Estimating the Proportion of True Null Hypotheses for Multiple Comparisons

by

H. Jiang
Northwestern University

R.W. Doerge
Purdue University

Technical Report #07-06

Department of Statistics
Purdue University

June 2007

Estimating the proportion of true null hypotheses for multiple comparisons

Hongmei Jiang[1] and R.W. Doerge[2]

[1] Department of Statistics
Northwestern University
2006 Sheridan Road
Evanston, IL 60208 USA
phone: (847) 467-1087
email: hongmei@northwestern.edu

[2] Department of Statistics
Purdue University
250 N. University Street
West Lafayette, IN 47907 USA
phone: (765) 494-6030
email: doerge@purdue.edu

Abstract

Whole genome microarray investigations (eg, differential expression, differential methylation, ChIP Chip) provide opportunities to test thousands of features in a genome. Traditional multiple comparison procedures such as familywise error rate (FWER) controlling procedures are too conservative. Although false discovery rate (FDR) procedures have been suggested as having greater power, the control itself is not exact and depends on the proportion of true null hypotheses. Because this proportion is unknown, it has to be accurately (small bias, small variance) estimated, preferably using a simple calculation that can be made accessible to the general scientific community. We propose an easy-to-implement method for estimating the proportion of true null hypotheses. This estimate has relatively small bias and small variance as demonstrated by (simulated and real data) comparing it with four existing procedures. Although presented here in the context of microarrays, this estimate is applicable for many multiple comparison situations.

Keywords: False discovery rate; Multiple comparisons; Type I error rate; Microarray.

1 Introduction

Genomic technologies are producing vast amounts of biological data that are the basis for investigations that require repetitive testing of the same hypothesis. Because the number of tests performed (eg, differential expression) is so large, sometimes the multiple comparison procedures that control the familywise error rate are too strict for biological applications (eg, differential methylation). In fact, many biologists would rather experience several more false positives (ie, type I errors; false rejections of the null hypothesis) than lose important information. In an attempt to address the multiple comparison issue Benjamini & Hochberg (1995) introduced an error rate measure called False Discovery Rate (FDR) (ie, the expected proportion of false rejections among all the rejected hypotheses). Specifically, a family of m hypothesis tests is considered, of which m_0 are true. The proportion of erroneously rejected null hypotheses among all the rejected null hypotheses can be captured by the random variable $Q = V/R$, where R is the number of rejected hypotheses and V is the number of false rejections (type I errors). Benjamini & Hochberg (1995) formally define the FDR to be the expected proportion of falsely rejected hypotheses among all the rejections,

$$\text{FDR} = E(Q) = E(V/R), \tag{1}$$

where $Q = 0$ when $R = 0$ (no rejections).

Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered p-values and $H_{(i)}$ be the null hypothesis corresponding to $p_{(i)}$. In Benjamini and Hochberg's (BH) FDR controlling procedure (Benjamini & Hochberg 1995), K is considered to be the largest k such that $p_{(k)} \leq (k/m)\alpha$. If a such K exists, all null hypotheses $H_{(i)}, i = 1, \cdots, K$ are rejected. If no such K exists, then no hypotheses are rejected. The BH FDR controlling procedure controls the FDR at exactly the level $(m_0/m)\alpha \leq \alpha$, and hence conservatively at α, for independent test statistics and for any configuration of false null hypotheses (Benjamini & Yekutieli 2001, Storey, et al. 2004). Benjamini & Hochberg (2000) proposed an adaptive procedure which provides more power than the original FDR controlling procedure by comparing each $p_{(k)}$ with $(k/\widehat{m}_0)\alpha$ where $\widehat{m}_0$ is an estimate of m_0. If the estimated value of m_0 is such that $\widehat{m}_0 \geq m_0$ with probability one, then the adaptive BH FDR controlling procedure will lead to FDR $= \frac{m_0}{m}(\frac{m}{\widehat{m}_0}\alpha) = \frac{m_0}{\widehat{m}_0}\alpha \leq \alpha$. Because the accuracy and variation of the estimate of m_0, or $\pi_0 = m_0/m$, directly affects the performance of the adaptive FDR controlling procedure our focus is on the estimation and effect of π_0.

