A diagnostic test approach for multitesting problems

Pablo Martínez-Camblor[1], Sonia Pérez-Fernández[2], and Norberto Corral[2]

Abstract In the last decades, multiple-testing problems have received much attention. Many different methods have been proposed in order to deal with this relevant issue. Most of them are focused on controlling some weak version of the Type I error such that the False Discovery Rate. Type II error is frequently forgotten. In this work, the multitesting problem is treated from a diagnostic test approach in which the p-values play the role of the studied predictive marker. In this context, the receiver operating characteristic, ROC, curve is estimated. Several Monte Carlo simulations help for a better understanding of the problem. Finally, a real dataset studying the relationship between atosomal CpG sites and characteristic of hepatocellular carcinoma is considered.

1 Introduction

Modern science frequently produces data on thousands of different features. Probably, the -omics technologies (genomics, transcriptomics, proteomics, etc.) stand for the most relevant examples although other fields like astrophysics, brain imaging or spatial epidemiology have also increased substantially the size of the collected data. Conventionally, statistical analyses of those data often include a huge number of hypotheses to be tested at the same time. In this context, standard statistical concepts, like the p-value, lose their original probabilistic interpretation. Notice that, for any fixed nominal level, the probability of spurious effects, or false positives, greatly increases when a massive number of null hypotheses are simultaneously tested. Classical multiple comparison procedures focus on controlling the probability of committing any Type I error i.e., to control the family wise error rate (FWER). Unfortunately, this objective is too ambitious in this context and more liberal criteria must be used (see Farcomeni [8] for a recent and extensive review of modern multiple hypothesis testing methods).

[1]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Hanover, NH, USA, Pablo.Martinez.Camblor@Dartmouth.edu ·[2]Department of Statistics and Operational Research and Mathematics Didactics, University of Oviedo, Oviedo, Spain, {perezsonia,norbert}@uniovi.es

In the multiple testing context, conventionally N (null) hypotheses are contrasted simultaneously from adequate tests. The classical Table 1 depicts schematically the possible practical situations. Of course, in practice, only the total number of hypotheses, N, and the total number of rejections, r, are really known.

Table 1 Number of mistakes committed when N null hypotheses are simultaneously tested.

	Reject	**Not Reject**	**Total**
Null true	$r - \boldsymbol{R}$	$\boldsymbol{U}$	N_0
Alternative true	$\boldsymbol{R}$	$u - \boldsymbol{U}$	$\boldsymbol{N} - N_0\,(= N_1)$
Total	r	$\boldsymbol{N} - r\,(= u)$	$\boldsymbol{N}$

Notice that controlling the FWER is equivalent to control the probability $\mathcal{P}\{\boldsymbol{r - R} > 0\}$. Seeger [16] introduced and later Benjamini and Hochberg [1] revised and popularized the false discovery rate (FDR), defined as the expected proportion of spurious effects i.e., FDR$= E[(r - \boldsymbol{R})/(r \vee 1)]$ ($a \vee b = b$ if $a \leq b$). The FDR is a frequentist well established definition for the multiple hypothesis testing error and, undoubtedly, the most used procedure.

Several generalizations for the FWER and the FDR criteria and different procedures to implement them have been proposed (see, for instance, Sarkar [15] and references therein). Most recent works deal with the problem of controlling the tail probability of false positives. Genovese and Wasserman [11] proposed to control the tail probability of the false discovery proportion (FDP) by the so labeled FDX (false discovery exceedance) i.e., for a fixed α, to control $\mathcal{P}\{1 - \boldsymbol{R}/r > \alpha\}$. In addition, we highlight the sequential goodness of fit (SGoF) strategy, proposed by Carvajal-Rodríguez *et al.* [2] and deeply explored by de Uña-Álvarez [5]. The SGoF method rejects an amount of null hypotheses equal to the difference between the observed and the expected amounts of p-values below a given threshold under the assumption that all nulls are true (we denote by $\boldsymbol{H}_0 = \cap_{i=1}^{\boldsymbol{N}} H_{0,i}$, in bold, this hypothesis). The outcome of most of those methods is a cutoff point; p-values below this threshold are declared significant while p-values above it are declared non-significant.

