EMPIRICAL BAYES ANALYSIS OF A MICROARRAY EXPERIMENT

by

Bradley Efron
Robert Tibshirani
John D. Storey
Virginia Tusher

Technical Report No. 216
May 2001

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California 94305

EMPIRICAL BAYES ANALYSIS OF A MICROARRAY EXPERIMENT

by

BRADLEY EFRON
Department of Statistics
Stanford University

ROBERT TIBSHIRANI
Department of Statistics
Stanford University

JOHN D. STOREY
Department of Statistics
Stanford University

VIRGINIA TUSHER
Department of Biochemistry
Stanford University

Technical Report No. 216
May 2001

This research was supported in part by National Institute of Health grant 2R01 CA72028 and 2R01 CA59039, National Science Foundation grant DMS-9971405 and DMS-0072360

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California 94305

Empirical Bayes Analysis of a Microarray Experiment

Bradley Efron [*]
Robert Tibshirani [†]
John D. Storey, [‡] and Virginia Tusher [§]

April 25, 2001

Abstract

Microarrays are a new technology that enables the simultaneous measurement of thousands of gene expression levels. A typical microarray experiment can produce millions of data points, raising serious problems of data reduction and simultaneous inference. We consider one such experiment in which oligonucleotide arrays were employed to assess the genetic effects of ionizing radiation on seven thousand human genes. A simple nonparametric empirical Bayes model is introduced that is used to guide the efficient reduction of the data to a single summary statistic per gene, and also to make simultaneous inferences concerning which genes were affected by the radiation. The empirical Bayes inferences are closely related to the frequentist False Discovery Rate criterion.

Acknowledgment We are very grateful to Dr. Gilbert Chu of the Stanford Biochemistry Department for sharing his ideas and data with us.

[*]Department of Statistics and Division of Biostatistics, Stanford University, Stanford CA 94305; brad@stat.stanford.edu

[†]Division of Biostatistics and Department of Statistics, Stanford University, Stanford CA 94305; tibs@stat.stanford.edu

[‡]Department of Statistics, Stanford University, Stanford CA 94305; jstorey@stat.stanford.edu

[§]Department of Biochemistry, Stanford University, Stanford CA 94305; tusher@cmgm.stanford.edu

1 Introduction

Through the use of DNA microarrays, a new technology, it is now possible to obtain quantitative measurements of the expression of thousands of genes present in a biological sample. DNA microarrays have now been used to monitor changes in gene expression during important biological processes (e.g. cellular replication and the response to changes in the environment), and to study variation in gene expression across collections of related samples (e.g. tumor samples from patients with cancer). A major statistical task is to understand the structure of the data from such studies, which often consist of measurements on thousands of genes in dozens of conditions.

This paper concerns the use of microarrays in a comparative experiment, where it is desired to compare gene expression under Treatment versus Control conditions. We wish to identify which of several thousand candidate genes have had their expression levels changed, either positively or negatively, by the Treatment. Answering this question requires an efficient data reduction strategy since microarrays deliver megabytes of information, and also statistical inference techniques that deal with the difficulties of simultaneous inference on thousands of genes. We discuss both problems here, working in the context of an experiment on radiation sensitivity discussed below.

The statistics literature for microarrays, still in its infancy and with much of it unpublished, has tended to focus on frequentist data-analytic devices such as cluster analysis, bootstrapping, and linear models, see Li and Wong (2000), Kerr and Churchill (2000), Black and Doerge (2000), Van del Laan et al. (2000), and Eisen et al. (1998). Parametric Bayesian modeling was featured in Newton et al. (2000) and to a lessor extent in Lee et al. (2000). Multiple comparison techniques, designed to control error rates in thousands of simultaneous hypotheses tests, were explored in Dudoit et al. (2000). Tusher et al. (2000) approach the simultaneity problem through the method of False Discovery Rates, as discussed in Section 5.

Our inferences here will be based on a simple nonparametric empirical Bayes model described in Section 3. The model produces useful *a posteriori* probabilities of effect for the individual genes, with a minimum of prior assumptions. It also connects nicely with Benjamini and Hochberg's theory of False Discovery Rates, (1995), as discussed in Section 5. Besides being useful in its own right, the empirical Bayes model helps select among competing data reduction schemes, in Section 4.

Here is some background on microarrays in general and the specific experiment analyzed in this paper. Virtually all living cells contain chromosomes, large pieces of DNA containing hundreds or thousands of genes, each of which specifies the composition and structure of a single protein. Proteins (polymers of amino acids), are the workhorse molecules of the cell, responsible, for example, for cellular structure, producing energy and important biomolecules like DNA and proteins, and for reproducing the cells chromosomes. Every cell in an organism has nearly the same set of chromosomes, and thus contains the same repertoire of proteins. However cells have remarkably distinct properties, such as the differences between human eye cells, hair cells and liver cells, distinctions which are the result of differences in the abundance, distribution and state of the cell proteins. One of the seminal discoveries of molecular biology was that these changes in protein abundance are determined in part by changes in the levels of messenger RNA (mRNA), small and relatively unstable nucleic acid polymers that shuttle information from chromosomes to the cellular machines that synthesize new proteins. Thus there is a logical connection between the state of a cell and the details of its protein and mRNA composition.

While it remains difficult to measure the abundances of a cell's proteins, the recently developed DNA microarray makes it possible to quickly and efficiently measure the relative representation of

each mRNA species in the total cellular mRNA population, or in more familiar terms to measure gene expression levels.

There are two major kinds of microarrays. In an oligonucleotide array, the kind studied in this paper, there are 20 probe pairs (pm, mm) for each gene. The pm (perfect match) probe is designed to match a small subsequence of the gene about 25 bases long. The mm (mismatch) probe is a control, being identical to pm except with the middle base flipped to its complement. An experimental sample is hybridized on the microarray, and the RNA expression of the gene is estimated by the difference in signal pm-mm averaged over the 20 probe pairs. There is some concern that subtracting mismatch numbers may actually degrade the inferences, a question we consider in this paper.

In a spotted cDNA microarray, the other major variety, one base sequence matching all or part of a gene is printed on a glass slide. The experimental sample is labeled with red dye and hybridized on the slide. As a control, a reference sample is labeled with green dye and hybridized on the same slide. Using a fluorescent microscope the log (red/green) intensities of RNA hybridization at each site are measured. The red/green microarray is featured in much of the recent literature, see Newton et al. (2000), Dudoit et al. (2000), and Lee et al. (2000). Our discussion, like that in Li & Wong (2000) centers in the Affymetrix oligonucleotide microarray, but similar analysis problems arise for both types of array. An example extending the empirical Bayes analysis to a cDNA microarray experiment appears in Remark D of Section 6, as an example of how our methods can be applied to other experimental situations.

From either type of microarray we obtain several thousand expression values, one or many for each gene. Microarrays in current use measure anywhere from 1,000 to 25,000 genes; larger ones will soon be available. In a typical study, a number of experimental samples are each hybridized to a different microarray, in order to learn about gene expression differences across different conditions. For example Alizadeh et al. (2000) studied gene expression patterns from tissue sample from a number of lymphoma patients, and related gene expression to patient survival. Clustering methods (Eisen et al. 1998) were the main tool used in that paper, and in a number of other similar studies. Here we will be interested in the more familiar statistical task of comparing Treatment and Control arrays, though carried out in an unfamiliar setting.

Our particular dataset comes from a set of 8 oligonucleotide microarrays in an experiment designed by Professor Gilbert Chu of the Stanford Biochemistry Department to study transcriptional responses to ionizing radiation. Some cancer patents have severe life-threatening reactions to radiation treatment. It is important to understand the genetic basis of this sensitivity, so that such patients can be identified before the treatment is given. The 8 microarrays were labeled

$$(U1A, U1B, I1A, I1B, U2A, U2B, I2A, I2B), \tag{1.1}$$

the labels indicating the following experimental design: RNA was harvested from two wild type human lymphoblastoid cell lines, designated "1" and "2", growing in an unirradiated state "U", or in irradiated state "I". RNA samples were labeled and divided into two identical aliquots for independent hybridizations, "A" and "B". Each microarray provided expression estimates for 6810 genes. Further experimental details appear in Remark A of Section 6.

2 The Data

Microarray experiments produce enormous amounts of data, more than two million feature numbers in the relatively small experiment we are discussing here. The statistical task is to effi-

ciently reduce these numbers to simple summaries of the genes' activities. One goal in this paper is to provide a method for comparing the statistical efficiency of different data reduction strategies.

