Universidade de Vigo

The Beta-Binomial SGoF method for multiple dependent tests

Jacobo de Uña Álvarez

Report 12/01

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 de Vigo

The Beta-Binomial SGoF method for multiple dependent tests

Jacobo de Uña Álvarez

Report 12/01

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita:

Universidade de 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

The Beta-Binomial SGoF method for multiple dependent tests

Jacobo de Uña-Álvarez, University of Vigo

February 2012

Abstract

In this paper a correction of SGoF multitesting method for dependent tests is introduced. The correction is based in the beta-binomial model, and therefore the new method is called Beta-Binomial SGoF (or BB-SGoF). Main properties of the new method are established, and its practical implementation is discussed. BB-SGoF is illustrated through the analysis of two different real data sets on gene/protein expression levels. The performance of the method is investigated through simulations too. One of the main conclusions of the paper is that SGoF strategy may have much power even in the presence of possible dependences among the tests.

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, the family-wise error rate (FWER) and the false discovery rate (FDR), among other measures, have been proposed as suitable significance criteria to perform the multiple testing. See Benjamini and Hochberg (1995), Nichols and Hayasaka (2003) or Dudoit and Laan (2008) for basic definitions and reviews of existing literature.

As a drawback of the FWER- and FDR-based methods, their power may be rapidly decreased as the number of tests grows, 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 (from Sequential Goodness-of-Fit), which focuses on the difference between the observed proportion of p-values below a given significance threshold (the γ parameter) and the expected one under the complete null of no effects (γ). This relates to the notion of second-level significance testing or *higher criticism* introduced by Tukey in 1976, and further extended in Donoho and Jin

(2004). SGoF 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; Carvajal-Rodríguez and de Uña-Álvarez, 2011a; de Uña-Álvarez, 2011). The theoretical statistical properties of SGoF were investigated in detail in de Uña-Álvarez (2011). It was illustrated that, with a large number of tests, the critical region provided by SGoF is wider than that of FDR-based methods in most scenarios. Both SGoF original method and its extensions provide reliable inference when the multiple tests are independent. In this paper, a correction of SGoF for possibly dependent tests is introduced.

In practice, dependences among the tests are found in many situations; see for example Efron (2007, 2010), Romano et al. (2008), or Carvajal-Rodríguez and de Uña-Álvarez (2011b). This dependence may be provoked by the existence of k blocks of tests which share the same within-block probability π_j ($j = 1, ..., k$) of reporting a significative p-value: the larger the variance of the π_j's, the greater the within-block correlation (see Section 2). In Figure 1 we report the density of the π_j's estimated for the two real data sets considered in Section 3; in this Figure, the expected value of π_j when all the null hypotheses are true is $\gamma = 0.05$. Hedenfalk data (Figure 1, top) provide a within-block correlation of 0.027, the mean and variance of the π_j's being 0.186 and 0.004 respectively. On the other hand, Diz data (Figure 1, bottom) give a correlation of 0.035, with mean and variance of the within-block probability of 0.099 and 0.003. The within-block correlation is significative at level 0.01 for Hedenfalk data (p=4.86e-11), but not for Diz data (p=0.013), see Sections 2 and 3 for details. In other words, at level 0.01 it is accepted that the density in Figure 1, bottom, degenerates at the mean, while a significative variance is present in Figure 1, top. The location of Hedenfalk data's density to the right of Diz data's suggests the presence of a larger amount of true effects or features along the performed tests; however, this should be carefully assessed at the light of the possible existing correlation.

While FDR-based strategies are robust in dependence scenarios, the same is not true for SGoF, which crucially depends on the correct estimation of the variance associated to the number of discoveries. In most practical situations with dependent tests, the final number of discoveries reported by SGoF will be too liberal, because it will be based on an underestimated variance (Owen, 2005). As an example, for the two real data sets of Figure 1, SGoF reports 428 (Hedenfalk data) and 8 (Diz data) discoveries; however, with the correction of SGoF for dependence introduced in this paper, these numbers translate into 389 (Hedenfalk) and 1-2 (Diz) discoveries respectively. Therefore, taking the possible dependences into account may have a big impact in the final decision of the researcher. Still, it will be noted that these corrected amounts of discoveries are much larger than when willing to control the FDR at level 5%.

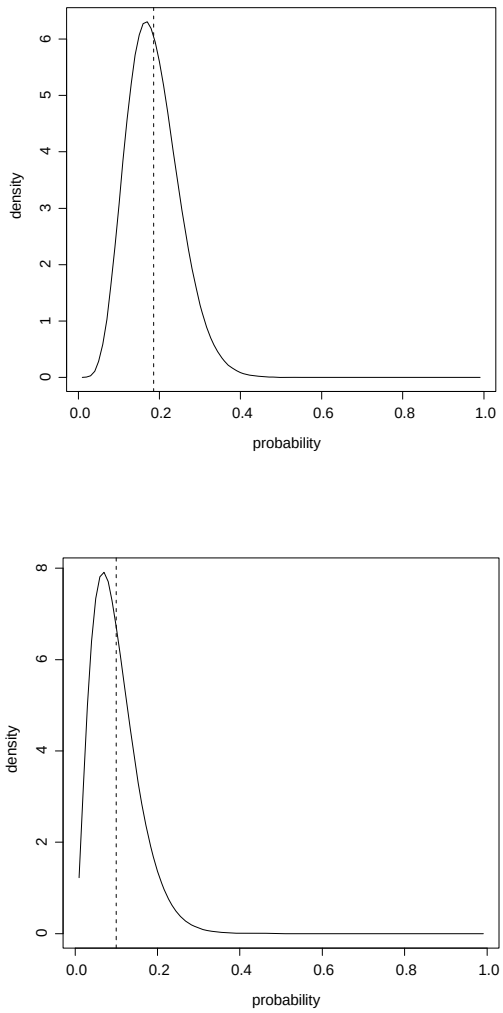

Figure 1. Beta density fitted to the probability of reporting a p-value below threshold 0.05. Top: Hedenfalk data with 266 blocks. Bottom: Diz data with 4 blocks. The dashed line corresponds to the mean probability.

The rest of the paper is organized as follows. In Section 2 we introduce the correction of SGoF to deal with dependent tests. This correction is based on the beta-binomial extension of the binomial model, so some notation and existing results for this model will be needed. Main properties of the method and its practical implementation will be discussed. The new method is used to revisit two real data sets in Section 3, which were previously analyzed under the assumption of independence (Storey and Tibshirani, 2003; Carvajal-Rodríguez

and de Uña-Álvarez, 2011a; de Uña-Álvarez, 2011). As announced, it will be seen that allowing for dependences may influence the results. A simulation study is reported in Section 4. Finally, Section 5 contains the main conclusions of this work and also discusses some open problems and questions which are left for future research.

