BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 5 2005, pages 660–668
doi:10.1093/bioinformatics/bti063

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 π_0, which is the probability for a gene to be unmodified. In this paper, we propose a novel family of estimators for π_0 that allows the calculation of FDR.

Results: The very simple method for estimating π_0 called LBE (Location 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/lbe

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 Γ, which is the same for every test,

$$\text{pFDR}(\Gamma) = \Pr(H = 0 | T \in \Gamma) = \frac{\pi_0 \Pr(T \in \Gamma | H = 0)}{\Pr(T \in \Gamma)}, \quad (1)$$

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 π_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 π_0.

In this paper, we provide a class of estimators for an upper bound of π_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 SPLOSH for obtaining a conservatively biased estimator for π_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 π_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 SPLOSH. 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:

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p). \tag{2}$$

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 π_0 based on (2). Thus, an unbiased estimator of $f(1)$ provides a conservatively biased estimator of π_0. As seen below, the procedures QVALUE, BUM and SPLOSH are based on this latter estimator whereas our procedure is based on the expectation of transformed P-values.

A widely used estimator for π_0 is the one proposed by Storey and Tibshirani (2003). Using a tuning parameter $\lambda \in [0, 1]$, π_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 λ, 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:

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

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 $\hat{\pi}_0 = \hat{f}(1)$.

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 π_0 is based upon the expectation of P under the model (2) that can be expressed as:

$$\frac{E(P)}{E_0(P)} = \pi_0 + (1 - \pi_0)\frac{E_1(P)}{E_0(P)},$$

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 π_0 leading to a conservatively biased estimator of π_0 is simply

$$\hat{\pi}_0 = 2\frac{1}{m}\sum_{i=1}^{m} P_i \tag{3}$$

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 φ any function defined on $[0, 1]$:

$$\frac{E[\varphi(P)]}{E_0[\varphi(P)]} = \pi_0 + (1 - \pi_0)\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]}. \tag{4}$$

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

$$(1 - \pi_0)\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]} \leq (1 - \pi_0)\frac{E_1(P)}{E_0(P)},$$

that is:

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

Intuitively, functions φ 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 φ 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* $\lim_{x \to 1}(f_1/f_0)(x) \leq 1$), *and let φ a real continuous function defined on* $[0, 1]$ *verifying the following conditions:*

(i) $\lim_{x \to 1} \varphi(x) = +\infty$

Table 1. Mean of all estimates for each simulated configuration with the methods QVALUE, BUM, SPLOSH and LBE with $n = 1$ and $n = 2$

m	π_0	Conf.	QVALUE	BUM	SPLOSH	LBE ($n = 1$)	LBE ($n = 2$)
100	0.2	(a)	0.249 306	0.320 796	0.209 010	0.336 732	0.277 032
		(b)	0.197 418	0.142 823	0.120 499	0.201 282	0.199 170
		(c)	0.225 291	0.199 298	0.196 262	0.270 588	0.244 248
	0.5	(a)	0.523 161	0.451 669	0.400 958	0.583 062	0.546 987
		(b)	0.506 183	0.364 379	0.338 784	0.503 977	0.505 531
		(c)	0.512 346	0.428 324	0.392 608	0.542 959	0.524 695
	0.8	(a)	0.784 258	0.744 977	0.556 002	0.839 616	0.831 779
		(b)	0.773 398	0.703 399	0.446 574	0.801 924	0.806 553
		(c)	0.761 632	0.743 969	0.503 310	0.812 241	0.796 493
500	0.2	(a)	0.251 933	0.321 501	0.236 210	0.338 584	0.283 288
		(b)	0.197 976	0.142 870	0.156 950	0.203 072	0.199 344
		(c)	0.223 192	0.197 882	0.230 946	0.270 274	0.239 555
	0.5	(a)	0.536 112	0.440 906	0.486 246	0.586 282	0.552 493
		(b)	0.495 076	0.365 937	0.418 347	0.500 949	0.497 814
		(c)	0.513 054	0.430 535	0.479 036	0.543 493	0.524 678
	0.8	(a)	0.806 984	0.748 141	0.671 984	0.832 719	0.819 043
		(b)	0.800 555	0.705 589	0.553 156	0.801 681	0.802 838
		(c)	0.808 455	0.749 179	0.634 703	0.817 921	0.812 607
2000	0.2	(a)	0.252 816	0.320 720	0.253 153	0.337 829	0.281 443
		(b)	0.198 622	0.142 818	0.170 251	0.202 982	0.200 213
		(c)	0.225 407	0.197 825	0.255 394	0.270 831	0.241 148
	0.5	(a)	0.533 424	0.436 461	0.524 105	0.586 090	0.550 845
		(b)	0.499 708	0.366 203	0.473 210	0.501 963	0.499 902
		(c)	0.515 855	0.431 369	0.526 475	0.544 224	0.526 171
	0.8	(a)	0.810 799	0.751 605	0.739 347	0.834 050	0.818 360
		(b)	0.797 784	0.705 598	0.588 101	0.800 280	0.799 031
		(c)	0.803 206	0.750 330	0.708 212	0.816 658	0.807 745

