LJAST
Libyan Journal of Applied Science and Technology
مجلة ليبيا للعلوم التطبيقية والتقنية

Volume 10 Issue 02
December 2022

Adjusted likelihood approach to estimate the proportion of true null hypotheses in multiple tests

Adel Ewhida[1], Ali Alammari[2] and Iman Ihwil[3]

[1,2,3]Tripoli University, Faculty of Science, Department of Statistics

P.O. Box 13219, Tripoli, Libya

Abstract

A number of methods have been established to estimate the proportion true null hypotheses in multiple testing under the assumption of independency. On other hand, the test statistics are either discrete or continuous. In this paper, we will review an existing likelihood approach for estimating the proportion of true nulls π_0 in controlling the false discovery rates when the test statistics are continuous (see Hualing and Hanfeng, 2021). We therefore present an extension of these method that can successfully make some improvement of the performance. Simulation study demonstrates that the new estimator performs very well.

Key words: Multiple testing, likelihood approach, false discovery rates.

1. Introduction

Multiple hypotheses tests have played an important role for large scale big data. Multiplicity occurs when more tests are made simultaneously, the more test done, the more errors occur. There are several statistical technique have been developed to prevent this erroneous from happening. On other hand, the population distribution of the observed p-values $p_1, \dots, p_m$ can be described as a finite mixture with mixing proportion π_0 of the uniform distribution on $(0,1)$ and another non-uniform distribution with the pdf, say $h(x)$

$$f(p/\pi_0.h) = \pi_0 + (1-\pi_0)h(p), \quad 0 \le p \le 1.$$

The mixture model approach has been widely adopted and several other estimates have been proposed (see Langaas, Lindqvist and Ferkingstad, 2005; Wu, Guan and Zhao, 2006; Jiang and Doerge, 2008; Zhao et al., 2012; Cheng, Gao and Tong, 2015; Tong et al., 2013 and Oluyemi and Hanfeng, 2016). Nevertheless, when these estimates aim to improve some aspects of the Storey's estimator, the biasedness remains significant even with m as large as 2000, especially when π_0 is not close to 1 (see Storey, 2002). Motivated by the histogram approach (see Mosig et al., 2001 and Nettleton et al., 2006) a new estimator is proposed via a likelihood approach with h being approximated by a modified

histogram pdf, where Akaike information criterion is used to determine the number of categories in histogram construction (see Hualing and Hanfeng, 2021). In this paper, we present an extension of this method that can successfully make some improvement of the performance. By using the Shimazaki and Shinomoto method to select the number of categories in histogram construction (see Shimazaki and Shinomoto, 2007). Simulation study demonstrates that the new estimator performs very well.

2. Methods

2.1 Existing Likelihood Estimating Method (see Hualing and Hanfeng, 2021)

Let $p_1, p_2, \cdots, p_m$ be a random sample of size m from the pdf

$$f\,(p/\,\pi_0.h\rangle \;=\; \pi_0 \;+\; (1-\,\pi_0)h\,(p), \qquad 0 \le p \le 1.$$

The method proposes π_0 to be estimated by the Maximum Likelihood Estimating (MLE) when h is subject to a histogram type approximation. Motivated by the histogram approach (see Mosig et al., 2001 and Nettleton et al., 2006), a histogram approximation to the alternative pdf $h(p)$ is proposed as follows. Let $k > 2$ be an integer. Define

$$\hat{h}(p) = \begin{cases} kq_j & if \quad (j-1)/k \le p < j/k \quad 1 \le j \le k-1 \\ k^2 q_{k-1}(1-p) & if \qquad (k-1)/k \le p \le 1 \end{cases}$$

where $0 \le q_j \le 1$ with $q_1 + q_2 + \cdots + q_{k-1} + q_{k-1}/2 = 1, q_k = \frac{q_{k-1}}{2}$), where k pre-specified using Akaike information criterion (AIC) method (see Akaike, 1974). The AIC selection of k is to choose estimate to k such that $2\hat{l}_k - 2(k-1)$ is maximized, i.e.,

$$\hat{k} = argmax\{2\hat{l}_k - 2(k-1)\}$$

So, the final estimate for π_0 is $\hat{\pi}_0(\hat{k})$.

