Research Article

Estimating the Proportion of True Null Hypotheses in Multiple Testing Problems

Oluyemi Oyeniran¹ and Hanfeng Chen²

¹Manufacturing, Toxicology and Applied Statistical Sciences, Janssen Research & Development, Spring House, PA 19002, USA

²Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA

Correspondence should be addressed to Hanfeng Chen; hchen@bgsu.edu

Received 26 July 2016; Revised 19 October 2016; Accepted 8 November 2016

Academic Editor: Shein-chung Chow

Copyright © 2016 O. Oyeniran and H. Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The problem of estimating the proportion, π_0 , of the true null hypotheses in a multiple testing problem is important in cases where large scale parallel hypotheses tests are performed independently. While the problem is a quantity of interest in its own right in applications, the estimate of π_0 can be used for assessing or controlling an overall false discovery rate. In this article, we develop an innovative nonparametric maximum likelihood approach to estimate π_0 . The nonparametric likelihood is proposed to be restricted to multinomial models and an EM algorithm is also developed to approximate the estimate of π_0 . Simulation studies show that the proposed method outperforms other existing methods. Using experimental microarray datasets, we demonstrate that the new method provides satisfactory estimate in practice.

1. Introduction

Estimating the proportion π_0 of true null hypotheses in a multiple testing setup is very crucial in wanting to assess and/or control false discovery rate, which is quite significant in genomics, disease discovery, and cancer discovery. Langaas et al. [1] remarked “An important reason for wanting to estimate π_0 is that it is a quantity of its own right. In addition, a reliable estimate of π_0 is important when we want to assess or control multiple error rates, such as the false discovery rate FDR of Benjamini and Hochberg [2].” In the case of testing for differential expression in DNA microarrays, the proportion of differentially expressed genes is $1 - \pi_0$, and it is important to know whether 5% or 35% of the genes, for example, are differentially expressed, even if we cannot identify these genes (see Langaas et al. [1]). Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. A common feature in genomes studies is the analysis of a large number of simultaneous measurements in a small number of samples. One must decide whether the findings are truly causative correlations or just the byproducts

of multiple hypothesis testing (Gyorffy et al. [3]). If one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected by chance alone may be unduly large.

In a multiple hypothesis testing problem, m null hypotheses are tested simultaneously; that is, we test

$$H_{0i} \text{ versus } H_{1i}, \quad (1)$$

for $i = 1, 2, \dots, m$, simultaneously. Assume that the m tests are constructed based on the observed p values, $p_1, \dots, p_m$, respectively. The unknown quantity π_0 to be estimated is the proportion of the true null hypotheses among $H_{01}, \dots, H_{0m}$. Introduce the i.i.d Bernoulli random variables $H_1, \dots, H_m$ with $P(H_i = 0) = \pi_0$. Then H_i can be interpreted in terms of the multiple testing problems as follows:

$$H_i = \begin{cases} 0, & \text{if } H_{0i} \text{ is true,} \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

for $i = 1, \dots, m$.

TABLE 1: Outcomes from m null hypotheses tests.

	Not rejected	Rejected	Total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	W	R	m

We assume the p values, $p_1, \dots, p_m$, are continuous and independent random variables, so that the p values are independently and identically distributed as $\text{Unif}(0, 1)$ when the null hypotheses are all true. One chooses to reject or fail to reject each null hypothesis based on the corresponding p value. Consequences of the tests are summarized in Table 1.

In Table 1 m_0 is the number of null hypotheses; R is the observable random variable representing the number of hypotheses rejected. Note that all other random variables U, V, S , and T in Table 1 are unobservable.

