Empirical Bayes Estimates for Large-Scale Prediction Problems

Bradley Efron[*†]

Abstract

Classical prediction methods such as Fisher's linear discriminant function were designed for small-scale problems, where the number of predictors N is much smaller than the number of observations n. Modern scientific devices often reverse this situation. A microarray analysis, for example, might include $n = 100$ subjects measured on $N = 10,000$ genes, each of which is a potential predictor. This paper proposes an empirical Bayes approach to large-scale prediction, where the optimum Bayes prediction rule is estimated employing the data from all the predictors. Microarray examples are used to illustrate the method. The results show a close connection with the *shrunken centroids* algorithm of Tibshirani et al. (2002), a frequentist regularization approach to large-scale prediction, and also with false discovery rate theory.

Keywords: microarray prediction, empirical Bayes, shrunken centroids, effect size estimation, correlated predictors, local fdr.

1 Introduction

An important class of prediction problems begins with the observation of n independent vectors,

$$(\boldsymbol{x}_j, y_j) \qquad j = 1, 2, \ldots, n. \tag{1.1}$$

Here $\boldsymbol{x}_j$ is a N-vector of predictors, while y_j is a real-valued response, taken to be dichotomous in most of what follows. For example, $\boldsymbol{x}_j$ might include age, height, weight, gender, etc. for person j, while y_j indicates whether or not that person later developed cancer. Given a newly observed N-vector $\boldsymbol{X}$, we would like to predict its corresponding Y value. Our task is to use the "training data" (1.1) to construct an effective prediction rule.

Classic prediction methods, such as Fisher's linear discriminant function, were fashioned for problems where N is much smaller than n, that is, where the number of predictors is less than the number of training cases. Current high-throughput scientific technology tends to produce just the opposite situation, with $N \gg n$; modern equipment may permit thousands of measurements on a single individual, but recruiting new subjects remains as difficult as ever.

Microarrays offer the iconic example. Here $\boldsymbol{x}_j$ is a vector of genetic expression measurements on subject j, one for each of N genes, where N is typically several thousand. In the *prostate cancer data* (Singh et al., 2002) we will use for motivation, there are $N = 6033$ genes measured

[*] *Department of Statistics, Stanford University*

[†] This work was supported in part by NIH grant 8R01 EB002784 and NSF grant DMS0505673.

on each of $n = 102$ men, $n_1 = 50$ healthy controls and $n_2 = 52$ prostate cancer patients. Given a new microarray measuring the same 6033 genes, we would like to predict whether or not that man develops prostate cancer.

Let t_i be the two-sample t-statistic comparing sick versus healthy subjects for gene i,

$$t_i = c_0 \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\hat{\sigma}_i} \qquad \left(c_0 = \sqrt{n_1 n_2 / n}\right), \tag{1.2}$$

where $\bar{x}_{i1}$ and $\bar{x}_{i2}$ are the mean expression levels on gene i for the healthy and sick subjects, and $\hat{\sigma}_i$ is the usual pooled estimate of standard deviation. For easier discussion later, we transform the t_i's to a normal scale,

$$z_i = \Phi^{-1}\left(F_{n-2}(t_i)\right), \tag{1.3}$$

with Φ and F_{n-2} the standard normal and t_{n-2} cumulative distribution functions (cdf), so that under the classical null hypothesis, z_i has a standard normal distribution,

$$H_0 : z_i \sim \mathcal{N}(0, 1). \tag{1.4}$$

Figure 1 shows the histogram of all 6033 z-values. The theoretical $\mathcal{N}(0, 1)$ null distribution fits the center of the histogram reasonably well, which makes sense since, presumably, most of the N genes have nothing to do with prostate cancer. However the histogram's heavy tails suggest some "non-null" genes that express themselves differently in sick and healthy subjects, and those are the ones that should be useful for prediction. Just how to fashion a prediction rule from them is the subject of this paper.

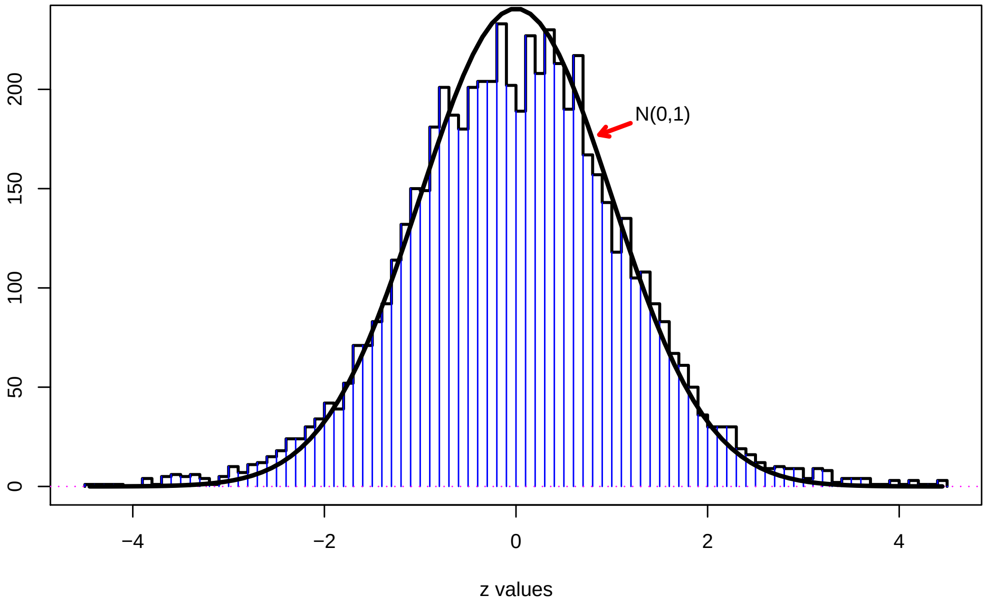

Figure 1: 6033 z-values from the prostate cancer study (Singh et al., 2002). A standard $\mathcal{N}(0, 1)$ density fits the histogram center, while the heavy tails indicate the presence of non-null genes that may be useful for prediction.

Large-scale prediction problems suffer from a surfeit of possible predictors, 6033 of them in this case, most of which are useless. Even the genuinely non-null cases appear to us in exaggerated form. *Selection bias*, the fact that we can only identify interesting possible predictors at the extremes of the N cases, means that an observed value at say, $z_i = 4$, probably corresponds to a true effect considerably nearer the null hypothesis.

This paper uses empirical Bayes methods both to select useful predictors and to undo selection bias in the evaluation of their predictive power. It was suggested by the "shrunken centroids" method of Tibshirani et al. (2002), described in Section 2.

A simple model is introduced in Section 2, which, if we knew the parameter values, would lead to an optimum prediction rule. Section 3 discusses Bayes estimation of the optimum rule, using a model of Brown (1971) and Stein (1981) to assist the calculations (and showing a connection with the theory of local false discovery rates). An empirical Bayes algorithm for approximating the Bayes solution is developed in Section 4. Section 5 modifies the empirical Bayes algorithm to allow for correlation among the predictors. A different problem is considered in Section 6: the estimation of effect sizes for those cases found to be non-null, where our empirical Bayes approach provides an alternative to the False Coverage Rate theory of Benjamini & Yekutieli (2005). Most of the paper concerns dichotomous responses y_j, but the results are extended to general response variables, for example survival times, in Section 7. Section 8 concludes the paper with Remarks that expand on some of the technical points and ideas.

A healthy literature on large-scale prediction has grown up around innovative computer-intensive techniques such as support vector machines, lasso and ridge regression regularization methods, the singular value decomposition and sparse data representation. Chapter 18 of Hastie et al. (2008) provides a nice overview. A main goal here, besides presenting some new methodology, is to trace the inferential connections between Bayesian theory, regularization methods like shrunken centroids, false discovery rates, and large-scale prediction.

2 A Simple Model

Motivation for our empirical Bayes prediction rules comes from a simple idealized probability model for a vector of predictors $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$. We assume that the individual predictors X_i are independently normal, with (location, scale) parameters (μ_i, σ_i), and with possibly different expectations in the two subject classes,

$$\frac{X_i - \mu_i}{\sigma_i} \overset{\text{ind}}{\sim} \mathcal{N}\left(\pm\frac{\delta_i}{2c_0}, 1\right) \quad \begin{cases} \text{"}-\text{"} & \text{healthy class} \\ \text{"}+\text{"} & \text{sick class} \end{cases}, \tag{2.1}$$

with $c_0 = (n_1 n_2/n)^{1/2}$ as in (1.2). (Here the classes have been labeled 'healthy' and 'sick' in deference to the prostate example. Section 7 discusses non-dichotomous response variables.) Null cases have $\delta_i = 0$, indicating no difference between the two classes; non-null cases, particularly those with large values of $|\delta_i|$, are promising ingredients for effective prediction.

Let

$$W_i \equiv (X_i - \mu_i)/\sigma_i \qquad i = 1, 2, \ldots, N \tag{2.2}$$

be the standardized versions of X_i in (2.1). The optimal prediction rule is based on the weighted sum

$$S = \sum_{i=1}^{N} \delta_i W_i \sim \mathcal{N}\left(\pm\|\boldsymbol{\delta}\|^2/2c_0, \|\boldsymbol{\delta}\|^2\right), \tag{2.3}$$

$\|\boldsymbol{\delta}\|^2 = \sum_1^N \delta_i^2$, with "±" indicating the two classes as in (2.1). We predict

$$\begin{aligned} &\text{"healthy" if } S < 0, \\ &\text{"sick" if } S > 0. \end{aligned} \tag{2.4}$$

Prediction error rates of the first and second kinds, confusing healthy with sick or vice versa, both equal

$$\alpha = \Phi\left(-\|\boldsymbol{\delta}\|/2c_0\right). \tag{2.5}$$

Effective prediction requires a large $\boldsymbol{\delta}$ vector. In what follows, prediction error will be called simply "α".

Rule (2.3), (2.4) is Fisher's linear discriminant function applied to situation (2.1) (Hastie et al., 2008), assuming equal prior probabilities for the two classes. Remark B of Section 8 discusses the case of unequal probabilities. Section 5 considers a more realistic version of (2.1) that allows correlations among the predictors X_i.

In practice we need to estimate the parameters

$$(\mu_i, \sigma_i, \delta_i), \qquad i = 1, 2, \ldots, N \tag{2.6}$$

entering into $S = \sum \delta_i W_i$. This is where the training data

$$\begin{aligned} &\boldsymbol{x} = (x_{ij}), \qquad i = 1, 2, \ldots, N \quad \text{and} \quad j = 1, 2, \ldots, n \\ \text{and} \qquad &\boldsymbol{y} = (y_1, y_2, \ldots, y_n), \end{aligned} \tag{2.7}$$

with y_j equal $+1$ or -1 depending on the dichotomous classification of subject j, comes in. `Ebay`, the algorithm used for the numerical calculations here, employs standard estimates for (μ_i, σ_i):

$$\hat{\mu}_i = \frac{\bar{x}_{i1} + \bar{x}_{i2}}{2}, \qquad \hat{\sigma}_i = \left(\frac{SS_{i1} + SS_{i2}}{n-2}\right)^{1/2}, \tag{2.8}$$

$\bar{x}_{i1}$ and SS_{i1} the mean and within-group sum of squares for gene i measurements in the healthy subjects, and likewise $\bar{x}_{i2}$ and SS_{i2} for the sick subjects.