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

(iii) φ *is convex*

(iv) $\varphi(E_0(P)) \geq E_0(P)$

Then:

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

The proof of the theorem is given in Appendix 2.

In the following, we denote S the set of functions verifying (i) to (iv), and the general class of estimators proposed for an upper bound of π_0 is:

$$\hat{\pi}_0 = \frac{(1/m)\sum_{i=1}^{m} \varphi(p_i)}{E_0[\varphi(P)]}, \quad \varphi \in S.$$

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

$$\hat{\pi}_0 \sim N\left(\frac{E[\varphi(P)]}{E_0[\varphi(P)]}, \frac{1}{E_0[\varphi(P)]^2}\frac{\sigma^2}{m}\right),$$

where $E[\varphi(P)]/E_0[\varphi(P)]$ is an upper bound of π_0 and σ^2 is the variance of the random variable $\varphi(P)$. Despite σ^2 is unknown, we can obtain an upper bound of this variance as follows.

Denote σ_0^2 the variance of the random variable $\varphi(P)$ under the null hypothesis and let $\Phi(P) = \{\varphi(P) - E[\varphi(P)]\}^2$.

Since, $\lim_{x\to 1}(\Phi(x)) = \infty$, $\lim_{x\to 0}(\Phi(x)) < \infty$ and f_0 and f are two non-increasing pdf such as $\lim_{x\to 1}\left[\frac{f_1}{f_0}(x)\right] \leq 1$, following the lemma given in Appendix 2:

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

$$E\{\{\varphi(P) - E[\varphi(P)]\}^2\} - E_0\{\{\varphi(P) - E[\varphi(P)]\}^2\}$$

$$\leq E(P) - E_0(P)$$

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

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:

$$\frac{1}{E_0[\varphi(P)]^2}\frac{\sigma_0^2}{m}.$$

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

4 PROPOSED ESTIMATOR

Let $\varphi(x) = -\ln(1 - x)$. This function φ belongs to the class S and we can show that $\forall n \in \mathbb{N}$, $E_1(\varphi(P)^{n+1})/E_0(\varphi(P)^{n+1}) \leq$

