Universidade Vigo

Power, FDR and conservativeness of BB-SGoF method for multiple dependent tests: a simulation study

Irene Castro-Conde and Jacobo de Uña-Álvarez

Report 13/03

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

Power, FDR and conservativeness of BB-SGoF method for multiple dependent tests: a simulation study

Irene Castro-Conde and Jacobo de Uña-Álvarez

Report 13/03

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

Power, FDR and conservativeness of BB-SGoF method for multiple dependent tests: a simulation study

Irene Castro-Conde and Jacobo de Uña-Álvarez

September 13, 2013

Abstract

Beta-Binomial SGoF (or BB-SGoF) method for multiple hypotheses testing has been recently proposed as a suitable modification of the Sequential Goodness-of-Fit (SGoF) multitesting method when the tests are correlated in blocks. In this paper we investigate the power, the false discovery rate, and the conservativeness of BB-SGoF in an intensive Monte Carlo simulation study. Important features such as automatic selection of the number of existing blocks and preliminary testing for independence are explored. Our study reveals that (a) BB-SGoF method maintains the properties of original SGoF in the dependent case; (b) BB-SGoF weakly controls for FDR even when the Beta-Binomial model is violated and the number of blocks k is unknown; and that (c) the loss of power of the automatic selector for the number of blocks relative to the benchmark method which uses the true k varies depending on the proportion and the type (strong, intermediate or weak) of the effects, being strongly influenced by the within-block correlation too.

1 Introduction

Multiple testing procedures have become more and more important in the last years due to the increasing availability of information in fields like genomics, transcriptomics, or proteomics. In many occasions the goal is to control for the number of type I errors (or false positives) along a sequence of hundreds, thousands or tens of thousands of hypotheses which are tested simultaneously. Methods traditionally used in this setting control for the family-wise error rate (FWER) or for the false discovery rate (FDR) at a pre-specified level α. FWER-controlling procedures ensure that no type I error will be committed with probability at least $1-\alpha$; on the other hand, FDR-based methods control at level α the expected proportion of false positives among the null hypotheses being rejected. FWER methods include, among others, Bonferroni and Holm (step-down) procedures, while the standard approach for controlling the FDR is the

Benjamini-Hochberg method (Benjamini and Hochberg, 1995). See Nichols and Hayasaka (2003) or Dudoit and van der Laan (2008) for a deeper introduction to this area.

Unfortunately, it has been quoted by several authors that FWER and FDR controlling methods sometimes exhibit a low power. Here, the power of a multitesting method is defined as the proportion of non-true nulls which are rejected. This has motivated the introduction of alternative decision criteria. Generalized FWER criterion was studied in van der Laan et al. (2004) or Lehmann and Romano (2005), among others. Genovese and Wasserman (2002) suggested to minimize the loss function $FNR + \lambda FDR$, where FNR stands for the false non-discovery rate (which is the expected proportion of non-true null hypotheses among the accepted ones), while λ is a pre-specified penalty. Storey (2003) proposed as a possible thresholding criterion to minimize a weighted average of the positive FDR and FNR, where the choice of the weight is left to the researcher who must proceed according to the importance of the rate of false discoveries relative to that of false non-discoveries. Other approaches based on p-value thresholding are reviewed in Genovese and Wasserman (2004).

More recently, Carvajal-Rodríguez *et al.* (2009) introduced a sequential goodness-of-fit (SGoF) test to make a decision on the number and allocation of non-true nulls. Carvajal-Rodríguez et al.'s SGoF method starts by comparing the observed amount of p-values below an initial threshold γ to the expected amount under the intersection or complete null (i.e., under the assumption that all the null hypotheses are true). Such a comparison is performed through a binomial test at level α; the excess of observed p-values below γ with respect to the critical point at level α in the binomial test is then used to identify the non-true nulls. Statistical properties of this approach were explored in de Uña-Álvarez (2011). SGoF's procedure controls for FWER (and FDR) at level α, but only in the weak sense (i.e. under the complete null), which makes a difference with other, more standard procedures. That is, SGoF method is liberal with respect to the strong control of FWER or FDR, which are not a priori bounded when some of the nulls are false. However, SGoF multitest controls at the pre-specified level α the probability that the number of false positives exceeds the number of false non-discoveries with p-value below γ (de Uña-Álvarez, 2012). This property of conservativeness is unique to the SGoF approach. SGoF method has become in short time a popular tool for applied scientists; as an indicator, we mention that, according to the Web of Knowledge, the seminal paper Carvajal-Rodríguez *et al.* (2009) has been cited 19 times only in one year (2012).

In practice, the test statistics along the multiple tests may be dependent. An adaptation of SGoF method to the case of dependent tests based on the beta-binomial model was introduced by de Uña-Álvarez (2012); this beta-binomial SGoF (or BB-SGoF) shares the main properties of original SGoF while taking the serial dependence of the p-values into account. Unlike for SGoF, the practical performance of its beta-binomial extension has not been extensively investigated so far. For example, evaluation of the FDR and the power of BB-SGoF is presently missing; similarly, it is still unclear how BB-SGoF will

perform when the underlying assumptions for the beta-binomial model are violated. Although de Uña-Álvarez (2012) reported a simulation study, it was restricted to the beta-binomial case; besides, due to the design of that study, it did not allow to distinguish the p-values coming from true and non-true nulls and, therefore, only the total amount of rejections and the family-wise rejection rate (rather than FDR or power) could be computed. This paper aims to fill these gaps.

In this paper we investigate the power, FDR and conservativeness of BB-SGoF methods through an intensive Monte Carlo simulation study. The organization of the paper is as follows. In Section 2 we briefly revisit SGoF and BB-SGoF procedures. In Section 3 the simulated scenarios are described. Simulation results are reported and commented in Section 4. Finally, in Section 5 we give the main conclusions of our research.

