covariance matrix that would apply if the z-values were independent,

$$C_0 = \mathrm{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}'/N = N[\mathrm{diag}(\boldsymbol{\pi}_{\cdot}) - \boldsymbol{\pi}_{\cdot}\boldsymbol{\pi}_{\cdot}'] \tag{13}$$

and

$$C_1 = \left(1 - \frac{1}{N}\right) \mathrm{diag}(\boldsymbol{\nu})\boldsymbol{\delta}\,\mathrm{diag}(\boldsymbol{\nu}),$$

$$\text{with } \delta_{k\ell} = \gamma_{k\ell\cdot}/\pi_{k\cdot}\pi_{\ell\cdot} - 1, \tag{14}$$

where diag indicates a diagonal matrix and $\boldsymbol{\pi}_{\cdot} = (\pi_{1\cdot}, \pi_{2\cdot}, \ldots, \pi_{K\cdot})'$.

Proof. Let $I_k(i)$ be the indicator function for event $z_i \in \mathcal{Z}_k$, so that $y_k = \sum_{i=1}^{N} I_k(i)$ and, for $k \neq \ell$,

$$E\{y_k y_\ell\} = E\left\{\sum_{i \neq j} I_k(i) I_\ell(j)\right\} = \sum_{i \neq j} \gamma_{k\ell}(i,j)$$

$$= N(N-1)\gamma_{k\ell\cdot}. \tag{15}$$

Then

$$\begin{aligned}\operatorname{cov}(y_k, y_\ell) &= N(N-1)\gamma_{k\ell\cdot} - N^2\pi_{k\cdot}\pi_{\ell\cdot}\\ &= -N\pi_{k\cdot}\pi_{\ell\cdot} + N(N-1)(\gamma_{k\ell\cdot} - \pi_{k\cdot}\pi_{\ell\cdot}).\end{aligned} \tag{16}$$

Similarly,

$$\operatorname{var}(y_k) = N(\pi_{k\cdot} - \pi_{k\cdot}^2) + N(N-1)(\gamma_{kk\cdot} - \pi_{k\cdot}^2), \tag{17}$$

verifying Lemma 1.

If z_i and z_j are independent in (5), then $\gamma_{k\ell}(i,j) = \pi_k(i)\pi_\ell(j)$. Independence of all z-values implies that all of the elements of matrix $\boldsymbol{\delta}$ in (14) are 0, leaving $\operatorname{cov}(\mathbf{y}) = C_0$. Conversely, the amount of correlation between the z-values determines the size of $\boldsymbol{\delta}$ and the increase of $\operatorname{cov}(\mathbf{y})$ above C_0.

To approximate $\boldsymbol{\delta}$, we add the assumption of bivariate normality for any pair of z-values, $\operatorname{cov}(z_i, z_j) \equiv \rho_{ij}$, so that, as in (10),

$$\gamma_{k\ell}(i,j) \doteq \frac{\Delta^2}{2\pi\sqrt{1-\rho_{ij}^2}} \times \exp\left\{-\frac{1}{2}\frac{z[k]^2 - 2\rho_{ij}z[k]z[\ell] + z[\ell]^2}{1-\rho_{ij}^2}\right\}. \tag{18}$$

Letting $g(\rho)$ indicate the empirical density of the $N(N-1)$ correlations ρ_{ij}, and using (10)–(18) yields a useful approximation:

Lemma 2. Under the bivariate normal approximation (18), the matrix $\boldsymbol{\delta}$ in (14) has elements

$$\delta_{k\ell} \doteq \int_{-1}^{1}\left[\frac{1}{\sqrt{1-\rho^2}}\exp\left(\frac{\rho}{2(1-\rho^2)} \times \left\{2z[k]z[\ell] - \rho(z[k]^2 + z[\ell]^2)\right\}\right) - 1\right]g(\rho)\,d\rho. \tag{19}$$

This compares well with theorem 1 of Owen (2005); the assumptions there imply the bivariate normal condition (18).

Application of Lemmas 1 and 2 reqires estimation of the correlation density $g(\rho)$, which we can obtain from observed correlations between the rows of X. As described in Remark A of Section 7, this gives

$$\text{breast cancer:}\quad g(\rho) \mathrel{\dot\sim} \mathrm{N}(0, .153^2) \tag{20}$$

for the breast cancer data and, more roughly,

$$\text{HIV:}\quad g(\rho) \mathrel{\dot\sim} \mathrm{N}(0, .42^2) \tag{21}$$

for the HIV study. It is no accident that $g(\rho)$ has mean near 0; in both cases, the data matrix X had its columns standardized to mean 0 and variance 1 (but not quantile normalized), a usual practice for negating "brightness" disparities between microarrays (see Bolstad, Irizarry, Astrand, and Speed 2003; Qui et al. 2005a,b). This forces the sum of covariances, and nearly the sum of correlations, to be 0. The normality assumed in (20) is not crucial; see Remark B. Section 3 shows that the standard deviation .153 is the vital number. Remark E discusses what happens in the absence of standardization.

Approximation (20) indicates a substantial amount of global correlation among genes in the breast cancer study, and even more correlation for the HIV study [eq. (21)]. The five examples in Owen's (2005) table 1 had standard deviations of $g(\rho)$ between .17 and .26, as did Qui et al.'s (2005a,b) main example. It is not surprising that correlations of this magnitude can undercut standard inference techniques, the key message of Qui et al. The goal here, made explicit in Section 5, is to understand and correct correlational problems.

Table 2. Standard Deviations and Correlations for Central and Tail Counts (Y_0, Y_1) [eq. (23)]; $z_i \sim N(0, 1)$, $i = 1, 2, \ldots, 3{,}226$; for z_i's Independent or z_i's Correlated as in C_{norm} [eq. (22)]; and Permutation Covariance C_{perm} [eq. (41)]

	Independent	C_{norm}	C_{perm}	*Poisson*
sd(Y_0)	26.4	**171.4**	176.0	176.4
sd(Y_1):	4.5	**16.1**	14.9	14.7
cor(Y_0, Y_1)	(−.12)	**(−.89)**	(−.81)	(−.90)

NOTE: C_{norm} and C_{perm} produce much larger standard deviations and much more negative correlations. "Poisson" is calculated from the Poisson approximation model (38).

