

# Correlation and Large-Scale Simultaneous Significance Testing

Bradley Efron

## Abstract

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 paper 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.

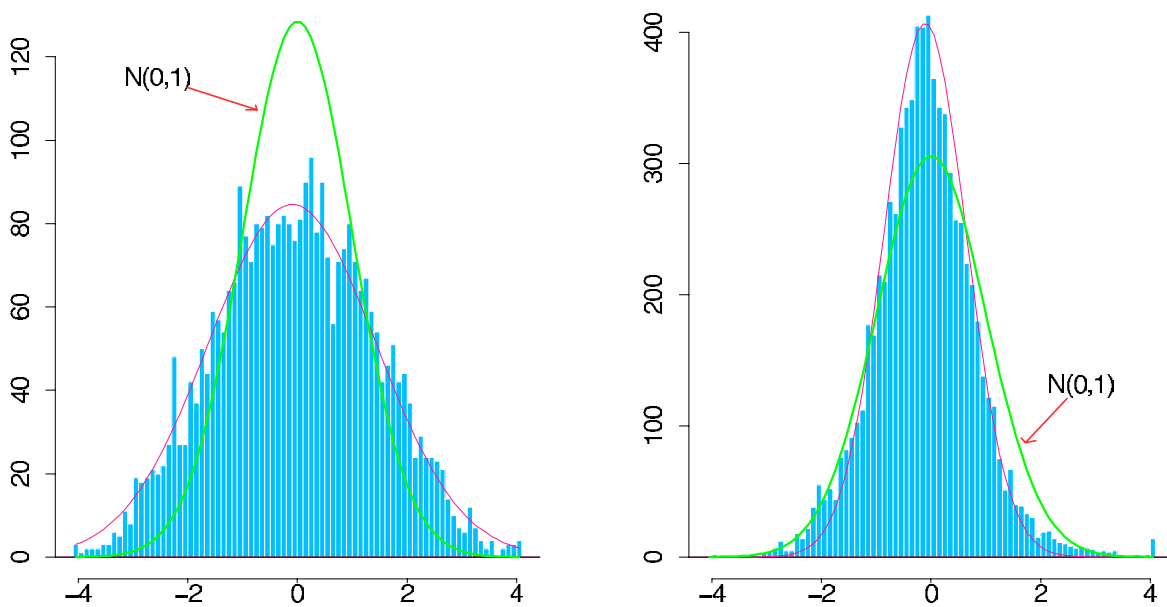
Key Words: Microarrays, Fdr, Empirical null, Correlated Processes

**1. Introduction** Modern computing machinery and improved scientific equipment have combined to revolutionize experimentation in fields such 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 paper 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, reported two-sample  $t$ -statistics “ $t_i$ ” comparing expression levels under two different conditions for  $N$  genes,  $N = 3226$  for the breast cancer study in the left panel, and  $N = 7680$  for the HIV experiment on the right;  $t_i$  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,

$$z_i = \Phi^{-1}(G_0(t_i)), \quad i = 1, 2, \dots, N, \quad (1.1)$$

where  $\Phi$  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’s  $t$  cdf with appropriate



**Figure 1:** Histograms of  $z$ -values from two microarray experiments . *Left panel:* breast cancer study, 3226 genes; *Right panel* HIV study, 7680 genes. *Heavy curves* indicate  $N(0, 1)$  theoretical null densities; *Light curves* indicate empirical null densities fit to central  $z$ -values, as in Efron (2004). The theoretical null distributions are too narrow in the left panel and too wide in the right. Both effects can be caused by correlations among the null  $z$ -values. Data from Hedenfalk et al. (2001), left panel, and van't Wout et al. (2003), right.

degrees of freedom for the HIV study, while a permutation method described in Section 4 provided  $G_0$  for the breast cancer experiment (also nearly a student's  $t$  cdf.) Assuming  $G_0$  is the correct null distribution for  $t_i$ , transformation (1.1) yields

$$z_i \sim N(0, 1) \tag{1.2}$$

for the null cases, called the *theoretical null* in what follows. Form (1.2) for the null distribution is convenient for general discussion, and can be achieved, or at least approximated, in most testing situations via transformations like (1.1).

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 non-null 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 paper concerns two related results:

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

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 pair-wise correlations, several million of them for Figure 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 (1.2) is individually correct for all the cases, while negative  $A$  narrows it, as on the right side of Figure 1. The random variable  $A$  is identically zero if the  $z_i$ 's are independent, but correlation allows  $A$  to vary around zero with positive variance  $\alpha^2$ , as summarized by the Theorem in Section 3, the more correlation the bigger  $\alpha$ . 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 is discussed in Section 5, particularly in terms of false discovery rates (FDR). 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 the use of 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,

$$\text{breast cancer} : N(-0.09, 1.55^2) \quad \text{HIV} : N(-0.11, 0.75^2). \quad (1.3)$$

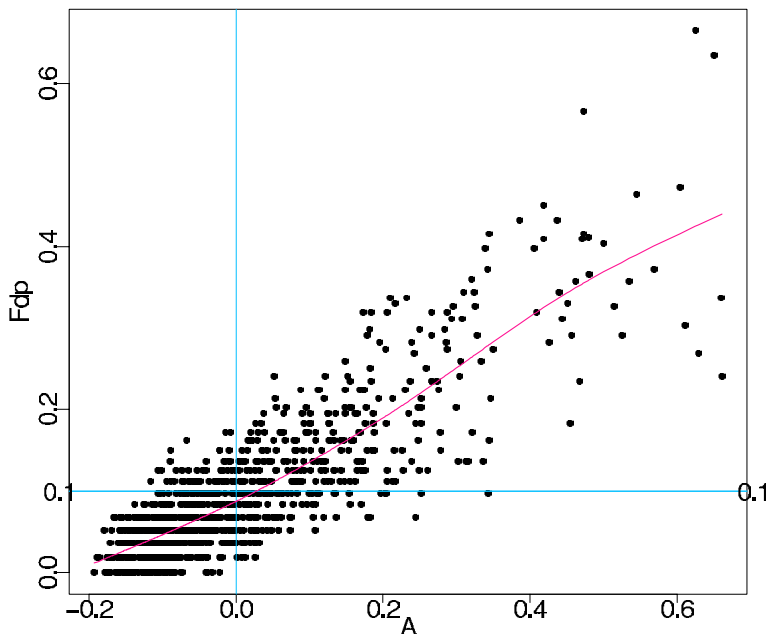
Looking ahead to Section 5, Figure 2 illustrates the disturbing effects of correlation on large-scale simultaneous inference. The simulation model involved  $N = 3000$  genes, with 95% obeying the theoretical null distribution  $z_i \sim N(0, 1)$  and the other 5% non-null,  $z_i \sim N(2.5, 1.25)$ ; the null  $z_i$ 's were correlated about the same amount as in the breast cancer data,  $\alpha = 0.15$ .

For each of a thousand replications, a standard Benjamini-Hochberg false discovery rate analysis (i.e. based on the theoretical  $N(0, 1)$  null distribution) was run at control level  $q = 0.10$ , and used to identify a set of "significant" genes. Each of the 1000 points in Figure 2 has coordinates

$$(A, \text{Fdp}), \quad (1.4)$$

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 0.096, close to the theoretical control level 0.10, but we see a strong dependence on the dispersion variate  $A$ : Fdp averaged 0.34 for the upper 5% of  $A$  realizations, and only 0.03 for the lower 5%. The true proportion of false discoveries reported by the “0.10” procedure easily varies by a factor of 10.



**Figure 2:** Benjamini-Hochberg FDR controlling procedure,  $q = 0.10$ , run for 1000 simulation trials; true false discovery proportion Fdp plotted versus dispersion variate  $A$ . Overall Fdp averaged 0.096, close to  $q$ , but with a strong dependence on  $A$ , as shown by smooth regression curve.

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, wide peaks positive ones. 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 (1.2) with an empirical null (1.3) automatically carries out the conditioning process, though in somewhat noisy fashion.