2 The method

2.1 SGoF revisited

Let $u_1, ..., u_n$ be a set of p-values corresponding to n tests which are performed in a simultaneous way, and let F_n be the empirical distribution function of the u_i's, that is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(u_i \leq x),$$

where $I(A)$ is the indicator function of the event A. Let $\gamma \in (0, 1)$ be an initial significance level, usually taken as $\gamma = 0.05$. Under the complete (or intersection) null that all the n null hypotheses are true (i.e., no effects), the expected amount of p-values below γ is just $n\gamma$. On the other hand, when $nF_n(\gamma)$ 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 (Carvajal-Rodríguez et al., 2009; de Uña-Álvarez, 2011) starts by performing a standard one-sided binomial test for $H_0 : F(\gamma) = \gamma$ versus the alternative $H_1 : F(\gamma) > \gamma$, 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 where z_α is the $1-\alpha$ quantile of the standard normal. Here, $\alpha = \gamma$ is usually taken. If H_0 is accepted, then there is no evidence against the complete null, and no effect is declared by SGoF. If H_0 is rejected, the number of effects declared by SGoF is given by

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

which is the excess in the number of observed p-values below threshold γ when compared to the expected amount, beyond the critical point z_α. More specifically, SGoF ends by claiming that the effects correspond to the $N_\alpha^{(0)}(\gamma)$ smallest p-values in the sequence $u_1, ..., u_n$. In this metatest, the FWER is controlled at level α in the weak sense (Carvajal-Rodríguez et al., 2009), which means that, under the complete null, the probability of rejecting one or more than one true nulls is not larger than α.

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

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

and where $Var^{(1)}(F_n(\gamma)) = F_n(\gamma)(1 - F_n(\gamma))/n$. In most practical applications we will have $\gamma < F_n(\gamma) < 1/2$ and hence $Var^{(0)}(F_n(\gamma)) < Var^{(1)}(F_n(\gamma))$. In other words, typically the variance estimated under the complete null will be smaller than the variance estimated without any restriction, the latter leading to a less liberal decision. The value $n^{-1}N_\alpha^{(1)}(\gamma)$ may be regarded as the lower limit of a one-sided $100(1-\alpha)\%$ confidence interval for $F(\gamma) - \gamma$. In its turn, under a mixture model this quantity $F(\gamma) - \gamma$ is a lower bound for the proportion of true effects (i.e. non-true nulls) with p-value smaller than γ, which gives an interesting alternative interpretation of SGoF multitest (de Uña-Álvarez, 2011). Unfortunately, when the tests are dependent, none of the variances in $N_\alpha^{(0)}(\gamma)$ and $N_\alpha^{(1)}(\gamma)$ above are correct, and therefore SGoF must be re-defined.

SGoF method is based on the binomial distribution, which serves as a null model for the test statistic $nF_n(\gamma)$ when the tests are independent. An extension of the binomial model which allows for correlated Bernoulli outcomes is the beta-binomial distribution (see e.g. Johnson and Kotz, 1970). The beta-binomial model is the basis for the correction of SGoF introduced in the next Section.

2.2 The Beta-Binomial SGoF

Before introducing the Beta-Binomial SGoF (BB-SGoF) method, we recall some basic definitions and expressions related to the beta-binomial model. The beta-binomial model states that the number of successes S among n trials is conditionally ruled by

$$P(S = s|\pi) = \binom{n}{s}\pi^s(1-\pi)^{n-s}, \qquad s = 0, ..., n,$$

where π is a random variable following the Beta(a, b) density ($a > 0$, $b > 0$)

$$g(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)}, \qquad 0 < \pi < 1,$$

and where $B(.,.)$ is the beta function. The resulting unconditional law for S is

$$P(S = s) = \binom{n}{s}B(a+s, n+b-s)/B(a,b), \qquad s = 0, ..., n.$$

The beta-binomial model has been used in several applications, including the analysis of point quadrat data (Kemp and Kemp, 1956), the consumer purchasing behavior (Chatfield and Goodhart, 1970), the household distribution of

incidence of disease (Griffiths, 1973), toxicological experiments (Williams, 1975) and, more recently, in proteomics (Pham et al., 2010). See Johnson and Kotz (1970) for illustrations of the model.

In this model, the expected probability of success $p = E(\pi)$ and the correlation between the outcomes of two different trials ρ are given respectively by $p = a/(a+b)$ and $\rho = 1/(a+b+1)$. This correlation comes from the fact that two outcomes will share the same random π and, numerically, it equals the variance of the beta density ($Var(\pi)$) relative to that of the expected binomial ($p(1-p)$). Certainly, assume that $S = \sum_{i=1}^{n} X_i$ where $X_i \in \{0,1\}$ is the outcome of the i-th Bernoulli trial ($X_i|\pi \sim Ber(\pi)$), and where $X_1, ..., X_n$ are conditionally independent given π. Then, for $i \neq j$,

$$E\left(X_i X_j\right) = E\left[E(X_i X_j|\pi)\right] = E\left[E(X_i|\pi)E(X_j|\pi)\right] = E\left(\pi^2\right)$$

and hence

$$\begin{aligned} Cov(X_i, X_j) &= E\left(X_i X_j\right) - E\left(X_i\right)E(X_j) = E\left(\pi^2\right) - p^2 = Var(\pi) \\ &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

On the other hand, since $E(X_i^2|\pi) = p$, we have

$$Var(X_i) = E(X_i^2) - E(X_i)^2 = E\left[E(X_i^2|\pi)\right] - p^2 = p(1-p).$$

Thus, from $p = a/(a+b)$ we obtain

$$\rho = Cor(X_i, X_j) = \frac{Var(\pi)}{p(1-p)} = \frac{1}{a+b+1}.$$

It also happens that the mean and variance of the number of successes are $E(S) = np$ and $Var(S) = np(1-p)(1+(n-1)\rho)$. Therefore, the beta-binomial model allows for a variance larger than binomial whenever $\rho > 0$, while it reduces to the standard binomial when $\rho = 0$. Often, the alternative parametrization (p, θ) of this model has been suggested, where $\theta = 1/(\alpha+\beta)$, the case $\theta = 0$ corresponding again to no correlation (Pham et al., 2010). Note that $\rho = \theta/(\theta+1)$.

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, ρ), where $n_1, ..., n_k$ may be distinct. In this setting, $p = E(\pi) = F(\gamma)$ represents the average proportion of p-values falling

below γ, which under the complete null is just γ; while ρ is the correlation between two different indicators X_i and X_j inside the same block (i.e. the within-block correlation).

Obviously, the expected probability of success $p = E(\pi)$ may be estimated by $p_n = \sum_{j=1}^{k} s_j / \sum_{j=1}^{k} n_j$. A simple estimator for the correlation ρ is given by $\rho_n = \sigma^2_{\widehat{p}_j}/(p_n(1-p_n))$, where $\sigma^2_{\widehat{p}_j}$ is the sample variance of

$$\widehat{p}_j = s_j/n_j, \qquad j = 1, ..., k.$$

Tarone (1979) introduced a test for the binomial model $H_0^T : \rho = 0$ against the beta-binomial alternative $H_1^T : \rho > 0$, which in the case of equal n_j's is based on the $Z-$statistic

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

where (recall) $n = \sum_{j=1}^{k} n_j$, rejecting H_0^T for large values of Z. That is, significant positive correlation is found when ρ_n is large relative to its expected value under the binomial (k/n). See Tarone (1979) for details on the testing procedure when the block sizes are unequal.

Model-based estimators for p and ρ may be derived by maximum-likelihood principles under the beta-binomial assumption. Explicitly, the log-likelihood of the (s_j, n_j)'s is given by (cfr. Tarone, 1979)

$$\begin{aligned} L(p,\rho) &= a_k + \sum_{j=1}^{k} \{s_j \log p + (n_j - s_j)\log q\} + \\ &\quad + \sum_{j=1}^{k} \log\left[1 + \frac{\rho}{2p^2q^2}\left\{(s_j - n_j p)^2 + s_j(2p-1) - n_j p^2\right\}\right] \end{aligned}$$

where $q = 1 - p$ and where a_k is a constant involving only the observations. As usual with maximum-likelihood estimates, the maximizer $(\widehat{p}, \widehat{\rho})$ of L on the $[0,1] \times [0,1]$ rectangle is an efficient, asymptotically normal estimator of (p, ρ). In practice, these estimators and their standard errors may be computed from existing free software; to this end, the function vglm of the R package VGAM has been used in Sections 3 and 4. The main goal of BB-SGoF is to provide inferences on the value of $p = F(\gamma)$, while allowing for dependences among the tests ($\rho > 0$). More specifically, BB-SGoF aims to construct a one-sided confidence interval for the excess of significant cases $\tau_n(\gamma) = n(p - \gamma)$, similarly as original SGoF does but considering the possible existing correlation. This confidence interval may be constructed from the asymptotic normality of $\widehat{p}$. We give now the details.

