Gene expression

A simple procedure for estimating the false discovery rate

Cyril Dalmasso, Philippe Broët* and Thierry Moreau

INSERM U472, Faculté de Médecine Paris-Sud, 16 Avenue Paul Vaillant-Couturier, 94807 Villejuif Cedex, France

Received on April 7, 2004; revised on June 23, 2004; accepted on July 24, 2004

Advance Access publication October 12, 2004

ABSTRACT

Motivation: The most used criterion in microarray data analysis is nowadays the false discovery rate (FDR). In the framework of estimating procedures based on the marginal distribution of the \( P \)-values without any assumption on gene expression changes, estimators of the FDR are necessarily conservatively biased. Indeed, only an upper bound estimate can be obtained for the key quantity \( \pi_0 \), which is the probability for a gene to be unmodified. In this paper, we propose a novel family of estimators for \( \pi_0 \) that allows the calculation of FDR.

Results: The very simple method for estimating \( \pi_0 \) called LBE (Locational Based Estimator) is presented together with results on its variability. Simulation results indicate that the proposed estimator performs well in finite sample and has the best mean square error in most of the cases as compared with the procedures QVALUE, BUM and SPLOSH. The different procedures are then applied to real datasets.

Availability: The R function LBE is available at http://ifr69.vjf.insERM.fr/ibe

Contact: broet@vjf.insERM.fr

1 INTRODUCTION

New transcriptome-oriented biotechnologies make nowadays possible the comparative analysis of thousands of genes expression in parallel for selecting relevant genes the transcriptional changes of which are related to a clinical or biological outcome (Schena, 2000). In such a case, a major multiple testing problem arises due to the fact that a large number of statistical tests are performed simultaneously (Hochberg and Tamhane, 1987). Until now, statistical procedures devoted to this multiple testing problem mostly focused on controlling or estimating false positive error criteria.

For cDNA microarray experiments, the most used criterion nowadays is the false discovery rate (FDR) introduced by Benjamini and Hochberg (1995). The FDR is the expected proportion of false discoveries among all discoveries. Noting \( V \) the random variable representing the number of false discoveries and \( R \) the number of significant results obtained from a particular multiple testing procedure, Benjamini and Hochberg defined the FDR by \( \text{FDR} = E(V/R) \) if \( R > 0 \), and 0 otherwise. In large-scale hypotheses generating studies such as microarray experiments, the FDR seems more relevant than the Family Wise Error Rate (FWER) defined by the probability of committing at least one false discovery (Hochberg and Tamhane, 1987). In this setting, the purpose of this paper is to propose a novel procedure for estimating the FDR.

*To whom correspondence should be addressed.

In their seminal paper, Benjamini and Hochberg (1995) presented a step up method in order to control the FDR and discussed another criterion, later called the positive FDR (pFDR) by Storey (2001). This criterion is defined as \( \text{pFDR} = E(V/R)|R > 0 \). However, Benjamini and Hochberg did not consider this criterion due to the fact that it cannot be controlled since under the complete null hypothesis (all null hypotheses tested are true), all significant results (if there are significant ones) are necessary false discoveries. Then, \( \text{pFDR} = 1 \) and it is impossible to insure that \( \text{pFDR} < \alpha \) for a given \( \alpha \neq 1 \).

Storey (2001) demonstrated that if the test statistics are independent and identically distributed, for a fixed rejection region \( \Gamma \), which is the same for every test,

\[
\text{pFDR}(\Gamma) = \Pr(H = 0|T \in \Gamma) = \frac{\pi_0 \Pr(T \in \Gamma|H = 0)}{\Pr(T \in \Gamma)},
\]

where \( H \) is the variable such as \( H = 0 \) if the null hypothesis \( H_0 \) is true, \( H = 1 \) if the alternative hypothesis \( H_1 \) is true, \( \pi_0 = \Pr(H = 0) \) is the probability of not being modified and \( T \) is the test statistic used for all tested hypotheses.

From its definition, the pFDR is obviously related to the FDR through \( \text{pFDR} = \text{FDR}/\Pr(R > 0) \). Since \( \Pr(R > 0) \) tends to one when the number of tested hypotheses tends to infinity, these two criteria are asymptotically equivalent and, in the following, we will note FDR for both of them.

Storey and Tibshirani (2003) proposed a method (implemented in R function QVALUE) for obtaining a conservatively biased estimator for the pFDR based on the marginal distribution of the \( P \)-values without making any assumption on the distribution related to the modified genes. In practice, from (1), estimating the FDR is based on the separate estimation of the following three terms \( \Pr(T \in \Gamma) \), \( \Pr(T \in \Gamma|H = 0) \) and \( \pi_0 \) where only an upper bound estimator of the latter quantity can be obtained.

