An Empirical Bayes Adjustment to Multiple p-values For the Detection of Differentially Expressed Genes in Microarray Experiments

Somnath Datta[1] **and Susmita Datta**[2]

[1]Department of Statistics
University of Georgia,
Athens, GA 30602, USA
[2]Department of Mathematics and Statistics
and Department of Biology
Georgia State University,
Atlanta, GA 30303, USA

datta@stat.uga.edu

Abstract

In recent microarray experiments thousands of gene expressions are simultaneously tested in comparing samples (e.g., tissue types or experimental conditions). Application of a statistical test, such as the t-test, would lead to a p-value for each gene that reflects the amount of statistical evidence present in the data that the given gene is indeed differentially expressed. We show how to use these p-values across the genes using the method of empirical Bayes estimation so that each gene in turn borrows evidence of differential expression (or non-differential expression, whatever the case may be) from all other genes on the microarray. A new set of accept/reject decisions are reached for the differential expressions using the empirical Bayes adjusted p-values through a resampling based step-down p-value calculation that protects the analyst against the overall (familywise) type 1 error rate. The utility of incorporating the empirical Bayes adjustment is illustrated via a number of simulation experiments where we compute various performance measures such as sensitivity, specificity, false discovery rate and false non-discovery rate of the overall testing mechanism with and without the empirical Bayes adjustment.

Keywords: Differentially expressed genes, p-values, multiple testing, microarray, empirical Bayes

1 Introduction

In a typical microarray experiment, expression levels of thousands of genes are simultaneously measured and compared in two (or more) tissue types. A gene is declared to be differentially expressed if its average difference in expression in the two tissue types is judged to be ‘statistically significant’. Statistical significance is typically achieved by using a t-test for comparing the mean expression levels of each genes in the two groups. Since the number of genes involved is huge, setting the type of error rate at 5% for each gene would lead to a large number of false positives in the entire experiment. One way to address this issue is to control the overall or familywise error rate, say at 5%, which would mean that there is a 5% chance that the procedure would declare at least one gene to be differentially expressed in the two tissue types when indeed no gene is differentially expressed. Controlling the familywise error rate in the context of microarray experiment is achieved by adjusting the p-values (observed levels of significance) of each gene via a permutation based step-down algorithm (Dudoit et al., 2002). Very recently, Datta *et al.* (2003) advocated the use of an empirical Bayes adjustment of all the t-test statistics prior to applying the step-down p-value adjustment algorithm. The empirical Bayes idea is to exploit the similarity in the structure of the statistical testing problems for each gene and "borrow evidence" for or against differential expression from other genes besides the data for a given gene. In this paper, we propose another novel empirical Bayes adjustment that applies to the p-values of multiple tests rather than the tests themselves. It has the following three distinct advantages over the earlier procedure in Datta *et al.* (2003): (i) since it applies to the p-values, the tests don't have to be t-tests; in particular they could be F-tests which might arise in certain ANOVA formulation with expression data (Kerr *et al.*, 2000), (ii) the empirical Bayes adjustment uses nonparametric techniques to estimate the marginal density of the p-values rather than using a parametric model for the prior distribution and is therefore robust against model mis-specification, (iii) since the null (marginal) distribution of each p-value is uniform, the step-down procedure may simplify in certain situation. The rest of the paper is organized as follows. The details, including the motivation of the empirical Bayes adjustment is explained in Section 2. This section also contains a description of the step-down p-value computation algorithm and when a particular gene will be considered to be differentially expressed. In Section 3, we report the results of a number of simulation studies where the performance of our procedure is compared with the step-down p-value alone. In all cases, the empirical Bayes adjustment led to an increase in the overall sensitivity (a commonly employed global measure of performance in multiple testing). In many cases the increase is more substantial than achieved by the earlier proposed empirical Bayes adjustments (Datta *et al*, 2003). The paper ends with a discussion section.