Consider the reparametrization of the beta-binomial model given by the logit transformation of p and ρ, that is $\beta_1 = \log(p/(1-p))$, or $p = \exp(\beta_1)/(1+$

$\exp(\beta_1))$, and $\beta_2 = \log(\rho/(1-\rho))$, or $\rho = \exp(\beta_2)/(1+\exp(\beta_2))$. With this reparametrization, an unrestricted maximization of the likelihood can be performed. Let $\widehat{\beta}_i$, $i = 1, 2$, the maximizers of the likelihood. The following $100(1-\alpha)\%$ confidence intervals for β_1 and β_2 can be computed:

$$I(\beta_i) = \left(\widehat{\beta}_i \pm se(\widehat{\beta}_i) z_{\alpha/2}\right), \qquad i = 1, 2,$$

where $se(\widehat{\beta}_i)$ denotes the estimated standard error of $\widehat{\beta}_i$, and where $z_{\alpha/2}$ stands for the $(1-\alpha/2)$-quantile of the standard normal distribution. Respectively, confidence intervals for p and ρ may be obtained by logit-backtransforming the limits of $I(\beta_i)$,

$$low_i = \widehat{\beta}_i - se(\widehat{\beta}_i) z_{\alpha/2}, \qquad upp_i = \widehat{\beta}_i + se(\widehat{\beta}_i) z_{\alpha/2},$$

as follows:

$$I(p) = \left(\exp(low_1)/(1+\exp(low_1)), \exp(upp_1)/(1+\exp(upp_1))\right),$$

$$I(\rho) = \left(\exp(low_2)/(1+\exp(low_2)), \exp(upp_2)/(1+\exp(upp_2))\right).$$

As mentioned, of particular interest is the $100(1-\alpha)\%$ one-sided confidence interval for $\tau_n(\gamma) = n(p-\gamma)$, since this parameter represents the excess of significant features (at level γ) with respect to the expected amount under the complete (or intersection) null. Therefore, we consider the interval

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

where low_1 is as above but with α in the place of $\alpha/2$. Note that a one-sided rejection region at level α for the complete null $H_0 : p = \gamma$ against the alternative $H_1 : p > \gamma$ is given by $\{0 \notin I(\tau_n(\gamma))\}$. 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, an according to the asymptotic normality of $\widehat{\beta}_1$, BB-SGoF weakly controls the FWER at level α when the number of tests n is large. It is also clear that the number of declared effects $N_\alpha^{BB}(\gamma; k)$ will grow with the number of tests n. This is because $se(\widehat{\beta}_1)$ goes to zero at a $\sqrt{n}$-rate, and therefore the lower limit $low_1 = \widehat{\beta}_1 - se(\widehat{\beta}_1) z_\alpha$ is shifted-up towards $\widehat{\beta}_1$ as $n \to \infty$. In practice, this translates into a power of BB-SGoF which increases with the number of tests, a property which is not shared by other multiple tests adjustments as e.g. FDR-controlling procedures (Carvajal-Rodríguez et al., 2009). Another consequence of the definition of BB-SGoF method is that the influence of the FWER-controlling parameter α is small or even negligible when

the number of tests is large; that is, moving from $\alpha = 0.05$ to *e.g.* $\alpha = 0.001$ will have almost no impact in $N_\alpha^{BB}(\gamma; k)$ when n is large since the normal quantile z_α will be divided by $\sqrt{n}$. Finally, it is also interesting that the threshold p-value reported by BB-SGoF, i.e. $F_n^{-1}(N_\alpha^{BB}(\gamma; k)/n)$, will be approximately $F^{-1}(F(\gamma) - \gamma)$ as n grows; this threshold is below the initial significance level γ regardless the shape of the cumulative distribution of the p-values F. All these properties of BB-SGoF were also indicated for the original SGoF formulation for independent tests (de Uña-Álvarez, 2011).

A crucial practical issue is how to choose the value of k; once k is fixed, the n_j's may be computed as $n_j = n/k$, $j = 1, ..., k$, so every block has the same size. Few independent blocks (k small) implies a strong correlation structure in which many pairs (X_i, X_j) will be correlated. In this situation, the value of $N_\alpha^{BB}(\gamma; k)$ may be much smaller than $N_\alpha^{(0)}(\gamma)$ or $N_\alpha^{(1)}(\gamma)$. On the contrary, a large number of blocks (k large) implicitly states weak dependence, leading to a value of $N_\alpha^{BB}(\gamma; k)$ which may be close to the number of effects declared by original SGoF for independent tests. See Figures 2 and 4, bottom, in which values of $N_\alpha^{BB}(\gamma; k)$ for several k's are reported for two real data sets. A possible suggestion is to display first a set of results corresponding to different decisions on k. Values of $\widehat{p}$, $\widehat{\rho}$, $N_\alpha^{BB}(\gamma; k)$ and p-values of Tarone's test may be explored to get information on the correlation structure and the possible number of existing effects when k varies. On the other hand, if one wills to select k in an automatic way, several criteria are possible. 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. Clearly, this k^N ensures the weak control of FWER at the nominal level α as long as the number of existing blocks falls between $k_{\min}$ and $k_{\max}$. In the next Section we have used $k_{\min} = 2$ and $n_{\min} \approx 6$.

3 BB-SGoF in practice

In this Section we provide a detailed application of BB-SGoF to two data sets. Both data sets contain sequences of p-values corresponding to tests performed on gene or protein expression levels.

3.1 Hedenfalk data

Our first illustrative example concerns 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.

The 3,170 p-values u_i, $i = 1, ..., n$, were transformed into a 0-1 sequence according to the value of the indicator $X_i = I(u_i \leq \gamma)$, where $\gamma = 0.05$ was taken as preliminary significance level. The proportion of significative tests among the 3,170 was $F_n(\gamma) = 0.1912$. Assuming independence among the tests, the number of effects declared by SGoF at level $\alpha = 0.05$ was $N_\alpha^{(0)}(\gamma) = 428.32$, i.e. about 428 discoveries. We obtained $N_\alpha^{(1)}(\gamma) = 412.08$ effects when using the conservative version of SGoF which estimates the variance without any restriction.

The independence assumption among the tests was checked through the runs test for randomness of a dichotomous (binary) sequence (cfr. Siegel and Castellan, 1988), giving a two-sided p-value of 0.002654. The number of runs was smaller than expected (Z=-3.0052) indicating a significative positive dependence among the tests (i.e. significative genes tend to be followed by a significative gene). Under dependence, inferences provided by SGoF above are not valid and, therefore, the number of significant genes must be re-evaluated.