Relying on the same framework, two procedures named BUM (Pounds and Morris, 2003) and SPLOSH (Pounds and Cheng, 2004) have been recently proposed. In practice, all these three methods are based on the marginal distribution of the \( P \)-values and provide a conservatively biased estimator for the FDR resulting from the overestimation of \( \pi_0 \).

In this paper, we provide a class of estimators for an upper bound of \( \pi_0 \) based on the expectation of the transformed \( P \)-values and from which we can obtain results on the asymptotic distribution. As for QVALUE, BUM and SPLOSH, our procedure do not make any assumption on the distribution related to modified genes. From our
proposed estimators, we can easily obtain estimators of the FDR or other quantities such as the \(q\)-values (Storey, 2003).

The paper is organized as follows: in Section 2, we present the general framework of the procedures QVALUE, BUM and SPLISH for obtaining a conservatively biased estimator for \(\pi_0\) based on the marginal distribution of the \(P\)-values. In Section 3, we present a general class of estimators for an upper bound of \(\pi_0\) with results on its asymptotic distribution. In Section 4, we propose a particular family of estimators and give guidelines for choosing one estimator in the family depending on the experimental setup and the accuracy needed. In Section 5, we present results from a simulation study that compares proposed estimators to those provided by QVALUE, BUM and SPLISH. In Section 6, we apply the different methods on real datasets and we conclude with a discussion.

2 GENERAL FRAMEWORK FOR PROCEDURES BASED ON THE MARGINAL DISTRIBUTION OF THE \(P\)-VALUES

Data can be modeled following a two components mixture model (McLachlan and Peel, 2000) whereby the population of genes can be considered as composed of two subpopulations of genes, those for which the null hypothesis is true (unmodified genes), and those for which the alternative hypothesis is true (modified genes). Let \(p_i, i = 1, \ldots, m\) be the \(P\)-values calculated for the \(m\) tested hypotheses. Let \(P\) be the random variable for which the \(P\)-values are the observations and let \(f\) be the marginal probability density function (pdf) of \(P\). Denote \(f_0\) the conditional pdf of \(P\) under the null hypothesis and \(f_1\) the conditional pdf of \(P\) under the alternative hypothesis. Then:

\[
f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p).
\] (2)

Under the null hypothesis (and if the assumption for the distribution of the test statistic under the null hypothesis is true) the \(P\)-values are uniformly distributed on \([0, 1]\) so that \(f_0(p) = 1_{[0,1]}(p)\) and the relation (2) is: \(f(p) = \pi_0 + (1 - \pi_0) f_1(p)\) where the conditional density \(f_1\) is unknown. Since \((1 - \pi_0) f_1(p)\) is non-negative and assuming that \(f\) (or \(f_1\)) is non-increasing for \(p \in [0, 1]\), then \(f(1)\) is the smallest upper bound for \(\pi_0\) based on (2). Thus, an unbiased estimator of \(f(1)\) provides a conservatively biased estimator of \(\pi_0\).

As seen below, the procedures QVALUE, BUM and SPLISH are based on this latter estimator whereas our procedure is based on the expectation of transformed \(P\)-values.

A widely used estimator for \(\pi_0\) is the one proposed by Storey and Tibshirani (2003). Using a tuning parameter \(\lambda \in [0, 1]\), \(\pi_0\) is estimated by:

\[
\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, m\}}{m(1 - \lambda)}.
\]

As argued by Storey and Tibshirani, there is a trade-off between bias (which decreases when \(\lambda \to 1\)) and variance (which increases when \(\lambda \to 1\)). Considering \(\hat{\pi}_0\) as a function of \(\lambda\), Storey and Tibshirani proposed to use a cubic spline based method to estimate the quantity \(\lim_{\lambda \to 1} \hat{\pi}_0(\lambda)\).

Actually, noting \(F\) the marginal cumulative distribution function (cdf) of \(P\), Storey and Tibshirani’s estimator can be viewed such as:

\[
\hat{\pi}_0(\lambda) = 1 - \hat{F}(\lambda) = \frac{1 - \hat{F}(\lambda)}{1 - \lambda}.
\]

Then, the estimated quantity is:

\[
\lim_{\lambda \to 1} \hat{\pi}_0(\lambda) = \lim_{\lambda \to 1} \frac{1 - \hat{F}(\lambda)}{1 - \lambda} = \frac{d\hat{F}}{d\lambda}(1) = \hat{f}(1).
\]

Pounds and Morris (2003) have proposed a parametric method assuming that the marginal distribution of the \(P\)-values arises from a beta-uniform mixture distribution. The model parameters are estimated using the maximum-likelihood method, and \(\pi_0 = f(1)\).

More recently, Pounds and Cheng (2004) have proposed a method also based on the marginal distribution of the \(P\)-values, but applying a local regression method (LOESS; Loader, 1999) to obtain a smooth estimate of \(f\) in a transformed space (for more details on the transformation used, see Pounds and Cheng, 2004).