The problem of estimating the proportion π_0 has naturally arisen in assessing or controlling an overall false rejection rate in a simultaneous hypotheses testing problems. A reliable estimate of π_0 is crucial when we want to control and/or assess the false discovery rate (FDR) proposed by Benjamini and Hochberg [2], defined as

$$\text{FDR} = E \{QI(R > 0)\}, \quad (3)$$

where $Q = V/R$ and

$$I(R > 0) = \begin{cases} 1 & \text{if } R > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Benjamini and Hochberg [2] prove that Simes' procedure (Simes [4]) has the FDR controlled at level α if the underlying test statistics and the corresponding p values are continuous and identically and independently distributed. Specifically, they show

$$\text{FDR} \leq \frac{m_0}{m} \alpha = \pi_0 \alpha. \quad (5)$$

Consequently, if γ is the FDR level that one wishes to achieve and if π_0 can be estimated efficiently, say by $\hat{\pi}_0$, then α can be chosen to be $\alpha = \gamma/\hat{\pi}_0$ to gain additional testing powers in the multiple testing problem with FDR being under control. If $\hat{\pi}_0$ is overestimated substantially, however, the value of α is underestimated substantially, leading to significantly narrower simultaneous confidence intervals for multiple comparisons and significant reduction of testing power for the multiple testing problems. On the other hand, if α is chosen via some other procedure, say the Bonferroni method in the multiple comparison problem, assessing FDR accurately via estimating π_0 efficiently is then of great interest.

The paper is organized as follows. Section 2 is a review of existing estimating methods; Section 3 introduces the new estimating procedure; Section 4 contains simulation results and Section 5 presents application of the new estimating procedure to real life examples.

2. Existing Estimating Methods

2.1. Mixture Model Framework. A mixture model can be used to fit the p values in order to estimate the proportion, π_0 , of the true hypotheses, where large scale parallel hypotheses are performed independently. The estimate for π_0 can be based on the mixture model for the common density f of the p values described as follows:

$$f = \pi_0 + (1 - \pi_0) h, \quad (6)$$

where $h(p)$ is the conditional probability density function of the p value under an alternative (see Langaas et al. [1]). Using this mixture representation, we are able to characterize the maximum likelihood estimate for f . The estimation method is derived under the assumption of independent and identically distributed p values. The null p values are uniformly distributed on $[0, 1]$. We should note that $h(p)$ describes the configuration of true alternative populations among the m underlying populations; it seems that a nonparametric approach to the estimation of π_0 is more appealing. Without loss of generality, we define $H = 1$ as the alternative hypothesis, which will be used throughout this section. Three nonparametric estimators proposed recently by other authors are described and discussed in the sequel subsections.

2.2. Storey's Method. Consider the common marginal mixture density of the p values, for any $\lambda \in (0, 1)$:

$$P(p > \lambda) = \pi(1 - \lambda) + (1 - \pi)P(p > \lambda | H = 1). \quad (7)$$

On the basis that $P(p > \lambda | H = 1)$ is typically small, a large majority of the p values in the interval $[\lambda, 1]$ should be corresponding to the true null hypothesis and thus uniformly distributed on the interval $(0, 1)$, of which most should be close to 1, so that $(1 - \pi)P(p > \lambda | H = 1)$ is approximately zero. Let $W(\lambda) = \#\{p_i : p_i > \lambda\}$. Note that $W(\lambda)$ should be approximately equal to the product of $m\pi_0$ and the length of the interval $[\lambda, 1]$; that is, $E(W(\lambda)) \approx m\pi_0(1 - \lambda)$. Hence, Storey [6] proposes that the proportion of true null hypotheses, π_0 , is estimated by (with an appropriately chosen λ)

$$\hat{\pi}_0^s(\lambda) = \frac{W(\lambda)}{(1 - \lambda)m}. \quad (8)$$

The value of λ has impacts on the behavior of $\hat{\pi}_0^s(\lambda)$. $\hat{\pi}_0^s(\lambda)$ has a large bias and small variance when λ is small and a small bias and large variance when λ is big, respectively. Since both extreme values of λ have a bias-variance trade-off, Storey et al. [7] propose bootstrapping, which is a resampling technique, to choose λ when estimating $\hat{\pi}_0^s(\lambda)$ so as to minimize the mean square error of $\hat{\pi}_0^s(\lambda)$. The resulting estimator is denoted by $\hat{\pi}_0^s$.