There is a lot at stake here. Table 1 shows the number of gene discoveries identified by the Benjamini-Hochberg two-sided procedure, FDR control level  $q = 0.10$ , for the two studies of Figure 1. The HIV results look much more dramatic using the empirical null distribution  $N(-0.11, 0.75^2)$ . In fact a null standard deviation of  $\sigma_0 = 0.75$  is quite believable given the

amount of correlation, while Figure 1 argues against the theoretical null. The breast cancer data has been used in the microarray literature to compare analysis techniques, under the presumption that better techniques will produce more discoveries, recently for instance in Pawitan et al. (2005) and Storey et al. (2005). Table 1 suggests caution in this interpretation, where using the empirical null negates any discoveries at all. (Efron (2004) gives reasons other than correlation for distrusting the theoretical null in large-scale testing problems, but only correlation effects will be discussed here.)

	breast cancer	HIV
Theoretical Null:	107	22
Empirical Null:	0	180

**Table 1:** Number of genes identified as significant discoveries by two-sided Benjamini-Hochberg procedure, 0.10 control level. *Top row* based on theoretical  $N(0, 1)$  null distribution; *Bottom row* using empirical null distribution (1.3).

Five pertinent references are discussed in what follows: Qui, Klebanov, and Yakovlev (2005), Qui et al. (2005), Owen (2005), Dudoit, van der Laan, and Pollard (2004), and Hsu (1993, 1996). Permutations and correlated  $z$ -values also play a role in Westfall and Young’s theory of adjusted  $p$ -values (1993), 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 paper.

**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,

$$z_i \sim N(0, 1) \text{ for } i = 1, 2, \dots, N, \tag{2.1}$$

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 non-null 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 7 BRCA1 and 8 BRCA2 women, each providing a microarray of expression levels on the same  $N = 3226$  genes. The usual two-sample  $t$ -statistic “ $t_i$ ” comparing BRCA2 versus BRCA1 for the 15 gene  $i$  expression levels gave  $z_i$  as in (1.1), with  $G_0$  nearly a student’s  $t$ -distribution with 13 degrees of freedom, see Section 4. Similarly, the HIV study compared four HIV positive patients versus four HIV negative controls,  $N = 7680$  genes per microarray, van’t Wout et al. (2003). In this case  $G_0$  was taken to be student’s  $t$  with 6 degrees of freedom. These data sets are discussed further in Efron (2004, 2005).

Let  $X$  represent the full data set, for example an  $N$  by  $n$  matrix for the breast cancer study, having  $N = 3226$  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 (1.1), with  $\mathbf{z}$  representing the vector of all  $N$   $z_i$ ’s. **Note** It is not necessary that the  $z_i$ ’s be obtained from  $t$ -tests, only that null distribution (2.1) can be achieved or approximated. For example, each of the original  $N$  cases might involve a separate linear regression, with the  $i$ th 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  $\Delta = 0.1$ , running from -4.1 to 4.1. We denote the count vector by  $\mathbf{y}$ ,

$$y_k = \#\{z_i \text{ in } k^{\text{th}} \text{ bin}\}, \quad k = 1, 2, \dots, K; \quad (2.2)$$

$\mathbf{y}$  is essentially the order statistic of  $\mathbf{z}$ , exactly so if we let  $\Delta \rightarrow 0$ . False Discovery Rate 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,

$$\mathcal{Z} = \cup_{k=1}^K \mathcal{Z}_k, \quad (2.3)$$

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

$$\pi_k(i) = Pr\{z_i \in \mathcal{Z}_k\}, \quad \pi_{k\cdot} = \frac{\sum_{i=1}^N \pi_k(i)}{N}, \quad (2.4)$$

and

$$\gamma_{k\ell}(i, j) = Pr\{z_i \in \mathcal{Z}_k \text{ and } z_j \in \mathcal{Z}_\ell\}, \quad \gamma_{k\ell\cdot} = \frac{\sum_{i \neq j} \gamma_{k\ell}(i, j)}{N(N-1)}. \quad (2.5)$$

Because of assumption (2.1) all the  $\pi_k(i)$  values are determined by  $\varphi(z)$ , the standard normal density, with Taylor approximation around centerpoint  $z[k]$

$$\pi_{k\cdot} = \pi_k(i) \doteq \Delta \cdot \varphi(z[k]) \quad (\varphi(z) = e^{-\frac{1}{2}z^2}/\sqrt{2\pi}). \quad (2.6)$$

The *expectation vector*  $\boldsymbol{\nu} = (v_1, v_2, \dots, v_K)'$  of  $\mathbf{y}$  is determined by (2.1),

$$\boldsymbol{\nu} = N\boldsymbol{\pi}_\cdot \doteq (\dots, N\Delta\varphi(z[k]), \dots)'. \quad (2.7)$$

Definitions (2.4-2.5) lead to a convenient expression for the covariance matrix of  $\mathbf{y}$ ;

*Lemma 1*

$$\text{Cov}(\mathbf{y}) = C_0 + C_1 \quad (2.8)$$

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

$$C_0 = \text{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}'/N = N[\text{diag}(\boldsymbol{\pi}_\cdot) - \boldsymbol{\pi}_\cdot\boldsymbol{\pi}_\cdot'], \quad (2.9)$$

and

$$C_1 = \left(1 - \frac{1}{N}\right) \text{diag}(\boldsymbol{\nu})\boldsymbol{\delta} \text{diag}(\boldsymbol{\nu}) \quad \text{with } \delta_{k\ell} = \gamma_{k\ell}/\pi_{k\cdot}\pi_{\ell\cdot} - 1. \quad (2.10)$$

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

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

$$\begin{aligned} E\{y_k y_\ell\} &= E\left\{ \sum_{i \neq j} I_k(i) I_\ell(j) \right\} = \sum_{i \neq j} \gamma_{k\ell}(i, j) \\ &= N(N-1)\gamma_{k\ell}. \end{aligned} \quad (2.11)$$

Then

$$\begin{aligned} \text{cov}(y_k, y_\ell) &= N(N-1)\gamma_{k\ell} - N^2\pi_{k\cdot}\pi_{\ell\cdot} \\ &= -N\pi_{k\cdot}\pi_{\ell\cdot} + N(N-1)(\gamma_{k\ell} - \pi_{k\cdot}\pi_{\ell\cdot}). \end{aligned} \quad (2.12)$$

Similarly,

$$\text{var}(y_k) = N(\pi_{k\cdot} - \pi_{k\cdot}^2) + N(N-1)(\gamma_{kk} - \pi_{k\cdot}^2), \quad (2.13)$$

verifying Lemma 1 ■

If  $z_i$  and  $z_j$  are independent in (2.1) then  $\gamma_{k\ell}(i, j) = \pi_k(i)\pi_\ell(j)$ ; independence of all  $z$ -values implies that all the elements of matrix  $\boldsymbol{\delta}$  in (2.10) are zero, leaving  $\text{cov}(\mathbf{y}) = C_0$ . Conversely, the amount of correlation between the  $z$ -values determines the size of  $\boldsymbol{\delta}$  and the increase of  $\text{cov}(\mathbf{y})$  above  $C_0$ .

To approximate  $\delta$ , we add the assumption of bivariate normality for any pair of  $z$ -values,  $\text{cov}(z_i, z_j) \equiv \rho_{ij}$ , so that as in (2.6),

$$\gamma_{k\ell}(i, j) \doteq \frac{\Delta^2}{2\pi\sqrt{1-\rho_{ij}^2}} \exp\left\{-\frac{1}{2}\frac{z[k]^2 - 2\rho_{ij}z[k]z[\ell] + z[\ell]^2}{1-\rho_{ij}^2}\right\}, \quad (2.14)$$

Letting  $g(\rho)$  indicate the empirical density of the  $N(N-1)$  correlations  $\rho_{ij}$ , and using (2.6)-(2.14) yields a useful approximation:

*Lemma 2* Under the bivariate normal approximation (2.14), the matrix  $\delta$  in (2.10) has elements

$$\delta_{k\ell} \doteq \int_{-1}^1 \left[ \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{\rho}{2(1-\rho^2)}\{2z[k]z[\ell] - \rho(z[k]^2 + z[\ell]^2)\}\right) - 1 \right] g(\rho) d\rho. \quad (2.15)$$

(This compares with Theorem 1 in Owen (2005); the assumptions there imply the bivariate normal condition (2.14).)