Considering each test as a sampling unit and its p-value as a marker of the null hypothesis credibility, the multiple hypothesis testing problem has a clear connection to the classification theory (this point of view was briefly explored by Storey [18]). In this paper, assuming that p-values follow a mixture distribution (see Section 2), multitesting problem is dealt with from a diagnostic test approach; in particular, the well-known receiver-operating characteristic (ROC) curve is derived. This curve does not provide a cut-off

point but graphical information about the real diagnostic capacity of the studied marker, in this case, the p-value. One of the main goals of this work is pointing out the limitation of the process. Notice that, depending on the power of the study, the p-value is not always a good *diagnostic marker*. And even when it is a good diagnostic marker, the classification could be difficult depending on the prevalence of untrue nulls. In Section 3, some particularities of the ROC curve when it is applied on multitesting problems are considered. Section 4 is devoted to the ROC curve estimation. The performance of the proposed method is empirically studied via Monte Carlo simulations. In Section 5, a real data problem is analyzed. Finally, in Section 6, we present our main conclusions.

2 The mixture model

When it is assumed that there exist N independent hypotheses ($H_{0,i}$, $1 \leq i \leq N$) which are going to be tested from adequate tests and that F_0 and F_1 are the cumulative distribution functions (CDFs) for the p-values when the null is true and untrue, respectively, then, the CDF of the p-values will be the mixture distribution

$$\mathbb{G}_N(t) = \pi_0 \cdot F_0(t) + (1 - \pi_0) \cdot F_1(t),$$

where π_0 ($= N_0/N$) is the true null proportion. In this case, for each $t \in [0, 1]$, the function $\mathbb{G}_N(t)$ represents the probability that the p-value associated with a randomly selected hypothesis will be less or equal to t. However, it should be noted that this model assumes that all true nulls follow the same distribution (F_0), which can be plausible, but also that all untrue nulls follow the same distribution (F_1), which is a quite more unrealistic proviso. Without this assumption, the probability that a p-value from an untrue hypothesis will be less or equal to t depends on the particular hypothesis from which this p-value has been drawn. Although depending on the particular experiment studied a random effects or hierarchical model can be more appropriate (see Efron *et al.* [6]), we adopt the most common situation in which the set of hypotheses to be tested are previously fixed, therefore $F_1 = N_1^{-1} \cdot \sum_{i \in J_1} F_{1,i}$, where $J_1 \subseteq \{1, \ldots, N\}$ stands for the set of indices in which the null is untrue. For each $t \in [0, 1]$, $\mathbb{G}_N(t)$ stands for the average probability that the p-value associated with a randomly selected hypothesis will be less or equal to t. This interpretation is still valid in the presence of dependency structures.

Notice that, on the usual assumption that, under the null, the CDF of each individual p-value is $t \cdot \mathbb{I}_{[0,1]}(t)$ ($\mathbb{I}_A$ stands for the standard indicator function on the set A), i.e., each individual p-value follows a uniform distribution within $[0, 1]$, it is derived

$$\mathbb{G}_N(t) = \pi_0 \cdot t \cdot \mathbb{I}_{[0,1]}(t) + (1 - \pi_0) \cdot F_1(t). \tag{1}$$

Remark 1. Although it is reasonable to assume that, under the null, each single p-value is uniformly distributed on $[0, 1]$, and this proviso is true when the null is simple and the distribution of the test statistic is continuous and known, the true p-value distribution is only stochastically dominated by the uniform if its distribution is discrete or the p-value is estimated by a resampling method (see, for instance, Farcomeni [8]).

