Master's Thesis

PERFORMANCE OF BETA-BINOMIAL SGoF MULTITESTING METHOD UNDER DEPENDENCE: A SIMULATION STUDY

AUTHOR: Irene Castro Conde

DIRECTOR: Jacobo de Uña Álvarez

Master in Statistical Techniques

University of Vigo

January 7, 2013

Jacobo de Uña Álvarez, Catedrático de Universidad del Departamento de Estadística e Investigación Operativa de la Universidad de Vigo.

HACE CONSTAR

Que el presente trabajo titulado *Performance of Beta-Binomial SGoF multitesting method under dependence: a simulation study* ha sido realizado por Irene Castro Conde bajo su dirección para su presentación como Trabajo Fin de Máster del *Máster en Técnicas Estadísticas.*

Vigo, 07 de enero de 2013.

Fdo.: Jacobo de Uña Álvarez.

Abstract

In a recent paper (de Uña-Álvarez, 2012) a correction of SGoF multitesting method for possibly dependent tests was introduced. This correction enhanced the field of applications of SGoF methodology, initially restricted to the independent setting, to make decisions on which hypotheses are to be rejected in a multiple hypothesis testing problem involving dependence.

In this work we make a contribution to that topic through an intensive Monte Carlo simulation study of that correction, called BB-SGoF (from Beta-Binomial). In these simulations several number of blocks, within-block correlation values, effect levels, and proportion of true effects are considered. The allocation of the true effects is taken to be random. False discovery rate, power, and conservativeness of the method (with respect to the number of existing effects with p-values below the given significance threshold) are computed along the Monte Carlo trials. Comparison to the original SGoF and Benjamini-Hochberg adjustments is provided. In de Uña-Álvarez (2012) FDR and power weren't reported so this implies a new contribution to the study of BB-SGoF procedure. Another contribution of this work is the development of an `R` code for the implementation of BB-SGoF method.

Part of this work is included in the forthcoming publication Castro Conde and de Uña-Álvarez J (2013).

Contents

Chapter 1

Introduction

1.1 Multiple Hypothesis Testing

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. Nowadays, there exist many statistical inference problems in areas such genomic and protenomics which involve the simultaneous test of thousands, or tens of thousands, of null hypotheses producing as a result a number of significant p-values or effects. Moreover, these hypotheses may have complex and unknown dependence structure among themselves. See e.g. Dudoit and Van der Laan (2008) for an introduction to this area.

One of the main problems in multiple hypotheses testing is that, if one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected may be overly large. So, in the multitesting setting, a specific procedure for deciding which null hypotheses should be rejected is needed.

The decision to reject or not the null hypotheses is usually based on tests statistics, defined as functions of the data which provide rejection regions for each of the n hypotheses. Some multitesting methods (SGoF, BB-SGoF, Bonferroni, Benjamini-Hochberg...) do not make a direct use of test statistics; rather they use the p-values to decide which hypotheses are to be rejected. Let us define formally the concept of unadjusted p-value in this multitest setting (Dudoit and Van der Laan, 2008).

Definition 1.1.1 (Unadjusted p-value) *The unadjusted p-value p_i, for the single test of null hypothesis H_{0i}, is defined as*

$$p_i \equiv inf\{\alpha \in [0,1] : \text{ Reject } H_{0i} \text{ at single test nominal level } \alpha\}, i = 1, ..., n.$$

That is, the unadjusted p-value p_i, for null hypothesis H_{0i}, is the *smallest nominal Type I error level* of the *single hypothesis testing procedure* at which one would reject H_{0i}. The smaller the unadjusted p-value p_i, the stronger evidence against the corresponding null hypothesis H_{0i}.

Specifically, null hypothesis H_{0i} is rejected at single test nominal Type I error level α if $p_i \leq \alpha$. That is, the set of rejected null hypotheses at single test nominal Type I error level α is

$$R_n(\alpha) = \{i : p_i \leq \alpha\}.$$

In any testing problem, two types of error can be committed. A Type I error, or false positive, is committed by rejecting a true null hypothesis. A Type II error, or false negative, is committed by failing to reject a false null hypothesis.

