Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 18

On the Statistical Properties of SGoF Multitesting Method

Jacobo de Uña-Alvarez*

*University of Vigo, jacobo@uvigo.es

On the Statistical Properties of SGoF Multitesting Method*

Jacobo de Uña-Alvarez

Abstract

In this paper we establish the statistical properties of SGoF multitesting method under a mixture model. It is assumed that the available set of p-values is statistically independent. Special attention is paid to the huge dimension problem in which the number of tests goes to infinity. Formulae for the power and the rate of false discoveries/non-discoveries of SGoF are given, so the role of the gamma-parameter of SGoF is understood. The existing connection between SGoF and a test of significance for the proportion of non-true nulls below gamma is explored. This connection suggests a possible modification of SGoF which may improve the power of the method. Simulation studies and a real data illustration are included.

KEYWORDS: false discovery rate, genomics, multiple testing, p-value

---

*Comments and suggestions from the Editor and two anonymous reviewers are acknowledged. The author thanks Antonio Carvajal-Rodriguez for careful reading of the paper, fruitful discussion and suggestions. Financial support from the grants 10PXIB300068PR of the Xunta de Galicia and MTM2008-03129 of the Spanish Ministry of Science and Innovation is acknowledged.

1 Introduction

Multiple-testing problems have received much attention since the advent of the -omic technologies: genomics, transcriptomics, proteomics, etc. They usually involve the simultaneous testing of thousands of hypotheses, or nulls, producing as a result a number of significant p-values or effects (that is, an increase in gene expression, or RNA/protein levels). In this setup, several criteria have been proposed to control for type-I errors (false positives). The family-wise error rate (FWER), defined as the probability of committing at least one type-I error through the several hypotheses under consideration, works as a substitute for the significance level in the traditional (single hypothesis) context. Typically FWER control is required in the strong sense, i.e. independently of the amount of true and false hypotheses.

Unfortunately, methods controlling the FWER have a remarkable lack of power, that is, they are unable to detect a reasonable amount of effects (Benjamini and Hochberg, 1995). As a more flexible strategy, the false discovery rate (FDR) criterion persecutes to maintain the proportion of false discoveries below a given level. Both FDR and FWER criteria coincide when all the nulls are true, but in general FDR will be smaller than FWER, so bounding the former allows for some improvement in the power (Benjamini and Hochberg, 1995). Traditional FWER- and FDR- based methods are nicely reviewed by Nichols and Hayasaka (2003) as well as by Dudoit and Laan (2008).

As a drawback of the FWER- and FDR-based methods, their power may be rapidly decreased as the number of tests grow, being unable to detect even one effect in particular situations (Carvajal-Rodríguez et al., 2009). This typically happens in situations with a large number of tests, when the effect in the non-true nulls is weak relative to the sample size (same reference). Storey (2003) suggested as a possible solution a weighted criterion in which both the FDR and the false non-discovery rate (FNR) are penalized. This issue was also explored in Cheng et al (2004), who proposed to evaluate the distance between the empirical and the uniform quantile processes, penalizing for the number of false discoveries. Further developments of FDR-based methods were given by Storey and Tibshirani (2003), Storey et al. (2004) and Nguyen (2004), among others.

Carvajal-Rodríguez et al. (2009) introduced a new multitesting strategy, SGoF, with focus on the existing number of effects with p-value below a given significance threshold (the γ parameter). SGoF method respects the FWER but only in the weak sense, that is, when all the nulls are true. This new approach provides a reasonable compromise between false discoveries and power (Carvajal-Rodríguez et al., 2009), and several enhancements of the method have been proposed (de Uña-Álvarez and Carvajal-Rodríguez, 2010). However, the theoretical statistical properties of SGoF have not been investigated in much detail so far. This paper fills this

erties of SGoF have not been investigated in much detail so far. This paper fills this gap.

The rest of the manuscript is organized as follows. In Section 2 we introduce the notations and we describe the method SGoF (Section 2.1) as well as a modification SGoF_1 (Section 2.2) with further power improvements. In Section 2.3 we compare the SGoF and FDR-based thresholding criteria through an illustrative example. Theoretical results are formally presented in Section 3. Specifically, we give expressions for the FDR, the FNR and the power of SGoF and SGoF_1 strategies. The influence of the SGoF's γ parameter is investigated. All these results correspond to the situation of independent tests. Simulation studies to compare the several multitesting adjustments are performed in Section 4, while a real data example is considered in Section 5. Section 6 provides a final discussion, including the issue of dependent p-values. Technical derivations are collected in the Appendix.

2 Methods

2.1 The mixture model

It is assumed that the n p-values $p_1, \ldots, p_n$ at hand constitute a random sample from a mixture distribution function (df) $F(x) = \pi_o F_0(x) + (1-\pi_0)F_1(x)$. Note that this implies that the p-values are statistically independent; see Section 6 for a brief discussion of the dependent case. Here, π_0 is the proportion of true nulls in the population, $F_0(x) = x$ is the df of the p-values corresponding to true nulls, and F_1 stands for the unknown df of the p-values corresponding to non-true nulls (or effects). The p-value p_i is attached to a null hypothesis H_{0i} which may be true or false. The *a priori* probability of being true for H_{0i} is π_0. Given that H_{0i} is true, p_i is distributed as F_0 (i.e. uniformly in the $(0,1)$ interval); on the contrary, conditionally on H_{0i} is false, p_i follows the df F_1.

In practice, both π_0 and F_1 must be estimated, and several estimators have been proposed in the related literature. See Cheng et al. (2004) and references therein. The problem is to decide which p_i come from F_1. If the density of F_1, $f_1 = F_1'$, exists and if it is monotone decreasing, traditional classification theory (i.e. minimizing Bayes' error) suggests the rule $\mathscr{R}(p^*) = \{p_i \leq p^*\}$ to detect the non-true nulls, where p^* is the solution to equation

$$f_1(p) = \frac{c(1|0)\pi_0}{c(0|1)(1-\pi_0)},$$

and where $c(0|1)$ is the penalty when we accept a non-true null, and $c(1|0)$ is the penalty when we reject a true null. As an example, in the special case $c(1|0) =$

DOI: 10.2202/1544-6115.1659

$c(0|1)$, we come up with equation $f_1(p) = \pi_0/(1-\pi_0)$, this is $f(p) = 2\pi_0$, which does not need to have a solution. Besides, even when the solution p^* exists, it may not be meaningful as a threshold of significance when it is too large. So this idea does not end with the classification problem. Similar comments hold true for the weighted FDR-FNR criterion of Storey (2003), see our Section 3 for further details.

2.2 SGoF revisited

Recently, Carvajal-Rodríguez et al. (2009) proposed a new method for p-value thresholding in multitesting problems. This method, called SGoF (from Sequential-Goodness-of-Fit), can be summarized as follows. Let F_n be the empirical df of the p_i, and let γ be an initial significance level, typically chosen as a small probability ($\gamma = 0.05$, 0.01, 0.001 and so on). Consider the intersection (or complete) null hypothesis $H_0 = \bigcap_{i=1}^n H_{0i}$. Under H_0 the p-values should fit well to a uniform distribution; more specifically, $F = F_0$ and hence the expected proportion of p-values below γ is just $F_0(\gamma) = \gamma$. If some of the null hypotheses are false, one expect a proportion $F(\gamma)$ above γ; so SGoF performs a one-sided (meta-)test at level α for

$$H_0(\gamma) : F(\gamma) = \gamma \quad \textit{versus} \quad H_1(\gamma) : F(\gamma) > \gamma,$$

with rejection region given by $\{nF_n(\gamma) \geq b_{n,\alpha}(\gamma)\}$, where the critical point $b_{n,\alpha}(\gamma)$ is defined through $b_{n,\alpha}(\gamma) = \inf\{b : P(Bin(n,\gamma) \geq b) \leq \alpha\}$. In the original formulation of SGoF (Carvajal-Rodríguez et al., 2009), $\gamma = 0.05$ and $\alpha = \gamma$, but no one of these conditions is essential, so we investigate the general case in which γ is an arbitrary (but known) value in the open interval $(0,1)$, and α is possibly distinct from γ. Interestingly, under the mixture model, the intersection null H_0 holds iff $H_0(\gamma)$ holds for some given γ (see Lemma 1 in the Appendix for the details).

As n grows, $b_{n,\alpha}(\gamma)$ can be approximated by

$$b_{n,\alpha}(\gamma) \approx n\gamma + \sqrt{n\gamma(1-\gamma)}z_\alpha \tag{1}$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ stands for the $(1-\alpha)$-quantile of the standard normal. By using this normal approximation, we see that $H_0(\gamma)$ is rejected iff

$$\frac{F_n(\gamma) - \gamma}{\sqrt{\gamma(1-\gamma)/n}} > z_\alpha,$$

which is the traditional (asymptotic) formulation of a one-sided test for a proportion. Rejection of $H_0(\gamma)$ in this meta-test means that there is at least one effect (i.e. a non-true null) among the n tests.