Here is a description of the data in the radiation experiment, and the notation we will use to describe it. Expression levels were recorded for 6810 different genes,

$$genes: \quad i = 1, 2, \ldots, n = 6810. \tag{2.1}$$

(There were actually 7129 genes, 319 of which had some missing data. For convenience this paper considers only the 6810 genes having complete data. The various analyses were also carried out on all 7129 genes, with nearly identical results.) Each gene on each plate was represented by 20 oligonucleotide "probes",

$$probes: \quad j = 1, 2, \ldots, J = 20 \tag{2.2}$$

Finally there were 8 plates, (the individual microarrays) representing the eight experimental conditions of the experiment described in the Introduction, (U1A, U1B, I1A, I1B, U2A, U2B, I2A, I2B),

$$plates: \quad k = 1, 2, \ldots, K = 8 \tag{2.3}$$

Two features were recorded for each probe of each gene on each plate, a "perfect match number" pm_{ijk} and a "mismatch number" mm_{ijk}, the latter referring to a deliberately distorted version of the oligonucleotide included as a control. Table 1 shows the 20 pairs of numbers for gene $i = 2715$ on plate $k = 1$.

We will investigate three separate stages of data reduction: "probe reduction", the mapping which takes the 20 probe pair numbers into a single expression value "M_{ik}" for gene i on plate k,

$$probe\ reduction: \quad \{(pm_{ijk}, mm_{ijk}), j = 1, 2, \ldots, 20\} \to M_{ik}; \tag{2.4}$$

"gene reduction", the mapping that takes the $K = 8$ expression values M_{ik} for gene i into a single expression score "Z_i",

$$gene\ reduction: \quad \{M_{ik}, k = 1, 2, \ldots, 8\} \to Z_i; \tag{2.5}$$

and finally an *inference mapping* that re-expresses Z_i in terms of a statistical inference concerning gene i's activity. The nonparametric empirical Bayes analysis of Section 3 will provide inferences of the form Prob$\{\text{Event}_i | Z_i\}$, where Event_i is an event of interest such as "gene i's activity was affected by radiation". Section 5 connects these probabilities with the frequentist False Discovery Rate criteria of Benjamini and Hochberg (1995).

There are of course an unlimited selection of possible data reductions from the original data, 320 numbers per gene in the radiation experiment, to the expression scores Z_i. For reasons explained in Section 4 the empirical Bayes analysis will lead us to prefer the following choices: For the probe reduction let

$$M_{ik} = \text{mean}\{\log(pm_{ijk}) - .5 \cdot \log(mm_{ijk}), j = 1, 2, \ldots, 20\}. \tag{2.6}$$

For the gene reduction, first compute the 4 differences $(D_{i1}, D_{i2}, D_{i3}, D_{i4})$ between the irradiated and unirradiated values within the same wildtype sample and aliquot, e.g.

$$D_{i1} = M_{i3} - M_{i1}, \tag{2.7}$$

Table 1: *The 20 pairs of perfect match and mismatch feature numbers for gene $i = 2715$ on plate $k = 1$ (U1A).*

probe	1	2	3	4	5	6	7	8	9	10
pm	1054	3242	1470	4050	1356	1476	561	606	1307	1057
mm	793	2333	826	1912	561	558	942	526	699	1060
probe	11	12	13	14	15	16	17	18	19	20
pm	974	1584	802	1399	1670	2514	2096	6592	5662	2244
mm	829	1771	601	569	840	950	700	8717	1484	668

the difference between the I1A and U1A values M_{ik}. Then take

$$Z_i = \bar{D}_i/(a_0 + S_i) \tag{2.8}$$

where $\bar{D}_i$ is the average of the 4 differences, S_i is their sample standard deviation, and a_0 is the 90th percentile of the 6810 S values. Specifications (2.6)-(2.8) will be used as a comparison point in all of our numerical examples. They will be compared with other choices in Section 4, including the current one included in the Affymetrix software.

3 Empirical Bayes Inferences

Besides analyzing the radiation data, our goal here is to provide data analytic techniques useful in a variety of microarray situations. With generality in mind we will avoid highly specified models, relying instead on a simple inference model that is likely to apply to most comparative experiments: that a gene is either affected on unaffected by the treatment of interest, radiation in our case, giving two possible distributions for the expression score "Z", (2.5). Lee et al. (2000) use a normal theory version of this idea, as, less directly, do Li & Wong (2000). Newton et al. (2000) focus on Gamma models. Here we will avoid parametric assumptions. The resulting nonparametric empirical Bayes analysis, which provides *a posteriori* probabilities of effect for the various genes, is further justified in Sections 4-6.

Let

$$\begin{aligned} p_1 &= \text{probability that a gene is affected} \\ p_0 &= 1 - p_1 = \text{probability unaffected,} \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} f_1(z) &= \text{the density of } Z \text{ for affected genes} \\ f_0(z) &= \text{the density of } Z \text{ for unaffected genes.} \end{aligned} \tag{3.2}$$

Then

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \tag{3.3}$$

is the mixture density of the two populations. In our situation we can estimate $f(z)$ directly from the 6810 expression scores Z_i obtained from the data reduction (2.4), (2.5).

In the absence of strong parametric assumptions such as normality, model (3.3) is useless without an estimate of the "null density" $f_0(z)$. Fortunately it is easy to obtain such estimates. What follows is the method we used to estimate $f_0(z)$ in the radiation experiment. Section 6 discusses variants of this method applicable more generally.

The 6810×8 matrix $\mathbf{M}$ of expression values (2.4), one value for each gene on each plate, gives a 6810×4 matrix $\mathbf{D}$ of differences between the irradiated and unirradiated expression values, as in (2.7). Let $\mathbf{M}_k$ indicate the kth column of $\mathbf{M}$, a 6810 vector. With the plates ordered as before, (U1A, U1B, I1A, I1B, U2A, U2B, I2A, I2B), the "difference matrix" $\mathbf{D}$ is

$$\mathbf{D} = (\mathbf{M}_3 - \mathbf{M}_1, \mathbf{M}_4 - \mathbf{M}_2, \mathbf{M}_7 - \mathbf{M}_5, \mathbf{M}_8 - \mathbf{M}_6) \tag{3.4}$$

Symbolically, the vector $\mathbf{Z}$ of expression scores (2.5) is obtained via

$$\begin{array}{ccccccc} \{\text{original data}\} & \rightarrow & \mathbf{M} & \rightarrow & \mathbf{D} & \rightarrow & \mathbf{Z}. \\ 6810 \times 20 \times 2 \times 8 & & 6810 \times 8 & & 6810 \times 4 & & 6810 \end{array} \tag{3.5}$$

Now let the "null difference matrix" $\mathbf{d}$ be the 6810×4 matrix obtained by differencing within the aliquot splits,

$$\mathbf{d} = (\mathbf{M}_2 - \mathbf{M}_1, \mathbf{M}_4 - \mathbf{M}_3, \mathbf{M}_6 - \mathbf{M}_5, \mathbf{M}_8 - \mathbf{M}_7), \tag{3.6}$$

so for example the first column of $\mathbf{d}$ records differences between the B and A splits of the unirradiated wildtype 1 experiments. We define "null scores" $\mathbf{z} = (z_1, z_2, \ldots, z_{6810})'$ by

$$\{\text{original data}\} \rightarrow \mathbf{M} \rightarrow \mathbf{d} \rightarrow \mathbf{z}, \tag{3.7}$$

with the understanding that except for the substitution of $\mathbf{d}$ for $\mathbf{D}$, the arrows in (3.7) indicate the same mappings as in (3.5).

We will use the empirical distribution of the null scores $\{z_i\}$ to estimate the null density $f_0(z)$ in (3.3). One could just as well take $\mathbf{M}_1 - \mathbf{M}_2$ as $\mathbf{M}_2 - \mathbf{M}_1$ in (3.6), etc., and in fact our numerical algorithm employs random sign permutations of the columns of $\mathbf{d}$ to improve the estimation of f_0. The basic idea here, that we can recover the "null hypothesis" from differences that negate treatment effects, shows up in one form or another in many of the microarray references, being essentially unavoidable in a comparative experiment. Further discussion appears in Section 6, which describes strategies that might be used for estimating f_o in situations less well-balanced than the radiation experiment.