Substituting (20) into (19) and then into Lemma 1 gives estimated null hypotheses covariance matrix C_{norm},

$$\operatorname{cov}(\mathbf{y}) = C_{\text{norm}}, \tag{22}$$

for the breast cancer data. The correlation term C_1 in $\operatorname{cov}(\mathbf{y}) = C_0 + C_1$ [eq. (12)] dominates the independence term C_0. As an informative example that we use later in the article, define

$$Y_0 = \#\{z_i \in [-1, 1]\} \qquad \text{and} \qquad Y_1 = \#\{z_i \geq 2.5\}, \tag{23}$$

which are central and tail counts for a hypothetical null vector $\mathbf{z}$ satisfying (5). Table 2 gives standard deviations and correlations for Y_0 and Y_1 if the z_i's are independent, or if they are correlated such that $\operatorname{cov}(\mathbf{y}) = C_{\text{norm}}$. Table 2 can be computed from $\operatorname{cov}(\mathbf{y})$, because Y_0 and Y_1 are linear functions of $\mathbf{y}$. Results for C_{perm}, the permutation estimate from Section 4, agree with those for C_{norm}. The cutoffs ±1 and 2.5 in (23) were chosen for convenient exposition and could just as well be replaced by similar values, say ±.75 and 3.0.

We see that gene correlations have a powerful effect on the counts that would be observed under null hypothesis (5); standard deviations are multiplied several fold (this being the main point in Owen 2005), whereas the negative correlation between the central and tail counts is driven toward −1. Tail null counts play a crucial role in computing FDRs, as discussed in Section 5, where the extreme negative correlation between Y_0 and Y_1 is used to "condition" FDR estimates. Section 3 provides an explanation for the negative correlation.

3. FIRST EIGENVECTOR

Lemmas 1 and 2 decompose the covariance matrix of the count vector $\mathbf{y}$ into an independence term C_0 and an additional term C_1 that accounts for correlation among the z-values [eq. (12)]. This section presents a simple approximation to C_1 in terms of its first eigenvector, which we use in Sections 4 and 5 to explain the effects of correlation on simultaneous inference.

The smooth curve in Figure 3 is the first eigenvector of C_{norm}, (20)–(22) the normal-theory estimate of $\operatorname{cov}(\mathbf{y})$ for the breast cancer data, whereas the jagged curve is the corresponding quantity for the permutation-based estimate C_{perm} of Section 4. The dots indicate the first eigenvector of C_{norm} for the HIV data, using (21). All three curves exhibit the same "wing-shaped" form. This will turn out to be proportional to

$$w(z) \equiv \varphi(z)\frac{z^2 - 1}{\sqrt{2}} = \varphi''(z)/\sqrt{2}, \tag{24}$$

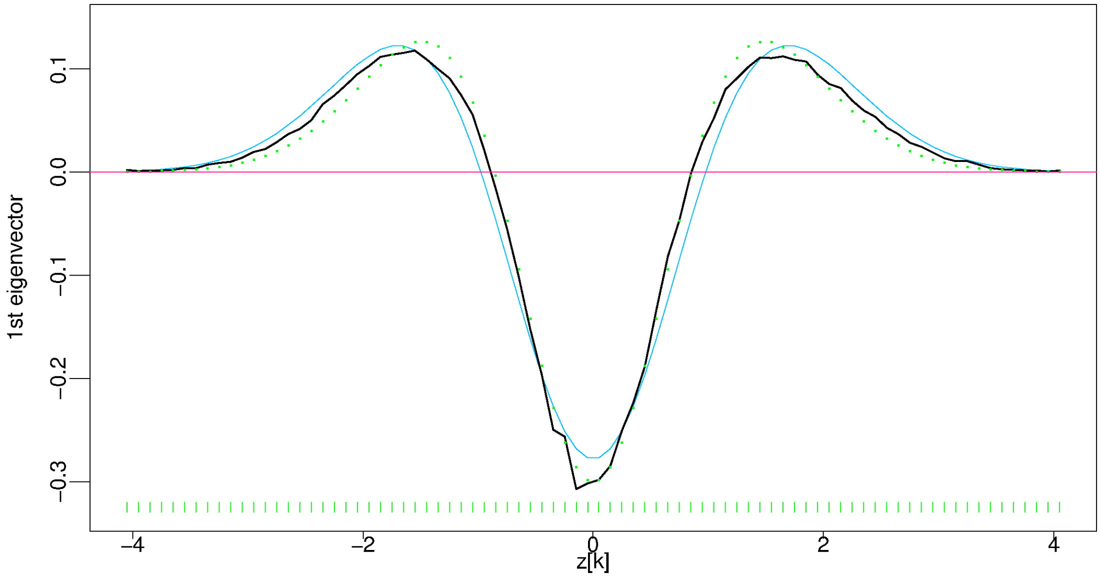

*Figure 2. First Eigenvectors of Three Different Estimates of cov(**y**): Normal-Theory Estimate (22) for Breast Cancer Data (smooth curve); Normal-Theory Estimate for HIV Data (21) (dots), and Permutation Estimate for Breast Cancer Data (41) (jagged curve). The striking "wing-shaped" form is proportional to the second derivative of the standard normal density. Dashes indicate bin midpoints z[k].*

where $\varphi(z)$ is the standard normal density [eq. (10)].

We use notation (7)–(10) and work in the discretized framework of Lemmas 1 and 2.

Lemma 3. Suppose that $g(\rho)$, the correlation density in (19), has mean 0 and standard deviation

$$\alpha = \left[\int_{-1}^{1} \rho^2 g(\rho)\, d\rho\right]^{1/2}. \tag{25}$$

Then the matrix $\boldsymbol{\delta}$ in (14) is approximated by the outer product

$$\boldsymbol{\delta} \doteq \alpha^2 \mathbf{q}\mathbf{q}', \quad \text{where } q_k = \frac{z[k]^2 - 1}{\sqrt{2}}, \tag{26}$$

with $z[k]$ the centerpoint of the kth histogram bin, $k = 1, 2, \ldots, K$. Approximation (26) becomes exact as $\alpha \to 0$, with $\boldsymbol{\delta}/\alpha^2 \to \mathbf{q}\mathbf{q}'$.

Proof. Let $R_{k\ell}(\rho)$ be the integrand in (19),

$$R_{k\ell}(\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{\frac{\rho}{1-\rho^2} z[k]z[\ell] - \frac{1}{2}\frac{\rho^2}{1-\rho^2}\left(z[k]^2 + z[\ell]^2\right)\right\} - 1. \tag{27}$$

Expanding $R_{k\ell}(\rho)$ in a Taylor series around $\rho = 0$, and ignoring terms of order ρ^3 or higher, gives

$$R_{k\ell}(\rho) \doteq \rho z[k]z[\ell] + \rho^2 q_k q_\ell. \tag{28}$$

Then, because $g(\rho)$ has mean 0,

$$\delta_{k\ell} = \int_{-1}^{1} R_{k\ell}(\rho) g(\rho)\, d\rho \doteq \alpha^2 q_k q_\ell, \tag{29}$$

which is (26).

Combining the three lemmas yields a useful approximation for the null covariance matrix of the count vector **y** under the bivariate normal assumptions of Section 2.