In case of rejection, SGoF concludes that the number of effects is given by the 'excess of significant cases' in the meta-test, this is $N_\alpha(\gamma) = nF_n(\gamma) - b_{n,\alpha}(\gamma) + 1$ (Carvajal-Rodríguez et al., 2009; de Uña-Álvarez and Carvajal-Rodríguez, 2010). Consequently, the $N_\alpha(\gamma)$ smallest p-values are declared as true effects. We denote the corresponding threshold p-value by $p^*_{n,\alpha}(\gamma)$, that is, $N_\alpha(\gamma) = nF_n(p^*_{n,\alpha}(\gamma))$. Note that, from (1),

$$\begin{aligned} N_\alpha(\gamma) &= nF_n(\gamma) - b_{n,\alpha}(\gamma) + 1 \\ &\approx n\left[F_n(\gamma) - \gamma\right] - \sqrt{n\gamma(1-\gamma)}z_\alpha + 1 \end{aligned}$$

and

$$\frac{N_\alpha(\gamma)}{n} = F_n(p^*_{n,\alpha}(\gamma)) \approx F_n(\gamma) - \gamma - \sqrt{\frac{\gamma(1-\gamma)}{n}}z_\alpha + \frac{1}{n},$$

or equivalently

$$\begin{aligned} p^*_{n,\alpha}(\gamma) &= F_n^{-1}\left(F_n(\gamma) - n^{-1}b_{n,\alpha}(\gamma) + n^{-1}\right) \\ &\approx F_n^{-1}\left(F_n(\gamma) - \gamma - \sqrt{\frac{\gamma(1-\gamma)}{n}}z_\alpha + \frac{1}{n}\right) \end{aligned} \tag{2}$$

where $F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$ is the so-called empirical quantile function. Since $b_{n,\alpha}(\gamma) \geq 1$ whenever $\alpha < 1$, and since F_n^{-1} is nondecreasing, we have $p^*_{n,\alpha}(\gamma) \leq \gamma$. This ensures that only p-values below γ may enter the set of hypotheses declared to be significant by SGoF. When only p-values below α are considered as candidates for effects, one should use the correction $N^c_\alpha(\gamma) = \min\{N_\alpha(\gamma), nF_n(\alpha)\}$ (cfr. de Uña-Álvarez and Carvajal-Rodríguez, 2010).

Since $F_n \to F$ as $n \to \infty$, the threshold p-value of SGoF $p^*_{n,\alpha}(\gamma)$ approaches to $p^*(\gamma) = F^{-1}(F(\gamma) - \gamma)$ as the number of tests n grows (and, in particular, $p^*(\gamma) \leq \gamma$). This is formally established as a Proposition in the next section (Proposition 1A). Note also that the limit threshold $p^*(\gamma)$ does not depend on the level α of the meta-test. This is because as $n \to \infty$ we have almost perfect knowledge on the value of $F(\gamma)$ and hence the influence of the level vanishes. For finite n, however, $p^*_{n,\alpha}(\gamma)$ grows with α and hence more effects are declared for a larger α. It is worthwhile recalling that the chosen α is controlling the FWER of SGoF in the weak sense, that is, under the intersection null $H_0 = \bigcap_{i=1}^n H_{0i}$ (Carvajal-Rodríguez

DOI: 10.2202/1544-6115.1659

et al., 2009). Since under H_0 the FWER coincides with the FDR, SGoF is also controlling the FDR in this case. It has been also referred that the power of SGoF increases with the number of tests (Carvajal-Rodríguez et al., 2009). The explanation of this property is found in the negative term $-n^{-1/2}\sqrt{\gamma(1-\gamma)}z_\alpha$ which turns the p-value threshold down (i.e. less power) when n is small. In Section 3 we will also see that $p^*(\gamma)$ increases with γ up to a maximum value, and then decreases.

SGoF's metatest statistic is related to the notion of second-level significance testing or *higher criticism* introduced by Tukey in 1976. However, for the best of our knowledge, this notion was not considered for testing individual hypotheses as in the SGoF method. As an extension of Tukey's idea, Donoho and Jin (2004) (see also Hall and Jin 2008, 2010) considered the 'higher criticism' test statistic

$$HC_n^* = \max_{1/n\leq\gamma\leq 1/2} \frac{F_n(\gamma)-\gamma}{\sqrt{\gamma(1-\gamma)/n}}$$

to detect a small fraction of nonnull hypotheses in what they called 'sparse heterogeneous mixtures'. In a related context, Donoho and Jin (2008) proposed a p-value thresholding method for feature selection based on a statistic similar to HC_n^*, where the maximum is taken within a given fraction of the top ranking features. Simulations reported in that paper suggest that this 'higher criticism thresholding' may be close to the ideal threshold which minimizes the missclassification rate when selecting features. See also Ahdesmäki and Strimmer (2010). Under SGoF's strategy, an initial threshold γ is fixed by the researcher, according to the level at which FWER is to be controlled. Hence, despite the possible similarities, no one of the referred methods performs a multitest adjustment in the same way as SGoF do.

There exists and interesting connection between the amount $F(\gamma)-\gamma$ and the proportion of non-true nulls with p-value falling below γ, $P_1(\gamma)=P(H_{0i}=1, p_i\leq\gamma)$, where $H_{0i}=1$ means "H_{0i} is false". Note that $P_1(\gamma)=(1-\pi_0)F_1(\gamma)=F(\gamma)-\pi_0\gamma\geq F(\gamma)-\gamma$, so $F(\gamma)-\gamma$ is a lower bound for $P_1(\gamma)$. Hence, as n grows, since $N_\alpha(\gamma)\approx n\,(F(\gamma)-\gamma)$, the number of effects declared by SGoF ($N_\alpha(\gamma)$) can be regarded as a lower bound for the number of true effects below the initial threshold γ (i.e. for $nP_1(\gamma)$). This also suggests a possible modification of SGoF which is expected to increase the power. This modification is introduced in the following subsection.

2.3 SGoF$_1$: testing for the number of effects

Consider the following null and alternative hypotheses:

$$H_0^1(\gamma): P_1(\gamma)=0 \quad \textit{versus} \quad H_1^1(\gamma): P_1(\gamma)>0.$$

Note that, under the mixture model, $H_0^1(\gamma)$ holds iff $\pi_0 = 1$, i.e. there exists a perfect coincidence between the intersection null H_0 and $H_0^1(\gamma)$ (whatever the value of γ is). Now, estimate $P_1(\gamma)$ by $P_{1,n}(\gamma) = F_n(\gamma) - \pi_{0,n}\gamma$ where $\pi_{0,n}$ is some consistent estimator of π_0. Then, provided that $P_{1,n}(\gamma)$ is normally distributed for large n, $H_0^1(\gamma)$ is rejected iff

$$\frac{P_{1,n}(\gamma)}{\sqrt{Var(P_{1,n}(\gamma))}} > z_\alpha,$$

where $Var(P_{1,n}(\gamma))$ stands for the variance of $P_{1,n}(\gamma)$ under $H_0^1(\gamma)$. In this case, the modified SGoF method, which we call SGoF$_1$, declares as effects the $N_{1,\alpha}(\gamma) = nP_{1,n}(\gamma) - n\sqrt{Var(P_{1,n}(\gamma))}z_\alpha + 1$ smallest p-values. As SGoF, this method weakly controls the FWER at level α. We can be more precise if e.g. we use the simple estimator of π_0 proposed by Dalmasso et al. (2005), namely

$$\pi_{0,n} = -\frac{1}{n}\sum_{i=1}^{n}\log(1-p_i).$$

It happens $E[\pi_{0,n}] \geq \pi_0$ so $\pi_{0,n}$ will result in some overestimation of the proportion of true nulls. Under (the intersection null) $H_0^1(\gamma)$ it can be seen that

$$Var(P_{1,n}(\gamma)) = \frac{\gamma[1-2(1-\gamma)\log(1-\gamma)]}{n}$$

(see Lemma 2 in the Appendix for a proof). This is the version of SGoF$_1$ we consider in this paper, although less conservative versions could be constructed. Introduce the threshold p-value of SGoF$_1$, which is given by

$$\begin{aligned} p^*_{1,n,\alpha}(\gamma) &= F_n^{-1}(n^{-1}N_{1,\alpha}(\gamma)) \\ &= F_n^{-1}\left(P_{1,n}(\gamma) - \sqrt{Var(P_{1,n}(\gamma))}z_\alpha + \frac{1}{n}\right). \end{aligned} \quad (3)$$

Note that (similarly as for $p^*_{n,\alpha}(\gamma)$) we have $p^*_{1,n,\alpha}(\gamma) \leq F_n^{-1}(F_n(\gamma) - \pi_{0,n}\gamma) \leq \gamma$; $p^*_{1,n,\alpha}(\gamma)$ grows with α; and $p^*_{1,n,\alpha}(\gamma)$ tends to increase (and hence the power increases) with the number of tests n. On the other hand, since $P_{1,n}(\gamma) \to P_1(\gamma)$ as $n \to \infty$, we have that $p^*_{1,n,\alpha}(\gamma)$ converges to $p^*_1(\gamma) = F^{-1}(P_1(\gamma)) = F^{-1}(F(\gamma) - \pi_0\gamma)$ as $n \to \infty$, which is greater than $p^*(\gamma)$. This means that asymptotically SGoF$_1$ will

DOI: 10.2202/1544-6115.1659

have more power than SGoF for any given γ, and accordingly a larger FDR. We have confirmed through simulations that this modification results in more power also in the finite sample case, specially for large values of γ. See our Section 4 below. However, unlike for SGoF, the asymptotic power of SGoF$_1$ increases with γ all along the $(0,1)$ interval.

There is a nice interpretation of SGoF$_1$'s threshold p-value, because (asymptotically) it is just the point for which the cumulative proportion of p-values equals the proportion of true effects below the initial significance level γ (i.e. $P_1(\gamma)$). For finite samples, a similar interpretation holds, but in this case one should refer (rather than to $P_1(\gamma)$) to the lower limit of significance (at level α) for $P_1(\gamma)$, namely

$$P_{1,n}(\gamma) - \sqrt{Var(P_{1,n}(\gamma))} z_\alpha + \frac{1}{n}.$$

Remark. The result $p^*_{1,n,\alpha}(\gamma) \to p^*_1(\gamma)$ as $n \to \infty$ (and hence the asymptotic interpretation of SGoF$_1$ method) still holds when α is replaced by a sequence α_n of FWER-controlling levels such that $n^{-1/2} z_{\alpha_n} = n^{-1/2}\Phi^{-1}(1-\alpha_n) \to 0$ as $n \to \infty$, see Proposition 1B in Section 3. For example, the choice $\alpha_n = \alpha/n$ is interesting because it introduces a conservative version of SGoF$_1$'s strategy, which may compensate an 'excess in power' (with a high associated FDR) when the number of tests n is very large. Same comments apply to SGoF. See our real data example in Section 5 for illustration.