We propose a simple and easy-to-implement method for estimating the proportion of true null hypotheses. The performance of this estimate is compared to existing methods via simulated and real data. Specifically, Benjamini & Hochberg (2000) estimated the number of true hypotheses from the observed p-values using the Lowest SLope (LSL) estimator. Their approach was based on a modification of the graphical method of Schweder & Spjotvoll (1982). Alternatively, Storey (2002) proposed an estimate of π_0 by assuming the p-values corresponding to true null hypotheses are uniformly distributed on the interval (0,1) and selecting a reasonable tuning parameter $0 \leq \lambda < 1$. Finally, Langaas, et al. (2005) derived estimators based on nonparametric maximum likelihood estimation of the p-value density, under the restriction of decreasing and convex decreasing densities. Although Benjamini and Hochberg's original and adaptive FDR controlling procedure are developed for independent statistics these procedures can also be applied to some dependence structures (Benjamini & Yekutieli 2001). Simulations have also demonstrated that they can be used for situations where there is a weak correlation structure among the genes (Storey et al. 2004). However, because of the small number of biological replicates used in most microarray experiments, it is very difficult to measure the correlation structure among a set or family of genes. Reiner, et al. (2003) proposed a procedure for the general case, but it is conservative when compared to the adaptive FDR controlling procedures.

2 Methods

2.1 Storey's approach

Our approach is motivated by the work of Storey (2002), where the proportion of true null hypotheses, π_0, is estimated by

$$\hat{\pi}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)m}, \tag{2}$$

where $W(\lambda) = \#\{p_i : p_i > \lambda\}$, and $0 \leq \lambda < 1$ is a tuning parameter. The rationale for this estimate is that p-values corresponding to true null hypotheses are uniformly distributed on the interval $(0, 1)$, of which most should be close to 1. Thus, for a reasonable λ, there are about $m_0(1-\lambda)$ such p-values in the interval $(\lambda, 1]$ such that $W(\lambda) \approx m_0(1-\lambda)$. Black (2004) pointed out that Equation (2) is an unbiased estimate of π_0 for all values of λ if all the null hypotheses are true and the p-values have a uniform

distribution on the interval (0,1). However, there is an upward bias when the p-values come from both true null and true alternative hypotheses. As it turns out, choosing the tuning parameter λ in Equation (2) is very important since there is a bias-variance trade-off. When $\lambda \to 0$, the variance of $\widehat{\pi}_0(\lambda)$ becomes smaller, and the bias of this estimate increases. When $\lambda \to 1$, the bias of $\widehat{\pi}_0(\lambda)$ becomes smaller, and the variance of this estimate increases. To address this point, Storey et al. (2004) proposed a bootstrap method that automatically chooses λ when estimating $\widehat{\pi}_0(\lambda)$.

Instead of choosing one specific λ, Storey & Tibshirani (2003) proposed an estimate of π_0 using $\lim_{\lambda \to 1} \widehat{\pi}_0(\lambda)$ so that the bias is small and there is a balance between both bias and variance. For this approach, $\widehat{\pi}_0(\lambda)$ is plotted over a range of $\lambda = 0, 0.05, 0.10, \cdots, 0.90$, and then a natural cubic smoothing spline is fit to these data for the purpose of estimating the overall trend of $\widehat{\pi}_0(\lambda)$ as $\lambda \to 1$. In the QVALUE (http://faculty.washington.edu/$\sim$ jstorey/) software, the predicted value of $\widehat{\pi}_0(\lambda)$ at $\lambda = 0.90$ is chosen as the estimate of π_0.