Theorem. If $g(\rho)$ has mean 0 and standard deviation α, then

$$\operatorname{cov}(\mathbf{y}) \doteq [\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}'/N] + \left(1 - \frac{1}{N}\right)(\alpha\mathbf{W})(\alpha\mathbf{W})'. \tag{30}$$

Here $\boldsymbol{\nu} = E\{\mathbf{y}\}$ as in (11), whereas $\mathbf{W}$ has components

$$W_k = N\Delta\varphi(z[k])\frac{z[k]^2 - 1}{\sqrt{2}} = N\Delta w(z[k]), \tag{31}$$

where $w(\cdot)$ is the wing-shaped function (24), N is the number of cases, and Δ is the bin width.

The theorem helps explain Figure 2; the second term in (30) dominates cov(**y**) in our two examples, making its first eigenvector nearly proportional to $w(z)$.

Table 3 relates to the accuracy of the Theorem. It shows the proportion of variance explained by the first eigenvector (i.e., the first eigenvalue divided by the sum of eigenvalues) for C_1, the correlation term in (12), and also for $\operatorname{cov}(\mathbf{y}) = C_0 + C_1$, assuming that $g(\rho) \sim \mathrm{N}(0, \alpha^2)$. For the breast cancer value $\alpha = .153$, the proportions are 98% for C_1 [the crucial quantity for the accuracy of (30)] and 45% for $C_{\text{norm}} = \operatorname{cov}(\mathbf{y})$. Although $N = 3{,}226$ in Table 2, this choice has little effect on the numbers. The 98% proportion indicates the theorem's substantial accuracy in the breast cancer context. For the HIV data, the proportion was 86%, still quite adequate.

The Theorem summarizes the effect of **z**'s entire correlation structure in a single parameter α. This permits a relatively

Table 3. Proportion of Total Variance Explained by the First Eigenvector, as a Function of α, for C_1, the Correlation Term in (12), and Also for C_{norm}, Assuming That $g(\rho) \sim N(0, \alpha^2)$ and N = 3,226

	α						
	0	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*
C_1	**1.00**	**1.00**	**.99**	**.98**	**.97**	**.95**	**.90**
C_{norm}	.04	.10	.27	.45	.59	.68	.72

NOTE: Proportions for C_1 determine the accuracy of approximation (30).

simple analysis of the inferential effects of correlation in what follows. Note that Hsu (1992, 1996) gave a thorough discussion of "one-factor" approximations to a correlation matrix and their use in simplifying simultaneous estimation calculations, just as in our Theorem. Here the technique is applied to the order statistic of $\mathbf{z} = (z_1, z_2, \ldots, z_N)'$ (in the form of the binned counts $\mathbf{y}$), with the surprising benefit that the first principle factor must lie near the wing-shaped function $\mathbf{W}$.

Poisson process considerations lead to a somewhat rough but evocative interpretation of the Theorem. Let $\mathbf{y} \sim \text{Po}(\mathbf{u})$ indicate a vector of independent Poisson variates, $y_k \overset{\text{ind}}{\sim} \text{Po}(u_k)$ for $k = 1, 2, \ldots, K$, whereas $\mathbf{y} \sim (\boldsymbol{\nu}, \Gamma)$ denotes that vector $\mathbf{y}$ has mean $\boldsymbol{\nu}$ and covariance Γ.

It is convenient here to consider the number of cases N to be a Poisson variate, say

$$N \sim \text{Po}(N_0), \qquad (32)$$

with $N_0 = 3{,}226$ in the breast cancer study. This simplifies (30) slightly, to

$$\text{cov}(\mathbf{y}) \doteq \text{diag}(\boldsymbol{\nu}) + (\alpha\mathbf{W})(\alpha\mathbf{W})', \qquad (33)$$

with $\boldsymbol{\nu}$ and $\mathbf{W}$ as in (11), (31) except that N_0 replaces N. If the z-values are independent, then (32) makes the counts, y_k, independent Poisson variates,

$$\mathbf{y} \sim \text{Po}(\boldsymbol{\nu}), \qquad (34)$$

agreeing with (33) at $\alpha = 0$.

A hierarchical model generalizes (34) to incorporate dependence. We assume that $\mathbf{y}$ depends on a mean vector $\mathbf{u}$, itself random,

$$\mathbf{y}|\mathbf{u} \sim \text{Po}(\mathbf{u}), \quad \text{where } \mathbf{u} \sim (\boldsymbol{\nu}, \Gamma), \qquad (35)$$

so that the components of $\mathbf{y}$ are conditionally independent given $\mathbf{u}$ but marginally dependent, with mean and covariance

$$\mathbf{y} \sim (\boldsymbol{\nu}, \text{diag}(\boldsymbol{\nu}) + \Gamma). \qquad (36)$$

To match (33), we need to set

$$\Gamma = (\alpha\mathbf{W})(\alpha\mathbf{W})'. \qquad (37)$$

Formulas (36) and (37) suggest a hierarchical Poisson structure for the count vector $\mathbf{y}$,

$$\mathbf{y} \sim \text{Po}(\mathbf{u}), \quad \text{where } \mathbf{u} = \boldsymbol{\nu} + A\mathbf{W}, \text{ with } A \sim (0, \alpha^2). \qquad (38)$$

If $\alpha = 0$, then this reduces to the independence case (34); otherwise, the Poisson intensity vector $\boldsymbol{\nu}$ is modified by the addition of an independent random multiple A of $\mathbf{W}$ with standard deviation α.

Model (38) can be only an approximation, because $\mathbf{u}$ may have negative coordinates, but it nicely explains phenomena like the extreme negative correlations between Y_0 and Y_1 shown in Table 2. Vector $\mathbf{W}$ is negative in $[-1, 1]$ and positive elsewhere, as in Figure 3, so positive A in (38) decreases the central counts and increases the tail counts [eq. (23)]. The opposite occurs when A is negative. [The "Poisson" column of Table 3 was calculated from model (38), $\alpha = .153$ and $N = 3{,}226$, except that the components of $\mathbf{u}$ were truncated at 0.]

Examining model (38) more carefully, the mean vector $\mathbf{u}$ turns out to be roughly proportional to a scaled normal density,

$$u_k \doteq \frac{N\Delta}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{z[k]^2}{2\sigma_0^2}\right\}, \quad \text{with } \sigma_0^2 = 1 + \sqrt{2}A; \qquad (39)$$

see Remark D. Model (39) implies that positive A makes the counts behave in an overdispersed normal fashion compared with the theoretical N(0, 1) density, and conversely for negative A. This confirms the first main point made in Section 1: Even if the null z-values are individually N(0, 1), correlation can make the ensemble $\mathbf{z}$ behave as $\text{N}(0, \sigma_0^2)$, with σ_0 substantially different than 1. Section 4 shows the same phenomenon in a permutation analysis. This point is refined in Section 5, where it is shown that A can be estimated and used to condition simultaneous hypothesis tests.