2.4 Relationship to FDR-based methods

It is interesting to relate the p-value threshold of SGoF (resp. SGoF$_1$) to that provided by other multitesting strategies, such as FDR-based methods, in the asymptotic setting. Bonferroni's method takes the threshold $p^*_{B,n,\alpha} = \alpha/n$, where α is controlling the FWER in the strong sense. In this case, it is clear that the power is lost in the limit, since $Pow(p^*_{B,n,\alpha}) = F_1(\alpha/n) \to 0$ as $n \to \infty$ (see Section 3 for a formal definition of the power). The same is not true for FDR-controlling strategies. For example, Benjamini and Hochberg (1995) proposed as threshold

$$p^*_{BH,n,\alpha}(\pi_0) = \max\left\{ p_i : p_i \leq \frac{\alpha}{\pi_0} F_n(p_i) \right\},$$

where α is controlling the FDR and where in practice π_0 is replaced by 1 or by some conservative estimator when it is unknown (as it happens in most of the cases). As

n grows, F_n converges to the true mixture df F, and hence $p^*_{BH,n,\alpha}(\pi_0) \to p^*_{BH,\alpha}(\pi_0)$ as $n \to \infty$, where $p^*_{BH,\alpha}(\pi_0)$ is the solution of

$$p = \frac{\alpha}{\pi_0} F(p).$$

In some instances, this solution will be too small to detect even a single true effect from the sequence of p-values. Of course, this problem gets worse when using the conservative version with $\pi_0 = 1$, $p^*_{BH,\alpha}(1)$. In order to illustrate this, we consider the following example. In the simulations section more evidence on the lack of power of BH is reported.

Example 1. Consider the mixture df given by $F(x) = \pi_0 x + (1-\pi_0)F_1(x)$ where $F_1(x) = x^\theta$ for some $\theta \in (0,1)$. Since $\theta < 1$, the density associated to F_1 is monotone decreasing. The effect size is controlled by the θ parameter; stronger effects are obtained for smaller values of θ. It is straightforward to see that the solution of $p = \alpha F(p)/\pi_0$ is given by

$$p^*_{BH,\alpha}(\pi_0;\theta) = \left[\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}\right]^{1/(1-\theta)},$$

with an associated power of $Pow(p^*_{BH,\alpha}(\pi_0;\theta)) = F_1(p^*_{BH,\alpha}(\pi_0;\theta)) = p^*_{BH,\alpha}(\pi_0;\theta)^\theta$. Note that $p^*_{BH,\alpha}(\pi_0;\theta)$ is less than 1 provided that $\pi_0 > \alpha$. Take $\pi_0 = 0.9$, $\theta = 0.5$ and $\alpha = 0.05$. Then, $p^*_{BH,\alpha}(\pi_0;\theta) = 0.0000342$ and $Pow(p^*_{BH,\alpha}(\pi_0;\theta)) = 0.005847953$, which means that with $n = 1000$ tests (and hence $n_1 = 100$ true effects on average) the BH threshold would be unable to detect even a single effect. One could argue that the asymptotic expression for $p^*_{BH,\alpha}(\pi_0;\theta)$ is not valid for $n = 1000$. To explore this, we have simulated 10,000 Monte Carlo trials from the model and we have computed the number of simulations for which no effect was found; the proportion was 41.79%, while the average number of rejections was 1.23 (standard deviation = 1.5). This shows that the problem of no power is also present in the 'finite sample' case ($n < \infty$).

Example 1 (see also the simulations in Section 4) shows that, in practice, FDR-controlling criteria may be too strict to detect true effects. However, in some occasions it will be the case that $p^*_{BH,\alpha}(\pi_0) > p^*(\gamma)$, or $p^*_{BH,\alpha}(\pi_0) > p^*_1(\gamma)$, and thus the BH method will be asymptotically more powerful than SGoF procedure. Example 1 is also useful to illustrate this (see Section 4 below for further illustration). The asymptotic p-value threshold of SGoF and SGoF$_1$ are given respectively by

$$p^*(\gamma) = p^*(\gamma,\theta) = F^{-1}((1-\pi_0)(\gamma^\theta - \gamma))$$

DOI: 10.2202/1544-6115.1659

and

$$p_1^*(\gamma) = p_1^*(\gamma, \theta) = F^{-1}((1-\pi_0)\gamma^\theta)$$

where $F(x) = \pi_0 x + (1-\pi_0)x^\theta$. In Figure 1 we report these values for $\gamma = 0.05$ in the case $\pi_0 = 0.9$. For comparison, we also report the BH threshold p-values $p^*_{BH,\alpha}(\pi_0;\theta)$ in the case $\alpha = 0.05$. From this Figure 1 it is seen that $p^*_{BH,\alpha}(\pi_0;\theta) > p^*(\gamma,\theta)$ for $\theta < 0.04$, and that $p^*_{BH,\alpha}(\pi_0;\theta) > p_1^*(\gamma,\theta)$ for $\theta < 0.02$; in this case,

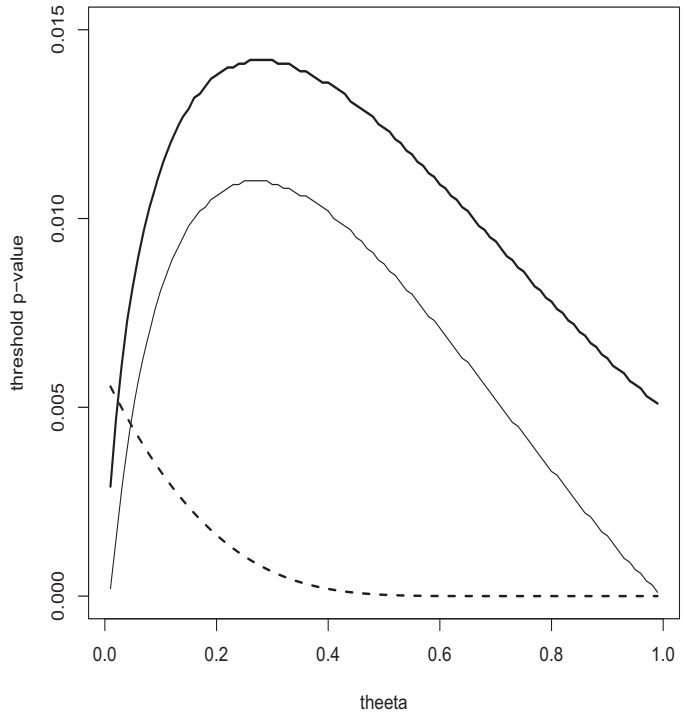

Figure 1. Asymptotic threshold p-value for SGoF(0.05) and $\text{SGoF}_1(0.05)$ with FWER=0.05 (thin and bold solid lines respectively) and for BH with FDR=0.05 (dashed line) for Example 1. The proportion of true nulls is 90%.

the FDR of SGoF procedure is below α, which results in a smaller power. However, as θ grows, the p-values corresponding to the non-true nulls are more mixed with

those belonging to the true nulls, and BH p-value threshold goes down to zero. The opposite behavior is found in the SGoF (respectively SGoF_1) threshold (which does not respect the 5% FDR level and hence it is more liberal) at least up to $\theta \approx 0.25$, showing the compromise between FDR and power which is achieved by using this method.

The situation with $p^*_{BH,\alpha}(\pi_0;\theta) > p^*_1(\gamma,\theta)$ (corresponding to extremely well-separated null and alternative p-values) can be interpreted as a case in which there is no reason to pay a FDR as high as 5% (as BH do). This is at least what it can be said under SGoF_1's view with $\gamma = 0.05$, because this method tells that the number of true effects with p-value below $\gamma = 0.05$ is less than the number of p-values falling below $p^*_{BH,\alpha}(\pi_0;\theta)$ (and hence declared as significant features by BH).

3 Theoretical considerations

Before establishing the main theoretical results, we recall some important definitions associated to multitesting problems. The false discovery rate (FDR) is defined as the expected proportion of true nulls among the rejected hypotheses. A related measure is the positive false discovery rate (PFDR), which is just the FDR given that there is at least one rejection. For the rejection region $\mathscr{R}(p) = \{p_i \leq p\}$, these quantities are functions of the threshold p-value p, so we write $FDR(p)$ and $PFDR(p)$. According to Theorem 1 in Storey (2003), and putting $H_{0i} = 0$ for "H_{0i} is true", we have

$$\begin{aligned} PFDR(p) &= P(H_{0i} = 0 | p_i \leq p) \\ &= \frac{\pi_0 P(p_i \leq p | H_{0i} = 0)}{P(p_i \leq p)} \\ &= \frac{\pi_0 p}{F(p)}. \end{aligned}$$

Besides,

$$\begin{aligned} FDR(p) &= PFDR(p) P\left(\bigcup_{i=1}^{n} \{p_i \leq p\}\right) \\ &= PFDR(p)\left[1 - (1 - F(p))^n\right], \end{aligned}$$

which explains why in general $FDR(p) \leq PFDR(p)$, and why the FDR approaches to the PFDR as the number of tests grows. Note that, for finite n, the PFDR is

DOI: 10.2202/1544-6115.1659

not necessarily controlled at level α even when using a multitest adjustment than controls FDR at that level. Proposition 1A below gives the FDR and PFDR of SGoF, while Proposition 1B states the corresponding results for SGoF_1. For SGoF_1, we assume throughout that $\pi_{0,n}$ is any strongly consistent estimator of π_0.