Then, the finite mixture distribution can be expressed as:

$$f\,(p/\,\pi_0.h\rangle \;=\; \pi_0 \;+\; (1-\,\pi_0)\{\textstyle\prod_{j=1}^{k}(kq_j)^{\omega_{ij}}\}\,\{k^2 q_{k-1}(1-p_i)\}^{\omega_{ik}}$$

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

$$\omega_{ij} = \begin{cases} 1 & if\ (j-1)/k \le p_i < j/k \\ 0 & otherwise \end{cases}$$

for $= 1,2,\dots,m$, $j = 1,2,\dots,k$. Note that for $1 \le i \le m$, $\sum_{j=1}^{k} \omega_{ij} = 1$, and $\sum_{i=1}^{m}\sum_{j=1}^{k} \omega_{ij} = m$.

The log-likelihood of the parameter π_0 of interest and the new nuisance parameter q becomes

$$l(\pi_0, q) = \sum_{i=1}^{m} \log\{\pi_0 + (1-\pi_0)\left\{\prod_{j=1}^{k-1}(kq_j)^{\omega_{ij}}\right\}\{k^2 q_{k-1}(1-p_i)\}^{\omega_{ik}}$$

So, maximizing the nonlinear log-likelihood function can be complicating. However, the Expectation-Maximization algorithm (EM algorithm) can be used to obtain an approximation to the MLE, $\hat{\pi}_0(k)$, easily. To do that, they introduce a latent Bernoulli variable z_i be a binary random variable with $z_i = 1$ if p belongs to the first mixture component U (0, 1) if and only if the null hypothesis is true, and $z_i = 0$ if p belong to the second mixture component $h(q1,\dots,qk)$, for $i = 1,\dots,m$ when the exact number of observations within each mixture component is fixed, Then the complete data likelihood function of (π_0, q) is:

$$\prod_{i=1}^{m} f(p_i|\pi_0)^{z_i} f(p_i|\pi_0, q_1 \dots q_k)^{1-z_i}$$

$$= \prod_{i=1}^{m} \pi_0{}^{z_i}\{(1-\pi_0)\left\{\prod_{j=1}^{k-1}(kq_j)^{\omega_{ij}}\right\}\{k^2 q_{k-1}(1-p_i)\}^{\omega_{ik}}\}^{1-z_i}$$

and the log likelihood function for complete data is:

$$l^*(\pi_0, q) = z_. \log \pi_0 + (m - z_.)\log(1-\pi_0) + \sum_{j=1}^{k-1}(\omega_{.j}{}^*)\log(kq_j) + \sum_{i=1}^{m}\omega_{ik}{}^*\log[k^2 q_{k-1}(1-p_i)]$$

where $\omega_{ij}^* = (1-z_i)\omega_{ij}$ $z_. = \sum_{i=1}^{m} z_i$, and $\omega_{.j} = \sum_{i=1}^{m} \omega^*{}_{ij}$

Using the current iterate of parameters $\theta^t = \pi_0^t, q_1^t, \dots, q_k^t$ at iteration, the next approximation $(\theta^{t+1} = \pi_0^{t+1}, q_1^{t+1}, \dots, q_k^{t+1})$ is given by the EM algorithm in two steps:

E-Step: conditional expectation of z_i given p

$$Q(\pi_0, q) = E(l^*(\pi_0, q)|p; \theta^t)$$

$$= E_{\theta^t}(z_.|P)\log \pi_0 + \left(m - E_{\theta^t}(z_.|P)\right)\log(1-\pi_0) + \sum_{j=1}^{k-1}\left(E_{\theta^t}(\omega_{.j}{}^*|P)\right)\log kq_j$$

$$+\sum_{i=1}^{m}(E_{\theta^t}(\omega_{ik}{}^{*}|P))\log(k^2 q_{k-1}(1-p_i))$$

M-step: In the M-step, $Q(\pi_0, q)$ is maximized to yield the next approximation $\theta^{t+1} = \pi_0^{t+1}, q_1^{t+1}, \ldots, q_k^{t+1}$

