Universidade Vigo

Adjusted p-values for SGoF multitesting method

Irene Castro-Conde and Jacobo de Uña-Álvarez

Report 13/05

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade Vigo

Adjusted p-values for SGoF multitesting method

Irene Castro-Conde and Jacobo de Uña-Álvarez

Report 13/05

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: Universidade Vigo

Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Adjusted p-values for SGoF multitesting method

Irene Castro-Conde and Jacobo de Uña-Álvarez

December 2, 2013

Abstract

In the field of multiple comparison procedures, adjusted p-values are an important tool to evaluate the significance of a test statistic while taking the multiplicity issue into account. In this paper we introduce adjusted p-values for Carvajal-Rodríguez et al. (2009)'s Sequential Goodness-of-Fit (SGoF) multitesting method. Main properties of the adjusted p-values are established. Several real data applications are performed to illustrate the practical usage of the adjusted p-values.

1 Introduction

In multiple hypotheses testing, a number of null hypotheses (or nulls) is tested in a simultaneous way, so some multiple significance criterion is used to control the error probability along the tests being performed. Traditional multiple testing procedures control for the familywise error rate (FWER) or for the false discovery rate (FDR), or for proper modifications and generalizations of these criteria. Usually strong control of the FWER or of the FDR is demanded, which means that the error criterion must be fulfilled under any configuration of the true and non-true nulls (Dudoit and van der Laan, 2008). In many situations, such methods exhibit a poor power, leading to a small amount of rejected nulls (Carvajal-Rodríguez et al., 2009). Weak control of the FWER allow for a greater power, at the extent of reporting bounds for FWER or FDR which are only valid when all the nulls are true (intersection or complete null). One of such methods is the Sequential Goodness-of-Fit (SGoF) procedure, introduced in Carvajal-Rodríguez et al. (2009). SGoF method looks for significance when comparing the observed and the expected amounts of p-values below an initial threshold γ, where the expectation is taken under the intersection or complete null. In that sense, it relates to the notion of second-level significance testing or higher criticism introduced by Tukey in 1976 and further extended by Donoho and Jin (2004).

Main properties of SGoF procedure were analyzed in more detail by de Uña-Álvarez (2011, 2012). In particular, it was quoted that SGoF gives flexibility to the FDR, by imposing a bound (α) merely under the complete null, with the resulting increase in power; and that the attained FDR is not too large in

the sense that the number of false positives remains smaller than the number of false negatives with large probability ($\geq 1-\alpha$). The goal of this paper is to introduce adjusted p-values for SGoF multitesting procedure. To this end, we will recall the precise definition of SGoF method, for which some notation is needed.

Given n nulls $H_{0i}, i = 1, ..., n$, let $p_1, ..., p_n$ be a sequence of independent p-values corresponding to the application n specific test statistics. Put F_n for the empirical distribution function of the p_i's. Let γ be an initial significance threshold. SGoF procedure looks for significance in the amount of p-values falling below γ, $nF_n(\gamma)$, at level α; if this amount is 'too large', the p-values are smaller than expected, and the complete null $H_0 = \cap_{i=1}^{n} H_{0i}$ is rejected. Under H_0, $nF_n(\gamma)$ follows a $Bin(n, \gamma)$ distribution; therefore, H_0 is rejected at level α if and only if $nF_n(\gamma) \geq b_{n,\alpha}(\gamma)$, where

$$b_{n,\alpha}(\gamma) = \inf \{b \in \{0, ..., n\} : P(Bin(n, \gamma) \geq b) \leq \alpha\}$$

is the $(1-\alpha)$-quantile of the $Bin(n, \gamma)$ distribution. When H_0 is rejected, the number of nulls rejected by SGoF is $N_{n,\alpha}(\gamma) = nF_n(\gamma) - b_{n,\alpha}(\gamma) + 1$. More specifically, the nulls corresponding to the $N_{n,\alpha}(\gamma)$ smallest p-values are declared as non-true; this means that a given p-value p_i is declared as a positive if and only if its rank is smaller than $N_{n,\alpha}(\gamma)$, that is, if $nF_n(p_i) \leq N_{n,\alpha}(\gamma)$ (assuming no ties). In the presence of ties among the p-values, $N_{n,\alpha}(\gamma)$ should only be regarded as an upper bound for the number of rejections (see the example in Section 4.1 for illustration of this issue).

Let p^* denote the maximum p-value declared as a positive by SGoF, that is, $N_{n,\alpha}(\gamma) = nF_n(p^*)$; since $b_{n,\alpha}(\gamma) \geq 1$ (unless $\alpha = 1$), we have $N_{n,\alpha}(\gamma) \leq nF_n(\gamma)$ and, from this, $p^* \leq \gamma$. This shows that the rejected p-values form a subset (proper or not) of the set of p-values falling below γ. Indeed, it becomes clear from the definition of $N_{n,\alpha}(\gamma)$ that the amount of p-values rejected by SGoF corresponds to the excess of significant cases in the binomial test. When n is large, the binomial quantile $b_{n,\alpha}(\gamma)$ is approximated by $n\gamma + \sqrt{n\gamma(1-\gamma)}z_\alpha$, where z_α is the $1-\alpha$ quantile of the standard normal; de Uña-Álvarez (2011) suggested as a more conservative rule to reject the $N^*_{n,\alpha}(\gamma) = nF_n(\gamma) - b^*_{n,\alpha}(\gamma) + 1$ nulls attached to the smallest p-values, where $b^*_{n,\alpha}(\gamma) = n\gamma + \sqrt{nF_n(\gamma)(1-F_n(\gamma))}z_\alpha$. $N^*_{n,\alpha}(\gamma)$ gives an asymptotic $(1-\alpha)$-lower confidence bound for the number of non-true nulls with p-value below γ. In its turn, this ensures asymptotically that the probability of the undesirable event that the number of false positives exceeds the number of false negatives (among the p-values smaller than γ) is bounded by α. This property of SGoF method is not shared by other commonly used multitesting procedures such as those based on the FDR or FWER control. See de Uña-Álvarez (2012) for more details.

In this paper, adjusted p-values for SGoF multitesting method are introduced. For this, we do not take the adjusted p-value of p_i as the minimum α for which SGoF procedure rejects p_i for the given γ, as implemented in SGoF+ software (http://webs.uvigo.es/acraaj/SGoF.htm). This, although possible, does not eliminate the dependency of the adjusted p-value on the initial threshold γ,

something which may not be desirable. For example, with $n = 1$ the adjusted p-value of p_1 corresponding to this definition gives γ if $p_1 \leq \gamma$ and 1 otherwise. With a single null to be tested ($n = 1$), one would expect that the adjusted p-value (for the multiplicity of tests) would give just the original p-value, something which is not achieved in this way. In the following section we define adjusted p-values for SGoF multitesting method satisfying such a requirement, by taking the γ and α parameters to be equal. In this way, the provided definition avoids the problem of dependencies on γ. Overall, adjusted p-values as defined here are a useful tool when looking for significance under the viewpoint of Carvajal-Rodríguez et al. (2009)'s Sequential Goodness-of-Fit Test.