Proposition 1A. (a) The PFDR and the FDR of $SGoF(\gamma)$ at weak FWER control α are respectively given by

$$PFDR(p^*_{n,\alpha}(\gamma)) = \frac{\pi_0 p^*_{n,\alpha}(\gamma)}{F(p^*_{n,\alpha}(\gamma))}$$

and

$$FDR(p^*_{n,\alpha}(\gamma)) = \frac{\pi_0 p^*_{n,\alpha}(\gamma)}{F(p^*_{n,\alpha}(\gamma))}\left[1-(1-F(p^*_{n,\alpha}(\gamma)))^n\right]$$

where $p^*_{n,\alpha}(\gamma) = F_n^{-1}(F_n(\gamma) - n^{-1}b_{n,\alpha}(\gamma) + n^{-1})$.

(b) As $n \to \infty$ we have $n^{-1}b_{n,\alpha}(\gamma) = \gamma + O(z_\alpha n^{-1/2})$, $p^*_{n,\alpha}(\gamma) \to p^*(\gamma) = F^{-1}(F(\gamma)-\gamma)$, and both $PFDR(p^*_{n,\alpha}(\gamma))$ and $FDR(p^*_{n,\alpha}(\gamma))$ converge with probability 1 to

$$PFDR(p^*(\gamma)) = \frac{\pi_0 p^*(\gamma)}{F(p^*(\gamma))} = \frac{\pi_0 F^{-1}(F(\gamma)-\gamma)}{F(\gamma)-\gamma}.$$

Proof. Statement (a) follows by evaluating $PFDR(p)$ and $FDR(p)$ at $p = p^*_{n,\alpha}(\gamma)$. For (b) just use (1), representation (2), and the strong consistency of both the empirical df and the empirical quantile function.

Proposition 1B. (a) The PFDR and the FDR of $SGoF_1(\gamma)$ at weak FWER control α are respectively given by

$$PFDR(p^*_{1,n,\alpha}(\gamma)) = \frac{\pi_0 p^*_{1,n,\alpha}(\gamma)}{F(p^*_{1,n,\alpha}(\gamma))}$$

and

$$FDR(p^*_{1,n,\alpha}(\gamma)) = \frac{\pi_0 p^*_{1,n,\alpha}(\gamma)}{F(p^*_{1,n,\alpha}(\gamma))}\left[1-(1-F(p^*_{1,n,\alpha}(\gamma)))^n\right]$$

where $p^*_{1,n,\alpha}(\gamma) = F_n^{-1}(P_{1,n}(\gamma) - n^{-1}b_{1,n,\alpha}(\gamma) + n^{-1})$, and where $b_{1,n,\alpha}(\gamma) = n\sqrt{Var(P_{1,n}(\gamma))}z_\alpha$.

(b) As $n \to \infty$ we have $n^{-1}b_{1,n,\alpha}(\gamma) = O(z_\alpha n^{-1/2})$, $p^*_{1,n,\alpha}(\gamma) \to p^*_1(\gamma) = F^{-1}(P_1(\gamma)) = F^{-1}(F(\gamma) - \pi_0\gamma)$, and both $PFDR(p^*_{1,n,\alpha}(\gamma))$ and $FDR(p^*_{1,n,\alpha}(\gamma))$ converge with probability 1 to

$$PFDR(p^*_1(\gamma)) = \frac{\pi_0 p^*_1(\gamma)}{F(p^*_1(\gamma))} = \frac{\pi_0 F^{-1}(F(\gamma) - \pi_0\gamma)}{F(\gamma) - \pi_0\gamma}.$$

Proof. Statement (a) follows by evaluating $PFDR(p)$ and $FDR(p)$ at $p = p^*_{1,n,\alpha}(\gamma)$. For (b) just use representation (3), and the strong consistency of both $P_{1,n}(\gamma)$ and the empirical quantile function.

The situation in which very few null hypotheses are true is problematic for FDR control. Note that the rate of false discoveries approaches to zero as $\pi_0 \to 0$ and hence controlling the FDR at a nominal level α will not be meaningful. In order to illustrate this, consider our Example 1 in Section 2; in this case, the asymptotic p-value threshold given by the optimal or 'benchmark' BH method (which perfectly estimates π_0) is above 1 whenever $\pi_0 < \alpha$. This means that one would declare all the null hypotheses as false regardless their associated p-values. The FDR resulting from this decision is just π_0, which will be much smaller than the nominal α as $\pi_0 \to 0$. At the same time, one may not feel comfortable with a method which reports a discovery when the associated p-value is large. Interestingly, the asymptotic threshold of SGoF and SGoF$_1$ when π_0 approaches zero (i.e. $F \approx F_1$) are given by $p^*(\gamma) = F_1^{-1}(F_1(\gamma) - \gamma)$ and $p^*_1(\gamma) = \gamma$ respectively, both of them below the initial significance level γ (and, indeed, the second one equal to γ). This is not surprising since (as discussed above) SGoF strategy answers the question 'how many effects are there below threshold γ?' In this way, SGoF and SGoF$_1$ exhibit large or even full power below γ in the case $\pi_0 \approx 0$ while providing reasonable p-value thresholds. See our simulations section for further illustration.

A natural counterpart of the FDR is the false non-discovery rate (FNR), that is, the expected proportion of non-true nulls among those being accepted. If we condition the expectation to accepting at least one hypothesis, we come up with the positive false non-discovery rate (PFNR), which is related to the FNR through

$$\begin{aligned} FNR(p) &= PFNR(p)P\left(\bigcup_{i=1}^{n}\{p_i > p\}\right) \\ &= PFNR(p)(1 - F(p)^n). \end{aligned}$$

Putting $H_{0i} = 1$ for "H_{0i} is false", it happens (Storey, 2003)

$$PFNR(p) = P(H_{0i} = 1 | p_i > p)$$

DOI: 10.2202/1544-6115.1659

$$\begin{aligned} &= \frac{(1-\pi_0)(1-F_1(p))}{1-F(p)} \\ &= 1-\frac{\pi_0(1-p)}{1-F(p)}. \end{aligned}$$

Propositions 2A and 2B give the PFNR and the FNR of SGoF and SGoF_1 multitesting methods respectively.

Proposition 2A. (a) The PFNR and the FNR of $SGoF(\gamma)$ at weak FWER control α are respectively given by

$$PFNR(p^*_{n,\alpha}(\gamma)) = 1-\frac{\pi_0(1-p^*_{n,\alpha}(\gamma))}{1-F(p^*_{n,\alpha}(\gamma))}$$

and

$$FNR(p^*_{n,\alpha}(\gamma)) = \left[1-\frac{\pi_0(1-p^*_{n,\alpha}(\gamma))}{1-F(p^*_{n,\alpha}(\gamma))}\right](1-F(p^*_{n,\alpha}(\gamma))^n).$$

(b) As $n\to\infty$ both $PFNR(p^*_{n,\alpha}(\gamma))$ and $FNR(p^*_{n,\alpha}(\gamma))$ converge with probability 1 to

$$\begin{aligned} PFNR(p^*(\gamma)) &= 1-\frac{\pi_0(1-p^*(\gamma))}{1-F(p^*(\gamma))} \\ &= 1-\frac{\pi_0(1-F^{-1}(F(\gamma)-\gamma))}{1-F(\gamma)+\gamma}. \end{aligned}$$

Proposition 2B. (a) The PFNR and the FNR of $SGoF_1(\gamma)$ at weak FWER control α are respectively given by

$$PFNR(p^*_{1,n,\alpha}(\gamma)) = 1-\frac{\pi_0(1-p^*_{1,n,\alpha}(\gamma))}{1-F(p^*_{1,n,\alpha}(\gamma))}$$

and

$$FNR(p^*_{1,n,\alpha}(\gamma)) = \left[1-\frac{\pi_0(1-p^*_{1,n,\alpha}(\gamma))}{1-F(p^*_{1,n,\alpha}(\gamma))}\right](1-F(p^*_{1,n,\alpha}(\gamma))^n).$$

(b) As $n\to\infty$ both $PFNR(p^*_{1,n,\alpha}(\gamma))$ and $FNR(p^*_{1,n,\alpha}(\gamma))$ converge with probability 1 to

$$PFNR(p^*_1(\gamma)) = 1-\frac{\pi_0(1-p^*_1(\gamma))}{1-F(p^*_1(\gamma))}$$

$$= \quad 1 - \frac{\pi_0(1 - F^{-1}(F(\gamma) - \pi_0\gamma))}{1 - F(\gamma) + \pi_0\gamma}.$$

When the density f_1 is monotone decreasing, the limit $p^*(\gamma) = F^{-1}(F(\gamma) - \gamma)$ of the SGoF's threshold $p^*_{n,\alpha}(\gamma)$ is an increasing-decreasing function of the γ parameter. As mentioned, typically one will be interested in choosing a small value for γ, which plays the role of an initial significance level. By letting the γ parameter vary, de Uña-Álvarez and Carvajal-Rodríguez (2010) introduced a significance trace for the number of effects declared by SGoF. Note that, on the contrary, the asymptotic threshold of SGoF$_1$ $p^*_1(\gamma) = F^{-1}(F(\gamma) - \pi_0\gamma)$ increases with γ, attaining its maximum value $F^{-1}(1 - \pi_0)$ at $\gamma = 1$.

Propositions 1A,B and 2A,B give as particular cases the asymptotic PFDR (or FDR) and PFNR (or FNR) of $SGoF(\gamma)$ and $SGoF_1(\gamma)$. Note that the function $PFDR(p)$ (resp. $PFNR(p)$) is monotone increasing (resp. decreasing) if f_1 is a decreasing density. In such a case, from $p^*(\gamma) \leq p^*_1(\gamma)$ it follows that the FDR (resp. FNR) of SGoF is smaller (resp. larger) than that of SGoF$_1$. Also, if γ^* is the maximizer of the function $\gamma \mapsto F(\gamma) - \gamma$, we have the bounds

$$\begin{aligned} PFDR(p^*(\gamma)) &\leq PFDR(p^*(\gamma^*)) \\ &= \frac{\pi_0 F^{-1}(F(\gamma^*) - \gamma^*)}{F(\gamma^*) - \gamma^*} \end{aligned}$$

and

$$\begin{aligned} PFNR(p^*(\gamma)) &\geq PFNR(p^*(\gamma^*)) \\ &= 1 - \frac{\pi_0(1 - F^{-1}(F(\gamma^*) - \gamma^*))}{1 - F(\gamma^*) + \gamma^*}. \end{aligned}$$

These are asymptotic upper and lower bounds for the FDR and FNR committed by SGoF. The upper bound for FDR can be seen as the maximum proportion of false discoveries one could accept in a reasonable way when thresholding the p-values. On the other hand, the lower bound for FNR says that there exists a particular choice of the γ parameter which minimizes the proportion of true effects among the nulls accepted by SGoF. This choice γ^* can be easily estimated in practice by maximizing the function $\gamma \mapsto F_n(\gamma) - \gamma$. Put γ^*_n for such a maximizer. We have the following result.