If σ_i were known, then

$$z_i = c_0 \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\sigma_i} \sim \mathcal{N}(\delta_i, 1) \tag{2.9}$$

would provide an obvious estimate of δ_i, say $\bar{\delta}_i = z_i$. With σ_i unknown, we convert the t-statistic t_i to the normal scale as in (1.2), (1.3). Remark F considers this transformation more carefully, but for now we will ignore it, and use the approximation $z_i \sim \mathcal{N}(\delta_i, 1)$ for our actual z-values (1.3).

Selection bias makes the $\bar{\delta}_i$ values overinflated estimates of the true δ_i's. Suppose that for the prostate data we decided to employ the genes having the 51 largest values of $|\bar{\delta}_i|$ for prediction. The vector of 51 $\bar{\delta}_i$'s has $\|\bar{\boldsymbol{\delta}}\| = 27.3$, suggesting $\alpha = .003$ in (2.5) (using $c_0 = (50 \cdot 52/102)^{1/2} = 5.05$). The empirical Bayes calculations of Section 3 show that a more realistic estimate for the actual 51-vector's length is 19.8, giving $\alpha = .025$.

The *shrunken centroids* algorithm of Tibshirani et al. (2002) counteracts selection bias by shrinking the estimates $\bar{\delta}_i = z_i$ toward zero according to a soft thresholding rule,

$$\hat{\delta}_i = \operatorname{sign}(z_i) \cdot (|z_i| - \lambda)_+. \tag{2.10}$$

In words, each value $\bar{\delta}_i = z_i$ is shrunk toward zero by amount λ, under the restriction that shrinking never goes past zero. A range of possible shrinkage parameters λ is tried, and for each one a prediction rule like (2.3) is formed, using

$$\hat{S}_\lambda = \sum \hat{\delta}_i \hat{W}_i \qquad \left[\hat{W}_i = (X_i - \hat{\mu}_i)/\hat{\sigma}_i\right] \tag{2.11}$$

for prediction as in (2.4). Cross-validation is then employed to estimate α_λ, the true error rate. (This description takes some liberties with the details of the shrunken centroids procedure.)

Notice that only cases having $|z_i| > \lambda$ enter into the prediction statistic $\hat{S}_\lambda$. This is a favorable property: prediction is easier to implement and understand when the number of predictors is small.

Table 1 shows a shrunken centroids analysis for the prostate data, carried out using `pamr`, a CRAN algorithm in the R language. Cross-validation suggests $\lambda = 2.16$ as the best shrinkage parameter (so for instance $z_i = 4$ yields $\hat{\delta}_i = 1.84$ in (2.11)) with estimated error rate $\hat{\alpha}_{\text{CV}} = .09$. 377 of the 6033 genes are involved in $\hat{S}_\lambda$.

shrinkage value λ	# nonzero genes	CV error rate
0.00	6033	0.34
0.54	3763	0.33
1.08	1931	0.23
1.62	866	0.12
2.16	**377**	**0.09**
2.70	172	0.10
3.24	80	0.16
3.78	35	0.30
4.32	4	0.41
4.86	1	0.48
5.29	0	0.52

Table 1: Shrunken centroids prediction for the prostate data (2.10), (2.11) using R program `pamr`, CRAN. The shrinkage parameter $\lambda = 2.16$ yields the smallest cross-validated error estimate, $\hat{\alpha}_{\text{CV}} = .09$. Prediction statistic $\hat{S}_\lambda$ involves 377 of the 6033 genes.

Looking at Table 1, it seems we should use $\lambda = 2.16$ in our prediction rule. There is, however, a subtle danger lurking here: because cross-validation is involved in the choice of "best" λ, the estimated rate .09 may be downwardly biased. It would take a second level of cross-validation to correct this bias.

A small simulation study was run with $N = 1000$, $n_1 = n_2 = 10$, and all $x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, 1)$. In this case $\delta_i = 0$ for every i in (2.1), so $\alpha = .50$ at (2.5); but the minimum cross-validated error rates observed in 100 repetitions of this set-up had median .30 with standard deviation ±.16.

This is an extreme example. Usually the downward bias is less severe, particularly when good prediction is possible. Nevertheless we will try to avoid such biases in what follows by using rules where the cross-validation calculations are not involved in the choice of tuning parameters.

3 Bayesian Prediction

Suppose we had a Bayesian prior distribution for the parameters in model (2.1) that enabled us to calculate posterior expectations for the δ_i's, say

$$\tilde{\delta}_i = E\{\delta_i | \boldsymbol{z}\}. \tag{3.1}$$

Bayes estimates are immune to selection bias: even if z_i were selected because it was the largest of the N z-values ($z_i = 5.29$ for gene $i = 610$ in the prostate data), $\tilde{\delta}_i$ would still be the correct Bayes estimate for δ_i. We could, for example, use the 50 largest values of $|\tilde{\delta}_i|$ to form $\tilde{S} = \sum \tilde{\delta}_i \hat{W}_i$, as in (2.3) or (2.11), while maintaining at least some confidence in the error rate estimate $\tilde{\alpha} = \Phi(-\|\tilde{\boldsymbol{\delta}}\|/2c_0)$. See Senn (2008) and Dawid (1994) for discussions of the "paradox" of Bayesian immunity to selection effects, including its dangers.

Brown (1971) and Stein (1981) developed a Bayesian model that is especially convenient for calculating $\tilde{\delta}_i$ in (3.1). For any (δ, z) pair we supposed that δ has a prior density $g(\delta)$,

$$\delta \sim g(\cdot) \quad \text{and} \quad z|\delta \sim \mathcal{N}(\delta, 1), \tag{3.2}$$

so that z has marginal density

$$f(z) = \int_{-\infty}^{\infty} \varphi(z-\delta) g(\delta)\, d\delta \qquad \left[\varphi(z) = e^{-z^2/2}/\sqrt{2\pi}\right]. \tag{3.3}$$

Theorem 1. *Under model* (3.2)*, the posterior density of δ given z is*

$$\begin{aligned} g(\delta|z) &= e^{\delta z - \psi(z)} \left[e^{-\delta^2/2} g(\delta)\right], \\ \textit{with} \qquad & \\ \psi(z) &= \log\left(f(z)/\varphi(z)\right). \end{aligned} \tag{3.4}$$

Proof. According to Bayes theorem,

$$g(\delta|z) = \varphi(z-\delta) g(\delta)/f(z), \tag{3.5}$$

which reduces immediately to (3.4). ■

Form (3.4) represents an exponential family having sufficient statistic δ, natural (canonical) parameter z, and cumulant generating function (cgf) $\psi(z)$. Therefore the conditional cumulants of δ given z can be obtained by differentiating ψ with respect to z:

Corollary 1.

$$E\{\delta|z\} = \psi'(z) \quad \textit{and} \quad \textit{Var}\{\delta|z\} = \psi''(z). \tag{3.6}$$

(Brown and Stein used multivariate versions of (3.6), differently derived, in their exploration of high-dimensional estimation theory.)

The advantage of Corollary 1 is that $\psi(z)$, and the cumulants of δ given z, are obtained directly from the marginal density $f(z)$, without requiring specific calculation of the prior $g(\delta)$, finessing the usual difficulties of deconvolution.

The algorithm `Ebay` described in Section 4 approximates $E\{\delta|z\}$ and $\text{Var}\{\delta|z\}$ by substituting a smoothed estimate $\hat{\psi}(z)$ into (3.6). Figure 2 displays the `Ebay` output $\hat{E}\{\delta|z\}$ for the prostate data, comparing it to the shrunken centroids curve (2.10) for $\lambda = 2.16$, the preferred choice in Table 1. $\hat{E}$ is better matched to the choice $\lambda = 1.42$ in (2.10), suggesting that less shrinking is better here.

Suppose we add to the Brown–Stein model (3.2) the assumption that the prior distribution of δ has a discrete atom of probability at $\delta = 0$ (see Remark C, Section 8),

$$p_0 = \text{Prob}\{\delta = 0\}. \tag{3.7}$$

Then Bayes theorem yields

$$\text{fdr}(z) \equiv \text{Prob}\{\delta = 0|z\} = p_0 \varphi(z)/f(z), \tag{3.8}$$

fdr(z) being the "local false discovery rate," Efron (2008). Comparing this with (3.4), (3.6) gives

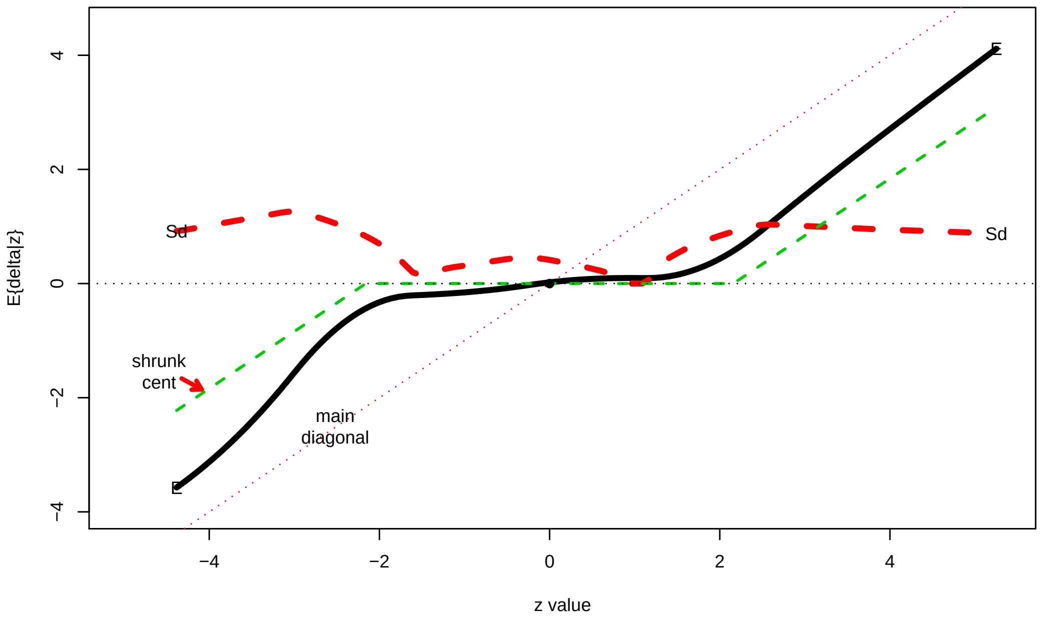

Figure 2: Heavy curve is $\hat{E}\{\delta|z\}$ for prostate data, `Ebay` algorithm, Section 4; compared with best shrunken centroids curve (2.10), $\lambda = 2.16$. Also shown, $\widehat{\mathrm{SD}} = \widehat{\mathrm{Var}}\{\delta|z\}^{1/2}$. At $z = 4$, $\hat{E} = 2.74$, shrunken centroid $= 1.84$, $\widehat{\mathrm{SD}} = .98$. Remark I explains the slight positive slope of $\hat{E}\{\delta|z\}$ for z in $(-2, 1.5)$.