2.2 Average estimate approach

As mentioned previously, the estimate $\widehat{\pi}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)m}$ where $0 \leq \lambda < 1$, has a large bias and small variance when λ is small, and a small bias and large variance when λ is big. Suppose for each λ_i, where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n < 1$, we compute $\widehat{\pi}_0(\lambda_i)$ as in Equation (2), then

$$E[\widehat{\pi}_0(\lambda_i)] = \pi_0 + \varepsilon_i,$$

where $E[\varepsilon_i] = \delta_i$, $\delta_i \geq \delta_{i+1}$, $Var[\varepsilon_i] = \sigma_i^2$, and $\sigma_i^2 \leq \sigma_{i+1}^2$. Therefore, a natural choice is to consider the average of $\widehat{\pi}_0(\lambda)$ over the values of λ_i,

$$\widehat{\pi}_0 = \frac{1}{n} \sum_{i=1}^{i=n} \widehat{\pi}_0(\lambda_i).$$

The bias of $\widehat{\pi}_0$, $1/n \sum_{i=1}^{i=n} \delta_i$, is smaller than δ_1 (the bias of the estimate of π_0 at $\lambda = \lambda_1$), and at the same time, $\widehat{\pi}_0$ has a smaller variance. Considering the average of $\widehat{\pi}_0(\lambda)$ over a range of λ to estimate π_0 reduces the problem to choosing the range of λ.

Define $0 = t_1 < t_2 < \cdots < t_B < t_{B+1} = 1$ as equally spaced points in the interval $[0, 1]$ such that the interval $[0, 1]$ is divided into B small intervals with equal length $1/B$. Specifically, $t_i = (i-1)/B$. For example, when $B = 10$,

$t_1 = 0, t_2 = 0.1, \cdots, t_{10} = 0.9$. For each t_i $(i = 1, \cdots, B)$, $\hat{\pi}_0(t_i)$ is an estimate of π_0 via Equation (2) with $\lambda = t_i$. The goal then becomes finding a subset of t_i's such that a new estimate of π_0 is obtained by taking the average of the corresponding values of $\hat{\pi}_0(t_i)$. Let NB_i denote the number of p-values which are greater than or equal to t_i, and let NS_i represent the number of p-values in the interval of $[t_i, t_{i+1})$. Therefore,

$$NB_i = \#\{p_k : p_k \geq t_i\}, \tag{3}$$

$$\hat{\pi}_0(t_i) = \frac{NB_i}{(1-t_i)m}, \tag{4}$$

$$NS_i = \#\{p_k : t_i \leq p_k < t_{i+1}\}, \tag{5}$$

where $i = 1, \cdots, B$.

If the NB_i p-values come from the null distribution, then on average there are $\frac{NB_i}{B-i+1}$ p-values in each of the $(B-i+1)$ small intervals on $[t_i, 1]$, ie, there are $\frac{NB_i}{B-i+1}$ p-values in each small interval $[t_j, t_{j+1})$ for $i \leq j \leq B$. Since the p-values corresponding to the true alternative hypotheses are smaller than those corresponding to the true null hypotheses, there are more p-values in the intervals $[t_i, t_{i+1})$ with small index i. For small i, NS_i is usually greater than $\frac{NB_i}{B-i+1}$. Therefore, initiating from $i = 1$, we find the first i such that $NS_i \leq \frac{NB_i}{B-i+1}$. If such i exists, t_i can be considered as the change point and we assume all the p-values bigger than t_i come from the true null hypotheses. Then π_0 can be estimated by

$$\widehat{\pi}_0(B) = \frac{1}{B-i+1} \sum_{j=i}^{j=B} \hat{\pi}_0(t_j) \tag{6}$$

$$= \frac{1}{B-i+1} \sum_{j=i}^{j=B} \frac{NB_j}{(1-t_j)m} \tag{7}$$

where $i = \min\{i : NS_i \leq \frac{NB_i}{B-i+1}\}$. In order to find the range of λ, only a lower bound of λ is required. The large values of t_i are used so that it ensures the estimate of π_0 has small bias. This is equivalent to fitting a straight line with slope 0 in the right bottom part of a $\widehat{\pi}_0(t_i)$ versus t_i plot, such that the intercept provides the estimate of π_0. A simple modification of this approach is to estimate π_0 by taking the average of $\hat{\pi}_0(t_j)$ from $j = i-1$ to B, that is,

$$\widehat{\pi}_0(B) = \frac{1}{B-i+2} \sum_{j=i-1}^{j=B} \frac{NB_j}{(1-t_j)m}. \tag{8}$$

where $i = \min\{i : NS_i \leq \frac{NB_i}{B-i+1}\}$. This ensures that the upward bias increases and the variance decreases, as $\hat{\pi}_0(t_{i-1})$ has smaller variance and bigger bias than $\hat{\pi}_0(t_j)$ for $j = i, \cdots, B$.