From the mixture model, the traditional BH procedure for controlling the FDR at level α proposed by Benjamini and Hochberg [1] can be seen as a plug-in method for estimating the threshold (Genovese and Wasserman [10]),

$$T_{\mathrm{BH}}(\alpha) = \sup\{t \in [0, 1] : \mathbb{G}_N(t) \geq t/\alpha\}, \tag{2}$$

and, assuming independence among tests, the SGoF method (taking $\gamma = \alpha$) tries to estimate the threshold (de Uña-Álvarez [5]) as

$$T_{\mathrm{SGoF}}(\alpha) = \mathbb{G}_N^{-1}\left(\mathbb{G}_N(\alpha) - \alpha - z_\alpha \cdot \sqrt{\alpha \cdot (1 - \alpha)/N + N^{-1}}\right), \tag{3}$$

where $\mathbb{G}_N^{-1}(t) = \inf\{s : \mathbb{G}_N(s) \geq t\}$ and $z_\alpha = \Phi^{-1}(1 - \alpha)$ (Φ stands for the standard normal CDF).

3 Receiver operating characteristic curve in the multitesting problem

The receiver-operating characteristic, ROC, curve is a popular graphical method frequently used in order to study the diagnostic capacity of continuous markers. It represents in a plot the true-positive rate (TPR) against the false-positive one (FPR) of all thresholds of the marker. Both practical and theoretical aspects of the receiver operating characteristic curve have been extensively studied in the specialized literature (see Zhou *et al.* [19] for a recent review). Assuming that smaller values of the marker indicate larger confidence that a given subject is positive, let χ and ξ be two continuous random variables representing the marker values for the negative and positive subjects, respectively. Therefore, for a fixed point $t \in [0, 1]$, the ROC curve is defined as follows,

$$\mathcal{R}(t) = F_\xi(F_\chi(t)) = \mathcal{P}\{\xi \leq F_\chi(t)\} = \mathcal{P}\{F_\chi^{-1}(\xi) \leq t\} = F_{F_\chi^{-1}(\xi)}(t), \tag{4}$$

where F_χ and F_ξ denote the CDFs for the variables χ and ξ, respectively. In the current context, p-values play the role of marker values and the true and false nulls are the negative and positive subjects, respectively. Assuming the mixture model (1), the problem is simplified; in this case, the ROC curve

stands for the CDF for the untrue nulls; i.e., $\mathcal{R}(t) = F_1(t)$ $(t \in [0,1])$. Notice that the sensitivity (TPR) must be interpreted as the *average* probability that an untrue null hypothesis will be correctly classified as untrue i.e., the average probability of rejecting an untrue null hypothesis (power). Figure 1 depicts the real ROC curve for $F_1(t) = t^a \cdot \mathbb{I}_{[0,1]}(t)$ with a such that the average sensitivity is 0.8 when the specificity (1-FPR) is 0.95. In order to give more relevance to high specificities, the scale of the x-axis has been modified. Remember that ROC curve is equal to F_1 and the main diagonal (dotted gray line; since the scale of x-axis is modified) is equal to F_0.

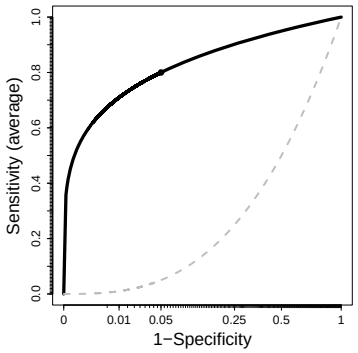

Fig. 1 ROC curve under the mixture model for $F_1(t) = t^a \cdot \mathbb{I}_{[0,1]}(t)$ with $a = log(0.8)/log(0.05)$. Remember that, in this context, $\mathcal{R}(t) = F_1(t)$. Notice that the x-axis is not in linear scale.

Due to the fact that the ROC curve does not depend on the prevalence of the studied characteristic, $1 - \pi_0$, this graphical method provides valuable information about the real capacity of the marker (p-value) to identify the subjects (tests) as positives (rejecting) or negatives (no rejecting). Conventionally, the aforementioned thresholds strongly depend on this prevalence; particularly, for the above situation, $T_{\mathrm{BH}}(0.1)$ is 0.0202, 0.0160 and 0.0076 for $\pi_0 = 0.75$, 0.80 and 0.90, respectively; with these cut-off points, the obtained powers would be 0.749, 0.736 and 0.697. The SGoF method also depends on the number of tests, N; when $N = 10,000$, $T_{\mathrm{SGoF}}(0.1)$ takes the values 0.0086, 0.0074 and 0.0043 for $\pi_0 = 0.75$, 0.80 and 0.90; leading to powers of 0.703, 0.696 and 0.668. The Youden index, Y, is achieved at point 0.06 (Youden index is often used in diagnostic tests in order to obtain an optimal threshold, it is defined as $Y = \max_{t \in \mathbb{R}}\{\mathrm{TPR}(t) - \mathrm{FPR}(t)\}$); at this point, the obtained FPR obviously is 0.060 and the average power 0.811 (= TPR); i.e., by using this cut-off point, in average 6% of the true nulls would be declared false and around 19% of the untrue nulls would not be rejected.