4. PERMUTATION METHODS

The previous results depend on the assumption of bivariate normality for every pair of z-values. Permutation methods lead to a direct empirical estimate, C_{perm}, for $\text{cov}(\mathbf{y})$. Carried out here for the breast cancer data, C_{perm} agrees well with the normal-theory estimate C_{norm} [eq. (22)] and lends support to the inferential theory of Section 5.

Let X represent the $3{,}226 \times 15$ matrix of observed expressions. Each row of X (i.e., each gene) provides a two-sample t-statistic comparing the eight BRCA2 and seven BRCA1 expressions, with $\mathbf{t}$ representing the vector of all 3,226 t-values. Repeating the computation after a random permutation of the columns of X (i.e., after a random division of the patients into groups of seven and eight) yields permuted matrix X^* and t-vector $\mathbf{t}^*$. A total of 1,000 such permutations were used to estimate a permutation null distribution G_0 for the t-values (the empirical distribution of all $1{,}000 \cdot 3{,}226 t_i$'s), which turned out to be slightly shorter-tailed than a standard t variate with 13 degrees of freedom; the z-values for Figure 1 were the calculated as in (1), $z_i = \Phi^{-1}(G_0(t_i))$.

The permutations contain information beyond the composite marginal distribution of the t_i^*'s. Each permuted data matrix X^* produces $\mathbf{t}^*$, corresponding z-value vector $\mathbf{z}^*$, and count vector $\mathbf{y}^*$ as in (6),

$$\underset{3{,}226\times 15}{X^*} \to \underset{3{,}226}{\mathbf{t}^*} \to \underset{3{,}226}{\mathbf{z}^*} \to \underset{82}{\mathbf{y}^*}. \qquad (40)$$

The sample covariance matrix of the 1,000 $\mathbf{y}^*$'s,

$$C_{\text{perm}} = \sum_{b=1}^{1000} (\mathbf{y}^{*b} - \mathbf{y}^{*\cdot})(\mathbf{y}^{*b} - \mathbf{y}^{*\cdot})'/999$$

$$[\mathbf{y}^{*\cdot} = \textstyle\sum \mathbf{y}^{*b}/1{,}000], \qquad (41)$$

is a nonparametric estimate of the null covariance matrix for $\mathbf{y}$. By permuting entire microarrays, we preserve the correlation structure of the genes while nullifying any actual BRCA1–BRCA2 differences. Note that the matrix X used to form X^* in (40) had each gene's BRCA1 or BRCA2 average subtracted from its corresponding expression values to eliminate any genuine group differences from the permutation results. Table 2 and Figure 3 demonstrate the similarity of C_{perm} and C_{norm}.

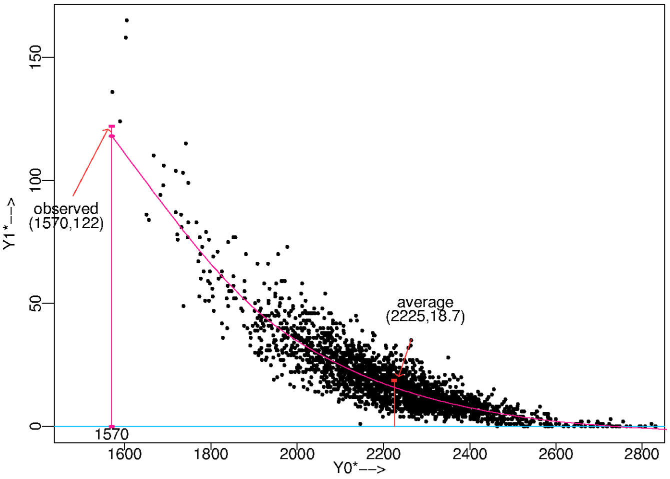

Figure 3. Central and Tail Counts (Y_0^*, Y_1^*) as in (23), for 4,000 Columnwise Permutations of the Breast Cancer Data. Y_1^* has unconditional expectation 18.7 but conditional expectation about 118 given Y_0^* equal to the observed central count 1,570. The smooth curve represents the fitted smoothing spline.

Each permutation vector $\mathbf{z}^*$ gave central and tail counts (Y_0^*, Y_1^*) as in (23). Figure 4 plots Y_1^* versus Y_0^*, now for 4,000 permutations. The unconditional average of Y_1^* is 18.7, but this does not take into account the powerful covariate Y_0^*. In particular, for Y_0^* equaling the observed count of 1,570, a conditional expectation of about 118 is predicted. In Section 5 we discuss why the observed central count might be so atypical of the permutation values in Figure 4.

In the actual breast cancer data, 122 genes have z_i's exceeding 2.5. Does this collection of 122 genes deserve to be reported as "mostly nonnull"? The answer obviously depends on whether the expected number of null genes having $z_i \geq 2.5$ is 18.7 or 118. In Section 5 we investigate this question, which bears on the second main point of Section 1, the effect of correlation on simultaneous inference.

The central count Y_0^* provides a convenient estimate of standard deviation for $\mathbf{z}^*$, a permutation vector of N z-values,

$$\widehat{\sigma}_0^* = 1\Big/\Phi^{-1}\left(\frac{1+Y_0^*/N}{2}\right). \qquad (42)$$

If we assume that the elements of $\mathbf{z}^*$ are normally distributed with mean 0 and some variance σ_0^{*2},

$$z_i^* \sim \mathrm{N}(0, \sigma_0^{*2}), \qquad (43)$$

then $E\{Y_0^*\} = N \cdot [2\Phi(1/\sigma_0^*) - 1]$, so (42) is a method-of-moments estimate for σ_0^* that depends only on the central count Y_0^*.

In Figure 5, the 4,000 permutation $\mathbf{y}^*$ vectors have been averaged in groups having about the same central spread $\widehat{\sigma}_0^*$ [eq. (42)]; curve 1.2 is the log of the average of those $\mathbf{y}^*$ vectors with $1.15 \leq \widehat{\sigma}_0^* \leq 1.25$, and so on. To first order, curves fall off as $-z^2/(2\sigma_0^2)$, with $\sigma_0 = 1.2$, 1, and .8, that is, as the log of a $\mathrm{N}(0, \sigma_0^2)$ density. This agrees with theoretical result (39), again showing how correlation can widen or narrow the effective null distribution; those $\mathbf{y}^*$'s with large central spread put more null cases into the far tails, and vice versa. Section 5 discusses how this phenomenon affects simultaneous significance testing, particularly FDR methods.