An application of Bayes' rule to the mixture model (3.3) gives the *a posteriori* probabilities $p_1(Z)$ and $p_o(Z)$ that a gene with score Z was affected or unaffected by the treatment,

$$p_1(Z) = 1 - p_o f_o(Z)/f(Z) \quad \text{and} \quad p_o(Z) = p_o f_o(Z)/f(Z). \tag{3.8}$$

The ratio $f_0(Z)/f(Z)$ can be estimated directly from the $\{Z_i\}$ and $\{z_i\}$ empirical distributions. The probabilities p_0 and $p_1 = 1 - p_0$, are unidentifiable without strong parametric assumptions, but this will turn out to be less problematic than it might seem. The constraint that $p_1(Z)$ be nonnegative for all Z does restrict p_0 and p_1,

$$p_1 \geq 1 - \min_Z\{f(Z)/f_0(Z)\} \quad \text{and} \quad p_o \leq \min_Z\{f(Z)/f_o(Z)\} \tag{3.9}$$

A more stable bound for p_1 and p_o is given in Remark G of Section 6.

Figure 1 displays the Bayesian inference curve $p_1(Z) = \text{Prob}\{\text{Event}|Z\}$ obtained from the probe and gene data reductions (2.6), (2.8). It was constructed as follows (skipping some technical details that appear in Section 6):

(a) The 6810 scores $\{Z_i\}$ were computed according to (3.5), using probe reduction (2.6) and gene reduction (2.8).

(b) The null scores $\{z_i\}$ were computed in the same way, but beginning with (3.7) rather than (3.5). (Actually 20 versions of the $\{z_i\}$ were generated, based on 20 independent row-wise sign permutations of **d**, see Remark D.)

(c) A logistic regression technique was used to estimate the ratio $f_0(z)/f(z)$ based on the relative densities of the $\{Z_i\}$ to the $\{z_i\}$.

(d) Relationship (3.9) gave an estimated upper bound for p_o, $p_o \leq .811$.

(e) The solid curve $\text{Prob}\{\text{Event}|Z\}$ is (3.8), with f_0/f estimated from the logistic regression, and p_0 equaling its estimated maximum value .811.

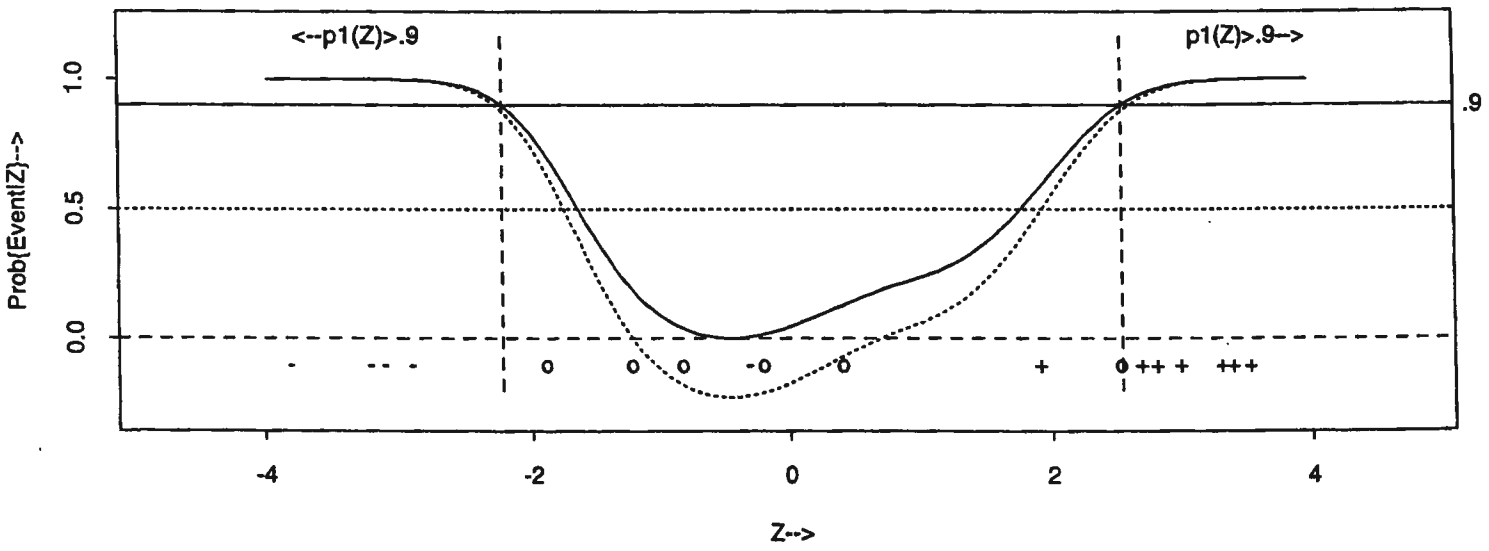

Figure 1: Solid curve: Bayesian inference mapping $\text{Prob}\{\text{Event}_i|Z_i\}$ from data reductions (2.6), (2.8); Event_i is "gene i affected by radiation". Symbols show Z values for 18 genes separately analyzed by Northern Blot: "+" positively affected, "-" negatively affected", "o" not affected. Dotted curve is lower bound (3.10).

Our Bayesian analysis is actually "empirical Bayes" in the sense that the crucial ratio $f_0(Z)/f(Z)$ in (3.8) is estimated from the data rather than from *a priori* assumptions. Newton et al. (2000) carry out a similar analysis, but using specific Bayesian modeling assumptions beyond (3.1)-(3.3).

The *a posteriori* probability of being affected is seen to increase as Z or $-Z$ grows large. The positive end of the Z axis corresponds to genes turned *on* by the radiation, with their expression values increased, while negative Z's indicate decreased expression under radiation. 127 of the 6810 genes had $p_1(Z)$ exceeding 0.90, more on the negative than positive end of the Z scales.

Eighteen of the 6810 genes were independently assessed by a Northern Blot analysis, a pre-microarray assay that serves here as a gold standard for gene expression. Seven of these, indicated by "+" in Figure 1, were deemed "affected positively by radiation", five indicated by "-" were "affected negatively", and six indicated by "o" were "not affected". There is good agreement between the Northern Blot assessments and the probabilities assigned in Figure 1. The full results, given in Section 6, show a high correlation between the gold standard and our results.

In comparing different data reductions, it is convenient to always have the same marginal distribution for Z. To this end, the raw scores $\{Z_i\}$ from (2.8) were monotonically transformed to have a nearly perfect $N(0,1)$ distribution, say by transformation $m(Z)$, and then the null scores were transformed according to the same $m(z)$. Notice that the crucial ratio $f_0(z)/f(z)$ remains the same under such transformations, so that $p_1(Z)$ and $p_o(Z)$ in (3.8) are transformation invariant. We will always make the empirical distribution of the $\{Z_i\}$ almost perfectly $N(0,1)$, using a normal scores transformation, implying for example that $42 = 6810 \cdot (1 - \Phi(2.5))$ of the 6810 genes have $Z_i > 2.5$, with Φ the standard normal cumulative distribution function.

Figure 2 shows the estimates of f_0, f_1, and f contributing to Figure 1; $f(Z)$ is a standard $N(0,1)$ density, by construction, while $f_0(z)$ is a less dispersed density. This is what we hoped for of course: the Z's should be more dispersed than the z's since they reflect the disturbing effects of the radiation treatment. The large values of Prob{Event$|Z$} in the tails of Figure 1 come from (3.8), and the small ratio of $f_0(z)$ to $f(Z)$. A good choice of data reductions makes $f_0(z)/f(z)$ small for $|z|$ large, and we will use this criteria to guide our choices of the probe and gene reductions in Section 4.

Looking again at (3.8),

$$p_1(Z) \geq 1 - f_0(Z)/f(Z), \tag{3.10}$$

since this corresponds to $p_0 = 1$, the largest possible value. The dotted curve in Figure 1 is $1 - f_0(Z)/f(Z)$. This is not much less than the solid curve for large values of $|Z|$, giving 106 genes with $p_1(Z) \geq .90$.