DOI: 10.2202/1544-6115.1659

Proposition 3. As $n \to \infty$ we have

$$FDR(p^*(\gamma_n^*)) \to \frac{\pi_0 F^{-1}(F(\gamma^*)-\gamma^*)}{F(\gamma^*)-\gamma^*}$$

and

$$FNR(p^*(\gamma_n^*)) \to 1-\frac{\pi_0(1-F^{-1}(F(\gamma^*)-\gamma^*))}{1-F(\gamma^*)+\gamma^*}.$$

We should point out that the 'automatic choice' of γ given by γ_n^* may have some drawbacks. Firstly, the value reported by γ_n^* may not be meaningful if it is too large, unless the researcher is decided to be definitively liberal. On the other hand, by minimizing the FNR one may end with an undesirably high FDR. Note that this strategy is useless for SGoF_1 because minimization of FNR with respect to γ always reports $\gamma = 1$.

Storey (2003) proposed as a possible thresholding criterion to minimize a weighted average of the PFDR and PFNR, namely

$$\begin{aligned} W_u(p) &= (1-u)PFDR(p)+uPFNR(p) \\ &= (1-u)\frac{\pi_0 p}{F(p)}+u\left[1-\frac{\pi_0(1-p)}{1-F(p)}\right]. \end{aligned}$$

Recall that PFDR increases with p, while the opposite is true for PFNR. The choice of u is left to the researcher who must proceed according to the importance of the rate of false discoveries relative to that of false nondiscoveries. In the case of SGoF, and when n is large enough, this criterion looks for the minimizer of $\gamma \mapsto W_u(p^*(\gamma))$, which will be closer to γ^* as $u \to 1$. However, smaller values of γ will appear as minimizers as u moves away from one. Importantly, one should keep in mind that the function $W_u(p)$ may not reach a minimum inside the interval $(0,1)$, and hence this criterion is not always useful.

A measure somehow connected with (but not equal to) the FNR is the power. The power is defined as the expected proportion of true effects which are detected by a multitesting strategy. For a rejection region of type $\mathscr{R}(p) = \{p_i \leq p\}$ we have

$$Pow(p) = E\left[\frac{1}{n_1}\sum_{i=1}^{n} I(p_i \leq p, H_{0i}=1)\right]$$

where $n_1 = \sum_{i=1}^{n} I(H_{0i}=1)$ is the (random) number of non-true nulls among the n nypotheses. By using the iid assumption and that, conditionally on $n_1 = k$,

$\sum_{i=1}^{n} I(p_i \leq p, H_{0i} = 1) \sim Bin(k, F_1(p))$, we obtain

$$Pow(p) = F_1(p);$$

that is, the power of $\mathscr{R}(p)$ is just the cumulative df of the non-true p-values evaluated at threshold p. Obviously, the power increases as p increases, and hence more power is associated with a higher FDR and a smaller FNR. Next two results provide the power associated to SGoF and SGoF$_1$ procedures. In particular, it is shown that the power of SGoF$_1$ is larger than that of SGoF.

Proposition 4A. The power of $SGoF(\gamma)$ at weak FWER control α is given by

$$Pow(p^*_{n,\alpha}(\gamma)) = F_1\left(p^*_{n,\alpha}(\gamma)\right).$$

As $n \to \infty$ we have $Pow(p^*_{n,\alpha}(\gamma)) \to F_1(p^*(\gamma)) = F_1(F^{-1}(F(\gamma) - \gamma))$ with probability 1.

Proposition 4B. The power of $SGoF_1(\gamma)$ at weak FWER control α is given by

$$Pow(p^*_{1,n,\alpha}(\gamma)) = F_1\left(p^*_{1,n,\alpha}(\gamma)\right).$$

As $n \to \infty$ we have $Pow(p^*_{1,n,\alpha}(\gamma)) \to F_1(p^*_1(\gamma)) = F_1(F^{-1}(F(\gamma) - \pi_0\gamma))$ with probability 1.

The maximum power of SGoF is asymptotically achieved by γ^*, which corresponds to the minimum FNR as discussed. On the other hand, since $F_1 \geq F$, lower bounds for the power are easily obtained from Proposition 4A,B, namely $F_1(p^*(\gamma)) \geq F(\gamma) - \gamma$ and $F_1(F^{-1}(F(\gamma) - \pi_0\gamma)) \geq F(\gamma) - \pi_0\gamma$. This means that at least the $100(F(\gamma) - \gamma)\%$ (resp. the $100(F(\gamma) - \pi_0\gamma)\%$) of the true effects will be detected by $SGoF(\gamma)$ (resp. by $SGoF_1(\gamma)$) when n is large.

We end this Section by recalling Property 3 of SGoF multitest in Carvajal-Rodríguez et al. (2009). This Property 3 says that the probability of rejecting the intersection null with SGoF multitest increases with the number of tests n. Under the mixture model, it is straightforward to see that the probability of rejection equals (with Φ the cumulative df of a standard normal)

$$r_n = P\left(\frac{F_n(\gamma) - \gamma}{\sqrt{\gamma(1-\gamma)/n}} > z_\alpha\right)$$

DOI: 10.2202/1544-6115.1659

$$= \quad 1-\Phi\left(z_\alpha\sqrt{\frac{\gamma(1-\gamma)}{F(\gamma)(1-F(\gamma))}}-\sqrt{n}\frac{F(\gamma)-\gamma}{\sqrt{F(\gamma)(1-F(\gamma))}}\right),$$

which satisfies $r_n \to 1$ as $n \to \infty$ whenever $F(\gamma) > \gamma$. Note that $F(\gamma)-\gamma = (1-\pi_0)(F_1(\gamma)-\gamma)$ is large when the proportion of non-true nulls is large or when $F_1(\gamma)-\gamma$ is large (strong effects).

4 Simulation studies

In this Section we investigate the finite sample behavior of the proposed methods in a simulated scenario. The number of hypotheses being tested is $n = 100$, 1000, or 10000, sampled from a population with a proportion of true nulls given by $\pi_0 = 1$, 0.9, or 0.7. The p-values corresponding to true nulls are generated according to a uniform distribution on the $(0,1)$ interval. For the generation of p-values under the alternative, we consider a one-sided test for the mean μ of a normal population (with unknown variance σ^2) based on a sample of size s, where $H_0 : \mu = 0$ and $H_1 : \mu > 0$. The test statistic is distributed as a Student's t with $s-1$ degrees of freedom. Sample sizes $s = 5$, 20, and 80 are considered, leading to effect levels

$$w = \sqrt{s}\frac{\mu-0}{\sigma_{s-1}}$$

where σ^2_{s-1} is the sampling variance which is distributed as $\sigma^2\chi^2_{s-1}/(s-1)$. This means that the non-true p-values are generated as

$$p_i = 1-\Phi_{s-1}\left(\frac{Z_i}{\sqrt{A_i/(s-1)}}+\sqrt{s}\frac{\mu-0}{\sqrt{\sigma^2 A_i/(s-1)}}\right)$$

where $Z_i \sim N(0,1)$ and $A_i \sim \chi^2_{s-1}$ are independent random variables, and Φ_{s-1} stands for the cumulative df of a t variable with $s-1$ degrees of freedom. We always take $\mu = 1$ under H_1, and the variance is fixed to $\sigma^2 = 6$. Note that, as the sample size s increases, we get stronger relative effects w; specifically, the given values for s (5, 20, and 80) lead to w around 1.13 (weak effects), 1.90 (moderate), and 3.68 (strong effects) respectively.

In Tables 1 to 3 we report the FDR and the power reached by the several methods along 1000 Monte Carlo trials. The methods under comparison are SGoF with $\gamma = 0.05$, $\gamma = 0.01$, and $\gamma = 0.001$; SGoF$_1$ with the same values for γ, and BH. To favor the BH method, we have used the true value of π_0 in its implementation;

note that this value is unknown in practice. Level $\alpha = 0.05$ was taken in all the cases; recall that this α controls the FDR in BH but only the FWER in the weak sense for SGoF and SGoF_1 methods. Besides the FDR and the power, we report the quotient power/FDR as a sensible measure of performance, which we call here the 'relative power'. Note that results with higher relative power reveal benefits when increasing the p-value threshold, because for those cases the FDR is increased but at a rate lower than that of the power. From Tables 1 to 3 we see that:

(a) Under the intersection null all the methods respect the nominal FDR (FWER) quite well, although some anticonservativeness is appreciated for SGoF and SGoF_1 methods when applied with a small γ value (specially for small to moderate number of tests). Besides, SGoF seems to be a bit more liberal than SGoF_1.

(b) When some of the hypotheses are false, both versions of SGoF present a power which, unlike for BH, increases with the number of tests. An exception is Table 3, in which the power of BH does not decrease or even slightly increases with the sample size. This can be justified because when $s = 80$ the p-values corresponding to the non-true nulls are extremely well-separated from those associated to the true nulls, and hence a 5% of FDR is more than enough to detect a large proportion of effects.