5. LARGE–SCALE SIGNIFICANCE TESTING

The scientific world is fond of significance testing because it requires a minimum of probabilistic modeling, no more than the specification of a null hypothesis distribution. However, as described in Sections 2–4, a disturbing danger arises in large-scale testing situations: Correlations among the test statistics may substantially widen or narrow the effective null distribution. This section discusses the consequences of correlation effects on FDRs and other simultaneous testing techniques, as well as methods for correcting their inferences.

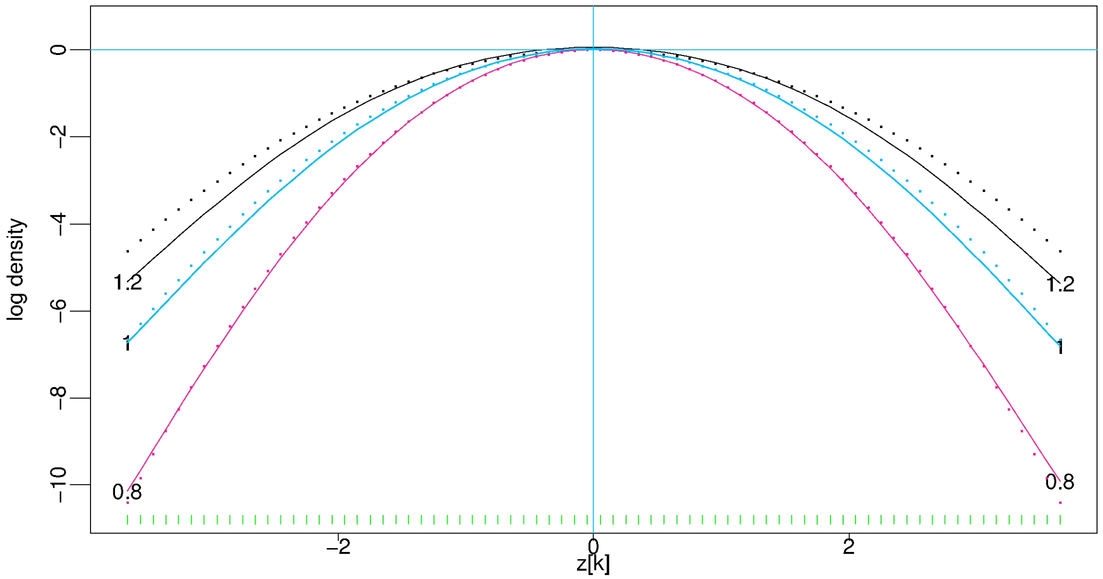

Figure 4. Log Average Densities for the 4,000 $\mathbf{y}^*$ Vectors Contributing to Figure 3. Here "1.2" graphs log $\mathbf{y}_k^*$, from an average of $\mathbf{y}^*$'s having $\widehat{\sigma}_0^*$ in [1.15, 1.25], versus z[k], and so on. Dotted curves are $-.5\ z[k]^2/\sigma_0^2$ for $\sigma_0^2 = 1.2$, 1.0, .8.

Suppose for a moment that we knew which z_i's among the full set of z-values correspond to null cases. For a given choice of x, define

$$Y(x) = \#\{\text{null } z_i \geq x\} \qquad \text{and} \qquad T(x) = \#\{z_i \geq x\}. \tag{44}$$

Y_1 in (24) is $Y(2.5)$ in this notation, and $T(2.5) = 122$ for the breast cancer study. Lehmann and Romano (2005) defined the *false discovery proportion* (FDP) to be

$$\text{FDP}(x) = Y(x)/T(x), \tag{45}$$

assuming that we are searching for "discoveries" only in the right tail. If FDP(x) were known (but the identity of the null cases was not), say $\text{FDP}(2.5) = 20/122 = .16$, then the group of 122 genes could be reported as "mostly significant," with the assurance of producing only 16% false discoveries.

In practice, $Y(x)$ is unobservable, as is FDP(x). A useful tactic is to replace $Y(x)$ by its expectation, as done by Benjamini and Hochberg (1995), giving an estimated FDR defined as

$$\text{FDR}(x) = E\{Y(x)\}/T(x). \tag{46}$$

Benjamini and Hochberg's procedure actually involves the expected *ratio*,

$$\text{FDR}(x) = E\{Y(x)/T(x)\}, \tag{47}$$

ingeniously prescribing a data-based choice of x that controls the FDR below some preset value.

Our calculations focus on FDR(x) [eq. (46)], an observable ratio that is important and useful in its own right. FDR(x) is an empirical Bayes estimate of the a posteriori probability that case i is null given $z_i \geq x$ (Storey 2002; Efron and Tibshirani 2002), amounting to a version of Storey's "q-value." Because $T(x)$ *is* observable, FDR(x) has intuitive interpretation as the expected proportion of null cases among those having $z_i \geq x$.

Formula (46) glosses over the fact that $E\{Y(x)\}$ itself is not directly calculable. Benjamini and Hochberg's original procedure replaced $E\{Y(x)\}$ with its upper bound assuming that all N cases were null, as in (5), where the theoretical null gives

$$E\{Y(x)\} = N \cdot \bar{\Phi}(x) \qquad [\bar{\Phi}(x) \equiv 1 - \Phi(x)]. \tag{48}$$

Improvements on (48) are possible through estimation of p_0, the proportion of null cases (Langaas and Lindquist 2005; Storey, Taylor, and Siegmund 2004). For p_0 near 1.0, the preferred situation in microarray studies, where the goal is to discover a small number of genuinely interesting genes, (48) is a good starting point for the discussion of correlation effects.

The hierarchical structure (38) gives conditional expectation

$$E\{\mathbf{y}|A\} = \boldsymbol{\nu} + A\mathbf{W}, \tag{49}$$

using only $E\{\mathbf{y}|\mathbf{u}\} = \mathbf{u}$, not the full Poisson assumptions. Letting the bin width $\Delta \to 0$ in (11) and (31) produces a conditional version of (49),

$$E\{Y(x)|A\} = N\bar{\Phi}(x)\left[1 + A\frac{x\varphi(x)}{\sqrt{2}\bar{\Phi}(x)}\right]; \tag{50}$$

see Remark H. The term multiplying A equals 4.99 at $x = 2.5$, giving conditional expectations $N\bar{\Phi}(x) \cdot (1.75)$ for $A = .15$ and $N\bar{\Phi}(x) \cdot (.25)$ for $A = -.15$.