2 SGoF and BB-SGoF revisited

2.1 SGoF method

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 distribution of the n p-values attached to the null hypotheses being tested, and let γ be an initial significance level, typically $\gamma = 0.05$. Under the complete null that all the n null hypotheses are true, the expected amount of p-values below γ is just $n\gamma$ and therefore if $nF_n(\gamma)$ is much larger than $n\gamma$, one gets evidence about the existence of a number of non-true nulls, or effects, among the n tests. Let F be the underlying distribution function of the p-values; SGoF multitest starts by performing a standard one-sided binomial test for $H_0 : F(\gamma) = \gamma$ versus the alternative $H_1 : F(\gamma) > \gamma$ at level α, based on the critical region

$$\frac{F_n(\gamma) - \gamma}{\sqrt{Var^{(0)}(F_n(\gamma))}} > z_\alpha,$$

where $Var^{(0)}(F_n(\gamma)) = \gamma(1-\gamma)/n$ and z_α is the $1-\alpha$ quantile of the standard normal. Here, the Gaussian distribution is used as an approximation to the binomial model since, in practice, the number of hypotheses n will be large (hundreds, thousands, etc). If H_0 is rejected, then the number of effects declared by SGoF is given by

$$N_\alpha^{(0)}(\gamma) = n[F_n(\gamma) - \gamma] - n\sqrt{Var^{(0)}(F_n(\gamma))}z_\alpha + 1,$$

which is the excess in the number of observed p-values below the threshold γ when compared to the expected amount, beyond the critical point z_α. Then, SGoF claims that the effects correspond to the $N_\alpha^{(0)}(\gamma)$ smallest p-values. In this metatest, the FWER and the FDR are controlled at level α in the weak sense (Carvajal-Rodríguez *et al*, 2009). SGoF method relates to the notion of

second level significance testing (or higher criticism) introduced by Tukey in 1976, and further explored by Donoho and Jin (2004, 2008).

A more conservative version of SGoF is obtained when declaring as true effects the $N_\alpha^{(1)}(\gamma)$ smallest p-values, where

$$N_\alpha^{(1)}(\gamma) = n[F_n(\gamma) - \gamma] - n\sqrt{Var^{(1)}(F_n(\gamma))}z_\alpha + 1,$$

and where $Var^{(1)}(F_n(\gamma)) = F_n(\gamma)(1 - F_n(\gamma))/n$. In this conservative version, the variance is estimated without any restriction, which has two important consequences. First, since often $\gamma < F_n(\gamma) < 0.5$, it turns out that $Var^{(1)}(F_n(\gamma)) > Var^{(0)}(F_n(\gamma))$ and, therefore, $N_\alpha^{(1)}(\gamma) < N_\alpha^{(0)}(\gamma)$, leading to a smaller amount of rejections compared to original SGoF. Second, the value $N_\alpha^{(1)}(\gamma)$ may be regarded as the lower bound of a $100(1-\alpha)\%$ confidence interval for $n(F(\gamma) - \gamma)$, which in its turn is smaller than the expected number of non-true nulls with p-value below γ; indeed, by using the conservative version of SGoF, one ensures that the number of false discoveries among the p-values below γ is smaller than the number of non-discoveries with probability $1 - \alpha$, which is a reasonable error criterion (de Uña-Álvarez, 2012).

Unlike for other multitesting procedures, the power of SGoF increases with the number of tests n. The reason for this is in the $-n\sqrt{Var}$ term appearing in the number of rejection, which decreases as n grows. Besides, since SGoF imposes no strong control of FWER nor FDR, in many instances its power is often greater than FWER- or FDR-controlling methods. Simulations and examples provided in Carvajal-Rodríguez *et al* (2009) and de Uña-Álvarez (2011, 2012) indicate that this is indeed the case when the number of tests is large, and there is a relatively small to moderate proportion of weak effects. Summaryzing, SGoF provides a flexible criterion of significance for multitesting problems, offering a good balance between error control and power. Unfortunately, SGoF method (in both its original and conservative versions) is very sensitive to correlation among the tests and, indeed, it may be very anticonservative (it tends to reject more than it should) in dependent scenarios, where it loses its weak FDR control (de Uña-Álvarez, 2012). This motivates the correction of SGoF reviewed in the next section.

2.2 BB-SGoF method

BB-SGoF (from Beta-Binomial SGoF, de Uña-Álvarez, 2012) is a correction of SGoF for correlated tests. It assumes that there exist k independent blocks of correlated p-values, where k is unknown. As SGoF, BB-SGoF makes a decision on the number of effects with p-values smaller than γ, but depending on the number of blocks k and the within-block correlation.

Given the initial significance threshold γ, BB-SGoF starts by transforming the initial set of p-values $u_1, ..., u_n$ into n realizations of a Bernoulli variable: $X_i = I_{\{u_i \leq \gamma\}}, i = 1, ..., n$. Then, by assuming that there are k independent blocks of p-values of sizes $n_1, ..., n_k$ (where $n_1 + ... + n_k = n$), the number of

successes s_j within each block j, $j = 1, ..., k$, is computed. Here, $X_i = 1$ is called *success*. After that, a set of independent observations $\{(s_j, n_j), j = 1, ..., k\}$ is available, where s_j $(j = 1, ..., k)$ is assumed to be a realization of a beta-binomial variable with parameters (n_j, p, η). In this setting, $p = F(\gamma)$ represents the average proportion of p-values falling below γ, which under the complete null is just γ; while η is the correlation between two different indicators X_i and X_j inside the same block (i.e., the within-block correlation between indicators). The beta-binomial model can be regarded as a frailty model in which the random probability π of the event $X_i = 1$ is shared within each block of tests; a beta distribution is used for π, and the variance of this beta distribution is responsible for the within-block correlation η (de Uña-Álvarez, 2012).

