Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 21 Issue 5 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Estimating the Proportion of True Null Hypotheses: A Likelihood Approach

By Hualing Zhao & Hanfeng Chen
Wuhan University of Technology

Abstract- Many estimators for the proportion π_0 of the true null hypotheses in a multiple testing problem have been proposed in literature. Motivated from the work on the histogram approach, in this article we propose a new estimator based on the likelihood function with an approximating alternative histogram. AIC is used to select the number of bins for the histogram. Simulation study demonstrates that the new estimator outperforms and substantially improves existing methods including Storey estimators, convex density estimator, and histogram estimator. The new method is applied to a real-life data set of breast cancer.

Keywords: akaike information criterion; false discovery rate; finite mixture model; multiple comparisons.

GJSFR-F Classification: MSC 2010: 97K80

Strictly as per the compliance and regulations of:

Estimating the Proportion of True Null Hypotheses: A Likelihood Approach

Hualing Zhao [α] & Hanfeng Chen [σ]

Abstract- Many estimators for the proportion π_0 of the true null hypotheses in a multiple testing problem have been proposed in literature. Motivated from the work on the histogram approach, in this article we propose a new estimator based on the likelihood function with an approximating alternative histogram. AIC is used to select the number of bins for the histogram. Simulation study demonstrates that the new estimator outperforms and substantially improves existing methods including Storey estimators, convex density estimator, and histogram estimator. The new method is applied to a real-life data set of breast cancer.

Keywords: *akaike information criterion; false discovery rate; finite mixture model; multiple comparisons.*

I. Introduction

Consider the problem of estimating the proportion π_0 of true null hypotheses in a collection of m tests, given the observed p-values $p_1, \cdots, p_m$ for the collection of m tests. This problem has attracted a lot of attentions in statistical literatures, attributing to its important role in dealing with multiple testing procedures, since the nominal paper Storey (2002). It has naturally arisen in assessing or controlling an overall false rejection rate, i.e., the false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) in multiple null hypothesis testing problems and multiple comparisons. Benjamini and Hochberg proved that if $p_1, \cdots, p_m$ are independent with continuous distributions, the popular Sime 's multiple testing procedure (Sime 1986), where a null hypothesis is rejected whenever the observed p-value is less than α, results in an FDR controlled by $\pi_0\alpha$. Therefore, a reliable estimate is essential to control the FDR in the Sime's multiple testing procedure and improve its testing power. On the other hand, the proportion π_0 is a quantity of interest in its own right. For example, researchers may wish to estimate the proportion of genes that are not differentially expressed in DNA microarray experiments (Parker et. al. 1988).

Store (2002) proposed to approach the estimating problem from a Bayesian point of view by treating $p_1, \cdots, p_m$ as a random sample of size m from a mixture distribution of a uniform distribution and a non-uniform distribution with mixing proportion π_0. Specifically, consider a population consisting of all outcomes of the p-values $p_1, \cdots, p_m$ in testing m null hypotheses. There are two types of p-values: true-null p-values and false-null p-values. Following Tong, Feng, Hilton and Zhao (2013)'s terminology, by true-null p-value we mean the p-value is observed from a true null hypothesis and by false-null p-value we mean it is observed from a false null hypothesis. Thus, the population distribution can be described as a finite mixture with mixing proportion π_0 of the uni-

Author α: Department of Statistics, School of Sciences, Wuhan University of Technology, Wuhan, Hubei 430070, China.
e-mail: zhaohualing2011@whut.edu.cn
Author σ: Department of Mathematics and Statistics, Bowling Green State University, Bowling, Green, OH 43403, USA.
e-mail: hchen@bgsu.edu

form distribution on $[0,1]$ and another non-uniform distribution with the pdf, say $h(x)$, with respect to the Lebesque measure dx. Denote the mixture by $f(x|\pi_0,h)$, i.e.,

$$f(x|\pi_0,h)=\pi_0+(1-\pi_0)h(x),\quad 0\le x\le 1. \tag{1}$$

By Storey's approach, the observed p-values are considered to be a random sample of size m from $f(x|\pi_0,h)$. As a consequence, the mixing proportion π_0 in the mixture model (1) represents the proportion of true-null p-value subpopulation and an estimate for π_0 based on a random sample from (1) thus yields an estimate for the proportion of true null hypotheses among the m null hypotheses in the multiple testing problem.