Even such relatively modest values of A greatly affect the *conditional FDR*,

$$\text{FDR}(x|A) = E\{Y(x)|A\}/T(x), \tag{51}$$

which can be expressed as

$$\text{FDR}(x|A) = \text{FDR}_0(x)\left[1 + A\frac{x\varphi(x)}{\sqrt{2}\bar{\Phi}(x)}\right], \tag{52}$$

where $\text{FDR}_0(x)$ is the standard unconditional estimate $N\bar{\Phi}(x)/T(x)$ based on the theoretical N(0, 1) null. For $x = 2.5$, FDR($x|A$) varies by a factor of 7 as A ranges from $-.15$ to $+.15$. A principal point of this article is that conditional FDR estimates are available in situations like those of Figure 1, whereas the unconditional estimates can produce grossly misleading results, as shown in Figure 2.

The idea in what follows is that A in (52), or some equivalent parameter, can be estimated from the central spread of the histogram of z-values and then used to condition inferences, as in Figure 2. Generalizing definition (23), for $x_0 > 0$, let

$$Y_0 = \#\{z_i \in [-x_0, x_0]\}, \tag{53}$$

and define

$$P_0 = 2\Phi(x_0) - 1 \qquad \text{and} \qquad Q_0 = \sqrt{2}x_0\varphi(x_0). \tag{54}$$

Then (49) and definitions (11) and (31) give $E\{Y_0|A\} \doteq N[P_0 - AQ_0]$, suggesting that

$$\widehat{A} = \frac{P_0 - \widehat{P}_0}{Q_0} \qquad [\widehat{P}_0 = Y_0/N] \tag{55}$$

as an estimate of A. Remark H shows that $x_0 = 1$ is a reasonable choice and derives the approximate standard error given A, yielding estimates

$$\begin{aligned} &\text{breast cancer:} \quad \widehat{A} = .57 \pm .04 \qquad \text{and} \\ &\text{HIV:} \quad \widehat{A} = -.21 \pm .03 \end{aligned} \tag{56}$$

for our two examples. For the breast cancer data, (56) implies $E\{Y(2.5)|\widehat{A}\} = 77$ and $\text{FDR}(2.5)|\widehat{A}) = 77/122 = .63$, which are underestimates according to our later calculations.

Permutation methods permit model-free estimates of the conditional FDR, as in Figure 4, which suggests that $E\{Y(2.5)|Y_0\} \doteq 118$, with a corresponding estimated FDR of $118/122 = .97$. Both of these approaches depend on the same basic idea: We use the observed central count Y_0 to condition the estimate of $Y(x)$, the unobserved null tail count. This is similar in spirit to Fisher's exact test for a 2×2 table, where the observed table margins, playing the role of Y_0, are used to establish the appropriate conditional null distribution.

At this point, we could substitute $\widehat{A}$ from (55) into (52) to obtain a conditional FDR estimate,

$$\text{FDR}(x|\widehat{A}) = \text{FDR}_0(x)\left[1 + \widehat{A}\frac{x\varphi(x)}{\sqrt{2}\bar{\Phi}(x)}\right], \tag{57}$$

where FDR_0 is the unconditional estimate based on the theoretical null. In situations like that of Figure 2, the goal would be to accurately assess our position on the A axis, to better estimate FDP rather than estimating the unconditional average of FDP.

Figure 6 reports on a small simulation experiment comparing conditional and nonconditional FDR estimates. The simulation

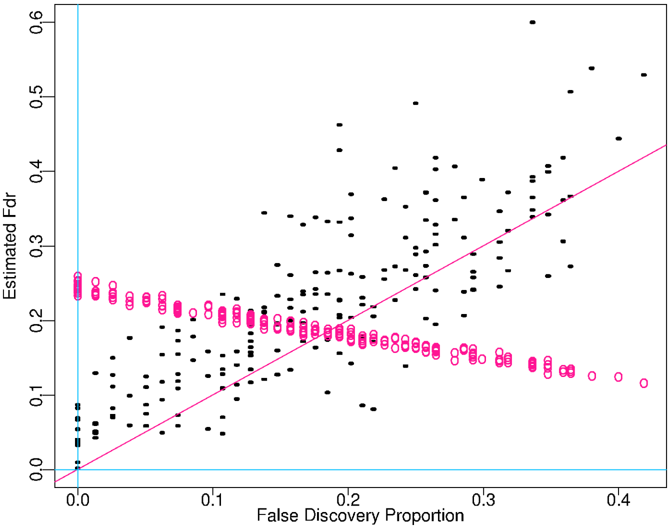

Figure 5. Simulation Experiment Comparing Conditional FDR Estimates (solid points), With Unconditional Estimates (open circles); N = 3,000, p_0 = .95. Null counts are generated as in (38), α = .15; nonnull counts from z ∼ N(2.5, 1.25). The horizontal axis is the actual FDP, FDP(x), x = 2.5, for each of the 200 trials. The unconditional estimate based on the theoretical null distribution declines as actual FDP increases.

involved 200 trials, each with $N = 3{,}000$ z-values, proportion $p_0 = .95$ null. The null counts were generated from the Poisson model (38) with $\alpha = .15$, whereas the 150 nonnull z's followed a N(2.5, 1.25) distribution.

For each of the 200 trials, conditional and unconditional estimates of the FDP(x), $x = 2.5$ [definition (5.2)], are plotted versus FDP(x). Strikingly, the unconditional estimate goes in the wrong direction, declining as the actual FDP increases. This yields misleading inferences at both ends of the FDP scale. The conditional FDR estimate correctly tracks FDP, although with a considerable amount of noise. Improved forms of (57) that included estimates of p_0, the null proportion, were used for both the conditional and unconditional FDR estimates in Figure 5, based on a version of the *locfdr* algorithm (Efron 2006), available as an R function from the Comprehensive R Archive Network; *locfdr* gave the empirical null estimates (3).

Figure 6 supports the second main claim of Section 1, that correlation effects can and sometimes must be taken into account in the analysis of large-scale simultaneous testing problems. Not doing so may yield dangerously misleading estimates of actual FDPs.

To reiterate the basic idea, even assuming that null cases individually follow the theoretical null distribution $z_i \sim \mathrm{N}(0, 1)$, correlation effects can make the ensemble null distribution behave more like $\mathrm{N}(0, \sigma_0^2)$ with σ_0 surprisingly far from 1. Ignoring this effect can undercut any simultaneous testing procedure, not just FDRs; almost by definition, all hypothesis testing procedures require an accurate notion of "null."

Other factors besides z-value correlations can affect the null distribution. In earlier work (Efron 2004, 2006), we suggested two other possibilities that support the use of empirical nulls: unidentified covariates in an observational study and correlations *across* microarrays. In fact, Figure 4 shows that it is unlikely that correlation effects alone are responsible for the extreme central overdispersion. These factors, as well as correlation, argue against uncritical use of theoretical null distributions in large-scale testing problems.