In Figure 2, top, we report the p-values of Tarone (1979)'s goodness-of-fit test for the binomial model against the beta-binomial alternative, when adding the binary outcomes $X_i = I(u_i \leq \gamma)$ into k blocks of equal size, $k = 2, ..., 501$. The choice $k_{\max} = 501$ gives a minimum block size of $n_{\min} = n/k_{\max} \approx 6$. The number of effects declared by BB-SGoF based on each corresponding beta-binomial alternative is also reported (Figure 2, bottom), where the cases $k = 2, 3, 5, 8$ were deleted for easier visualization, since they provided too large values ($k = 2, 8$) or negative values ($k = 3, 5$). Negative values of $N_\alpha^{BB}(\gamma; k)$ can be indeed ignored in this case, since they come from an unreliable fit of the beta-binomial model, identified because the standard error of $\widehat{\beta}_1$ was more than 25 times the median standard error along k. As explained in Section 2-3, BB-SGoF declares

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

effects, where $low_1 = \widehat{\beta}_1 - se(\widehat{\beta}_1)z_\alpha$; $N_\alpha^{BB}(\gamma; k)$ is just a $100(1-\alpha)\%$ lower limit for the excess of significant cases $\tau_n(\gamma) = n(F(\gamma) - \gamma)$ and hence it acts as a substitute for $N_\alpha^{(i)}(\gamma)$, $i = 0, 1$, under dependence.

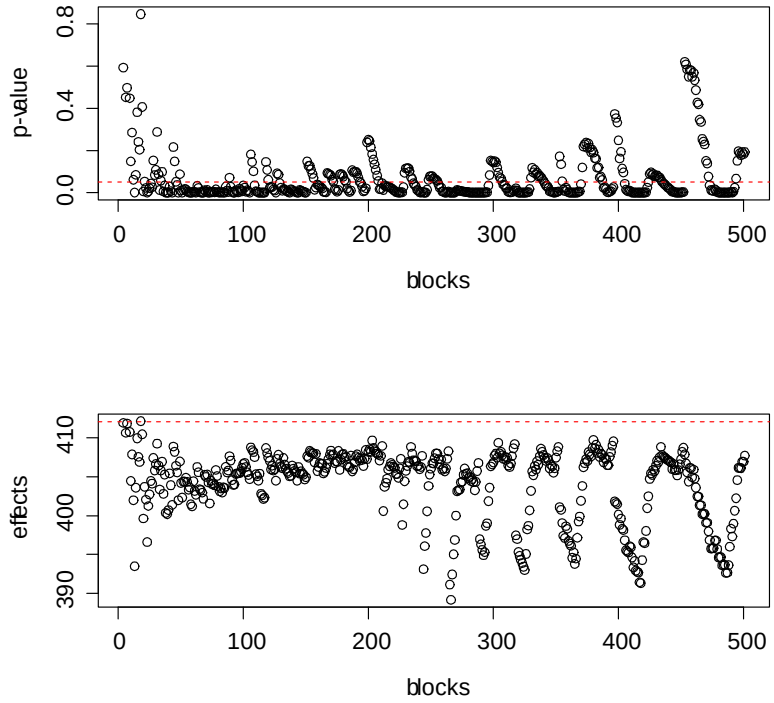

Figure 2. P-values of Tarone (1979)'s test (top) and number of effects declared by BB-SGoF (bottom) along the number of blocks k. Hedenfalk data, $\alpha = \gamma = 0.05$. The horizontal dashed line at the bottom figure corresponds to $N_\alpha^{(1)}(\gamma)$.

Figure 2, top, shows that there exists significance to reject the binomial model in favour of the beta-binomial alternative for a large range of k values. This is in agreement of the previous application of the runs test. In particular, the minimum value of $N_\alpha^{BB}(\gamma; k)$ along k is obtained for $k = 266$, namely $N_\alpha^{BB}(\gamma; k^N) = 389.1544$ or about 389 declared effects. This value of k also corresponds to the minimum p-value of Tarone's test (p=4.86e-11). This is smaller than the 412 or the 428 effects declared by the binomial SGoFs for independent tests. The reason for this is that the variance in the estimation of $p = F(\gamma)$ is larger when the tests are dependent; moreover, for the Hedenfalk data it happens that the value of $F(\gamma)$ estimated under the beta-binomial model is smaller than $F_n(\gamma)$ for most of the values of k ($\widehat{F(\gamma)} = 0.1857$ for $k = 266$), so this also provokes a decrease in $N_\alpha^{BB}(\gamma)$ when compared to $N_\alpha^{(i)}(\gamma)$. In order to illustrate this, in Figure 3 we report the estimated proportion of p-values below threshold γ ($\widehat{F(\gamma)}$) provided by the beta-binomial model along k, and the standard errors of their logit transformations, $se(\widehat{\beta}_1)$. From this Figure 3 it is

seen that the minimum $N_\alpha^{BB}(\gamma; k^N)$ corresponds to a small value of $\widehat{F(\gamma)}$ and to a local pick in the standard error of $\widehat{\beta}_1$.

It is interesting to point out that the most conservative decision provided by BB-SGoF (389 discoveries) at level $\alpha = 0.05$ is still much more powerful than that obtained from standard methods which control the FDR at 5%. Indeed, Benjamini-Hochberg FDR-based method at that level gives for this data set only 157 discoveries (based on a preliminary estimation of the proportion of effects of 28.33%, see de Uña-Álvarez, 2011), which are less than half the discoveries declared by $N_\alpha^{BB}(\gamma; k^N)$. The reason for this is that BB-SGoF only controls for FWER in the weak sense, being liberal about the proportion of false discoveries otherwise. Results reported previously have quantified in about 13% the FDR corresponding to a number of discoveries of around 400 (Tables 4 and 5 in de Uña-Álvarez, 2011).

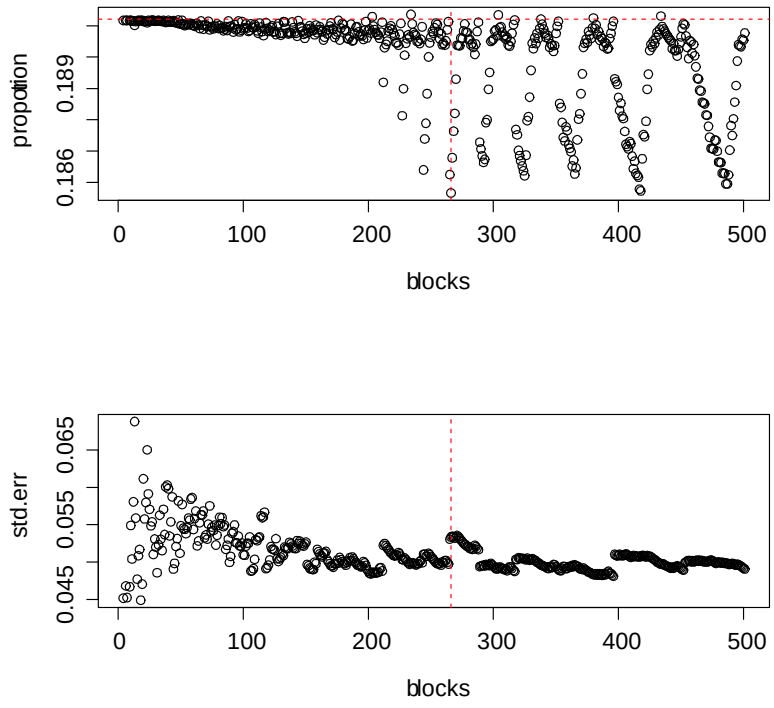

Figure 3. Estimated proportion of significative genes (top) and standard error of $\widehat{\beta}_1$ (bottom) along the number of blocks k. Hedenfalk data, $\gamma = 0.05$. Case $k^N = 266$ and $p_n = F_n(\gamma)$ are highlighted.