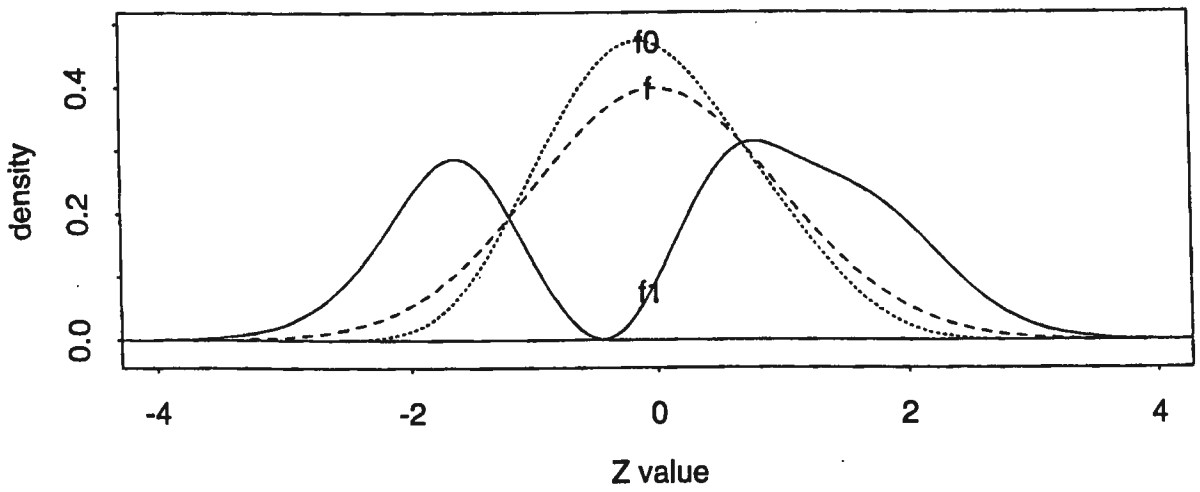

Figure 2: Estimates of $f(\cdot)$, $f_o(\cdot)$ and $f_1(\cdot)$ for the situation of Figure 1, model (3.1)-(3.3); $p_o = .811$, its estimated maximum from (3.9), used in the construction of f_1.

4 Efficient Estimation of Gene Effects

The data reductions (2.6)-(2.8) leading to Figure 1 were chosen to maximize the "power" of the analysis, that is the chance of detecting genes affected by the radiation treatment. This section concerns the rationale for these choices, with further discussion coming in Sections 4-6.

Let $\mathbf{v}_i$ represent all of the data for gene i, so $\mathbf{v}_i$ is a 320-vector in the radiation experiment. As in (3.2)-(3.3) let

$$\begin{aligned} f_1^{\mathbf{v}}(\mathbf{v}) &= \text{density of } \mathbf{v} \text{ for affected genes} \\ f_o^{\mathbf{v}}(\mathbf{v}) &= \text{density of } \mathbf{v} \text{ for unaffected genes} \end{aligned} \tag{4.1}$$

and

$$f^{\mathbf{v}}(\mathbf{v}) = p_o f_o^{\mathbf{v}}(\mathbf{v}) + p_1 f_1^{\mathbf{v}}(\mathbf{v}), \tag{4.2}$$

the latter being the mixture density. Notice that p_o and p_1 have the same meaning here as in (3.1).

Defining the likelihood ratio statistic

$$R^{\mathbf{v}}(\mathbf{v}) = f^{\mathbf{v}}(\mathbf{v})/f_o^{\mathbf{v}}(\mathbf{v}), \tag{4.3}$$

Bayes' theorem gives the analogue of (3.8),

$$p_1^{\mathbf{v}}(\mathbf{v}_i) = 1 - p_o/R^{\mathbf{v}}(\mathbf{v}_i) \quad \text{and} \quad p_o^{\mathbf{v}}(\mathbf{v}_i) = p_o/R^{\mathbf{v}}(\mathbf{v}_i) \tag{4.4}$$

for the *a posteriori* probability of gene i being affected or unaffected by the treatment. An informative microarray experiment is one that produces some large values of $R^{\mathbf{v}}(\mathbf{v}_i)$, i.e. values of $p_i^{\mathbf{v}}(\mathbf{v}_i)$ near 1, and a familiar omnibus information measure is the *Kullback-Leibler distance*

$$KL^{\mathbf{v}} = \int f^{\mathbf{v}}(\mathbf{v}) \log R^{\mathbf{v}}(\mathbf{v}). \tag{4.5}$$

In our situation it isn't practical to estimate the high-dimensional densities $f^{\mathbf{v}}(\mathbf{v})$ and $f_o^{\mathbf{v}}(\mathbf{v})$, at least not without extensive modeling. However we can easily estimate the corresponding densities for a one-dimensional statistic $\mathbf{Z} = s(\mathbf{v})$, as was done in Section 3. Unless Z is sufficient, information will be lost in going from $\mathbf{v}$ to Z. Writing $\mathbf{v} = (Z, \mathbf{u})$, where $\mathbf{u}$ is "everything else besides Z" in the specification of $\mathbf{v}$, a standard calculation describes the loss of Kullback-Leibler distance by

$$KL^{\mathbf{v}} - KL^{Z} = \int f(Z) KL^{\mathbf{u}|Z}, \tag{4.6}$$

$KL^{\mathbf{u}|Z}$ being the conditional KL distance for $\mathbf{u}$ given Z, and

$$KL^{Z} = \int p(Z) \log R(Z) \quad \text{with} \quad R(Z) = f(Z)/f_o(Z), \tag{4.7}$$

$R(Z)$ the likelihood ratio based on Z.

With this background in mind we searched for mappings $Z = s(\mathbf{v})$ that produced large values of $R(Z)$, i.e. good separation between $f(Z)$ and $f_o(Z)$ as in Figure 2. Figure 3 shows part of the search. The curve marked "90" is equivalent to the dashed curve in Figure 1, the difference here being that the vertical axis is plotted on the logit scale, $\log p_1(Z)/(1 - p_1(Z))$, to emphasize

differences in the tails. Keeping probe reduction (2.6) fixed, Figure 3 compares 5 different choices of a_o in the gene reduction (2.8): a_o equal to the 90th percentile of the 6810 S_i values; the 50th percentile; the 5th percentile; $a_0 = 0$; and $a_0 \to \infty$. The choice $a_0 = 0$ makes Z_i in (2.8) proportional to the one-sample t-statistic for the four differences $(D_{i1}, D_{i2}, D_{i3}, D_{i4})$, while $a_0 \to \infty$ makes Z_i equivalent to the numerator $\bar{D}_i$. The plotted curves are the logits of (3.10), the conservative lower bound for $p_1(Z)$, (taking $p_o = 1$ in (3.8).)

Figure 3 shows that the best choice for a_0 is the one we used before, a_0 the 90th percentile. This manifests itself as higher values of Prob{Event|Z} at both ends of the Z scale. The density $f_0(z)$ in Figure 2 is more concentrated around zero than it is say for the disastrous choice $a_0 = 0$, raising $f(Z)/f_o(Z)$ in the tails and thus $p_1(Z)$, (3.10). The numbers N90 in figure 3 indicate the number of genes having lower bound (3.10) for $p_1(Z_i)$ greater than .90. These range downward from 106 for $a_o = 90$ to 0 for $a_o = 0$. For our purposes, N90 is a more relevant indicator of performance than KL^Z, (4.7) better indicating large tail values of $R(Z)$.