Remark 2. As it is well-known, the area under the ROC curve (AUC) is one of the most commonly used global index of diagnostic accuracy (Faraggi and Reiser [7]). It ranges between 1/2, when the marker does not contribute to a correct classification, and 1, if the marker can classify perfectly all subjects. The AUC has a direct probabilistic interpretation: in particular, it is the

probability that the value of the marker in a randomly chosen negative subject will be higher than the value of the marker in a randomly chosen positive subject. In the current context it can be read as the average of the power when the Type I error varies between 0 and 1. However, large Type I errors are not interesting in practice; hence, limiting the range of the Type I error and considering the partial area under the curve, pAUC (Ma *et al.* [13]), seems to be a more adequate measure. In this case, the AUC is 0.931 and the pAUC between 0 and 0.05 is 0.0372 (0.745 in a $0-1$ scale).

4 Receiver operating characteristic curve estimation

In practice, we observe N p-values, $\{p_1, \ldots, p_N\}$, corresponding to N nulls, $\{H_{0,1}, \ldots, H_{0,N}\}$; Γ will denote the subset of untrue nulls. Assuming independence between the p-values, the Liapunov's Central Limit Theorem (see, for instance, Ibarrola *et al.* [12]) implies the following result:

Theorem 1. *Let $\{p_1, \cdots, p_N\}$ be an independent random sample where for each $i \in 1, \ldots, N$, p_i was drawn from $F_{*,i}$. For each $t \in [0,1]$, let be $\mathbb{G}_N(t) = (1/N) \cdot \sum_{i=1}^{N} F_{*,i}(t)$ and $\hat{\mathbb{G}}_N(t) = (1/N) \cdot \sum_{i=1}^{N} \mathbb{I}_{(-\infty,t]}(p_i)$, then*

$$N \cdot \frac{\hat{\mathbb{G}}_N(t) - \mathbb{G}_N(t)}{\sqrt{\sum_{i=1}^{N} F_{*,i}(t) \cdot (1 - F_{*,i}(t))}} \xrightarrow[N]{\mathcal{L}} \mathcal{N}(0,1). \tag{5}$$

This Central Limit Theorem for non identically distributed variables implies that, if N is sufficiently large, the empirical CDF estimator provides a good approximation of the real distribution function. In Genovese and Wasserman [10] this convergence is deeply considered from a stochastic process approach.

On the other hand, the π_0 estimation has been previously considered in the specialized literature. For instance, Dalmasso *et al.* [4] took advantage of the logarithmic function properties and defined the family of estimators

$$\hat{\pi}_{0,k} = \frac{(1/N) \sum_{i=1}^{N} [-\log(1 - p_i)]^k}{k!}, \text{ with } k \in \mathbb{N}. \tag{6}$$

Once π_0 is estimated, from the mixture model and result (5), the estimation of the ROC curve is direct. However, assuming that for any fixed nominal level the probability of rejecting an untrue null hypothesis is higher than the probability of rejecting a true null hypothesis, and taking into account that $\mathcal{R}$ is a non-decreasing function, then

$$\hat{\mathcal{R}}(t) = \max \left\{ \sup_{s \in [0,t]} \{\hat{F}_1(s)\}, \, t \cdot \mathbb{I}_{[0,1]}(t) \right\}, \tag{7}$$

where $\hat{F}_1(s) = \min\{(\hat{\mathbb{G}}_N(s) - \hat{\pi}_{0,2} \cdot s \cdot \mathbb{I}_{[0,1]}(s)) \cdot (1 - \hat{\pi}_{0,2})^{-1}, 1\}$, is a more appropriate estimator for the ROC curve. Of course, the estimator is still valid by using other π_0 approximations.