Application of Lemmas 1 and 2 requires an estimate 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 gave

$$\text{breast cancer} : \quad g(\rho) \sim N(0, 0.153^2) \quad (2.16)$$

for the breast cancer data, and, more roughly,

$$\text{HIV} : \quad g(\rho) \sim N(0, 0.42^2), \quad (2.17)$$

for the HIV study. It is no accident that  $g(\rho)$  has mean near zero; in both cases the data matrix  $X$  had its columns standardized to mean zero and variance one (but not quantile normalized), usual practice to negate “brightness” disparities between microarrays, see Bolstad et al. (2003) and Qui et al. (2005). This forces the sum of covariances, and, nearly, the sum of correlations, to be zero. The normality assumed in (2.16) is not crucial, see Remark B. Section 3 shows that the standard deviation 0.153 is the vital number. Remark E discusses what happens in the absence of standardization.

Approximation (2.16) indicates a substantial amount of global correlation among genes in the breast cancer study, and even more correlation for the HIV study (2.17). The five examples in Owen’s (2005) Table 1 had standard deviations of  $g(\rho)$  between 0.17 and 0.26, as did Qui et al.’s (2005) main example. It is not surprising that correlations of this magnitude



	Independent	$C_{\text{norm}}$	$C_{\text{perm}}$	Poisson
$sd(Y_0)$ :	26.4	<b>171.4</b>	176.0	176.4
$sd(Y_1)$ :	4.5	<b>16.1</b>	14.9	14.7
$\text{cor}(Y_0, Y_1)$ :	(-0.12)	<b>(-0.89)</b>	(-0.81)	(-0.90)

**Table 2:** Standard deviations and correlations for central and tail counts  $(Y_0, Y_1)$ , (2.19);  $z_i \sim N(0, 1)$   $i = 1, 2, \dots, 3226$ ; for  $z_i$ 's independent or  $z_i$ 's correlated as in  $C_{\text{norm}}$ , (2.18); also for permutation covariance  $C_{\text{perm}}$ , (4.2).  $C_{\text{norm}}$  and  $C_{\text{perm}}$  produce much larger standard deviations and much more negative correlations. "Poisson" calculated from Poisson approximation model (3.15).

can undercut standard inference techniques, the key message in Qui et al. The goal here, made explicit in Section 5, is to understand and correct correlational problems.

Substituting (2.16) into (2.15) and then into Lemma 1, gives estimated null hypotheses covariance matrix " $C_{\text{norm}}$ ",

$$\text{cov}(\mathbf{y}) = C_{\text{norm}} \tag{2.18}$$

for the breast cancer data. The correlation term  $C_1$  in  $\text{cov}(\mathbf{y}) = C_0 + C_1$ , (2.8), dominates the independence term  $C_0$ . As an informative example we will use later, define

$$Y_0 = \#\{z_i \in [-1, 1]\} \quad \text{and} \quad Y_1 = \#\{z_i \geq 2.5\}, \tag{2.19}$$

these being central and tail counts for a hypothetical null vector  $\mathbf{z}$  satisfying (2.1). Table 2 shows standard deviations and correlation for  $Y_0$  and  $Y_1$  if the  $z_i$ 's are independent, or if they are correlated such that  $\text{cov}(\mathbf{y}) = C_{\text{norm}}$ . (Table 2 can be computed from  $\text{cov}(\mathbf{y})$  since  $Y_0$  and  $Y_1$  are linear functions of  $\mathbf{y}$ ). Results for  $C_{\text{perm}}$ , the permutation estimate from Section 4, agree with  $C_{\text{norm}}$ . The cutoffs  $\pm 1$  and 2.5 in (2.19) were chosen for convenient exposition, and could just as well be replaced by similar values, say  $\pm 0.75$  and 3.0.

We see that gene correlations have a powerful effect on the counts that would be observed under null hypothesis (2.1): standard deviations are multiplied several fold (this being the main point in Owen (2005)), while the negative correlation between the central and tail counts is driven toward -1. Tail null counts play a crucial role in computing false discovery rates, 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

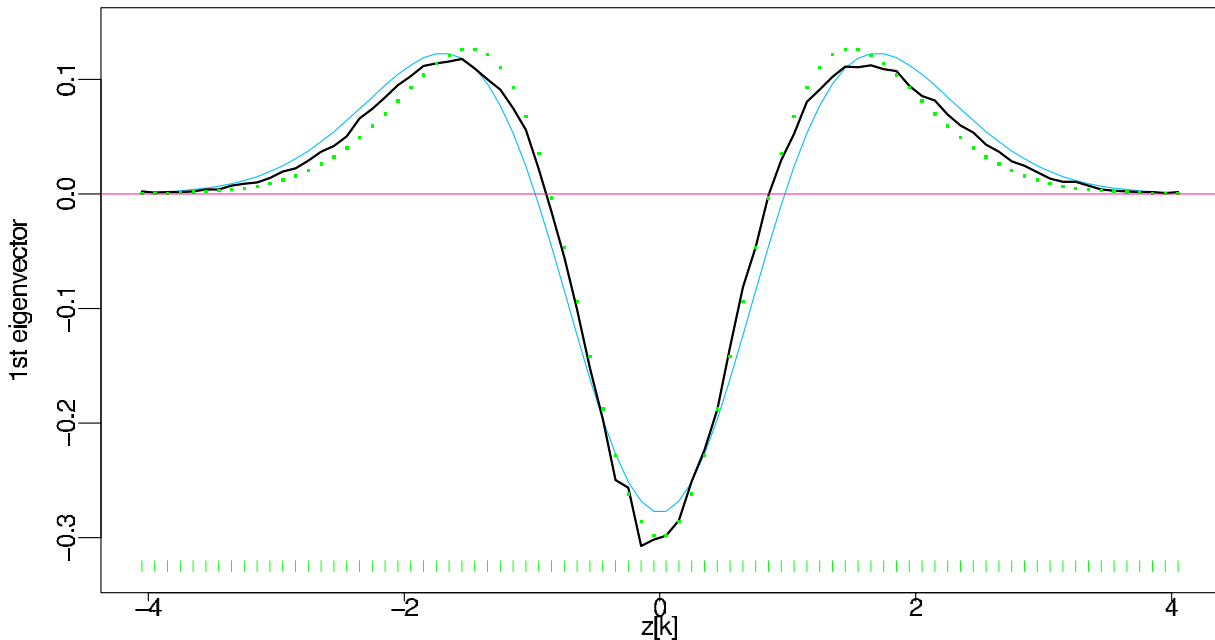
Lemmas 1 and 2 decompose the covariance matrix of the count

vector  $\mathbf{y}$  into an independence term  $C_0$  and an additional term  $C_1$  that accounts for correlation among the  $z$ -values, (2.8). This section presents a simple approximation to  $C_1$  in terms of its first eigenvector, which will be used 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  $C_{\text{norm}}$ , (2.16)-(2.18) the normal-theory estimate of  $\text{cov}(\mathbf{y})$  for the breast cancer data, while the jagged curve is the corresponding quantity for the permutation-based estimate  $C_{\text{perm}}$  of Section 4. The dots indicate the first eigenvector of  $C_{\text{norm}}$  for the HIV data, using (2.17). All three curves exhibit the same “wing-shaped” form. This will turn out to be proportional to

$$w(z) \equiv \varphi(z) \frac{z^2 - 1}{\sqrt{2}} = \varphi''(z) / \sqrt{2}, \quad (3.1)$$

where  $\varphi(z)$  is the standard normal density, (2.6).