6. DISCUSSION

Massive datasets like those from the breast cancer and HIV studies can be misleadingly comforting in their suggestion of great statistical accuracy. Correlation considerations produce a more sobering picture. A single degree of freedom, embodied by the random variable A in (38), can dominate variability as it does in Table 2. Qui et al. (2005b) emphasized the harmful effect of correlation, using it as an argument against empirical Bayes microarray analyses such as those of Storey (2002) or Efron, Tibshirani, Storey, and Tusher (2001). Their arguments also could be used against all other popular microarray analysis techniques.

The results presented in this article are more optimistic. The correlation effect, although perhaps very large, is shown to manifest itself through the simple wing-shaped function of Section 3. This enables the statistician to identify and remove much of it. In Figure 4, for example, the tail count Y_1 has standard deviation 15.5, much larger than the value 4.5 applying to independent z_i's, but the residual standard deviation reduces to 6.5 after prediction from the central counts. This is the idea behind the empirical null (Efron 2004, 2006), whose most important job is predicting the conditional expectation of the tail null counts.

That being said, Qui et al.'s concern for the consequences of correlation on microarray analyses (nicely summarized in their sec. 7) remains pertinent, especially in high-correlation settings like the HIV study. Improved biomedical methods in exposure, registration, and background control of arrays may alleviate the problem, as may new array designs that incorporate greater gene duplication. Purely statistical improvements can also reduce correlations, for instance, by more extensive standardization techniques Qui et al. (2005a). However, none of this will help if microarray correlations are inherent in the way genes interact at the DNA level rather than a limitation of current methodology.

van der Laan and coworkers (van der Laan and Hubbard 2005; Dudoit et al. 2004) have developed a comprehensive resampling theory for correlated testing problems. This theory is much more general than that presented here. However, this generality comes at a price in terms of assumptions and applicability; the columns of the $N \times n$ data matrix X are considered to be iid. N-vectors, and the main results are justified by asymptotic bootstrap arguments, as the number of microarrays n goes to infinity. Examples like those in Figure 1, with $n \ll N$, raise legitimate concerns about the relevance of $n \to \infty$ asymptotics.

Not all large-scale testing situations involve microarrays. Correlation may be less of a problem in other scientific venues, such as functional magnetic resonance imaging or time-of-flight spectroscopy. In any case, it seems worthwhile to obtain some overall measure of correlation such as α [eq. (25)]. Large α's suggest the use of correlation-resistant analysis techniques like the FDR/empirical null combination.

7. REMARKS

Remark A (Sec. 2). Empirical correlation distributions (20) and (21) were obtained from the row-wise correlations of the original data matrix X, with X 3,226 × 15 and 7,680 × 8 in our two examples. Let $\widehat{\rho}_{ij}$ be the sample correlation between rows i and j of X, after first subtracting off each gene's average response within each treatment group (to nullify any genuine treatment differences); $g(\rho)$ is essentially the empirical distribution of all $N(N-1)/2$ $\widehat{\rho}_{ij}$ values (as in Owen 2005), but when dealing with small numbers of microarrays (only 15 or 8 points per correlation in our two examples), some care must be taken to remove the variability added by sampling error. This was done by transforming to

$$\widehat{\xi}_{ij} = \frac{1}{2}\log\frac{1+\widehat{\rho}_{ij}}{1-\widehat{\rho}_{ij}}, \tag{58}$$

assuming a translation model $\widehat{\xi}_{ij} = \xi_{ij} + \epsilon$ on this scale, estimating the distribution of ϵ by repeating the calculations beginning with matrices X^* in which the entries within columns of X were independently permuted, inferring the ξ distribution by deconvolution, and retransforming back to the ρ scale. These calculations apply to the correlations within each column of X. Assuming independent columns, it is easy to demonstrate by simulation that nearly the same $g(\rho)$ distribution applies to the z-values (1).

Remark B (Sec. 2). Owen's examples support normality for $g(\rho)$ as in (20), but a wide range of other distributions fit reasonably well. Table 4 gives the best distribution supported on just three ρ values. "Best" here is defined in terms of numerically minimizing a chi-squared discrepancy between the empirical distribution of the $\widehat{\xi}_{ij}$ values and the model $\widehat{\xi} = \xi + \epsilon$, taking the ϵ distribution as earlier. The solution turned out to have mean 0 and standard deviation $\alpha = .153$, as in (20), and gave about the same estimate of cov($\mathbf{y}$) as (22).

Remark C (Sec. 2). The amount of genewise correlation represented by (20) is enormous. For comparison, suppose that there were actually 10 equal-sized groups of genes with independence across groups but $\rho_{ij} = .50$ for all genes within groups. After standardization of X, this gives $\alpha = .15$, about the same as (20).

Remark D (Sec. 3). The components of $\mathbf{u} = \boldsymbol{\nu} + A\mathbf{W}$ in (38) are

$$u_k \doteq N\Delta f_A(z[k]), \quad \text{where } f_A(z) = \varphi(z)\cdot[1 + Aq(z)] \tag{59}$$

and $q(z) = (z^2-1)/\sqrt{2}$ [eqs. (11), (31)]. Here $f_A(z)$ is symmetric around 0, with even moments easily obtained from Hermite polynomial calculations,

$$\int_{-\infty}^{\infty} f_A(z)\,dz = 1,$$
$$\int_{-\infty}^{\infty} f_A(z)z^2\,dz = 1 + \sqrt{2}A, \qquad \text{and} \tag{60}$$
$$\int_{-\infty}^{\infty} f_A(z)z^4\,dz = 3 + \frac{12}{\sqrt{2}}A.$$

Table 4. Best Three-Point Distribution Estimate for the Breast Cancer Correlation Density g(ρ)

ρ	−.250	0	.444
$g(\rho)$	.131	.793	.076

This supports approximation (39), which can be improved on using higher-order Edgeworth terms.

Remark E (Sec. 3). The vectors $\boldsymbol{\nu}$ and $\mathbf{W}$ in (38) relate to the 0th and second Hermite polynomials. Standardization of the columns of X suppresses the first polynomial, with important consequences here. Without standardization, the first eigenvector of cov($\mathbf{y}$), divided by $\varphi(z)$, may be proportional to z, the first polynomial, instead of the second polynomial $z^2 - 1$.

Remark F (Sec. 4). The permutation calculations at the beginning of Section 4 provided 1,000 z_i^* values for each index i, after which the empirical distribution of all 3,226 × 1,000 values provided the "permutation null" G_0. This kind of calculation can ignore correlation among the z_i's because it depends only on marginal permutation distributions. It is worth restating that permutation null distributions as typically computed tend to resemble theoretical nulls and do *not* automatically compensate for correlation effects. The sophisticated permutation algorithms of Westfall and Young (1993) and Westfall (1997) do involve genewise correlations, but applied to different purposes than specified in this article. Their "step-down max-T" algorithm gave results similar to using a N(0, 1) null for the breast cancer and HIV studies.