Consider the problem of testing simultaneously n null hypotheses, of which h_0 are true and R_n is the number of hypotheses rejected. Table 1.1 summarizes the number of errors committed:

	Non rejected, R_n^c	Rejected, R_n	total
H_0 True	W_n	V_n	h_0
H_1 True	U_n	S_n	h_1
total	n-R_n	R_n	n

Table 1.1: Classification of errors committed in multiple testing.

According to this classification, V_n is the number of null hypotheses which are rejected (Type I errors) while U_n are the false nulls which are not rejected (Type II errors). Ideally, one would like to simultaneously minimize the probability of both errors.

But this is not feasible and one seeks for a trade-off between the two types of error. This trade-off typically involves the minimization of the probability of Type II error, i.e., maximization of power, subject to a Type I error constraint.

Traditionally great importance is given to the control of Type I error. The problem that arises when multiple tests are performed is that this Type I error control gets lost. Hence, there is a need for multitest correction methods aiming to control for this error. However when the number of tests is very large, control of Type I error usually entails an important increase of the Type II error i.e. a great loss of statistical power. As a result, in recent years many efforts have been made to improve methods of multitest correction looking for a balance between control of Type I error and power.

1.2 Error criteria. Type I error rates

When testing multiple hypotheses, there are many possible definitions for the Type I error rate of a testing procedure. Accordingly, we define a *Type I error rate* as a function of the number of Type I errors V_n and rejected hypotheses R_n.

1.2.1 Type I error rates based on the distribution of the number of Type I errors

Definition 1.2.1 (FWER) *The **family-wise error rate** is the probability of having at least one Type I error,*

$$FWER \equiv P(V_n > 0).$$

The family-wise error rate (FWER), defined as the probability of committing at least one Type I error through the several hypotheses under consideration, works as a substitute for the significance level in the traditional (single hypothesis) context. One can differentiate between two types of control for FWER:

- A procedure controls the FWER in the **weak sense** if the FWER control at level α

is guaranteed only when all null hypotheses are true (complete, intersection or global null hypothesis).

- A procedure controls the FWER in the **strong sense** if the FWER control at level α is guaranteed for any configuration of true and non-true null hypotheses (including the global null hypothesis).

Typically FWER control is required in the strong sense, i.e. independently of the amount and location of true and false hypotheses. Unfortunately, methods controlling the FWER have a remarkable lack of power, that is, they are unable to detect a reasonable amount of effects (Benjamini and Hochberg, 1995). In fact, the power to detect a specific hypothesis while controlling the FWER is greatly reduced when the number of hypotheses in the family increases. A relaxation of FWER is the generalized FWER criterion.

Definition 1.2.2 (gFWER) *The* ***generalized family-wise error rate*** *for a user-supplied integer $k \in \{0, ..., n-1\}$, is the probability of having at least $k+1$ Type I errors. That is,*

$$gFWER(k) \equiv P(V_n > k).$$

When k=0, the gFWER reduces to usual family-wise error rate, FWER.

Most multiple testing procedures focus on control of the FWER, e.g., Bonferroni procedure (See section 1.5.1).

1.2.2 Type I error rates based on the distribution of the proportion of Type I errors among the rejected hypotheses

Definition 1.2.3 (FDR) *The* ***false discovery rate*** *is the expected value of the proportion of Type I errors among the rejected hypotheses,*

$$FDR \equiv E\left[\frac{V_n I_{\{R_n>0\}}}{R_n}\right] = E\left[\frac{V_n}{R_n}|R_n > 0\right] P_r(R_n > 0)$$

where I_A is the indicator function of an event A.

Under the complete null hypothesis, all R_n rejected hypothesis are Type I errors, hence $V_n/R_n = 1$ whenever $R_n > 1$ and $FDR = FWER = P(V_n > 0)$. FDR controlling procedures therefore also control the FWER in the weak sense. In general, because $V_n/R_n \leq 1$ ($I_{V_n \geq 1} \geq V_n/R_n \Rightarrow E(I_{V_n \geq 1}) \geq E(V_n/R_n) \Rightarrow P(V_n > 0) \geq E(V_n/R_n)$), we have $FDR \leq FWER$ for any given multiple testing procedure. Thus, procedures controlling the FWER are typically more conservative, i.e., they lead to fewer rejected hypotheses, than those controlling the FDR.

The family-wise error rate (FWER) and the false discovery rate (FDR) have been proposed as suitable significance criteria for multiple testing. See Benjamin and Hochberg (1995), Nichols and Hayasaka (2003) or Dudoit and Van der Laan (2008).