Figure 1, top, in Section 1, displays the beta density fitted to Hedenfalk data when taking $k = 266$. The a and b parameters for the beta model were estimated by maximum likelihood from the clustered data, by maximizing the beta-binomial likelihood given in Section 2.2. We obtained $\widehat{a} = 6.66$ and $\widehat{b} = 29.21$. These values correspond to a beta-binomial with parameters $\widehat{p} = .1857$ and $\widehat{\rho} = .0271$. Therefore, a point estimate for the excess of significant cases $\tau_n(\gamma) = n(p-\gamma)$ is given by 430.05 (the lower limit of a 95% confidence interval is 389.15 as indicated above).

3.2 Diz data

As a second example, we consider a list of 261 p-values coming from protein expression experiments in eggs of the marine mussel *Mytilus edulis* (Diz et al., 2009). In that study, *M. edulis* female protein expression profiles of two lines differing in sex ratio of their progeny were compared. In this case, the number of p-values falling below threshold $\gamma = 0.05$ was $nF_n(\gamma) = 26$, with an estimated proportion of significant tests of $F_n(\gamma) = 0.0996$. The application of the original SGoF and its conservative version (both for independent tests) gave $N_\alpha^{(0)}(\gamma) = 8.16$ and $N_\alpha^{(1)}(\gamma) = 5.99$ respectively, where $\alpha = 0.05$ was used. The runs test for randomness did not reject the hypothesis of independence among the tests (p=0.6812). In any case, for illustrative purposes, we performed the BB-SGoF method to this data set.

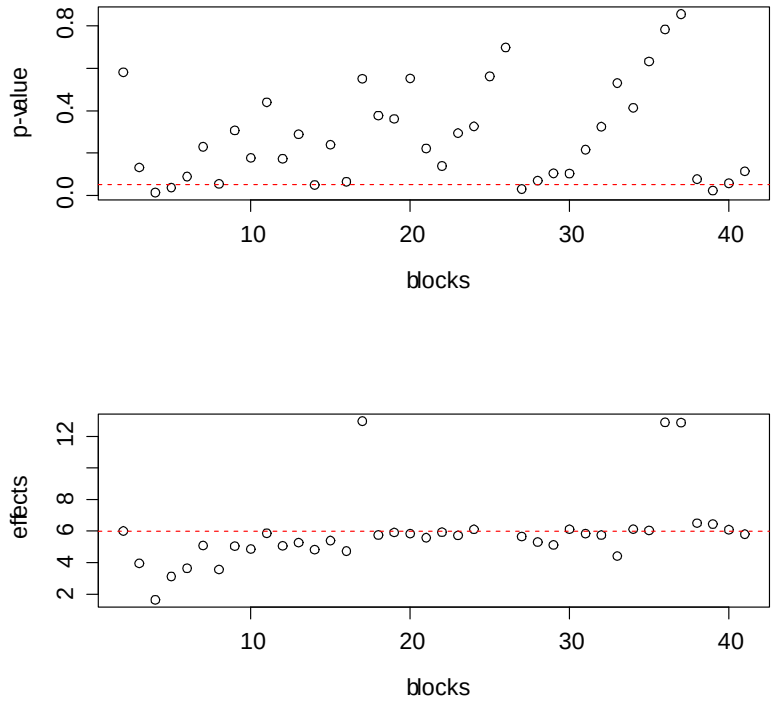

Figure 4. P-values of Tarone (1979)'s test (top) and number of effects declared by BB-SGoF (bottom) along the number of blocks k. Diz data, $\alpha = \gamma = 0.05$. The horizontal dashed line at the bottom figure corresponds to $N_\alpha^{(1)}(\gamma)$.

In Figure 4 we report the p-values of Tarone's test and the number of effects declared by BB-SGoF along the grid $k = 2, ..., 41$ (so again we have a minimum number of tests per block of about 6). As for Hedenfalk data, the unreliable situations with standard error of $\widehat{\beta}_1$ greater than 25 times the median standard error were deleted (this excluded cases $k = 25, 26$ which are associated to negative values of $N_\alpha^{BB}(\gamma; k)$). In this case, most of the p-values of Tarone's test are not significative (they are above 0.05), according to the result of the runs test for randomness. The most conservative version of BB-SGoF corresponds to $k^N = 4$ and $N_\alpha^{BB}(\gamma; k^N) = 1.65$ declared effects. Interestingly, FDR-controlling strategies are unable to detect a single effect even when rising the FDR to 20% (Carvajal-Rodríguez and de Uña-Álvarez, 2011a). For this choice of k, the binomial model is accepted at level 0.01 against the beta-binomial alternative (p-value=0.0127). For the other choices of k the p-value of Tarone's test was even greater. Three values of k reported a very large value of $N_\alpha^{BB}(\gamma; k)$; this was because the standard error of $\widehat{\beta}_1$ was very small (Figure 5). The estimated

values of ρ for these three cases were almost null (smaller than 1.16e-23).

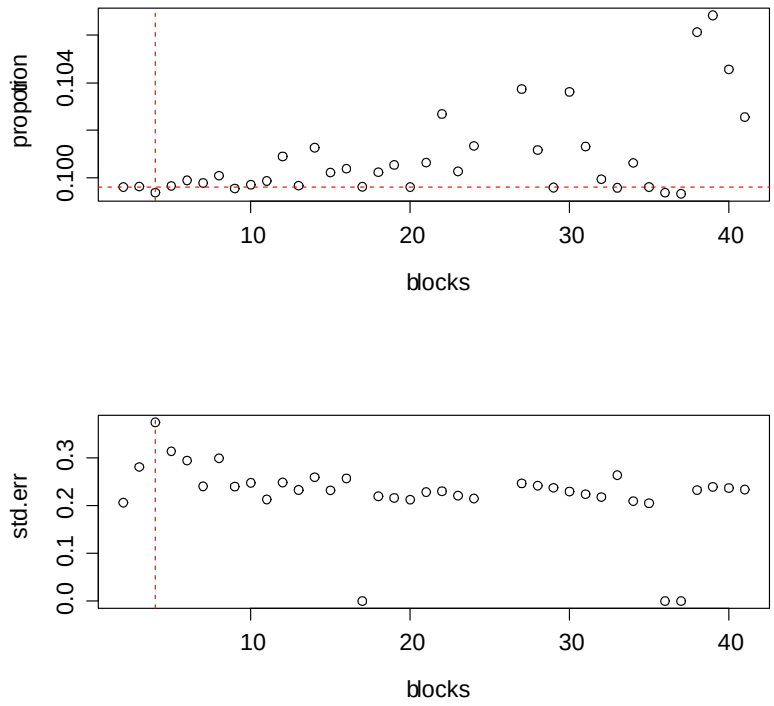

Figure 5. Estimated proportion of significative genes (top) and standard error of $\widehat{\beta}_1$ (bottom) along the number of blocks k. Diz data, $\gamma = 0.05$. Case $k^N = 4$ and $p_n = F_n(\gamma)$ are highlighted.

Figure 1, bottom, in the Introduction displays the beta density fitted to Diz data when taking $k = 4$. The a and b parameters for the beta model were estimated by maximum likelihood from the clustered data, by maximizing the beta-binomial likelihood given in Section 2.2. We obtained $\widehat{a} = 2.72$ and $\widehat{b} = 24.65$. These values correspond to a beta-binomial with parameters $\widehat{p} = .0994$ and $\widehat{\rho} = .0353$. Therefore, a point estimate for the excess of significant cases $\tau_n(\gamma) = n(p - \gamma)$ is given by 12.89 (the lower limit of a 95% confidence interval is 1.65 as indicated above).