Corollary 2. *Under model* (3.2), (3.7),

$$E\{\delta|z\} = -\frac{d}{dz}\log\left(fdr(z)\right) \quad and \quad Var\{\delta|z\} = -\frac{d^2}{dz}\log\left(fdr(z)\right). \tag{3.9}$$

It seemingly makes sense that only genes with low false discovery rates should be utilized in prediction rules. The corollary shows that this is roughly true, but in a rather surprising manner: large values of $\tilde{\delta}_i = E\{\delta_i|z_i\}$ depend on the rate of change of $\log(\mathrm{fdr}(z_i))$, not on $\mathrm{fdr}(z_i)$ itself. Small values of $\mathrm{fdr}(z_i)$ usually correspond to large values of $|\tilde{\delta}_i|$, but this doesn't have to be the case. Usually $\log(\mathrm{fdr}(z_i))$ is nearly constant around $z = 0$, where $\mathrm{fdr}(z) = 1$. This forces both $\hat{E}$ and $\widehat{\mathrm{SD}}$ to be small, as seen in Figure 2 (see Remark I).

4 Empirical Bayes Prediction

The `Ebay` algorithm that produced Figure 2 employs empirical Bayes methods to construct effective prediction rules. That is, it uses $\boldsymbol{z}$, the vector of all N z-values, to estimate the Bayes prediction rule (2.3), (2.4). Here is a schematic description of `Ebay`'s operation:

(1) A target error rate α_0 is selected (default $\alpha_0 = .025$).

(2) An estimate $\hat{f}(z)$ for the marginal density $f(z)$, (3.3), is obtained using Poisson regression on $\boldsymbol{z}$; see Remark D.

(3) The estimated cumulative generating function $\hat{\psi}(z) = \log(\hat{f}(z)/\varphi(z))$, (3.4), is numerically differentiated to give

$$\hat{\delta}_i = \hat{\psi}'(z_i) = \hat{E}\{\delta_i | z_i\}, \tag{4.1}$$

as in (3.6).

(4) Letting $\hat{\boldsymbol{\delta}}_I$ be the vector of I largest $\hat{\delta}_i$'s (in absolute value), I is chosen to be the smallest integer such that the nominal error rate $\Phi(-\|\hat{\boldsymbol{\delta}}_I\|/2c_0)$, (2.5), is less than α_0; that is, I is the minimum choice yielding

$$\|\hat{\boldsymbol{\delta}}_I\| \geq 2c_0\Phi^{-1}(1-\alpha_0). \tag{4.2}$$

(5) The empirical Bayes prediction rule is based on the sign of

$$\hat{S} = \sum_I \hat{\delta}_i \left(\frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i} \right), \tag{4.3}$$

$(\hat{\mu}_i, \hat{\sigma}_i)$ as in (2.8).

(6) Repeated 10-fold cross-validation is used to furnish an unbiased estimate of the rule's prediction error; see Remark G.

Table 2 shows a portion of `Ebay`'s output for the prostate data. It's prediction rule employs genes with the 51 largest values of $|\hat{\delta}_i|$, at which point (4.2) is first satisfied (compared with 377 genes for the apparently best shrunken centroids rule in Table 1). An unbiased error estimate, based on 20 randomized 10-fold cross-validation runs, was .092, the same as the minimum error seen in Table 1; see Remark G.

Step	Index	z-value	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{cor}}$
1	610	5.29	4.30	0.335	0.335
2	1720	4.83	3.78	0.285	0.281
3	364	−4.42	−3.70	0.250	0.250
4	3940	−4.33	−3.64	0.222	0.222
5	4546	−4.29	−3.58	0.199	0.215
6	4331	−4.14	−3.40	0.182	0.189
7	332	4.47	3.34	0.167	0.181
8	914	4.40	3.24	0.154	0.166
9	1068	4.25	3.06	0.144	0.148
10	4088	−3.88	−3.05	0.135	0.149
⋮	⋮	⋮	⋮	⋮	⋮
45	4154	−3.38	−2.26	0.029	0.050
46	2	3.57	2.25	0.028	0.050
47	2370	3.56	2.24	0.028	0.049
48	3282	3.56	2.23	0.027	0.048
49	3505	−3.33	−2.18	0.026	0.046
50	905	3.51	2.18	0.025	0.047
51	**4040**	**−3.33**	**−2.17**	**0.025**	**0.048**

Table 2: `Ebay` prediction rule for the prostate data; rule uses genes with 51 largest $|\hat{\delta}_i|$ values, $\hat{\alpha} = \Phi(-\|\hat{\boldsymbol{\delta}}\|/2c_0) = .025$. Cross-validation error rate $.092 \pm .004$. Column $\hat{\alpha}_{\text{cor}}$ explained in Section 5.

There are, potentially, many reasons why the nominal error rate .025 might be over-optimistic: $(\hat{\mu}_i, \hat{\sigma}_i)$ in (2.8) does not equal (μ_i, σ_i); the X_i are not normally distributed; the X_i are not independent (see Section 5); the empirical Bayes estimates $\hat{\delta}_i$ differ from the actual Bayes estimates (3.1).

This last point can cause particular trouble at the extremes of the z scale, just where $|\hat{\delta}(z)|$ is largest but there are fewest z_i's for the estimation of $\hat{\delta}$. Figure 3 concerns the following artificial situation, using notation similar to that for the prostate data and model (2.1):

$$\begin{aligned}
&\bullet\ N = 5000,\ n_1 = n_2 = 20, && \\
&\bullet\ \delta_i \overset{\text{ind}}{\sim} \mathcal{N}(1.5, 1) && \text{for } i = 1, 2, \dots, 250, \\
&\bullet\ \delta_i = 0 && \text{for } i = 251, 252, \dots, 5000, \\
&\bullet\ x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\pm\delta_i/2c_0, 1) && \text{for all } i \text{ and } j,\ c_0 = \sqrt{20^2/40}.
\end{aligned} \tag{4.4}$$

This results in

$$z_i \overset{\text{ind}}{\sim} \mathcal{N}(\delta_i, 1) \tag{4.5}$$

at (2.9), with $\delta_i \sim \mathcal{N}(1.5, 1)$ for the first 250 genes, and 0 otherwise.

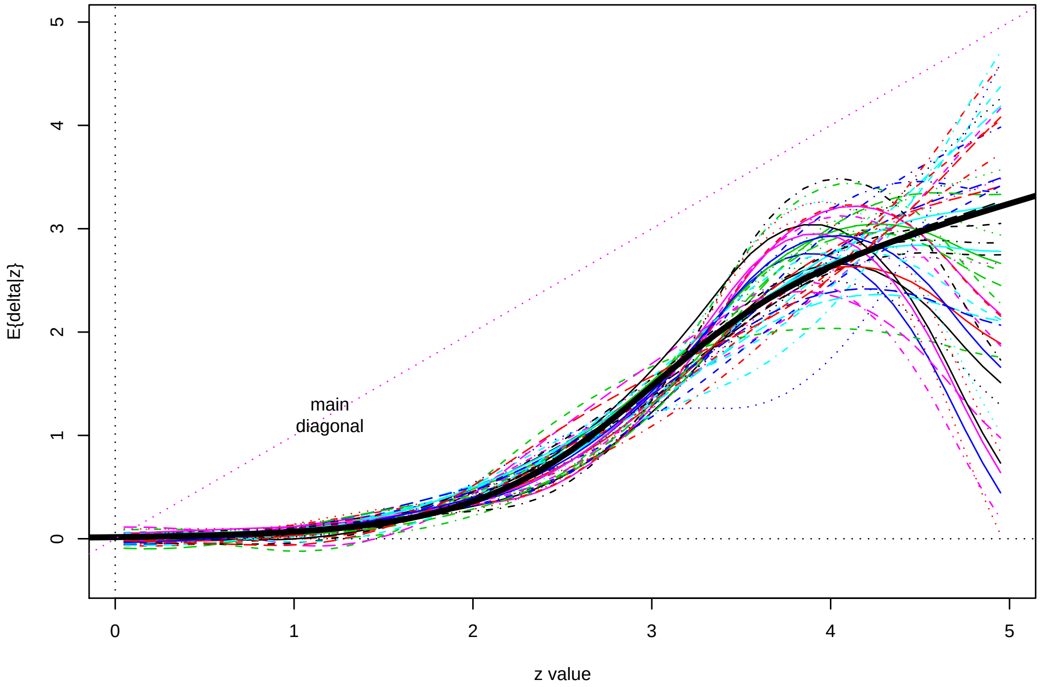

Figure 3: True curve $E\{\delta|z\}$ (heavy), compared to $\hat{E}\{\delta|z\} = \hat{\psi}(z)'$, 50 simulations from model (4.4). Estimates $\hat{E}$ reasonably accurate for $z < 4$, but fall apart for larger z values.

Figure 3 compares $\hat{E}\{\delta|z\}$ from the `Ebay` algorithm with the true curve $E\{\delta|z\}$. The estimates are reasonably accurate up to $z = 4$, but degenerate beyond that. Remark E derives a delta-method formula for the standard error of $\hat{E}\{\delta|z\}$ that predicts this behavior.

An option in Ebay allows for truncation of the $\hat{\delta}$ estimation procedure at some number "k_{trunc}" of observations in from the extremes. With $k_{\text{trunc}} = 5$ for instance, $\hat{\delta}_i$ for the five largest z_i values is set equal to $\max\{\hat{\delta}_i : i \leq N - 5\}$, and similarly at the negative end of the z scale.

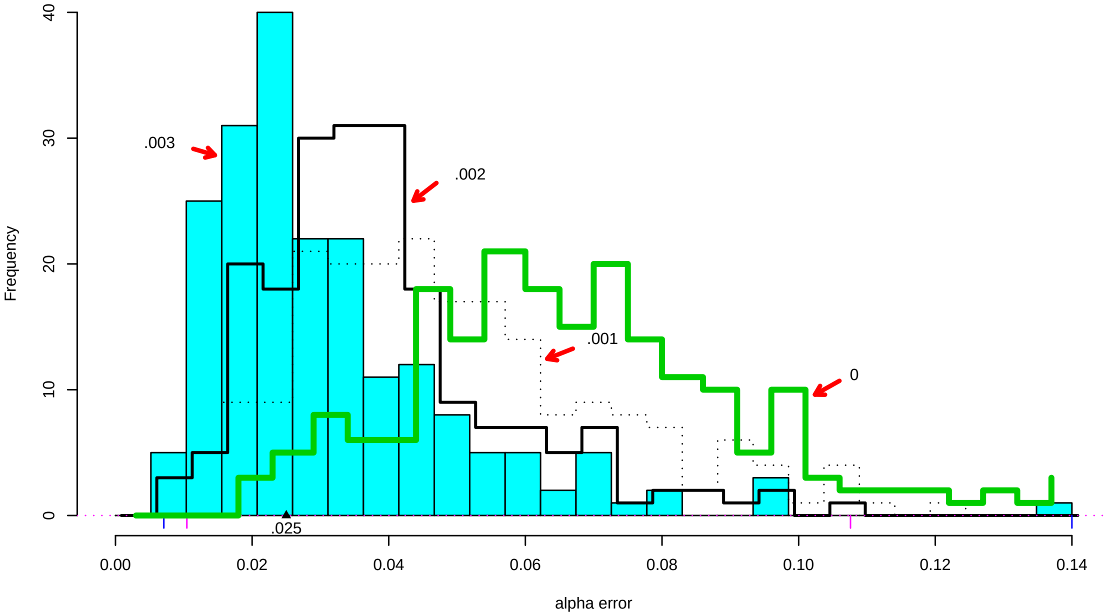

Figure 4: Actual prediction errors α of Ebay rule with nominal $\alpha_0 = .025$; 200 simulations from model (4.4). As truncation parameter increases from 0 (rightmost histogram) to 15 (leftmost), actual errors decrease toward nominal α_0.