1.2.3 Other Type I error rate criteria

As we have said, there are many criteria for the control of Type I error rates. Table 1.2 shows a classification of different measures of Type I error in multiple testing problems. In this table, F_{ξ}^{-1} stands for the quantile function of a random variable ξ.

1.3 Power

Definition 1.3.1 (Power) *The probability that no Type II error occurs is called test power. That is, the power is a measure of the skill of the test to detect an effect which is present.*

In the setting of multiple testing one needs to adjust the usual definition of power. Then, the **power** is defined as the *expected value of the proportion of rejected hypotheses among the true effects*. For a rejection region of type $R_n(\alpha) = \{i : p_i \leq \alpha\}$ we have:

$$Power(\alpha) = E\left[\frac{1}{h_1}\sum_{i=1}^{n} I_{\{p_i \leq \alpha, H_{0i}=1\}}\right]$$

where $h_1 = \sum_{i=1}^{n} I_{\{H_{0i}=1\}}$ is the unknown number of non-true nulls among the n hypotheses and $H_{0i} = 1$ indicates that the null hypothesis H_{0i} is false.

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 el 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. Otherwise, Benjamin and Hochberg (1995) demonstrated that the direct control of FDR increase considerably the statistical power of multitest adjustment.

Type I error rate	Definition
Family-wise error rate	$FWER = P(V_n > 0)$
Generalized family-wise error rate	$gFWER(k) = P(V_n > k)$
Per-comparison error rate	$PCER = E[V_n]/n$
Per-family error rate	$PFER = E[V_n]$
Median-based per-family error rate	$mPFER = F_{V_n}^{-1}(1/2)$
Quantile number of false positives	$QNFP(\delta) = F_{V_n}^{-1}(\delta)$
Tail probability for the proportion of false positives	$TPPFP(q) = P\left(\frac{V_n}{R_n} > q\right)$
False discovery rate	$FDR = E[V_n I_{\{R_n>0\}}/R_n]$
Proportion of expected false positives	$PEFP = E[V_n]/E[R_n]$
Quantile proportion of false positives	$QPFP(\delta) = F_{V_n/R_n}^{-1}(\delta)$
Generalized tail probability error rate	$gTP(q, g) = P(g(V_n, R_n) > q)$
Generalized expected error rate	$gEV(g) = E[g(V_n, R_n)]$

Table 1.2: Commonly-used Type I error rates. Taken from Dudoit S. and Van der Laan M.J. (2008).

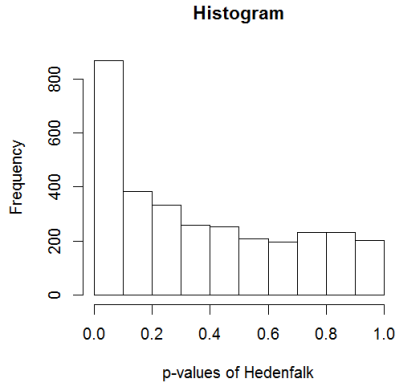
1.4 An example

As an illustrative example, we consider the micro array study of hereditary breast cancer of Hedenfalk et al. (2001). Many cases of hereditary breast cancer are due to mutations in either the BRCA1 or the BRCA2 gene. The histopathological changes in these cancers are often characteristic of the mutant gene. They hypothesized that the genes expressed by these two types of tumors are also distinctive, perhaps allowing to identify cases of hereditary breast cancer on the basis of gene-expression profiles.

The patients consisted of 23 with BRCA1 mutations, 17 with BRCA2 mutations, 20 with familial breast cancer, 19 with possibly familial breast cancer and 34 with sporadic breast cancer to determine whether there are distinctive patterns of global gene expression in these three kinds of tumors.

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, 56 genes were eliminated. This left n =3,170 genes.

Figure 1.1 shows the histogram of these 3170 p-values:



It can be seen from Figure 1.1 that the p-values present an asymmetry and they are concentrated around the zero, which possibly indicates the existence of effects (non-true null hypotheses). The set of p-values is available in the library *qvalue* of the software *R* (R Development Core Team, 2008).

1.5 Methods

In this section we describe some of the existing multitesting methods, which are based only on the p-values and aim to control Type I error and power in a simultaneous way.