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

m	π_0	Conf.	QVALUE	BUM	SPLOSH	LBE ($n = 1$)	LBE ($n = 2$)
100	0.2	(a)	0.144 446	0.020 800	0.083 601	0.056 776	0.111 253
		(b)	0.123 129	0.005 259	0.046 003	0.045 034	0.103 125
		(c)	0.134 242	0.010 784	0.062 438	0.052 016	0.114 992
	0.5	(a)	0.200 141	0.044 201	0.120 371	0.076 975	0.163 175
		(b)	0.192 907	0.029 282	0.062 260	0.068 850	0.158 191
		(c)	0.199 905	0.046 662	0.106 579	0.073 373	0.155 267
	0.8	(a)	0.190 442	0.067 848	0.152 469	0.092 223	0.206 415
		(b)	0.201 031	0.030 335	0.125 538	0.089 315	0.205 440
		(c)	0.200 243	0.046 509	0.140 019	0.090 272	0.195 778
500	0.2	(a)	0.067 970	0.009 311	0.054 648	0.026 380	0.056 836
		(b)	0.055 716	0.002 369	0.017 699	0.019 997	0.043 296
		(c)	0.061 759	0.003 785	0.034 406	0.023 189	0.049 990
	0.5	(a)	0.091 064	0.020 922	0.069 278	0.033 387	0.072 376
		(b)	0.091 746	0.013 303	0.034 630	0.032 180	0.072 119
		(c)	0.092 356	0.020 756	0.068 751	0.034 396	0.074 581
	0.8	(a)	0.109 255	0.031 400	0.108 568	0.039 787	0.088 537
		(b)	0.109 257	0.013 100	0.082 031	0.039 990	0.090 144
		(c)	0.113 246	0.019 601	0.102 903	0.041 502	0.093 249
2000	0.2	(a)	0.034 272	0.004 592	0.030 933	0.013 288	0.027 721
		(b)	0.028 699	0.001 148	0.007 631	0.010 340	0.023 457
		(c)	0.030 548	0.002 085	0.016 979	0.011 907	0.024 628
	0.5	(a)	0.045 018	0.010 728	0.031 426	0.016 578	0.035 501
		(b)	0.043 649	0.006 624	0.021 714	0.015 435	0.034 110
		(c)	0.045 579	0.009 997	0.034 938	0.016 281	0.035 719
	0.8	(a)	0.055 588	0.016 036	0.069 977	0.020 032	0.044 740
		(b)	0.056 874	0.006 834	0.042 342	0.020 287	0.046 171
		(c)	0.056 032	0.009 539	0.065 325	0.020 202	0.044 640

$(E_1(\varphi(P)^n))/(E_0(\varphi(P)^n))$ (Appendix 3). Then, the set of functions $\varphi(x)^n$ leads to a family of estimators for which the bias for π_0 is decreasing with n.

It is worth noting that, under the null hypothesis, $\varphi(P)$ follows an exponential distribution with parameter 1. Then, using this variable change, $E_0(\varphi(P)^n) = n!$ (Appendix 4) and, for $n \in \mathbb{N}$, the proposed estimator is:

$$\hat{\pi}_{0(n)} = \frac{(1/m)\sum_{i=1}^{m}[-\log(1-p_i)]^n}{n!}. \tag{6}$$

Following results stated in the previous section,

$$\hat{\pi}_{0(n)} \sim N\left[\frac{E(\varphi(P)^n)}{n!}, \frac{1}{(n!)^2}\frac{\sigma^2_{(n)}}{m}\right]$$

where $\sigma^2_{(n)}$ is the variance of the random variable $\varphi(P)^n$.

An upper bound of $\sigma^2_{(n)}$ is $\sigma^2_{0(n)} = [E_0(\varphi(P)^{2n})] - [E_0(\varphi(P)^n)]^2 = (2n)! - (n!)^2$. Then,

$$\mathrm{Var}(\hat{\pi}_{0(n)}) \leq \frac{1}{(n!)^2}\frac{(2n)! - (n!)^2}{m} = \frac{\binom{n}{2n} - 1}{m}. \tag{7}$$

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 π_0, it is important to preserve oneself from the risk to underestimate π_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 l for the variance's upper bound such as $n = \max\left[1,\ \max\left(n \in \mathbb{N}^* | \frac{\binom{n}{2n}-1}{m} \leq l\right)\right]$. Other rules may obviously be considered.

5 SIMULATIONS

In order to compare the proposed estimator of π_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 and 2000). In each case, all values were independently sampled from a normal distribution, $X_{i,j,k} \sim N(\mu_{ij}, 1)$. For the first condition, all the data were simulated with $\mu_{i1} = 0$. For the second condition, a proportion π_0 of genes were simulated with $\mu_{i2} = 0$ (unmodified genes) whereas modified genes were simulated using three different configurations: (a) $\mu_{i2} = 1$ for all modified genes; (b) $\mu_{i2} = 2$ for all modified genes; (c) the first half with $\mu_{i2} = 1$, the second half with $\mu_{i2} = 2$. Different π_0 values were considered ($\pi_0 = 0.2$, 0.5 and 0.8).