A remaining issue is how to choose B. Specifically, how many λ's should be used in the interval [0,1]. Recall that a motivating factor of the proposed average estimate approach is to balance the bias and variance. The natural way to measure both the bias and variance is the mean-squared error, $E[\widehat{\pi}_0(B) - \pi_0]^2$. Since the true value of π_0 is unknown, and the theoretical result is intractable, we take a bootstrap approach in the following way:

1. For each $B \in I$, $I = \{5, 10, 20, 50, 100\}$, compute $\widehat{\pi}_0(B)$ as in Equation (8).

2. Form N bootstrap samples of the p-values, and compute the bootstrap estimates $\widehat{\pi}_0^{*b}(B)$ for $b = 1, \cdots, N$ and $B \in \{5, 10, 20, 50, 100\}$.

3. For each $B \in I$, estimate its respective mean-squared error as

$$\widehat{\mathrm{MSE}}(B) = \frac{1}{N} \sum_{b=1}^{N} [\widehat{\pi}_0^{*b}(B) - \bar{\pi}_0]^2,$$

   where,

$$\bar{\pi}_0 = \mathrm{average}_{B' \in I}\{\widehat{\pi}_0(B')\}.$$

4. Let $\widehat{B} = \arg\min_{B \in I} \widehat{\mathrm{MSE}}(B)$, then the estimate of π_0 is $\widehat{\pi}_0 = \widehat{\pi}_0(\widehat{B})$.

Notice that in step three the value of π_0 is estimated by the average of the $\widehat{\pi}_0(B)$ over a range of B.

3 Results

3.1 Simulation studies

To investigate the performance of the proposed average estimate approach, a simulation study was performed. Taking $m = 1,000$ (ie, 1,000 genes are tested for differential expression), let π_0 vary over a wide range, say $\pi_0 = 0.50, 0.60, \cdots, 0.90$ which are reasonable for microarray experiments. Hypotheses, H_0: $\mu = 0$ versus H_a: $\mu > 0$, are tested for independent random variables Z_i $(i = 1, \cdots, m)$ from null distribution N(0,1) and alternative distribution N(2,1) (ie, $m\pi_0$ and $m(1 - \pi_0)$ random variables have mean 0 and 2, respectively). For each test, the p-value is computed as $p_i = P(Z > z_i)$,

where Z is a random variable of standard normal distribution N(0,1) and z_i is the observed value of Z_i. For each value of π_0, $l = 1,000$ data sets were simulated.

For the choice of B, we have either B being fixed (ie, $B = 5$, 10, 20, 50, and 100) or being chosen by the proposed bootstrap approach. For each of the $l = 1,000$ simulated data sets, when B is fixed, the estimate of π_0 is computed via Equation (8), that is, $\widehat{\pi}_0 = \frac{1}{B-i+2} \sum_{j=i-1}^{j=B} \widehat{\pi}_0(t_j)$ where $i = \min\{i : NS_i \leq \frac{NB_i}{B-i+1}\}$. If such i does not exist, π_0 is estimated by the average of $\widehat{\pi}_0(t_{B-1})$ and $\widehat{\pi}_0(t_B)$. For the bootstrap approach to automatically choose B, the range of B is $5, 10, 20, 50, 100$.

For completion the performance of the proposed average estimate approach is compared with several existing procedures. Specifically,

1. Benjamini and Hochberg's lowest slope estimate (LSL) (Benjamini & Hochberg 2000),

2. Storey's bootstrap estimate ($\text{Storey}_{\text{boot}}$) (Storey et al. 2004),

3. Storey and Tibshirani's smoother estimate ($\text{ST}_{\text{smoother}}$) (Storey & Tibshirani 2003),

4. Langass *et al.*'s nonparametric maximum likelihood estimate (convest) (Langaas et al. 2005).

For procedures 2 and 3, the QVALUE software was downloaded from the website http://faculty.washington.edu/ $\sim$ jstorey/. For procedure 4, the R function convest was downloaded from the R library limma as part of the Bioconductor project at http://www.bioconductor.org.