1.5.1 Bonferroni

The Bonferroni correction is a method used to counteract the problem of multiple comparisons. It is considered the simplest and most conservative method to control the familywise error rate.

The correction is based on the idea that, if an experimenter is testing n dependent or independent hypotheses on a set of data, then one way of maintaining the familywise error rate is to test each individual hypothesis at a statistical significance level of $1/n$ times what it would be if only one hypothesis were tested. So, if it is desired that the significance level for the whole family of tests should be (at most) α, then the Bonferroni correction would be to test each of the individual tests at a significance level of α/n.

The Bonferroni correction states that rejecting all p_i smaller than $\frac{\alpha}{n}$ will control the FWER at level α. The proof follows from Boole's inequality:

$$FWER = P\left\{\bigcup_{i_0}(p_i \leq \frac{\alpha}{n})\right\} \leq \sum_{i_0} P(p_i \leq \frac{\alpha}{n}) = h_0\frac{\alpha}{n} \leq n\frac{\alpha}{n} = \alpha$$

where i_0 represents the index going along all the true null hypotheses and h_0 is the number of true null hypotheses.

This result does not require the tests to be independent.

1.5.2 BH

The Benjamini–Hochberg procedure (BH step-up procedure) controls the false discovery rate (at level α). The procedure works as follows:

i) For a given α, let k be the largest i for which $p_{(i)} \leq \frac{i}{n}\alpha$, where $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(n)}$

are the ordered p-values.

ii) Then reject (i.e. declare positive discoveries) all $H_{0(i)}$ for $i = 1, 2, \ldots, k$, where $H_{0(i)}$ is the null hypotheses attached to $p_{(i)}$.

The BH procedure controls the FDR al level α when the n tests are independent, and also in various scenarios of dependence (see Benjamini and Yekutieli, 2001). It also satisfies the inequality:

$$FDR \leq \frac{h_0}{n}\alpha \leq \alpha.$$

So the above procedure controls the FDR at level $\frac{h_0}{n}\alpha$.

It becomes clear from last inequalities that knowledge on h_0 is relevant for improving the BH procedure. However, if an estimator of h_0 is inserted into the BH procedure in a obvious way, the FDR control at the desired level is no longer guaranteed.

The BH procedure was proven to control the FDR by Benjamini and Hochberg (1995). Simes (1986) introduced the same procedure in order to control the FWER in the weak sense i.e. under the intersection or complete null hypothesis. Hommel (1988) showed that Simes procedure does not control the FWER in the strong sense. Based on the Simes procedure, Hochberg (1988) proposed a step-up procedure which does control the FWER in the strong sense.

In Storey (2003), a modified version of the FDR called the 'positive false discovery rate' (pFDR) was introduced. Moreover, a new quantity called the 'q-value' was introduced and investigated. It was motivated as a natural 'Bayesian posterior p-value', or rather the pFDR analogue of the p-value.

The q-value is defined to be the FDR analogue of the p-value. The q-value of an individual hypothesis test is the minimum FDR at which the test may be called significant. One approach is to directly estimate q-values rather than fixing a level at which to control the FDR.

1.5.3 SGoF

Recently, Carvajal-Rodríguez et al. (2009) proposed a new method for p-value thresholding in multitesting problems. This method, called SGoF (from Sequential- Goodness-of-

Fit), can be summarized as follows. Let F_n be the empirical distribution of the p-values, and let γ be an initial significance level, typically $\gamma = 0.05$. 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 z_α is the $1-\alpha$ quantile of the standard normal. Here, $\alpha = \gamma$ is usually taken. If H_0 is rejected, 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 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 is controlled at level α in the weak sense (Carvajal-Rodríguez et al., 2009), but not in the strong sense. Besides, SGoF does not control FDR at any level, being liberal to this regard. We denote the corresponding threshold p-value by $p^*_{n,\alpha}(\gamma)$, that is, $N_\alpha(\gamma) = nF_n(p^*_{n,\alpha}(\gamma))$. We have that $p^*_{n,\alpha}(\gamma) \leq \gamma$.