Table 3. Mean square error for each simulated configuration with the methods QVALUE, BUM, SPLOSH and LBE with $n = 1$ and $n = 2$

m	π_0	Conf.	QVALUE	BUM	SPLOSH	LBE ($n = 1$)	LBE ($n = 2$)
100	0.2	(a)	0.023 275	0.015 024	0.007 063	0.021 916	0.018 299
		(b)	0.015 152	0.003 297	0.008 435	0.002 028	0.010 624
		(c)	0.018 642	0.000 117	0.003 908	0.007 685	0.015 168
	0.5	(a)	0.040 552	0.004 288	0.024 284	0.012 818	0.028 807
		(b)	0.037 214	0.019 250	0.029 863	0.004 751	0.025 030
		(c)	0.040 074	0.007 313	0.022 881	0.007 224	0.024 693
	0.8	(a)	0.036 479	0.007 626	0.082 759	0.010 066	0.043 574
		(b)	0.041 081	0.010 251	0.140 654	0.007 973	0.042 207
		(c)	0.041 529	0.005 301	0.107 611	0.008 291	0.038 303
500	0.2	(a)	0.007 313	0.014 849	0.004 295	0.019 901	0.010 164
		(b)	0.003 105	0.003 270	0.002 166	0.000 409	0.001 873
		(c)	0.004 349	0.000 019	0.002 140	0.005 476	0.004 061
	0.5	(a)	0.009 588	0.003 929	0.004 984	0.008 558	0.007 989
		(b)	0.008 434	0.018 150	0.007 865	0.001 035	0.005 201
		(c)	0.008 691	0.005 256	0.005 162	0.003 074	0.006 166
	0.8	(a)	0.011 974	0.003 674	0.028 164	0.002 652	0.008 194
		(b)	0.011 925	0.009 085	0.067 654	0.001 600	0.008 126
		(c)	0.012 883	0.002 967	0.037 901	0.002 042	0.008 845
2000	0.2	(a)	0.003 963	0.014 594	0.003 781	0.019 173	0.007 401
		(b)	0.000 825	0.003 271	0.000 943	0.000 116	0.000 549
		(c)	0.001 577	0.000 009	0.003 357	0.005 159	0.002 299
	0.5	(a)	0.003 142	0.004 152	0.001 567	0.007 686	0.003 844
		(b)	0.001 903	0.017 945	0.001 189	0.000 242	0.001 162
		(c)	0.002 327	0.004 810	0.001 920	0.002 221	0.001 959
	0.8	(a)	0.003 204	0.002 599	0.008 570	0.001 560	0.002 337
		(b)	0.003 237	0.008 958	0.046 692	0.000 412	0.002 131
		(c)	0.003 147	0.002 558	0.012 688	0.000 685	0.002 051

In each case, the P-values, calculated under the null hypothesis H_0: $\mu_{i1} = \mu_{i2}$, were obtained from the Student's statistic. Then, we estimated π_0 from QVALUE, BUM, SPLOSH and LBE.

In the previous section, we provide a method to select n for the LBE estimator according to the experimental setup and a chosen threshold l for the variance. Using this rule with $l = 0.05^2$ for the variance, the selected value is $n = 1$ for $m = 100$ and $m = 500$ and $n = 2$ for $m = 2000$. However, for completeness, we considered the LBE estimation with $n = 1$ and $n = 2$ in each case.

For each setup, 1000 iterations were performed. The mean, the standard deviation and the mean square error of each estimator were estimated over 1000 iterations.

Table 1 displays the means of the five estimators (for each simulated configuration with the different methods). It shows that even if all the estimators are supposed to be conservatively biased, BUM and SPLOSH procedures dramatically underestimate π_0 in most of the simulated configurations. As an example, under configuration (b) and with $\pi_0 = 80\%$ and $m = 500$, the estimates mean for SPLOSH and BUM procedures are $\hat{\pi}_0 = 55\%$ and $\hat{\pi}_0 = 71\%$. For a few cases, the estimates mean for QVALUE is less than π_0, particularly for a small number of genes and high value of π_0. Nevertheless, the greatest underestimation of QVALUE estimator is only of 3.8% [for $\pi_0 = 80\%$, $m = 100$ and configuration (c)]. The proposed estimator with $n = 1$ provides an upper bound for π_0 in all the cases. For $n = 2$, the mean of $\hat{\pi}_0$ over 1000 simulations is less than π_0 with a small difference ($<3 \times 10^{-3}$) in only six cases, which can be explained by the variability of the estimates mean. The estimations provided by LBE are greater than those provided by QVALUE in all cases except one. However, the difference is not >8.7% for $n = 1$ and 4.8% for $n = 2$.