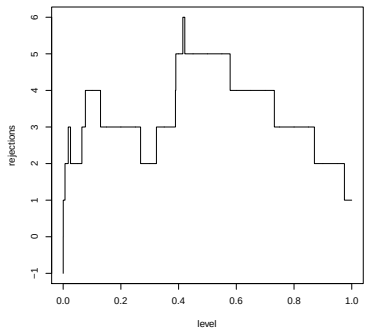
The paper is organized as follows. In Section 2, we introduce the definition of the adjusted p-values of SGoF multitest and we give several examples. The main properties of these adjusted p-values are given in Section 3 while in Section 4 we show three real data applications to illustrate their practical usage. Finally, in Section 5 we give the main conclusions of our research. Technical Lemmas are deferred to the Appendix.

2 SGoF adjusted p-values

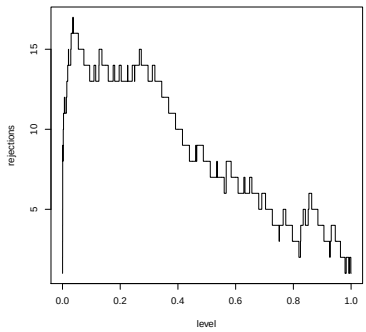
Let γ be an initial p-value threshold. Consider SGoF multitesting criterion based on the special choice $\alpha = \gamma$. This can be regarded as a fair application of SGoF in which no prominent role is given to any of the two significance thresholds. We term this value $\alpha = \gamma$ as the level of the test. For simplicity, we denote $b_{n,\alpha}(\alpha)$ and $N_{n,\alpha}(\alpha)$ by $b_n(\alpha)$ and $N_n(\alpha)$ respectively. In practice, the number of nulls rejected by SGoF $N_n(\alpha) = nF_n(\alpha) - b_n(\alpha) + 1$ is roughly an increasing-decreasing function of α; this is because the amount of rejections crucially depends on the distance between the observed and the expected proportions of p-values falling below $\alpha = \gamma$, $F_n(\alpha) - \alpha$, which has an increasing-decreasing shape. For illustration purposes, in Figure 1 we depict this function $N_n(\alpha)$ for particular simulated sequences of $n = 10$, 50 and 100 p-values. In this Figure 1, α is restricted to fall in the set of p-values. The p-values come from the application of the two-sided Student's test to compare samples of 7 and 8 individuals (i.e. 13 degrees of freedom) drawn from Gaussian populations, along n positions ('genes') for each individual, with respective means 0 (group 1) and μ (group 2), where $\mu = 0$ for the true nulls and $\mu = \pm 2$ for the non-true nulls, representing intermediate effects, and where the variance-covariance matrix is the identity. The percentage of non-true nulls in the simulations is 33%. We note that, unlike for SGoF method with fixed γ, when linking α and γ through $\alpha = \gamma$ an increasing value of α does not necessarily result in a less stringent multitest. This will be important when interpreting the adjusted p-values to be introduced.

It is informative to look at the function $N_n(\alpha)$ for small values of n. We consider as particular examples the cases with $n \leq 3$. When needed, the ordered p-values are denoted by $p_{(1)} \leq ... \leq p_{(n)}$.

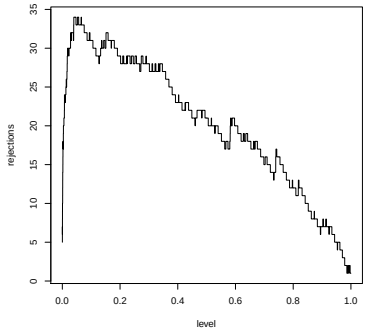
Example 1. The case $n = 1$. Clearly, $b_n(\alpha) = 1$ in this case (unless



(a) $n = 10$ simulated p-values



(b) $n = 50$ simulated p-values



(c) $n = 100$ simulated p-values

Figure 1: Simulated function of the number of rejections of SGoF ($N_n(\alpha)$) in three different scenarios

$\alpha = 1$). Therefore, the (single) null hypothesis is rejected by SGoF if and only if $nF_n(\alpha) \geq 1$, that is, when $p_1 \leq \alpha$. Note that this is just the ordinary way in which a single test is performed at level α. Obviously, in case of rejection, $N_n(\alpha) = 1$.

Example 2. The case $n = 2$. Since $P(Bin(2, \alpha) \geq 2) = \alpha^2 \leq \alpha$ and $P(Bin(2, \alpha) \geq 1) = 2\alpha(1 - \alpha) + \alpha^2 > \alpha$, we have $b_n(\alpha) = 2$ in this case. Then, SGoF rejects the intersection null if and only if both p-values fall below α; in that case, the number of rejected nulls is $N_n(\alpha) = nF_n(\alpha) - b_n(\alpha) + 1 = 1$, so SGoF rejects the null attached to the smallest p-value $p_{(1)}$. As a function of α, $N_n(\alpha)$ takes the value 0 for $0 < \alpha < p_{(2)}$, and the value 1 for $p_{(2)} \leq \alpha < 1$.

Example 3. The case $n = 3$. Straightforward calculations give that $b_n(\alpha) = 3$ if $\alpha > 0.5$, while $b_n(\alpha) = 2$ if $\alpha \leq 0.5$. To help the discussion, assume at this point that $\alpha \leq 0.5$ (this will be the case in most practical cases). Then, the intersection null is rejected if and only if $p_{(2)} \leq \alpha$ and, if this happens, $N_n(\alpha) = 2$

or 1 depending on whether $p_{(3)} \leq \alpha$ or $p_{(3)} > \alpha$.