Setting $dQ/d\pi_0 = 0$, we have

$$\pi_0^{t+1} = \frac{E_{\theta^t}(z_.|P)}{m},$$

where

$$E_{\theta^t}(z_.|P) = \sum_{i=1}^{m} E_{\pi_0^t, q_j^t}(z_i|P) = E_{\pi_0^t, q_j^t}(z_i|P) = p_{\pi_0^t, q_j^t}(z_i = 1|P) = \widehat{z_i}$$

$$= \frac{\pi_0^t}{\pi_0^t + (1-\pi_0^t)\left\{\prod_{j=1}^{k-1}(k q_j{}^t)^{\omega_{ij}}\right\}\{k^2 q_{k-1}{}^t(1-p_i)\}^{\omega_{ik}}}$$

and second to approximate q_j, by $dE(z_i|p;\theta^t)/dq_j = 0$ we have

$$q_j{}^{t+1} = \frac{\omega_{.j}}{m - E_{\pi_0^t, q_j^t}(z_.|P)} = \frac{\omega_{.j}}{m(1-\pi_0^{t+1})}$$

To sum up, let $\pi_0^t, q_1^t, \ldots, q_k^t$ be the t^{th}approximations to the maximum likelihood, and the $\pi_0^{t+1}, q_1^{t+1}, \ldots, q_k^{t+1}$ is approximation with EM algorithm is given by

$$\pi_0^{t+1} = \frac{\sum_{i=1}^{m} \frac{\pi_0^t}{\pi_0^t + (1-\pi_0^t)\left\{\prod_{j=1}^{k-1}(k q_j{}^t)^{\omega_{ij}}\right\}\{k^2 q_{k-1}{}^t(1-p_i)\}^{\omega_{ik}}}}{m}$$

$$q_j{}^{t+1} = \frac{\sum_{i=1}^{m}(1-\widehat{z_i})\omega_{ij}}{\widehat{m}}$$

$$q_{k-1}{}^{t+1} = \frac{2\sum_{i=1}^{m}(1-\widehat{z_i})(\omega_{i(k-1)} + \omega_{i(k)})}{3\widehat{m}}$$

$$q_k{}^{t+1} = q_{k-1}{}^{t+1}/2$$

where

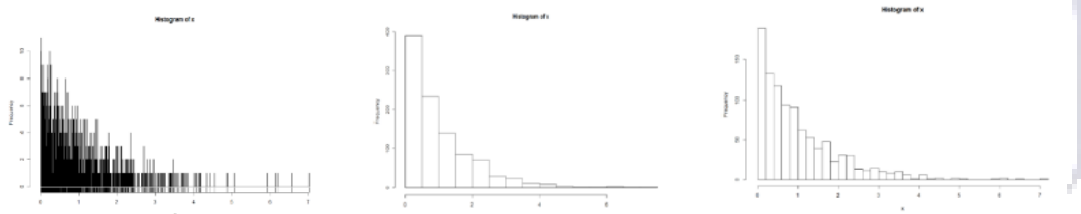
$$\hat{m} = \sum_{j=1}^{k}\sum_{i=1}^{m}(1-\hat{z}_i)\omega_{ij} = m - \sum_{i=1}^{m}\hat{z}_i$$

repeat until

$|Q(\theta^{t+1}) - Q(\theta^t)| < \varepsilon; \quad t = 0,1,2,\ldots$

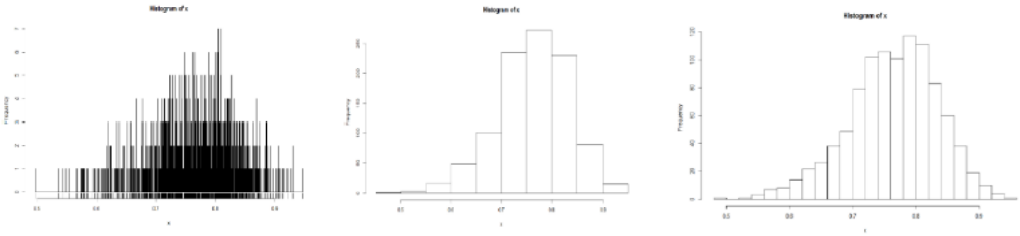
2.2 Shimazaki and Shinomoto Method

Shimazaki S and Shinomoto H, 2007 proposed a method (SH) to estimate how the optimal size of bins decreases when more experimental trials are added to the data. The method provides the cost function for number of sequences. However, we have applied this method and compared it with Surge's role and Freedman method for right skewed exponential distribution, left skewed beta distribution and normal distributions as shown in Figures 1, 2 and 3. By using R language to simulate data with a sample size of 10000. The Figures explained how the performance of SH method compared with the other two methods and how it was very well.