4 Simulation study

In order to further explore the performance of BB-SGoF method, we have carried out the following simulation study. The number of tests was $n = 500$ or $n = 1000$. The number of independent blocks of tests k_0 was chosen to get $n/k_0 =$

$25, 50$ tests per block (i.e. $k_0 = 20, 40$ for $n = 500, 1000$ respectively in the case $n/k_0 = 25$, and $k_0 = 10, 20$ for $n = 500, 1000$ respectively in the case $n/k_0 = 50$). The beta-binomial model was generated as follows: $\pi_1, ..., \pi_{k_0} \sim Beta(a, b)$ and, given π_j, the total number of successes in block j was generated as $s_j \sim Bin(n_j, \pi_j)$, where $n_j = n/k_0$, $j = 1, ..., k_0$. Given s_j, a random permutation of s_j 1's and $n_j - s_j$ 0's was taken, $j = 1, ..., k_0$. Note that, in this manner, the sequence of indicators $X_1 = I(u_1 \leq \gamma), ..., X_n = I(u_n \leq \gamma)$ rather than the sequence of p-values $u_1, ..., u_n$ is generated. We took $a = (1 - \rho)p/\rho$ and $b = (1 - \rho)(1 - p)/\rho$ where $p = 0.05, 0.1, 0.2$ and $\rho = 0.1, 0.05$. We always took $\gamma = \alpha = 0.05$. Note that $\gamma = 0.05$ implies that $p = 0.05$ corresponds to the complete null of no effects, while $p = 0.1, 0.2$ represent situations in which a smaller ($p = 0.1$) or larger ($p = 0.2$) proportion of effects are present. More specifically, the excess of significant cases $\tau_n(\gamma) = n(p - \gamma)$ takes the following values in our simulations: 0, 25 and 75 for $n = 500$, and 0, 50 and 150 for $n = 1000$.

In Tables 1-4 we report the average number of effects declared by BB-SGoF when based on different decisions for the number of existing blocks. Specifically, we took $k^* = k_0/2, k_0, 2k_0$ (corresponding to an underestimation, correct estimation, or overestimation of the number of blocks, respectively), and $k^N = \arg\min_k N_\alpha^{BB}(\gamma; k)$, which is the automatic choice corresponding to the most conservative decision (the smallest number of effects declared by BB-SGoF). Minimization was performed on the grid $k = 4, ..., 101$. For comparison purposes, we also report the average number of effects declared by original SGoF and its conservative version ($N_\alpha^{(0)}(\gamma)$ and $N_\alpha^{(1)}(\gamma)$ respectively), both corresponding to the independent setting. The averages were computed along 250 Monte Carlo simulations. Standard deviations for the number of rejected nulls are reported too. Averages and standard deviations of the automatic number of blocks k^N are also displayed in Tables 1-4. Furthermore, we give the familywise rejection rate (FWRR), defined as the proportion of trials for which one or more than one effect was declared; note that, in the case $p = 0.05$ (complete null), this is just the FWER or the FDR.

$n = 500$	$n/k_0 = 50$		$\rho = .1$			$\rho = .05$	
$k_0 = 10$		Mean	SD	FWRR	Mean	SD	FWRR
$p = .05$	$N_\alpha^{(0)}$	2.55	6.33	.2480	1.30	3.27	.2000
	$N_\alpha^{(1)}$	2.02	5.35	.2320	0.95	2.58	.1840
	N_{α,k_0}^{BB}	0.38	1.88	.0600	0.20	0.93	.0640
	$N_{\alpha,k_0/2}^{BB}$	0.57	2.56	.0720	0.38	1.54	.1080
	$N_{\alpha,2k_0}^{BB}$	0.73	2.96	.0840	0.43	1.44	.1130
	N_{α,k^N}^{BB}	0.24	1.39	.0400	0.03	0.26	.0200
	k^N	34.88	36.35	-	41.06	35.14	-
$p = .10$	$N_\alpha^{(0)}$	17.19	14.75	.8160	17.98	11.63	.9200
	$N_\alpha^{(1)}$	14.58	13.15	.8000	15.13	10.38	.9080
	N_{α,k_0}^{BB}	6.55	8.69	.5840	9.10	8.28	.7960
	$N_{\alpha,k_0/2}^{BB}$	7.90	10.85	.6000	10.69	9.79	.8200
	$N_{\alpha,2k_0}^{BB}$	9.31	10.37	.6920	11.82	9.24	.8760
	N_{α,k^N}^{BB}	4.33	7.07	.4320	4.10	6.50	.4080
	k^N	27.62	32.26	-	39.22	34.23	-
$p = .20$	$N_\alpha^{(0)}$	68.09	23.02	1.000	67.22	16.13	1.000
	$N_\alpha^{(1)}$	61.54	21.76	1.000	60.65	15.56	1.000
	N_{α,k_0}^{BB}	45.87	18.86	.9960	50.62	15.27	.9800
	$N_{\alpha,k_0/2}^{BB}$	47.54	22.47	.9800	51.91	18.66	.9640
	$N_{\alpha,2k_0}^{BB}$	51.76	19.78	1.000	53.33	17.40	.9680
	N_{α,k^N}^{BB}	37.55	23.15	.8600	29.98	25.40	.6520
	k^N	24.33	30.62	-	34.38	33.98	-

Table 1. Mean and SD of the number of rejected nulls and familywise rejection rate (FWRR) for SGoF and BB-SGoF multitesting methods along 250 Monte-Carlo trials, $n = 500$. Mean and SD for the automatic number of blocks k^N is also reported. The average proportion of p-values below $\gamma = 0.05$ is p, the number of blocks is $k_0 = 10$, and the within-block correlation is ρ. FWER-control parameter is $\alpha = 0.05$.

$n = 500$	$n/k_0 = 25$		$\rho = .1$			$\rho = .05$	
$k_0 = 20$		Mean	SD	FWRR	Mean	SD	FWRR
$p = .05$	$N_\alpha^{(0)}$	1.95	4.55	.2600	0.86	2.73	.1440
	$N_\alpha^{(1)}$	1.47	3.73	.2320	0.61	2.16	.1200
	N_{α,k_0}^{BB}	0.28	1.42	.0800	0.22	1.08	.0600
	$N_{\alpha,k_0/2}^{BB}$	0.42	1.89	.0920	0.21	0.94	.0680
	$N_{\alpha,2k_0}^{BB}$	0.83	2.69	.1600	0.55	1.94	.1200
	N_{α,k^N}^{BB}	0.16	1.15	.0400	0.02	0.22	.0160
	k^N	33.52	32.80	-	42.18	34.09	-
$p = .10$	$N_\alpha^{(0)}$	16.88	10.60	.9560	16.99	9.75	.9680
	$N_\alpha^{(1)}$	14.10	9.49	.9400	14.17	8.74	.9640
	N_{α,k_0}^{BB}	8.09	7.36	.7960	10.85	7.92	.8960
	$N_{\alpha,k_0/2}^{BB}$	8.72	7.85	.8000	11.15	8.36	.8720
	$N_{\alpha,2k_0}^{BB}$	11.36	8.64	.8960	13.93	23.97	.9040
	N_{α,k^N}^{BB}	5.58	6.54	.6000	4.84	7.27	.4400
	k^N	26.43	28.41	-	34.95	30.91	-
$p = .20$	$N_\alpha^{(0)}$	67.06	17.01	1.000	67.80	12.80	1.000
	$N_\alpha^{(1)}$	60.50	17.01	1.000	61.16	12.09	1.000
	N_{α,k_0}^{BB}	50.52	14.05	1.000	55.22	13.50	.9920
	$N_{\alpha,k_0/2}^{BB}$	50.90	15.38	.9880	54.28	17.36	.9560
	$N_{\alpha,2k_0}^{BB}$	55.83	15.25	1.000	56.58	16.60	.9640
	N_{α,k^N}^{BB}	41.93	19.57	.9000	30.83	25.96	.6200
	k^N	23.06	27.13	-	35.06	33.83	-

Table 2. Mean and SD of the number of rejected nulls and familywise rejection rate (FWRR) for SGoF and BB-SGoF multitesting methods along 250 Monte-Carlo trials, $n = 500$. Mean and SD for the automatic number of blocks k^N is also reported. The average proportion of p-values below $\gamma = 0.05$ is p, the number of blocks is $k_0 = 20$, and the within-block correlation is ρ. FWER-control parameter is $\alpha = 0.05$.