2 Method

2.1 The Empirical Bayes Formulation

Suppose, we have a number of tests of similar structure with associated p-values denoted $\widehat{p}_i$, $1 \le i \le M$. In microarray studies, M would equal the total number of genes (probe sets etc.) on a microarray and for the ith gene $\widehat{p}_i$ might be the observed level of significance for a test that compares its average expression levels in two tissue types, say normal versus cancer cells. The p-values indicate evidence against the null hypotheses in the sense that the smaller a p-value, the more significant the evidence is that the gene is indeed differentially expressed. In general, it is defined as the chance of observing a value of the test statistic that is as extreme as (e.g., as large as) the value of the test statistic for the sample at hand, when indeed the gene is not differentially expressed. Thus, it is always a function of the sample test statistic and hence a random variable. Under the null hypothesis of no differential expression, $\widehat{p}_i$ is a uniformly distributed on the interval $(0,1)$. In the empirical Bayes formulation, we embed these distributions in a larger family of parametric distributions which also support the alternative hypotheses. One such model would be to consider a uniform distribution with a scale parameter θ_i, for the ith gene. Since our goal is to identify genes with "small" p-values, we would test a new set of hypotheses $H_0^i : \theta_i = 1$ versus $H_1^i : \theta_i < 1$ in this model. Since we are faced with simultaneous testing of a (large) number of hypotheses H^i, $1 \le i \le M$, of similar structure, we might do better by combining data of all genes using an empirical Bayes approach (Robbins, 1964; Efron and Morris, 1975). To that end, assume a common, but unknown prior distribution G, say, for each θ_i on $(0,1)$. A Bayes test of H_0^i against H_1^i would reject H_0^i for small values of $\widehat{\theta}_i^B$, that is, the posterior mean $E(\theta_i|\widehat{p}_i)$ of θ_i. Since the prior distribution G is unknown, this posterior mean needs to be estimated through nonparametric function estimation techniques from data across all genes that share this common prior distribution which we now pursue. The resulting estimated posterior mean would be called an empirical Bayes estimate of $\widehat{\theta}_i$, and is denoted $\widehat{\theta}_i^{EB}$.

2.1.1 Construction of $\widehat{\theta}_i^{EB}$

It follows from straightforward calculation (Datta, 1991) that the posterior mean $E(\theta_i|\widehat{p}_i)$ of θ_i with respect to a prior G is given by

$$\widehat{\theta}_i^B = \frac{\int I_{\{\theta_i > \widehat{p}_i\}} dG(\theta_i)}{\int I_{\{\theta_i > \widehat{p}_i\}} (\theta_i)^{-1} dG(\theta_i)},$$

$$= \widehat{p}_i + \frac{\int I_{\{\theta_i > \widehat{p}_i\}} f_G(\theta_i) d\theta_i}{f_G(\widehat{p}_i)},$$

where f_G is the common marginal density of the $\widehat{p}$. Therefore, we could estimate f_G by a kernel density estimator based on the $\widehat{p}$

$$\widehat{f}_G(t) = \frac{1}{Mh} \sum_{j=1}^{M} K\left(\frac{t - \widehat{p}_j}{h}\right),\ t \ge 0,$$

where K is a given kernel (density) and $h \approx 0$ is a bandwidth. Both K and h are user selectable, and some empirical guidelines are provided in subsection 2.3.

Next note that, for any t,

$$\int I_{\{\theta_i > t\}} f_G(\theta_i) d\theta_i = E\left(I_{\{\widehat{p} > t\}}\right),$$

which, in turn, can be estimated by its empirical counterpart

$$M^{-1} \sum_{j=1}^{M} I_{\{\widehat{p}_j > t\}}.$$

Combining these pieces we now obtain an empirical Bayes estimate of θ_i given by

$$\widehat{\theta}_i^{EB} = \widehat{p}_i + \frac{\sum_{j=1}^{M} I_{\{\widehat{p}_j > \widehat{p}_i\}}}{\sum_{j=1}^{M} K_h\left(\widehat{p}_i - \widehat{p}_j\right)}, \qquad (1)$$

where $K_h(\,\cdot\,) = h^{-1} K(\,\cdot\,/h)$.

2.2 The Step Down P-value Calculation

After calculating the empirical Bayes estimates $\widehat{\theta}_i^{EB}$, for all genes i, the next step would be to compute the corresponding step-down adjusted p-values $\widetilde{p}_i$ following the general algorithm of Westfall and Young (1993).

Step 1: Find the rank orders r_i such that $|\widehat{\theta}_{r_1}^{EB}| \le \cdots \le |\widehat{\theta}_{r_M}^{EB}|$ and let $u_i = |\widehat{\theta}_{r_i}^{EB}|$, $1 \le i \le M$ be the ordered values by their magnitudes.

Step 2: Generate a collection of random variables $\widehat{p}_i^*$, $1 \le i \le M$ from the (approximate) null distribution of the original $\widehat{p}_i$, $1 \le i \le M$.