4.1 Simulation study

The behavior of the proposed estimator is empirically studied via Monte Carlo simulations. Two different strategies were considered: in the first one, the p-values were directly drawn from different theoretical mixture models; in the second one, the whole problem is simulated, i.e., one sample is drawn and the corresponding p-values are computed from an adequate test where the null hypothesis is $\mu = 0$.

In the first scenario we run independent random samples of size N from the distribution $\mathbb{G}_N(t) = \pi_0 \cdot t \cdot \mathbb{I}_{[0,1]}(t) + (1-\pi_0) \cdot F_1(t)$ where $F_1 = (1/N_1) \sum_{i=1}^{N_1} t^{a_i} \cdot \mathbb{I}_{[0,1]}(t)$ with $a_i = L + (i/N_1) \cdot U$ and two parameters (L and U) selected in order to obtain different power averages; particularly, at level 0.05, power averages ($\bar{\beta}_{0.05}$) of 0.6 and 0.8 were studied.

In the second considered scenario the whole problem is simulated. We consider situations where the nulls to be tested are $H_{0,i} : \mu_i = 0$, where μ_i stands for the expected value of the ith population ($1 \leq i \leq N$). Under the null, we drawn independent samples (with size $n = 25$) from a standard normal distribution while, under the alternative, the samples were drawn from a $\mathcal{N}(\mu^*, 1)$ distribution where μ^* was taken such that the power average was the desired one ($\bar{\beta}_{0.05} = 0.6$ and 0.8 were considered). Then, the p-values were computed by using the Student t-test (parametric).

Figure 2, left, depicts the approximate shape of the involved curves in the first scenario; notice that the real one depends on the number of untrue nulls (N_1). At right, the shape of the respective ROC curve in the second scenario is displayed.

Table 2 shows the observed results for both scenarios on 1,000 Monte Carlo iterations. Particularly, we report mean $\pm$ standard deviation for the absolute difference between the real and the estimated proportion of true nulls (π_0), as well as the integrated absolute error committed by the ROC curve estimator proposed in (7). In addition, information about the committed errors (measured by $E[\hat{T}] = |\hat{T} - T|/T$) for $T_{\mathrm{BH}}(0.05)$ and $T_{\mathrm{SGoF}}(0.05)$ are also reported. For $N = 1,000$, the observed results are disappointing; both the mean and the standard deviation of the ROC curve estimates were really large. However they decrease for $N = 10,000$. It is worth to notice that the observed results in the second scenario were a bit better than the previous ones but, in any case, really similar to them.

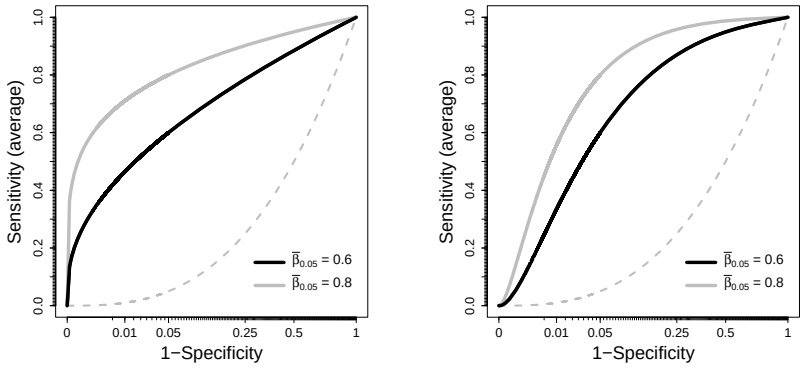

Fig. 2 Left, ROC curve under the mixture model ($F_1(t) = (1/N_1) \cdot \sum_{i=1}^{N_1} t^{a_i} \cdot \mathbb{I}_{[0,1]}(t)$ with $a_i = L + (i/N_1) \cdot U$ and $L = 0.05$) when p-values are obtained from the first described scenario. Right, ROC curve when p-values are obtained from the second described scenario.