$n = 1000$	$n/k_0 = 50$		$\rho = .1$			$\rho = .05$	
$k_0 = 20$		Mean	SD	FWRR	Mean	SD	FWRR
$p = .05$	$N_\alpha^{(0)}$	3.68	8.78	.2880	1.55	4.17	.1916
	$N_\alpha^{(1)}$	3.07	7.76	.2640	1.21	3.50	.1822
	N_{α,k_0}^{BB}	0.44	3.65	.0440	.15	1.04	.0327
	$N_{\alpha,k_0/2}^{BB}$	0.44	3.24	.0440	.23	1.43	.0467
	$N_{\alpha,2k_0}^{BB}$	0.94	4.75	.1160	.38	1.81	.0794
	N_{α,k^N}^{BB}	0.31	2.78	.0280	.08	.59	.0187
	k^N	17.87	18.19	-	24.07	25.98	-
$p = .10$	$N_\alpha^{(0)}$	39.24	22.04	.9640	41.13	19.22	.9840
	$N_\alpha^{(1)}$	35.16	20.45	.9600	36.86	17.88	.9840
	N_{α,k_0}^{BB}	19.68	15.76	.8680	27.03	15.86	.9600
	$N_{\alpha,k_0/2}^{BB}$	20.24	16.52	.8800	27.74	16.44	.9480
	$N_{\alpha,2k_0}^{BB}$	25.57	17.46	.9400	31.22	16.69	.9720
	N_{α,k^N}^{BB}	16.31	15.24	.7880	21.72	16.30	.8520
	k^N	14.92	14.36	-	20.46	23.33	-
$p = .20$	$N_\alpha^{(0)}$	136.64	30.81	1.000	140.10	23.31	1.000
	$N_\alpha^{(1)}$	127.38	29.60	1.000	130.67	22.40	1.000
	N_{α,k_0}^{BB}	103.48	26.41	1.000	115.61	21.18	1.000
	$N_{\alpha,k_0/2}^{BB}$	104.40	27.73	1.000	115.77	23.93	.9920
	$N_{\alpha,2k_0}^{BB}$	112.84	27.39	1.000	121.51	22.95	.9960
	N_{α,k^N}^{BB}	97.17	27.19	.9960	103.67	32.97	.9440
	k^N	13.20	11.02	-	16.54	18.23	-

Table 3. Mean and SD of the number of rejected nulls and familywise rejection rate (FWRR) for SGoF and BB-SGoF multitesting methods along 250 Monte-Carlo trials, $n = 1000$. Mean and SD for the automatic number of blocks k^N is also reported. The average proportion of p-values below $\gamma = 0.05$ is p, the number of blocks is $k_0 = 20$, and the within-block correlation is ρ. FWER-control parameter is $\alpha = 0.05$.

$n = 1000$	$n/k_0 = 25$		$\rho = .1$			$\rho = .05$	
$k_0 = 40$		Mean	SD	FWRR	Mean	SD	FWRR
$p = .05$	$N_\alpha^{(0)}$	1.73	4.20	.2200	0.93	3.32	.1189
	$N_\alpha^{(1)}$	1.35	3.49	.2120	0.73	2.76	.1049
	N_{α,k_0}^{BB}	0.17	1.03	.0360	0.27	1.43	.0420
	$N_{\alpha,k_0/2}^{BB}$	0.20	1.27	.0400	0.28	1.50	.0420
	$N_{\alpha,2k_0}^{BB}$	0.55	2.14	.1080	0.57	2.44	.0839
	N_{α,k^N}^{BB}	0.10	0.81	.0240	0.13	0.93	.0280
	k^N	24.33	22.93	-	28.98	26.68	-
$p = .10$	$N_\alpha^{(0)}$	38.54	17.60	.9920	39.41	13.58	1.000
	$N_\alpha^{(1)}$	34.44	16.35	.9920	35.20	12.64	1.000
	N_{α,k_0}^{BB}	24.23	14.08	.9600	29.38	11.61	.9960
	$N_{\alpha,k_0/2}^{BB}$	24.37	14.65	.9560	29.56	12.18	.9840
	$N_{\alpha,2k_0}^{BB}$	29.26	15.69	.9720	32.39	12.45	.9840
	N_{α,k^N}^{BB}	18.95	13.99	.8680	21.06	14.11	.8040
	k^N	22.73	19.99	-	26.62	24.68	-
$p = .20$	$N_\alpha^{(0)}$	143.35	23.23	1.000	140.42	20.68	1.000
	$N_\alpha^{(1)}$	133.79	22.34	1.000	130.97	19.86	1.000
	N_{α,k_0}^{BB}	118.25	20.61	1.000	121.94	19.31	1.000
	$N_{\alpha,k_0/2}^{BB}$	119.18	20.88	1.000	122.27	19.65	1.000
	$N_{\alpha,2k_0}^{BB}$	126.68	22.26	1.000	125.67	21.51	.9960
	N_{α,k^N}^{BB}	109.62	25.70	.9800	104.68	39.52	.9000
	k^N	20.93	17.69	-	26.78	19.47	-

Table 4. Mean and SD of the number of rejected nulls and familywise rejection rate (FWRR) for SGoF and BB-SGoF multitesting methods along 250 Monte-Carlo trials, $n = 1000$. Mean and SD for the automatic number of blocks k^N is also reported. The average proportion of p-values below $\gamma = 0.05$ is p, the number of blocks is $k_0 = 40$, and the within-block correlation is ρ. FWER-control parameter is $\alpha = 0.05$.

From Tables 1-4 the following features are appreciated.

BB-SGoF based on the true number of blocks k_0 controls the FWER at the nominal level $\alpha = 0.05$ fairly well. However, when the chosen number of blocks k^* is different from k_0, the control of FWER is lost. This is more clear for $k^* = 2k_0$, when the assumed dependence structure is weaker than the true one; in this case, the FWER may be up to three times the nominal. In practice, the value of k_0 will be unknown, so a realistic version of BB-SGoF is that based on the automatic choice k^N. As expected, the FWER associated to this automatic

criterion $N_\alpha^{BB}(\gamma; k^N)$ is smaller than 0.05; in all the performed experiments this FWER ranges between 0.016 and 0.04, indicating the conservativeness of this approach. On the other hand, the results corresponding to original SGoF method for independent tests were clearly anticonservative. This issue is more evident for a larger number of tests, a larger within-block correlation, and a larger number of tests per block. For original SGoF ($N_\alpha^{(0)}(\gamma)$), FWER is up to 7 times that of BB-SGoF based on k_0; the conservative version of original SGoF ($N_\alpha^{(1)}(\gamma)$) offers similar results to this regard, being unable to cope with the existing dependences among the tests (as expected).