As mentioned, compared to FWER and FDR controlling procedures, SGoF method exhibits a greater power in many instances, particularly when the number of tests is large. However, it has been pointed out that FDR-based methods may be more powerful than SGoF when the p-values concentrate around zero (the so-called situation with strong effects). To explore this point with small n, consider the application of Benjamini-Hochberg (1995) FDR method (BH) to the case $n = 3$. To bound the FDR by α, BH rejects the nulls correponding to the p-values below $p^*_{BH} = \max\{p_i : p_i \leq \alpha F_n(p_i)\}$. With $n = 3$, this implies checking if $p_{(3)} \leq \alpha$ (3 rejected nulls), $p_{(3)} > \alpha$ but $p_{(2)} \leq \alpha/2$ (2 rejections), $p_{(3)} > \alpha, p_{(2)} > \alpha/2$ but $p_{(1)} \leq \alpha/3$ (1 rejection), or if $p_{(i)} > \alpha(i/n)$ for $i = 1, 2, 3$ (no rejection). Summarizing, when $p_{(3)} \leq \alpha$ or when $p_{(2)} > \alpha$ SGoF will reject a smaller amount of nulls compared to BH (2 vs 3 or 0 vs 0-1 respectively), but when $p_{(2)} \leq \alpha < p_{(3)}$ the opposite (namely 1 vs. 0) may happen. For example, if the sequence of p-values is 0.02, 0.04, 0.06, then BH is unable to reject any null when controlling the FDR at 5%, while SGoF applied with $\gamma = \alpha = 0.05$ rejects the null pertaining to the smallest p-value. SGoF finds significance in this case because the amount of p-values falling below 0.05 reaches the 5% critical point of the one-sided binomial test ($b_n(\alpha) = 2$).

Now we formally introduce the adjusted p-values for SGoF.

Definition 1. Let $p_1, ..., p_n$ be a sequence of p-values corresponding to n nulls $H_{01}, ..., H_{0n}$. The SGoF adjusted p-value of p_i $(1 \leq i \leq n)$ is defined as

$$\widetilde{p}_i = \inf\{\alpha \in (0, 1) : nF_n(p_i) \leq N_n(\alpha)\}$$

if the set $\{\alpha \in (0, 1) : nF_n(p_i) \leq N_n(\alpha)\}$ is not empty. Otherwise, $\widetilde{p}_i = 1$.

Remark 1. The open interval $(0, 1)$ in the Definition may be replaced by the close interval $[0, 1]$ as long as the smallest p-value is strictly positive.

In words, the adjusted p-value of p_i is the smallest α for which SGoF multitest with $\alpha = \gamma$ rejects the null corresponding to p_i. For illustration, we give now the adjusted p-values for Examples 1-3.

Example 1 (continued). In the case $n = 1$, SGoF rejects the (single) null hypothesis at level α if and only if $p_1 \leq \alpha$. Therefore, $\widetilde{p}_1 = p_1$.

Example 2 (continued). In the case $n = 2$, SGoF rejects at most $p_{(1)}$, and this only happens when both p-values fall below the level α. Therefore, $\widetilde{p}_{(1)} = p_{(2)}$ and $\widetilde{p}_{(2)} = 1$.

Example 3 (continued). The case $n = 3$ is more involved. For $\alpha > 0.5$ we have $N_n(\alpha) = nF_n(\alpha) - 2$, while for $\alpha \leq 0.5$ we have $N_n(\alpha) = nF_n(\alpha) - 1$. This implies that at most two nulls will be rejected, and therefore $\widetilde{p}_{(3)} = 1$. If $p_{(3)} > 0.5$ one has $nF_n(\alpha) \leq 2$ for $\alpha \leq 0.5$ and hence $p_{(2)}$ is not rejected at any level; so $\widetilde{p}_{(2)} = 1$ in this case. However, when $p_{(3)} \leq 0.5$, one rather has $\widetilde{p}_{(2)} = p_{(3)}$. Finally, $\widetilde{p}_{(1)} = p_{(2)}$ or $\widetilde{p}_{(1)} = p_{(3)}$ depending on whether $p_{(2)} \leq 0.5$ or not. As a particular example, consider as above the sequence of p-values 0.02, 0.04, 0.06; the corresponding sequence of adjusted p-values for SGoF is 0.04, 0.06, 1.

Main properties of the adjusted p-values for arbitrary n are given in the next Section.

3 Main results

The first desirable property of adjusted p-values for multiple hypotheses testing is that they generalize the concept of p-value for a single test. SGoF adjusted p-values have this property, as discussed in Section 2. Another important property is that adjusted p-values should be greater than the corresponding original p-values, since they are introduced to protect the researcher against the multiplicity of tests. For SGoF adjusted p-values this is formally stated as follows.

Property 1. It holds $\widetilde{p}_i \geq p_i$ for $i = 1, ..., n$.

Proof: If there is no α such that $nF_n(p_i) \leq N_n(\alpha)$ then $\widetilde{p}_i = 1$ and the result holds. Therefore, assume that $\widetilde{p}_i = \inf\{\alpha \in (0,1) : nF_n(p_i) \leq N_n(\alpha)\}$. Note first that $N_n(\alpha) \leq nF_n(\alpha)$ (as discussed in the Introduction). We thus have $nF_n(p_i) \leq N_n(\widetilde{p}_i) \leq nF_n(\widetilde{p}_i)$, where for the first inequality we use Lemma 2 in the Appendix. Since p_i is a jump point of F_n, the result follows.□

The next property states that adjusted p-values correspond to a (non strictly) monotone transformation of the original p-values.

Property 2. If $p_i > p_j$ then $\widetilde{p}_i \geq \widetilde{p}_j$.

Proof: Take two p-values such that $p_i > p_j$. Assume $\widetilde{p}_i < 1$ (otherwise there is nothing to prove). We will show that $nF_n(p_j) \leq N_n(\widetilde{p}_i)$, which is enough to conclude. But this follows from $nF_n(p_j) \leq nF_n(p_i) \leq N_n(\widetilde{p}_i)$.□

Given a significance threshold $\alpha \in (0,1)$, a natural question is if the number of nulls rejected by SGoF method at level α (i.e. $N_n(\alpha)$) coincides with the amount of adjusted p-values not greater than α. This will not be the case in general, since $N_n(\alpha)$ has an increasing-decreasing shape as a function of α and hence, rejection of p_i at some level α' does not ensure rejection of that p-value for $\alpha \geq \alpha'$. Still, the number of $\widetilde{p}_i$'s not greater than α is an upper bound for $N_n(\alpha)$. This property is formally given below.

Property 3. It holds

$$N_n(\alpha) \leq \sum_{i=1}^{n} I(\widetilde{p}_i \leq \alpha).$$

On the other hand, if $\widetilde{p}_i \leq \alpha$, then there exists $\alpha' \leq \alpha$ such that the null attached to p_i is rejected at level α' by SGoF.