Note that the idea leading to Storey's estimator $\hat{\pi}_0^s$ is to treat the addition $(1 - \pi)P(p > \lambda | H = 1)$ term as zero to get rid of the complication caused by the unknown alternative distribution. As a consequence, Storey's estimator $\hat{\pi}_0^s$ will tend to overestimate π_0 in a size of $(1 - \pi_0)P(p > \lambda | H = 1)/(1 - \lambda)$, at least theoretically. However, the anticipated size

of overestimate becomes less significant when π_0 is closer to 1; that is, $1 - \pi_0$ is closer to 0. The biasedness of $\hat{\pi}_0^c$ as an estimator for $P(p > \lambda)/(1 - \lambda)$ can be dominating, when π_0 is close to 1, as evident in the simulation results in Section 4 and as observed by other authors as well, for example, Langaas et al. [1].

2.3. Convex Method. Let f and h be defined in (6). Langaas et al. [1] prove that if $h(p)$ is twice differentiable, convex, and decreasing, f can be expressed as

$$f(p) = \int_0^1 f_\theta(p) \mu(d\theta), \quad (9)$$

where the kernel density

$$f_\theta(p) = \begin{cases} \frac{2(\theta - p)}{\theta^2 I_{(0,\theta)}(p)} & \text{if } \theta \in (0, 1], \\ I_{[0,1]}(p) & \theta = 0, \end{cases} \quad (10)$$

with μ being any probability measure on $[0, 1]$. Thus Langaas et al. [1] are able to characterize the nonparametric maximum likelihood estimate for f for the density $f(p)$ as

$$\hat{f}(p) = \int_0^1 f_\theta(p) \hat{\mu}(d\theta), \quad (11)$$

where $\hat{\mu}$ is the nonparametric maximum likelihood estimate of μ . Let p values $p_{(1)}, \dots, p_{(m)}$ be the ordered statistics of the p values. The nonparametric maximum likelihood estimator of $f(\cdot)$ is given by

$$\hat{f} = \arg \max \prod_{i=1}^m \hat{f}(p_{(i)}). \quad (12)$$

Langaas et al. [1] then propose to estimate the proportion π_0 by

$$\hat{\pi}_0^c = \hat{f}(1). \quad (13)$$

It is noted that as the estimator $\hat{\pi}_0^c$ is constructed via a density estimate $\hat{f}(p)$ at or about the upper bound of support, namely, $\hat{f}(1)$, it can be conservative and overestimate π_0 when the assumption $h(1) = 0$ is questionable or $h(p) \rightarrow 0$ slowly as $p \rightarrow 1$. The overestimating problem can be even more severe when π_0 is not large; that is, $1 - \pi_0$ is not so small. However, the remarks in the end of Section 2.2 are applicable to $\hat{\pi}_0^c$ as well.

2.4. Average Estimate Approach. The average estimate approach was motivated by Storey's method. Jiang and Doerge [8] observe (and many other authors point out as well) that Storey's estimator has a large bias and small variance when λ is small and a small bias and large variance when λ is big. Since both extremes of λ have a bias-variance trade-off, Jiang and Doerge [8] propose to combine Storey's estimate with different values of λ that vary from a small extreme to a large extreme.

Let $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n < 1$, and suppose that, for each λ_i , we have Storey's estimate

$$\hat{\pi}_0(\lambda_i) = \frac{W(\lambda_i)}{(1 - \lambda_i)m}. \quad (14)$$

Jiang and Doerge [8] propose to estimate π_0 by the average of $\hat{\pi}_0(\lambda)$ over the values of λ_i ; that is,

$$\hat{\pi}_0^a = \frac{1}{n} \sum_{i=1}^n \hat{\pi}_0(\lambda_i). \quad (15)$$

This estimate aims to minimize both the bias and the variance at the same time, if λ_i 's are selected appropriately.