In contrast, Table 2 which displays the standard error estimation for each method, shows that the standard error of the proposed estimator for $n = 1$ is always less than the standard error of QVALUE (the least difference is 1.8%). As expected, the proposed estimator's mean decreases with n (in almost all cases) and variance increases in all cases with n. However, for $n = 2$, there are only two cases for which the proposed estimator standard error is greater than QVALUE's standard error.

The estimated standard errors for LBE with $n = 1$ and $n = 2$ are less than the upper bounds calculated from (7) for the standard error. Indeed, for $m = 100$, 500 and 2000 the calculated values are 0.1, 0.045, 0.022 (for $n = 1$) and 0.224, 0.1, 0.05 (for $n = 2$).

Table 3 presents the mean square error for each estimator. Compared to QVALUE, Table 3 shows that for $m = 100$ and $m = 500$, the proposed estimator with $n = 1$ has the lowest mean square error in 16 cases out of 18, and for $m = 2000$, 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 π_0.

As an example, Figure 1 presents the histogram of the different estimators for the four methods in one case [$m = 2000$, configuration

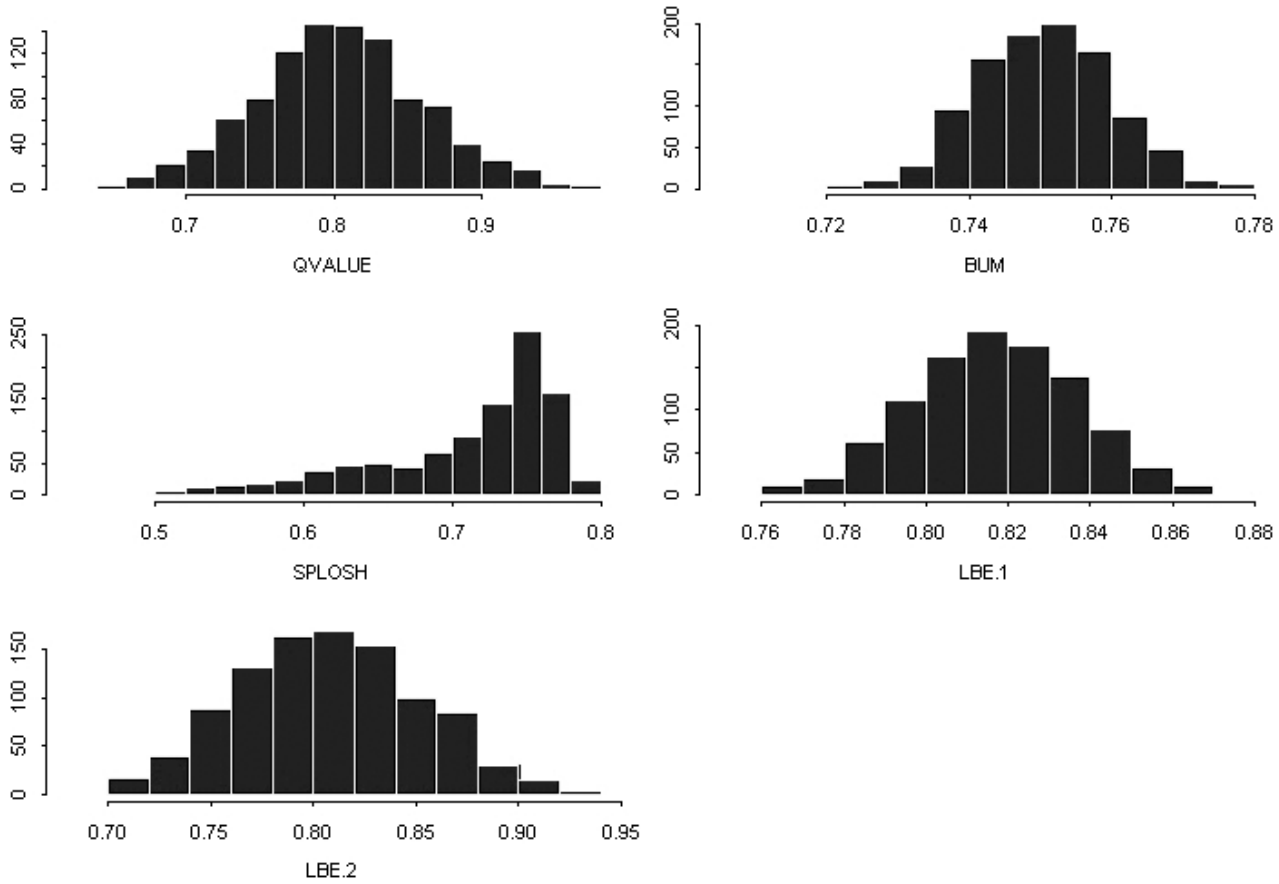

Fig. 1. Estimates distribution for QVALUE, BUM, SPLOSH and LBE with $n = 1$ and $n = 2$ in the case: $m = 2000$, configuration (c) and $\pi_0 = 0.8$.

(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 π_0.

Concerning QVALUE and LBE, simulation results have shown that the upper bound for π_0 estimated by both methods is closer to the true value as π_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 π_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 π_0.

The estimates obtained for π_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 1.65×10^{-3}, 1.64×10^{-3} and 8.03×10^{-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 $(\hat{\pi}_0 mt/\#\{p_i \leq t\})$. When selecting all genes so that the FDR is $<$10%, for the three experiments, QVALUE leads to select 290, 1206 and 9 genes, respectively, and our proposed method leads to select 282, 1187 and 9 genes, respectively. BUM and SPLOSH procedures generally led to select larger numbers of genes but as shown by the simulation study, these procedures led to underestimate π_0 in many cases and the true FDR may be quite $>$10%.

7 DISCUSSION

In this paper, we propose a novel procedure for estimating the FDR that proceeds, as QVALUE, BUM and SPLOSH, from the marginal distribution of the P-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 the FDR.

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 expectation of the transformed P-values. Moreover, we provide results on their asymptotic distribution under the assumption that the P-values are independent. From these estimators, FDR and q-values are easy to obtain.