The first well-studied estimator for π_0 was proposed by Storey (2002) as follows:

$$\hat{\pi}_0^s(\lambda)=W(\lambda)/\{m(1-\lambda)\},$$

where λ is an appropriately chosen value close to 1 and $W(\lambda)$ is the total number of the p-values greater than λ. This estimator is motivated by the fact that false-null p-values are typically small so that with λ close to 1, $1-\lambda$ is close to the expected proportion of true-null p-values falling into $(\lambda,1]$. Since the proportion of the p-values falling into $(\lambda,1]$ is a mixture of $1-\lambda$ and $\int_\lambda^1 h(x)dx$ with mixing proportions π_0, $\hat{\pi}_0^s$ is obtained mathematically by simply setting the proportion of false-null p-values falling into $(\lambda,1]$ to be zero, i.e., $\int_\lambda^1 h(x)dx=0$. As a consequence, $\hat{\pi}_0^s$ tends to overestimate π_0, as $\int_\lambda^1 h(x)dx$ is typically positive. The biasedness can be too great to be acceptable when π_0 belongs to the lower part of $[0,1]$. Storey (2002) proposed a bootstrap procedure to choose λ that minimizes an upper bound of the mean square error of the resulting positive FDR estimator $\hat{Q}(\lambda)=\alpha\hat{\pi}_0^s(\lambda)/R(\alpha)$ where $R(\alpha)=\#\{p_i\le\alpha\}$. See more discussion and some other work on how to select λ, see Storey, Taylor and Siegmund (2004), and Nettleton, Hwang, Caldo and Wise (2006).

Since the publication of Storey (2002), the mixture model approach described above has been adopted widely and many other estimators have been proposed. They include: Langaas, Lindqvist and Ferkingstad (2005)'s by a histogram approach, Wu, Guan and Zhao (2006)'s polynomial-type estimator, Jiang and Doerge (2008)'s estimator averaging over different λ-values with Storey (2002)'s estimator, Zhao, Wu, Zhang and Chen (2012)'s estimator in exponential mixture model, Cheng, Gao and Tong (2015)'s estimator with reduction of estimating bias and standard deviation. For more references and discussion, see Cheng, Gao and Tong (2015), and Tong, Feng, Hilton and Zhao (2013) that deals with possibly dependent p-values. Nevertheless, when these estimators are intended to improve some aspects of Storey's estimator, the biasedness remains significant even with m as large as 2,000, especially when π_0 is not close to 1. Motivated by the histogram approach (see Mosig et al. 2001 and Nettleton et al. 2006), a new estimator is proposed in this paper via a likelihood approach with h being approximated by a modified histogram pdf. Akaike information criterion is used to select the number of categories in histogram construction. Simulation study demonstrates that the new estimator significantly improves popularly cited existing methods such as Storey (2002), Langaas, Lindqvist and Ferkingstad Langaas (2005), Jiang and Doerge (2008) and Nettleton et al.(2006).

The paper is organized as follows. The new method along with discussion of computing issues is presented in Section 2. Simulation results are reported in Section 3. A real-life example is given in Section 4.

II. Method

Let $p_1, ..., p_m$ be a random sample of size m from the pdf $f(x|\pi_0, h)$ defined in (1). We propose π_0 to be estimated by the MLE when h is subject to a histogram-type approximation. We shall proceed with a study on the identification problem in the mixture model.

Notes

a) Identification

Let Θ be the parameter space consisting of all (π_0, h)'s with $\pi_0 \in (0, 1)$ and h being a probability density function on [0,1] and continuous in a neighborhood of 1with $h(1-) = 0$, where

$$h(1-) = \lim_{x \to 1-} h(x).$$

For convenience and confirmation, when a pdf is used as a parameter throughout the paper, we should refer to the probability distribution, say the CDF specified by the pdf.

Lemma 1 *The distribution $f(x|\pi_0, h)$ in the model (1) is identifiable in $(\pi_0, h) \in \Theta$.*

Proof. Let two parameter sets (π_1, h_1) and (π_2, h_2) in Θ define the same distribution, i.e., for any $x \in [0, 1]$,

$$\pi_1 x + (1 - \pi_1)H_1(x) = \pi_2 x + (1 - \pi_2)H_2(x), \tag{2}$$

where H_i is the CDF of h_i, $i = 1, 2$. Since h_1 and h_2 are continuous in a common neighborhood of 1, (2) implies that for any x in a neighborhood of 1,

$$\pi_1 + (1 - \pi_1)h_1(x) = \pi_2 + (1 - \pi_2)h_2(x).$$

So $\pi_1 = \pi_2$ because $h_1(1-) = h_2(1-) = 0$. This in turn implies $H_1 = H_2$ in (2) because $\pi_1 < 1$ and $\pi_2 < 1$. The lemma is proved.