Figure 4 shows the actual misclassification error probabilities α for 200 simulations from model (4.4), each time using the Ebay prediction rule with nominal error rate $\alpha_0 = .025$. As the truncation parameter k_{trunc} increases from 0 to 15, the actual prediction errors α decrease toward the target value .025. Table 3 displays the means and standard deviations for the data in Figure 4.

k_{trunc}:	15	10	5	0
Mean:	.032	.038	.051	.066
SD:	.019	.018	.022	.025

Table 3: Means and standard deviations for actual prediction errors α in simulation experiment for Figure 4.

Truncation had a less dramatic effect on the prostate data: for $k_{\text{trunc}} = 0, 5, 10, 15$, the cross-validated error estimates were .092, .085, .070, .077. Lowering the target rate from $\alpha_0 = .025$ to .01 gave corresponding error estimates .070, .062, .061, .058. Correlation among the predictors is part of the problem here; see Section 5.

Our original error estimate $\hat{\alpha} = .092$ is "honest", i.e., nearly unbiased for the Ebay rule produced with $(\alpha_0, k_{\text{trunc}}) = (.025, 0)$. So are the $\hat{\alpha}$ estimates for the other $(\alpha_0, k_{\text{trunc}})$ combinations. Choosing the combination with the smallest $\hat{\alpha}$, however, again raises the possibility of over-optimism, as discussed at the end of Section 2.

More elaborate "honest" selection criteria, beyond the current capabilities of `Ebay`, might involve minimizing a linear combination of nominal error rate and number of predictors, say

$$\Phi(-\|\hat{\boldsymbol{\delta}}_I\|/2c_0) + C \cdot I \tag{4.6}$$

over all choices of I; accounting for correlation as in Section 5, adjusting for non-normality; using theoretical or data-based techniques to choose the truncation parameter, etc.

Some "snooping" into the cross-validation estimates seems inevitable in applications. Nevertheless, I believe that holding snooping to a minimum is good practice for honest prediction assessment, and that empirical Bayes methods, perhaps further refined, can be sufficiently accurate to allow for a nearly-honest practical methodology.

5 Correlation Corrections

The assumption of case-wise independence in model (2.1) is likely to be untrue, perhaps spectacularly untrue, in many applications. Suppose that the vector $\boldsymbol{W}$ of standardized predictors $W_i = (X_i - \mu_i)/\sigma_i$, (2.2), actually has covariance matrix Σ. Then both error probabilities in (2.5) become

$$\alpha = \Phi(-\Delta_0 \cdot \eta) \quad \text{where} \begin{cases} \Delta_0 & = \|\boldsymbol{\delta}\|/2c_0 \\ \eta & = (\boldsymbol{\delta}^t\boldsymbol{\delta}/\boldsymbol{\delta}^t\Sigma\boldsymbol{\delta})^{1/2}. \end{cases} \tag{5.1}$$

Here Δ_0 is the independence value, while η is a correction factor, usually less than 1, that increases the error rate α.

If we can estimate Σ we can estimate correction factor η,

$$\hat{\eta} = (\hat{\boldsymbol{\delta}}^t\hat{\boldsymbol{\delta}}/\hat{\boldsymbol{\delta}}^t\hat{\Sigma}\hat{\boldsymbol{\delta}})^{1/2}. \tag{5.2}$$

According to (2.1), $\text{cov}(\boldsymbol{W}) = \Sigma$ has diagonal elements 1 in both classes, so the off-diagonal elements $\rho_{ii'}$ are correlations. Notice that we need estimate these for only the I cases selected by the `Ebay` algorithm at (4.2), not for all N cases. For the prostate data we need to estimate a 51×51 correlation matrix Σ, from the 51×102 data submatrix $\boldsymbol{x}_I$ of the full 6033×102 matrix $\boldsymbol{x}$ whose rows are indexed by the first column of Table 2.

The last column of Table 2 in Section 4 shows $\hat{\alpha}_{\text{cor}}$, obtained from (5.1), (5.2), with $\hat{\Sigma}$ the usual sample correlation matrix. Correlation degrades the nominal error probability from .025 to .048 (closer to the cross-validation estimate .092). Much of the degradation is due to three large correlations,

$$r_{34,19} = .97, \quad r_{36,15} = .65, \quad r_{42,28} = .92, \tag{5.3}$$

the subscripts referring to the steps in Table 2.

Table 4 concerns a microarray study having more severe correlation problems, the Michigan lung cancer study discussed in Subramanian et al. (2005). There are $N = 5217$ genes, $n = 86$ subjects, $n_1 = 62$ "good outcomes" and $n_2 = 24$ "poor outcomes". Here the `Ebay` algorithm stopped after 200 steps, without $\hat{\alpha}$ reaching the target value $\alpha_0 = .025$. The correlation-corrected errors $\hat{\alpha}_{\text{cor}}$ are much more pessimistic, actually increasing after the first 6 steps, eventually to $\hat{\alpha}_{\text{cor}} = .360$. A cross-validation error rate of .37 confirmed the pessimism. Restricting `Ebay` to use at most $I = 10$ predictors reduced the cross-validated error rate to .29, as suggested by Table 4 (an example of the kind of "snooping" disparaged at the end of Section 4, unless the decision to use the $I = 10$ `Ebay` prediction rule was made *before* the cross-validation calculations).

Sample correlation matrices tend toward overdispersion when n is small compared to the number of variates. `Ebay` includes an option for empirical Bayes shrinkage of the elements of $\hat{\Sigma}$; see Remark H.

Step	Index	z-value	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{cor}}$
1	3144	4.62	3.683	0.3290	0.329
2	2446	4.17	3.104	0.2813	0.307
3	4873	4.17	3.104	0.2455	0.256
4	1234	3.90	2.686	0.2234	0.225
5	621	3.77	2.458	0.2072	0.213
6	676	3.70	2.323	0.1942	0.228
7	2155	3.69	2.313	0.1824	0.230
8	3103	3.60	2.140	0.1731	0.236
9	1715	3.58	2.103	0.1647	0.240
10	452	3.54	2.028	0.1574	0.243
⋮	⋮	⋮	⋮	⋮	⋮
193	3055	2.47	0.499	0.0519	0.359
194	1655	−2.21	−0.497	0.0518	0.359
195	2455	2.47	0.496	0.0517	0.359
196	3916	2.47	0.496	0.0516	0.359
197	4764	2.47	0.495	0.0515	0.359
198	1022	−2.20	−0.492	0.0514	0.359
199	1787	−2.19	−0.490	0.0513	0.360
200	901	−2.18	−0.486	0.0512	0.360

Table 4: `Ebay` output for the Michigan lung cancer study. Correlation error estimates $\hat{\alpha}_{\text{cor}}$ are much more pessimistic, as confirmed by cross-validation.

6 Effect Size Estimation

Current developments in large-scale simultaneous inference have focused on hypothesis testing, where the goal is to identify a small number of non-null cases among a large number of potential candidates. See Dudoit et al. (2003) for a nice review. Benjamini & Yekutieli (2005) address a more ambitious goal: to assess the effect sizes for the non-null cases, that is, to estimate how far away they lie from the null hypothesis. The empirical Bayes theory of Section 4 provides an alternative approach to effect size estimation.

We begin with assumptions (3.2), (3.7), that

$$z_i \sim \mathcal{N}(\delta_i, 1) \qquad i = 1, 2, \ldots, N, \tag{6.1}$$

and that proportion p_0 of the effects δ_i equal 0,

$$p_0 = \text{Prob}\{\delta_i = 0\}, \tag{6.2}$$

these being the uninteresting null cases. The local false discovery rate $\text{fdr}(z) = p_0\varphi(z)/f(z)$, (3.8), is the Bayes posterior probability $\text{Prob}\{\delta_i = 0|z_i\}$. If $\widehat{\text{fdr}}(z_i)$, an estimate of $\text{fdr}(z_i)$, is suitably small, then case i can be reported as "probably non-null", and we would like to put some sort of confidence limits on the effect size δ_i. The prior $g(\delta)$ in (3.2) is now of the mixed form

$$g(\delta) = p_0 I_0(\delta) + (1 - p_0) g_1(\delta), \tag{6.3}$$

where $I_0(\delta)$ is a delta-function at 0, and $g_1(\delta)$ indicates the density of the non-null cases (see Remark C). Then the mixture density $f(z)$, (3.3), becomes

$$f(z) = p_0\varphi(z) + (1 - p_0) f_1(z)$$

where

$$f_1(z) = \int_{-\infty}^{\infty} \varphi(z-\delta) g_1(\delta) d\delta. \tag{6.4}$$

Theorem 2. *Under model* (6.1), (6.2), *the posterior density of effect size δ given z and given that $\delta \neq 0$ is*

$$g_1(\delta|z) = e^{\delta z - \psi_1(z)} \left[e^{-\delta^2/2} g_1(\delta) \right]$$

where

$$\psi_1(z) = \log \left\{ \frac{1-\text{fdr}(z)}{\text{fdr}(z)} \Big/ \frac{1-p_0}{p_0} \right\}. \tag{6.5}$$

Proof. Bayes rule says that $g_1(\delta|z) = \varphi(z-\delta)g_1(\delta)/f_1(z)$, yielding

$$g_1(\delta|z) = e^{\delta z - \log\{f_1(z)/\varphi(z)\}} \left[e^{-\delta^2/2} g_1(\delta) \right]. \tag{6.6}$$

An equivalent form of (3.8) is

$$1 - \text{fdr}(z) = \text{Prob}\{\delta \neq 0 | z\} = (1-p_0) f_1(z)/f(z), \tag{6.7}$$

from which we obtain, using (6.4),

$$\frac{f_1(z)}{\varphi(z)} = \frac{p_0}{1-p_0} \frac{1-\text{fdr}(z)}{\text{fdr}(z)}. \tag{6.8}$$

Combining (6.8) and (6.6) verifies Theorem 2. ■

As in (3.6), the conditional moments of a non-null δ (one for which $\delta \neq 0$) given z are obtained by differentiating $\psi_1(z)$,

$$E_1\{\delta|z\} = \psi_1'(z) \quad \text{and} \quad \text{Var}_1\{\delta|z\} = \psi_1''(z), \tag{6.9}$$

where the subscript "1" indicates conditioning on $\delta \neq 0$. Some calculation gives E_1 and Var_1 in terms of $E\{\delta|z\}$ and $\text{Var}\{\delta|z\}$ in (3.6):

Corollary 3. *Under model* (6.1), (6.2),

$$E_1\{\delta|z\} = E\{\delta|z\} / \left[1 - \text{fdr}(z)\right]$$

and

$$\text{Var}_1\{\delta|z\} = \frac{1}{1-\text{fdr}(z)} \left[\text{Var}(\delta|z) - \frac{\text{fdr}(z)}{1-\text{fdr}(z)} E\{\delta|z\}^2 \right]. \tag{6.10}$$

Note. Since $\delta = 0$ with probability fdr(z), we have

$$E\{\delta^j|z\} = [1-\text{fdr}(z)] \cdot E_1\{\delta^j|z\}. \tag{6.11}$$

Using (6.11) with $j = 1$ and 2 leads to a quick verification of (6.10).