Table 2 Mean ± standard deviation for the absolute difference between π_0 and $\hat{\pi}_{0,2}$; the integrated absolute difference between the ROC curve estimation and its target; the errors $E[\hat{T}_{\mathrm{BH}}(0.05)]$ and $E[\hat{T}_{\mathrm{SGoG}}(0.05)]$ obtained from 1,000 Monte Carlo iterations for both described scenarios.

N	π_0	$\bar{\beta}_{0.05}$	$\lvert\hat{\pi}_{0,2} - \pi_0\rvert$	$\int \lvert\hat{\mathcal{R}} - \mathcal{R}\rvert$	$E[\hat{T}_{\mathrm{BH}}(0.05)]$	$E[\hat{T}_{\mathrm{SGoF}}(0.05)]$
Scenario I						
1,000	0.95	0.60	0.056 ± 0.04	0.178 ± 0.09	0.190 ± 0.14	1.871 ± 2.21
		0.80	0.054 ± 0.04	0.182 ± 0.14	0.103 ± 0.07	4.287 ± 7.61
1,000	0.85	0.60	0.059 ± 0.05	0.109 ± 0.07	0.093 ± 0.07	0.350 ± 0.26
		0.80	0.052 ± 0.04	0.074 ± 0.06	0.052 ± 0.04	0.546 ± 0.44
10,000	0.95	0.60	0.019 ± 0.01	0.109 ± 0.06	0.056 ± 0.04	0.257 ± 0.19
		0.80	0.018 ± 0.01	0.075 ± 0.06	0.032 ± 0.02	0.381 ± 0.29
10,000	0.85	0.60	0.032 ± 0.02	0.088 ± 0.04	0.031 ± 0.02	0.091 ± 0.07
		0.80	0.020 ± 0.02	0.045 ± 0.02	0.016 ± 0.01	0.130 ± 0.10
Scenario II						
1,000	0.95	0.60	0.054 ± 0.04	0.168 ± 0.11	2.404 ± 2.50	0.520 ± 0.40
		0.80	0.056 ± 0.04	0.189 ± 0.15	0.446 ± 0.33	0.450 ± 0.35
1,000	0.85	0.60	0.055 ± 0.04	0.082 ± 0.06	0.455 ± 0.33	0.158 ± 0.12
		0.80	0.055 ± 0.04	0.069 ± 0.07	0.137 ± 0.11	0.164 ± 0.12
10,000	0.95	0.60	0.018 ± 0.01	0.080 ± 0.05	0.865 ± 0.70	0.114 ± 0.08
		0.80	0.017 ± 0.01	0.063 ± 0.06	0.150 ± 0.11	0.112 ± 0.08
10,000	0.85	0.60	0.020 ± 0.02	0.044 ± 0.02	0.159 ± 0.12	0.044 ± 0.03
		0.80	0.015 ± 0.01	0.027 ± 0.02	0.043 ± 0.03	0.047 ± 0.03

5 A real-world example: the Shen data

The Shen data contains information of 62 Taiwanese cases of hepatocellular carcinoma (HCC) on which tumor and adjacent non-tumor tissues were analyzed using Illumina methylation arrays (Illumina, Inc., San Diego, CA) that screen 26,538 autosomal CpG sites. The reader is referred to Shen *et al.* [17] for a complete information about the original study. The data are publicly available at the Gene Expression Omnibus (GEO) page, with access number GSE37988 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37988). Probes corresponding to the X and Y chromosomes were removed from the dataset in order to eliminate the X-inactivation effects. We considered the raw data (without any previous quality control). Due to the fact that each CpG is measured on the same subjects, the parametric Student t-test or the non-parametric Wilcoxon test for paired samples can be used in order to check the null of equality between, let us abuse, distributions.

The total number of CpG sites with a p-value less than 0.05 (usual nominal level) was 12,394 (46.7%) using the t-test, and 12,592 (47.4%) using the Wilcoxon test. Figure 3 shows the p-value histograms and $\mathbb{G}_N$ function estimates for the Student (black) and the Wilcoxon (gray) test. Table 3 shows estimations of π_0 using different k-values of the estimator defined in (6). Results are influenced by the p-values close to 1, specially for largest k.