(c) In most of the cases SGoF and SGoF_1 present more power than BH. Exceptions to this are (again) well-separated distributions (Table 3), and intermediate effect sizes (Table 2) when the γ parameter is taken too small. Of course, larger powers are associated to an increase of the FDR. In general, the power (also the FDR) decreases as γ decreases (this is because the considered values of γ are below the γ^* for which the power is maximum). But this is not true for Table 3 in the case $\pi_0 = 0.9$, where γ^* seems to be around 0.01. In relative terms, the choice $\gamma = 0.05$ could be considered as optimal because it leads to the maximum relative power (but this depends on the situation, since more relative power is sometimes achieved at $\gamma = 0.001$). However, the FDR level for $\gamma = 0.05$ could be taken as too large in some applications (specially for large n), so one may prefer the more conservative $\gamma = 0.01$ or even $\gamma = 0.001$ in special instances.

(d) SGoF_1 exhibits a power at least as good as SGoF, with a better relative performance for larger values of γ. This is in agreement with the theoretical results. Both versions of SGoF are almost equivalent for $\gamma = 0.001$. On the other hand, in Table 1 the relative power of SGoF_1 is larger than that of SGoF, while the opposite is true in Table 3 -even when in all the cases the power of SGoF_1 is above that of SGoF.

(e) SGoF and SGoF_1's FDR levels under a 30% of effects may be smaller or greater than those corresponding to $\pi_0 = 0.9$, depending on the choice of γ.

DOI: 10.2202/1544-6115.1659

(f) The relative power of both versions of SGoF is in general above that pertaining to BH. The situation is the opposite in Table 2 with $n = 100$, and in the same Table 2 with $\pi_0 = 0.7$ for the smallest value of γ.

Table 1. FDR, power, and relative power of Benjamini-Hochberg (BH), SGoF, and SGoF_1 methods performed at 5% level, with weak effect level ($w = 1.13$, or $s = 5$), see text. The number of hypotheses is n, the proportion of true nulls is π_0, and $\pi_{0,n}$ stands for its estimated value. Averages along 1,000 Monte Carlo trials (standard deviations for $\pi_{0,n}$ between brackets).

	$\pi_0 = 1$		$\pi_0 = 0.9$			$\pi_0 = 0.7$	
	FDR	FDR	Pow	RP	FDR	Pow	RP
$n = 100$;π_{0n} :	0.996 (0.094)		0.946 (0.097)			0.825 (0.092)	
BH	0.052	0.044	0.003	0.069	0.049	0.006	0.119
$SGoF(0.05)$	0.063	0.126	0.016	0.125	0.180	0.038	0.212
$SGoF_1(0.05)$	0.059	0.133	0.017	0.125	0.210	0.046	0.220
$SGoF(0.01)$	0.076	0.100	0.008	0.083	0.104	0.015	0.141
$SGoF_1(0.01)$	0.076	0.100	0.008	0.083	0.104	0.015	0.141
$SGoF(0.001)$	0.085	0.067	0.006	0.089	0.064	0.006	0.088
$SGoF_1(0.001)$	0.085	0.067	0.006	0.089	0.064	0.006	0.088
$n = 1000$;π_{0n} :	1.002 (0.032)		0.942 (0.031)			0.825 (0.030)	
BH	0.046	0.049	1×10^{-4}	0.008	0.052	0.001	0.012
$SGoF(0.05)$	0.043	0.391	0.017	0.042	0.341	0.068	0.199
$SGoF_1(0.05)$	0.045	0.443	0.022	0.051	0.350	0.085	0.242
$SGoF(0.01)$	0.069	0.253	0.005	0.019	0.297	0.017	0.056
$SGoF_1(0.01)$	0.041	0.242	0.005	0.019	0.310	0.019	0.061
$SGoF(0.001)$	0.092	0.105	0.001	0.012	0.131	0.002	0.013
$SGoF_1(0.001)$	0.092	0.105	0.001	0.012	0.131	0.002	0.013
$n = 10000$;π_{0n} :	1.000 (0.010)		0.941 (0.010)			0.824 (0.009)	
BH	0.068	0.043	3×10^{-5}	0.001	0.048	5×10^{-5}	0.001
$SGoF(0.05)$	0.046	0.651	0.036	0.056	0.346	0.084	0.242
$SGoF_1(0.05)$	0.044	0.656	0.045	0.069	0.352	0.101	0.287
$SGoF(0.01)$	0.067	0.609	0.008	0.013	0.317	0.022	0.069
$SGoF_1(0.01)$	0.065	0.628	0.010	0.015	0.320	0.026	0.080
$SGoF(0.001)$	0.081	0.268	0.001	0.002	0.298	0.002	0.007
$SGoF_1(0.001)$	0.049	0.268	0.001	0.002	0.299	0.002	0.008

(g) The quantity $\pi_{0,n}$ overestimates the proportion of true nulls, specially for a small sample size s (i.e. weaker relative effects) in the case $\pi_0 = 0.7$. Hence, there is still some possibilities of enhancing the power of SGoF_1 by considering more accurate estimators.

Table 2. FDR, power, and relative power of Benjamini-Hochberg (BH), SGoF, and SGoF_1 methods performed at 5% level, with moderate effect level ($w = 1.90$, or $s = 20$), see text. The number of hypotheses is n, the proportion of true nulls is π_0, and $\pi_{0,n}$ stands for its estimated value. Averages along 1,000 Monte Carlo trials (standard deviations for $\pi_{0,n}$ between brackets).

	$\pi_0 = 1$		$\pi_0 = 0.9$			$\pi_0 = 0.7$	
	FDR	FDR	Pow	RP	FDR	Pow	RP
$n = 100$;π_{0n} :	0.995 (0.096)		0.909 (0.091)			0.739 (0.096)	
BH	0.045	0.055	0.075	1.355	0.048	0.175	3.633
$SGoF(0.05)$	0.058	0.172	0.169	0.982	0.110	0.349	3.173
$SGoF_1(0.05)$	0.051	0.178	0.176	0.988	0.117	0.374	3.207
$SGoF(0.01)$	0.072	0.138	0.133	0.967	0.069	0.202	2.904
$SGoF_1(0.01)$	0.072	0.138	0.133	0.967	0.069	0.202	2.904
$SGoF(0.001)$	0.090	0.070	0.071	1.017	0.022	0.068	3.092
$SGoF_1(0.001)$	0.090	0.070	0.071	1.017	0.022	0.068	3.092
$n = 1000$;π_{0n} :	1.000 (0.031)		0.913 (0.030)			0.739 (0.028)	
BH	0.053	0.060	0.026	0.426	0.052	0.144	2.763
$SGoF(0.05)$	0.063	0.263	0.279	1.059	0.122	0.402	3.277
$SGoF_1(0.05)$	0.058	0.277	0.300	1.085	0.133	0.433	3.249
$SGoF(0.01)$	0.093	0.187	0.173	0.924	0.071	0.224	3.167
$SGoF_1(0.01)$	0.065	0.187	0.173	0.928	0.072	0.230	3.183
$SGoF(0.001)$	0.075	0.097	0.051	0.533	0.034	0.062	1.833
$SGoF_1(0.001)$	0.075	0.097	0.051	0.533	0.034	0.062	1.833
$n = 10000$;π_{0n} :	1.000 (0.010)		0.913 (0.010)			0.739 (0.010)	
BH	0.061	0.057	0.012	0.207	0.050	0.141	2.824
$SGoF(0.05)$	0.051	0.295	0.325	1.103	0.129	0.422	3.279
$SGoF_1(0.05)$	0.052	0.309	0.347	1.120	0.140	0.453	3.248
$SGoF(0.01)$	0.054	0.202	0.190	0.942	0.072	0.231	3.212
$SGoF_1(0.01)$	0.056	0.206	0.196	0.951	0.074	0.239	3.227
$SGoF(0.001)$	0.091	0.107	0.057	0.535	0.031	0.065	2.067
$SGoF_1(0.001)$	0.053	0.107	0.057	0.535	0.031	0.065	2.078

DOI: 10.2202/1544-6115.1659

Table 3. FDR, power, and relative power of Benjamini-Hochberg (BH), SGoF, and $SGoF_1$ methods performed at 5% level, with strong effect level ($w = 3.68$, or $s = 80$), see text. The number of hypotheses is n, the proportion of true nulls is π_0, and $\pi_{0,n}$ stands for its estimated value. Averages along 1,000 Monte Carlo trials (standard deviations for $\pi_{0,n}$ between brackets).

	$\pi_0 = 1$		$\pi_0 = 0.9$			$\pi_0 = 0.7$	
	FDR	FDR	Pow	RP	FDR	Pow	RP
$n = 100$;π_{0n} :	1.000 (0.101)		0.909 (0.095)			0.704 (0.095)	
BH	0.060	0.051	0.833	16.42	0.049	0.945	19.25
$SGoF(0.05)$	0.066	0.026	0.572	22.31	0.015	0.812	55.83
$SGoF_1(0.05)$	0.06	0.027	0.576	21.15	0.020	0.844	42.64
$SGoF(0.01)$	0.086	0.033	0.742	22.35	0.014	0.843	59.33
$SGoF_1(0.01)$	0.086	0.033	0.742	22.35	0.014	0.843	59.33
$SGoF(0.001)$	0.112	0.011	0.671	59.25	0.004	0.681	137.6
$SGoF_1(0.001)$	0.112	0.011	0.671	59.25	0.004	0.681	137.6
$n = 1000$;π_{0n} :	0.998 (0.031)		0.900 (0.031)			0.703 (0.032)	
BH	0.049	0.050	0.838	16.68	0.050	0.942	18.99
$SGoF(0.05)$	0.062	0.036	0.784	21.63	0.019	0.871	45.33
$SGoF_1(0.05)$	0.045	0.047	0.816	17.26	0.030	0.907	30.19
$SGoF(0.01)$	0.086	0.040	0.810	20.14	0.016	0.859	52.54
$SGoF_1(0.01)$	0.063	0.041	0.811	19.89	0.017	0.865	49.68
$SGoF(0.001)$	0.080	0.012	0.654	55.54	0.003	0.667	196.1
$SGoF_1(0.001)$	0.080	0.012	0.654	55.54	0.003	0.667	196.1
$n = 10000$;π_{0n} :	1.000 (0.010)		0.901 (0.010)			0.702 (0.010)	
BH	0.045	0.050	0.840	16.87	0.050	0.942	18.87
$SGoF(0.05)$	0.049	0.053	0.844	16.06	0.024	0.892	37.09
$SGoF_1(0.05)$	0.048	0.067	0.872	12.52	0.038	0.926	24.31
$SGoF(0.01)$	0.051	0.046	0.831	18.08	0.018	0.866	48.91
$SGoF_1(0.01)$	0.050	0.049	0.837	17.12	0.019	0.874	45.16
$SGoF(0.001)$	0.081	0.012	0.661	55.18	0.003	0.669	200.4
$SGoF_1(0.001)$	0.051	0.012	0.661	55.18	0.003	0.670	199.9

In summary, one may say that both SGoF and $SGoF_1$ approaches will be often more powerful than BH, but the situation may change if the null and the alternative p-value distributions are very well separated (large effect sizes); that the

relative power of SGoF's is in general above that of BH; and that $SGoF_1$ slightly outperforms SGoF in the sense of the power but not necessarily in relative power (e.g. SGoF may be preferable to $SGoF_1$ for strong effects).