Proof. If the null attached to p_i is rejected at level α by SGoF, then $nF_n(p_i) \leq N_n(\alpha)$. Therefore, $\{\alpha' \in (0,1) : nF_n(p_i) \leq N_n(\alpha')\}$ is non-empty, and $\widetilde{p}_i = \inf\{\alpha' \in (0,1) : nF_n(p_i) \leq N_n(\alpha')\} \leq \alpha$. This shows the given inequality. Conversely, if $\widetilde{p}_i \leq \alpha$, then $nF_n(p_i) \leq N_n(\alpha')$ for some $\alpha' \leq \alpha$ and hence p_i is rejected at level α' by SGoF.□

Remark 2. Unlike for Properties 1 and 2, Property 3 may fail in the presence of ties. When ties are present, it is desirable to prevent the researcher from making different decisions for null hypotheses sharing the same (tied) p-value. To this end, the corrected amount of rejections given by $\widetilde{N}_n(\alpha) = \min\{N_n(\alpha), nF_n(q_n(\alpha)^-)\}$, where $q_n(\alpha) = F_n^{-1}(n^{-1}N_n(\alpha))$, is suggested. Here, $F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$ denotes the empirical quantile function. Property 3 is satisfied even in the presence of ties with such prevention. Note that the

Definition 1 of adjusted p-value is unchanged when $N_n(\alpha)$ is replaced by $\widetilde{N}_n(\alpha)$. On the other hand, $\widetilde{N}_n(\alpha)$ collapses to $N_n(\alpha)$ when there are no ties. See the example in Section 4.1 for illustration of this correction.

An interesting property of the adjusted p-values as defined here is that they fall within the set of original p-values or they take the value 1. This is not obvious from the definition; however, the property can be obtained from the fact that the function $\alpha \mapsto b_n(\alpha)$ is non-decreasing (Lemma 2 in Appendix A).

Property 4. It holds

$$\widetilde{p}_i = \min \{p_j : nF_n(p_i) \leq N_n(p_j)\},$$

with the convention 1 if the set is empty.

Proof. If $\widetilde{p}_i = 1$ then both $\{\alpha \in (0,1) : nF_n(p_i) \leq N_n(\alpha)\}$ and $\{p_j : nF_n(p_i) \leq N_n(p_j)\}$ are empty sets, and the equality follows by the convention. Since $\widetilde{p}_i \geq p_i$ (Property 1), $\widetilde{p}_i$ is larger than some p-value. Assume $p_{(k)} \leq \widetilde{p}_i < p_{(k+1)}$ for some $k \in \{1, ..., n\}$, where $p_{(n+1)} \equiv 1$. We will show that $\widetilde{p}_i = p_{(k)}$. We have $nF_n(p_i) \leq N_n(\widetilde{p}_i) = nF_n(\widetilde{p}_i) - b_n(\widetilde{p}_i) + 1$. Since $b_n(\alpha)$ is non-decreasing and $F_n(\widetilde{p}_i) = F_n(p_{(k)})$, we have $N_n(\widetilde{p}_i) \leq N_n(p_{(k)})$, and the proof is complete.□

Property 4 is useful for the calculation and implementation of the adjusted p-values, since the search for the infimum along a continuous interval may be restricted to a finite set (the original p-values themselves). Also, since some of the adjusted p-values will take the value 1, it is relevant from a practical viewpoint to identify them in a first step. A characterization of such p-values is given in the following statement.

Property 5. Let $N^{[n]} = \max_{1 \leq j \leq n, p_j < 1} N_n(p_j)$, and let $p^{[n]} = \max \left\{p_j : nF_n(p_j) \leq N^{[n]}\right\}$. Then, $p_i > p^{[n]}$ if and only if $\widetilde{p}_i = 1$.

Proof. Take $p_i > p^{[n]}$. Then, $nF_n(p_i) > N^{[n]} \geq N_n(p_j)$ for all $p_j < 1$. This entails that $\{p_j : nF_n(p_i) \leq N_n(p_j)\}$ is an empty set (or it reduces to $\{1\}$) and therefore $\widetilde{p}_i = 1$. Conversely, if $\widetilde{p}_i = 1$ then the set $\{p_j : nF_n(p_i) \leq N_n(p_j)\}$ is empty (Case 1), or $\{p_j : nF_n(p_i) \leq N_n(p_j)\} = \{1\}$ (Case 2). In Case 1, we have that $nF_n(p_i) > N^{[n]}$ and therefore $p_i > p^{[n]}$. In Case 2, if we assume $p_i \leq p^{[n]}$ then we have $nF_n(p_i) \leq N^{[n]}$ and, therefore, there exists $p_j < 1$ such that $nF_n(p_i) \leq N_n(p_j)$, which is a contradiction, and the proof is complete.□

Remark 3. If there are no ties, we have $p^{[n]} = p_{(N^{[n]})}$. When ties are present, one needs to change $N_n(\alpha)$ by its correction $\widetilde{N}_n(\alpha)$ (see Remark 2) and, by defining $\widetilde{N}^{[n]}$ and $\widetilde{p}^{[n]}$ similarly to $N^{[n]}$ and $p^{[n]}$, the Property 5 still holds and $p^{[n]} = p_{(\widetilde{N}^{[n]})}$. This identity will allow us to identify this threshold p-value easily. Example in Section 4.1 illustrates the differences between N_n and $\widetilde{N}_n$ in the presence of ties. We recall that $N_n(\alpha)$ and $\widetilde{N}_n(\alpha)$ will report the same value when there are no ties, unless for $\alpha = 1$ (see Section 4.2 for an illustration of this situation).

As we mentioned in the Introduction, a conservative version of the SGoF metatest was proposed in de Uña-Álvarez (2011) when n is large. This conservative rule rejects the $N^*_{n,\alpha}(\gamma) = nF_n(\gamma) - b^*_{n,\alpha}(\gamma) + 1$ nulls attached to the smallest p-values, where $b^*_{n,\alpha}(\gamma) = n\gamma + \sqrt{nF_n(\gamma)(1 - F_n(\gamma))}z_\alpha$. Adjusted p-values for Conservative SGoF may be introduced as for original SGoF. For this,

we take $\alpha = \gamma$ as above and we denote $b^*_{n,\alpha}(\alpha)$ and $N^*_{n,\alpha}(\alpha)$ by $b^*_n(\alpha)$ and $N^*_n(\alpha)$ respectively.

Definition 2. Let $p_1, ..., p_n$ be a sequence of p-values corresponding to n nulls $H_{01}, ..., H_{0n}$. The conservative SGoF adjusted p-value of p_i $(1 \leq i \leq n)$ is defined as

$$\widetilde{p}^*_i = \inf \{\alpha \in (0,1) : nF_n(p_i) \leq N^*_n(\alpha)\}$$

if the set $\{\alpha \in (0,1) : nF_n(p_i) \leq N^*_n(\alpha)\}$ is not empty. Otherwise, $\widetilde{p}^*_i = 1$.