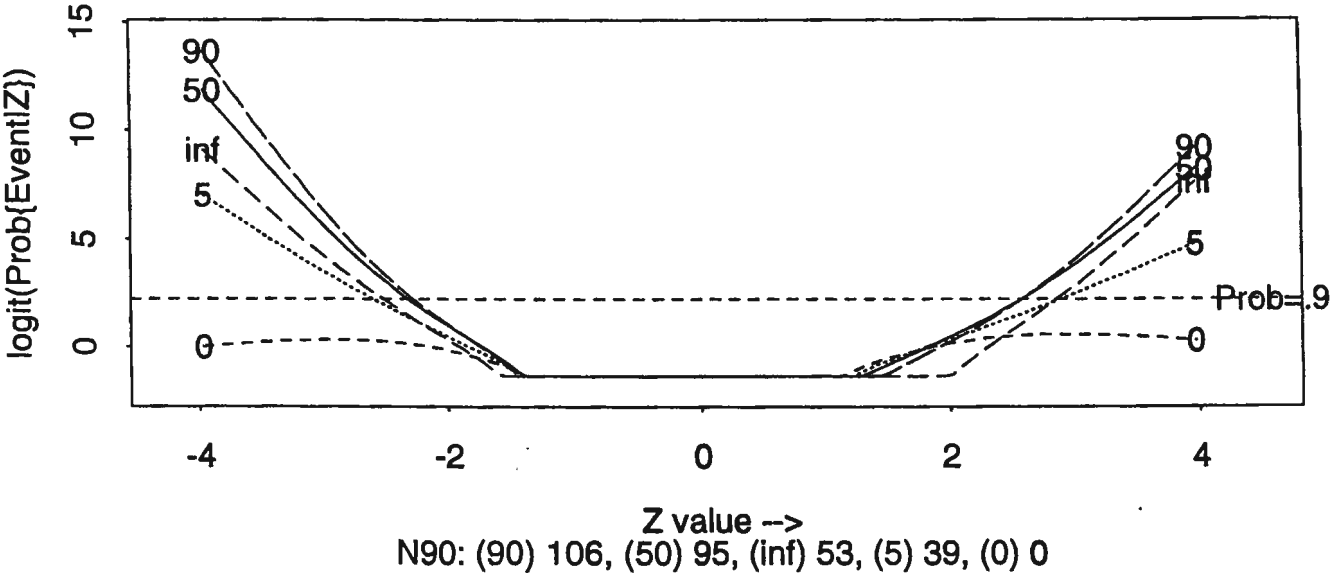

Figure 3: Choice of a_o in the gene mapping $Z_i = \bar{D}_i/(a_o + S_i)$, (2.8); vertical axis is logit of Prob{Event|Z}, estimated as in (3.8) with $p_o = 1$; "90" indicates a_o equaling 90th percentile of the 6810 S_i values, etc.; "inf" is limit as $a_o \to \infty$. We see that 90 is the best choice in terms of maximizing prob{Event|Z} for large $|Z|$; $a_o = 0$ is worst. All choices used probe reduction (2.6). The vertical axis is truncated at lower bound Prob{Event|Z} = .20. **N90** is the number genes having Prob{Event|Z} $\geq$ .90.

Now keeping the gene reduction fixed as in (2.6), $a_o = .90$, Figure 4 compares probe reductions of the form

$$M_{ik} = \underset{j}{\text{mean}}\{s(pm_{ijk}) - c \cdot s(mm_{ijk})\}, \tag{4.8}$$

with s either the log function or the identity function. For example curve 2 in the left panel uses $M_{ik} = \underset{j}{\text{mean}}\{pm_{ijk} - mm_{ijk}\}$ while the dotted curve in the right panel uses $M_{ik} = \underset{j}{\text{mean}}\{\log(pm_{ijk})\}$. Our preferred choice (2.6)-(2.8) is curve 1, "$c = .5$ & logs". The "Affy" curve in the left panel was based on the algorithm provided by Affymetrix, which is similar to the "$c = 1$ no logs" choice, but with a provision for removing apparent outliers among the 20 $pm_{ijk} - mm_{ijk}$ differences before averaging.

Figure 4 indicates a substantial advantage to taking logs, and a mild advantage to using $c = .5$ rather than $c = 1$ or $c = 0$. The comparison between $c = .5$ and $c = 1$ is close on the log scale, but other comparisons, reported in Efron et al. (2000), reinforce the superiority of $c = .5$. We also tried using various L-estimators in (4.7), including trimmed means. When applied on the log scale this form of robustification made almost no difference to our results.

Some comments are in order:

- There is no claim that the mapping $Z = s(\mathbf{v})$ described by (2.6), (2.8) is "correct", only that it is relatively efficient in preserving the information in $\mathbf{v}$. The estimated curve $p_1(Z)$ still is meaningful as the *a posteriori* probability of effect given the insufficient statistic Z, and also has the FDR interpretation of Section 5. (Efron et al. (2000) show that in fact a better "$s(\mathbf{v})$" can be obtained in the radiation experiment by removing plates U1A and I1A from the data set (1.1); a processing error appears to have degraded the results from I1A.) Our tactic of choosing the Z mapping to maximize $p_1(Z)$ is nearly equivalent to minimizing FDR, which was the approach taken in Tusher et al. (2000).

- There is also no claim that (2.6), (2.8) enjoys general superiority. The equivalents of Figures 3 and 4 might point to a different choice of $s(\mathbf{v})$ in another data set. Section 6 discusses how our methodology can be applied to other comparative microarray experiments.

- Overfitting is not a threat in a genuine Bayesian framework, where results like those from Figure 3 and 4 can be thought of as just a computer-based attempt to numerically solve a probabilistic maximization problem. However in our empirical Bayes framework too much data-based maximization could in fact lead to overfitting. Two forms of bootstrapping were employed as a check on our results: "gene resampling", in which the rows of the 6810×8 matrix $\mathbf{M}$ were resampled to give $\mathbf{M}^*$; and "row resampling" in which row i of $\mathbf{M}^*$ was obtained as the average of 20 resampled rows from the 20×8 matrix $\mathbf{x}_i$ having entries

$$x_{ijk} = \log(pm_{ijk}) - .5 \cdot \log(mm_{ijk}). \tag{4.9}$$

The bootstrap results indicated that the differences seen in Figure 3 and 4 were much greater than the standard errors of the curves, so that overfitting was not a threat. For example row resampling showed that the difference between the $a_o = .90$ and $a_o = .50$ curves at $Z = -3$, which looks suspiciously small in Figure 3, had point estimate and standard error 0.68 ± 0.13.

5 False Discovery Rates

The empirical Bayes analysis of Section 3 is closely related to Benjamini and Hochberg's False Discovery Rate (FDR) criterion. For a collection of simultaneous hypothesis tests, FDR is the expected proportion of type I errors made using a given rejection rule. Define the *local false discovery rate* to be

$$\text{fdr}(Z) = p_o f_o(Z)/f(Z), \tag{5.1}$$

so $\text{fdr}(Z)$ is the *a posteriori* probability (3.8) that a gene with score Z is unaffected. It will be shown that (5.1) has a natural FDR interpretation. We begin with a numerical example.

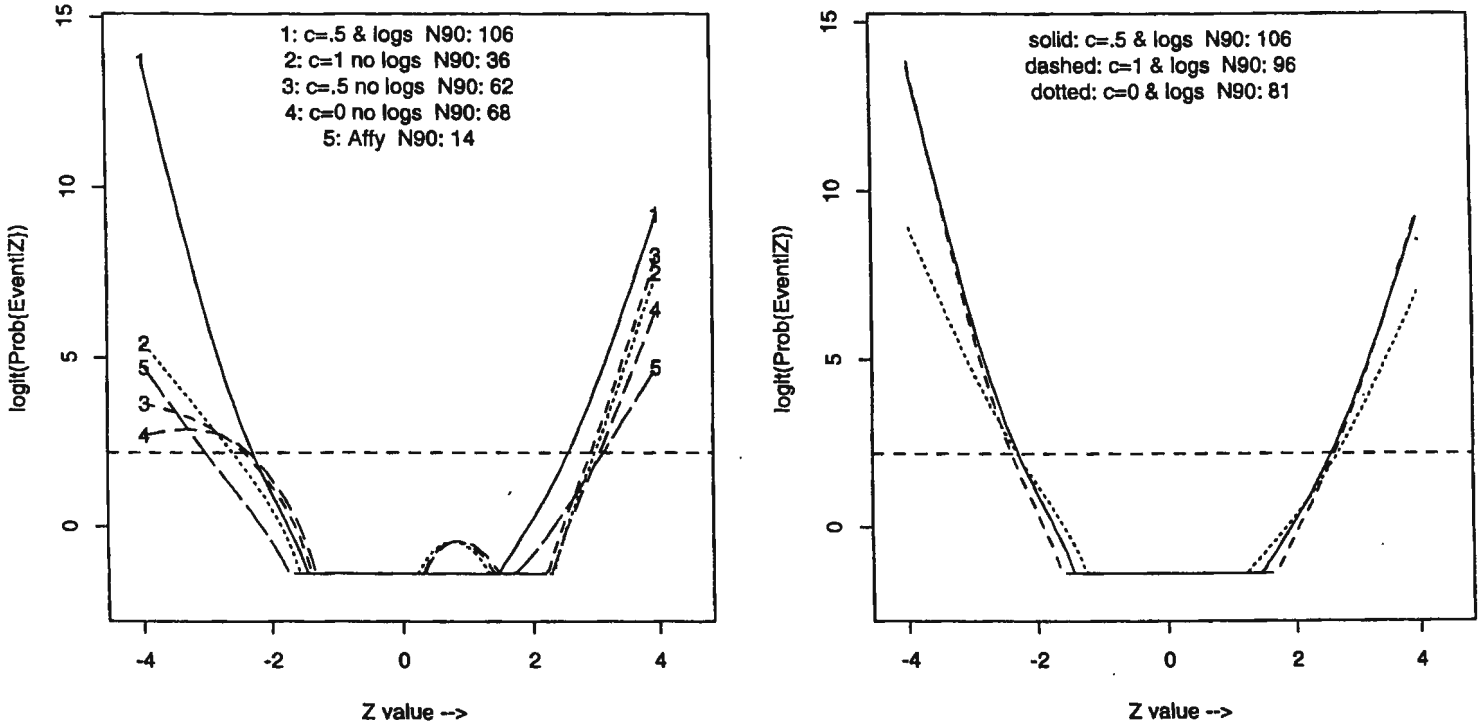

Figure 4: Comparison of various probe reductions (gene reduction fixed as in (2.8), a_0 = 90th percentile). The solid curve in both panels is the choice (2.6) used previously; constant "c" is multiple of mm level subtracted from pm level, e.g. "$c = 1$ no logs" uses $M_{ik} = \text{mean}\{pm_{ijk} - mm_{ijk}\}$. "Affy" based on the probe reduction software provided with the Affymetrix Genechip.