Remark G (Sec. 4). Permutation methods produced unstable results for the HIV data. In part this reflects small sample sizes, with only 34 distinct permutations available. Of more concern, there seem to be secular effects systematically disturbing expression levels *across* microarrays. A version of Remark A based on random subsamples of the 7,680 genes gave (21).

Remark H (Sec. 5). Linear functions of the count vector $\mathbf{y}$ yield useful estimates of A. For $\mathbf{m} = (m_1, m_2, \ldots, m_K)'$, define

$$\theta_m = \sum_k m_k u_k/N \qquad \text{and} \qquad \widehat{\theta}_m = \sum_k m_k y_k/N \tag{61}$$

in Poisson model (38). It is convenient to work with a continuous version of $\mathbf{m}$, say $m(z)$, where $m(z[k]) = m_k$. Letting

$$P_m = \int_{-\infty}^{\infty} m(z)\varphi(z)\,dz \quad \text{and} \quad Q_m = -\int_{-\infty}^{\infty} m(z)q(z)\varphi(z)\,dz \tag{62}$$

gives

$$\theta_m \doteq \int_{-\infty}^{\infty} m(z)f_A(z)\,dz = P_m - AQ_m, \tag{63}$$

as in (59). If $m(z)$ is the indicator function of (x, ∞), then (63) becomes (50).

Because $E\{\widehat{\theta}_m|A\} = \theta_m$, (63) suggests

$$\widehat{A} = \frac{P_m - \widehat{\theta}_m}{Q_m} \tag{64}$$

as a method-of-moments estimator for A; (55) is (64), where $m(z)$ is the indicator function of $(-x_0, x_0)$, Model (38) then yields $\text{var}\{\widehat{\theta}_m|A\} \doteq \int f_A(z)m(z)^2\,dz/N$ and

$$\text{var}\{\widehat{A}_m|A\} \doteq \frac{1}{N}\frac{\int_{-\infty}^{\infty} f_A(z)m(z)^2\,dz}{(\int_{-\infty}^{\infty}\varphi(z)q(z)m(z)\,dz)^2}. \tag{65}$$

This formula, with A equaling .57 and −.21 for the two studies, gave the standard errors in (56).

Suppose that we wish to minimize (65) among functions $m(z)$ supported on $(-x_0, x_0)$. The formula is linear in A and in fact does not vary much across reasonable values of A. At $A = 0$, standard theory says that choosing $m(z)$ proportional to $q(z) = (z^2 - 1)/\sqrt{2}$ within $(-x_0, x_0)$ is optimal, with the minimum variance equaling $[N \int_{-x_0}^{x_0} \varphi(z) q(z)^2 \, dz]^{-1}$. For $x_0 = 1$, this gives $\mathrm{var}\{\widehat{A}_m | A = 0\} = 5.03/N$, compared with $1/N$ for $x_0 = \infty$ (the ideal choice but an unallowable one given the possibility of biasing $\widehat{A}_m$ with nonnull data). Taking $m(z)$ as the indicator of $(-1, 1)$ gives $5.83/N$. Reducing x_0 to .80 provides slightly smaller variance when $m(z)$ is the indicator, $5.36/N$. The more important point is that *conditional* variance estimates, as in (65), are both convenient and appropriate for the calculations here.

[Received July 2005. Revised July 2006.]

REFERENCES

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.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003), "A Comparison of Normalization Methods for High-Density Oligonucleotide Array Data Based on Variance and Bias," *Bioinformatics*, 19, 185–193.

Dudoit, S., van der Laan, M., and Pollard, K. (2004), "Multiple Testing, Part I. Single-Step Procedures for Control of General Type I Error Rates," *Statistical Applications in Genetics and Molecular Biology*, 3, article 13.

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

——— (2005), "Local False Discovery Rates," available at *http://www-stat.stanford.edu/∼rbrad/papers/False.pdf*.

——— (2006), "Size, Power, and False Discovery Rates," available at *http://www-stat.stanford.edu/∼brad/papers/Size.pdf*; *The Annals of Statistics*, to appear.

Efron, B., and Tibshirani, R. (2002), "Empirical Bayes Methods and False Discovery Rates for Microarrays," *Genetic Epidemiology*, 23, 70–86.

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.

Ge, Y., Dudoit, S., and Speed, T. (2003), "Resampling-Based Multiple Testing for Microarray Data Analysis" (with comments), *Test*, 12, 1–77.

Hedenfalk, I., Duggen, D., Chen, Y., et al. (2001), "Gene Expression Profiles in Hereditary Breast Cancer," *New England Journal of Medicine*, 344, 539–548.

Hsu, J. C. (1992), "The Factor Analytic Approach to Simultaneous Inference in the General Linear Model," *Journal of Graphical and Computational Statistics*, 1, 151–168.

——— (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.

Ishwaran, H., and Rao, J. S. (2003), "Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection," *Journal of the American Statistical Association*, 98, 438–455.

Langaas, M., and Lindquist, B. (2005), "Estimating the Proportion of True Null Hypotheses, With Application to DNA Microarray Data," *Journal of the Royal Statistical Society*, Ser. B, 67, 555–572.

Lehmann, E., and Romano, J. (2005), "Generalizations of the Familywise Error Rate," *The Annals of Statistics*, 33, 1138–1154.

Owen, A. (2005), "Variance of the Number of False Discoveries," *Journal of the Royal Statistical Society*, Ser. B, 67, 411–426.

Pawitan, Y., Murthy, K., Michiels, S., and Ploner, A. (2005), "Bias in the Estimation of False Discovery Rate in Microarray Studies," *Bioinformatics*, 21, 3865–3872.

Qui, X., Brooks, A., Klebanov, L., and Yakovlev, A. (2005a), "The Effects of Normalization on the Correlation Structure of Microarray Data," *BMC Bioinformatics*, 6.

Qui, X., Klebanov, L., and Yakovlev, A. (2005b), "Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology in Microarray Data Analysis," *Statistical Applications in Genetics and Molecular Biology*, 4, paper 34.

Storey, J. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

Storey, J., Dai, J., and Leek, J. (2005), "The Optimal Discovery Procedure, II: Applications to Comparative Microarray Experiments," available at *http://www.bepress.com/uwbiostat/paper260*.

Storey, J., Taylor, J., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Ser. B, 66, 187–205.

van der Laan, M., and Hubbard, A. (2005), "Quantile Function–Based Null Distribution in Resampling-Based Multiple Testing," available at *http://www.bepress.com/ucbiostat/paper198*.

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.

Westfall, P. (1997), "Multiple Testing of General Contrasts Using Logical Constraints and Correlations," *Journal of the American Statistical Association*, 92, 299–306.

Westfall, P., and Young, S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*, New York: Wiley.