Define $0 = t_1 < t_2 < \dots < t_B < t_{B+1} = 1$ 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, $t_i = (i - 1)/B$. Let $b_i = \#\{p_k : p_k \geq t_i\}$ and $c_i = \#\{p_k : t_i \leq p_k < t_{i+1}\}$ for $i = 1, \dots, B$. Define

$$i^* = \min \left\{ i : c_i \leq \frac{b_i}{(B - i + 1)} \right\}. \quad (16)$$

Jiang and Doerge [8] then propose to estimate π_0 by

$$\hat{\pi}_0(B) = \frac{1}{B - i^* + 1} \sum_{j=1}^B \hat{\pi}_0(t_j). \quad (17)$$

To apply the estimate, one has to choose a value of B . Jiang and Doerge [8] develop a bootstrapping algorithm to pick the optimal B . It should be noted that, as an average of Storey's estimates, this estimate is expected to inherit conservativeness from Storey's method.

3. New Method

We propose a finite mixture model of the uniform distribution and a multinomial distribution $M(k, q_1, \dots, q_k)$ with a mixing proportion π_0 to fit the p values. Denote $UM(\pi_0; k, q_1, \dots, q_k)$ for this finite mixture distribution. By this approach, the alternative distribution h defined in (6) is restricted to the multinomial distribution family, $M(k, q_1, \dots, q_k)$. Alternatively, the multinomial distribution can be viewed as a parametric approximation to the nonparametric unknown density h , similarly to the idea of the empirical likelihood method (see Owen [9]). So this procedure is considered as nonparametric.

To apply the approach, we need to settle two things first: (a) convert the continuous-type observations $p_1, \dots, p_m$ into discrete data with k categories and (b) select an integer k .

3.1. Selection of k . It is often the case in applications that the p values are highly skewed (see Storey and Tibshirani [10], Zhao et al. [11], and Markitsis and Lai [12]). In this case, we recommend Doane's modification of Sturges' rule for selection of k as follows, to count for skewness of the mixture distribution (6) (see Doane [13] and Sturges [14]):

$$\hat{k} = \log_2(1 + \hat{y}), \quad (18)$$

where $\hat{\gamma}$ is an estimate of the skewness coefficient. In the case of symmetry, we recommend to adapt Sturges' rule to determine the selection of k :

$$\hat{k} = 1 + \log_2 m. \quad (19)$$

3.2. Transformation of p_i 's. To transform p_i 's, partition the unit interval into $\hat{k}$ subintervals with equal width $1/\hat{k}$. Define

$$\xi_{ij} = \begin{cases} 1, & \text{if } \frac{(j-1)}{\hat{k}} \leq p_i < \frac{j}{\hat{k}}, \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

for $i = 1, \dots, m, j = 1, \dots, \hat{k}$. Keep in mind that

$$\begin{aligned} \sum_{j=1}^{\hat{k}} \xi_{ij} &= 1, \quad i = 1, \dots, m, \\ \sum_{i=1}^m \sum_{j=1}^{\hat{k}} \xi_{ij} &= m. \end{aligned} \quad (21)$$

From Storey's Bayesian interpretation for the multiple testing problem, given that the alternative is true with a probability of $1 - \pi_0$, the p value follows the distribution h . In the same way, the transformed data ξ_{ij} 's can be interpreted as follows. Given that the alternative is true with a probability of $1 - \pi_0$, $\xi_i = (\xi_{i1}, \dots, \xi_{i\hat{k}})'$ is a multinomial random vector and is distributed as $M(k, q_1, \dots, q_k)$, for $i = 1, \dots, m$. Therefore, $\xi_1, \dots, \xi_m$ are independently and identically distributed as the finite mixture distribution $UM(\pi_0; k, q_1, \dots, q_k)$ and the maximum likelihood estimate $\hat{\pi}_0$ for π_0 thus results from the transformed data ξ_{ij} 's. Explicitly, $\hat{\pi}_0$ together with $(\hat{q}_1, \dots, \hat{q}_k)$ maximizes the log-likelihood function

$$l(\pi_0, q_1, \dots, q_k) = \sum_{i=1}^m \log \left\{ \pi_0 + (1 - \pi_0) \prod_{j=1}^k q_j^{\xi_{ij}} \right\}. \quad (22)$$