Step 3: Convert the $\widehat{p}^*$ to the corresponding empirical Bayes estimates $\widehat{\theta}^{*EB}$ by formula (1), with $\widehat{p}^*$ in place of $\widehat{p}$ throughout and let $u_i^* = |\widehat{\theta}_{r_i}^{*EB}|$, $1 \le i \le M$ (Note that the ordering r_i is not changed during resampling) and monotonize them by $u_i^* = \min(u_i^*, u_{i+1}^*)$, $i = (M-1)$, $\cdots$, 1.

Step 4: Repeat Steps 2 and 3 a large number of times, say B, and denote the u_i^* values by $u_i^*(1), \cdots, u_i^*(B)$.

Step 5: Compute

$$\widetilde{p}_{r_i} = B^{-1}\sum_{l=1}^{B} I(\, u_i^*(l) \le u_i)$$

and monotonize them as $\widetilde{p}_{r_i} = \max(\widetilde{p}_{r_i}, \widetilde{p}_{r_{i-1}})$, for $i = 2, \cdots, M$.

For any given $0 < \alpha < 1$, (e.g., $\alpha = 0.05$) representing the familywise error rate control, declare genes $r_1, \cdots, r_{k_\alpha}$ to be differentially expressed, where

$$k_\alpha = \max\{1 \le k \le M : \widetilde{p}_{r_k} \le \alpha\}.$$

Step 2 above can be carried out in a variety of ways depending on the situation. In the simplest case, if all the tests (genes) are assumed independent then $\widehat{p}$s can be generated by independently sampling from a $U(0,1)$. In the context of a two sample problem (e.g., t-tests), $\widehat{p}$s could be obtained by calculating the observed level of significance of the test statistics calculated using randomly resampled or permuted vectors of observations of all gene expressions from the original data (Dudoit *et al.* 2002). Datta *et al.* (2003) suggested creating pseudo datasets by resampling the residuals in an ANOVA model for the gene expression in multiple tissue types. This is described in more detail in Section 3.

2.3 Bandwidth and Kernel Selection

The values near zero would typically correspond to the differentially expressed genes and, in microarray studies, often there will only be a handful of them. It is thus important to preserve the narrow pick near zero. As a result, it is best to undersmooth the estimated density by selecting a smaller bandwidth h than usual. Figure 1 shows the two density estimates for two choices of bandwidths for a typical sample from a simulation study. Based on empirical studies we recommend using h as small as 0.001. We have used the standard normal kernel for our simulations.

3 Results

In this section, we report the results of a number of simulation studies where we compute various performance measures for multiple p-values with and without the empirical Bayes adjustment. They are

(i) Sensitivity: proportion amongst differentially expressed genes that were declared significant,

(ii) Specificity: proportion amongst non-differentially expressed genes that were not declared significant,

(iii) False discovery rate (FDR): proportion amongst genes declared significant that were not-differentially expressed,

(iv) False non-discovery rate (FNR): proportion amongst genes declared not significant that were differentially expressed.

3.1 ANOVA Models for Expression Data

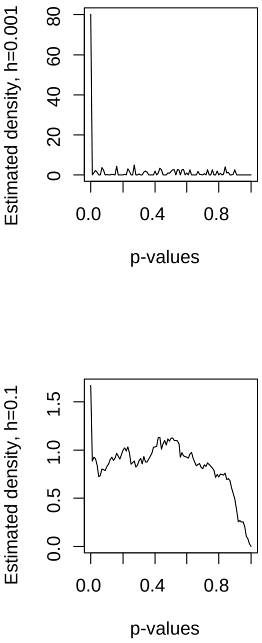

Figure 1: Estimated density of $\widehat{p}$ for the same sample using two different bandwidths

Linear models (ANOVA) have been successfully used to describe the (log-transformed) expression levels of genes in experiments involving multiple tissue types (Kerr *et al.*, 2000, Kerr and Churchill, 2001 and Kerr *et al.*, 2002, Datta *et al.*, 2003, etc.). Consider an experiment in which we have measured the expression level X (appropriately normalized and transformed) for r individuals, g genes and J tissue types (varieties). Consider a design where a single microarray consisted of the expression levels of all genes for an individual in a given tissue type. We model X as

$$X_{ijk} = \mu + I_k + G_i + V_j + (IG)_{ik} + (GV)_{ij} + \epsilon_{ijk;}$$

here $1 \le i \le g$, $1 \le j \le J$ and $1 \le k \le r$ index genes, tissue types (variety) and individuals, respectively. The ϵ_{ijk}'s denote mean zero random errors, which will be generated from a normal distribution. In this model μ represents the overall or mean expression level; main effects I_k, G_i and V_j reflect the overall differences in the expression levels for individuals, genes and varieties, respectively and the interaction term $(IG)_{ik}$ accounts for the variability of expression of the ith gene among individuals. Our primary interest lies in the gene × tissue-type interaction $(GV)_{ij}$ which measures the effect of gene i in tissue type j. The null hypothesis of no differential expression of gene i in two tissue types j_1 and j_2 is expressed as $H_0^{i;j_1,j_2} : (GV)_{ij_2} - (GV)_{ij_1} = 0$.