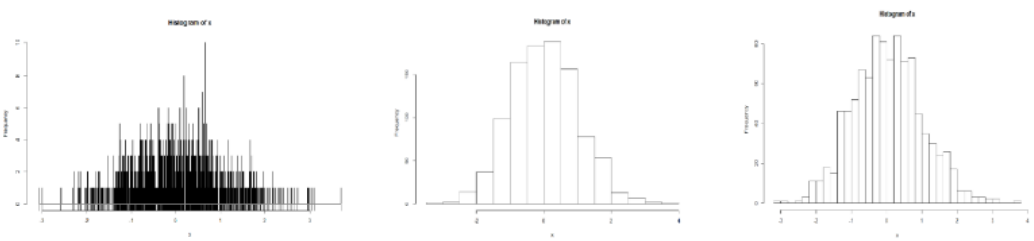
SH method Sturge's method Freedman method

Figure 1: Displaying the performance of three methods for right skewed exponential distribution .



SH method Sturge's method Freedman method

Figure 2: Displaying the performance three methods for left skewed beta distribution



SH method Sturge's method Freedman method

Figure 3: Displaying the performance of the three methods in normal distribution

The Algorithm of Shimazaki and Shinomoto method (Shimazaki and Shinomoto, 2007)

1. First, we should start with determine the sample size of the Histogram distribution which is n. we used R language to cluster the variables.
2. Divide the data range into B bins width λ, and count the number of events L_i that enter to the *i*-th bin.
3. Calculate the mean and the variance of the number of events L as:

$$L = \sum_{i=1}^{B} \mathrm{Li}$$

and

$$\mathrm{V} = 1/B \sum_{i=1}^{B} (\mathrm{Li} - \mathrm{L})^2$$

4. Compute a formula (cost function)

$$C(\lambda) = \left(\frac{2L - V}{\lambda^2}\right)$$

5. Repeat until change λ , then find λ^* that minimize C (λ)

2.3 Adjusted likelihood Method

We have applied the SH method in the likelihood approach. Therefore, the EM iteration process geos as shown below: inputs $\omega_{ij}, i = 1, \ldots, M\ , j = 1, \ldots, k$ transformed from the observed p-value, where k was the estimated using SH method. The output is the estimate of $\theta = (\pi_0, q_1 \ldots q_k)$

- *begin*
- *initialization: set* $\pi_0 = \pi_0^{(0)}, q_j = q_j^{(0)}$
- *repeat*
- *set:* $\pi_0 = \pi_0^{(t)}, q_j = q_j^{(t)}$ *the current approximation*
- *Compute*

$$\pi_0^{t+1} = \frac{\sum_{i=1}^{m} \frac{\pi_0^t}{\pi_0^t + (1-\pi_0^t)\{\prod_{j=1}^{k-1}(kq_j{}^t)^{\omega_{ij}}\}\{k^2 q_{k-1}{}^t(1-p_i)\}^{\omega_{ik}}}}{m}$$

$$q_j{}^{t+1} = \frac{\sum_{i=1}^{m}(1-\hat{z}_i)\omega_{ij}}{\hat{m}}$$

$$q_{k-1}{}^{t+1} = \frac{2\sum_{i=1}^{m}(1-\hat{z}_i)(\omega_{i(k-1)} + \omega_{i(k)})}{3\hat{m}}$$

$$q_k{}^{t+1} = q_{k-1}{}^{t+1}/2$$

$$\hat{m} = m - \sum_{i=1}^{m} \hat{z}_i$$

- *Until* $|Q(\theta^{t+1}) - Q(\theta^t)| < \varepsilon; \quad t = 0,1,2,\ldots$

  then $\hat{\pi}_0 = \pi_0^{t+1}$

3 Simulation study

To investigate the properties and performance of these methods, a simulation study was performed taking three different values of sample size m, with variety of replicate of each case (Independent, weak dependent, moderate dependent and strong dependent cases), with different true values π_0. The generation were drawn from multivariate normal distribution where μ is the mean vector of the sample and Σ is the covariance matrix. The generation of simulated data, calculation of the estimators, the calculation of the estimated average and the

empirical standard deviation were all done in R language version 3.6.2 and the online source package cp4p.