Remarks.

(a) First of all, note the simple fact that $h(1-) = 0$ implies that h is non-uniform. Of course, identification of the parameter π_0 of interest requires that h is non-uniform.

(b) When $\pi_0 = 1$, f is un-identifiable in h.

(c) The assumption that h is continuous in a neighborhood of 1 with $h(1-) = 0$ is critical. Without it, the lemma can fail. For a counterexample, let $\pi_1 = 1/3$ and $h_1(x) = 1/4 + (3/2)x$, and $\pi_2 = 1/2$ and $h_2(x) = 2x$. We have

$$\int_0^1 h_1(x)dx = 1/4 + 3/4 = 1, \ \int_0^1 h_2(x)dx = 1,$$

and for any x in $[0, 1]$

$$\pi_1 + (1 - \pi_1)h_1(x) = 1/3 + (2/3)[1/4 + (3/2)x] = 1/2 + x,$$

and

$$\pi_2 + (1-\pi_2)h_2(x) = 1/2 + (1/2)(2x) = 1/2 + x.$$

That is, $\pi_1 + (1-\pi_1)h_1(x) = \pi_2 + (1-\pi_2)h_2(x)$ for all x in $[0,1]$, but $\pi_1 \neq \pi_2$ and $h_1 \neq h_2$.

b) A histogram approximation to h

Motivated by the histogram approach (see Mosig et al. (2001), a histogram approximation to the alternative pdf $h(x)$ is proposed as follows. Let $k > 2$ be an integer. Define

$$\tilde{h}(x) = \begin{cases} kq_j, & \text{if } (j-1)/k \leq x < j/k,\ 1 \leq j \leq k-1 \\ k^2 q_{k-1}(1-x), & \text{if} (k-1)/k \leq x \leq 1 \end{cases}$$

where $0 \leq q_1 \leq 1, \cdots, 0 \leq q_{k-1} \leq 1$ satisfying $q_1 + \ldots + q_{k-1} + q_{k-1}/2 = 1$. The linear modification over the right most subinterval $(1-1/k, 1]$ warrants $\hat{h} \in \Theta$ so that $f(x|\pi_0, \tilde{h})$ is identifiable in π_0 and $q = (q_1, \cdots, q_{k-2})$. As Storey(2002) and Tong et al. (2013) remarked, the false-null p-values that follow the distribution h are typically small. In other words it is often the case in application that the alternative pdf h is highly skewed to the right. Therefore, with the linear modification on the rightmost subinterval $(1-1/k, 1]$, the effect on the accuracy of the estimate for Storey (2002), Jiang and Doerge (2008) and some other methods assume that h is zero in a neighborhood of 1.

c) The new estimator with fixed k

Suppose that $k > 2$ is specific for now (a selection procedure will be described later). When h is approximated by or restricted to $\tilde{h}$, with a random sample $p_1, \cdots, p_m$, the log-likelihood of the parameter π_0 of interest and the new nuisance parameter q becomes

$$\begin{aligned} l(\pi_0, q) &= \sum_{i=1}^{m} \log f(p_i|\pi_0, \tilde{h}) \\ &= \sum_{i=1}^{m} \log \left\{ \pi_0 + (1-\pi_0)\{\prod_{j=1}^{k-1}(kq_j)^{\xi_{ij}}\}\{k^2 q_{k-1}(1-p_i)\}^{\xi_{ik}} \right\}, \end{aligned} \tag{3}$$

where ξ_{ij} is the indicator whether p_i falls into the j-th category of the histogram with k bins or not, i.e.,

$$\xi_{ij} = \begin{cases} 1, & \text{if } (j-1)/k \leq p_i < j/k, \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

for $i = 1, \cdots, m$, $j = 1, \cdots, k$.