In the calculations for Figure 1, $N = 74$ of the 6810 genes had Z scores in the interval $Z \in [1.9, 2.1]$, while the twenty permuted null score data sets $\{z_i\}$ had 676 falling into $[1.9, 2.1]$, an average of $33.8 = 676/20$ per set. Taking p_o to be its estimated maximum .811, this suggests that among the $N = 74$ binned Z values, the expected number of "unaffecteds" is $27.4 = .811 \cdot 33.8$. If we now declare all genes with Z in $[1.9, 2.1]$ to be affected, our expected proportion of false discoveries is

$$27.4/74 = 37\%. \tag{5.2}$$

Notice that (5.2) is the obvious estimate of (5.1) for $Z = 2$,

$$\widehat{\text{fdr}}(2) = \widehat{p}_o \widehat{f}_o(2)/\widehat{f}(2) \qquad \left[\widehat{p}_o = .811, \quad \widehat{f}_o(2) = \frac{33.8}{6810}, \quad \widehat{f}(2) = \frac{74}{6810}\right]. \tag{5.3}$$

In general if we bin the genes into small intervals on the Z scale, then a bin declared "affected" will have a False Discovery Rate of about $\text{fdr}(Z)$, (5.1), the equality becoming exact as the number of genes goes to infinity. This last statement can be rigorously verified under modest ergodic conditions that preclude extremely high correlations among the Z's or the z's. (A possible difficulty with (5.1) is discussed in Remark F of Section 6.)

Figure 5 reports on a simulation experiment used to check the accuracy of $\text{fdr}(Z)$ as an estimate of FDR. A 6810×8 matrix $\mathbf{M}$ was constructed in a way that mimicked the radiation experiment,

$$M_{ik} = \theta_i t_k + \epsilon_{ik} \qquad [\epsilon_{ik} \overset{\text{ind}}{\sim} N(0,2)], \tag{5.4}$$

$(t_1, t_2, \ldots, t_8) = (0,0,1,1,0,0,1,1)$; 681 of the "gene effects" θ_i were chosen from a $N(-1.5, 1)$ distribution 681 from $N(1.5, 1)$, and the remaining 5448 set at zero. In other words 80% of the genes were unaffected and 20% were affected, 10% in each direction.

The matrix $\mathbf{M}$ was processed into a $\mathbf{Z}$ vector according to (3.5) and also twenty $\mathbf{z}$ vectors according to (3.7), using (2.8) for the mappings from $\mathbf{D} \to \mathbf{Z}$ and $\mathbf{d} \to \mathbf{z}$. Following the same algorithm that lead to Figure 1, these gave an estimated fdr(Z) curve, (5.1), that in fact looked much like the one for the actual experiment.

Figure 5 reports on two different choices of a_o in (2.8), $a_o = .90$ in the left panel and $a_o \to \infty$ on the right. There are 100 points in each panel, corresponding to a binning of the fdr axis in units of .01 from 0 to 1, with the jth point plotted at $\text{fdr}_j = j/100$ and

$$\text{FDR}_j = \text{proportion of } j^{\text{th}} \text{ bin with } \theta_i = 0, \tag{5.5}$$

the empirical False Discovery Rate for that bin. The FDR_j values are averages over 50 simulations, each of the individual simulations being noisy versions of the same picture. If formula (5.1) is actually the local false discovery rate then the points should lie near the main diagonal FDR = fdr as they do. The slight conservative bias $\text{FDR}_j \leq \text{fdr}_j$, came from the fact that the upper bound for p_o used in (5.1) (calculated as in Remark G of Section 6) substantially overestimated the true value $p_o = .80$.

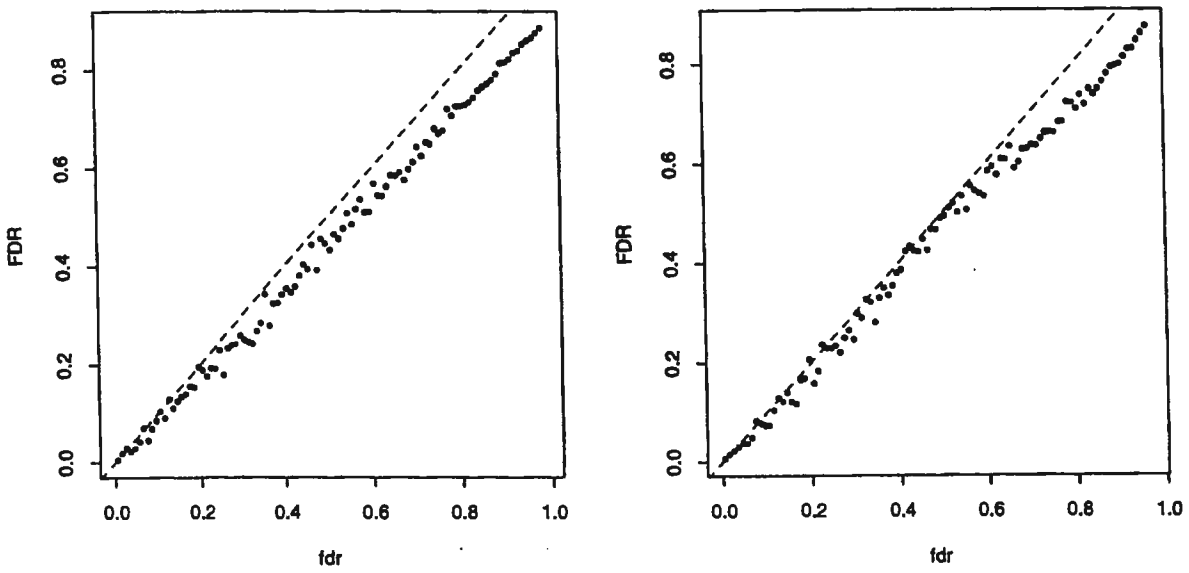

Figure 5 Simulations comparing empirical Bayes formula fdr, (5.1) with actual False Discovery Rate FDR, as explained in text. *Left panel* $a_o = .90$ in mapping (2.8). *Right panel* $a_o \to \infty$.

In the artificial situation (5.4), taking $a_o \to \infty$ in (2.8) gives a more efficient choice of Z_i than $a_o = .90$, doubling both $N90$ and the Kullback-Leibler distance (4.7). Nevertheless the inefficient choice $a_o = .90$ still is accurately calibrated: $\text{FDR}_j \doteq \text{fdr}_j$.

False Discovery Rates are usually defined for an entire rejection region, e.g. for

$$\mathcal{R} = \{Z : R(Z) \geq r_o\} \qquad [R(Z) = f(Z)/f_o(Z)]. \tag{5.6}$$

rather than locally as in (5.1). We can think of this as replacing our original choice of summary statistic $Z = s(\mathbf{v})$ with

$$\widetilde{Z} = \begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} Z \in \mathcal{R} \\ Z \notin \mathcal{R}\,. \end{matrix} \tag{5.7}$$

Assuming that genes having $\widetilde{Z} = 1$ are declared effected, the empirical Bayes formula (5.1) now becomes

$$\widetilde{\text{fdr}}(1) = p_o \widetilde{f}_o(1)/\widetilde{f}(1), \tag{5.8}$$

with straightforward estimate

$$\widehat{p}_o \, \frac{\text{proportion}\{z_i \in \mathcal{R}\}}{\text{proportion}\{Z_i \in \mathcal{R}\}}. \tag{5.9}$$

The heuristic argument proceeding (5.2) is more obvious here: (5.9) estimates the proportion of genes in the "effected" region $\mathcal{R}$ that are actually unaffected, and in this sense estimates Benjamini and Hochberg's FDR.