Our prediction algorithm in Section 4 requires only the estimation of $E\{\delta|z\}$. Effect size estimation is more difficult, requiring $\mathrm{Var}\{\delta|z\}$ and $\mathrm{fdr}(z)$ as well. The plug-in estimate of $\mathrm{Var}_1\{\delta|z\}$ in (6.10) may be particularly unstable, in which case we can conservatively replace it with $\widehat{\mathrm{Var}}\{\delta|z\}$, as shown next.

Rearranging (6.10) yields

$$\frac{\mathrm{Var}_1}{\mathrm{Var}} = \frac{1-\mathrm{fdr}(z)Q(z)}{1-\mathrm{fdr}(z)} \quad \text{where} \quad Q(z) = \frac{E^2}{(1-\mathrm{fdr}(z))\cdot \mathrm{Var}}, \tag{6.12}$$

(with Var= $\mathrm{Var}\{\delta|z\}$, etc.) so that

$$\mathrm{Var}_1 \le \mathrm{Var} \qquad \text{if } \mathrm{Var} \le E^2/(1-\mathrm{fdr}(z)). \tag{6.13}$$

Since Var is usually near 1, this last condition is satisfied whenever $\hat{\delta} = E\{\delta|z\}$ gets large enough to be interesting; in the case of the prostate data, for $z \ge 2$.

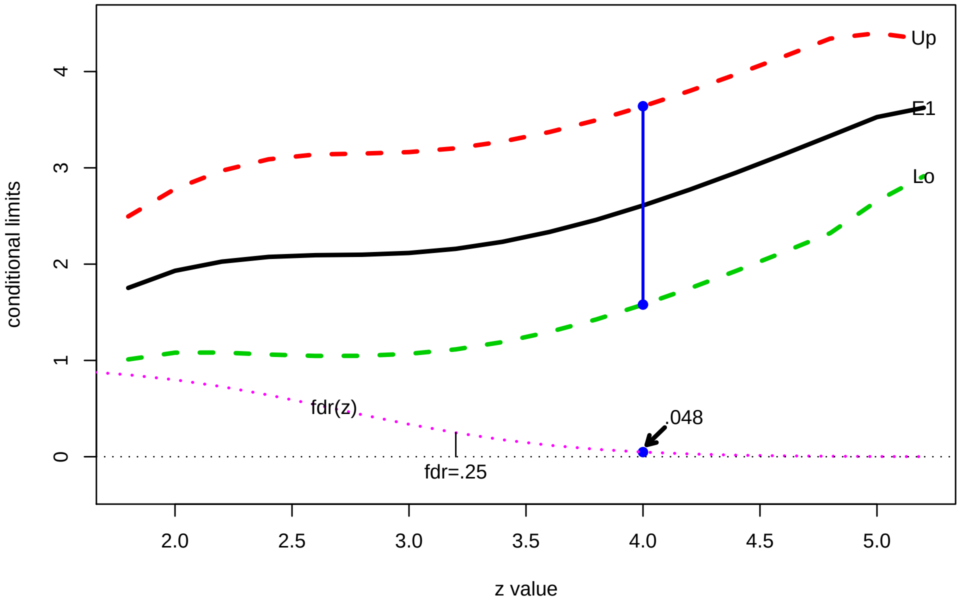

Figure 5: Effect size estimation for the prostate data. Band is approximate 68% interval for δ given z and given $\delta \neq 0$; also showing $\widehat{\mathrm{fdr}}(z)$, estimated probability $\delta = 0$ given z. At $z = 4$, $\widehat{\mathrm{fdr}}(z) = .048$, with interval $[1.58, 3.64]$ if δ is non-null.

Figure 5 demonstrates effect size estimation for the prostate data. The Poisson GLM estimate of $f(z)$ described in Remark D provides estimates of $\widehat{\mathrm{fdr}}(z)$, $\hat{E}\{\delta|z\}$, and $\widehat{\mathrm{Var}}\{\delta|z\}$, as in Figure 2. The curved band in Figure 5 follows

$$\hat{E}\{\delta|z\}/\left[1-\widehat{\mathrm{fdr}}(z)\right] \pm \widehat{\mathrm{Var}}\{\delta|z\}^{1/2} \tag{6.14}$$

as in Corollary 3, showing approximate 68% intervals for δ given z and given $\delta \neq 0$ — made more conservative by replacing $\widehat{\mathrm{Var}_1}$ with $\widehat{\mathrm{Var}}$. At $z = 4$ for example, we estimate that either $\delta = 0$ with probability $\widehat{\text{fdr}}(4) = .048$, or, if $\delta \neq 0$, it lies in the interval $[1.58, 3.64)$ with estimated posterior probability exceeding .68. (Remember that δ, as defined in (2.1), is the number of standard deviations separating the two class means, *multiplied by* c_0.)

Benjamini & Yekutieli's (2005) False Coverage Rate algorithm provides conservative frequentist confidence bounds on the cases declared non-null by an FDR testing procedure, assuming independence of the z_i's. There is, however, a heavy price to pay: the bounds tend to be very wide. For $z = 4$ in the prostate example, their 68% interval is $[1.36, 6.64)$ (using their Definition 1, with $q = .32$). Part of the problem, as discussed in Section 7 of Efron (2008), is that the Bejamini/Yekutieli procedure does not split off an atom of probability at $\delta = 0$, though splitting seems natural in the hypothesis testing framework of (3.7) or (6.3).

The approximate 68% non-null limits (6.14) were calculated for 25 replications of simulation model (4.4). They appear in Figure 6, along with the true Bayesian posterior limits $(z + 1.5)/2 \pm 1/\sqrt{2}$. Using $\widehat{\text{Var}}$ instead of $\widehat{\text{Var}}_1$ in (6.14) makes the intervals too wide, but their overall performance is acceptable as rough estimates of effect size.

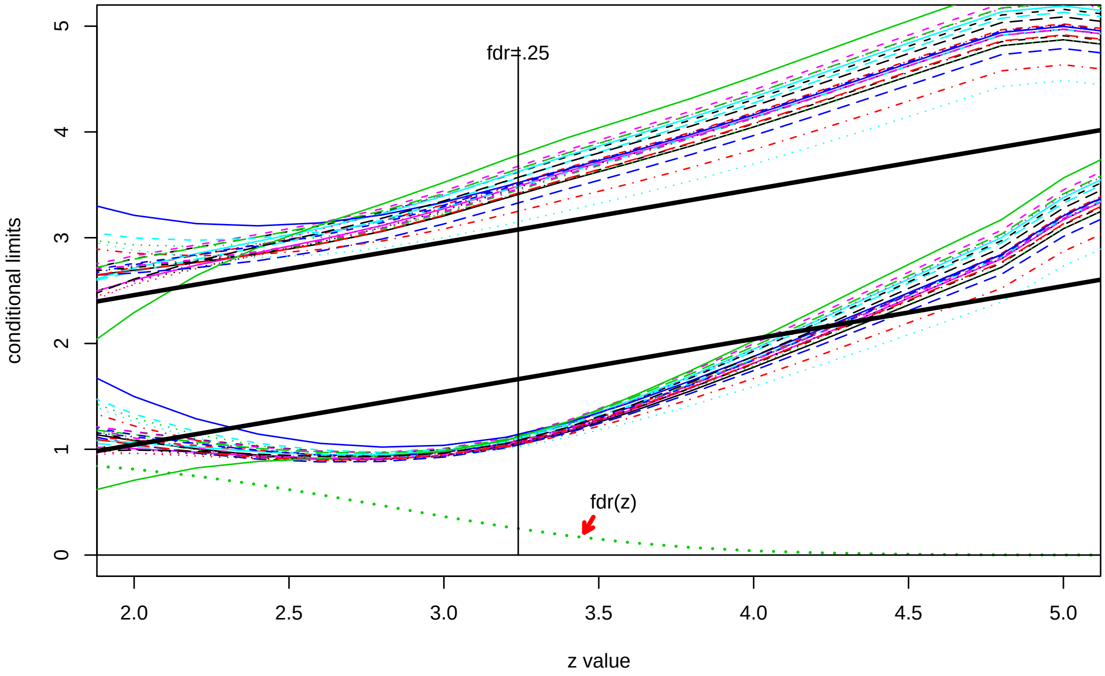

Figure 6: Approximate effect size limits (6.14) for 25 replications of simulation model (4.4). Heavy straight lines are actual 68% Bayes posterior limits for non-null cases.

7 Other Response Variables

The development so far has concerned dichotomous response variables: healthy versus sick in the prostate example. This section extends the empirical Bayes prediction methodology to general univariate responses.

Let Y be a univariate response of interest, for example a survival time that we wish to predict from $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)'$ as in Section 2. For convenience we assume that Y has been standardized to have mean 0 and variance 1, denoted

$$Y \sim (0, 1), \tag{7.1}$$

though this will play no role in the actual methodology.

We suppose that Y influences the standardized variable $W_i = (X_i - \mu_i)/\sigma_i$, (2.1), through linear regression,

$$W_i = \beta_i Y + \epsilon_i, \qquad i - 1, 2, \ldots, N, \tag{7.2}$$

$\text{Var}(\epsilon_i) = 1$, where the vector of errors $\boldsymbol{\epsilon}$ is uncorrelated with Y. In the dichotomous situation of (2.1), $Y = -1$ or 1 and $\beta_i = \delta_i/2c_0$. Effective prediction of Y depends upon discovering those X_i's with large values of $|\beta_i|$.

The joint distribution of Y and $\boldsymbol{W}$ has mean vector and covariance matrix

$$\begin{pmatrix} Y \\ \boldsymbol{W} \end{pmatrix} \sim \left[\begin{pmatrix} O \\ \boldsymbol{O} \end{pmatrix}, \begin{pmatrix} 1 & \boldsymbol{\beta}^t \\ \boldsymbol{\beta} & \boldsymbol{\beta}\boldsymbol{\beta}^t + \Sigma \end{pmatrix} \right], \tag{7.3}$$

Σ indicating the covariance matrix of $\boldsymbol{\epsilon}$. The best linear predictor of Y from $\boldsymbol{W}$ is

$$\begin{aligned} Y^\dagger &= \boldsymbol{\beta}^t(\boldsymbol{\beta}\boldsymbol{\beta}^t + \Sigma)^{-1}\boldsymbol{W} \\ &= \frac{1}{1 + \Delta^2}\boldsymbol{\beta}^t \Sigma^{-1}\boldsymbol{W}, \end{aligned} \tag{7.4}$$

where Δ^2 is the squared Mahalanobis distance

$$\Delta^2 = \beta^t \Sigma^{-1} \beta. \tag{7.5}$$

If Σ is the identity, as assumed in (2.1), then $Y^\dagger = \text{constant} \cdot \boldsymbol{\beta}^t \boldsymbol{W}$, similarly to (2.3).