The main aim of BB-SGoF method is to construct a one-sided confidence interval for $\tau_n(\gamma) = n(p-\gamma) = n(F(\gamma)-\gamma)$, similarly as SGoF does but considering the possible existing correlation. This confidence interval may be constructed from the asymptotic normality of the maximum-likelihood estimator (MLE) $\hat{p}$ of p. In practice, in order to perform an unrestricted optimization of the beta-binomial likelihood function, the logit reparametrization $\beta_1 = log(p/(1-p))$ and $\beta_2 = log(\eta/(1-\eta))$ is used. Introduce the $100(1-\alpha)\%$ one-sided confidence interval for $\tau_n(\gamma)$

$$I(\tau_n(\gamma)) = (n(\exp(low_1)/(1+\exp(low_1)) - \gamma), \infty)$$

where $low_1 = \hat{\beta}_1 - se(\hat{\beta}_1)z_\alpha$, with $se(\hat{\beta}_1)$ the standard error of the MLE $\hat{\beta}_1$ of β_1. Formally, BB-SGoF acts as follows. If $0 \in I(\tau_n(\gamma))$, the complete null is accepted and no effect is declared. On the contrary, if $0 \notin I(\tau_n(\gamma))$, then BB-SGoF declares as effects the smallest $N_\alpha^{BB}(\gamma; k)$ p-values, where

$$N_\alpha^{BB}(\gamma; k) = n(\exp(low_1)/(1+\exp(low_1)) - \gamma).$$

By definition, and according to the asymptotic normality of $\hat{\beta}_1$, BB-SGoF weakly controls the FWER at level α when the number of tests n is large (see de Uña-Álvarez , 2012, for details.) A crucial issue of this method is how to choose the value of k, because in practice it will be unknown. Another point to consider is the size of the blocks. A reasonable automatic choice for k is $k_N = arg\ min_k N_\alpha^{BB}(\gamma; k)$, corresponding to the most conservative decision of declaring the smallest number of effects along k. In this criterion, minimization may be performed along a grid $k = k_{min}, ..., k_{max}$ where k_{min} is the smallest number of existing blocks (i.e. the strongest allowed correlation), and $k_{max} = n/n_{min}$, where n_{min} is the smallest allowed amount of tests in each block. Here, for simplicity, it is assumed that all the blocks have the same size. This k_N ensures the weak control of FWER at the nominal level α when the number of blocks is unknown, as long as it falls between k_{min} and k_{max}.

Of course, the application of the automatic criterion k_N to choose the number of blocks entails a certain loss of power. This issue was somehow illustrated in a preliminary simulation study (Uña-Álvarez , 2012). Therefore, in practice a preliminary test for independence is recommended; if the p-values are not correlated, one may apply SGoF procedure for independent tests to increase

the power. In the setting of the beta-binomial model, Tarone (1979) introduced a procedure for testing $H_0^T : \eta = 0$ against $H_1^T : \eta > 0$; if H_0^T is true, then the beta-binomial model collapse to the binomial model and SGoF multitesting method may be applied. In the case of equal n_j's, Tarone's test is based on the Z–statistic

$$Z = \frac{n\eta_n - k}{\sqrt{2k}},$$

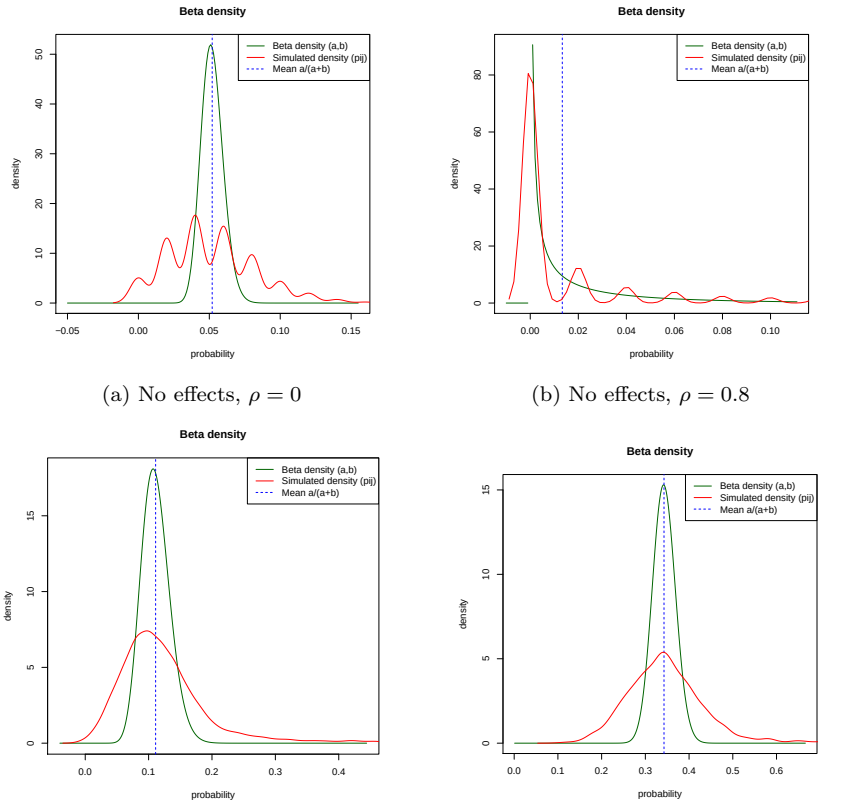
where (recall) $n = \sum_{j=1}^{k} n_j$ and η_n is a estimator of the correlation η, rejecting H_0^T for large values of Z. That is, significant positive correlation is found when η_n is large relative to its expected value under the binomial model (k/n). The ability of this test to detect dependencies in our setting is explored through simulations below.

3 Simulated scenarios

We have designed a simulated scenario similar to the study of Hedenfalk data (Hedenfalk , Duggan, et al., 2001), where the mean expression levels of about 3000 genes in two different groups A and B of individuals (with sample sizes of 7 and 8) were compared. In order to study the influence of the number of null hypotheses in the performance of the multitesting procedures, we considered the cases $n = 500$, $n = 1000$, and $n = 3000$. Hedenfalk's sample sizes of 7 and 8 were taken for groups A and B respectively. The samples were drawn from n-variate Gaussian populations with different correlation structures. The 2-sample t-test was applied to test for each null hypothesis, the sequence of n p-values coming from the computation of two-sided tails of the Student's t distribution with 13 degrees of freedom. To summarize numerical results, 1000 Monte Carlo trials were performed.