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 $l = 0.05^2$ for the variance upper bound of the estimator, $n = 1$ for $2 \leq m < 2000$, $n = 2$ for $2000 \leq m < 7500$ and $n = 3$ for $m \geq 7500$. However, this threshold l is arbitrary and should be chosen according to the accuracy needed.

As seen in the simulation study, BUM and SPLOSH procedures underestimate π_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 π_0. However, our proposed estimator has the smallest variance, so that the risk to underestimate π_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 these results have to be taken cautiously when looking at simulation results.

Although the proposed method is dedicated to the FDR, the estimate of π_0 can be used with other criteria such as the local FDR (Efron *et al*., 2001).

In conclusion, the proposed method for estimating an upper bound of π_0 appears to be very useful for calculating the FDR and should be recommended for its nice properties and its simplicity.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Broët,P., Lewin,A., Richardson,S., Dalmasso,C. and Magdelenat,H. (2004) A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* (Epub ahead of print).

Callow,M.J., Dudoit,S., Gong,E.L., Speed,T.P. and Rubin,E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, **10**, 2022–2029.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 457, 77–87.

Efron,B., Tibshirani,R., Storey,J. and Tusher,V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Guterson,B., Esteller,M., Kallioniemi,O.P. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **22**, 539–548.

Hochberg,Y. and Tamhane,A. (1987) *Multiple Comparison Procedures*. Wiley.

Loader,C. (1999) *Local Regression and Likelihood*. Springer-Verlag, NY.

McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. Wiley, NY.

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737–1745.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Schena,M. (2000) Microarray biochip technology. *Biotechniques* (in press).

Storey,J.D. (2001) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B*, **64**, 479–498.

Storey,J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.*, **31**, 2013–2035.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

8 APPENDIX

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

Assuming that f, 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,

$$\left\{ \begin{array}{l} E(P) = 1 - \int_0^1 F(x)dx \\ E_0(P) = 1 - \int_0^1 F_0(x)dx \end{array} \right\}$$

$$\Rightarrow E(P) \leq E_0(P) \Rightarrow E(\hat{\pi}_0) = 2E(P) \leq 2E_0(P) = 1.$$

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* $\lim_{x\to 1}\frac{f_1}{f_0}(x) \leq 1$) *and let φ a continuous function defined on* $[0,1]$ *such as* (i) $\lim_{x\to 1}\varphi(x) = +\infty$ *and* (ii) $\lim_{x\to 0}\varphi(x) < +\infty$.
Then, $E_1[\varphi(P)] - E_0[\varphi(P)] \leq E_1(P) - E_0(P)$.