Combining (7.4) with (7.2) produces a simple expression for the conditional mean and variance of $Y^\dagger$ given Y,

$$Y^\dagger | Y \sim \left(\frac{\Delta^2}{1 + \Delta^2} Y, \frac{\Delta^2}{1 + \Delta^2} \right), \tag{7.6}$$

from which (7.1) gives

$$\text{cor}(Y, Y^\dagger) = \Delta / \sqrt{1 + \Delta^2}. \tag{7.7}$$

Effective prediction of Y from $\boldsymbol{W}$ requires a large value of $\Delta = (\beta^t \Sigma \beta)^{1/2}$. (In the context of Section 2, where $\Sigma = I$ and $\boldsymbol{\beta} = \boldsymbol{\delta}/2c_0$, we have $\Delta = \|\boldsymbol{\delta}\|/2c_0$, so the error probability α equals $\Phi(-\Delta)$ at (2.5)).

To bring empirical Bayes methods to bear on the estimation of $Y^\dagger$ we need to estimate posterior expectations for the regression coefficients β_i from the training data (1.1): the $N \times n$ matrix $\boldsymbol{x}$ and the n-vector of responses $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$. Let $\boldsymbol{x}_i^t$ indicate the ith row of $\boldsymbol{x}$. Applying model (7.2) independently to each column of $\boldsymbol{x}$ gives a linear model for the rows,

$$\boldsymbol{x}_i = \mu_i \mathbf{1}_n + \sigma_i(\beta_i \boldsymbol{y} + \boldsymbol{\epsilon}_i), \tag{7.8}$$

$\mathbf{1}_n$ a vector of n 1's, where the components of $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in})^t$ are independent and identically distributed, with mean 0 and variance 1. Ordinary least squares applied to (7.8) provides familiar estimates of μ_i, σ_i and β_i. In the dichotomous setting of (2.1), $\hat{\mu}_i$ and $\hat{\sigma}_i$ are as given in (2.8) while $2c_0\hat{\beta}_i$ equals $\bar{\delta}_i = z_i$ in (2.9).

If we assume that the errors ϵ_i are normally distributed, then the t-statistic "t_i" for testing $\beta_i = 0$ in (7.8) has a non-central t distribution with $n-2$ degrees of freedom and non-centrality parameter proportional to β_i,

$$t_i \sim t_{n-2}(\delta_i), \qquad \left[\delta_i \equiv 2c_0\beta_i \quad \text{with} \quad c_0^2 = \sum_1^n (y_i - \bar{y})^2/4\right]. \tag{7.9}$$

In usual practice, (7.9) remains a reasonable approximation as long as the ϵ_{ij} distribution does not have heavy tails. With dichotomous y_i, $c_0 = (n_1 n_2/n)^{1/2}$ as before.

We can transform t_i to a z-value via

$$z_i = \Phi^{-1}\left(F_{n-2}(t_i)\right), \tag{7.10}$$

with Φ and F_{n-2} the standard normal and central t_{n-2} cdf's. If n is large then (7.9) gives

$$z_i \dot{\sim} \mathcal{N}(\delta_i, 1) \tag{7.11}$$

as in (2.9). Remark F improves upon approximation (7.11), but we will take it as given here.

We can now proceed as in Section 4:

1. $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)^t$ provides $\hat{f}(z)$, an estimate of the marginal density of the z-values (Remark D) and $\hat{\psi}(z) = \hat{f}(z)/\varphi(z)$.

2. We then calculate
$$\hat{\delta}_i = \hat{\psi}'(z_i) = \hat{E}\{\delta_i | z_i\} \qquad \text{for } i = 1, 2, \ldots, N. \tag{7.12}$$

3. $\hat{\boldsymbol{\delta}}_I$, the vector of I largest $\hat{\delta}_i$'s in absolute value, gives
$$\hat{\Delta}_I = \|\hat{\boldsymbol{\delta}}_I\|/2c_0 \quad \text{and} \quad \widehat{\text{cor}}_I = \hat{\Delta}_I \Big/ \sqrt{1 + \hat{\Delta}_I^2}. \tag{7.13}$$

4. We continue increasing I until either $\widehat{\text{cor}}_I$ reaches some target value or I reaches a preselected upper bound, and use
$$\hat{Y}^\dagger = \frac{1}{1 + \hat{\Delta}_I^2} \sum_{i=1}^{I} \frac{\hat{\delta}_i}{2c_0} \left(\frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i}\right), \tag{7.14}$$
from (7.4), to predict Y from $\boldsymbol{X}$.

Steps (3) and (4) assume uncorrelated X_i's, i.e., $\not\Sigma$ the identity matrix, but correlation can be incorporated as in Section 5.

These steps were carried out for an ongoing lung cancer microarray study involving $n = 100$ patients each measured on $N = 16,000$ genes. All patients received the same new drug. The response variable "Y" was a categorical assessment of improvement, adjusted for two covariates, running from -2 (worst) to $+2$ (best).

Figure 7 shows $\hat{E}\{\delta|z\}$, calculated by Steps (1) and (2) above. It seems clear that any power of the microarray expression measurements to predict Y must come from those genes having z_i less than -2. Table 5 shows this to be true. Predictive power is modest here, with theoretical correlation only .48 after $I = 50$ steps, asymptoting to .57 at $I = 16,000$.

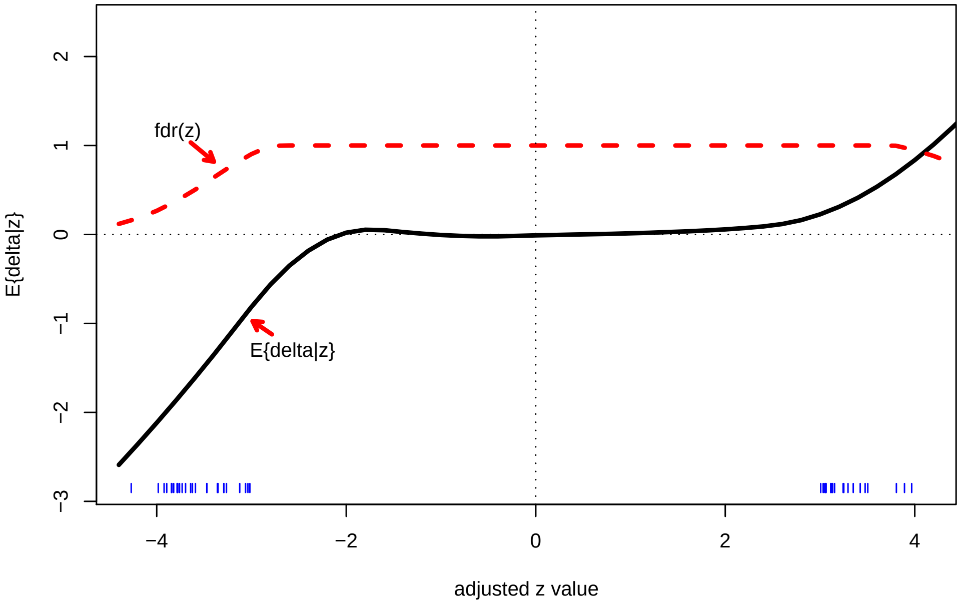

Figure 7: Lung cancer microarray study, $N = 16,000$ genes, $n = 100$ patients, ordered categorical response variable Y. Heavy curve $\hat{E}\{\delta|z\}$, (7.14). Dashes indicate those z-values exceeding 3 in absolute value.

8 Remarks

The following remarks expand on some of the questions and technical points raised earlier.

A. Centroids Interpretation Prediction rule (4.3), which depends on the sign of $\hat{S} = \sum \hat{\delta}_i \hat{W}_i$, $\hat{W}_i = (X_i - \hat{\mu}_i)/\hat{\sigma}_i$, can be stated in more conventional centroid terminology: letting

$$D_1 = \|\hat{\boldsymbol{W}} + \hat{\boldsymbol{\delta}}/2c_0\| \quad \text{and} \quad D_2 = \|\hat{\boldsymbol{W}} - \hat{\boldsymbol{\delta}}/2c_0\|, \tag{8.1}$$

we predict "healthy" if $D_1 < D_2$ and "sick" if $D_2 < D_1$; so $\hat{\boldsymbol{\delta}}/2c_0$ and $-\hat{\boldsymbol{\delta}}/2c_0$ are the standardized centroids. An alternative statement refers to the hyperplane $\hat{\mathcal{L}}$ passing through the origin of N-space orthogonal to the line segment connecting $\hat{\boldsymbol{\delta}}/2c_0$ with $-\hat{\boldsymbol{\delta}}/2c_0$: we predict healthy or sick depending on which side of $\hat{\mathcal{L}}$ the point $\hat{\boldsymbol{W}}$ falls.

B. Unequal Prior Probabilities Prediction rule (4.3) tacitly assumes that our dichotomous response variable has equal prior probabilities on the two categories, irrespective of the observed frequencies n_1 and n_2 in the training set. Suppose that the prior probabilities are actually π_1 and π_2. Starting with model (2.1), calculations involving Fisher's linear discriminant function imply the following change from Remark A: the prediction boundary $\hat{\mathcal{L}}$ is translated to intersect the orthogonal line segment at directed distance

$$\frac{c_0}{\|\hat{\boldsymbol{\delta}}\|} \log\left(\frac{\pi_1}{\pi_2}\right) \tag{8.2}$$

from the origin. (The definition of $\hat{\boldsymbol{W}}$ is still $(\boldsymbol{X} - \hat{\boldsymbol{\mu}})/\hat{\boldsymbol{\sigma}}$, with $(\hat{\mu}_i, \hat{\sigma}_i)$ as given in (2.8).)

Step	Index	z-value	$\hat{\delta}$	$\hat{\beta}$	$\hat{\Delta}_I$	$\widehat{\text{cor}}_I$
1	12404	-4.27	-2.44	-0.20	0.04	0.04
2	6342	-3.98	-2.10	-0.17	0.07	0.07
3	2516	-3.92	-2.02	-0.16	0.10	0.10
4	488	-3.89	-1.99	-0.16	0.12	0.12
5	8471	-3.84	-1.93	-0.16	0.15	0.15
6	25	-3.84	-1.92	-0.16	0.17	0.17
7	2872	-3.82	-1.90	-0.15	0.20	0.19
8	300	-3.78	-1.85	-0.15	0.22	0.21
9	545	-3.78	-1.85	-0.15	0.24	0.23
10	12448	-3.78	-1.84	-0.15	0.26	0.26
⋮	⋮	⋮	⋮	⋮	⋮	⋮
45	10905	-2.83	-0.60	-0.05	0.54	0.47
46	390	-2.83	-0.59	-0.05	0.54	0.47
47	1498	-2.82	-0.59	-0.05	0.54	0.48
48	10317	-2.81	-0.57	-0.05	0.54	0.48
49	7894	-2.79	-0.56	-0.05	0.55	0.48
50	13263	-2.79	-0.55	-0.04	0.55	0.48

Table 5: Right column shows $\widehat{\text{cor}}_I$, (7.13), for lung cancer data; $I = 1$ to 50. Final value $\widehat{\text{cor}}_{16000} = .57$.