Properties of $\tilde{p}^*_i$ can be derived similarly as for $\tilde{p}_i$. This is clear for Properties 2 and 3, where the arguments used for $\tilde{p}_i$ are still valid. However, since the function $\alpha \to b^*_n(\alpha)$ is not necessarily non-decreasing, some care is needed for Properties 1, 4 and 5. For example, in order to prove $\tilde{p}^*_i \geq p_i$, $i = 1, ..., n$ (Property 1), we note that $N^*_n(\alpha) \leq nF_n(\alpha)$ is not ensured in general and, therefore, one can not follow the steps in the proof for $\tilde{p}_i$. It can be seen, however, that $b^*_n(\alpha) > 0$ holds (and thus $N^*_n(\alpha) \leq nF_n(\alpha)$ follows) in most practical cases; more specifically, if the maximum p-value is smaller than 0.999, the result is valid for $n > 2$. When avoiding degenerated cases, Property 1 may be established for $\tilde{p}^*_i$ similarly as for $\tilde{p}_i$ by using the right-continuity of $N^*_n(\alpha)$. On the other hand, the formal derivation of Property 4 for $\tilde{p}^*_i$ is more involved (if true). Still, by using the fact that $\alpha \in [p_{(j)}, p_{(j+1)}) \to b^*_n(\alpha)$ is increasing (at least) when $1-\Phi(\sqrt{(log(2n/\pi))}) < p_{(j)} < p_{(j+1)} \leq 1-\Phi(-\sqrt{(log(2n/\pi))})$ (here Φ stands for the cumulative distribution function of a standard normal), one may argue that the discrete approximation given in Property 4 will work exactly also for $\tilde{p}^*_i$ (at least) when $\tilde{p}^*_i \in [p_{(j)}, p_{(j+1)})$ for some $p_{(j)}$ satisfying the given inequalities. This would exclude in principle an exact result for small and large adjusted p-values. An alternative to overcome this issue would be to re-define $\tilde{p}^*_i$ as $\min\{p_j : nF_n(p_i) \leq N^*_n(p_j)\}$; this is indeed the way in which adjusted p-values for conservative SGoF are implemented in the R (R Core team, 2013) package *sgof* (Castro-Conde and de Uña-Álvarez, 2013). The application of conservative SGoF is illustrated in the real data example of Section 4.3.

4 Examples of application

4.1 Needleman data

Needleman et al. (1979) studied the neuropsychologic effects of unidentified childhood exposure to lead by comparing various psychological and classroom performances between two groups of children differing in the lead level observed in their shed teeth. While there is no doubt that high levels of lead are harmful, Needleman's findings regarding exposure to low lead levels were controversial. Needleman's study was attacked on the ground of methodological flaws. One of the methodological flaws pointed out is control of multiplicity. Needleman et al. (1979) present three families of endpoints, and comment on the results of separate multiplicity adjustments within each family. Benjamini and Yekutieli (2011) discussed results of BH procedure for Needleman's data when controlling for multiplicity both separately and jointly. For illustration of SGoF multitesting

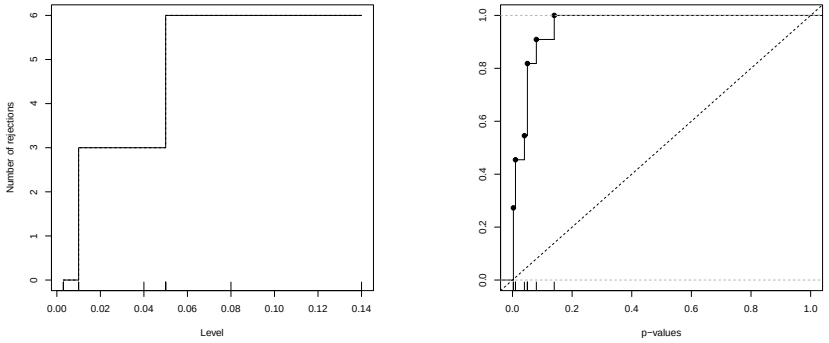
procedure, we will focus on the first family of endpoints, which corresponds to the teacher's behavioral ratings.

Table 1: Summary of the results obtained for Needleman data

		Adjusted p-values				
	p-value (p_i)	Hochberg	BH	SGoF	$N_n(p_i)$	$\widetilde{N}_n(p_i)$
Distractible	0.003	0.027	0.011	0.01	2	0
Does not follows sequence of directions	0.003	0.027	0.011	0.010	2	0
Low overall functioning	0.003	0.027	0.011	0.010	2	0
Impulsive	0.010	0.070	0.022	0.050	4	3
Daydreamer	0.010	0.070	0.022	0.050	4	3
Easily frustrated	0.040	0.140	0.061	0.050	4	3
Not persistent	0.050	0.140	0.061	1.000	7	6
Dependent	0.050	0.140	0.061	1.000	7	6
Does not follow simple directions	0.050	0.140	0.061	1.000	7	6
Hyperactive	0.080	0.140	0.088	1.000	8	6
Disorganized	0.140	0.140	0.140	1.000	8	6

Using Hochberg method (aimed to control the FWER; Hochberg, 1988) at 0.05 level, three hypotheses are rejected. By applying the BH procedure at 0.05 FDR level, two more nulls are rejected (see Table 1 in Benjamini and Yekutieli, 2011). On the other hand, SGoF is able to reject 7 null hypotheses when taking $\alpha = \gamma = 0.05$, but a correction of the number of rejections is needed because ties are present in this set of p-values. Note that it is not possible to reject exactly 7 p-values without making different decisions for the null hypotheses sharing the tied value $p_i = 0.05$. In order to avoid that, we compute the corrected number of rejections introduced in Remark 2 and we obtain $\widetilde{N}_n(\alpha) = 6$. Under this correction, the number of SGoF adjusted p-values smaller than $\alpha = 0.05$ is also 6 (note that, unlike for $\widetilde{N}_n(\alpha)$, $N_n(\alpha)$=7 violates Property 3 in this case due to ties). In this example, the largest SGoF adjusted p-value smaller than 1 is $p^{[n]} = \widetilde{p}_{(6)}$, where 6 corresponds to the maximum number of rejections coming from $\widetilde{N}_n$ (see Remark 3). Table 1 shows the adjusted p-values reported by Hochberg (computed with *MuToss* package, MuToss Coding Team et al., 2012), BH and SGoF procedures, together with the number of rejections (the original, N_n, and the correction for ties, $\widetilde{N}_n$) given by SGoF, when taking as level $\alpha = \gamma$ each original p-value p_i. It is seen that BH and SGoF adjusted p-values are smaller than those of Hochberg (only true for the six smallest p-values for SGoF), revealing the greater power of the former methods. On the other hand, SGoF adjusted p-values may be smaller or larger than those of the BH procedure. This indicates that, for this dataset, SGoF method may reject

an amount of nulls larger or smaller than that of BH depending on the level. For example, by inspection of the adjusted p-values, one may see that SGoF and BH reject 3 and 0 nulls respectively when taking $\alpha = 0.01$, while these figures change to 3 and 5 respectively when considering $\alpha = 0.025$.



