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

Volume 13 Issue 01
June 2025
ISSN 2958-6119

Resampling adjusted likelihood and average estimate approaches for microarray data analysis

Adel Ewhida[1] and Iman Ihwil[1]
[1]Tripoli University, Faculty of Science, Department of Statistics,
P.O. Box 13219, Tripoli, Libya
Running title: Estimating the proportion of true null hypotheses.

Abstract

The burgeoning field of genomics has reinvigorated interest in multiple testing methods as it introduces new methodological and computational challenges. Whole genome microarray studies (e.g., differential expression, differential methylation, ChIP-chip) offer the possibility to test millions of traits in one genome. This article discusses the preferment of Adjusted likelihood approach and Average estimate approach (see Ewhida, Alammari and Ihwil, 2022; Jiang and Doerge, 2008) for estimating the proportion of true nulls π_0 for CHIP-on-chip experiment data. Which has been done by Elnfati, Iles and Miller 2016, to express where the protein and DNA are bound, and the sample was taken from Drosophila fly (Fruit flies) in three replicates to determine the effect of HISTONE protein on fetus growth. The study demonstrates that, the Adjusted likelihood approach estimator performs very well.

Key words: Multiple testing, likelihood approach and Average estimate approach.

1. Introduction

Genomic technologies generate vast amounts of biological data that form the basis for studies that require repeated testing of the same hypothesis. Because the number of tests performed is so large, the multiple comparison procedures that control the familywise error rate are sometimes too stringent for biological applications (Jiang and Doerge, 2008). In fact, the main aims are to present some modern methods of estimating the proportion of true hypotheses where it plays a main role on false discovery rate (FDR) control, denoted by π_0 (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). In this paper, we carried out a simulation and real data experiment to compute estimating of the

proportion of true nulls with independent structures using the adjusted likelihood and average estimate approaches (see Ewhida, Alammari and Ihwil, 2022; Jiang and Doerge, 2008).

2. Methods

2.1 Adjusted likelihood approach (see Ewhida, Alammari and Ihwil, 2022)

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

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

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

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

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 Shimazaki and Shinomoto method (see Shimazaki and Shinomoto, 2007). The Algorithm of Shimazaki and Shinomoto method is,

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

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

and

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

- Compute a formula (cost function)

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

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

Then we have applied the SH method in the likelihood approach (see Hualing and Hanfeng, 2021).

So, the finite mixture distribution can be expressed as:

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

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

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

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:

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

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

and the log likelihood function for complete data is:

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

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

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

E-Step: conditional expectation of z_i given p

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Average estimate approach (see Jiang and Doerge, 2008)

This approach is motivating the work of Storey, 2002, where the $\hat{\pi}_0$ estimated by

$$\hat{\pi}(\lambda) = \frac{W(\lambda)}{(1-\lambda)m} \qquad (1)$$

where $W(\lambda) = \#\{p_i; p_i > \lambda\}$ and $0 \le \lambda < 1$ is a tuning parameter. However, this estimator has large bias and small variance when λ is small and a small bias and large variance when λ is big. Therefore, Jiang and Doerge, 2008 proposed an estimate of $\hat{\pi}_0(\lambda)$ as the average of $\hat{\pi}_0(\lambda)$ over the values of λ_i

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

where the approach aimed to balance the bias and variance. Define $0 = t_1 < t_2 < \cdots < t_B < t_{B+1} = 1$ is equally spaced points in the interval [0, 1], where divided into B small intervals with equal length 1/B. Specifically, $t_i = (i-1)/B$. For each t_i, $\hat{\pi}_0(t_i)$ is an estimate of π_0 by equation (1) with $\lambda = t_i$. Let $Nb_i = \#\{p_k; p_k \ge t_i\}$ and $Ns_i = \#\{p_k; t_i \le p_k < t_{i+1}\}$ for all $i = 1,\ldots,B$. If the Nb_i p-values come

from the null distribution, from this can be estimated by

$$\hat{\pi}_0(B) = \frac{1}{B-i+1}\sum_{j=1}^{j=B}\hat{\pi}_0(t_j) = \frac{1}{B-i+1}\sum_{j=1}^{j=B}\frac{Nb_i}{(i-t_j)m},$$

where $i = \min\{i;\ Ns_i \leq \frac{Nb_i}{B-i+1}\}$.

3. Simulation study

To investigate the properties and performance of these methods, 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 $\pi_0 = 0.50$, 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 $\hat{\pi}_0^{aveg}$ for the average method estimator and $\hat{\pi}_0^{adj}$ for the new Adjusted likelihood estimator.

m	π_0	$\hat{\pi}_0^{aveg}$	$\hat{\pi}_0^{adj}$
200	0.50	0.577 (0.2051)	0.572 (0.1493)
	0.75	0.868 (0.0768)	0.808 (0.1552)
	0.90	0.932 (0.0775)	0.936 (0.1526)
500	0.50	0.728 (0.1843)	0.562 (0.1449)
	0.75	0.801 (0.0707)	0.783 (0.1499)
	0.90	0.912 (0.0710)	0.918 (0.1479)
1500	0.50	0,565 (0.1530)	0.549 (0.1334)
	0.75	0.803 (0.0623)	0.796 (0.1352)
	0.90	0.901 (0.0556)	0.917 (0.1315)

4. Microarray data application

This CHIP-on-chip experiment has been done by Elnfati, Iles and Miller 2016, to express where the protein and DNA are bound, and the sample was taken from Drosophila fly (Fruit flies) in three

replicates to determine the effect of HISTONE protein on fetus growth, and the result was good. However, the adjusted likelihood approach gives a much lower estimate of $\pi_0$than the average estimate approach as shown in Table 1. From this real data analysis, adjusted likelihood approach provides a slightly larger estimate than average approach as shown in Table 2.

Table 2: The estimate of the proportion of true null hypotheses $\pi_0$using two methods: the adjusted likelihood approach and average approach with B chosen via the bootstrapping procedure (Bboot) applied to drosophila fly data.

Adjusted Likelihood approach	Average estimate approach
0.19642	0.11

5. Conclusion

In this work, we have used two methods for estimating the proportion of true null hypotheses (π_0). The adjusted likelihood method gives a much lower estimate of $\pi_0$than the average estimate approach.

References

1. Cheng Y, Gao D, Tong T. 2015. Boas and variance reduction in estimating the proportion of true null hypotheses. Biostatistics; 16: 189-204.
2. Elnfati AH, Iles D, Miller D. 2016. Nucleosomal chromatin in the mature sperm of Drosophila melanogaster. Genomics Data; 7: 175–177
3. Ewhida A, Alammari A and Ihwil E. 2022. Adjusted likelihood approach to estimate the proportion of true null hypotheses in multiple tests.
4. 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
5. Jiang H, Doerge R. 2008. Estimating the proportion of true null hypotheses for multiple comparisons. Cancer Informatics; 6: 26-32.
6. 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.

7. Oluyemi O, Hanfeng C.2016. Estimating the proportion of true null hypotheses in multiple testing problems. Journal of Probability statistics, Article ID 3937056.
8. 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 Poolingin a Daughter Design, Using an Adjusted False Discovery Rate Criterion. Genetics; 157: 1683 -1698.
9. Nettleton D, Hwang JTG, Caldo RA, Wise RP. 2006. Estimating the Number of true null hypotheses from a Histogram of p-values. Biological, and Envirinmental statistics: 11: 337 - 356.
10. 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.
11. Shimazaki H, Shinomoto S. 2007. A Method for Selecting the Bin Size of a Time Histogram. Neural Computation 19(6):1503-27.
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.