PROOF OF THE LEMMA

(1)
$$\begin{aligned}
&\forall a \in [0,1],\\
&\{E_1[\varphi(P)] - E_0[\varphi(P)]\} - \{E_1(P) - E_0(P)\}\\
&\quad = \int_0^1 (\varphi(x)-x)(f_1-f_0)(x)dx\\
&\quad = \int_0^a (\varphi(x)-x)(f_1-f_0)(x)dx\\
&\qquad + \int_a^1 (\varphi(x)-x)(f_1-f_0)(x)dx
\end{aligned}$$

(2)
$$\begin{aligned}
&f_0 \text{ and } f_1 \text{ pdf}\\
&\Rightarrow \int_0^1 (f_1-f_0)(x)dx = 0\\
&\Rightarrow \forall a \in [0,1],\\
&\quad \times \int_0^a (f_1-f_0)(x)dx = -\int_a^1 (f_1-f_0)(x)dx
\end{aligned}$$

(3)
$$\begin{aligned}
&\lim_{x\to 1}\left[\frac{f_1}{f_0}(x)\right] \leq 1\\
&\quad \Rightarrow \exists a^* \in [0,1] | \forall x \in [a^*,1], \quad (f_1-f_0)(x) \leq 0\\
&\quad \Rightarrow \int_0^{a^*} [\varphi(x)-x](f_1-f_0)(x)dx\\
&\qquad \leq \sup_{x\in[0,a^*]}[\varphi(x)-x]\int_0^{a^*}(f_1-f_0)(x)dx\\
&\qquad\quad \times \int_{a^*}^1 [\varphi(x)-x](f_1-f_0)(x)dx\\
&\qquad \leq \inf_{x\in[a^*,1]}[\varphi(x)-x]\int_{a^*}^1 (f_1-f_0)(x)dx\\
&\quad \Rightarrow \int_0^{a^*} [\varphi(x)-x](f_1-f_0)(x)\,dx\\
&\qquad + \int_{a^*}^1 [\varphi(x)-x](f_1-f_0)(x)dx\\
&\qquad \leq \left[\sup_{x\in[0,a^*]}[\varphi(x)-x] - \inf_{x\in[a^*,1]}[\varphi(x)-x]\right]\\
&\qquad\quad \times \int_0^{a^*}(f_1-f_0)(x)dx
\end{aligned}$$

(4)
$$\left\{\begin{array}{c}\lim_{x\to 0}\varphi(x) < +\infty \text{ [condition (ii)]}\\ \varphi \text{ continuous}\end{array}\right\}$$
$$\Rightarrow \forall a \in [0,1], \varphi(a) - a < +\infty$$

(5) Let $A = \sup_{x\in[0,a^*]}[\varphi(x)-x]$.
$$\begin{aligned}
&\lim_{x\to 1}\varphi(x) = +\infty \text{ [condition (i)]}\\
&\quad \Rightarrow \lim_{x\to 1}\varphi(x) - x = +\infty\\
&\quad \Leftrightarrow \forall B > 0, \exists \eta > 0 | \forall x \in [0,1[,\\
&\qquad \{1-x < \eta \Rightarrow \varphi(x) - x > B\}\\
&\quad \Rightarrow \exists a^{**} > a^* | \forall x \in [0,1[,\\
&\qquad \{x > a^{**} \Rightarrow \varphi(x) - x > A\}\\
&\quad \Rightarrow \sup_{x\in[0,a^{**}]}[\varphi(x)-x] \leq \inf_{x\in[a^{**},1]}[\varphi(x)-x]\\
&\quad \Rightarrow \left[\sup_{x\in[0,a^{**}]}(\varphi(x)-x) - \inf_{x\in[a^{**},1]}(\varphi(x)-x)\right] \leq 0\\
&\quad \Rightarrow \left[\sup_{x\in[0,a^{**}]}(\varphi(x)-x) - \inf_{x\in[a^{**},1]}(\varphi(x)-x)\right]\\
&\qquad \times \int_0^{a^{**}}(f_1-f_0)(x)dx \leq 0
\end{aligned}$$