3.3. EM Algorithm. Note that maximizing the nonlinear log-likelihood function (22) can be complicating. The EM algorithm may be used easily to obtain an approximation to $\hat{\pi}_0$. In order to do so, we introduce a latent Bernoulli variable w that indicates the component membership in the finite mixture distribution $UM(\pi_0; k, q_1, \dots, q_k)$. That is, $P(w = 1) = \pi_0$ and given $w = 0$, ξ_1 follows $M(k, q_1, \dots, q_k)$. Note that (w, ξ_1) has the distribution

$$\begin{aligned} g(w, \xi_1 | \pi_0, q_1, \dots, q_k) \\ = \pi_0^w \left\{ (1 - \pi_0) q_1^{\xi_{11}} q_2^{\xi_{12}} \dots q_k^{\xi_{1k}} \right\}^{1-w}, \end{aligned} \quad (23)$$

for $w = 0$ or 1 and $\xi_{1j} = 0$ or $1, j = 1, \dots, k$.

Let $(w_i, \xi_{i1}, \dots, \xi_{ik}), i = 1, \dots, m$, be a random sample of size m from model (23). In the present problem, ξ_{ij} 's are only available data for analysis and w_i 's are unobservable and

so considered as missing values. This defines a missing-value model. The log-likelihood with the complete data is given by

$$\begin{aligned} l^*(\pi_0, q_1, \dots, q_k) &= T \log \pi_0 + (m - T) \log (1 - \pi_0) \\ &\quad + \sum_{j=1}^k (m - T) \xi_{.j} \log q_j, \end{aligned} \quad (24)$$

where $T = \sum_{i=1}^m w_i$, and $\xi_{.j} = \sum_{i=1}^m \xi_{ij}$, for $j = 1, \dots, k$. We are ready to describe the EM algorithm.

E-Step. Let $\pi_0^{(s)}$ and $q^{(s)}$ be the current approximations to the maximum likelihood estimates under the model $UM(\pi_0, k, q_1, \dots, q_k)$. For the next approximation, the E-step establishes the expected log-likelihood function as

$$\begin{aligned} Q(\pi_0, q) &= E_{\pi_0^{(s)}, q^{(s)}} \{ l^*(\pi_0, q_1, \dots, q_k | \eta_1, \dots, \eta_m) \} \\ &= E_{\pi_0^{(s)}, q^{(s)}} (T | \eta) \log \pi_0 \\ &\quad + (m - E_{\pi_0^{(s)}, q^{(s)}} (T | \eta)) \log (1 - \pi_0) \\ &\quad + \sum_{j=1}^k (m - E_{\pi_0^{(s)}, q^{(s)}} (T | \eta)) \xi_{.j} \log q_j \end{aligned} \quad (25)$$

M-Step. In the M-step, $Q(\pi_0, q)$ is maximized to yield the next approximation $\pi_0^{(s+1)}$ and $q^{(s+1)}$. Consider

$$\begin{aligned} \frac{\partial Q}{\partial \pi_0} &= 0 \implies \\ \pi_0^{(s+1)} &= \frac{E_{\pi_0^{(s)}, q^{(s)}} (T | \eta)}{m} = \sum_{i=1}^m E_{\pi_0^{(s)}, q^{(s)}} (w_i | \eta). \end{aligned} \quad (26)$$

where

$$\begin{aligned} E_{\pi_0^{(s)}, q^{(s)}} (w_i | \eta) &= P_{\pi_0^{(s)}, q^{(s)}} (w_i = 1 | \eta) \\ &= \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1 - \pi_0^{(s)}) \prod_{j=1}^k [q_j^{(s)}]^{\xi_{ij}}}. \end{aligned} \quad (27)$$