3.1 Independent case

In this case, we used randomly generated independent data with no dependence structure within or between the genes from mixture normal distribution. Each p-values were computed by the cumulative distribution function of standard normal, with true values of $\pi_0 = 0.25$, 0.50, 0.75 and 0.90. The leading diagonal of covariance matrix Σ contained all 1's. This procedure was replicated 100 times, for sample sizes (200, 500 and 1500 genes). The result of these two methods performances is shown in table 1 below.

Table 1: Empirical average of the estimates for the proportion π_0 with their standard deviations in Parentheses in independent data. Each of the entries is based on 100 replicates. Denote $\hat{\pi}_0$ for the existing likelihood method estimator and $\hat{\pi}_0^{adj}$ for the new Adjusted likelihood estimator.

m	π_0	$\hat{\pi}_0$	$\hat{\pi}_0^{adj}$
200	0.25	0.283	0.294
		(0.2954)	(0.1213)
	0.50	0.568	0.572
		(0.2031)	(0.1493)
	0.75	0.801	0.808
		(0.0781)	(0.1552)
	0.90	0.928	0.936
		(0.0883)	(0.1526)
500	0.25	0.314	0.371
		(0.2734)	(0.1321)
	0.50	0.556	0.562
		(0.1892)	(0.1449)
	0.75	0.775	0.783
		(0.0721)	(0.1499)
	0.90	0.915	0.918
		(0.0775)	(0.1479)
1500	0.25	0.307	0.333
		(0.2486)	(0.1330)
	0.50	0.553	0.549
		(0.1700)	(0.1334)
	0.75	0.781	0.796
		(0.0700)	(0.1352)
	0.90	0.913	0.917
		(0.0721)	(0.1315)

3.2 Weak dependent

In this case, each p-values were computed by the cumulative distribution function of standard normal, with true values of $\pi_0 =$ 0.15, 0.55 and 0.95. The leading diagonal of covariance matrix Σ contained all 1's and the ρ's in the Σ matrix were all 0.15. This procedure was replicated 30 times, for sample sizes (200, 500 and 1500 genes). The result of these two methods performances is shown in table 2 below.

Table 2: Empirical average of the estimates for the proportion π_0 with their standard deviations in Parentheses in weak dependent data. Each of the entries is based on 30 replicates. Denote $\hat{\pi}_0$ for the existing likelihood method estimator and $\hat{\pi}_0^{adj}$ for the new Adjusted likelihood estimator.

m	π_0	$\hat{\pi}_0$	$\hat{\pi}_0^{adj}$
200	0.15	0.310	0.362
		(0.1790)	(0.2219)
	0.55	0.693	0.711
		(0.1725)	(0.2191)
	0.95	0.975	0.982
		(0.1685)	(0.1731)
500	0.15	0.307	0.332
		(0.1541)	(0.2016)
	0.55	0.686	0.691
		(0.1665)	(0.1823)
	0.95	0.944	0.931
		(0.1433)	(0.1574)
1500	0.15	0.262	0.308
		(0.1288)	(0.1978)
	0.55	0.661	0.672
		(0.1249)	(0.1580)
	0.95	0.957	0.959
		(0.0)	(0.1219)

3.3 Moderate dependent

In this case, each p-values were computed by the cumulative distribution function of standard normal, with true values of π_0= 0.20, 0.60 and 0.80. The leading diagonal of covariance matrix Σ contained all 1's and the ρ's in the Σ matrix were all 0.55. This procedure was replicated 15 times, for sample sizes (200, 500 and 1500 genes). The result of these two methods performances is shown in table 3 below.