**Figure 3:** First eigenvectors of three different estimates of  $\text{cov}(\mathbf{y})$ : *smooth curve* normal-theory estimate (2.18) for breast-cancer data; *dots* normal-theory estimate for HIV data (2.17); *jagged curve* permutation estimate for breast cancer data (4.2); The striking “wing-shaped” form is proportional to the second derivative of the standard normal density. Dashes indicate bin midpoints  $z[k]$ .

We use notation (2.3)-(2.6), and work in the discretized framework of Lemmas 1 and 2:  
*Lemma 3* Suppose  $g(\rho)$ , the correlation density in (2.15) has mean zero and standard

deviation

$$\alpha = \left[ \int_{-1}^1 \rho^2 g(\rho) d\rho \right]^{\frac{1}{2}}. \quad (3.2)$$

Then the matrix  $\boldsymbol{\delta}$  in (2.10) is approximated by the outer product

$$\boldsymbol{\delta} \doteq \alpha^2 \mathbf{q}\mathbf{q}' \quad \text{where} \quad q_k = \frac{z[k]^2 - 1}{\sqrt{2}}, \quad (3.3)$$

with  $z[k]$  the centerpoint of the  $k$ th histogram bin,  $k = 1, 2, \dots, K$ ; approximation (3.3) becomes exact as  $\alpha \rightarrow 0$ , with  $\boldsymbol{\delta}/\alpha^2 \rightarrow \mathbf{q}\mathbf{q}'$ ,

*Proof* Let  $R_{k\ell}(\rho)$  be the integrand in (2.15),

$$R_{k\ell}(\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp \left\{ \frac{\rho}{1-\rho^2} z[k]z[\ell] - \frac{1}{2} \frac{\rho^2}{1-\rho^2} (z[k]^2 + z[\ell]^2) \right\} - 1. \quad (3.4)$$

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

$$R_{k\ell}(\rho) \doteq \rho z[k]z[\ell] + \rho^2 q_k q_\ell; \quad (3.5)$$

then, since  $g(\rho)$  has mean 0,

$$\delta_{k\ell} = \int_{-1}^1 R_{k\ell}(\rho) g(\rho) d\rho \doteq \alpha^2 q_k q_\ell, \quad (3.6)$$

which is (3.3). ■

Combining the three lemmas yields a useful approximation for the null covariance matrix of the count vector  $\mathbf{y}$  under the bivariate normal assumptions of Section 2:

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

$$\text{cov}(\mathbf{y}) \doteq [\text{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}'/N] + \left(1 - \frac{1}{N}\right) (\alpha\mathbf{W})(\alpha\mathbf{W})' \quad (3.7)$$

Here  $\boldsymbol{\nu} = E\{\mathbf{y}\}$  as in (2.7), while  $\mathbf{W}$  has components

$$W_k = N\Delta\varphi(z[k]) \frac{z[k]^2 - 1}{\sqrt{2}} = N\Delta w(z[k]), \quad (3.8)$$

with  $w(\cdot)$  the wing-shaped function (3.1),  $N$  the number of cases, and  $\Delta$  the bin width.

The Theorem helps explain Figure 2: the second term in (3.7) dominates  $\text{cov}(\mathbf{y})$  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. first eigenvalue divided by sum of eigenvalues) for  $C_1$ ,

the correlation term in (2.8), and also for  $\text{cov}(\mathbf{y}) = C_0 + C_1$ , assuming  $g(\rho) \sim N(0, \alpha^2)$ . For the breast cancer value  $\alpha = 0.153$ , the proportions are 98% for  $C_1$ , (the crucial quantity for the accuracy of (3.7)), and 45% for  $C_{\text{norm}} = \text{cov}(\mathbf{y})$ .  $N = 3226$  in Table 2, but 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.

$\alpha$ :	0	0.05	0.10	0.15	0.20	0.25	0.30
$C_1$ :	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>0.98</b>	<b>0.97</b>	<b>0.95</b>	<b>0.90</b>
$C_{\text{norm}}$ :	.04	0.10	0.27	0.45	0.59	0.68	0.72

**Table 3:** Proportion of total variance explained by first eigenvector, as a function of  $\alpha$ ; for  $C_1$ , the correlation term in (2.8), and also for  $C_{\text{norm}}$ ; assuming  $g(\rho) \sim N(0, \alpha^2)$  and  $N = 3226$ . Proportions for  $C_1$  determine accuracy of approximation (3.7).

The Theorem *summarizes the effect of  $\mathbf{z}$ ’s entire correlation structure in a single parameter  $\alpha$* . This permits a relatively simple analysis of the inferential effects of correlation in what follows. Note: Hsu (1992, 1996) gives 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  $\mathbf{z} = (z_1, z_2, \dots, z_N)'$ , (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  $\mathbf{y} \sim Po(\mathbf{u})$  indicate a vector of independent Poisson variates,  $y_k \stackrel{\text{ind}}{\sim} Po(u_k)$  for  $k = 1, 2, \dots, K$ , while  $\mathbf{y} \sim (\boldsymbol{\nu}, \Gamma)$  denotes that vector  $\mathbf{y}$  has mean  $\boldsymbol{\nu}$  and covariance  $\Gamma$ .

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

$$N \sim Po(N_0), \tag{3.9}$$

with  $N_0 = 3226$  in the breast cancer study. This slightly simplifies (3.7), to

$$\text{cov}(\mathbf{y}) \doteq \text{diag}(\boldsymbol{\nu}) + (\alpha \mathbf{W})(\alpha \mathbf{W})', \tag{3.10}$$

with  $\boldsymbol{\nu}$  and  $\mathbf{W}$  as in (2.7), (3.8) except that  $N_0$  replaces  $N$ . If the  $z$ -values are independent then (3.9) makes the counts  $y_k$  independent Poisson variates,

$$\mathbf{y} \sim Po(\boldsymbol{\nu}), \tag{3.11}$$

agreeing with (3.10) at  $\alpha = 0$ .

A hierarchical model generalizes (3.11) to incorporate dependence. We assume that  $\mathbf{y}$  depends on a mean vector  $\mathbf{u}$ , itself random,

$$\mathbf{y}|\mathbf{u} \sim \text{Po}(\mathbf{u}) \quad \text{where} \quad \mathbf{u} \sim (\boldsymbol{\nu}, \Gamma), \quad (3.12)$$

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

$$\mathbf{y} \sim (\boldsymbol{\nu}, \text{diag}(\boldsymbol{\nu}) + \Gamma). \quad (3.13)$$

To match (3.10) we need to set

$$\Gamma = (\alpha \mathbf{W})(\alpha \mathbf{W})', \quad (3.14)$$

Formulas (3.13)-(3.14) suggest a hierarchical Poisson structure for the count vector  $\mathbf{y}$ :

$$\mathbf{y} \sim \text{Po}(\mathbf{u}) \text{ where } \mathbf{u} = \boldsymbol{\nu} + A\mathbf{W}, \quad \text{with } A \sim (0, \alpha^2). \quad (3.15)$$

If  $\alpha = 0$  this reduces to the independence case (3.11); otherwise the Poisson intensity vector  $\boldsymbol{\nu}$  is modified by the addition of an independent random multiple “ $A$ ” of  $\mathbf{W}$  having standard deviation  $\alpha$ .

Model (3.15) can only be an approximation since  $\mathbf{u}$  may have negative coordinates, but it nicely explains phenomena like the extreme negative correlations between  $Y_0$  and  $Y_1$  seen in Table 2: vector  $\mathbf{W}$  is negative in  $[-1, 1]$  and positive elsewhere, as in Figure 3, so positive  $A$  in (3.15) decreases the central counts and increases the tail counts, (2.19), the opposite happening when  $A$  is negative. (The “Poisson” column of Table 3 was calculated from model (3.15),  $\alpha = 0.153$  and  $N = 3226$ , except that the components of  $\mathbf{u}$  were truncated at zero.)