From $\partial Q / \partial q = 0$, we have

$$q_j^{(s+1)} = \frac{\xi_{.j}}{m - E_{\pi_0^{(s)}, q^{(s)}} (T | \eta)} = \frac{\xi_{.j}}{m (1 - \pi_0^{(s+1)})}. \quad (28)$$

To sum up, let $\pi_0^{(s)}, q_1^{(s)}, \dots, q_k^{(s)}$ be the s th approximations to the maximum likelihood estimates of $\hat{\pi}_0, \hat{q}_1, \dots, \hat{q}_k$ that maximize the log-likelihood function defined in (22). Then the $(s + 1)$ th approximation with the EM algorithm is given by

$$\pi_0^{(s+1)} = \frac{1}{m} \sum_{i=1}^m \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1 - \pi_0^{(s)}) \prod_{j=1}^k [q_j^{(s)}]^{\xi_{ij}}}, \quad (29)$$

$$q_j^{(s+1)} = \frac{\xi_{.j}}{m (1 - \pi_0^{(s+1)})}.$$

So the EM iteration process goes as shown in Algorithm 1.

Input: $\xi_{ij}, i = 1, \dots, k, j = 1, \dots, m$ transformed from the observed p -values $p_1, \dots, p_m$

Output: Estimate of $\pi_0 : \hat{\pi}_0$

(1) **begin**

(2) Initialization: Set $\pi_0 = \pi_0^{(o)}, q_j = q_j^{(o)}, 1 \leq j \leq k$.

(3) **repeat**

(4) Set $\pi_0 = \pi_0^{(s)}, q_j = q_j^{(s)}$ be the s th (current) approximation.

Compute

$$\pi_0^{(s+1)} = \frac{1}{m} \sum_{i=1}^m \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1 - \pi_0^{(s)}) \prod_{j=1}^k [q_j^{(s)}]^{\xi_{ij}}},$$

$$q_j^{(s+1)} = \frac{\xi_{\cdot j}}{m(1 - \pi_0^{(s+1)})}$$

(5) **until** $l(\pi_0^{(s+1)}, q_1^{(s+1)}, \dots, q_k^{(s+1)}) - l(\pi_0^{(s)}, q_1^{(s)}, \dots, q_k^{(s)}) \leq \epsilon$;

(6) Then $\hat{\pi}_0 = \pi_0^{(s+1)}$.

ALGORITHM 1: EM iteration.

It is well known that each EM iteration gets closer to the maximum of log-likelihood but only in a linear convergency rate. If the components are similar in their densities, then the convergence is extremely slow. The convergence will also be slow when the maximum likelihood solution requires some of the weight parameters to be zero, because the algorithm can never reach such a boundary point. An additional and related problem is that of deciding when to stop the algorithm. One risk to a naive user is the natural tendency to use a stopping rule for the algorithm based on the changes in the parameters or the likelihood being sufficiently small. Taking small steps does not indicate that we are close to the solution.

To combat this problem, Lindsay [15] exploited the regularity of the EM algorithm process to predict, via the device known as Aitken acceleration, the value of the log-likelihood at the maximum likelihood solution. The Aitken's acceleration rule is usually recommended to predict the maximum efficiently and suitably whenever one is using a linearly convergent algorithm with a slow rate of convergence. If we let l_{i-2}, l_{i-1} , and l_i be the log-likelihood values for the three consecutive iterations, then the predicted final value is

$$l_i^\infty = l_{i-2} + \frac{l_{i-1} - l_{i-2}}{(1 - c_i)}, \quad (30)$$

where $c_i = (l_i - l_{i-1}) / (l_{i-1} - l_{i-2})$. Terminate the EM iteration when $l_i^\infty - l_i$ is sufficiently small.

4. Simulation Studies

In order to investigate the properties of the estimators described in Section 2 and compare the performance to that of the newly developed estimating procedure described in Section 3, we conducted simulation experiments. The generation of simulated data and the calculation of the estimates were both done in the language *R*. Simulation studies were conducted with the p values based on a one-sided z -test

in the finite normal mixture model $z \sim \pi_0 N(0, 1) + (1 - \pi_0) N(1, 1)$ with various values of m and π_0 for performance comparisons.