Table 1 summarizes the simulation results. Bias and the standard deviation of the estimates are estimated by

$$\widehat{\text{Bias}} = \frac{1}{l}\sum_{i=1}^{i=l}(\hat{\pi}_{0i} - \pi_0),$$

$$\widehat{\text{Std}} = \sqrt{\frac{1}{l-1}\sum_{i=1}^{i=l}(\hat{\pi}_{0i} - \frac{1}{l}\sum_{i=1}^{i=l}\hat{\pi}_{0i})^2},$$

where $\hat{\pi}_{0i}$ estimates π_0 for the ith simulation, and π_0 is the true value. As demonstrated, the LSL approach has the largest upward bias which guarantees that Benjamini and Hochberg's adaptive FDR controlling procedure

controls the FDR below a pre-chosen FDR level. However, the FDR can be much lower than the pre-chosen FDR level. The LSL approach also has the smallest variation. The last three approaches [2-4] all underestimate the proportion of true null hypotheses. The proposed average estimate approach provides estimates of π_0 that have upward but relatively small bias and relatively small variance regardless of whether B is fixed or automatically chosen via bootstrap procedure. When B increases, the bias increases and the variation decreases. Both the small upward bias and small variance provide evidence that the proposed average estimate approach has better properties when compared to the other approaches.

The FDR is also compared in this numerical study by applying Benjamini and Hochberg's adaptive FDR controlling procedure (Benjamini & Hochberg 2000) with π_0 estimated using the above mentioned five methods (Table 2). The FDR significance level was chosen as $\alpha = 0.05$. For the purpose of comparison, the original BH FDR controlling procedure (Benjamini & Hochberg 1995) and the adaptive FDR controlling procedure with the incorporation of the true value of π_0 were also applied to the p-values. It can be seen that the original BH FDR controlling procedure has the lowest FDR as expected. Because Benjamini and Hochberg's lowest slope approach overestimates π_0, the FDR is below, but much lower than, the pre-chosen α, although this approach has a bigger FDR than that of the BH procedure. Storey's bootstrap estimate, the smoother estimate, and convest estimate produce higher FDRs than the pre-chosen level because all three methods underestimate π_0. Our proposed average estimate approach overestimates π_0, its FDR is below but very close to the pre-chosen significance level $\alpha = 0.05$. Table 2 also demonstrates that the FDR for the proposed average estimate has the relatively small variation.

The power of the five adaptive FDR controlling procedures is compared (Table 3). The power of a procedure is measured by average power which is defined to be the ratio of average number of correct rejections of true alternative hypotheses to the total number of true alternative hypotheses. Formally, $\textit{average power} = E(S)/(m - m_0)$. As illustrated, the power decreases when π_0 increases for all of the FDR controlling procedures. The original BH procedure has the lowest power, while Benjamini and Hochberg's adaptive procedure has the second lowest power. It is not surprising that $\text{Storey}_{\text{boot}}$ procedure has the largest statistical power, because the FDR of this procedure exceeds the pre-chosen FDR significance level (Table 2).

Table 1: The estimate of the proportion of true null hypotheses is compared for: Benjamini and Hochberg's lowest slope approach (LSL), Storey's $\widehat{\pi}_0(\lambda)$ estimate with λ selected via bootstrapping ($\text{Storey}_{\text{boot}}$), Storey and Tibshirani's smoother method ($\text{ST}_{\text{smoother}}$), Langass's nonparametric maximum likelihood approach (convest), and the proposed average estimate approach with fixed values of $B = 5, 10, 20, 50, 100$ and with B chosen via the bootstrapping procedure (B_{boot}). There are 1,000 simulated data sets, each with a total of $m = 1,000$ hypothesis tests, for each value of π_0.