The global definition of FDR has the advantage of being easier to estimate. We can use the totally nonparametric estimator (5.9) rather than having to estimate the local ratio $f_o(Z)/f(Z)$ in (5.1). On the other hand (5.8) is a composite measure that assigns the same FDR to all the genes in $\mathcal{R}$ even though some of them have Prob{Event$|Z$} much greater than others. Comparing (5.1) with (5.8) it is easy to see that $\widetilde{\text{fdr}}(1)$ is the conditional expectation of fdr(Z) given $Z \in \mathcal{R}$,

$$\widetilde{\text{fdr}}(1) = E_f\{\text{fdr}(Z)|\mathcal{R}\}. \tag{5.10}$$

so that fdr(Z) is more precise than $\widetilde{\text{fdr}}(1)$. More on the connection between FDR and fdr appears in Storey (2001).

6 Remarks and Details

A. *The Experiment* Lymphoblastoid cell lines GM14660 and GM08925, (Coriell Cell Repositories, Camden New Jersey) were seeded at 2.5×105 cells/ml. The treatment consisted of 5 Gy of ionizing radiation. After 24 hours, RNA was isolated, labeled, and divided into two aliquots that were independently hybridized to the HuGeneFL Genechip microarray, Affymetrix Corporation.

B. *Northern Blot Analysis* Northern Blot Analysis produced a quantitative score "G_i" for each of the 18 genes indicated in Figure 1, G standing for Gold Standard. G scores exceeding 1.30 were taken to indicate a positive effect of radiation on gene activity, the "+" symbols in Figure 1; likewise "$-$" for $G_i < 0.70$ and "o" for $0.70 \le G_i \le 1.30$. Figure 6 compares the Z_i scores from Figure 1 with $\log G_i$ for the 18 test genes. We see a strong monotone relationship, correlation 0.87.

The agreement in Figure 6 is impressive, especially considering the magnitude of the sampling errors in the individual expression values. Our gold standard, the Northern Blot score, is not pure gold, itself being subject to experimental error. There is only one flagrant disagreement in Figure 6, the "$-$" gene at $Z_i = -0.31$. The vector of differences (3.4) was $\mathbf{D}_i = (-1.59, 0.55, 0, 88, -0.83)$ for this gene, so that both wildtypes yielded aliquots of opposing signs. In contrast the "o" point at $Z_i = 2.51$, lying just below the "+" cutoff value $G = 1.30$, was consistently positive, $\mathbf{D}_i = (4.54, 2.81, 1.64, 2.65)$, strengthening our belief that this gene was positively affected by the radiation.

C. *Debrightening and Desumming* Some microarray plates are "brighter" than others in that they produce systematically larger expression levels. Following probe reduction (2.4) we debrightened

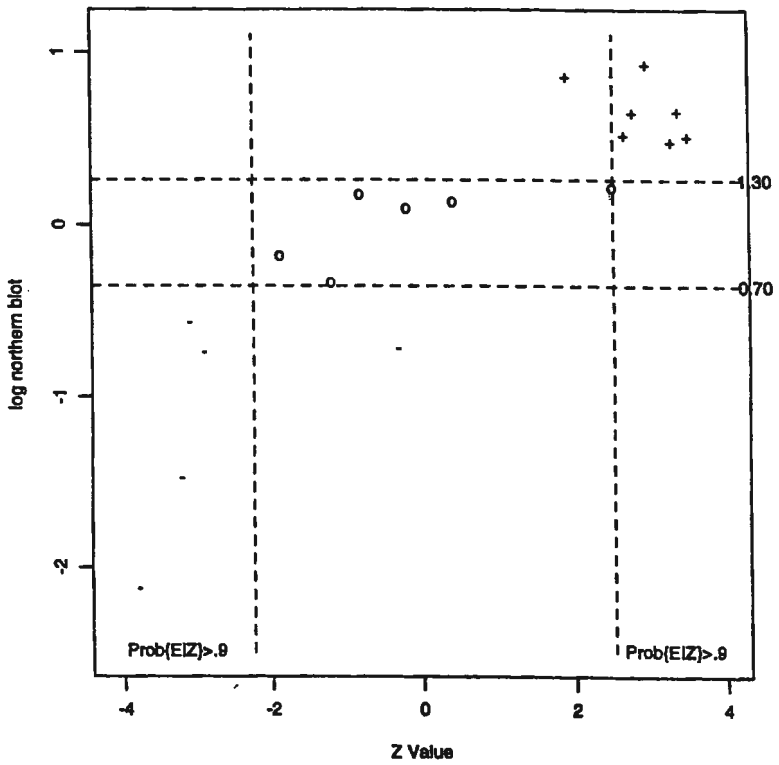

Figure 6: Comparison of Z scores from the analysis in Figure 1 with the logarithm of the Northern Blot results. Correlation 0.87. Z values outside the two vertical lines have $\text{Prob}\{\text{Event}_i|Z_i\} \geq .90$.

the data by separately standardizing the columns of $\mathbf{M}$. That is, each column of $\mathbf{M}$ was linearly transformed to have mean 0 and empirical standard deviation 1.

"Desumming" corrects for another type of data inhomogeneity. Corresponding to $\mathbf{D}$ (3.4), let

$$\mathbf{S} = (\mathbf{M}_3 + \mathbf{M}_1, \mathbf{M}_4 + \mathbf{M}_2, \mathbf{M}_7 + \mathbf{M}_5, \mathbf{M}_8 + \mathbf{M}_6) \tag{6.1}$$

A gene with larger $\mathbf{S}$ values tended to have larger values of $\mathbf{D}$, which undercut the exchangeability across genes implicit in our empirical Bayes analyses. (Newton et al. (2000) adjust their data for a similar problem.) After debrightening, the individual columns of $\mathbf{D}$ were desummed as follows: a linear regression $|D_{ik}| = a_0 + a_1|S_{ik}| + \text{error}$ was fit individually to each column, and then each D_{ik} was transformed to

$$D_{ik}/(\widehat{a}_0 + \widehat{a}_1|S_{ik}|). \tag{6.2}$$

Similar transformations were made on the columns of $\mathbf{d}$, (3.6). It was the transformed $\mathbf{D}$ and $\mathbf{d}$ matrices that were used to compute the scores $\mathbf{Z}$ and $\mathbf{z}$ via (2.8). Desumming made almost no difference to the results in Figure 1, but the exchangeability issue is important general point of concern for the empirical Bayes analysis, see Remark F.

D. *Logistic Regression Estimate of* $f_0(z)/f(z)$. The ratio $f_0(z)/f(z)$ in (3.8) was estimated by logistic regression. Given $B = 20$ replications of $\mathbf{z}$, all $n \cdot (1 + B) = 6810 \cdot 21$ scores Z_i and z_i were plotted on a line, with Z_i's considered as "successes" and $\mathbf{z}$'s as "failures". The probability $\pi(z)$ of a success at point z is given in terms of the densities (3.2),

$$\pi(z) = f(z)/(f(z) + Bf_0(z)), \tag{6.3}$$

so that (3.8) becomes

$$p_1(Z) = 1 - p_0\frac{1 - \pi(Z)}{B\pi(Z)}. \tag{6.4}$$

With $n = 6810$ genes, the normal scores transformation resulted in $\max\{Z_i\} = -\min\{Z_i\} = 3.80$, while the null scores $\{z_i\}$ were confined to a smaller range, as in Figure 2. Our algorithm divided the range $[-4, 4]$ into 139 equal intervals, counted the number of Z_i's and z_i's in each interval, and estimated $\pi(z)$ by logistic regression, for use in (6.4). The regression function was a natural spline with 5 degrees of freedom, called by the Splus command $ns(x, df = 5)$, x being the 139 center points of the intervals. The choice of $B = 20$ **z** replications was based on an analysis like that in Figure 3, which showed considerable improvement for B increasing from 1 to 10, but little gain past 20. Other methods of estimating $f_o(z)/f(z)$ is possible, and in fact the details of the logistic regression method made little difference to our results. The "global" estimate (5.9) avoids regression entirely, at the expense of providing less specific results.