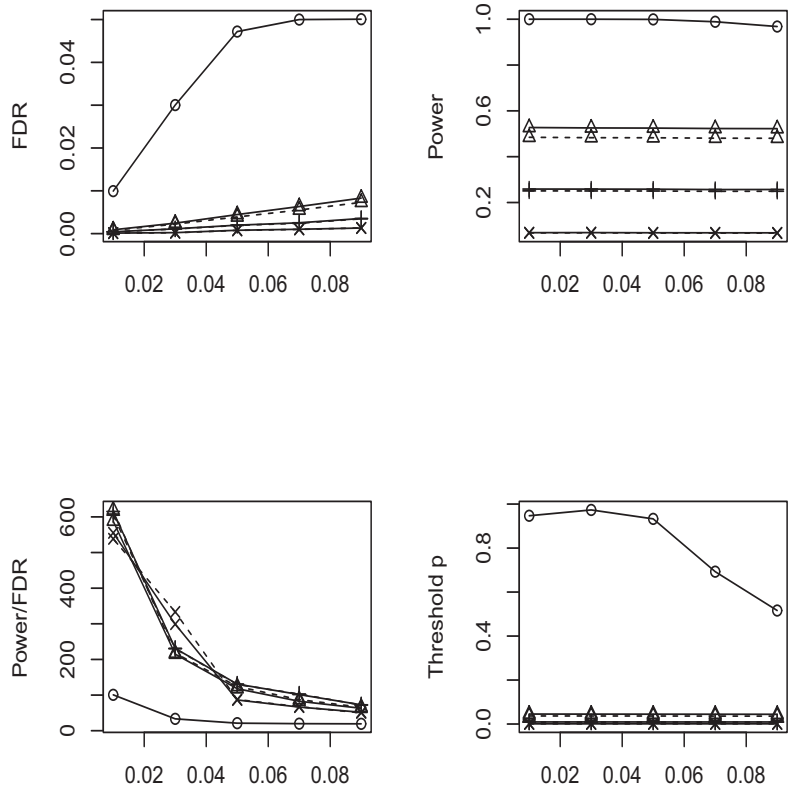

Figure 2. FDR, power, relative power, and threshold p-value as functions of the proportion of true nulls, for the several methods performed at 5% level, with moderate effect level ($s = 20$) and $n = 1000$ tests: BH ($\circ$), SGoF or $SGoF_1$ with $\gamma = 0.05$ ($\triangle$), 0.01 ($+$), and 0.001 ($\times$). Continuous lines for $SGoF_1$ and dashed lines for SGoF. Averages along 1,000 Monte Carlo trials.

In Section 3 we discussed the case in which $\pi_0 \approx 0$. It was mentioned that BH tend to declare all the null hypotheses as false in such a case, with a resulting

DOI: 10.2202/1544-6115.1659

FDR of $100\pi_0\%$ regardless the nominal α. For illustration purposes, here we simulate from our model the simultaneous testing of $n = 1000$ null hypotheses with $\pi_0 = 0.01$, 0.03, 0.05, 0.07, and 0.09, with moderate effect level ($s = 20$) under the alternative. Results for BH at nominal FDR of 5% and for SGoF and SGoF$_1$ methods at the same nominal (weak-control) FWER are reported in Figure 2. It is seen that the power of both SGoF's strategies is about half that of BH when using $\gamma = 0.05$, with worse results for smaller γ's. However, the relative power of SGoF and SGoF$_1$ is (up to six times) larger than that of BH for small π_0, indicating that paying a FDR as large as $100\pi_0\%$ (as BH do) is not justified. Figure 2 also reveals that BH identifies as true effects p-values close to 1 when $0.01 \leq \pi_0 \leq 0.05$, with a threshold p-value still as large as 0.52 for the largest π_0 ($\pi_0 = 0.09$). The threshold p-value of both SGoF versions is never above 0.045 in these simulations, according to the maximum value fixed for γ ($\gamma = 0.05$).

5 An example

As an illustrative example, we took the microarray study of hereditary breast cancer by Hedenfalk et al. (2001). One of the goals of this study was to find genes differentially expressed between BRCA1- and BRCA2-mutation positive tumors. Thus, for each of the 3,226 genes of interest, a p-value was assigned based on a suitable statistical test for the comparison. Following previous analysis of these data (Storey and Tibshirani, 2003), 56 genes were eliminated because they had one or more measurements exceeding 20. This left n =3,170 genes.

We have applied SGoF and SGoF$_1$ multitesting approaches to these data for several choices of γ, specifically $\gamma = 0.05$, $\gamma = 0.01$, $\gamma = 0.005$, and $\gamma = 0.001$. Besides the standard version with (weak) FWER control $\alpha = 0.05$, we have considered the conservative versions based on $\alpha/n = 1.577 \times 10^{-5}$. In Table 4 we report the number of discoveries with the corresponding threshold p-value for each of the eight different combinations. We see that the multitest correction becomes more conservative as the γ parameter decreases. The more conservative version among the eight metatests corresponds to SGoF with $\gamma = 0.001$ and FWER of α/n, which declares 64 tests as true effects, with a p-value threshold of 0.00071. The SGoF$_1$ version declares 67 tests as significative in this case; we can interpret these numbers by saying that there is statistical significance about the existence of 67 (SGoF$_1$) or more than 64 (SGoF) true effects with p-value smaller than 0.001.

The estimated rates of false discoveries, proportions of effects detected, and relative powers are provided in Table 5. These quantities were estimated by using

the following formulae:

$$FDR_n(p) = \frac{\pi_{0,n} p}{F_n(p)}, \quad \pi_{0,n} = -\frac{1}{n}\sum_{i=1}^{n} \log(1-p_i) = 0.7177,$$

and

$$Pow_n(p) = \frac{F_n(p) - \pi_{0,n} p}{1 - \pi_{0,n}}.$$

In terms of relative power, the best option among the eight considered metatests is performing SGoF with $\gamma = 0.005$ and $\alpha = 0.05$; this choice provides an estimated detection rate of more than 16.45%, with a FDR of 5.05% among the 155 discoveries (threshold p-value of 0.00341). For comparison, we have applied the Benjamini-Hochberg multitest correction (based on $\pi_{0,n} = 0.7177$) at 5% of FDR, resulting almost the same results: 157 discoveries and a threshold p-value of 0.00344. The BH's estimated power is 16.67% and its estimated relative power is 3.34. Hence, in this example the 'optimal' results (in the sense of the relative power) are achieved for a FDR of around 5% (the same happens in Table 2 of the simulation section, case $n = 100$ and $\pi_0 = 0.7$), but in general this does not need to be the case (as it becomes clear from our simulation section).

Note also that in general one may play with the values of γ and α in the application of SGoF and SGoF$_1$ to fit a given FDR. This is because the thresholds $p^*_{n,\alpha}(\gamma)$ and $p^*_{1,n,\alpha}(\gamma)$ may be increased or decreased by changing these parameters, as discussed in Section 2.

Table 4. Number of discoveries (threshold p-value between brackets) for SGoF and SGoF$_1$ when applied to Hendefalk data ($n = 3,170$).

FWER α :	0.05		$0.05/n$	
γ	SGoF	SGoF$_1$	SGoF	SGoF$_1$
0.05	428 (0.02538)	471 (0.03133)	395 (0.02260)	438 (0.02724)
0.01	225 (0.00731)	233 (0.00771)	208 (0.00641)	219 (0.00692)
0.005	155 (0.00341)	160 (0.00374)	143 (0.00306)	149 (0.00316)
0.001	71 (0.00078)	71 (0.00078)	64 (0.00071)	67 (0.00074)

DOI: 10.2202/1544-6115.1659

Table 5. FDR, power (Pow) and relative power (RP) estimated for SGoF and $SGoF_1$ when applied to Hedenfalk data.