π_0	0.5	0.6	0.7	0.8	0.9
	Estimates of π_0				
LSL	0.7151	0.7889	0.8561	0.9184	0.9683
$\text{Storey}_{\text{boot}}$	0.4814	0.5789	0.6765	0.7728	0.8660
$\text{ST}_{\text{smoother}}$	0.4951	0.5939	0.6980	0.7993	0.8973
convest	0.4963	0.5938	0.6947	0.792	0.8882
$B = 5$	0.5132	0.6113	0.7136	0.8086	0.9058
$B = 10$	0.5082	0.6084	0.7083	0.8045	0.9052
$B = 20$	0.5141	0.6128	0.7115	0.8076	0.9064
$B = 50$	0.5196	0.6175	0.7156	0.8106	0.9078
$B = 100$	0.5243	0.6210	0.7180	0.8122	0.9085
B_{boot}	0.5195	0.6175	0.7148	0.8113	0.9082
	Standard deviation of π_0 estimates				
LSL	0.0323	0.0269	0.0225	0.0155	0.0092
$\text{Storey}_{\text{boot}}$	0.0467	0.0491	0.0513	0.0522	0.0549
$\text{ST}_{\text{smoother}}$	0.0513	0.0570	0.0608	0.0654	0.0656
convest	0.0331	0.0364	0.0337	0.0321	0.0328
$B = 5$	0.0335	0.0356	0.0420	0.0428	0.0382
$B = 10$	0.0391	0.0390	0.0402	0.0412	0.0366
$B = 20$	0.0331	0.0343	0.0358	0.0371	0.0331
$B = 50$	0.0293	0.0309	0.0321	0.0334	0.0315
$B = 100$	0.0272	0.0291	0.0307	0.0321	0.0312
B_{boot}	0.0301	0.0301	0.0313	0.0313	0.0311

Table 2: Simulation results of the False Discovery Rate (FDR) at significance level $\alpha = 0.05$ for six procedures: Benjamini and Hochberg's FDR controlling procedure with incorporation of the true π_0 (BH_{π_0}), Benjamini and Hochberg's FDR controlling procedure (BH), Benjamini and Hochberg's adaptive approach with incorporation of the estimate of π_0 which is estimated by the proposed average estimate procedure where B is chosen via bootstrapping, Benjamini and Hochberg's lowest slope approach (LSL), Storey's bootstrapping approach ($Storey_{boot}$), Storey and Tibshirani's smoother method ($ST_{smoother}$), and Langass *et al.*'s nonparametric maximum likelihood estimate (convest), respectively. The total number of hypotheses tests is $m = 1,000$, and the size of simulation study 1,000 for each value of π_0.

π_0	0.5	0.6	0.7	0.8	0.9
	Estimate of FDR				
BH_{π_0}	0.0499	0.0501	0.0506	0.0507	0.0520
BH	0.0252	0.0301	0.0349	0.0408	0.0455
LSL	0.0352	0.0386	0.0409	0.0445	0.0474
$Storey_{boot}$	0.0521	0.0524	0.0526	0.0529	0.0542
$ST_{smoother}$	0.0527	0.0529	0.0528	0.0531	0.0546
convest	0.0506	0.0508	0.0512	0.0516	0.0531
B_{boot}	0.0479	0.0492	0.0486	0.0485	0.0493
	Standard deviation of the FDR estimates				
BH_{π_0}	0.0129	0.0166	0.0222	0.0328	0.0743
BH	0.0117	0.0151	0.0214	0.0332	0.0736
LSL	0.0126	0.0163	0.0220	0.0331	0.0740
$Storey_{boot}$	0.0144	0.0177	0.0226	0.0336	0.0742
$ST_{smoother}$	0.0146	0.0177	0.0226	0.0334	0.0742
convest	0.0139	0.0171	0.0224	0.0332	0.0754
B_{boot}	0.0134	0.0173	0.0219	0.0314	0.0625

Table 3: Simulation results for the evaluation of statistical power at significance level $\alpha = 0.05$ for six procedures: Benjamini and Hochberg's FDR controlling procedure with incorporation of the true π_0 (BH_{π_0}), Benjamini and Hochberg's FDR controlling procedure (BH), Benjamini and Hochberg's adaptive approach with incorporation of the estimate of π_0 which is estimated by the proposed average estimate procedure where B is chosen via bootstrapping, Benjamini and Hochberg's lowest slope approach (LSL), Storey's bootstrapping approach ($Storey_{boot}$), Storey and Tibshirani's smoother method ($ST_{smoother}$), and Langass *et al.*'s nonparametric maximum likelihood estimate (convest), respectively. The total number of hypotheses tests is $m = 1,000$, and the size of simulation study is 1,000 for each value of π_0.