$$\left[\begin{array}{l}
\text{since } a^{**} > a^*\\
\quad \Rightarrow \int_{a^*}^1 (f_1-f_0)(x)dx \leq 0\\
\quad \Rightarrow \int_0^{a^{**}} (f_1-f_0)(x)dx \geq 0\\
\quad \Rightarrow \{E_1[\varphi(P)] - E_0[\varphi(P)]\} - \{E_1(P) - E_0(P)\} \leq 0\\
\quad \Rightarrow E_1[\varphi(P)] - E_0[\varphi(P)] \leq E_1(P) - E_0(P)
\end{array}\right]$$

PROOF OF THE THEOREM

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

(2) From the lemma:
$$\begin{aligned}
&E_1[\varphi(P)] - E_0[\varphi(P)] \leq E_1(P) - E_0(P)\\
\text{Then: } &\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]} - 1 \leq \frac{E_1(P) - E_0(P)}{E_0[\varphi(P)]}\\
\Rightarrow &\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]} - 1 \leq \frac{E_1(P) - E_0(P)}{\varphi[E_0(P)]} \text{ [from (1)]}\\
\Rightarrow &\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]} - 1 \leq \frac{E_1(P) - E_0(P)}{E_0(P)}\\
&\text{[since } \varphi[E_0(P)] \geq E_0(P) \text{ (iv)]}\\
\Rightarrow &\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]} - 1 \leq \frac{E_1(P)}{E_0(P)} - 1\\
\Rightarrow &\frac{E_1[\varphi(P)]}{E_0[\varphi(P)]} \leq \frac{E_1(P)}{E_0(P)}
\end{aligned}$$

8.3 Proof of $\forall n \in \mathbb{N}$, $\frac{E_1(\varphi(P)^{n+1})}{E_0(\varphi(P)^{n+1})} \leq \frac{E_1(\varphi(P)^n)}{E_0(\varphi(P)^n)}$ [with $\varphi(P) = -(1-P)$]

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 ψ 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) ψ *is convex*

(iv) $\psi[E_0(Z)] \geq E_0(Z)$

Then:

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

Denote g_0 and g_1 the conditional pdf of the random variable $Z = \varphi(P)^n$ under the null hypothesis and under the alternative hypothesis, respectively. $\lim_{x\to\infty} \frac{g_1}{g_0}(x) \leq 1$. Indeed:

$$\begin{aligned}
\lim_{x\to+\infty} \frac{g_1}{g_0}(x) &= \lim_{x\to+\infty} \frac{f_1(1-\mathrm{e}^{-x^{1/n}}) \times \mathrm{e}^{-x^{1/n}}}{\mathrm{e}^{-x^{1/n}}} \\
&= \lim_{x\to+\infty} f_1(1-\mathrm{e}^{-x^{1/n}}) \\
&= \lim_{y\to 1} f_1(y) \quad \text{with } y = 1-\mathrm{e}^{-x^{1/n}} \\
&\leq 1
\end{aligned}$$

Let $\psi : [0,+\infty] \to \mathbb{R}$ such as $\psi(Z) = Z^{(n+1)/n}$.

(i) $\lim_{x\to+\infty} \psi(x) - x = \lim_{x\to+\infty} x^{(n+1)/n} - x = +\infty$.

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

(iii) $\psi''(x) = [(n+1)/n^2]x^{(1-n)/n} \geq 0 \Rightarrow \psi$ 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:

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

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

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

Let $X \sim \exp(1)$

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

$$E(X) = 1!$$

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

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

$$\begin{aligned}
E(X^{n+1}) &= \int_0^{+\infty} x^{n+1}\mathrm{e}^{-x}dx \\
&= [-x^{n+1}\mathrm{e}^{-x}]_0^{+\infty} + (n+1)\int_0^{+\infty} x^n\mathrm{e}^{-x}dx \\
&= (n+1)E(X^n) \\
&= (n+1)!
\end{aligned}$$