(a) Corrected number of SGoF rejections $\widetilde{N}_n(\alpha)$ for $\alpha \in \{p_1, ..., p_n\}$.

(b) Empirical cumulative distribution function of the p-values (solid) and uniform distribution (dashed line).

Figure 2: Needleman data.

In Figure 2 we depict the (corrected) number of SGoF rejections $\widetilde{N}_n$ (left) and the empirical cumulative distribution function of the p-values (right). From this Figure 2, it can be seen that the number of rejections increases as the level grows (in the range $0 - 0.14$), according to the increasing deviations of $F_n(\alpha)$ with respect to the uniform distribution. This illustrates the way in which SGoF method looks for significance.

4.2 Neuhaus data

In second place, we will analyze a vector of $n = 15$ p-values obtained from Neuhaus, von Essen et al. (1992). In this paper, the effects of two different treatments for acute myocardial infarction (improved thrombolysis (rt-PA), which has been reported to yield higher patency rates than those achieved with standard regimens of thrombolytic treatment, and anisoylated plasminogen streptokinase activator (APSAC)) were investigated in a randomized multicenter trial in 421 patients. Four families of hypotheses were identified in that study. We will focus on the results obtained for the study of cardiac and other events after the start of thrombolitic treatment (15 hypotheses related to reinfarction, recurrent ischemia, blood pressure decrease, bleeding, allergic reaction, cardiogenic shock and in-hospital death). In the original paper, there is no attention to the problem of multiplicity of tests and the authors concluded that the improved rt-PA treatment is more favorable than ASPAC with fewer bleeding compli-

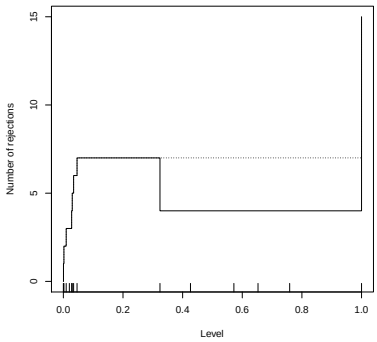
cations and a substantially lower in-hospital mortality rate. Full results are displayed in Table 2. Benjamini and Hochberg (1995) analyzed this dataset and they reached to the conclusion that Hochberg procedure at 5% level just rejects the 3 smallest p-values corresponding to allergic reaction and to two different aspects of bleeding, but no significance is found in the comparison of mortality. On the other hand, the BH method at FDR of 5% is able to reject one more null hypothesis which does indicate a mortality decrease (Table 2).

Table 2: Summary of the results obtained for Neuhaus data

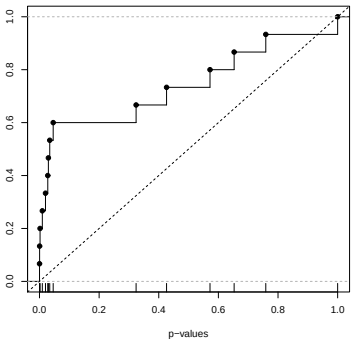
		Adjusted p-values				
	p-values (p_i)	Hochberg	BH	SGoF	$N_n(p_i)$	$\widetilde{N}_n(p_i)$
Allergic reaction 0 to 24 h	0.0001	0.0015	0.0015	0.0004	0	0
Bleeding puncture site	0.0004	0.0056	0.0030	0.0019	1	1
Bleeding overall	0.0019	0.0247	0.0095	0.0095	2	2
In hospital death	0.0095	0.1140	0.0356	0.0278	3	3
Bleeding transfusion	0.0201	0.2211	0.0603	0.0298	3	3
Cardiogenic shock 90 min to 48 h	0.0278	0.2682	0.0639	0.0344	4	4
Blood pressure decrease 0 to 90 min	0.0298	0.2682	0.0639	0.0459	5	5
In hospital death 0 to 48 h	0.0344	0.2752	0.0645	1.0000	6	6
Cardiogenic shock 0 to 90 min	0.0459	0.3213	0.0765	1.0000	7	7
Pericardial tamponade	0.3240	1.0000	0.4860	1.0000	4	4
Blood pressure decrease 90 min to 48 h	0.4263	1.0000	0.5813	1.0000	4	4
Cerebrovascular ischemia	0.5719	1.0000	0.7149	1.0000	4	4
Reinfarction	0.6528	1.0000	0.7532	1.0000	4	4
Bleeding recurrent ischemia	0.7590	1.0000	0.8132	1.0000	4	4
Bleeding cerebral	1.000	1.0000	1.0000	1.0000	16	15

If we apply SGoF ($\alpha = \gamma = 0.05$) to that sequence of 15 p-values we obtain 7 rejections which makes also significant another aspect of bleeding (bleeding transfusion), cardiogenic shock, and a blood pressure decrease. Moreover, the largest SGoF adjusted p-value smaller than 1 is $p^{[n]} = \widetilde{p}_{(7)}$, where 7 corresponds to $N^{[n]}$, the maximum number of rejections (excluding $p_i = 1$), coming from $N_n(\alpha)$ because, in this case, there are no ties in this data set (see Remark 3). Table 2 shows the adjusted p-values reported by Hochberg, BH and SGoF methods, and the number of rejections (the original, N_n, and the ones corrected for ties, $\widetilde{N}_n$) given by SGoF, when taking as level $\alpha = \gamma$ each original p-value p_i. Both sequences $N_n(p_i)$ and $\widetilde{N}_n(p_i)$ coincide but for the largest p-value, which takes the value 1. Note that the value of 16 reported by $N_n(1)$ makes no sense

since there are only 15 hypotheses under consideration.



(a) Solid line: Number of rejections of SGoF for $\alpha \in \{p_1, ..., p_n\}$. Dashed line: Number of SGoF adjusted p-values smaller than or equal to α.



(b) Empirical cumulative distribution function of the p-values (solid) and uniform distribution (dashed line).

Figure 3: Neuhaus data.

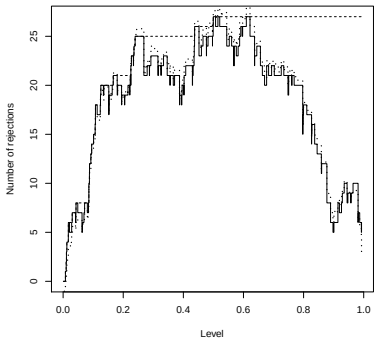
The values of $\widetilde{N}_n(\alpha)$ for $\alpha \in \{p_j, j = 1, ..., n\}$ and the empirical cumulative distribution function of the p-values are depicted in Figure 3 (left and right respectively). In this Figure 3 we see that $\widetilde{N}_n$ is an increasing-decreasing function of the level α, which reaches it maximum in the open $(0, 1)$ interval at $\alpha = p_{(9)}$, where the largest deviance between $F_n(\alpha)$ and α is encountered (Figure 3, right). Another point which is visible from Figure 3, left, is that one may construct a monotone rejection rule from the adjusted p-values, by considering the amount of $\tilde{p}_i$'s smaller than or equal to α (dashed line in the Figure).