3 A GENERAL CLASS OF ESTIMATORS

The proposed class of estimators for an upper bound of \(\pi_0\) is based upon the expectation of \(P\) under the model (2) that can be expressed as:

\[
E(P) = \pi_0 + (1 - \pi_0) E_1(P),
\]

where \(E_0\) and \(E_1\) are the expectations of the conditional distribution of \(P\) under the null and the alternative hypothesis, respectively.

Since under the null hypothesis, \(P \sim U[0, 1]\), \(E_0(P) = \frac{1}{2}\) so that the previous equation can be written as: \(2E(P) = \pi_0 + 2(1 - \pi_0) E_1(P)\).

It follows that an estimator of an upper bound of \(\pi_0\) leading to a conservatively biased estimator of \(\pi_0\) is simply

\[
\hat{\pi}_0 = \frac{1}{m} \sum_{i=1}^{m} \frac{P_i}{\pi_0(\lambda)}.
\] (3)

since \(E(\hat{\pi}_0) \leq 1\) (Appendix 1).

As shown below, a transformation of the random variable \(P\) can be considered in order to reduce the bias of this estimator. Noting \(\varphi\) any function defined on \([0, 1]\):

\[
\frac{E(\varphi(P))}{E_0(\varphi(P))} = \frac{\pi_0 + (1 - \pi_0) \varphi(P)}{E_0(\varphi(P))}.
\] (4)

A function \(\varphi\) leading to an estimator with a lower bias than (3) is such as

\[
(1 - \pi_0) \varphi(P) \leq (1 - \pi_0) E_1(\varphi(P)).
\]

that is:

\[
\frac{E_1(\varphi(P))}{E_0(\varphi(P))} \leq \frac{E_1(P)}{E_0(P)}.
\] (5)

Intuitively, functions \(\varphi\) that are well-suited for achieving the above inequality are such as that take on values which are greater for \(P\) close to \(1\) than for \(P\) close to \(0\). The following general theorem gives formal conditions on \(\varphi\) that leads to the required inequality (5).

**Theorem.** Let \(f_0\) and \(f_1\) be two non-increasing probability density functions of the random variable \(P\) defined on \([0, 1]\) (denote \(f_0\) the one such as \(\lim_{p \to 1}(f_1/f_0)(s) \leq 1\), and let \(\varphi\) a real continuous function defined on \([0, 1]\) verifying the following conditions:

(i) \(\lim_{x \to 1} \varphi(x) = +\infty\)
The variance of the random variable $\phi(P)$ can be obtained as follows. The proof of the theorem is given in Appendix 2.

$$E\{\pi_0\}$$ and the general class of estimators proposed for an upper bound

$$m\rightarrow\infty$$, and the general class of estimators proposed for an upper bound

Then:

$$\tilde{\pi}_0 = \left(1/m\right) \sum_{i=1}^{m} \phi(p_i) / E[\phi(P)]$$, \quad \phi \in S.

Assuming the independence of the $P$-values, we obtain results on the asymptotic distribution of $\tilde{\pi}_0$. Indeed, according to the central limit theorem, as $m$ tends to infinity:

$$\tilde{\pi}_0 \sim N\left( E[\phi(P)] / E[\phi(P)]^2, \sigma^2 \right)$$.

where $E[\phi(P)] / E[\phi(P)]$ is an upper bound of $\pi_0$ and $\sigma^2$ is the variance of the random variable $\phi(P)$. Despite $\sigma^2$ is unknown, we can obtain an upper bound of this variance as follows.

Denote $\sigma_0^2$ the variance of the random variable $\phi(P)$ under the null hypothesis and let $\Phi(P) = \{\phi(P) - E[\phi(P)]\}^2$.

Since, $\lim_{n \to \infty} (\phi(x)) = \infty$, $\lim_{n \to \infty} (\phi(x)) < \infty$ and $f_0$ and $f$ are two non-increasing pdf such as $\lim_{x \to +\infty} f(x)/f_0(x) \leq 1$, following the lemma given in Appendix 2:

$E[\phi(P)] - E_0[\phi(P)] \leq E(P) - E_0(P)$

$E[\phi(P)] - E_0[\phi(P)] = E(P) - E_0(P)$

$E[\phi(P)] - E_0[\phi(P)] \leq E(P) - E_0(P)$

$\sigma^2 - \sigma_0^2 \leq E(P) - E_0(P)$.

But, as stated previously (Appendix 1), $E(P) \leq E_0(P)$, then $\sigma^2 \leq \sigma_0^2$.

As the distribution of the $P$-values is known under the null hypothesis, we can obtain an upper bound of the asymptotic variance of the estimator:

$$\frac{1}{E[\phi(P)]^2} \sigma_0^2$$.