C. The Prior Density $g(\delta)$ In the Brown–Stein model (3.2), the prior density $g(\delta)$ can be extended to a general probability distribution $G(\delta)$ incorporating discrete atoms of probability as in (3.7). Theorem 1's statement is almost unchanged,

$$dG(\delta|z) = e^{\delta z - \psi(z)} e^{-\delta^2/2} dG(\delta). \tag{8.3}$$

The factor $e^{-\delta^2/2}$ guarantees that the exponential family has natural parameter space including all values of z, justifying Corollary 2 for all z. The same considerations apply to Theorem 2.

D. Estimating $f(z)$ `Ebay` estimates $f(z)$, the mixture density (3.3), by means of a Poisson generalized linear model (glm) applied to binned counts of the N z-values. In Figure 1, for example, there are $K = 90$ bins, each of width 0.1, ranging from -4.5 to 4.5. The counts

$$c_k = \#\{z_i \text{ in bin } k\}, \qquad k = 1, 2, \dots, K \tag{8.4}$$

are the heights of the histogram bars. Let $\boldsymbol{b}$ indicate the K-vector of bin midpoints. Then the estimate $\hat{\boldsymbol{f}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_K)^t$ of $f(z)$ at the points in $\boldsymbol{b}$ is obtained by Poisson regression of the counts on a natural spline function of the midpoints:

$$\hat{\boldsymbol{f}} = \texttt{glm}\,(\boldsymbol{c} \sim \texttt{ns}(\boldsymbol{b}, \texttt{df}), \texttt{poisson})\,\$\texttt{fit} \tag{8.5}$$

in R notation; default degrees of freedom df equals 7 in `Ebay`; $\hat{\boldsymbol{f}}$ is the discretized mle of $f(z)$ in the 7-parameter exponential family defined by the natural spline basis.

Estimate (8.5) is the same one employed by *locfdr*, the local false discovery rate algorithm described in Efron (2008). Applied to the prostate data, *locfdr* estimated $\hat{p}_0 = .93$ for the proportion of null genes (3.7), assuming that $f(z)$ is the correct null density.

E. Accuracy Formula for $\hat{E}\{\delta|z\}$ A closed-form delta-method expression for the variance of $\hat{\delta}_i = \hat{E}\{\delta_i|z_i\}$ can be derived if we are willing to assume that the z_i's are independent of each other.

Let M be the $K \times m$ structure matrix $\mathtt{ns}(\boldsymbol{b}, \mathtt{df})$ in (8.5), $K = 90$ and $m = 8$; $\text{diag}(\boldsymbol{c})$ the $K \times K$ diagonal matrix with diagonal entries the bin counts c_k; and $G = M^t \text{diag}(\boldsymbol{c}) M$. Section 5 of Efron (2007) employs the relationship

$$d\hat{\boldsymbol{\ell}} = MG^{-1}M^t d\boldsymbol{c} \tag{8.6}$$

for the derivative matrix of the K-vector $\hat{\boldsymbol{\ell}} = \log(\hat{\boldsymbol{f}})$ with respect to a continuized version of $\boldsymbol{c}$.

Let D be the $(K-2) \times K$ matrix whose kth row is

$$(0, 0, \dots, 0, -1, 0, 1, 0, 0, \dots)/d_0, \tag{8.7}$$

with -1 in the kth place: $D\hat{\boldsymbol{\ell}} = \hat{\boldsymbol{\ell}}'$, the numerical derivative of $\hat{\boldsymbol{\ell}}$. This gives

$$d\hat{\boldsymbol{\ell}}' = DMG^{-1}M^t d\boldsymbol{c}. \tag{8.8}$$

The Poisson estimate $\text{Cov}(\boldsymbol{c}) = \text{diag}(\boldsymbol{c})$ for the covariance matrix of $\boldsymbol{c}$ then yields $\text{Cov}(\hat{\boldsymbol{\ell}}') = DMG^{-1}M^tD^t$. But since

$$\psi'(z) = \frac{d}{dz} \log \{f(z)/\varphi(z)\} = z + \ell'(z), \tag{8.9}$$

we have $\hat{\delta}_{(k)} \equiv \hat{\psi}'(z = b_k) = b_k + \hat{\ell}'_k$ in (4.1), implying that

$$\text{Cov}(\hat{\boldsymbol{\delta}}) = \text{Cov}(\hat{\boldsymbol{\ell}}') = DMG^{-1}M^tD^t. \tag{8.10}$$

Table 6 shows estimates of standard error for $\hat{\delta}$ (square roots of the diagonal elements in (8.10)) calculated for the prostate data. As in Figure 3, we can see an explosive increase in variability as $|z|$ increases to 4.

z:	-4	-3	-2	-1	0	1	2	3	4
sd:	.41	.12	.09	.06	.04	.05	.09	.10	.33

Table 6: Delta-method standard errors for $\hat{\delta}(z) = \hat{E}\{\delta|z\}$, formula (8.10), for the prostate data.

F. Transforming t-values to z-values The ith row of $\boldsymbol{x}$ comprises n independent observations

$$x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i \pm \sigma_i \delta_i / 2c_0, \sigma_i^2) \qquad \text{for } j = 1, 2, \dots, n \tag{8.11}$$

in the notation of Section 2, with n_1 "$-$" values and n_2 "+" values. The corresponding two-sample t-statistic t_i follows a non-central t distribution with $n-2$ degrees of freedom and noncentrality parameter δ_i,

$$t_i = c_0 \frac{\bar{x}_{2i} - \bar{x}_{1i}}{\hat{\sigma}_i} \sim t_{n-2}(\delta_i). \tag{8.12}$$

Our previous discussion treated t_i as $z_i \sim \mathcal{N}(\delta_i, 1)$, but `Ebay` actually employs transformations that improve the accuracy of Corollary 3.

Let

$$z_i = \Phi^{-1}\left(F_{n-2}(t_i)\right), \tag{8.13}$$

as in (7.10), so if $\delta_i = 0$ then $z_i \sim \mathcal{N}(0,1)$. If $\delta_i \neq 0$, z_i is still surprisingly close to normal,

$$z_i \dot{\sim} \mathcal{N}\left(\zeta_i, \sigma^2(\zeta_i)\right), \qquad \left[\zeta_i = \Phi^{-1}\left(F_{n-2}(\delta_i)\right)\right], \tag{8.14}$$

with $\sigma(\zeta_i) < 1$. For example, with $\delta_i = 4$ and $n = 102$, z_i from (8.13) has (mean, standard deviation, skewness, kurtosis) equal $(3.845, .931, -.046, .010)$. A plot of (8.14) superimposed on (8.12) barely differentiates the two curves.

The computation of $\hat{\delta}_i$, (4.1) in the `Ebay` algorithm, is actually carried out using (8.14):

- The vector $\boldsymbol{t} = (t_1, t_2, \dots, t_N)^t$ is converted component-wise to $\boldsymbol{z} = (z_1, z_2, \dots, z_N)$, as in (8.13).
- An estimate $\hat{f}(z)$ is constructed from $\boldsymbol{z}$ as in Remark *D*.
- A modified version of Corollary 2, described below, provides empirical Bayes estimates $\hat{\zeta}_i$.
- Finally, transformation (8.14) is inverted to give

  $$\hat{\delta}_i = F_{n-2}^{-1}\left(\Phi(\hat{\zeta}_i)\right), \tag{8.15}$$

  after which `Ebay` proceeds as in Steps 4–6 in Section 4.

Suppose the Brown–Stein model (3.2) is modified to have $z|\delta \sim \mathcal{N}(\delta, \sigma^2)$. Then it is easy to show that

$$E\{\delta|z\} = z + \sigma^2 \ell'(z) \quad \text{and} \quad \text{Var}\{\delta|z\} = \sigma^2 + \sigma^4 \ell''(z), \tag{8.16}$$

where $\ell(z)$ is the log of the marginal density $f(z)$. The empirical Bayes estimate $\hat{\zeta}_i$ mentioned above is given by

$$\hat{\zeta}_i = z_i + \hat{\sigma}_i^2 \hat{\ell}'(z_i), \tag{8.17}$$

$\hat{\ell}(z) = \log(\hat{f}(z))$ and $\hat{\sigma}_i^2 = \sigma^2(z_i)$, where the variance function $\sigma^2(\cdot)$ in (8.14) is calculated numerically. None of this gave answers much different than using (4.1) directly, but the transformation effect becomes more important when n is smaller.

G. Cross-Validation Procedure Both `Ebay` and the shrunken centroids procedure default to 10-fold cross-validation replicated R times. Each replication randomly splits the N cases into 10 folds, with correctly proportional numbers of "healthy" and "sick" in each fold. As usual, the prediction rule is refit 10 times with the cases of each fold withheld from the training set in turn, the cross-validated rate $\hat{\alpha}_{\text{CV}}$ being the overall proportion of errors on the withheld cases averaged over all R replications. The R replications also provide a standard error for $\hat{\alpha}_{\text{CV}}$.

It is useful to remember that $\hat{\alpha}_{\text{CV}}$ is not an estimate of error for the specific prediction rule selected by `Ebay` or `pamr` (unlike the actual prediction errors in Figure 4, which were computed from knowledge of the simulation structure (4.4)). Rather, it is the expected error rate for rules selected according to the same recipe, as emphasized in Efron (1983). In this sense it differs from the ideal Bayesian estimate $\tilde{\alpha} = \Phi(-\|\tilde{\boldsymbol{\delta}}\|/2c_0)$ following (3.1), or its empirical Bayes version $\hat{\alpha} = \Phi(-\|\hat{\boldsymbol{\delta}}\|/2c_0)$, both of which apply directly to the prediction rule at hand.

H. Empirical Bayes Estimation of $\not\Sigma$ The histogram of off-diagonal elements $r_{ii'}$ of a sample correlation matrix will usually be more dispersed than the corresponding histogram of true correlations $\rho_{ii'}$, because sampling error adds a component of variance to the $r_{ii'}$ values. `Ebay` includes an empirical Bayes shrinkage option to account for overdispersion in the estimation of $\not\Sigma$, (5.2).

Let

$$\nu_{ii'} = \frac{\sqrt{n-4}}{2}\log\left(\frac{1+\rho_{ii'}}{1-\rho_{ii'}}\right) \quad \text{and} \quad v_{ii'} = \frac{\sqrt{n-4}}{2}\log\left(\frac{1+r_{ii'}}{1-r_{ii'}}\right) \tag{8.18}$$

denote Fisher's transform of $\rho_{ii'}$ and $r_{ii'}$ (where the usual constant $n-3$ has been reduced to $n-4$ since *two* separate means are subtracted off, for the healthy and sick subjects separately). A standard normal theory approximation (Johnson & Kotz, 1970, Chapt. 32, Sect. 4), says that

$$v_{ii'} \dot{\sim} \mathcal{N}(\nu_{ii'}, 1), \tag{8.19}$$

implying that the histogram of the $v_{ii'}$ values will have variance about one unit greater than that for the true $\nu_{ii'}$'s.

Suppose the ensemble of true $\nu_{ii'}$ values has (mean, variance) say (M, A), and that $v_{ii'} \sim (\nu_{ii'}, 1)$ as in (8.19), so that the $v_{ii'}$ ensemble $\dot{\sim} (M, A+1)$. Then

$$\tilde{\nu}_{ii'} = M\left(1 - \sqrt{C}\right) + \sqrt{C} v_{ii'}, \qquad [C = A/(A+1)] \tag{8.20}$$

is the linear function of $v_{ii'}$ having (mean, variance) $\dot{\sim} (M, A)$.