From Table 2 we see that adjusted p-values for Hochberg and BH methods are greater than for SGoF procedure, but for the 8 largest p-values for which SGoF provides an adjusted p-value of 1. As a result, for Neuhaus data SGoF method will reject more nulls than Hochberg and BH methods whenever the metatest is performed at level $\alpha = 0.05$ or smaller. Since the 'excess of significant cases' $N_n(\alpha) = nF_n(\alpha) - b_n(\alpha) + 1$ in the binomial metatest is never above 7, there is no chance for SGoF to reject any of the $15 - 7 = 8$ nulls with the largest p-values. Note that, for example, BH with FDR control of 10% rejects 9 nulls; therefore, just like for Needleman's data, the power of SGoF relative to BH may vary with the level.

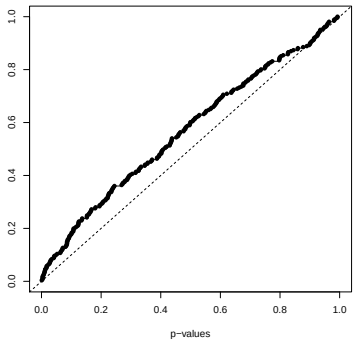
4.3 Diz data

In Diz et al. (2009), multiple comparison procedures were applied to a set of n =261 p-values coming from protein expression experiments in eggs of the marine mussel Mytilus edulis. In that study, the authors compared M. edulis

female protein expression profiles of two lines differing in sex ratio of their progeny. 26 out of the 261 p-values were smaller than 0.05. BH FDR-controlling procedure was applied, being unable to detect any significant feature even when allowing for a 20% of false discoveries. Indeed, the minimum adjusted p-value for BH procedure takes the value 0.2231. When applying SGoF method with $\alpha = \gamma = 0.05$, the seven null hypotheses corresponding to the minimum seven p-values were rejected ($p_{(7)} = 0.0077$).



(a) Solid line: Number of rejections of SGoF for $\alpha \in \{p_1, ..., p_n\}$. Dotted line : number of rejections of Conservative SGoF. Dashed line: number of SGoF adjusted p-values smaller than or equal to α.



(b) Empirical cumulative distribution function of the p-values (solid) and uniform distribution (dashed line).

Figure 4: Diz data.

The values of $N_n(\alpha)$ for $\alpha \in \{p_j, j = 1, ..., n\}$ are depicted in Figure 4. We do not consider the correction for ties $\widetilde{N}_n$ in this case since there is only one repetition with a negligible impact in the results. Since n is relatively large, we compute in this case the (asymptotic) conservative version of SGoF, $N_n^*(\alpha)$. Results are displayed in Figure 4, left, where it is seen that conservative SGoF rejects fewer nulls when the level $\alpha = \gamma$ is small. On the other hand, Figure 4, right, depicts the cumulative edf of the original p-values; this Figure shows a local maximum around $p_{(94)} = 0.244703$, which corresponds to the maximum distance between $F_n(\alpha)$ and α.

In Figure 5 we plot SGoF adjusted p-values versus BH adjusted p-values (left) and conservative SGoF adjusted p-values as defined in Definition 2 (right). It becomes clear from this Figure 5 that, for Diz et al. (2009)'s p-values, SGoF method entails a more powerful significance criterion compared to FDR controlling strategies. Moreover, we see that the adjusted p-values of SGoF and conservative SGoF are very similar, although the second one tends to be more conservative for small α's. This is in agreement with the relative values of $N_n(\alpha)$ and $N_n^*(\alpha)$ in Figure 4, left.

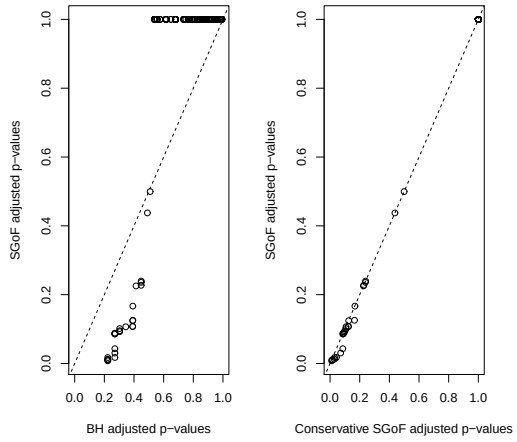


Figure 5: Left: Comparison between SGoF and BH adjusted p-values. Right: Comparison between SGoF and conservative SGoF adjusted p-values.

5 Main conclusions

In this paper adjusted p-values for SGoF multitesting procedure have been introduced, and their main properties have been established. These properties are crucial to understand the way in which SGoF adjusted p-values should be implemented and interpreted in practice. Adjusted p-values, as introduced here, allow for an automatic inspection of SGoF results, by letting the 'level' of the test $\alpha = \gamma$ to vary on the unit interval. Explicitly, the adjusted p-value is the smallest level at which SGoF procedure would still reject the given null hypothesis, while controlling for the multiplicity of tests.

The practical usage of SGoF adjusted p-values has been illustrated through the analysis of three biomedical or biological data sets. It has been seen that, unlike for other multitesting corrections, the adjusted p-values fall in the set of the original p-values. Since SGoF multitest is based on the excess of significance when comparing two proportions (the observed and the expected amounts of p-values below a given threshold), it turns out that the $n - N^{[n]}$ largest p-values are adjusted to be 1, where n is the number of tests and $N^{[n]}$ is the maximum number of nulls rejected by SGoF when letting the level vary. However, among the smallest $N^{[n]}$ p-values, SGoF method provides in many instances adjusted p-values smaller than those obtained by ordinary FWER or FDR controlling procedures. This is a result of the weak control of FWER which is imposed by SGoF strategy.

Two practical scenarios for which some care is needed have been discussed. These are the situations with tied p-values and the case in which the largest

p-value is equal to 1. Suitable correction for SGoF has been proposed and investigated to this regard. Also, it has been pointed up that the theoretical properties of SGoF adjusted p-values may not be so immediate to obtain (if true) when using the asymptotic normal approximation, as conservative SGoF method does. Another remark is that, since the amount of nulls rejected by SGoF is in general an increasing-decreasing function of the level, the number of adjusted p-values below α is only an upper bound for the number of rejections at that level α. In practice, when an adjusted p-value falls below α, one should only conclude that the corresponding null is rejected by SGoF at some level $\alpha' \leq \alpha$.