In the next section, we propose a particular family of functions $\phi$ belonging to the class $S$ and we provide a method to select one in the family.

4 PROPOSED ESTIMATOR

Let $\phi(x) = -\ln(1 - x)$. This function $\phi$ belongs to the class $S$ and we can show that $\forall n \in \mathbb{N}$, $E_i(\phi(P)^{n+1}) / E_0(\phi(P)^{n+1}) \leq \cdots$
The proposed estimator is an unbiased estimator for an upper bound of \( \pi_0 \) due to the dispersion of the estimator. In practice, for a specified number \( m \) of tested hypotheses, one can choose \( n \) according to a certain value for the variance’s upper bound such as \( n = \max \left \{ 1, \max \left \{ n \in \mathbb{N}^+ \mid \frac{(m-1)}{m} \leq \hat{\pi}_0 \right \} \right \} \). Other rules may obviously be considered.

### 5 SIMULATIONS

In order to compare the proposed estimator of \( \pi_0 \) named LBE (Location Based Estimator) to those provided by QVALUE, BUM and SPLOSH, we performed a simulation study as follows.

Data were generated to mimic a two class comparison study with normalized log-ratio measurements for \( m \) genes \((i = 1, \ldots, m)\) obtained from 20 experiments corresponding to two conditions \((j = 1, 2)\), each with 10 replicated samples \((k = 1, \ldots, 10)\). Three total numbers of genes were considered \((m = 100, 500 \text{ and } 2000)\). In each case, all values were independently sampled from a normal distribution, \( X_{i,j,k} \sim \mathcal{N}(\mu_{ij}, 1)\). For the first condition, all the data were simulated with \( \mu_{1} = 0 \). For the second condition, a proportion \( \pi_0 \) of genes were simulated with \( \mu_0 = 0 \) (unmodified genes) whereas \( \pi_0 \) modified genes were simulated using three different values \((\pi_0 = 0.2, 0.5 \text{ and } 0.8)\).

It is worth noting that, under the null hypothesis, \( \varphi(P) \) follows an exponential distribution with parameter 1. Then, using this variable \( n \) decreases with \( \varphi(P) \). As it can easily be seen, there is a balance between bias (decreasing as \( n \) increase) and variance (increasing as \( n \) increase). Even if the proposed estimator is an unbiased estimator for an upper bound of \( \pi_0 \), it is important to preserve oneself from the risk to underestimate \( \pi_0 \) due to the dispersion of the estimator.

---

### Table 2. Standard error for each simulated configuration with the methods QVALUE, BUM, SPLOSH and LBE with \( n = 1 \) and \( n = 2 \)