Monte Carlo data were simulated independently from $z \sim \pi_0 N(0, 1) + (1 - \pi_0) N(1, 1)$ and each p value was computed by $p_i = 1 - \Phi(z)$, where Φ is the cumulative distribution function of standard normal $N(0, 1)$. For the generation of simulated data, three true values of π_0 were considered, namely, 0.25, 0.5, 0.75, with the sample size $m = 200, 500$, and 1,000. In each case, 1,000 Monte Carlo trials were performed. In implementation of the EM algorithm to compute the proposed new estimate $\hat{\pi}_0$, the determination rule of the EM algorithm is $l_i^\infty - l_i \leq 10^{-4}$. The EM algorithm converged for all the simulated datasets. In each case, the performance of the proposed new estimate $\hat{\pi}_0$ was compared with those of several existing procedures through the same set of Monte Carlo data. Specifically, in the simulations, we considered the following existing methods:

- (1) Storey's bootstrap estimate denoted by $\hat{\pi}_0^b$
- (2) Langaas et al.'s convex estimate denoted by $\hat{\pi}_0^c$
- (3) Jiang and Doerge's average method estimate denoted by $\hat{\pi}_0^a$

Storey's estimate $\hat{\pi}_0^s$ was computed by using the *R* package *nFDR* and the convex estimate $\hat{\pi}_0^c$ by the *R* package *limma* with the function *convest*. Table 2 summarizes the simulation results. It is clearly shown that the new estimate outperformed substantially over the existing methods with comparable standard errors, while $\hat{\pi}_0^s(0.5)$ performs much worse than $\hat{\pi}_0^s$ as expected.

5. Application to Real Life Microarray Data

To further evaluate the performance of the new method in comparison to the three existing methods, consider the real life data from the DNA microarray experiments reported in Golub et al. [5] that can be downloaded from the *R* package *multtest*. The dataset was used by many authors (see [11, 16])

TABLE 2: Monte Carlo empirical estimates with standard deviations (in parentheses) by the new innovative approach along with existing estimators. The Monte Carlo size is 1,000 in each case. Denote $\hat{\pi}_0$ for the new estimator, $\hat{\pi}_0^s$ for Storey's estimator, $\hat{\pi}_0^s(0.5)$ for Storey's estimator with $\lambda = 0.5$, $\hat{\pi}_0^c$ for the convex method estimator, and $\hat{\pi}_0^a$ for the average method estimator.

m	π_0	$\hat{\pi}_0$	$\hat{\pi}_0^s$	$\hat{\pi}_0^s(0.5)$	$\hat{\pi}_0^c$	$\hat{\pi}_0^a$
Estimates of π_0						
200	0.25	0.235	0.30	0.500	0.314	0.430
		(0.016)	(0.021)	(0.035)	(0.022)	(0.03)
	0.5	0.489	0.613	0.660	0.584	0.626
		(0.035)	(0.043)	(0.047)	(0.041)	(0.044)
	0.75	0.78	0.717	0.79	0.685	0.773
		(0.055)	(0.051)	(0.056)	(0.048)	(0.054)
500	0.25	0.249	0.34	0.528	0.33	0.491
		(0.011)	(0.015)	(0.024)	(0.014)	(0.022)
	0.5	0.52	0.646	0.720	0.623	0.704
		(0.023)	(0.028)	(0.032)	(0.027)	(0.031)
	0.75	0.72	0.806	0.860	0.785	0.819
		(0.032)	(0.036)	(0.038)	(0.035)	(0.035)
1000	0.25	0.242	0.393	0.482	0.349	0.394
		(0.005)	(0.012)	(0.015)	(0.011)	(0.013)
	0.5	0.51	0.61	0.66	0.61	0.63
		(0.016)	(0.019)	(0.021)	(0.019)	(0.019)
	0.75	0.72	0.67	0.86	0.79	0.81
		(0.023)	(0.021)	(0.027)	(0.025)	(0.026)