Pursuing model (3.15) more carefully, the mean vector  $\mathbf{u}$  turns out to be roughly proportional to a scaled normal density,

$$u_k \doteq \frac{N\Delta}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{z[k]^2}{2\sigma_0^2}\right\} \quad \text{with} \quad \sigma_0^2 = 1 + \sqrt{2} A, \quad (3.16)$$

see Remark D; (3.16) implies that positive  $A$  makes the counts behave in an overdispersed normal fashion compared to the theoretical  $N(0, 1)$  density, and conversely for  $A$  negative. This confirms the first main point of the Introduction: even if the null  $z$ -values are individually  $N(0, 1)$ , correlation can make the ensemble  $\mathbf{z}$  behave as  $N(0, \sigma_0^2)$ , with  $\sigma_0$  substantially

different from 1. Section 4 shows the same phenomenon happening 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 upon the assumption of bivariate normality for every pair of  $z$ -values. Permutation methods lead to a direct empirical estimate “ $C_{\text{perm}}$ ” for  $\text{cov}(\mathbf{y})$ . Carried out here for the breast cancer data,  $C_{\text{perm}}$  agrees well with the normal-theory estimate  $C_{\text{norm}}$ , (2.18), and lends support to the inferential theory of Section 5.

Let  $X$  represent the  $3226 \times 15$  matrix of observed expressions. Each row of  $X$ , that is each gene, provides a two-sample  $t$ -statistic comparing the 8 BRCA2 and 7 BRCA1 expressions,  $\mathbf{t}$  representing the vector of all 3226  $t$ -values. Repeating the computation after a random permutation of the columns of  $X$ , that is after a random division of the patients into groups of 7 and 8, yields permuted matrix  $X^*$  and  $t$ -vector  $\mathbf{t}^*$ . 1000 such permutations were employed to estimate a permutation null distribution  $G_0$  for the  $t$ -values (the empirical distribution of all  $1000 \cdot 3226$   $t_i$ '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 (1.1),  $z_i = \Phi^{-1}(G_0(t_i))$ .

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 (2.2),

$$X^*_{3226 \times 15} \rightarrow \mathbf{t}^*_{3226} \rightarrow \mathbf{z}^*_{3226} \rightarrow \mathbf{y}^*_{82}. \quad (4.1)$$

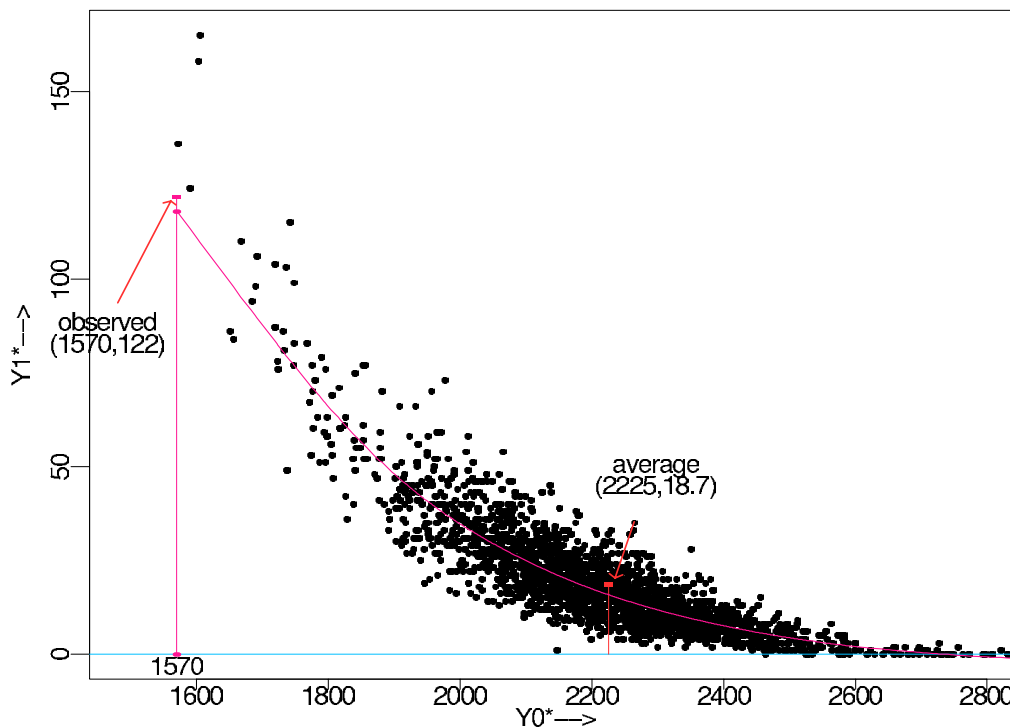
The sample covariance matrix of the 1000  $\mathbf{y}^*$ 's,

$$C_{\text{perm}} = \sum_{b=1}^{1000} (\mathbf{y}^{*b} - \mathbf{y}^{*\cdot})(\mathbf{y}^{*b} - \mathbf{y}^{*\cdot})' / 999 \quad [\mathbf{y}^{*\cdot} = \sum \mathbf{y}^{*b} / 1000], \quad (4.2)$$

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:* the matrix  $X$  used to form  $X^*$  in (4.1) had each gene's BRCA1 or BRCA2 average subtracted from its corresponding expression values in order to eliminate any genuine group differences from the permutation results.) Table 2 and Figure 3 demonstrate the similarity of  $C_{\text{perm}}$  to  $C_{\text{norm}}$ .

Each permutation vector  $\mathbf{z}^*$  gave central and tail counts  $(Y_0^*, Y_1^*)$  as in (2.19). Figure 4 plots  $Y_1^*$  versus  $Y_0^*$ , now for 4000 permutations. The unconditional average of  $Y_1^*$  is 18.7,

but that does not take into account the powerful covariate  $Y_0^*$ : in particular, for  $Y_0^*$  equaling the observed count 1570, a conditional expectation of about 118 is predicted. (Section 5 discusses why the observed central count might be so atypical of the permutation values in Figure 4.)



**Figure 4:** Central and tail counts  $(Y_0^*, Y_1^*)$  as in (2.19), for 4000 column-wise 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 1570. Smooth curve is fitted smoothing spline..

In the actual breast cancer data, 122 genes have  $z_i$  exceeding 2.5. Does this collection of 122 genes deserve to be reported as “mostly non-null”? The answer obviously depends on whether the expected number of null genes having  $z_i \geq 2.5$  is 18.7 or 118. Section 5 investigates this question, which bears on the second main point of the Introduction, 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,

$$\hat{\sigma}_0^* = 1/\Phi^{-1}\left(\frac{1 + Y_0^*/N}{2}\right); \quad (4.3)$$