<table>
<thead>
<tr>
<th>( m )</th>
<th>( \pi_0 )</th>
<th>Conf.</th>
<th>QVALUE</th>
<th>BUM</th>
<th>SPLOSH</th>
<th>LBE ((n = 1))</th>
<th>LBE ((n = 2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.2</td>
<td>(a)</td>
<td>0.144446</td>
<td>0.020800</td>
<td>0.083601</td>
<td>0.056776</td>
<td>0.111253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.123129</td>
<td>0.005259</td>
<td>0.046003</td>
<td>0.04503</td>
<td>0.103125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.134242</td>
<td>0.010784</td>
<td>0.062438</td>
<td>0.052016</td>
<td>0.114992</td>
</tr>
<tr>
<td>0.5</td>
<td>(a)</td>
<td>0.200141</td>
<td>0.044201</td>
<td>0.120371</td>
<td>0.076975</td>
<td>0.163175</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.192907</td>
<td>0.029282</td>
<td>0.062260</td>
<td>0.068850</td>
<td>0.158191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.199905</td>
<td>0.046662</td>
<td>0.106579</td>
<td>0.073373</td>
<td>0.155267</td>
</tr>
<tr>
<td>0.8</td>
<td>(a)</td>
<td>0.190442</td>
<td>0.067848</td>
<td>0.15249</td>
<td>0.092223</td>
<td>0.206415</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.201031</td>
<td>0.030335</td>
<td>0.125538</td>
<td>0.089315</td>
<td>0.205440</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.200243</td>
<td>0.046509</td>
<td>0.240109</td>
<td>0.090272</td>
<td>0.215778</td>
</tr>
<tr>
<td>500</td>
<td>0.2</td>
<td>(a)</td>
<td>0.067970</td>
<td>0.009311</td>
<td>0.054648</td>
<td>0.026380</td>
<td>0.056836</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.055716</td>
<td>0.002369</td>
<td>0.017699</td>
<td>0.019997</td>
<td>0.043296</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.067759</td>
<td>0.003785</td>
<td>0.034406</td>
<td>0.023189</td>
<td>0.059990</td>
</tr>
<tr>
<td>0.5</td>
<td>(a)</td>
<td>0.091064</td>
<td>0.020922</td>
<td>0.069278</td>
<td>0.033387</td>
<td>0.072376</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.091746</td>
<td>0.013303</td>
<td>0.034630</td>
<td>0.032180</td>
<td>0.072119</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.092356</td>
<td>0.020756</td>
<td>0.068751</td>
<td>0.034396</td>
<td>0.074581</td>
</tr>
<tr>
<td>0.8</td>
<td>(a)</td>
<td>0.109255</td>
<td>0.031400</td>
<td>0.108568</td>
<td>0.039787</td>
<td>0.088537</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.109257</td>
<td>0.013100</td>
<td>0.082031</td>
<td>0.039990</td>
<td>0.090144</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.113246</td>
<td>0.019601</td>
<td>0.102903</td>
<td>0.041502</td>
<td>0.093249</td>
</tr>
<tr>
<td>2000</td>
<td>0.2</td>
<td>(a)</td>
<td>0.034272</td>
<td>0.004592</td>
<td>0.030933</td>
<td>0.013288</td>
<td>0.027721</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.028699</td>
<td>0.001148</td>
<td>0.007631</td>
<td>0.010340</td>
<td>0.023457</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.030548</td>
<td>0.002085</td>
<td>0.016979</td>
<td>0.011907</td>
<td>0.024628</td>
</tr>
<tr>
<td>0.5</td>
<td>(a)</td>
<td>0.045018</td>
<td>0.010728</td>
<td>0.031426</td>
<td>0.016578</td>
<td>0.035501</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.043649</td>
<td>0.006624</td>
<td>0.021714</td>
<td>0.015435</td>
<td>0.034110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.045579</td>
<td>0.009997</td>
<td>0.034938</td>
<td>0.016281</td>
<td>0.035719</td>
</tr>
<tr>
<td>0.8</td>
<td>(a)</td>
<td>0.055588</td>
<td>0.016036</td>
<td>0.069977</td>
<td>0.020032</td>
<td>0.044740</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b)</td>
<td>0.056874</td>
<td>0.006834</td>
<td>0.042342</td>
<td>0.020287</td>
<td>0.046171</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>0.056032</td>
<td>0.009359</td>
<td>0.065325</td>
<td>0.020202</td>
<td>0.044640</td>
</tr>
</tbody>
</table>
In each case, the estimates mean for QVALUE is less than the greatest underestimation of QVALUE estimator is only of 3.8% for the mean of \( \pi \) and with the simulated configurations. As an example, under configuration variance, the selected value is the mean of \( \pi \) and BUM procedures are estimated over 1000 iterations. The estimated standard errors for LBE with \( n = 2 \) in each case.

For each setup, 1000 iterations were performed. The mean, the standard deviation and the mean square error of each estimator were calculated configuration with the different methods. It shows that even if all the estimators are supposed to be conservatively biased, BUM is the mean of \( \pi \) and with configuration \( n \). However, for completeness, we considered the LBE estimation with \( n = 2 \) in each case.

For each simulated configuration with the methods QVALUE, BUM, SPLOSH and LBE with \( n = 1 \) and \( n = 2 \). Table 3 displays the mean square error for each estimator. Table 3 presents the mean square error for each estimator. Compared to QVALUE, Table 3 shows that for \( n = 100 \) and \( n = 500 \) the proposed estimator with \( n = 1 \) has the lowest mean square error in 16 cases out of 18, and for \( n = 2000 \), the proposed estimator with \( n = 2 \) has the lowest mean square error in 6 cases out of 9. For 6 and 5 cases out of 27, SPLOSH and BUM have the lowest mean square error over the five estimators, respectively. However, it is quite difficult to interpret these results since it has been previously shown that these latter estimators tend frequently to underestimate \( \pi_0 \).

As an example, Figure 1 presents the histogram of the different estimators for the four methods in one case \( |n| = 2000 \), configuration
A simple estimator for the FDR

Fig. 1. Estimates distribution for QVALUE, BUM, SPLOSH and LBE with \( n = 1 \) and \( n = 2 \) in the case: \( m = 2000 \), configuration (c) and \( \pi_0 = 0.8 \). It illustrates that the proposed estimator seems to be normally distributed in finite samples, which appears to be roughly true for QVALUE, but not for BUM and SPLOSH. The graphic diagram also illustrates that the variance of QVALUE is higher than the variance of the proposed estimator, and that BUM and SPLOSH, in this case, underestimate \( \pi_0 \).

Concerning QVALUE and LBE, simulation results have shown that the upper bound for \( \pi_0 \) estimated by both methods is closer to the true value as \( \pi_0 \) is increasing and there is a large overlap between the distributions under the null and alternative hypothesis. This is not surprising, since from (1) and (4), the bias is depending on \( \pi_0 \) and the distribution of the \( P \)-values under the alternative hypothesis.