The t-statistic testing $H_0^{i;j_1,j_2}$ is $t_{i;j_1,j_2} =$

$$c\frac{(\overline{X}_{ij_2\cdot} - \overline{X}_{\cdot j_2\cdot} - \overline{X}_{ij_1\cdot} + \overline{X}_{\cdot j_1\cdot})}{\widehat{\sigma}}$$

with

$$\widehat{\sigma}^2 = \frac{\sum_{ijk}\left(X_{ijk} - \overline{X}_{ij\cdot} - \overline{X}_{i\cdot k} + \overline{X}_{i\cdot\cdot}\right)^2}{\nu},$$

$\nu = g(r-1)(J-1)$, and $c = \sqrt{gr/\{2(g-1\}}$. The $P-$value for testing differential expression of gene i between tissue types j_1 and j_2 is given by $\widehat{p}_{i;j_1,j_2} = 2[1 - \mathcal{P}^{t(\nu)}(|t_{i;j_1,j_2}|)]$, where $\mathcal{P}^{t(\nu)}$ is the cumulative distribution function of a central t distribution with ν degrees of freedom.

3.1.1 The Simulated Data

We used the same simulation setup as in Datta *et al.* (2003) with $g = 500$ genes for $r = 3$ individuals, each with $J = 3$ tissue types. Data were simulated using the ANOVA model introduced above, for a variety of differential expression patterns generated using the following gene-variety interaction terms.

$$(GV)_{i1} = 15\delta_i,\ (GV)_{i2} = -3\delta_i,\ (GV)_{i3} = -12\delta_i,$$
$$\text{for } 1 \le i \le 250,$$

and

$$(GV)_{i1} = -15\delta_i,\ (GV)_{i2} = 3\delta_i,\ (GV)_{i3} = 12\delta_i,$$
$$\text{for } 251 \le i \le 500.$$

The errors were generated as *i.i.d.* standard normal variates and all the main effect terms μ, I_k, G_i, V_j and the individual-gene interaction terms $(IG)_{ik}$ were set to zero.

In our four simulations, only a small number of genes (ranging from ten to thirty) were differentially expressed (i.e., $\delta_i \ne 0$). The values of δ_i for the differentially expressed genes are shown in Figure 2.

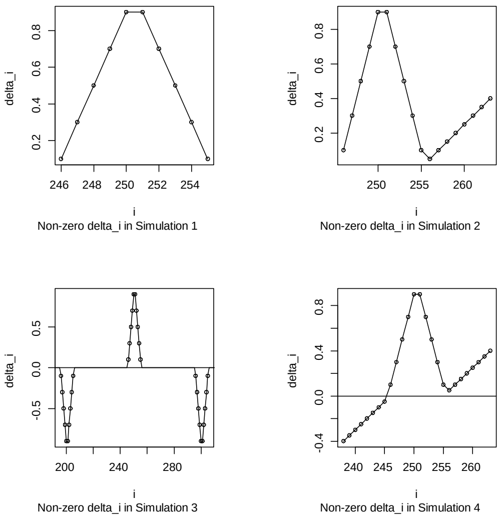

Figure 2: Patterns of differential expressions used in the simulated datasets

3.1.1 Performance

For each set of parameters (differential expression pattern) mentioned earlier, we generated thirty independent data sets. P-values for the ANOVA based t-tests are adjusted using a resampling-based step-down procedure, as in subsection 2.2, both with and without the proposed empirical Bayes modification for comparison. The resampling scheme used was the same as in Datta *et al.* (2003) and is usually referred to as a ‘model based bootstrap’ for a regression model in the bootstrap world. Under this scheme, one fits an ANOVA model and computes the model residuals. A bootstrap resample is obtained by randomly resampling the model residuals and adding the estimated main effects etc. but not the gene-variety interaction. See Datta *et al.* (2003) for further details. For each gene and each comparison, significance was declared if the adjusted P-value was less than $\alpha = 0.05$. The performance measures (proportions) were calculated for each sample and averaged over thirty independent samples for each simulation. The specificity in all cases was over 99% and the FNR was less than 4%, with slightly reduced rate for tests with the empirical Bayes adjustments. They are not reported further in Table 1.