Table 3: Empirical average of the estimates for the proportion π_0 with their standard deviations in Parentheses in moderate independent data. Each of the entries is based on 15 replicates. Denote $\hat{\pi}_0$ for the existing likelihood method estimator and $\hat{\pi}_0^{adj}$ for the new Adjusted likelihood estimator.

m	π_0	$\hat{\pi}_0$	$\hat{\pi}_0^{adj}$
200	0.20	0.378	0.396
		(0.1561)	(0.1972)
	0.60	0.665	0.676
		(0.1137)	(0.2025)
	0.80	0.932	0.931
		(0.1645)	(0.2213)
500	0.20	0.290	0.301
		(0.1298)	(0.1648)
	0.60	0.640	0.647
		(0.0947)	(0.1954)
	0.80	0.915	0.912
		(0.1211)	(0.1934)
1500	0.20	0.254	0.271
		(0.0874)	(0.1364)
	0.60	0.612	0.619
		(0.0547)	(0.1454)
	0.80	0.867	0.862
		(0.0737)	(0.1348)

3.4 Strong dependent

In this case, each p-values were computed by the cumulative distribution function of standard normal, with true values of π_0= 0.30, 0.50 and 0.70. The leading diagonal of covariance matrix Σ contained all 1's and the ρ's in the Σ matrix were all 0.90. This procedure was replicated 25 times, for sample sizes (750 and 1300 genes). The result of these two methods performances is shown in table 4 below.

Table 4: Empirical average of the estimates for the proportion π_0 with their standard deviations in Parentheses in strong dependent data. Each of the entries is based on 25 replicates. Denote $\hat{\pi}_0$ for the existing likelihood method estimator and $\hat{\pi}_0^{adj}$ for the new Adjusted likelihood estimator.

M	π_0	$\hat{\pi}_0$	$\hat{\pi}_0^{adj}$
750	0.30	0.401	0.469
		(0.1464)	(0.2363)
	0.50	0.614	0.583
		(0.1378)	(0.2036)
	0.70	0.771	0.790
		(0.1429)	(0.1973)
1300	0.30	0.347	0.384
		(0.0822)	(0.1093)
	0.50	0.582	0.557

		(0.0887)	(0.1632)
	0.70	0.721	0.739
		(0.0911)	(0.1389)

4 Conclusion

For all simulated types of data, it is clearly shown that the new adjusted likelihood estimate outperformed substantially over the existing method with comparable standard errors. We also observed that, the estimators approached the true values of π_0 as the sample size increased. In general, the higher of true values of π_0, the estimators more effective (see tables 1, 2, 3 and 4).

References

1. Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control; 19(6): 716 – 723.
2. Cheng Y, Gao D, Tong T. 2015. Boas and variance reduction in estimating the proportion of true null hypotheses. Biostatistics; 16: 189-204.
3. Hualing Z, Hanfeng C. 2021. Estimating the Proportion of True Null Hypotheses: a Likelihood Approach. Global Journal of Science Frontier Research: F Mathematics and Decision Sciences; 21(5): 2249-4626
4. Jiang H, Doerge R. 2008. Estimating the proportion of true null hypotheses for multiple comparisons. Cancer Informatics; 6: 26-32.
5. Langaas M, Lindqvist BH, Ferkingstad E. 2005. Estimating the proportion of true null hypotheses with application to DNA microarray data. J. R. Stat. Soc B; 67: 555-572.
6. Oluyemi O, Hanfeng C. 2016. Estimating the proportion of true null hypotheses in multiple testing problems. Journal of Probability statistics, Article ID 3937056.
7. Mosig MO, Lipkin E, Galina K, et al. 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.

8. Nettleton D, Hwang JTG, Galdo RA, Wise RP. 2006. Estimating the number of true null hypotheses from a histogram of p-values. Journal of Agricultural, Biological, and Environmental Statistics; 11: 337-356.
9. Shimazaki H, Shinomoto S. 2007. A Method for Selecting the Bin Size of a Time Histogram. Neural computation; 19(6):1503-27.
10. Storey JD, 2002. A direct approach to false discovery rates. J. R. Stat. Soc B; 64: 479-498.
11. Tong T, Feng Z, Hilton JS, Zhao H. 2013. Estimating the proportion of true null hypotheses using the pattern of observed p-value. J. Appl. Stat.; 40: 1949-1964.
12. Wu B, Guan Z, Zhao H. 2006. Parametric and nonparametric FDR estimation revisited. Biometrics; 62: 735- 744.
13. 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. Communication in statistics-simulation and Computation; 41: 1580- 1592.