The proportion of true nulls (i.e. 'genes equally expressed') Π_0 was 1 (complete null), 0.9 (10% of effects), or 0.67 (33% of effects). Mean was always taken as zero in group A, while in group B it was μ for 1/3 of the effects and $-\mu$ for the other 2/3 of effects, with $\mu = 1$ (weak effects), $\mu = 2$ (intermediate effects), or $\mu = 4$ (strong effects). Random allocation of the effects among the n tests ('genes') was considered. Within-block correlation levels of $\rho = 0, 0.2$ and 0.8 were taken, where $\rho = 0$ means independence and $\rho = 0.8$ indicates strong correlation. With regard to the number of blocks, we considered $k = 20$, so we had 25 tests per block when $n = 500$, 50 tests per block when $n = 1000$ and 150 tests per block when $n = 3000$. For random generation, the function `rmvnorm` of the `R` software (R Core Team,2013) was used.

In Figure 1, fitted beta densities are shown for particular Monte Carlo trials in four different situations of the case $n = 1000$. The beta density was estimated by maximizing the beta-binomial likelihood based on the true number of blocks $k = 20$ and the true numbers of block sizes (50 tests per block). For this, we just note that, if $P(u_i \leq \gamma)$ follows a $Beta(a, b)$ distribution, then $p = a/(a+b)$ and $\eta = 1/(a + b + 1)$ (in the notation of Section 2.2) and, therefore, the



(a) No effects, $\rho = 0$

(b) No effects, $\rho = 0.8$

(c) 10% of intermediate effects, $\rho = 0.8$

(d) 33% of intermediate effects, $\rho = 0.8$

Figure 1: Fitted Beta densities Vs. sampling densities for particular trials in four different situations ($n = 1000$)

MLE's of a and b are directly obtained from those of p and η. Regarding the simulated density, this was estimated by the kernel method by taking as initial sample the 20 within-block proportions of p-values falling below $\gamma = 0.05$. This Figure 1 shows that the simulated trials may not fit the beta model well; indeed, it is suggested that the beta-binomial assumption may entail some underdispersion compared to the simulated scenario. This is interesting, since theoretical properties of BB-SGoF are only valid when $P(u_i \leq \gamma)$ follows a beta distribution. Simulations reported in the next Section indicate however that BB-SGoF procedure may perform well even under departures of the beta model.

BB-SGoF method with $\gamma = \alpha = 0.05$ was applied under perfect knowledge of the true value of k but also when underestimating $(k/2)$ or overestimating $(2k)$ the true number of blocks. We also applied the automatic (data-driven)

choice of k (k_N) by minimizing the number of effects declared by BB-SGoF along the grid $k = 2, ..., 61$. Since the true number of blocks was $k = 20$, the grid somehow represents the uncertainty one may have in practice.

For each situation, we computed the FDR, the power (both averaged along the 1000 Monte Carlo trials), and the proportion of trials for which the number of declared effects was not larger than the number of effects with p-value below γ (this is just 1-FDR under the complete null); as indicated in de Uña-Álvarez (2012), under the beta-binomial model BB-SGoF guarantees that this proportion (labeled as COV in Tables below) is asymptotically (i.e. $n \to \infty$) larger than or equal to $1 - \alpha$, a property which is not shared by other multitesting methods. Since the simulated models are not beta-binomial (Figure 1), it is interesting to see to what extent COV differs from 95% in our simulations. Because of the same reason, there is not guarantee that the FDR of BB-SGoF will bounded by α under the complete null, even when using the true value of k (benchmark method). Computation of these quantities for the conservative SGoF method for independent tests and for the BH method (with a nominal FDR of 5%) was also included to compare. The results are given in the following section.

4 Simulation results

Tables 1 to 3 report the results of the 1000 Monte Carlo simulations for the case n=3000 (the results for $n = 500, 1000$ were similar and they are briefly discussed in Section 4.4). In each table we report the FDR, Power (POW) and the Coverage (COV) of six methods: conservative SGoF (denoted by SGoF), BH, BB-SGoF(k) (BB-SGoF with the real number of blocks, benchmark method), BB-SGoF(k/2) (BB-SGoF when underestimating the real number of blocks), BB-SGoF(2k) (BB-SGoF when overestimating the real number of blocks), and Auto BB-SGoF (the automatic BB-SGoF procedure based on k_N). In these tables we represent by Π_0 the proportion of true nulls (1-proportion of effects).

4.1 Complete null hypothesis

First we analyze the case of no effects ($\Pi_0 = 1$), i.e. we consider the complete null hypothesis. It should be recalled that, under the complete null, all the rejected null hypotheses are Type I errors and therefore $FDR = FWER$. Obviously, the power in all these situations is 100% since there are no effects. Moreover, the coverage coincides to $1 - FDR$ as indicated above. This explains why only figures corresponding to FDR are given in Table 1.

From Table 1 we see that all the methods respect the nominal FDR of 5% fairly well in the independent setting ($\rho = 0$). For example, SGoF, BH and BB-SGoF(k) report an FDR of 0.048, 0.057 and 0.039, respectively. The automatic BB-SGoF reports a FDR below nominal (0.003), something expected due to its conservativeness. As correlation grows, original SGoF for independent tests loses control of FWER; for example, when $\rho = 0.2$, $FDR = 0.178$ and when $\rho = 0.8$

Table 1: Simulation results of $n = 3000$ tests and proportion of true nulls $\Pi_0 = 1$

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$
SGoF	0.048	0.178	0.358
BH	0.057	0.051	0.036
BB-SGoF(k)	0.039	0.056	0.07
BB-SGoF(k/2)	0.033	0.056	0.033
BB-SGoF(2k)	0.044	0.094	0.135
Auto BB-SGoF	0.003	0.023	0.017

, $FDR = 0.358$, i.e., it is 7 times the nominal. Interestingly, BB-SGoF method adapts well to the correlated settings; this is particularly true for the benchmark method which uses the true k, and for BB-SGOF(k/2) which underestimates the number of existing blocks. For instance, in the case of strong correlation ($\rho = 0.8$) these methods report a FDR of 0.07 and 0.033, respectively. Again, the automatic BB-SGoF reports a FDR below nominal, revealing its conservative nature.