and the references therein) to illustrate the applications of proposed estimating methods for the proportion of true null hypotheses in multiple testing problems. The dataset consists of 38 bones marrow samples, 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples, obtained from acute leukemia patients at the time of diagnosis, before chemotherapy. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 7,129 human genes. But after pre-processing, there were only 3,051 genes accurately readable resulting in a microarray 3051×38 matrix. For comparison of gene expressions between two groups, let μ_1 and μ_2 be the true mean intensities for each gene, in groups 1 and 2, respectively, to determine whether the gene is differentially expressed, that is, to test

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2. \quad (31)$$

With each of 3,051 genes, a two-sample Welch t -statistic is computed and its two-sided p value is computed under a central t -student distribution with degree 36 of freedom. A histogram of these p values is displayed in Figure 1. It is clear that the p values are highly skewed to the right.

The four estimates for the proportion π_0 of nondifferentially expressed genes among 3,051 genes, π_0 , $\hat{\pi}_0$, $\hat{\pi}_0^s$, $\hat{\pi}_0^c$, and $\hat{\pi}_0^a$, are given in Table 3. It appears that the four estimates are significantly different, varying from as low as 0.4150 to as high as 0.4913. Noting that it has been commented by

FIGURE 1: Histogram of the p values for leukemia data reported in Golub et al. [5].

TABLE 3: The estimates of the proportion of true null hypotheses with the microarray data of Golub et al. [5].

$\hat{\pi}_0$	$\hat{\pi}_0^s$	$\hat{\pi}_0^c$	$\hat{\pi}_0^a$
0.4150	0.4643	0.4701	0.4913

many authors that the estimates $\hat{\pi}_0^s$, $\hat{\pi}_0^c$, and $\hat{\pi}_0^a$ are usually conservative, we conclude that the lower estimate of 0.42 by the proposed new estimating procedure appears as expected and might be closer to the true value of π_0 .

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 67, no. 4, pp. 555–572, 2005.
- [2] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 57, no. 1, pp. 289–300, 1995.
- [3] B. Györfy, A. Györfy, and Z. Tulassay, "The problem of multiple testing and solutions for genome-wide studies," *Orvosi Hetilap*, vol. 146, pp. 559–563, 2005.
- [4] R. J. Simes, "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 73, no. 3, pp. 751–754, 1986.
- [5] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

- [6] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 64, no. 3, pp. 479–498, 2002.
- [7] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 66, no. 1, pp. 187–205, 2004.
- [8] H. Jiang and R. W. Doerge, "Estimating the proportion of true null hypotheses for multiple comparisons," *Cancer Informatics*, vol. 6, pp. 25–32, 2008.
- [9] A. Owen, "Empirical likelihood for linear models," *The Annals of Statistics*, vol. 19, no. 4, pp. 1725–1747, 1991.
- [10] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.
- [11] H. Zhao, X. Wu, H. Zhang, and H. Chen, "Estimating the proportion of true null hypotheses in nonparametric exponential mixture model with application to the leukemia gene expression data," *Communications in Statistics. Simulation and Computation*, vol. 41, no. 9, pp. 1580–1592, 2012.
- [12] A. Markitsis and Y. Lai, "A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes," *Bioinformatics*, vol. 26, no. 5, pp. 640–646, 2010.
- [13] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, vol. 30, no. 4, pp. 181–183, 1976.
- [14] H. A. Sturges, "The choice of a class interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926.
- [15] B. G. Lindsay, "Mixture models: theory, geometry and applications," in *Proceedings of the NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute for Mathematical Statistics, 1995.
- [16] Z. Guan, B. Wu, and H. Zhao, "Application of Bernstein polynomial in the estimation of false-discovery-rate," *Statistica Sinica*, vol. 18, pp. 905–923, 2008.

Hindawi

Submit your manuscripts at
<http://www.hindawi.com>