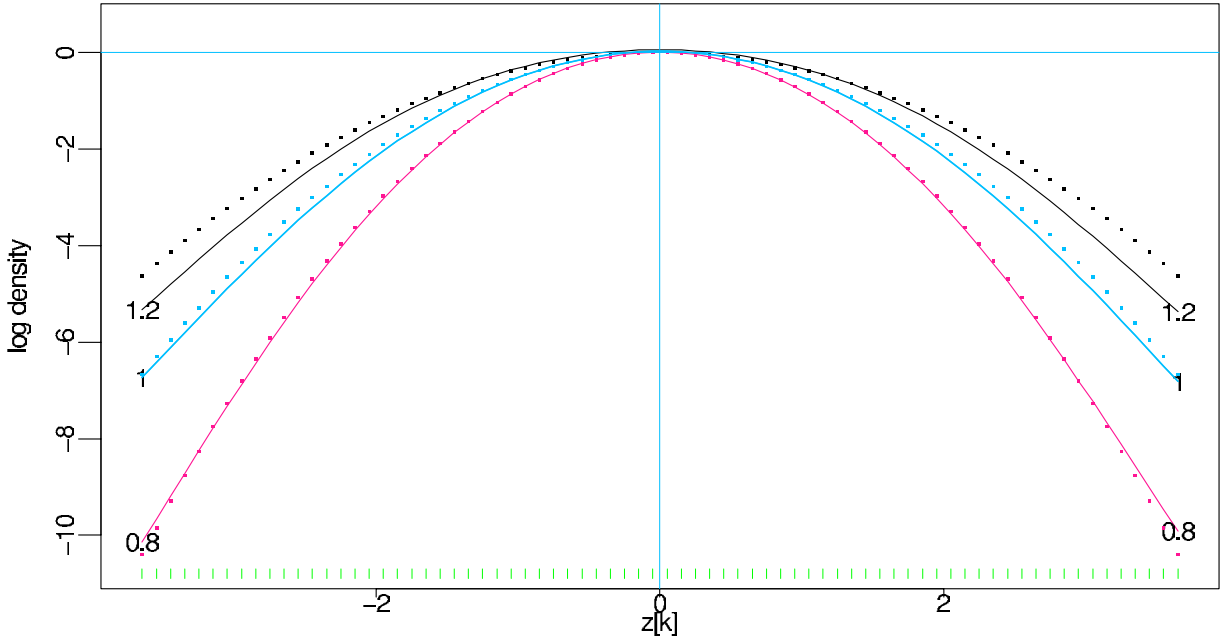
if we assume that the elements of  $\mathbf{z}^*$  are normally distributed with mean 0 and some variance

$\sigma_0^{*2}$ ,

$$z_i^* \sim N(0, \sigma_0^{*2}), \quad (4.4)$$

then  $E\{Y_0^*\} = N \cdot [2\Phi(1/\sigma_0^*) - 1]$  so (4.3) is a method of moment estimate for  $\sigma_0^*$  that depends only on the central count  $Y_0^*$ .

In Figure 5 the 4000 permutation  $\mathbf{y}^*$  vectors have been averaged in groups having about the same central spread  $\hat{\sigma}_0^*$ , (4.3); curve “1.2” is the log of the average of those  $\mathbf{y}^*$  vectors having  $1.15 \leq \hat{\sigma}_0^* \leq 1.25$ , etc. To first order the curves fall off as  $-z^2/(2\sigma_0^2)$ , with  $\sigma_0 = 1.2, 1, 0.8$ , i.e. as the log of a  $N(0, \sigma_0^2)$  density. This agrees with theoretical result (3.16), 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 conversely*. Section 5 discusses how this phenomenon affects simultaneous significance testing, particularly false discovery rate methods.



**Figure 5:** Log average densities for the 4000  $\mathbf{y}^*$  vectors contributing to Figure 3; “1.2” graphs  $\log \mathbf{y}_k^*$ , from average of  $\mathbf{y}^*$ 's having  $\hat{\sigma}_0^*$  in  $[1.15, 1.25]$ , versus  $z[k]$ , etc. Dotted curves are  $-.5z[k]^2/\sigma_0^2$  for  $\sigma_0^2 = 1.2, 1.0, 0.8$ .

**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 false discovery rates and other simultaneous testing techniques, as well as methods for correcting their inferences.

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

$$Y(x) = \#\{\text{null } z_i \geq x\} \quad \text{and} \quad T(x) = \#\{z_i \geq x\}. \quad (5.1)$$

$Y_1$  in (2.19) is  $Y(2.5)$  in this notation, and  $T(2.5) = 122$  for the breast cancer study. Lehmann and Romano (2005) define the *false discovery proportion* to be

$$\text{Fdp}(x) = Y(x)/T(x), \quad (5.2)$$

assuming here that we are searching for “discoveries” only in the right tail. If  $\text{Fdp}(x)$  were known (but not the identity of the null cases), say  $\text{Fdp}(2.5) = 20/122 = 0.16$ , 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, and likewise  $\text{Fdp}(x)$ . A useful tactic is to replace  $Y(x)$  by its expectation, as in Benjamini and Hochberg (1995), giving an estimated false discovery rate defined as

$$\text{Fdr}(x) = E\{Y(x)\}/T(x). \quad (5.3)$$

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

$$\text{FDR}(x) = E\{Y(x)/T(x)\}, \quad (5.4)$$

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

Our calculations will focus on  $\text{Fdr}(x)$ , (5.3), an observable ratio that is important and useful in its own right.  $\text{Fdr}(x)$  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”. Since  $T(x)$  is observable,  $\text{Fdr}(x)$  has intuitive interpretation as the expected proportion of null cases among those having  $z_i \geq x$ .

Formula (5.3) 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 (2.1) where the theoretical null gives

$$E\{Y(x)\} = N \cdot \bar{\Phi}(x) \quad [\bar{\Phi}(x) \equiv 1 - \Phi(x)]. \quad (5.5)$$

Improvements on (5.5) are possible via estimation of “ $p_o$ ”, the proportion of null cases, Langaas et al. (2005), Storey et al. (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, (5.5) is a good starting point for the discussion of correlation effects.

The hierarchical structure (3.15) gives conditional expectation

$$E\{\mathbf{y}|A\} = \boldsymbol{\nu} + A\mathbf{W} \quad (5.6)$$

(using only  $E\{\mathbf{y}|\mathbf{u}\} = \mathbf{u}$ , not the full Poisson assumptions). Letting the bin width  $\Delta \rightarrow 0$  in (2.7) and (3.8) produces a conditional version of (5.6), Remark H,

$$E\{Y(x)|A\} = N\bar{\Phi}(x) \left[ 1 + A \frac{x\varphi(x)}{\sqrt{2} \bar{\Phi}(x)} \right]. \quad (5.7)$$

The term multiplying  $A$  equals 4.99 at  $x = 2.5$ , giving conditional expectation  $N\bar{\Phi}(x) \cdot (1.75)$  for  $A = 0.15$ , and  $N\bar{\Phi}(x) \cdot (0.25)$  for  $A = -0.15$ .

Even such relatively modest values of  $A$  greatly affect the *conditional Fdr*

$$\text{Fdr}(x|A) = E\{Y(x)|A\}/T(x), \quad (5.8)$$

which can be expressed as

$$\text{Fdr}(x|A) = \text{Fdr}_0(x) \left[ 1 + A \frac{x\varphi(x)}{\sqrt{2} \bar{\Phi}(x)} \right], \quad (5.9)$$

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  $x = 2.5$ ,  $\text{Fdr}(x|A)$  varies by a factor of seven as  $A$  ranges from -0.15 to +0.15. A principal point of this paper is that conditional Fdr estimates are available in situations like those of Figure 1, while the unconditional estimates can produce grossly misleading results, as in Figure 2.

The idea in what follows is that “ $A$ ” in (5.9), 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 (2.19), for  $x_0 > 0$  let

$$Y_0 = \#\{z_i \in [-x_0, x_0]\}. \quad (5.10)$$

and define

$$P_0 = 2\Phi(x_0) - 1 \quad \text{and} \quad Q_0 = \sqrt{2}x_0\varphi(x_0). \quad (5.11)$$

Then (5.6) and definitions (2.7), (3.8) give  $E\{Y_0|A\} \doteq N[P_0 - AQ_0]$ , suggesting

$$\hat{A} = \frac{P_0 - \hat{P}_0}{Q_0} \quad [\hat{P}_0 = Y_0/N] \quad (5.12)$$

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

$$\text{breast cancer} : \hat{A} = 0.57 \pm 0.04, \quad \text{HIV} : \hat{A} = -0.21 \pm .03 \quad (5.13)$$

for our two examples. For the breast cancer data, (5.13) implies  $E\{Y(2.5)|\hat{A}\} = 77$  and  $\text{Fdr}(2.5|\hat{A}) = 77/122 = 0.63$ , underestimates according to our later calculations.

Permutation methods permit model-free estimates of the conditional Fdr, as in Figure 4, which suggests  $E\{Y(2.5)|Y_0\} \doteq 118$ , with corresponding estimated Fdr  $118/122 = 0.97$ .