It is worth noting that for practical use, investigator would probably truncate the estimator at one. However, simulations results (data not shown) have shown that if \( n \) is chosen according to the proposed rule, truncating or not the estimator provides very close results.

6 EXAMPLES

Our proposed estimator together with QVALUE, BUM and SPLOSH have been applied to the publicly available datasets from the breast study conducted by Hedenfalk et al. (2001), the leukemia study conducted by Golub et al. (1999) and the apolipoprotein AI (Apo AI) experiment conducted by Callow et al. (2000).

The aim of the study of Hedenfalk et al. (2001) was to examine breast cancer tissues from patients with BRCA1–BRCA2-related cancer and cases of sporadic breast cancer to determine global gene expression patterns in the different classes of tumors. The initial dataset consists of 3226 genes expression ratios corresponding to the fluorescent intensities from a tumor sample divided by those from a common reference sample. For each gene, a log-expression ratio was available. In this paper, we focus on the comparison of BRCA1 and BRCA2 with a subset of 3030 genes for which log-ratio values >0.1 and <10 and the data were normalized following a classical analysis of variance model [same as in Broët et al. (2004)].

The aim of the study of Golub et al. (1999) was to identify the differentially expressed genes between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The expression levels of 6817 genes were measured using Affymetrix high-density oligonucleotide chips. Data were pre-processed as described in Dudoit et al. (2002), leading to the analysis of 3051 genes.

The aim of the study of Callow et al. (2000) was to identify genes with altered expression in the livers of apo AI knock-out mice compared to inbred control mice. The considered dataset consists of 6384 genes expression values corresponding to the log of the fluorescent intensities from a mice sample divided by those from a common reference sample. We excluded genes having at least one fluorescent intensity equal to zero so that 6226 genes were retained and the data were standardized within arrays.

For each dataset, \( P \)-values were calculated for each gene from a two-sample \( t \)-test. Then, we applied the methods QVALUE, BUM, SPLOSH and LBE to these sets of \( P \)-values in order to estimate \( \pi_0 \).
The estimates obtained for \( \pi_0 \) by QVALUE, BUM, SPLOSH and LBE (with \( n = 2 \), that corresponds for the three datasets to a threshold \( l = 0.05^2 \) for the estimator's variance) are as follows. For the Hedenfalk et al. dataset: 0.669, 0.586, 0.622 and 0.688, respectively; for the Golub et al. dataset: 0.496, 0.453, 0.524 and 0.525, respectively; and for the Callow et al. dataset 0.901, 0.837, 0.830 and 0.895, respectively.

For each dataset, LBE and QVALUE estimates are very close, which is not surprising when looking at simulation results presented in the previous section. For the two first datasets, QVALUE estimate is lower than the LBE estimate, but for the third dataset, LBE estimate is lower.

As compared to QVALUE, we can obtain upper bounds for the variances, which are \( 1.65 \times 10^{-3}, 1.64 \times 10^{-3} \) and \( 8.03 \times 10^{-3} \) for the Hedenfalk et al. dataset, the Golub et al. dataset and the Callow et al. dataset, respectively. These variances correspond to standard errors of 4.06, 4.05 and 2.83% respectively.

As seen in Storey and Tibshirani (2003), FDR(\( t \)) is estimated by \( \hat{\pi}_0 \) is an estimate of \( \pi_0 \) that is computed using the \( \hat{\pi}_0 \) estimator, which is the least biased estimator of \( \pi_0 \) among the \( \hat{\pi}_0 \) estimators, and it is the estimator of \( \pi_0 \) that is used in most applications. When selecting all genes so that the FDR is \( \leq \pi_0 \), for the three experiments, the FDR is lower than the LBE estimate, but for the third dataset, LBE estimate is lower.

To understand why this happens, we need to consider the expectation of the transformed \( \hat{\pi}_0 \) values. For all these procedures, a key quantity is the probability for a gene of being unmodified. Estimating this latter quantity without making assumptions on the distribution of modified genes leads to a conservatively biased estimator of \( \pi_0 \) that is unbiased and should be chosen according to the accuracy needed.

In order to select one particular estimator among the proposed family, the following guidelines may be suggested. According to the experimental setup and a threshold \( l = 0.05^2 \), the FDR is estimated by \( \pi_0 \) is an estimate of \( \pi_0 \) that is computed using the \( \hat{\pi}_0 \) estimator, which is the least biased estimator of \( \pi_0 \) among the \( \hat{\pi}_0 \) estimators, and it is the estimator of \( \pi_0 \) that is used in most applications. When selecting all genes so that the FDR is \( \leq \pi_0 \), for the three experiments, the FDR is lower than the LBE estimate, but for the third dataset, LBE estimate is lower.