The number of effects declared by BB-SGoF method based on k_0 clearly increases with the number of tests, while it only varies slightly for different values of ρ and k_0 with fixed n. This property is shared by the automatic BB-SGoF based on k^N and also for BB-SGoF based on a wrong estimation of the number of blocks. This ability to increase the number of rejections with the number of tests is also evident for original SGoF, as previoulsy quoted in the independent setting (Carvajal-Rodríguez et al., 2009; de Uña-Álvarez, 2011). An interesting question here is the amount of power which is lost by the conservative, automatic BB-SGoF $N_\alpha^{BB}(\gamma; k^N)$ when compared with the benchmark $N_\alpha^{BB}(\gamma; k_0)$. The relative number of declared effects ranges between 0.4505 ($n = 500$, $p = 0.1$, $\rho = 0.05$, $k_0 = 10$) and 0.9390 ($n = 1000$, $p = 0.2$, $\rho = 0.1$, $k_0 = 40$). In general, it is seen that this relative power is smaller for a smaller number of tests, a smaller proportion of existing effects, and a smaller within-block correlation, while the influence of k_0 is of a smaller magnitude. Interestingly, the relative power of automatic BB-SGoF is never below 72% when $n = 1000$ regardless the other parameters in the simulation study (Tables 3 and 4).

Finally, we see from Tables 1-4 that the automatic selector k^N has a large variance, and that its average value seems to be independent of k_0 for $n = 500$, although not for $n = 1000$, where it grows with k_0. However, we should recall that the goal of k^N is not to provide an estimator for k_0; rather, it suggests a conservative decision to prevent any overestimation of the number of effects with respect to the benchmark $N_\alpha^{BB}(\gamma; k_0)$. More specifically, we see from Tables 1-4 that k^N tends to be larger than k_0 for $n = 500$, while the opposite is true for $n = 1000$.

5 Discussion

In this paper a correction of SGoF mutitesting procedure for dependent tests has been introduced. The correction is relevant, since the original SGoF does not respect the nominal FWER under dependence, because of the underestimated variance of the number of discoveries it is based on. Since SGoF for independent tests uses the binomial model, the introduced correction uses an extension of the binomial distribution (the beta-binomial) which allows for possible correlations

among the Bernoulli outcomes. The beta-binomial model represents the dependence structure among the indicators of reaching statistical significance at level γ for each individual test; therefore, it does not imply any rigid assumption on the joint distribution of the p-values themselves. In that sense, even when alternative extensions of the binomial model for dependent trials exist (e.g. Tarone, 1979), the simple beta-binomial approach provides a reasonable adaptation of SGoF multitesting method under dependence which works well in practice.

We point out that the unique input needed by BB-SGoF is the set of p-values; the crude data leading to these p-values are not used in the computations. Obviously, the p-values should not be sorted. Indeed, in order to detect possible dependences, no particular data-driven ordering of the p-values should be applied. Any initial guess of the researcher on the existing serial dependence should not be obtained from the p-values themselves. For Hedenfalk data, we followed the original order considered by Hedenfalk et al. (2001); we do not give any special interpretation to that. The proposed BB-SGoF method suggests that a number of blocks of dependent outcomes *in the given order* probably exist. Similarly, for Diz data we followed the ordering in which the authors provided the sequence of p-values. No evidence of correlation was found in this latter case.

The introduced BB-SGoF method conserves the main properties of original SGoF. In particular, the number of effects declared by BB-SGoF increases with the number of tests, a property which typically fails for other approaches (as those controlling the FDR at a given level). The reason behind this property is in the fact that BB-SGoF focus on the amount of p-values below threshold γ which are connected with nontrue null hypotheses; the metatest for this question is performed at a level α which controls the FWER in the weak sense, while a strong control of FWER or FDR is not imposed. Therefore, BB-SGoF approach is recommended in situations with dependent tests in which finding effects is difficult because of the large number of tests involved or the weakness of the effects. The practical advantages of BB-SGoF approach have been illustrated in two real data applications.

A critical point in the performance of BB-SGoF is the preliminary decision on the number of existing blocks of tests k. Roughly speaking, a small value of k establishes strong dependence and therefore the power of the method may be decreased. On the contrary, a large k may result in a too liberal decision. To overcome this issue, we have proposed the automatic decision k^N attached to the minimum number of discoveries provided by BB-SGoF along a grid $k = k_{\min}, ..., k_{\max}$. This ensures the weak control of FWER at the given nominal level α as long as the true number of blocks falls between $k_{\min}$ and $k_{\max}$. Interestingly, simulations have revealed that this very conservative strategy has a reasonable relative power, mainly when the number of tests is large.

An interesting question is the possibility of following a Bayesian approach in this context. Note that the excess of significant cases which BB-SGoF con-

centrates in, $\tau_n(\gamma) = n(p - \gamma)$, is just the mean value of the random variable $n(\pi - \gamma)$. Bayesian confidence sets for this variable based on the posterior distribution of π could be constructed to evaluate the number of existing effects. The benefits of this approach relative to those presented in this work are left for future studies.

Acknowledgements. The author thanks AP Diz for providing the set of p-values, and Antonio Carvajal-Rodríguez for fruitful discussions. Support from the projects MTM2011-23204 of the Spanish Ministry of Science and Innovation (FEDER support included) and 10PXIB300068PR of the Xunta de Galicia is acknowledged.

6 References

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-Rodríguez, de Uña-Álvarez J (2011a) Assessing Significance in High-Throughput Experiments by Sequential Goodness-of-Fit and q-Value Estimation. PLoS ONE 6(9), e24700.

Carvajal-Rodríguez A, de Una-Álvarez J (2011b) 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

Carvajal-Rodríguez A, de Una-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.

Chatfield, C, Goodhardt, GJ (1970). The beta-binomial model for consumer purchasing behaviour. Appl. Statist. 19, 240-50.

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.

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

de Uña-Álvarez and Carvajal-Rodríguez A (2010) SGoFicance Trace: Assessing Significance in High Dimensional Testing Problems. PLoS ONE 5(12), e15930.

Diz AP, Dudley E, MacDonald BW, Pina B, Kenchington EL, et al. (2009) Genetic variation underlying protein expression in eggs of the marine mussel Mytilus edulis. Mol Cell Proteomics 8: 132-144.

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

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

Efron B (2007) Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association 102, 93-103.

Efron B (2010) Correlated z-values and the accuracy of large-scale statistical estimates (with Discussion). Journal of the American Statistical Association 105, 1042-1069.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics 29, 637-48.

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.

Johnson NL, Kotz (1970). Distributions in statistics, continuous univariate distributions-2. Houghton Mifflin, Boston.

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.

Romano JP, Shaikh AM, Wolf M (2008) Control of the false discovery rate under dependence using the bootstrap and subsampling (with Discussion). Test, 417-471.

Siegel S, Castellan NJ (1988) Nonparametric Statistics for the Behavioural Sciences, 2nd edn, McGraw-Hill, New York.

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.

Tarone RE (1979) Testing the goodness-of-fit of the binomial distribution. Biometrika 66, 585-590.

Thang V. Pham., Sander R. Piersma, Marc Warmoes and Connie R. Jimenez (2010) On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics, Bioinformatics 26, 363-369.

Williams DA (1975) The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity. Biometrics 31, 949-952