π_0	0.5	0.6	0.7	0.8	0.9
	Estimate of power				
BH_{π_0}	0.5376	0.4351	0.3330	0.2307	0.1181
BH	0.3619	0.3138	0.2572	0.1915	0.1061
LSL	0.4460	0.3683	0.2881	0.2061	0.1095
$Storey_{boot}$	0.5479	0.4451	0.3422	0.2370	0.1238
$ST_{smoother}$	0.5508	0.4471	0.3435	0.2380	0.1244
convest	0.5400	0.4378	0.3353	0.2323	0.1198
B_{boot}	0.5274	0.4304	0.3296	0.2281	0.1234
	Standard deviation of the power estimates				
BH_{π_0}	0.0331	0.0382	0.0452	0.0530	0.0564
BH	0.0361	0.0380	0.0451	0.0504	0.0542
LSL	0.0399	0.0407	0.0466	0.0526	0.0553
$Storey_{boot}$	0.0408	0.0445	0.0490	0.0547	0.0588
$ST_{smoother}$	0.0409	0.0445	0.0492	0.0547	0.0590
convest	0.0378	0.0414	0.0466	0.0537	0.0574
B_{boot}	0.0367	0.0409	0.0449	0.0522	0.0591

Table 4: The estimate of the proportion of true null hypotheses and the number of statistically significant genes for the leukemai data (Golub, et al. 1999) at significance level $\alpha = 0.05$ after applying Benjamni and Hochberg's adaptive FDR controlling procedure with π_0 estimated using five methods: Benjamini and Hochberg's lowest slope approach (LSL), Storey's $\widehat{\pi}_0(\lambda)$ estimate with λ selected via bootstrapping ($\text{Storey}_{\text{boot}}$), Storey and Tibshirani's smoother method ($\text{ST}_{\text{smoother}}$), Langass's convest approach (convest), and the proposed average approach with B chosen via the bootstrapping procedure (B_{boot}). A two-sample t-test was used to compute the p-values.

Method	Estimate of π_0	Number of Significant genes
LSL	0.899	584
$\text{Storey}_{\text{boot}}$	0.595	787
$\text{ST}_{\text{smoother}}$	0.583	791
convest	0.595	787
B_{boot}	0.604	776

3.2 Microarray data application

The same five estimating π_0 methods were also applied to the training samples of the leukemia data of Golub, et al. (1999), which consist of 27 patients with acute lymphoblastic leukemia (ALL) and 11 patients with acute myeloid leukemia (AML). The samples were assayed using Affymetrix Hgu6800 chips and the gene expression data of 7129 genes (Affymetrix probes) are available from R library golubEsets. For each gene, a simple two-sample t-test was employed for testing differential gene expression and the p-value was computed. Table 4 gives the estimate of the proportion of true null hypotheses and the number of statistically significant genes.

From this real data analysis, it can be seen that the Benjamni and Hochberg's LSL approach conservatively overestimates π_0, hence it leads to lowest power in terms of the number of rejections. Our proposed average approach provides a slightly larger estimate than Storey's bootstrap approach, the smoother estimate, and the nonparametric maximum likelihood approach (convest), even though they end up with a similar number of rejections.