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  $2 \times 2$  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  $\hat{A}$  from (5.12) into (5.9) to obtain conditional Fdr estimate

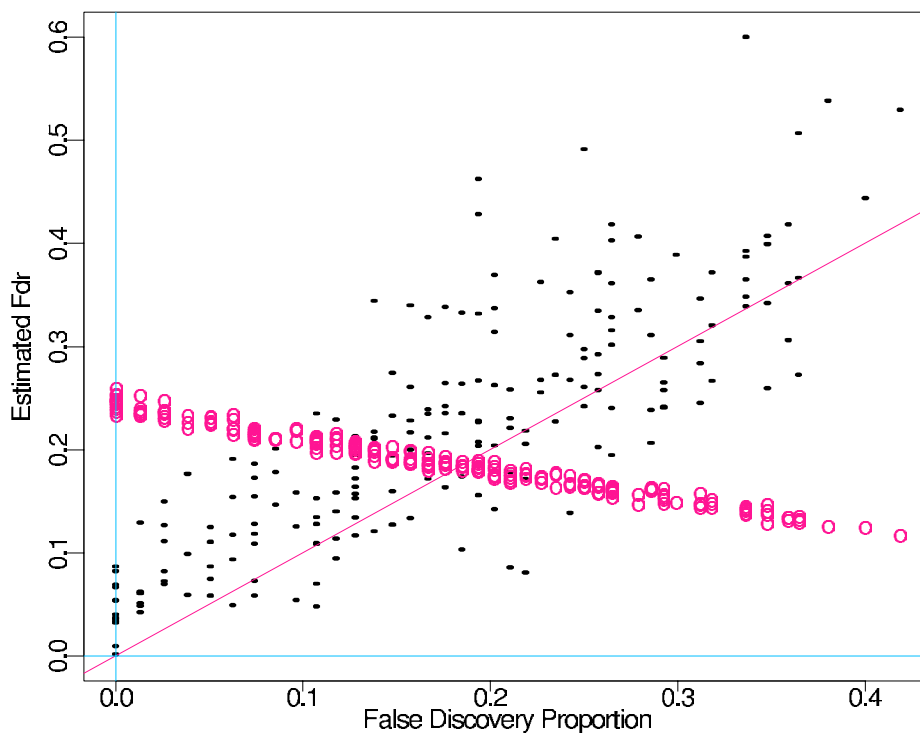
$$\text{Fdr}(x|\hat{A}) = \text{Fdr}_0(x) \left[ 1 + \hat{A} \frac{x\varphi(x)}{\sqrt{2}\Phi(x)} \right], \quad (5.14)$$

where  $\text{Fdr}_0$  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 in order 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 involved 200 trials, each with  $N = 3000$   $z$ -values, proportion  $p_0 = 0.95$  null. The null counts were generated from the Poisson model (3.15) with  $\alpha = 0.15$ , while the 150 non-null  $z$ 's followed a  $N(2.5, 1.25)$  distribution.

For each of the 200 trials, conditional and unconditional estimates of the False Discovery Proportion  $\text{Fdp}(x), x = 2.5$ , definition (5.2), are plotted versus  $\text{Fdp}(x)$ . 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, though with a considerable amount of noise. (Improved forms of (5.14) 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 (1.3.)



**Figure 6:** Simulation experiment comparing conditional Fdr estimates, solid points, with unconditional estimates, open circles;  $N = 3000$ ,  $p_o = 0.95$ . Null counts generated as in (3.15),  $\alpha = 0.15$ ; non-null counts from  $z \sim N(2.5, 1.25)$ . Horizontal axis is actual False Discovery Proportion  $Fdp(x)$ ,  $x = 2.5$ , for each of the 200 trials. The unconditional estimate, which is based on the theoretical null distribution, declines as actual Fdp increases.

Figure 6 supports the second main claim of the Introduction: 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 false discovery proportions.

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

Other factors besides  $z$ -value correlations can affect the null distribution. Efron (2004, 2006) suggests two other possibilities that support the use of empirical nulls: unidentified

covariates in an observational study, and correlations *across* microarrays. In fact, Figure 4 makes it seem 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 data sets 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 (3.15), can dominate variability as it does in Table 2. Qui, Klebanov, and Yakovlev (2005) emphasize the harmful effect of correlation, using it as an argument against empirical Bayes microarray analyses such as those in Storey (2002) or Efron et al. (2001). Their arguments could be used just as well against all other popular microarray analysis techniques.

The results presented in this paper are more optimistic. The correlation effect, though perhaps very large, is shown to manifest itself via 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 bigger 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 Section 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 might new array designs that incorporate greater gene duplication. Purely statistical improvements can also reduce correlations, for instance by more extensive standardization techniques as in Qui, Brooks, Klebanov, and Yakovlev (2005). None of this will help, however, 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 his coauthors, as in van der Laan and Hudson (2005) and Dudoit, van der Laan, and Pollard (2004), have developed a comprehensive resampling theory for correlated testing problems. This theory is much more general than the one presented here. However the 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 i.i.d.  $N$ -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, having  $n \ll N$ , raise legitimate concerns about the relevance of  $n \rightarrow \infty$  asymptotics.

Not all large-scale testing situations involve microarrays. Correlation may be less of a problem in other scientific venues like fMRI imaging or time of flight spectroscopy. In any case it seems worthwhile to obtain some overall measure of correlation such as  $\alpha$ , (3.2). Large  $\alpha$ 's suggest the use of correlation-resistant analysis techniques like the Fdr/empirical null combination.

## 7. Remarks

**A** (*Section 2*) Empirical correlation distributions (2.16)-(2.17) were obtained from the row-wise correlations of the original data matrix  $X$ , with  $X$   $3226 \times 15$  and  $7680 \times 8$  in our two examples. Let  $\hat{\rho}_{ij}$  be the sample correlation between rows  $i$  and  $j$  of  $X$ , after first subtracting off each gene's average response within each treatment group (in order to nullify any genuine treatment differences);  $g(\rho)$  is essentially the empirical distribution of all  $N(N-1)/2$   $\hat{\rho}_{ij}$  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 is needed to remove the variability added by sampling error. This was done by transforming to

$$\hat{\xi}_{ij} = \frac{1}{2} \log \frac{1 + \hat{\rho}_{ij}}{1 - \hat{\rho}_{ij}}; \quad (7.1)$$

assuming a translation model  $\hat{\xi}_{ij} = \xi_{ij} + \epsilon$  on this scale; estimating the distribution of  $\epsilon$  by repeating the calculations beginning with matrices  $X^*$  in which the entries within columns of  $X$  were independently permuted; inferring the  $\xi$  distribution by deconvolution; and retransforming back to the  $\rho$  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 (1.1).

**B** (*Section 2*) Owen's examples support normality for  $g(\rho)$  as in (2.16), but a wide range of other distributions fit reasonably well. Table 4 shows the best distribution supported on just three  $\rho$  values. "Best" here is defined in terms of numerically minimizing a chi-square discrepancy between the empirical distribution of the  $\hat{\xi}_{ij}$  values and the model  $\hat{\xi} = \xi + \epsilon$ , taking the  $\epsilon$  distribution as above. The solution turned out to have mean 0 and standard deviation  $\alpha = 0.153$ , as in (2.16), and gave about the same estimate of  $\text{cov}(\mathbf{y})$  as (2.18).

**C** (*Section 2*) The amount of gene-wise correlation represented by (2.16) is enormous.

$\rho$ :	-0.250	0.000	0.444
$g(\rho)$ :	0.131	0.793	0.076

**Table 4:** Best 3-point distribution estimate for the breast cancer correlation density  $g(\rho)$ .

For comparison, suppose that there were actually 10 equal sized groups of genes, with independence across groups but  $\rho_{ij} = 0.50$  for all genes within groups. After standardization of  $X$  this gives  $\alpha = 0.15$ , about the same as (2.16).