On the other hand, when the researcher overestimates the number of blocks (BB-SGoF(2k)), the FDR is above the nominal (FDR=0.094 for $\rho = 0.2$ and FDR=0.135 for $\rho = 0.8$); this is because BB-SGoF decision becomes more liberal as the assumed dependence structure gets weaker. The BH method respects the nominal FDR regardless the value of ρ, something expected since the well-known robustness property of Benjamini-Hochberg method in dependence settings.

Summarizing, the results for the benchmark BB-SGoF are relevant, since they suggest FWER control (in the weak sense) even when the simulated model is not beta-binomial. Besides, since in practice the true number of blocks will be unknown, it is interesting to see that its automatic version preserves the level well.

4.2 Case of $\Pi_0 = 0.9$: 10% of effects

In this section we focus on the case of 10% of effects ($\Pi_0 = 0.9$, Table 2). When the effects are weak ($\mu = 1$), we can see that the only method which respects the FDR al level α is BH. BB-SGoF(k) shows a FDR as large as 33%, although decreasing as the correlation ρ grows. The same features are seen for the automatic BB-SGoF method. This relatively large FDR is connected with a larger power; certainly, the power of automatic BB-SGoF relative to BH is above 43 under independence (POW=0.175 and 0.004 respectively), and above 41 with $\rho = 0.2$ (POW=0.167 and 0.004 respectively). With strong correlation ($\rho = 0.8$) this rate becomes smaller (about 9). In the independent setting, this poor power of BH procedure with a large number of tests and a small proportion of weak to moderate effects was reported in previous research (Carvajal-Rodríguez et al., 2009); interestingly, the automatic BB-SGoF method is able to detect about 17% of the existing non-true nulls in situations in which BH FDR-controlling method only reports less than 1% ($\rho = 0, 0.2$).

Table 2: Simulation results of $n = 3000$ tests and proportion of true nulls $\Pi_0 = 0.9$

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.8$		
		FDR	POW	COV	FDR	POW	COV	FDR	POW	COV
$\mu = 1$	SGoF	0.334	0.200	1	0.328	0.201	0.962	0.262	0.175	0.725
	BH	0.044	0.004	1	0.048	0.004	1	0.026	0.007	0.997
	BB-SGoF(k)	0.331	0.197	1	0.314	0.183	0.998	0.195	0.109	0.870
	BB-SGoF(k/2)	0.329	0.194	1	0.313	0.183	0.995	0.166	0.087	0.959
	BB-SGoF(2k)	0.332	0.198	1	0.321	0.191	0.987	0.221	0.136	0.815
	Auto BB-SGoF	0.310	0.175	1	0.300	0.167	0.998	0.119	0.061	0.983
$\mu = 2$	SGoF	0.079	0.731	1	0.083	0.731	0.978	0.097	0.689	0.739
	BH	0.045	0.607	1	0.045	0.606	1	0.039	0.605	0.986
	BB-SGoF(k)	0.078	0.725	1	0.075	0.713	0.998	0.066	0.612	0.876
	BB-SGoF(k/2)	0.076	0.721	1	0.075	0.712	0.995	0.051	0.579	0.957
	BB-SGoF(2k)	0.079	0.727	1	0.078	0.721	0.994	0.081	0.648	0.814
	Auto BB-SGoF	0.065	0.685	1	0.066	0.683	0.999	0.041	0.534	0.982
$\mu = 4$	SGoF	0.0004	0.848	1	0.001	0.849	0.978	0.041	0.794	0.740
	BH	0.046	0.999	0	0.045	0.999	0	0.042	0.999	0.151
	BB-SGoF(k)	0.0004	0.839	1	0.0005	0.822	0.998	0.018	0.706	0.873
	BB-SGoF(k/2)	0.0004	0.833	1	0.0005	0.821	0.995	0.004	0.660	0.967
	BB-SGoF(2k)	0.0004	0.842	1	0.0007	0.834	0.996	0.032	0.747	0.811
	Auto BB-SGoF	0.0002	0.786	1	0.0003	0.781	0.999	0.002	0.603	0.986

The situation with a 10% of strong effects ($\mu = 4$) is different. In Table 2 we see that the FDR of SGoF-type methods may be very small compared to γ or α (less evident with increasing correlation). For example, SGoF(k) and Auto BB-SGoF reported a FDR of 0.0004 and 0.0002 in the case of $\rho = 0$ and 0.018 and 0.002 in the case of strong correlation ($\rho = 0.8$), respectively. This is because, with such strong effects, the p-values corresponding to the non-true nulls concentrate aroung zero, being well separated from the p-values pertaining to the true nulls; BB-SGoF method is able to detect this and to automatically report a small rate of false discoveries. On the contrary, the nominal FDR of 5% imposed by BH method is too large when the effects are strong, as it can be seen by analyzing the values of COV; even when BH is detecting almost 100% of the effects, most of the times (100% for $\rho = 0, 0.2$, 85% for $\rho = 0.8$) this is done at the price of committing more false discoveries than false non-discoveries, as long as the interest is on the p-values below $\gamma = 0.05$. On the other hand, the power of benchmark BB-SGoF is never below 70%. In these situations, BB-SGoF method offers a good compromise between power and conservativeness.

In the case of intermediate effects ($\mu = 2$), the relative results achieved by the several multitesting methods are similar to those corresponding to weak or strong effects, although FDR and power take intermediate values. In particular, the FDR of BB-SGoF methods varies between 4% and 8% (depending on the decision on the number of blocks and the existing correlation), close to the nominal 5% of BH, and therefore the power is homogeneous along the several procedures (around $50 - 70$% depending on the correlation).