As compared to QVALUE, we can obtain upper bounds for the variances, which are \( 1.65 \times 10^{-3}, 1.64 \times 10^{-3} \) and \( 8.03 \times 10^{-3} \) for the Hedenfalk et al. dataset, the Golub et al. dataset and the Callow et al. dataset, respectively. These variances correspond to standard errors of 4.06, 4.05 and 2.83% respectively.

As seen in Storey and Tibshirani (2003), FDR(\( t \)) is estimated by \( \hat{\pi}_0 \) is an estimate of \( \pi_0 \) that is computed using the \( \hat{\pi}_0 \) estimator, which is the least biased estimator of \( \pi_0 \) among the \( \hat{\pi}_0 \) estimators, and it is the estimator of \( \pi_0 \) that is used in most applications. When selecting all genes so that the FDR is \( \leq \pi_0 \), for the three experiments, the FDR is lower than the LBE estimate, but for the third dataset, LBE estimate is lower.

In contrast to QVALUE, BUM and SPLOSH that proceed from an estimate of the marginal density evaluated at one with complex procedures, our proposed estimators are simply obtained from the marginal densities evaluated at one with complex procedures, our proposed estimators are simply obtained from the marginal densities evaluated at one with complex procedures. However, this threshold is arbitrary and should be chosen according to the accuracy needed.

As seen in the simulation study, BUM and SPLOSH procedures underestimate \( \pi_0 \) in most of the cases, leading to an anticonservatively biased estimator of the FDR. Simulations study has shown that LBE and QVALUE expectations are close, the latter one providing the less biased estimator of \( \pi_0 \). However, our proposed estimator has the smallest variance, so that the risk to underestimate \( \pi_0 \) is smaller with LBE than with QVALUE. Regarding the bias and variance trade-off, the mean square error of the proposed estimator is the smallest in most of the cases. Applying the four methods on a real dataset, QVALUE and LBE have provided very close results, which is in agreement with the simulation results. BUM and SPLOSH have led to select a greater number of genes, but as shown by the simulation study, these procedures led to select larger numbers of genes.

REFERENCES


APPENDIX

8.1 Proof of \( E(\hat{\pi}_0) \leq 1 \)

Assuming that \( f_k \), the marginal pdf is non-increasing and \( f_0 = 1_{[0,1]} \), F, the marginal cdf and \( F_0 \), then the conditional cdf under the null hypothesis, are such as \( F > F_0 \). Then,

\[
\begin{align*}
E(P) &= 1 - \int_0^1 F(x)dx \\
E_0(P) &= 1 - \int_0^{F_0(x)}dx \\
\Rightarrow E(P) &\leq E_0(P) \Rightarrow E(\hat{\pi}_0) = 2E(P) \leq 2E_0(P) = 1.
\end{align*}
\]
8.2 Proof of theorem

The proof of the theorem follows the lemma:

**Lemma.** Let $f_0$ and $f_1$ two non-increasing probability density function of the random variable $P$ defined on $[0, 1]$ (denote $f_0$ the one such as $\lim_{x \to 1} f_0(x) = 1$) and let $\phi$ a continuous function defined on $[0, 1]$ such as (i) $\lim_{x \to 1} \phi(x) = +\infty$ and (ii) $\lim_{x \to 0} \phi(x) < +\infty$. Then, $E_1(\phi(P)) - E_0(\phi(P)) \leq E_1(P) - E_0(P)$.

**Proof of the Lemma**

(1) \[
\forall a \in [0, 1], \\
\{E_1(\phi(P)) - E_0(\phi(P))\} - \{E_1(P) - E_0(P)\} \\
= \int_0^1 (\phi(x) - x)(f_1 - f_0)(x)dx \\
= \int_0^a (\phi(x) - x)(f_1 - f_0)(x)dx \\
+ \int_a^1 (\phi(x) - x)(f_1 - f_0)(x)dx \\
\Rightarrow \|\{\phi(x) - x\} \leq 1 \\
\Rightarrow \exists a^{*} \in [0, 1], [a^{*}, 1], (f_1 - f_0)(x) \leq 0 \\
\Rightarrow \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
\leq \sup_{x \in [0, 1]} [\phi(x) - x] \int_0^{a^{*}} (f_1 - f_0)(x)dx \\
\times \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
\leq \inf_{x \in [0, 1]} [\phi(x) - x] \int_0^{a^{*}} (f_1 - f_0)(x)dx \\
\Rightarrow \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
+ \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
\leq \sup_{x \in [0, a^{*}]} [\phi(x) - x] - \inf_{x \in [a^{*}, 1]} [\phi(x) - x] \\
\times \int_0^{a^{*}} (f_1 - f_0)(x)dx \\
\]