Model	Tissue types	without EB		with EB	
		Sen	FDR	Sen	FDR
1	1 vs. 2	0.80	0.01	0.83	0.12
	2 vs. 3	0.63	0.01	0.71	0.06
2	1 vs. 2	0.72	0.01	0.78	0.09
	2 vs. 3	0.44	0.01	0.57	0.06
3	1 vs. 2	0.81	0.00	0.84	0.06
	2 vs. 3	0.64	0.00	0.73	0.03
4	1 vs. 2	0.68	0.00	0.76	0.08
	2 vs. 3	0.36	0.01	0.51	0.07

Table 1. Sensitivity and FDR with and without the empirical Bayes adjustments in an ANOVA setting for gene expressions

Overall, we notice improvement in sensitivity (denoted ‘Sen’ in Table 1) in all cases and a moderate increase in the FDR, a fate shared by an earlier empirical Bayes adjustment proposed in Datta *et al.* (2003). Moreover, in comparison to the results in Datta *et al.* (2003), we notice substantial additional gain in applying the new procedure in most cases (e.g., up to forty percent improvement in sensitivity).

4 Discussion

In this paper, we propose a novel empirical Bayes modification to multiple test statistics. This adjustment works with the p-values of the original test statistics and in turn produces a new set of test statistics such that each member borrows ‘evidence’ from others along with the evidence in itself. The multiplicity adjusted step down p-values of the resulting empirical Bayes tests can be computed via resampling algorithms.

It is hoped that the empirical Bayes procedure would detect additional differentially expressed genes compared to the step-down procedure using the p-values without this adjustments, while maintaining a comparable level of familywise (or overall) type 1 error rate control. The increase in sensitivity in the simulation studies certainly gives such an indication. The simulation studies also show, however, that the increase in sensitivity comes at the cost of a modest increase in the false discovery rate. At present, we are attempting to combine FDR control with the empirical Bayes adjustments while maintaining an edge over sensitivity. One difficulty in achieving this goal is the lack of available procedure for precise control of the FDR, especially for dependent test statistics such as the empirical Bayes tests. In our experience, the currently available procedures are too conservative and are not suitable for our purpose.

Unlike our previously proposed empirical Bayes adjustment (Datta *et al.*, 2003) that applies to the studentized test statistics (such as the t-tests), the adjustment proposed here works on the p-values of multiple tests. Therefore the current procedures would have broader applicability to other types of tests such as the F-tests. For example, in a microarray experiment involving multiple tissue types (e.g., normal, adenoma and carcinoma) one would be able to detect genes that are differentially expressed amongst the various types of tissues (without restricting attention to a particular tissue pair).

In this paper, we have used the standard kernel estimator for estimating the marginal density of the p-values under the Bayes model. It will be interesting to investigate other nonparametric methods such as density estimators based on wavelets in this context.

We are currently applying the empirical Bayes adjusted p-values to a dataset on colon cancer studied in Datta *et al.* (2003) and the results will be forthcoming.

Acknowledgments

Susmita Datta's research was supported in part by the Distinguished Cancer Clinicians and Scientists Program of Georgia Cancer Coalition.

5 References

Datta, S. (1991): Nonparametric empirical Bayes estimation with O($n^{-1/2}$) rate of a truncation parameter. *Statistics and Decisions*, **9**: 45-61.

Datta, S., Satten, G. A., Benos, D. J., Xia, J., Heslin, M. J., and Datta, S. (2003): An empirical Bayes adjustment to increase the sensitivity of detecting differentially expressed genes in microarray experiments. *Bioinformatics*, to appear.

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

Efron, B. and Morris, C. (1975): Data analysis using Stein's estimator and its generalization. *J. Amer. Statist Assoc.*, **70**: 311-319.

Kerr, M. K. and Churchill, G. A. (2001): Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sc. USA*, **98**: 8961-8965.

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, N. W., and Churchill, G. A. (2002): Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, **12**: 203-217.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000): Analysis of variance for gene expression microarray data. *J. Comp. Biol.*, **7**: 819-837.

Robbins, H. (1964): The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, **35**: 1-20.

Westfall, P. H. and Young, S. S. (1993): *Resampling based multiple testing: examples and methods for p-value adjustment*. New York, Wiley.