SGoF adjusted p-values have been introduced and investigated under the assumption of independence among the tests. In many real life problems, however, the sequence of n p-values at hand will suffer from some kind of dependence. A correction of SGoF method for serially dependent p-values has been recently proposed (de Uña-Álvarez, 2012); such a correction basically takes into account that the SGoF rejection rule $N_n(\alpha)$ must be updated to deal with the increasing variance coming from the existing correlation. Therefore, in principle adjusted p-values for SGoF metatest under dependence could be introduced following the ideas in this paper. This problem is currently under investigation and the corresponding results will be provided as soon as they become available.

A Technical Lemmas

In order to prove our main Lemma 2 we need a preliminary result. For each n and $b = 1, ..., n$ introduce the function $f_b^{(n)}(\alpha) = P(Bin(n, \alpha) \geq b) - \alpha = \sum_{k=b}^{n} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} - \alpha$, $0 < \alpha < 1$. In Figure 7 these functions are displayed for the case $n = 5$.

Lemma 1. The function $\alpha \in (0, 1) \to f_b^{(n)}(\alpha)$ is strictly increasing in every α^* such that $f_b^{(n)}(\alpha^*) = 0$.

Proof. In first place we consider the cases $b = 1$ and $b = n$. In both cases, the only zeros of $f_b^n(\alpha)$ are 0 and 1, but we are considering that $\alpha \in (0, 1)$ and hence there is nothing to prove. On the other hand, for $b \in \{2, ..., n-1\}$, it is easily seen that $f_b^n(\alpha)$ takes the value zero when $\alpha = 0, 1$. Now, the first derivative of $f_b^n(\alpha)$ is given by $n\binom{n-1}{b-1}\alpha^{b-1}(1-\alpha)^{n-b} - 1$, which is -1 when $\alpha = 0, 1$. Besides, the second derivative of $f_b^n(\alpha)$ is $\alpha^{b-2}(1-\alpha)^{n-b-1}\{b-1-(n-1)\alpha\}$, which is positive for all $\alpha < \frac{b-1}{n-1}$ and negative for all $\alpha > \frac{b-1}{n-1}$. Summarizing, $f_b^n(\alpha)$ takes negative values for small α, decreases toward zero as α approaches to 1, and has a unique inflexion point; this allows to conclude that the function $f_b^n(\alpha)$ increases when crossing the line $y = 0$. See Figure 7 for an illustration in the case $n = 5$.□

Lemma 2. The function $\alpha \in (0, 1) \to b_n(\alpha)$ is non-decreasing.

Proof. Take $\alpha < \alpha' = \alpha + \epsilon$ for some $\epsilon > 0$ small enough. We will show that $b_n(\alpha) \leq b_n(\alpha')$. We assume $b_n(\alpha) \geq 1$ since otherwise the result is immediate. We will consider separately two different cases. First (case I), suppose that $b_n(\alpha)$

is such that $f^{(n)}_{b_n(\alpha)}(\alpha) < 0$; then, by continuity, we have that $f^{(n)}_{b_n(\alpha)}(\alpha') < 0$ and, again by continuity, $f^{(n)}_{b_n(\alpha)-1}(\alpha') > 0$ since $f^{(n)}_{b_n(\alpha)-1}(\alpha) > 0$ which follows from $b_n(\alpha) = \inf\{b \in \{0, ..., n\} : f^n_b(\alpha) \leq 0\}$. Hence $b_n(\alpha') = b_n(\alpha)$ in case I. Second (case II) if $f^{(n)}_{b_n(\alpha)}(\alpha) = 0$ then $f^{(n)}_{b_n(\alpha)}(\alpha') > 0$ by Lemma 1, and $f^{(n)}_b(\alpha') > 0$ $\forall b \leq b_n(\alpha)$. Hence, $b_n(\alpha') = \inf\{b \in \{0, ..., n\} : f^n_n(\alpha') \leq 0\} > b_n(\alpha)$ in case II, which concludes the proof. □

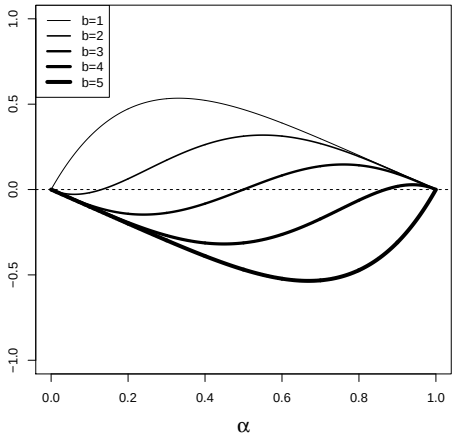


Figure 6: $f^n_b(\alpha)$ in the case $n = 5$ for the several possible values of b.

Acknowledgement

Work was supported by the Grant MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science and Innovation.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

[2] Benjamini, Y. and Yekutieli, D. (2001) The control of the False Discovery Rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

[3] Carvajal-Rodríguez, A., de Uña-Álvarez, J. and Rolán-Álvarez, E. (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics* **10**, 1–14.

[4] Castro-Conde, I. and de Uña-Álvarez, J. (2013). sgof: Multiple hypotheses testing. R package version 1.0. *http://CRAN.R-project.org/package=sgof*

[5] de Uña-Álvarez, J. (2011). On the statistical properties of SGoF multitesting method. *Statistical Applications in Genetics and Molecular* **10**, Article 18.

[6] de Uña-Álvarez, J. (2012). The Beta-Binomial SGoF method for multiple dependent tests. *Statistical Applications in Genetics and Molecular Biology*, **11**, Issue 3, Article 14.

[7] Diz, A.P., Dudley, E., MacDonald, B.W., Pina, B.. Kenchington, E.L., et al. (2009). Genetic variation underlying protein expression in eggs of the marine mussel Mytilus edulis. *Molecular & Cellular Proteomics* **8**, 132–144.

[8] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* **32**, 962–994.

[9] Dudoit, S. and van der Laan, M.J. (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

[10] Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* **75**, 800–802.

[11] MuToss Coding Team, Blanchard, G., et al. (2012). mutoss: Unified multiple testing procedures. R package version 0.1-7. *http://CRAN.R-project.org/package=mutoss*

[12] Needleman, H., Gunnoe, C., Leviton, A., Reed, R., Presie, H., Maher, C. and Barret, P. (1979). Deficits in psychologic and classroom performance of children with elevated dentine lead levels. *The New England Journal of Medicine* **300**, 689–695.

[13] Neuhaus, K.L., von Essen, R. et al. (1992). Improved thrombolysis in acute myocardial infarction with front-loaded administration of alteplase: results of the rt-PA-APSAC patency study (TAPS). *Journal of the American College of Cardiology* **19**, 885–891.

[14] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org/*.