(2) $f_0$ and $f_1$ pdf
\[
\Rightarrow \int_0^1 (f_1 - f_0)(x)dx = 0 \\
\Rightarrow \forall a \in [0, 1], \\
\times \int_0^a (f_1 - f_0)(x)dx = - \int_a^1 (f_1 - f_0)(x)dx \\
\]

(3) \[
\lim_{x \to 1} \left[ \frac{f_1}{f_0}(x) \right] \leq 1 \\
\Rightarrow \exists a^{*} \in [0, 1]|\forall x \in [a^{*}, 1], (f_1 - f_0)(x) \leq 0 \\
\Rightarrow \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
\leq \sup_{x \in [0, 1]} [\phi(x) - x] \int_0^{a^{*}} (f_1 - f_0)(x)dx \\
\times \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
\leq \inf_{x \in [0, 1]} [\phi(x) - x] \int_0^{a^{*}} (f_1 - f_0)(x)dx \\
\Rightarrow \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
+ \int_0^{a^{*}} [\phi(x) - x](f_1 - f_0)(x)dx \\
\leq \sup_{x \in [0, a^{*}]} [\phi(x) - x] - \inf_{x \in [a^{*}, 1]} [\phi(x) - x] \\
\times \int_0^{a^{*}} (f_1 - f_0)(x)dx \\
\]

**Proof of the Theorem**

(1) Note: As $\phi$ is convex (iii), following the Jensen inequality:
\[
E_0[\phi(P)] \geq \phi[E_0(P)] \\
\]

(2) From the lemma:
\[
E_1(\phi(P)) - E_0(\phi(P)) \leq E_1(P) - E_0(P) \\
\]

Then:
\[
\frac{E_1(\phi(P))}{E_0(\phi(P))} - 1 \leq E_1(P) - E_0(P) \\
\Rightarrow E_1(\phi(P)) - E_0(\phi(P)) - \phi[E_0(P)] \leq E_0(\phi(P)) \text{ from (1)} \\
\Rightarrow E_1(\phi(P)) - E_0(\phi(P)) - E_0(\phi(P)) \leq E_1(P) - E_0(P) \\
\text{[since $\phi[E_0(P)] \geq E_0(P)$ (iv)]} \\
\Rightarrow E_1(\phi(P)) - E_0(\phi(P)) \leq E_1(P) - E_0(P) - 1 \\
\Rightarrow E_1(\phi(P)) \leq E_1(P) - E_0(P) \\
\Rightarrow E_0(\phi(P)) \leq E_0(P) \\
\]

667
8.3 Proof of $\forall n \in \mathbb{N}, \frac{E_{1}(\psi(P)^n)}{E_{0}(\psi(P)^n)} \leq \frac{E_{1}(\psi(P)^{n+1})}{E_{0}(\psi(P)^{n+1})}$

[with $\psi(P) = -(1 - P)$]

Following the same argumentation as previously, the following variant of the theorem can easily be shown:

**Theorem.** Let $g_0$ and $g_1$ two non-increasing pdf of the random variable $Z$ defined on $[0, +\infty]$ denote $g_0$ the one such as $\lim_{x \to +\infty} \frac{g_1}{g_0}(x) \leq 1$, and let $\psi$ a real function defined on $[0, +\infty]$ verifying the following conditions:

(i) $\lim_{x \to +\infty} \psi(x) - x = +\infty$

(ii) $\lim_{x \to 0} \psi(x) < +\infty$

(iii) $\psi$ is convex

(iv) $E_0(Z) = n! \geq 1$ \Rightarrow $\psi[E_0(Z)] = E_0(Z)^{(n+1)/n} \geq E_0(Z)$

(Appendix 4)

Then, following the previous theorem:

$$\frac{E_1[\psi(Z)]}{E_0[\psi(Z)]} \leq \frac{E_1(Z)}{E_0(Z)}$$

$$\frac{E_1[\psi(P)^{n+1}]}{E_0[\psi(P)^{n+1}]} \leq \frac{E_1[\psi(P)^n]}{E_0[\psi(P)^n]}$$

8.4 Proof of $\psi(P) \sim \exp(1) \Rightarrow E_0[\psi(P)^n] = n!$

Let $X \sim \exp(1)$

The equality $E_0[\psi(P)^n] = n!$ is obviously true for $n = 1$ and $n = 2$:

$$E(X) = 1!$$

$$E(X^2) = 2!$$

Let's assume that $E(X^n) = n!$ and lets show that $E(X^{n+1}) = (n+1)!$:

$$E(X^{n+1}) = \int_0^{+\infty} x^{n+1} e^{-x} \, dx$$

$$= \left[-x^{n+1} e^{-x}\right]_0^{+\infty} + (n+1) \int_0^{+\infty} x^n e^{-x} \, dx$$

$$= (n+1) E(X^n)$$

$$= (n+1)!$$