**D (Section 3)** The components of  $\mathbf{u} = \boldsymbol{\nu} + A\mathbf{W}$  in (3.15) are

$$u_k \doteq N\Delta f_A(z[k]) \quad \text{where} \quad f_A(z) = \varphi(z) \cdot [1 + Aq(z)], \quad (7.2)$$

$q(z) = (z^2 - 1)/\sqrt{2}$ , (2.7), (3.8);  $f_A(z)$  is symmetric around zero, with even moments easily obtained from Hermite polynomial calculations,

$$\int_{-\infty}^{\infty} f_A(z) dz = 1, \quad \int_{-\infty}^{\infty} f_A(z) z^2 dz = 1 + \sqrt{2}A, \quad \int_{-\infty}^{\infty} f_A(z) z^4 dz = 3 + \frac{12}{\sqrt{2}} A. \quad (7.3)$$

This supports approximation (3.16), which can be improved upon using higher order Edgeworth terms.

**E (Section 3)** The vectors  $\boldsymbol{\nu}$  and  $\mathbf{W}$  in (3.15) relate to the zero<sup>th</sup> and second Hermite polynomials. Standardization of the columns of  $X$  suppresses the first polynomial, with important consequences here. Without standardization, the first eigenvector of  $\text{cov}(\mathbf{y})$ , divided by  $\varphi(z)$ , may be proportional to  $z$ , the first polynomial, instead of the second polynomial  $z^2 - 1$ .

**F (Section 4)** The permutation calculations at the beginning of Section 4 provided 1000  $z_i^*$  values for each index  $i$ , after which the empirical distribution of all  $3226 \times 1000$  values provided the “permutation null”  $G_0$ . This kind of calculation can ignore correlation among the  $z_i$ ’s since it only depends 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 gene-wise correlations, but applied toward different purposes than in this paper. Their “step-down max-T” algorithm gave results similar to using a  $N(0, 1)$  null for the breast cancer and HIV studies.

**G (Section 4)** Permutation methods produced unstable results for the HIV data. Partly 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 7680 genes gave (2.17).

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

$$\theta_m = \sum_k m_k u_k / N \quad \text{and} \quad \hat{\theta}_m = \sum_k m_k y_k / N \quad (7.4)$$

in Poisson model (3.15). It is convenient to work with a continuous version of  $\mathbf{m}$ , say  $m(z)$  where  $m(z[k]) = m_k$ . Letting

$$P_m = \int_{-\infty}^{\infty} m(z) \varphi(z) dz \quad \text{and} \quad Q_m = - \int_{-\infty}^{\infty} m(z) q(z) \varphi(z) dz \quad (7.5)$$

gives

$$\theta_m \doteq \int_{-\infty}^{\infty} m(z) f_A(z) dz = P_m - A Q_m \quad (7.6)$$

as in (7.2). If  $m(z)$  is the indicator function of  $(x, \infty)$  then (7.6) becomes (5.7).

Since  $E\{\hat{\theta}_m | A\} = \theta_m$ , (7.6) suggests

$$\hat{A} = \frac{P_m - \hat{\theta}_m}{Q_m} \quad (7.7)$$

as a method of moments estimator for  $A$ ; (5.12) is (7.7) where  $m(z)$  is the indicator function of  $(-x_0, x_0)$ , Model (3.15) then yields  $\text{var}\{\hat{\theta}_m | A\} \doteq \int f_A(z) m(z)^2 dz / N$  and

$$\text{var}\{\hat{A}_m | A\} \doteq \frac{1}{N} \frac{\int_{-\infty}^{\infty} f_A(z) m(z)^2 dz}{\left(\int_{-\infty}^{\infty} \varphi(z) q(z) m(z) dz\right)^2}. \quad (7.8)$$

This formula, with  $A$  equaling 0.57 and -0.21 for the two studies, gave the standard errors in (5.13).

Suppose we wish to minimize (7.8) 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, the minimum variance equaling  $[N \int_{-x_0}^{x_0} \varphi(z) q(z)^2 dz]^{-1}$ . For  $x_0 = 1$  this gives  $\text{var}\{\hat{A}_m | A = 0\} = 5.03/N$ , compared to  $1/N$  for  $x_0 = \infty$ , (the ideal choice but an unallowable one given the possibility of biasing  $\hat{A}_m$  with non-null data). Taking  $m(z)$  as the indicator of  $(-1, 1)$  gives  $5.83/N$ . Reducing  $x_0$  to 0.80 provides slightly smaller variance when  $m(z)$  is the indicator,  $5.36/N$ . The more important point is that *conditional* variance estimates, as in (7.8), are both convenient and appropriate for the calculations here.



**Acknowledgment** The author is very grateful to Trevor Hastie and Rob Tibshirani for suggesting the applicability of permutation methods to correlation calculations. I gratefully appreciate the support from National Institute of Health grant 8R01 EB002784 and National Science Foundation grant DMS0505673.

### References

- Benjamini Y and Hochberg Y, (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J.R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289-300.
- Bolstad B, Irizarry R, Astrand M, Speed T, (2003), “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”, *Bioinformatics* **19** (2):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”, *Stat. Application in Genetics and Molecular Biology*, **3**, No. 1, Article 13.
- Efron B, Tibshirani R, Storey J and Tusher V, (2001), Empirical Bayes analysis of a microarray experiment, *J. Amer. Statist. Assoc.* **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, (2006), “Size, power, and false discovery rates”,  
<http://www-stat.stanford.edu/~brad/papers/Size.pdf>.
- Ge Y, Dudoit S, 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 Engl. Jour. Medicine* **344**, 539-48.
- Hsu, JC (1992), “The factor analytic approach to simultaneous inference in the general linear model”, *Journal of Graphical and Computational Statistics* **1**:151-168.
- Hsu, JC (1996), *Multiple Comparisons: Theory and Methods*. Chapman and Hall, London.

- Ishwaran H and Rao JS (2003), “Detecting differentially expressed genes in microarrays using Bayesian model selection”, *JASA* **98**, 438-55.
- Lanaas M and Lindquist B (2005), “Estimating the proportion of true null hypotheses, with application to DNA microarray data”, *JRSS-B* **67**, 555-72.
- Lehmann E and Roman J (2005), “Generalizations of the Familywise Error Rate”, *Annals Stat.* **33**, 1138-1154.
- Owen A, (2005), “Variance of the number of false discoveries”, *JRSS-B* **67**, 411-26.
- Pawitan Y, Murthy K, Michiels S, and Ploner A (2005), “Bias in the estimation of false discovery rate in microarray studies”, *Bioinformatics* **21**, 3865-72.
- Qui, X, Klebanov L, Yakovlev A, (2005), “Correlation between gene expression levels and limitations of the empirical Bayes methodology in microarray data analysis”, *Statistical Applications in Genetics and Molecular Biology* **4**, Issue 1, Paper 34.
- Qui X, Brooks A, Klebanov L, Yakovlev A, (2005), “The effects of normalization on the correlation structure of microarray data”, *BMC Bioinformatics* **6:120** {b} 2005-05-16.
- Storey J, (2002), “A direct approach to false discovery rates”, *JRSS-B* **64**, 479-498.
- Storey J, Taylor J, Siegmund D, (2004), “Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach”, *JRSS-B* **66**, 187-205.
- Storey J, Dai J, and Leek J (2005), “The optimal discovery procedure II: applications to comparative microarray experiments”, <http://www.bepress.com/uwbiostat/paper260>.
- van der Laan, M and Hubbard A (2005), “Quantile function based null distribution in resampling based multiple testing”, <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<sup>4</sup>+ T-Cell lines”, *Journal of Virology* **77**, 1392-1402.
- Westfall P and Young S, (1993), *Resampling-based multiple testing: examples and methods for p-value adjustments*. Wiley, New York.
- Westfall P (1997), “Multiple testing of general contrasts using logical constraints and correlations”, *JASA* **92**, 299-306.