An important issue is that of the losing of power when using automatic BB-

SGoF compared to the benchmark method. From Table 2 we see that the power of the auotmatic method relative BB-SGoF(k) is above 85%, with the case of strong correlation and weak effects as an exception, when it breaks down to 56%. In this case, automatic BB-SGoF deals unsuccessfully with the uncertainty on k together with the poor expectatives on the power, which are a consequence of the closeness of the alternative hypotheses to the nulls and the large correlation. One should note that a large value of ρ will result in a relatively large variance and, consequently, in a lower number of rejections when applying BB-SGoF method. Finally, we see in Table 2 that coverage values of BB-SGoF are nicely large, although they become as low as 81% with strong correlation when the number of blocks is overestimated. For BB-SGoF(k), COV is always above 87% (99.8% when $\rho \leq 0.2$), without reaching the nominal 95% in the case $\rho = 0.8$ which holds asymptotically. Since the number of tests is large ($n = 3000$), one may wonder why the coverage of the benchmark method which makes use of the true number of blocks is below 95%. A possible explanation is found in the departure of the simulated scenarios with respect to the beta-binomial assumption (Figure 1, bottom); we also mention that, with large correlation, a larger sample size n could be needed to reflect the asymptotic behaviour of a given method. Coverages reported by the automatic BB-SGoF are above 98% regardless the correlation and, therefore, in practice it may be recommended as a conservative approach.

4.3 Case of $\Pi_0 = 0.67$: 33% of effects

In Table 3 the results corresponding to a 33% of weak effects ($\Pi_0 = 0.67$) are given. Compared to Table 2, it is seen that the FDR of all the methods decreases, while the power increases; this is because the existence of a larger amount of non-true nulls. As in Table 2, in Table 3 we see that nor SGoF neither BB-SGoF are controlling for FDR at any pre-specified level. Again, the FDR attained by BB-SGoF may be regarded as a suitable proportion of false discoveries given the situation at hand; BB-SGoF(k), for example, reports a FDR of about $10 - 13\%$ with weak effects, but it goes down to about 3% and 0.01% with intermediate and strong effects respectively. The power of BB-SGoF method increases with the effect level (from weak to strong) and it decreases as the correlation grows, similarly as in Table 2.

Compared to BH approach, BB-SGoF(k) reports a relative power of about 8-13 with weak effects, being about 7-12 when comparing automatic BB-SGoF to BH; this reveals once more that BB-SGoF strategy may represent a large gain in power when the effect level is weak (true and non-true p-values well mixed). When the effects are intermediate or strong, the situation is the opposite, according to the lower FDR of BB-SGoF. However, with strong effects for example, the power of automatic BB-SGoF relative to BH is always above 84%, which again indicates a good balance between conservativeness and ability to detect true alternative hypotheses. On the other hand, conservativeness of BB-SGoF procedure may be assessed through the attained coverages; in this sense, benchmark BB-SGoF coverages are above 96% in all the situations (improving

Table 3: Simulation results of $n = 3000$ tests and proportion of true nulls $\Pi_0 = 0.67$

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.8$		
		FDR	POW	COV	FDR	POW	COV	FDR	POW	COV
$\mu = 1$	SGoF	0.135	0.301	1	0.135	0.302	1	0.124	0.301	0.901
	BH	0.036	0.023	1	0.032	0.023	1	0.024	0.031	1
	BB-SGoF(k)	0.134	0.298	1	0.131	0.293	1	0.103	0.249	0.995
	BB-SGoF(k/2)	0.134	0.297	1	0.132	0.293	1	0.103	0.249	0.997
	BB-SGoF(2k)	0.135	0.299	1	0.133	0.297	1	0.111	0.269	0.979
	Auto BB-SGoF	0.129	0.286	1	0.128	0.284	1	0.097	0.233	0.997
$\mu = 2$	SGoF	0.032	0.826	1	0.032	0.826	1	0.034	0.823	0.932
	BH	0.033	0.831	1	0.033	0.831	1	0.031	0.832	0.961
	BB-SGoF(k)	0.031	0.823	1	0.031	0.821	1	0.026	0.795	0.992
	BB-SGoF(k/2)	0.031	0.820	1	0.031	0.819	1	0.026	0.794	0.993
	BB-SGoF(2k)	0.032	0.824	1	0.032	0.823	1	0.029	0.807	0.979
	Auto BB-SGoF	0.028	0.805	1	0.028	0.803	1	0.024	0.776	0.995
$\mu = 4$	SGoF	0.0001	0.908	1	0.0001	0.907	1	0.003	0.902	0.933
	BH	0.033	0.999	0	0.033	0.999	0	0.032	0.999	0.011
	BB-SGoF(k)	0.0001	0.903	1	0.0001	0.900	1	0.0004	0.864	0.994
	BB-SGoF(k/2)	0.0001	0.900	1	0.0001	0.898	1	0.0003	0.861	0.996
	BB-SGoF(2k)	0.0001	0.905	1	0.0001	0.903	1	0.001	0.881	0.978
	Auto BB-SGoF	0.0001	0.882	1	0.0001	0.879	1	0.0002	0.840	0.998

its results with 10% of effects, see Table 2), and this percentage increases to 99.5% when considering automatic BB-SGoF. These figures may be as low as 1% or even 0% for BH (strong effects), similarly as in Table 2, situations in which this method could be regarded as too anticonservative, at least as long as COV is concerned.

Regarding the power of automatic BB-SGoF relative to the benchmark BB-SGoF, from Table 3 we see that this rate is always above 94%, the worst situation being again the case with strongest correlation and weakest effects. This improves substantially the worst rate of 56% found from Table 2 and, therefore, the presence of a larger amount of non-true nulls is beneficial to the data-driven BB-SGoF method. This improvement could be explained by the fact that, with $\mu = 1$ and $\rho = 0.8$, the automatic number of blocks k_N tends to be larger with 33% of effects than with 10% (see Table 6 below) and, consequently, Auto BB-SGoF becomes more liberal.

4.4 Influence of the number of test (n)

As mentioned at the beginning of this Section, simulations with a lower number of tests $n = 500, 1000$ were performed. In Table 4 we report the results corresponding to $n = 500$ in the case of no effects (complete null). The results in this Table 4 are similar to those in Table 1 for the case $n = 3000$; all the methods respect the nominal FDR of 5% but SGoF procedure for independent tests (which fail in the presence of correlation) and BB-SGoF when overestimating the number of blocks (it is anticonservative when $\rho = 0.8$). Results for $n = 1000$ were roughly the same and they are not shown.