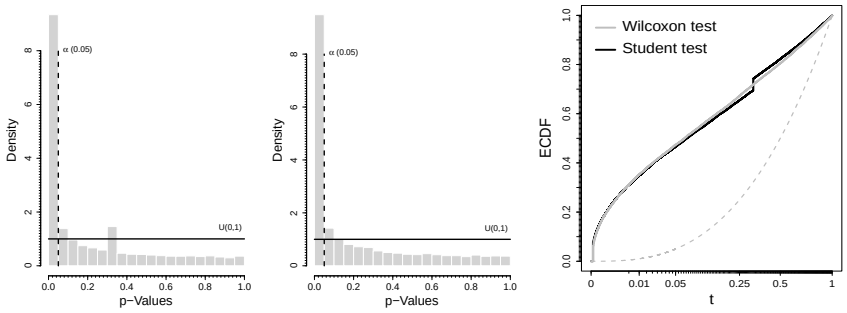

Fig. 3 Histogram for the p-values obtained from the t (left) and Wilcoxon (middle) tests, as well as $\hat{\mathbb{G}}_N(\cdot)$ for the t (black) and Wilcoxon (gray) tests.

	$\hat{\pi}_{0,1}$	$\hat{\pi}_{0,2}$	$\hat{\pi}_{0,3}$	$\hat{\pi}_{0,4}$
t-test	0.404	0.384	0.450	0.590
Wilcoxon test	0.424	0.427	0.541	0.791

Table 3 Values of $\hat{\pi}_{0,k}$ for different k-values.

The BH method, $\hat{T}_{BH}(0.05)$, declares significant the 10,451 smallest p-values (cut-off point of 0.019690) for the t-tests and the 10,585 smallest ones when the Wilcoxon test is used (cut-off point of 0.019943). At these points,

the estimated average sensibilities (using $\hat{\pi}_{0,2}$) were 0.627 and 0.681, for the t and the Wilcoxon test, respectively.

Figure 4 depicts the estimated ROC curves for the Student (left) and the Wilcoxon (right) test for the different considered π_0 estimates. Obviously, the results depend on the $\hat{\pi}_{0,k}$ value; however, it should be recalled that the x-axis scale was altered and the real difference is not as large as it seems.

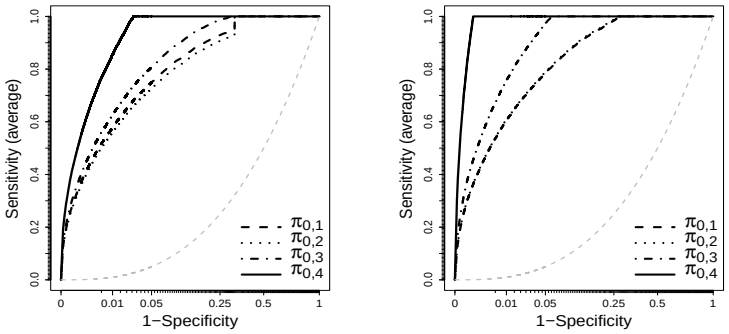

Fig. 4 ROC curve estimations for different $\hat{\pi}_{0,k}$ ($k \in \{1, 2, 3, 4\}$) for t-test (left) and Wilcoxon test (right).

6 Discussion

Currently, a number of researchers, mainly bioinformaticians, must often deal with multitesting problems. As a consequence, the development of adequate statistical tools in order to handle and control the involved Type I and Type II errors is a really hot topic in the specialized literature. Although some authors as Genovese and Wasserman [9] or Storey [18] have already considered the false non-discovery rate, most of the works are focused on trying to control, in some way, the false positive rate. However, the first quantity is crucial in order to know what the real capacity of detecting true effects is. Notice that, even when we know the exact number of false nulls, the p-value could not be an appropriate measure for separating those from the true nulls.