`Ebay` first obtains robust estimates of M and $A+1$ from the set of values $\{v_{ii'}\}$, and then substitutes $\hat{M}$ and $\hat{C} = \hat{A}/(\hat{A}+1)$ into (8.20) to give estimates $\tilde{\nu}_{ii'}$. In order to protect genuine outliers like those in (5.3), Efron & Morris' (1972) *limited translation rule* is enforced: $\tilde{\nu}_{ii'}$ is not allowed to shrink further than one unit away from $v_{ii'}$. Finally, $\tilde{\nu}_{ii'}$ gives $\tilde{\rho}_{ii'}$ by inverting transformation (8.18). ($\tilde{\not{\Sigma}}$ may no longer be a correlation matrix, but that is not required for use in (5.2).)

A small simulation experiment was run, comparing $\tilde{\not{\Sigma}}$ with the usual (unshrunk) estimate $\hat{\not{\Sigma}}$. It began with model (4.4), modified to instill correlation among the 5000 entries in any one column of X; the root mean square of true pair-wise correlations was set equal to 0.10, about triple that for the prostate study and half that for the Michigan lung cancer study of Table 4. Each of 200 replications yielded $\hat{\boldsymbol{\delta}}_I$ as in (4.2), the $I \times I$ sample correlation matrix $\hat{\not{\Sigma}}$, and its empirical Bayes counterpart $\tilde{\not{\Sigma}}$.

The corresponding estimates (5.2),

$$\hat{\eta} = \left(\hat{\boldsymbol{\delta}}_I^t \hat{\boldsymbol{\delta}}_I / \hat{\boldsymbol{\delta}}_I^t \hat{\not{\Sigma}} \hat{\boldsymbol{\delta}}_I\right)^{\frac{1}{2}} \quad \text{and} \quad \tilde{\eta} = \left(\hat{\delta}_I^t \hat{\delta}_I / \hat{\boldsymbol{\delta}}_I^t \tilde{\not{\Sigma}} \hat{\boldsymbol{\delta}}_I\right)^{\frac{1}{2}} \tag{8.21}$$

are compared with

$$\eta_{\text{true}} = \left(\hat{\delta}_I^t \hat{\delta}_I / \hat{\boldsymbol{\delta}}_I^t \not{\Sigma} \hat{\boldsymbol{\delta}}_I\right)^{\frac{1}{2}} \tag{8.22}$$

in Table 7; $\tilde{\eta}$ is seen to offer only minor improvement over $\hat{\eta}$. Robust estimates of standard deviation for $\tilde{\eta} - \eta_{\text{true}}$ compared to $\hat{\eta} - \eta_{\text{true}}$ were a little more decisive: .074 compared to .085. Root mean square errors for estimating all of the elements of $\not{\Sigma}$ strongly favored $\tilde{\not{\Sigma}}$ over $\hat{\not{\Sigma}}$, $\widetilde{\text{rms}} = 6.30$ versus $\widehat{\text{rms}} = 9.83$.

	η_{true}	$\hat{\eta}$	$\tilde{\eta}$	$\widehat{\text{rms}}$	$\widetilde{\text{rms}}$
mean:	.597	.588	.598	9.83	6.30
stdev:	.138	.171	.164	9.13	5.74

Table 7: Estimates $\hat{\eta}$ and $\tilde{\eta}$, (8.21), compared with true correlation correction factor η_{true}, (8.22); 200 replications of correlated model (4.4); rms values are root mean square errors for estimating the elements of $\not{\Sigma}$.

The $\hat{\alpha}_{\text{cor}}$ values in Table 2 and Table 4 were based on $\hat{\not{\Sigma}}$, `Ebay`'s default option. Using $\tilde{\not{\Sigma}}$ gave smaller estimates of the correlation effect in both cases. The choice is not crucial here since the current version of `Ebay` does not involve $\hat{\alpha}_{\text{cor}}$ in constructing the prediction rule, but either or both methods convey useful information on the effects of correlation among the predictions.

Regularized estimation of correlation matrices is a major subject in its own right (see Warton, 2008), and other methods might further improve on $\hat{\not{\Sigma}}$. However $\hat{\not{\Sigma}}$ performs relatively well in our context for two reasons: the dimension "I" of $\not{\Sigma}$ tends not to be too large, and more importantly, we need only estimate the function η, (5.1), not all of $\not{\Sigma}$. If we are willing to consider $\hat{\boldsymbol{\delta}}$ fixed in (5.2), then $\hat{\boldsymbol{\delta}}^t \not{\Sigma} \hat{\boldsymbol{\delta}}$ is a linear function of $\not{\Sigma}$'s elements, estimated almost unbiasedly by $\hat{\boldsymbol{\delta}}^t \hat{\not{\Sigma}} \hat{\boldsymbol{\delta}}$. The estimation of $\not{\Sigma}$ would be more crucial if we were attempting to implement the general linear discriminant function rather than the simplified version (2.3), (2.4).

I. Overdispersed z-Values The z-value histogram for the prostate data in Figure 1 is a little bit wider than $\mathcal{N}(0,1)$ near $z=0$: a fit to the center of the histogram gave $z \dot{\sim} \mathcal{N}(0,1.06^2)$ (using the *locdfr* algorithm, Efron (2008)). This discrepancy is reflected in Figure 2 by the slight upward slope of $\hat{E}\{\delta|z\}$ for z between -2 and 1.5. Theorem 1 and Corollary 2 in Section 3 depend on the assumption $z \sim \mathcal{N}(\delta,1)$. If actually $z \sim \mathcal{N}(\delta,\sigma^2)$, with $\sigma^2 > 1$, then the formula for $E\{\delta|z\}$ must be modified as in (8.16). We can compensate for overdispersion by using the values $\tilde{z}_i = z_i/1.06$ rather than z_i in the `Ebay` algorithm. Doing so flattened $\hat{E}\{\delta|z\}$ to zero between -2 and 1.5 in Figure 2, and shrank it slightly toward zero for larger $|z|$.

Figure 8 concerns a leukemia microarray study from Golub et al. (1999) where overdispersion is more severe. Here there are $N=7129$ genes measured on $n=72$ subjects in two subtypes, $n_1=45$ and $n_2=12$. Two-sample t-tests gave z-values z_i as in (1.2), (1.3). The histogram of z_i's corresponding to Figure 1 has $z \dot{\sim} \mathcal{N}(0.9,1.68^2)$ near its center.

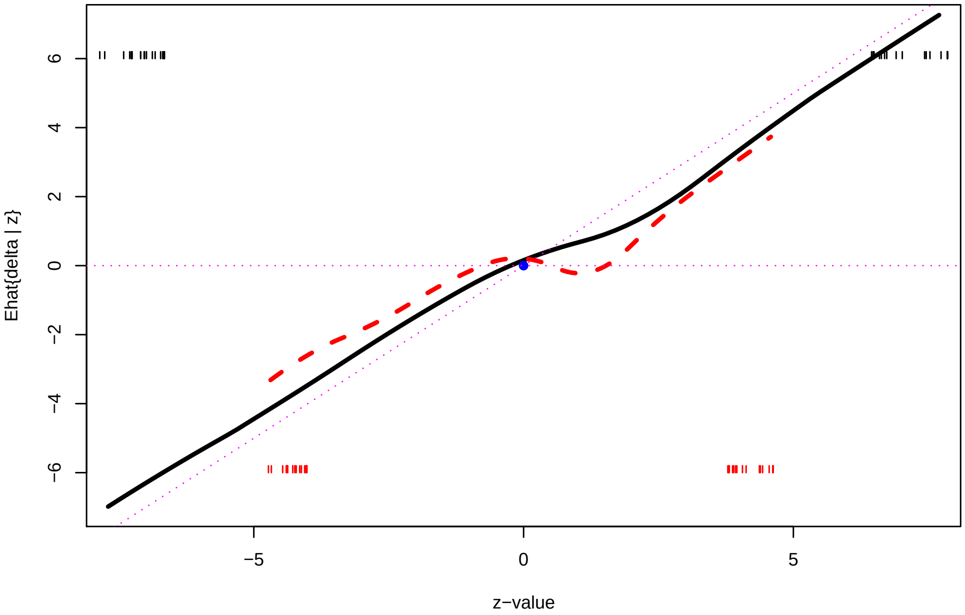

Figure 8: Solid curve is $\hat{E}\{\delta|z\}$ for leukemia data, Golub et al. (1999). Broken curve is $\hat{E}\{\delta|z\}$ based on standardized values $\tilde{z}_i = (z_i - .09)/1.68$. Top row of dashes indicate 40 most extreme z_i values; lower row 40 most extreme $\tilde{z}_i$ values.

Now the curve $\hat{E}\{\delta|z\}$ based on the standardized values $\tilde{z}_i = (z_i - .09)/1.68$ is much less optimistic than that based on the original z_i's, especially taking account of the decreased size of the $\tilde{z}_i$'s. Prediction looks extremely easy with the z_i's; many genes have $|\hat{\delta}_i|$ values, (4.1), exceeding 6. However $|\hat{\delta}_i|$ tops out below 4 for the $\tilde{z}_i$'s. `Ebay` required only $I=10$ genes to reach target error $\alpha_0 = .01$ using the z_i's, (4.2), compared with $I=34$ for the $\tilde{z}_i$'s.

Which prediction rule is better? The answer depends on the reason for the z_i's overdispersion. If in fact $z_i \sim \mathcal{N}(\delta_i,1)$ and the appearance of overdispersion is due to most of the δ_i's lying far from zero, then the $I=10$ rule should perform well. However, overdispersion may indicate ephemeral effects, for example due to unobserved covariates in an observational study, that won't help with future predictions, in which case the $\tilde{z}_i$ analysis is more realistic.

References

Benjamini, Y. & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, *100*(469), 71–93. With comments and a rejoinder by the authors.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.*, *42*, 855–903.

Dawid, A. P. (1994). Selection paradoxes of Bayesian inference. In *Multivariate analysis and its applications (Hong Kong, 1992)*, volume 24 of *IMS Lecture Notes Monogr. Ser.* (pp. 211–220). Hayward, CA: Inst. Math. Statist.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, *18*(1), 71–103.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, *78*(382), 316–331.

Efron, B. (2007). Size, power and false discovery rates. *Ann. Statist.*, *35*(4), 1351–1377.

Efron, B. (2008). Microarrays, empirical bayes, and the two-groups model. *Statist. Sci.*, *23*, 1–47. With comments and a rejoinder by the author.

Efron, B. & Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.*, *67*, 130–139.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, *286*(5439), 531–537.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning* (2nd ed.). Springer Series in Statistics. New York: Springer-Verlag. Data mining, inference, and prediction.

Johnson, N. L. & Kotz, S. (1970). *Distributions in statistics. Continuous univariate distributions. 2.* Boston, Mass.: Houghton Mifflin Co.

Senn, S. (2008). A note concerning a selection “paradox” of Dawid’s. *Amer. Statist.*, *62*, 206–210.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D’Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*(2), 203–209.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, *9*(6), 1135–1151.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, *102*(43), 15545–15550.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, *99*(10), 6567–6572.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Amer. Statist. Assoc.*, *103*(481), 340–349.