Table 4: Simulation results of $n = 500$ tests and proportion of true nulls $\Pi_0 = 1$

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$
SGoF	0.051	0.092	0.288
BH	0.056	0.055	0.037
BB-SGoF(k)	0.043	0.056	0.059
BB-SGoF(k/2)	0.043	0.058	0.045
BB-SGoF(2k)	0.041	0.070	0.136
Auto BB-SGoF	0.013	0.017	0.025

In Table 5 we report the FDR, the power and the coverage of the several methods when the proportion of effects is 10% and $n = 500$. The features one can appreciate here are similar to those in Table 2. However, due to the smaller number of tests, the FDR and power of BB-SGoF method are smaller. This is because the property of SGoF-type methods, for which the power is an increasing function of n. According to this, the coverages of BB-SGoF get better; for example, for BB-SGoF(k) they are always above 93%, and this increases to 98% for automatic BB-SGoF. These coverages may be very low for BH with strong effects (similarly as in the case $n = 3000$), ranging from 11.6% in the independent setting to 41.5% when $\rho = 0.8$.

Table 5: Simulation results of $n = 500$ tests and proportion of true nulls $\Pi_0 = 0.9$

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.8$		
		FDR	POW	COV	FDR	POW	COV	FDR	POW	COV
$\mu = 1$	SGoF	0.282	0.137	0.994	0.270	0.136	0.986	0.228	0.135	0.835
	BH	0.0425	0.012	1	0.050	0.013	1	0.023	0.016	0.997
	BB-SGoF(k)	0.269	0.130	0.998	0.255	0.125	0.992	0.162	0.084	0.943
	BB-SGoF(k/2)	0.268	0.128	0.996	0.249	0.125	0.994	0.135	0.074	0.967
	BB-SGoF(2k)	0.272	0.132	0.996	0.261	0.123	0.991	0.185	0.106	0.909
	Auto BB-SGoF	0.219	0.093	1	0.209	0.089	0.998	0.095	0.050	0.989
$\mu = 2$	SGoF	0.053	0.623	0.999	0.057	0.622	0.995	0.077	0.603	0.861
	BH	0.044	0.607	1	0.045	0.607	1	0.042	0.609	0.972
	BB-SGoF(k)	0.051	0.609	0.999	0.052	0.605	0.998	0.053	0.535	0.937
	BB-SGoF(k/2)	0.050	0.608	0.999	0.052	0.602	0.999	0.046	0.516	0.967
	BB-SGoF(2k)	0.051	0.613	0.999	0.054	0.611	0.997	0.062	0.566	0.913
	Auto BB-SGoF	0.039	0.537	1	0.040	0.539	0.999	0.037	0.455	0.982
$\mu = 4$	SGoF	0.0001	0.711	0.999	0.0006	0.711	0.995	0.022	0.699	0.869
	BH	0.044	0.999	0.116	0.045	0.999	0.130	0.042	0.999	0.415
	BB-SGoF(k)	0.0001	0.695	0.999	0.0003	0.688	0.998	0.008	0.619	0.936
	BB-SGoF(k/2)	0.0001	0.692	0.999	0.0002	0.686	0.998	0.003	0.592	0.972
	BB-SGoF(2k)	9.573e-05	0.698	1	0.0004	0.696	0.997	0.011	0.655	0.917
	Auto BB-SGoF	9.924e-05	0.611	1	0.0001	0.609	0.999	0.001	0.519	0.988

The power of BB-SGoF(k) is about 5-11 times that of BH with weak effects, but it may be as low as 0.6 with strong effects and strong correlation. Therefore, performance of BB-SGoF relative to BH is poorer with a smaller n; this

reinforces the fact that BB-SGoF is more suitable for multitesting problems in high dimensions. Interestingly, the power of automatic BB-SGoF relative to benchmark SGoF is above 60% in all the cases (the worst situation is again that with weak effects and $\rho = 0.8$). Compared to the figures in Table 2, it is seen that a smaller number of tests is not beneficial for the automatic BB-SGoF criterion, for which the power relative to that of BB-SGoF(k) is 87% with $n = 3000$ (averaging the nine simulated scenarios) but only 81% with $n = 500$.

Results for $n = 500$ and 33% of effects were also obtained, as well as results for the case $n = 1000$ with 10% or 33% of effects. These results (not shown) provided no other relevant evidences than those discussed above.

4.5 Automatic choice of the number of blocks

Automatic BB-SGoF implements a preliminary estimation of the number of blocks of dependent p-values. Since this estimation is performed on the basis of a conservative criterion (this is, to minimize the number of rejections), it does not lead in general to a precise approximation of the true k. In order to illustrate this point, we report in Table 6 the number of blocks detected on average (i.e. the mean of k_N) and its standard deviations (in brackets), in the case $n = 3000$. Recall that the true number of blocks is 20.

Table 6: Number of blocks detected on average and it standard deviations (in brackets), $n = 3000$