E. *Estimating the Null Distribution* The null density $f_0(z)$ is supposed to describe the distribution of expression scores for genes unaffected by the treatment of interest. Basing f_0 on **d** in (3.6) seems natural for the radiation experiment, but other choices are possible and may be necessary for other experimental designs. Table 2 shows a small portion (5 of 2638 genes) of the data from a microarray study comparing two different types of liver cancer, 37 Type I patients versus 36 Type II, with Type II having worse prognosis. Spotted cDNA arrays were used, the "red-green" variety, the tabled values being $1000 \cdot \log$ (red/green) intensity ratios. The full table corresponds to matrix **M** in (3.5), now 2638×73. Probe reduction (2.8) is very simple here, (red, green) $\to \log$(red/green), though more efficient reductions may be possible as shown in Dudoit et al. (2000). The analogue of Figure 3 indicated a preference for $a_o = 0$, i.e. for taking Z to be the two-sample t-statistic between the Types.

	TYPE 1						TYPE 2				
	pat1	pat2	pat3	pat4	pat5	...	pat38	pat39	pat40	pat41	pat42
GENE1	230.0	-1350	-1580.0	-400	-760	...	970	110	-50	-190.0	-200
GENE2	470.0	-850	-0.8	-280	120	...	390	-1730	-1360	-0.8	-330
GENE3	-920.0	-1070	1360.0	-510	-1120	...	70	-1150	340	-400.0	580
GENE4	0.1	380	730.0	180	-90	...	1040	180	1070	250.0	1880
GENE5	390.0	-1960	-210.0	200	230	...	530	-1170	670	0.7	890

Table 2 Some data from a microarray study comparing two types of liver cancer.

If both groups had had 36 patients we could obtain null scores quite easily; randomly split the Type I patients into two groups of 18 each, say "A" and "B", and likewise "C" and "D" for the Type II patients, and define the z's as t-statistics between groups AUC versus BUD. (In other words, use balanced permutations that put equal numbers of Type I's and Type II's into each of the two permuted groups; using unbalanced permutations would add an unwanted component of variance to the null scores.)

This was done for the actual data set by omitting a randomly selected 37th Type I patient in each permutation. To make the correspondence between $f(Z)$ and $f_o(z)$ exact, the definition of Z was changed in a similar way: to be the t-statistic between groups AUB versus CUD, with one Type I patient omitted. In this case we generated multiple realizations of the sets $\{Z_i\}$ as well as of the null scores $\{z_i\}$. The empirical Bayes analysis produced results similar to those in Figure 2, with Type II playing the role of the Treatment group.

As a simple but informative model for the radiation experiment, suppose that M_{ik} in (3.5) can

be expressed as

$$M_{ik} = \mu_i + \alpha_i w_k + \theta_i t_k + \epsilon_{ik}, \tag{6.5}$$

where $\mathbf{t}' = (0,0,1,1,0,0,1,1)$, $\mathbf{w}' = (-1,-1,-1,-1,1,1,1,1)$, and ϵ_{ik} is an independent noise term. Here θ_i represents the treatment effect while α_i is the differential response for gene i between the first and second wildtypes. Then Z_i in (2.8) is

$$Z_i = (\theta_i + e_i)/(a_o + S_i), \qquad S_i = \left[\sum_{\ell=1}^{4}(e_{i\ell} - e_i)^2/3\right]^{\frac{1}{2}}, \tag{6.6}$$

where each $e_{i\ell}$ is the difference of two ϵ_{ik}'s and e_i is the average of the four $e_{i\ell}$'s. The z_i's have the same expression except that $\theta_i = 0$ in (6.6). We can see that $f_o(z)$ is a legitimate null hypothesis comparator for $f(Z)$.

Suppose we had defined null scores by differencing across wildtypes instead of across aliquots: $\mathbf{d} = (\mathbf{M}_5 - \mathbf{M}_1,\ \mathbf{M}_6 - \mathbf{M}_2,\ \mathbf{M}_7 - \mathbf{M}_3,\ \mathbf{M}_8 - \mathbf{M}_4)$ replacing (3.6). Then z_i would pick up an additional term due to the gene/wildtype interaction α_i in (6.5), adding a component of variance to $f_o(z)$, and decreasing the likelihood ratio $f(z)/f_o(z)$. Models like (6.5) are helpful in guiding the choice of the Z and z mappings, even if we do not need them for the data-based estimation of f and f_o.

The additive model (6.5) gives every column of $\mathbf{d}$ the same distribution but we might not trust the Treatment differences to really have the same distribution as the Control, I2B-I2A compared to U2B-U2A for example. Empirically this turned out not to be a problem for the radiation experiment, but if it had we might have used only the first and third columns of $\mathbf{d}$ in (3.6).

F. *Exchangeability Considerations* The empirical Bayes formulas (5.1) or (3.8), $\text{fdr}(Z) = p_o f_o(Z)/f(Z)$, implicitly depend on an exchangeability assumption among the null scores. As a counterexample suppose that genes that are greatly affected by the treatment tend to have large values of $|z_i|$. In the numerical example (5.3) this could mean that *less* than 81.1% of the z scores in $[1.9, 2.1]$ were from "unaffecteds", implying that (5.1) was biased downward.

As a check on the radiation experiment, we regressed $|z_i|$ on Z_i. In fact the estimated regression was almost perfectly flat, indicating absence of the counterexample effect, but if it had been there we would have needed to introduce regression corrections into the calculation of $f_o(Z)/f(Z)$. The debrightening adjustment of Remark C was introduced in this spirit, though it turned out not to affect the estimation of $f_o(Z)/f(Z)$ in this case.

G. *Better Upper Bound Estimates for "p_o"* The upper bound (3.9), $p_o \le \min\{f(Z)/f_o(Z)\}$, can be poorly estimated by the choice $\min\{\widehat{f}(Z)/\widehat{f}_o(Z)\}$ used in Figure 1. More stable upper bounds can be constructed by integrating over an interval "$\mathcal{A}$" near $Z = 0$,

$$p_o \le \frac{\int_{\mathcal{A}}[f(Z)/f_o(Z)]f_o(Z)}{\int_{\mathcal{A}} f_o(Z)} = \frac{\int_{\mathcal{A}} f(Z)}{\int_{\mathcal{A}} f_o(Z)}. \tag{6.7}$$

Simulation showed that the choice $\mathcal{A} = [-.5, .5]$ performed better than $\min\{f(Z)/f_o(Z)\}$, particularly when the true p_o was near 1. The upper bound (6.7) is directly estimated by proportion $\{Z_i$'s in $\mathcal{A}\}$/proportion$\{z_i$'s in $\mathcal{A}\}$, avoiding the logistic regression estimate for $f(Z)/f_o(Z)$. This gave $p_o \le .825$ in the context of Figure 1, not much different than the previous estimate $p_o \le .811$.

References

Alizadeh, A., Eisen, M., Davis, R.E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., J., P., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R., Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000), "Identification of molecularly and clinically distinct substypes of diffuse large b cell lymphoma by gene expression profiling", *Nature* **403**, 503-511.

Benjamini, Y. & Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Jour. Royal Stat. Soc., B* **57**, 289-300.

Black, M. & Doerge, R. (2000), "Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change for cDNA microarrays". Dept. of Statistics, Purdue University.

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000), "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments". Technical Report, Dept. Statistics, U. Cal. Berkeley.

Efron, B., Tibshirani, R., Goss, V., & Chu, G. (2000), "Microarrays and their use in a comparative experiment". Stanford Technical Report #213.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Nat. Acad. Sci.* **95**, 14863-14868.

Kerr, K. & Churchill, G. (2000), "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments". To appear *Proc. Nat. Acad. Sci.*

Lee, M., Kuo, F., Whitmore, G. & Sklar, J. (2000), Importance of replication in microarray gene expression studies: statistical methods and evidence from a single cdna array experiment. To appear, *Proc. Nat., Acad. Sci.*

Li, C. & Wong, W.H. (2000), "Model-based analysis of oligonucleitide arrays: expression index computation and outlier detection". Unpublished.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. & Tsui, K. (2000), "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data". To appear, *J. Comp. Biology.*

Storey, J. (2001), "The False Discovery Rate: A Bayesian Interpretation and the q-value". Stanford Technical Report, jstorey@stat.stanford.edu.

Tusher, V., Tibshirani, R. & Chu, C. (2000), "Significance analysis of microarrays applied to transcriptional responses to ionizing radiation". To appear *Proc. Nat. Acad. Sci.*

Van del Laan, M. & Bryan, J. (2000), "Gene expression analysis with the parametric bootstrap". Report #81, Biostatistics Group, U. Cal. Berkeley.