Let $(\hat{\pi}_0(k), \hat{q})$ be the MLE of (π_0, q) with the log-likelihood function (3), subject to $0 < \pi_0 < 1$ and $0 < q_1 < 1, \cdots, 0 < q_{k-1} < 1$, $q_k = q_{k-1}/2$ such that $\sum_{j=1}^{k} q_j = 1$. This defines the new estimator $\hat{\pi}_0(k)$ for π_0 with specified k. A direct application of the standard theory of MLE yields the following theorem and hence a theory of consistency of the new estimator.

Theorem 1 *For any* $(\pi_0, h) \in \Theta$ *and* $k > 2$, *let*

$$q_j = \int_{(j-1)/k}^{j/k} h(x)dx$$

Notes

for $j = 1, \cdots, k-2$. *Put* $q = (q_1, \cdots, q_{k-2})$ *and* $\theta = (\pi_0, q)$. *Then* $\sqrt{m}(\hat{\pi}_0(k) - \pi_0, \hat{q} - q)$ *converges to* $N(0, \Sigma)$ *in distribution, as* $m \to \infty$, *where*

$$\Sigma = -\left[E\left(\frac{\partial^2}{\partial\theta\partial\theta}\log f(X|\pi_0, \tilde{h})\right)\right]^{-1}.$$

d) Computation of $\hat{\pi}_0(k)$

Maximizing the nonlinear log-likelihood function (3) can be complicating. However, the EM algorithm can be used to obtain an approximation to the MLE $\hat{\pi}_0(k)$ easily. To do so, introduce a latent Bernoulli variable w indicating the component-ship of the p-value in the finite mixture distribution. That is, let w be a binary random variable with $P(w = 1) = \pi_0$ and $P(w = 0) = 1 - \pi_0$, and let a random variable p follow the distribution as follows. Given $w = 1$, p follows the uniform distribution on $[0, 1]$ and given $w = 0$, $p \sim \tilde{h}$. It is clear that p follows the mixture distribution $f(x|\pi_0, \tilde{h})$. Thus $\mathcal{P} = \{p_1, \cdots, p_m\}$ can be viewed as the incomplete data of a random sample $(p_1, w_1), \cdots, (p_m, w_m)$ from (p, w) with missing values $w_1, \cdots, w_m$.

Note that with the complete data $(p_1, w_1), \cdots, (p_m, w_m)$, the likelihood function of (π_0, q) is

$$\prod_{i=1}^{m} \pi_0^{w_i} \left\{(1-\pi_0)[\prod_{j=1}^{k-1}(kq_j)^{\xi_{ij}}][k^2 q_{k-1}(1-p_i)]^{\xi_{ik}}\right\}^{1-w_i}, \quad (5)$$

where ξ_{ij}'s are defined in (4), and the log-likelihood function is thus

$$l^*(\pi_0, q) = w_\cdot \log \pi_0 + (m - w_\cdot)\log(1-\pi_0) + \sum_{j=1}^{k-1} \xi^*_{\cdot j} \log(kq_j) + \sum_{i=1}^{m} \xi^*_{ik} \log[k^2 q_{k-1}(1-p_i)], \quad (6)$$

where $\xi^*_{ij} = (1 - w_i)\xi_{ij}$, $w_\cdot = \sum_{i=1}^m w_i$, and $\xi^*_{\cdot j} = \sum_{i=1}^m \xi^*_{ij}$.

The EM algorithm can be easily implemented as follows. Let $(\pi_0^{(s)}, q^{(s)})$ be the current approximations to the MLE $(\hat{\pi}_0(k), \hat{q})$ with the log-likelihood function l given in (3). The next approximation $(\pi_0^{(s+1)}, q^{(s+1)})$ is given by the EM algorithm in two steps, the so-called E-step and M-step.

E-Step: Compute the conditional expectation of the log-likelihood function

$$Q(\pi_0, q) = E_{\pi_0^{(s)}, q^{(s)}}\{l^*(\pi_0, q)|\mathcal{P}\}$$

$$= E_{\pi_0^{(s)},q^{(s)}}(w_\cdot|\mathcal{P})\log(\pi_0) + (m - E_{\pi_0^{(s)},q^{(s)}}(w_\cdot|\mathcal{P}))\log(1-\pi_0)$$

$$+\sum_{j=1}^{k-1} E_{\pi_0^{(s)},q^{(s)}}(\xi^*_{\cdot j}|\mathcal{P}))\log(kq_j) + \sum_{i=1}^{m}\{E\{\xi^*_{ik}\log[k^2 q_{k-1}(1-p_i)]|\mathcal{P}\}.$$