4 Summary

As array technology improves, it is anticipated that the number of features per array will only increase, hence multiple comparisons will continue to be a challenging problem. Specific to microarrays, the false discovery rate (FDR) is preferred to familywise error rate (FWER) because the FDR controlling procedures have more statistical power than the FWER controlling procedures, even at the cost of a few more type I errors (ie, false positives). Since Benjamini & Hochberg (1995) proposed their FDR controlling procedure, a variety of methods have been proposed to estimate the π_0, the proportion of true null hypotheses. When our, and others, estimate of π_0 is incorporated into the Benjamini and Hochberg's FDR controlling procedure, the adaptive FDR controlling procedure has more power and an FDR close to the pre-chosen level. In this work, we have compared several methods for estimating π_0 via a numerical investigation. Benjamini Hochberg's lowest slope approach (Benjamini & Hochberg 2000) overestimates π_0. Storey's estimate $\widehat{\pi}_0(\lambda)$ (Storey 2002) also overestimates π_0 for any fixed value $0 \leq \lambda < 1$. When $\lambda \to 1$, the bias becomes smaller, and the variance becomes bigger. In order to find the optimal λ such that $\widehat{\pi}_0(\lambda)$ has small variation, Storey proposed a bootstrapping method (Storey et al. 2004). However, this method underestimates π_0 and the downward bias increases as the true value π_0 gets bigger. Storey & Tibshirani (2003) proposed a smoother method to estimate $\lim_{\lambda \to 1} \widehat{\pi}_0(\lambda)$ such that this estimate has small bias. Unfortunately, this method also underestimates π_0, although the bias is very small. Furthermore, the variation of this estimate is relatively large, which makes the adaptive FDR controlling procedure unstable. More recently, Langaas et al. (2005) proposed an estimate based on the nonparametric maximum likelihood function of the p-value density restricted to convex decreasing densities. However, this method also underestimates π_0, most likely because the distribution of the p-values is not decreasing for large p-values and tends to be flat. Using the limitations of the existing approaches for estimating π_0 as the motivation, we propose the average estimate approach by taking average of the estimates of π_0 over a range of equally spaced points on the interval $[0, 1]$. While our average estimate approach has a slightly larger bias, it also has smaller variation than any of the other methods. Furthermore, when compared to the other methods it is easy to implement (eg, Excel) when the number of points used in approach is fixed (say, $B = 10$), and can be automated to choose B via a bootstrap procedure (ie, R code available). When our proposed estimated value of π_0 is incorporated into Benjamini and Hochberg's adaptive FDR

controlling procedure, more statistical power is gained such that the FDR can be controlled below, yet extremely close to a desired level α.

References

Y. Benjamini & Y. Hochberg (1995). 'Controlling the false discovery rate: a practical and powerful approach to multiple testing'. *Journal of the Royal Statistical Society, Series B* **57**:289–300.

Y. Benjamini & Y. Hochberg (2000). 'On the adaptive control of the false discovery rate in multiple testing with independent statistics'. *Journal of Educational and Behavioral Statistics* **25**(1):60–83.

Y. Benjamini & D. Yekutieli (2001). 'The control of the false discovery rate in multiple testing under dependency'. *The Annals of Statistics* **29**:1165–1188.

M. A. Black (2004). 'A note on the adaptive control of false discovery rates'. *Journal of the Royal Statistical Society, Series B* **66**(2):297–304.

T. Golub, et al. (1999). 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring'. *Science* **286**:531–537.

M. Langaas, et al. (2005). 'Estimating the proportion of true null hypotheses, with application to DNA microarray data'. *Journal of the Royal Statistical Society, Series B* **67**:555–572.

A. Reiner, et al. (2003). 'Identifying differentially expressed genes using false discovery rate controlling procedures'. *Bioinformatics* **19**:368–375.

T. Schweder & E. Spjotvoll (1982). 'Plots of p-values to evaluate many tests simutaneously'. *Biometrika* **49**:493–502.

J. D. Storey (2002). 'A direct approach to false discovery rates'. *Journal of the Royal Statistical Society, Series B* **64**:479–498.

J. D. Storey, et al. (2004). 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach'. *Journal of the Royal Statistical Society, Series B* **66**:187–205.

J. D. Storey & R. Tibshirani (2003). 'Statistical significance for genomewide studies'. *Proceedings of the National Academy of Sciences, USA* **100**(16):9440–9445.