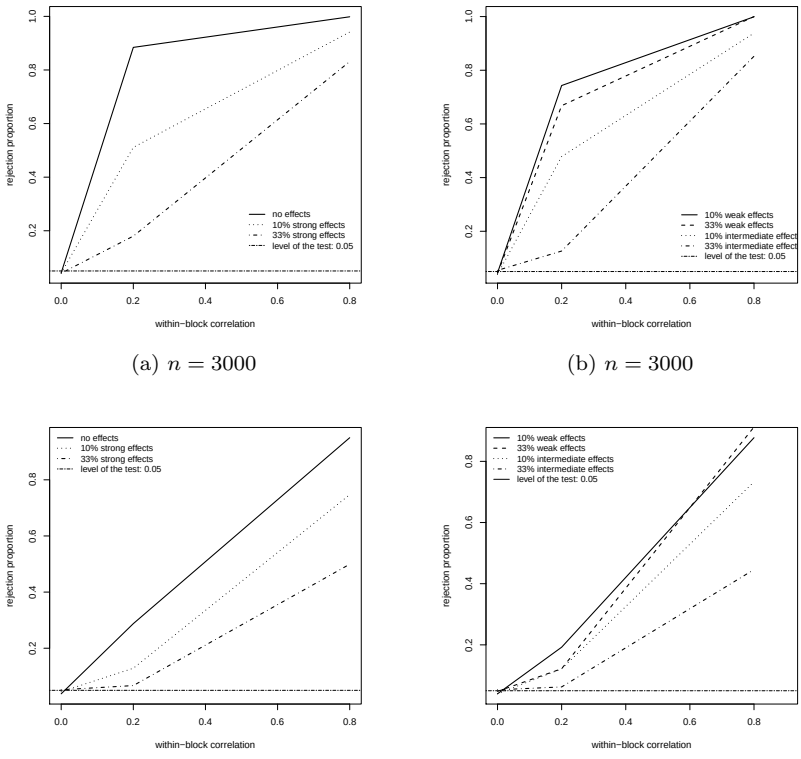
	$\mu = 1$			$\mu = 2$		$\mu = 4$	
	$\Pi_0 = 1$	$\Pi_0 = 0.9$	$\Pi_0 = 0.67$	$\Pi_0 = 9$	$\Pi_0 = 0.67$	$\Pi_0 = 0.9$	$\Pi_0 = 0.67$
$\rho = 0$	6.43(11.30)	3.67(8.09)	1.93(4.51)	2.93(6.67)	1.25(1.23)	2.59(6.41)	1.25(1.84)
$\rho = 0.2$	6.02(6.11)	4.63(5.45)	3.62(4.48)	3.02(4.51)	1.37(1.65)	3.04(4.19)	1.48(1.67)
$\rho = 0.8$	6.09(5.46)	4.74(4.51)	7.62(6.39)	4.42(4.29)	5.71(5.54)	4.32(3.82)	5.35(5.03)

Results in Table 6 indicate that k_N strongly underestimates the value of k, which is a result of the conservativeness of the underlying criterion. Note that fewer blocks represents a situation with a stronger dependence structure and, consequently, a smaller number of rejections when applying BB-SGoF. More specifically, under the complete null, the average of k_N is about 6, with a standard deviation which decreases as the correlation increases. With weak effects, this average varies depending on the proportion of effects and the correlation degree; the same happens with intermediate or strong effects. Roughly, it is seen that the average (also the standard deviation) of k_N decreases as the effect level changes from weak to strong, while it increases with the correlation. Therefore, it seems that k_N is protecting BB-SGoF against situations in which the amount of rejections could be too large, due to the strength of the effects or the low correlation. Interestingly, a larger proportion of effects results in a smaller value of k_N when $\rho = 0.2$ but the opposite is true for $\rho = 0.8$, so no general conclusion can be given to this regard. Overall, it can be said that k_N plays an important role when looking for conservativeness but it is a biased,

highly dispersed estimator of the true number of blocks.

4.6 Tarone test

As mentioned in Section 2, Tarone (1979) introduced a test for the binomial model $H_0^T : \eta = 0$ against the beta-binomial alternative $H_1^T : \eta > 0$. Here we denote by η the correlation between Bernoulli outcomes $I_{\{u_i \leq \gamma\}}$ sharing the same block, which is different from ρ, the correlation between the normally distributed 'gene expression levels' in our simulations (but we have $\eta = 0$ when $\rho = 0$). In Figure 2 we show the rejection proportion (along the 1000 simulations) of Tarone's test performed at level 0.05, in the case when the value of k is correctly specified, for several correlation degrees $\rho = 0, 0.2, 0.8$ (represented in the x axis) and $n = 3000, 500$ (top and bottom, respectively).



(a) $n = 3000$ (b) $n = 3000$

(c) $n = 500$ (d) $n = 500$

Figure 2: Proportion of rejections of the Tarone's test.

In Figure 2 we can see that, when $\rho = 0$, the rejection proportion is about 5%, indicating that Tarone's test respects the level well. As ρ departs from zero,

the rejection proportion grows, and the same happens when moving from the case $n = 500$ to $n = 1000$; both features were of course expected. In general, the power of Tarone's test decreases as the proportion and/or the level of the effects (non-true nulls) increase; this suggests that the presence of effects introduces noise when testing for correlation.

5 Main conclusions

BB-SGoF method may control the FWER in the weak sense even when the underlying model is not beta-binomial. This suggests that the beta-binomial model may have enough flexibility to represent the correlation structure among the tests in practice. BB-SGoF method is also robust with respect to miss-specification of the number of existing blocks, although it becomes too liberal when this parameter is overestimated. The automatic BB-SGoF procedure performs well, with only a moderate loss of power ($5 - 15\%$) with respect to the benchmark version in most of the cases. However, when there is a small proportion (10%) of weak effects, this loss of power may be as large as 44%, particularly when the correlation within the blocks of tests is strong. Therefore, more efforts are needed to select the unknown number of blocks in an automatic (data driven) way. Another interesting finding of our simulation study is the ability of Tarone's test to detect dependence in practice. When the null hypothesis of no correlation is accepted, application of original SGoF (rather than BB-SGoF) is recommended. As SGoF for independent tests, BB-SGoF method is liberal with respect to the FDR or the FWER, and this explains why it is able to exhibit a good power in difficult situations where FWER and FDR controlling procedures fail to detect non-true nulls (this is typically the case when the number of tests is large, and there is a small to moderate proportion of weak effects). Furthermore, conservativeness of BB-SGoF has been assessed; more specifically, BB-SGoF method ensures that, with large probability, the number of false discoveries will not exceed the number of false non-discoveries (at least when the focus is on the p-values below a given threshold), thus offering a good compromise between false discovery rate and power.

6 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] 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.

[3] de Uña-Álvarez, J. (2011). On the statistical properties of SGoF multitesting method. *Statistical Applications in Genetics and Molecular*, **10**, Issue 1, Article 18.

[4] 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.

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

[6] Donoho, D. and 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.

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

[8] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society B*, **64**, 499–518.

[9] Genovese, C.and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* **32**, 1038–1061.

[10] Lehmann, E.L. and Romano, J.P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, **33**, 1138–1154.

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

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

[13] Storey, J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value.*Annals of Statistics*, **31**, 2013–2035.

[14] Tarone, R.E.(1979). Testing the goodness of fit of the binomial distribution. *Biometrika*, **66**, 585–590.

[15] van der Laan, M.J., Dudoit, S. and Pollard, K.S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 15.