Note

$$\begin{aligned} E_{\pi_0^{(s)},q^{(s)}}(w_i|\mathcal{P}) := \hat{w}_i &= P_{\pi_0^{(s)},q^{(s)}}(w_i = 1|\mathcal{P}) \\ &= \frac{\pi_0^{(s)}}{\pi_0^{(s)} + (1-\pi_0^{(s)})\{\prod_{j=1}^{k-1}(kq_j^{(s)})^{\xi_{ij}}\}\{k^2 q_{k-1}^{(s)}(1-p_i)\}^{\xi_{ik}}\}}. \end{aligned}$$

We have

$$\begin{aligned} Q(\pi_0, q) &= (\sum_{i=1}^{m}\hat{w}_i)\log\pi_0) + (m - \sum_{i=1}^{m}\hat{w}_i)\log(1-\pi_0) + \sum_{j=1}^{k-1}\left[\sum_{i=1}^{m}(1-\hat{w}_i)\xi_{ij}\right]\log(q_j k) \\ &+ \sum_{i=1}^{m}(1-\hat{w}_i)\xi_{ik}\log[k^2 q_{k-1}(1-p_i)]. \end{aligned}$$

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

Setting $\partial Q/\partial\pi_0 = 0$, we immediately have

$$\pi_0^{(s+1)} = \frac{\sum_{i=1}^{m}\hat{w}_i}{m}.$$

By Lemma 2 in Appendix,

$$\begin{aligned} q_j^{(s+1)} &= \frac{1}{\hat{m}}\sum_{i=1}^{m}(1-\hat{w}_i)\xi_{ij},\ j = 1, \cdots, k-2 \\ q_{k-1}^{(s+1)} &= \frac{2}{3\hat{m}}\sum_{i=1}^{m}(1-\hat{w}_i)(\xi_{i(k-1)} + \xi_{ik}) \\ q_k^{(s+1)} &= q_{k-1}^{(s+1)}/2, \end{aligned}$$

The formulas in the EM algorithm above are similar to to those developed by Oluyemi (2016) and Oluyemi and Chen (2016) where there are some algebraic errors.

where

$$\hat{m} = \sum_{j=1}^{k}\sum_{i=1}^{m}(1-\hat{w}_i)\xi_{ij} = m - \sum_{i=1}^{m}\hat{w}_i\ .$$

Notes

e) Selection of k via AIC

In this subsection, we discuss about selection of k. As a histogram-type approximation to $h(x)$, the value of k can reveal different features of the data. Since k is the number of categories of the histogram fitting h, a larger value of k does a better fitting job and hence is expected to result in a more accurate estimate of π_0. However, a larger value of k causes a greater standard error in estimating π_0 because there are more nuisance parameters to handle. As the estimate $\hat{\pi}_0(k)$ is based on the likelihood function, the Akaike information criterion (AIC) can be a nature choice of criteria for selection of an appropriate value of k. Let $\hat{l}_k$ be the maximum value of the log-likelihood function l defined in (3), i.e., $\hat{l}_k = l(\hat{\pi}_0(k), \hat{q})$. Noting that we have total of $k-1$ free parameters in computing $\hat{l}_k$, the AIC selection of k is to choose $\hat{k}$ such that $2\hat{l}_k - 2(k-1)$ is maximized, i.e.,

$$\hat{k} = \arg\max\{2\hat{l}_k - 2(k-1)\}.$$

The final estimate $\hat{\pi}_0$ for π_0 is $\hat{\pi}_0 = \hat{\pi}_0(\hat{k})$.

III. Simulation Study

Simulation studies are conducted with the p-values based on one-sided z-test in finite normal mixture models to evaluate the performance of the new estimator $\hat{\pi}_0$ and compare it with some existing method. Specifically, consider the finite normal model $\pi_0 N(0,1) + (1-\pi_0)N(1,1)$ with five griding values of π: 0.2, 0.35, 0.50, 0.65 and 0.8 . The p-value is computed by $p = 1 - \Phi(z)$. Four sample sizes, $m = 500, 1000, 1500$ and 2000 are consider. With each combination, 2000 Monte Care trials are used.

In the simulation studies, the EM algorithm is used to approximate the MLE's and a linear algorithm is used to search for the AIC selection $\hat{k}$.