	SGoF			$SGoF_1$		
γ	FDR	Pow	RP	FDR	Pow	RP
FWER 0.05						
0.05	.1349	.4137	3.06	.1513	.4466	2.95
0.01	.0739	.2328	3.15	.0752	.2407	3.20
0.005	.0501	.1645	3.29	.0532	.1693	3.18
0.001	.0250	.0773	3.10	.0250	.0773	3.10
FWER $0.05/n$						
0.05	.1302	.3839	2.95	.1415	.4201	2.97
0.01	.0701	.2161	3.08	.0719	.2271	3.16
0.005	.0487	.1520	3.12	.0483	.1584	3.28
0.001	.0253	.0697	2.75	.0252	.0730	2.90

6 Discussion and concluding remarks

Traditional multitest adjustments aim to control the FWER or, when being more liberal, the FDR. This means that, given a (small) type-I error rate α, the proportion of true nulls among those being rejected (FDR) is maintained below α. In many applications, particularly when the number of tests is large, such criteria result in a small power, i.e., the proportion of rejected nulls among the false ones is low. The problem gets worse when using the more stringent FWER instead of the FDR. On the contrary, given a (small) γ, SGoF method performs a meta-test of significance at level α to determine the number of effects (non-true nulls) with p-value smaller than γ. SGoF controls the FWER at level α but only in the weak sense, that is, when all the nulls are true. We have shown that SGoF is able to detect at least the $100(F(\gamma)-\gamma)\%$ of the existing effects as the number of tests grows, independently of the α value. This result is achieved by declaring as significant features all the p-values falling below the threshold $p^*(\gamma) = F^{-1}(F(\gamma)-\gamma)$. The per comparison error rate (PCER) attached to the threshold $p^*(\gamma)$ is given by $\pi_0 p^*(\gamma)$, which is always bounded by $\pi_0\gamma \leq \gamma$, but possibly much smaller according to the distance $F(\gamma)-\gamma$ and the special shape of F (see Figure 1). This shows that SGoF is not equivalent to any procedure aiming to control the PCER at a given level. When the number of tests is relatively small, both the power and the threshold p-value of SGoF decrease when using a smaller α.

The fact that the FDR of SGoF is not controlled at any level can be considered as a drawback of the method, and this limitation should be taken into account. However, at the same time it has interesting consequences. The first one is the power improvement, as discussed. Secondly, it is useful to identify situations in which paying a $100\alpha\%$ of FDR for the fixed α could be too much. This may happen when the effects are strong relative to the sample size (e.g. Table 3). Then, SGoF meta-test may reveal that there is no statistical significance to declare as many effects as a FDR-based method would do. Finally, according to the provided theoretical results, the FDR of SGoF may be maintained as low as desired by an appropriate tuning of the γ parameter, the value of α being also important when the number of tests is moderate.

A modification of SGoF, SGoF_1, has been introduced and investigated in this paper. The theoretical results show that SGoF_1 is more powerful than SGoF, particularly for large values of γ, with a corresponding extra in the FDR. It is not always the case, however, that the performance of SGoF_1 is better in the sense of the relative power, i.e. the quotient Power/FDR. All these features have been analyzed through simulation studies and also by exploring real data.

Inside the SGoF family, there exists a member which minimize the false non-discovery rate, leading to an optimal power. This member corresponds to the choice $\gamma = \gamma^*$, where γ^* denotes the point at which the distance between the observed and expected p-value cumulative distribution functions $(F(\gamma) - \gamma)$ is maximum. Even when most practitioners will prefer to fix γ in advance, the choice γ^* could report interesting additional results to help the interpretation. In particular, the FDR attached to γ^* is the maximum FDR one should accept in a given multitesting problem, at least under SGoF's point of view.

The technical results stated in this paper are valid under the assumption that the available p-values are statistically independent. However, in many applications it will be the case that the multiple tests under consideration are dependent. As noted by Owen (2005), dependences may greatly affect the variance of the number of discoveries. This means that the denominator in SGoF's metatest statistic $\sqrt{\gamma(1-\gamma)/n}$ should be replaced by a more conservative quantity when analyzing dependent p-values. Indeed, Carvajal-Rodríguez and de Uña-Álvarez (2011) showed that SGoF's method, as defined in this paper, loses its FWER control when the p-values are highly correlated. Several corrections of the variance for dependences have been proposed in different multitesting scenarios (Owen, 2005; Efron, 2010), and this is without any doubt an exciting research area. We will investigate in our future work the possible application to SGoF method of such corrections for dependence.

DOI: 10.2202/1544-6115.1659

7 Appendix: Technical results

In this Section we collect some technical results which have been referred along the paper.

Lemma 1. Assume that the density of F_1, $f_1 = F_1'$, exists, and that it is a differentiable, monotone decreasing function. Then, the intersection null H_0 is characterized by the meta-null hypothesis $H_0(\gamma)$ tested by $SGoF(\gamma)$.

Proof. It is clear that the complete null hypothesis H_0 implies $H_0(\gamma)$, so we prove the converse. Assume $F(\gamma) = \gamma$ for a given $0 < \gamma \leq \gamma^*$ where γ^* stands for the maximizer of $\varphi(x) = F(x) - x$; note that $\varphi'(x) = f(x) - 1 = (1-\pi_0)(f_1(x) - 1)$ and that $\varphi''(x) < 0$ because f_1 is decreasing, so γ^* is the solution of $f_1(x) = 1$ which always exists. Then, for $x < \gamma$ we have

$$0 \leq F(x) - x \leq F(\gamma) - \gamma = 0$$

where the first inequality follows because $\varphi_1(x) = F_1(x) - x$ satisfies $\varphi_1(x) \geq \varphi_1(0) = 0$ under the given conditions. Hence, we get $F(x) = x$ for $0 \leq x \leq \gamma$ from which $F_1(x) = x$ in the same interval. This gives $f_1(x) = 1$ for $0 \leq x \leq \gamma$ and since f_1 is decreasing and integrates 1 on the interval $[0,1]$ we conclude $f_1 \equiv 1$. In the case $\gamma > \gamma^*$ we have $F(x) = x$ for $\gamma \leq x \leq 1$ from which $f_1(x) = 1$ for $\gamma \leq x \leq 1$, and the same argument applies.

Lemma 2. Let $P_{1,n}(\gamma) = F_n(\gamma) - \pi_{0,n}\gamma$ where

$$\pi_{0,n} = -\frac{1}{n}\sum_{i=1}^{n} \log(1 - p_i).$$

Under the intersection null we have

$$n^{1/2} \frac{P_{1,n}(\gamma)}{\sqrt{\gamma[1 - 2(1-\gamma)\log(1-\gamma)]}} \to N(0,1)$$

in distribution.

Proof. Put $\varphi(x) = -\log(1-x)$. We have

$$P_{1,n}(\gamma) = F_n(\gamma) - \pi_{0,n}\gamma = \frac{1}{n}\sum_{i=1}^{n} \left[I(p_i \leq \gamma) - \gamma\varphi(p_i) \right].$$

Now,

$$E\left[I(p_1 \leq \gamma) - \gamma\varphi(p_1) \right] = F(\gamma) - \gamma \int \varphi dF,$$

while

$$E\left[I\left(p_1 \leq \gamma\right) - \gamma\varphi\left(p_1\right)\right]^2 = F(\gamma) + \gamma^2 E\left[\varphi\left(p_1\right)^2\right] - 2\gamma E\left[I\left(p_1 \leq \gamma\right)\varphi\left(p_1\right)\right].$$

Under the intersection null we have

$$\int \varphi dF = -\int \log(1-x)dx = 1,$$

so $E\left[I\left(p_1 \leq \gamma\right) - \gamma\varphi\left(p_1\right)\right] = F(\gamma) - \gamma = 0$; besides, straightforward calculations give

$$E\left[I\left(p_1 \leq \gamma\right)\varphi\left(p_1\right)\right] = -\int_0^\gamma \log(1-x)dx = (1-\gamma)\ln(1-\gamma) + \gamma$$

and

$$E\left[\varphi\left(p_1\right)^2\right] = \int \log^2(1-x)dx = 2.$$

Summarizing, under the intersection null $E\left[I\left(p_1 \leq \gamma\right) - \gamma\varphi\left(p_1\right)\right] = 0$ and

$$E\left[I\left(p_1 \leq \gamma\right) - \gamma\varphi\left(p_1\right)\right]^2 = \gamma[1 - 2(1-\gamma)\log(1-\gamma)].$$

Apply the central limit theorem to conclude.

References

Ahdesmäki M, Strimmer K (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Annals of Applied Statistics* 4:503-519.

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

Carvajal-Rodriguez A, de Uña-Álvarez J, Rolan-Alvarez E (2009) A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics* 10:209.

DOI: 10.2202/1544-6115.1659

Carvajal-Rodríguez A, de Uña-Álvarez J (2011) A Simulation Study on the Impact of Strong Dependence in High-Dimensional Multiple-Testing I: The Case without Effects. *M.P. Rocha et al. (Eds.): 5th International Conference on PACBB, AISC 93*, pp. 241-246. Springer

Cheng C, Pounds SB, Boyett JM, Pei, D, Kuo ML, Roussel (2004) Statistical significance threshold criteria for analysis of microarray gene expression data. *Statistical Applications in Genetics and Molecular Biology* Vol. 3, Issue 1, Article 36.

Dalmasso C, Broet P, Moreau T (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics* 21:660-668.

de Uña-Álvarez J, Carvajal-Rodríguez A (2010) 'SGoFicance Trace': Assessing significance in high-dimensional testing problems. *PLoS ONE* 5(12): e15930. doi:10.1371/journal.pone.0015930

Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* 32:962-994.

Donoho D, Jin J (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of National Academy of Science* 105:14790-14795.

Efron B (2010) Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* 105:1042-1055.

Hall P, Jin J (2008) Properties of higher criticism under strong dependence. *Annals of Statistics* 36: 381-402.

Hall P, Jin J (2010) Innovated higher criticism for detecting spares signals in correlated noise. *Annals of Statistics* 38:1686-1732.

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

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, et al. (2001) Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine* 344: 539-548.

Nguyen D (2004) On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray sequences. *Computational Statistics & Data Analysis* 47:611-637.

Nichols T, Hayasaka S (2003) Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research* 12: 419-446.

Owen A (2005) Variance of the number of false discoveries. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 67:411-426.

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* 31: 2013-2035.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of National Academy of Science* 100: 9440-9445.

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 66:187-205.

DOI: 10.2202/1544-6115.1659