Actually, this work does not propose real practical solutions but it pays attention to an usually forgotten aspect of the multitesting problem. It intends to ponder the technical limitations which this complex issue provokes. Remember that, in most cases, we only have one sample drawn from an N-dimensional vector. By assuming independence among tests, one can perform some kind of correct inference; furthermore, the independence assumption is reasonable in a number of practical situations (see Clarke and Hall [3]), but it may be a source of serious mistakes and misleading conclusions. In the pres-

ence of arbitrary correlation structures, the limitations are clear. The main problem lies in the variability of the observed proportion of rejections under the null; while in the independent case, in the usual practical problems, this number is really close to the fixed nominal level, under dependent structures it can vary a lot (see Martínez-Camblor [14]) and both the π_0 and the $\mathcal{R}(\cdot)$ estimates are strongly dependent on that observed value.

The simulation studies show the limitation capacity to perform a correct estimation of the ROC curve in the multitesting context. In addition, the presence of p-values close to the extremes (zero and one) provokes precision problems in the obtained estimates; unfortunately, when the number of tests is large (most frequent case), this problem is not unusual. The accuracy problem, frequently ignored, gains relevance when the selected cut-off point strongly depends on the fifth or sixth decimal position.

In this report, we explore the interpretation of the sensibility and specificity in the multitesting problem by assuming the mixture model and, in this context, a ROC curve estimator is proposed. The explored methods allow us to give an estimation of the sensitivity average for each particular problem. Even taking into account the limitations of the procedures, in each particular problem, this quantity can help to have a better understanding of what the actual capacity of p-values to distinguish false from true null hypotheses is.

Acknowledgements The authors acknowledge support by the Grants MTM2014-55966-P and MTM2015-63971-P from the Spanish Ministerio of Economía y Competitividad and by BP16118 and FC-15-GRUPIN14-101 from the Principality of Asturias.

References

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statist Soc Ser B 57(1):289–300
2. Carvajal-Rodríguez A, de Uña-Álvarez J, Rolan-Álvarez E (2009) A new multitest correction (SGoF) that increase its statistical power when increasing the number of test. BMC Bioinformatics 10:209
3. Clarke S, Hall P (2009) Robustness of multiple testing procedures against dependence. Ann Statist 37(1):332–358
4. Dalmasso C, Broët P, Moreau T (2005) A simple procedure for estimating the false discovery rate. Bioinformatics 21(5:660–668
5. de Uña-Álvarez J (2011) On the statistical properties of SGoF multitesting method. Statist Appl Genet Molec Biol 10(1):Art18
6. Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Amer Statist Assoc 96:1151–1160
7. Faraggi D, Reiser B (2002) Estimation of the area under the ROC curve. Statist Med 21(20):3093–3106
8. Farcomeni A (2008) A Review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Statist Meth Med Res 17:347–388

9. Genovese CR, Wasserman L (2002) Operating characteristics and extensions of the procedure. J Royal Statist Soc Ser B 64:499–517
10. Genovese CR, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Statist 32(3):1035–1061
11. Genovese CR, Wasserman L (2006) Exceedance control of the false discovery proportion. J Amer Statist Assoc 201:1408–1417
12. Ibarrola P, Pardo L, Quesada V (1997) Teoría de la Probabilidad. Síntesis, Madrid.
13. Ma H, Bandos AI, Rockette HE, Gur D (2013) On use of partial area under the ROC curve for evaluation of diagnostic performance. Statist Med 32(20):3449–3458
14. Martínez-Camblor P (2014) On correlated z-values distribution in hypothesis testing. Comp Statist Data Anal 79:30–43
15. Sarkar SK (2007) Stepup procedures controlling generalized FWER and generalized FDR. Ann Statist 35(6):2405–2420
16. Seeger PA (1968) A note on a method for the analysis of significance *en masse*. Technometrics 10:586–593
17. Shen J, Wang S, Zhang YJ, Kappil M, Wu HC, Kibriya MG, Wang Q, Jasmine F, Ahsan H, Lee PH, Yu MW, Chen CJ, Santella RM (2012) Genome-wide DNA methylation profiles in hepatocellular carcinoma. Hepatology 55(6):1799–1808
18. Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Statist 31:2013–2035
19. Zhou X-H, Obuchowski NA, McClish DK (2002) Statistical Methods in Diagnostic Medicine. Wiley, New York