Four existing estimators popularly cited in the literature are considered for purpose of comparison. They are Storey's estimator with bootstrap method $\hat{\pi}_0^s$, the convex density estimator $\hat{\pi}_0^c$ by Langaas, Lindqvist and Ferkingstad (2005), the averaging Storey estimator $\hat{\pi}_0^a$ by Jiang and Doerge (2008) and the histogram estimator $\hat{\pi}_0^h$ by Nettleton, Hwang, Caldo, and Wise (2006). The R package **cp4p** is used for computations of these estimates; all the computations are done with default settings of the R package.

Simulation results are reported in Table 1. The following findings from the simulation studies are immediate:

1. The new estimator $\hat{\pi}_0$ performs very well. A clear converging pattern is demonstrated in each simulation model, as m increases. The convergence with lower values of π_0 is faster than with higher values.
2. The new estimator performs substantially better than all other four existing estimators. It is noted that the Storey's $\hat{\pi}_0^s$ performs somewhat better than the other three existing estimators in general and its best performance is on the higher values of π_0 as expected. See Storey (2002).

IV. Real Data Analysis

In this section, we apply the new estimate method to the real life data from Hedenfalk et al. (2001)where 3226 genes were studied with $n_1 = 7$ BRECA1 arrays and $n_2 = 8$ BRCA2 arrays. The example is to test with each gene the null hypothesis that there is no differential gene expression between BRCA1-mutation-positive tumors and BRCA2-mutation-positive tumors by using a

Table 1: Empirical average of the estimates for the proportion π_0 with their empirical standard deviations in parentheses. The true model for the p-value of one-sided test is: $p = 1 - \Phi(x)$ with $x \sim \pi_0 N(0,1) + (1-\pi_0)N(1,1)$. Each of the entries is based on 2,000 Monte Carlo trials.

m	π_0	$\hat{\pi}_0$	$\hat{\pi}_0^s$	$\hat{\pi}_0^c$	$\hat{\pi}_0^a$	$\hat{\pi}_0^h$
500	0.20	0.194	0.339	0.324	0.357	0.378
		(0.101)	(0.059)	(0.061)	(0.068)	(0.593)
500	0.35	0.327	0.454	0.569	0.488	0.507
		(0.116)	(0.066)	(0.066)	(0.067)	(0.096)
500	0.50	0.457	0.586	0.583	0.618	0.644
		(0.130)	(0.067)	(0.065)	(0.066)	(0.086)
500	0.65	0.605	0.710	0.712	0.746	0.773
		(0.123)	(0.069)	(0.064)	(0.060)	(0.068)
500	0.80	0.749	0.826	0.834	0.864	0.885
		(0.113)	(0.066)	(0.059)	(0.052)	(0.051)
1000	0.20	0.202	0.328	0.308	0.322	0.314
		(0.082)	(0.044)	(0.047)	(0.057)	(0.085)
1000	0.35	0.336	0.455	0.445	0.462	0.468
		(0.095)	(0.049)	(0.051)	(0.060)	(0.086)
1000	0.50	0.480	0.583	0.578	0.599	0.615
		(0.093)	(0.052)	(0.050)	(0.055)	(0.074)
1000	0.65	0.624	0.701	0.707	0.729	0.750
		(0.089)	(0.052)	(0.049)	(0.052)	(0.065)
1000	0.80	0.777	0.829	0.834	0.886	0.875
		(0.084)	(0.050)	(0.044)	(0.041)	(0.042)
1500	0.20	0.206	0.292	0.291	0.298	0.294
		(0.072)	(0.042)	(0.040)	(0.051)	(0.071)
1500	0.35	0.348	0.423	0.433	0.440	0.449
		(0.079)	(0.049)	(0.044)	(0.056)	(0.074)
1500	0.50	0.480	0.553	0.569	0.581	0.579
		(0.081)	(0.053)	(0.043)	(0.054)	(0.070)
1500	0.65	0.636	0.684	0.703	0.721	0.741
		(0.075)	(0.055)	(0.043)	(0.048)	(0.055)
1500	0.80	0.788	0.801	0.829	0.848	0.870
		(0.068)	(0.050)	(0.040)	(0.039)	(0.042)
2000	0.20	0.206	0.295	0.290	0.292	0.284
		(0.066)	(0.038)	(0.035)	(0.045)	(0.062)
2000	0.35	0.348	0.425	0.434	0.435	0.436
		(0.071)	(0.042)	(0.037)	(0.050)	(0.067)
2000	0.50	0.496	0.556	0.571	0.579	0.585
		(0.065)	(0.047)	(0.038)	(0.051)	(0.065)
2000	0.65	0.636	0.683	0.700	0.713	0.731
		(0.071)	(0.050)	(0.038)	(0.047)	(0.056)
2000	0.80	0.798	0.812	0.831	0.847	0.865
		(0.056)	(0.047)	(0.035)	(0.035)	(0.038)

Notes

Notes

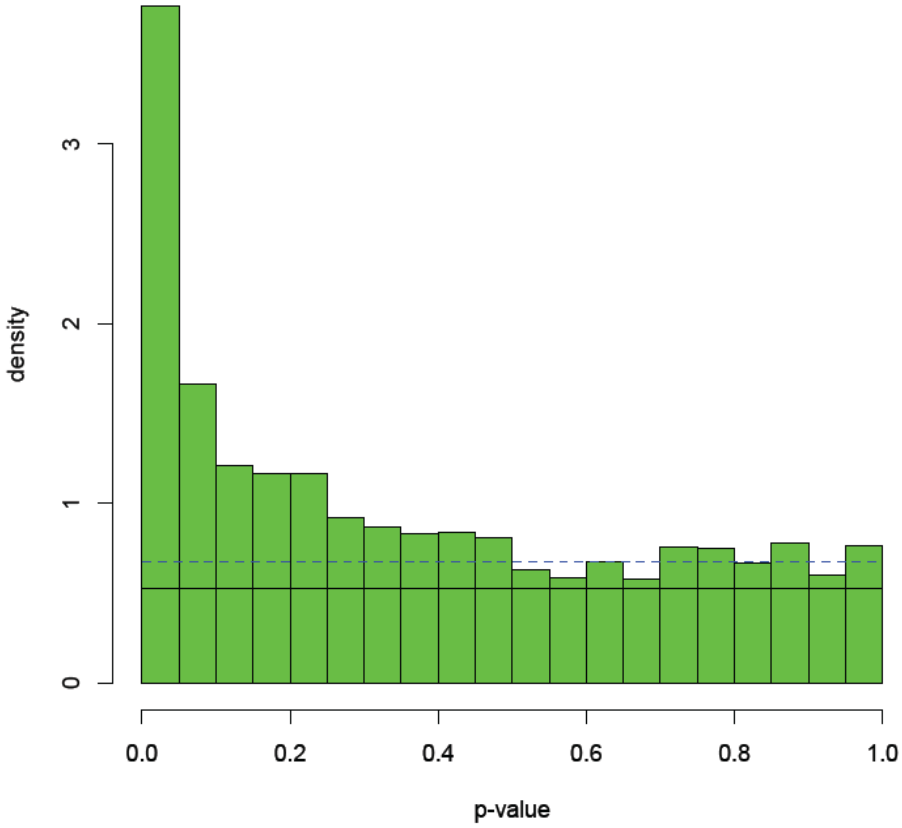

Figure 1: Histogram of the available 2752 p-values for detection of differently expressed genes with the data in Hedenfalk et al.(2001). The solid line has the intercept of $\hat{\pi}_0 = 0.523$ and the dash line has the intercept of $\hat{\pi}_0^h = 0.677$.

two-sample t-statistic. It was also analyzed by Storey and Tibshirani (2003). Following their instructions, we were able to download the p-values from *http://www.genomine.org/qvalue/results.txt*. Removing missing values left with m=2752 p-values.

The estimates for the proportion π_0 of true null hypotheses by the new estimator and the other four existing estimators are listed in Table 2. The AIC selection of k for the new estimator is $\hat{k} = 41$. In Figure 1 for the histogram of the 2752 available p-values with data in Hedenfalk et al. (2001), the solid horizontal line would be the expected bottom of the graph of the density function of the p-value under assumption $\inf h(x) = 0$ if the real value of π_0 were equal to the estimate $\hat{\pi}_0$ (0.523), whereas the dash line would indicate the expected bottom of the graph of the density function if the real value of π_0 were equal to the histogram estimate $\hat{\pi}_0^h = 0.677$. It is evident from Figure 1 that the existing estimators overestimates π_0.

Table 2: Estimates for the proportion π_0 of not differentially expressed genes with the break cancer data in Hedenfalk et al. (2001)

$\hat{\pi}_0$	$\hat{\pi}_0^s$	$\hat{\pi}_0^c$	$\hat{\pi}_0^a$	$\hat{\pi}_0^h$
0.523	0.689	0.682	0.704	0.677

References Références Referencias

1. Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* 64(3): 479-498.

2. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* 57(1): 289-300.
3. Simes, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 751-754.
4. Parker, R.A. & Rothenberg, RB. (1988) Identifying important results from multiple statistical tests. *Statistics in Medicine* 7: 1031-1043.
5. Tong, T., Feng, Z., Hilton, J.S. & Zhao, H. (2013) Estimating the proportion of true null hypotheses using the pattern of observed p-values. *J. Appl. Stat.* 40(9): 1949-1964.
6. Storey, J.D., Taylo, J.E. & Siegmund, D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates. *J. R. Stat. Soc. Ser. B* 66(1):187-205.
7. Nettleton, D., Hwang, J.T.G., Caldo, R.A. & Wise, R.P. (2006) Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics* 11(3): 337-356.
8. Zhao, H., Wu, X., Zhang, H. & Chen, H. (2012) 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* 41(9): 1580–1592.
9. Langaas, M. & Lindqvist, B. H. & Ferkingstad, E. (2005) Estimating the proportion of true hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. B* 67(4): 555-572.
10. Wu, B., Guan, Z. & a Zhao, H. (2006) Parametric and nonparametric FDR estimation revisited. *Biometrics* 62(3): 735-744.
11. Cheng, Y., Gao, D. & Tong, T. (2015) Bias and variance reduction in estimating the proportion of true-null hypotheses. *Biostatistics* 16(1): 189-204.
12. Mosig, M. O., Lipkin, E. Khutoreskaya, G. Tchourzyna, E., Soller, M. & and Friedmann, A. (2001) A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a Daughter design, using an adjusted false discovery rate criterion. *Genetics* 157: 1683-1698.
13. Jiang, H., & Doerge, R. (2008) Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Informatics* 6: 25-32.
14. Hedenfalk, I., Duggan, D. Chen, Y.D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, Borg, A., & Trent, I. (2001) Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine* 344(8): 539-548.
15. Oyeniran, O. (2016) Estimating the proportion of true null hypotheses in multiple testing problems. *Doctoral Dissertation, Bowling Green State University.*
16. Oyeniran, O. & Chen, H. (2016) Estimating the proportion of true null hypotheses in multiple testing problems. *Journal of Probability and Statistics*. Volume 2016, Article ID 3937056, http://dx.doi.org/10.1155/2016/3937056.
17. Storey, J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* 31(6): 2013-2035.
18. Storey, J.D. & Tibshirani, R. (2003) Statistical significance for genome wide studies. *Proc. Nat. Acad. Sci.s* 100(16): 9440-9445.

Notes

Appendix: Computation for the estimate of q

In M-step of the EM iteration algorithm, we compute the update for the estimate of q based on the following result.

Lemma 2 *Let* $g(q) = \sum_{j=1}^{k-1} x_j \log q_j$, *where* $x_j \geq 0$ *are constant and* $0 \leq q_j \leq 1$ *such that*

$$\sum_{j=1}^{k-2} q_j + (3/2)q_{k-1} = 1.$$

Then the maximum of g is attained at $\hat{q}_1 = x_1/\hat{m}, \cdots, \hat{q}_{k-2} = x_{k-2}/\hat{m}$, $\hat{q}_{k-1} = (2/3)x_{k-1}/\hat{m}$, *where*

$$\hat{m} = \sum_{j=1}^{k-1} x_j.$$

Proof. Let

$$G(q, \lambda) = g(q) + \lambda[1 - (\sum_{j=1}^{k-2} q_j + (3/2)q_{k-1})].$$

We have $\partial G/\partial\lambda = 1 - (\sum_{j=1}^{k-2} q_j + (3/2)q_{k-1})$ and

$$\frac{\partial g}{\partial q_j} = x_j/q_j - \lambda, j = 1, \cdots, k-2; \frac{\partial g}{\partial q_{k-1}} = x_{k-1}/q_{k-1} - (3/2)\lambda.$$

Setting all the partial derivatives to be zero yields

$$q_j\lambda = x_j, j = 1, \cdots, k-2; (3/2)q_{k-1}\lambda = x_{k-1}.$$

Adding all the equations above gives the solution of $\lambda = \hat{m}$ and so the solutions for q_j. The lemma is proved.

Notes