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 short, SGoF multitesting procedure makes a decision on the number of effects among those p-values smaller than a significance level γ. This method is an useful al-

ternative to FDR-based methods (e.g. Benjamini-Hochberg) when the number of tests is large, the proportion of effects is small, and the effects are weak to moderate. In such situation the power of BH methods is usually poor and will be improved by SGoF proceed. Unfortunately, SGoF 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. See Carvajal-Rodríguez et al. (2009) and de Uña-Álvarez (2011) for more information.

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 betabinomial 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.

1.5.4 BB-SGoF

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).

BB-SGoF (from Beta-Binomial SGoF, de Uña-Álvarez 2012) is a correction of SGoF for correlated tests. It is assumed 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 α, 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: $X_i = I_{\{u_i \leq \gamma\}}, i = 1, ..., n$. Here we change the notation for the p-values since p will be used to represent a population parameter. 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, $p = F(\gamma)$ 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).

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) $n = \sum_{j=1}^{k} n_j$ and ρ_n is a estimator of the correlation ρ, 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).

Model-based estimators for p and ρ may be derived by maximum-likelihood principles under the beta-binomial assumption. As usual with maximum-likelihood estimates, the maximizer $(\hat{p}, \hat{\rho})$ of the likelihood on the $[0, 1] \times [0, 1]$ rectangle is an efficient, asymptotically normal estimator of (p, ρ). 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 $\tau_n(\gamma) = n(p - \gamma) = n(F(\gamma) - \gamma)$, similarly as original SGoF does but considering the possible existing correlation. This confidence interval may be constructed from the asymptotic normality of $\hat{p}$.

Consider the reparametrization of the beta-binomial model given by the logit transformation of p and ρ, that is $\beta_1 = log(p/(1-p))$ and $\beta_2 = log(\rho/(1-\rho))$. With this reparametrization, an unrestricted maximization of the likelihood can be performed.

The following $100(1-\alpha)\%$ confidence intervals for β_1 and β_2 can be computed:

$$I(\beta_i) = (\hat{\beta}_i \pm se(\hat{\beta}i) z_{\alpha/2}), \quad i = 1, 2,$$

where $se(\hat{\beta}i)$ denotes the estimated standard error of $\hat{\beta}_i$, 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 $I(\beta_i)$.

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 $low_1 = \hat{\beta}_1 - se(\hat{\beta}_1)z_\alpha$.

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. It is also clear that the number of declared effects will grow with the number of tests. This is because $se(\hat{\beta}_1)$ goes to zero at a $\sqrt{n}$-rate, and therefore the lower limit $low_1 = \hat{\beta}_1 - se(\hat{\beta}_1)z_\alpha$ is shifted-up towards $\hat{\beta}_1$ 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. $F^{-1}(N_\alpha^{BB}(\gamma; k)/n)$, 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 of this method is how to choose the value of k; once k is fixed, the n_j's may be computed for example 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 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.

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}.

BB-SGoF in practice

In this section we refer a detailed application of BB-SGoF to Hedenfalk data set as provided in de Uña-Álvarez (2012). As mentioned in Section 1.4, this data set contains a sequence of 3170 p-values corresponding to tests performed on gene expression levels concerns to a study of hereditary breast cancer by Hedenfalk et al. (2001).

Assuming independence among the tests and taking $\gamma = 0.05$ as initial threshold, the number of effects declared by SGoF at level $\alpha = 0.05$ was $N_{\alpha}^{(0)}(\gamma) = 428.32$ and $N_{\alpha}^{(1)}(\gamma) = 412.08$ effects when using the conservative version of SGoF which estimates the variance of $F_n(\gamma)$ without any restriction. The independence assumption among the tests was checked through the runs test for randomness of a dichotomous (binary) sequence, giving a two-sided p-value of 0.002654.

Under dependence, inferences provided by SGoF above are not valid and, therefore, the number of significant genes must be re-evaluated.

The minimum value of $N_{\alpha}^{BB}(\gamma;k)$ along k ($k = 2, ..., 501$) 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. This is not surprising, since 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.

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, 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.

1.6 Contributions of this work

In the recent paper, *The Beta-Binomial SGoF method for multiple dependent tests* (de Uña-Álvarez, 2012), a simulation study of BB-SGoF was carried out. The study reported the average number of effects declared by BB-SGoF when based on different decisions for the number of existing blocks and the average number of effects declared by original SGoF and its conservative version, both corresponding to the independent setting. The averages were computed along 250 Monte Carlo simulations. Standard deviations for the number of rejected nulls were reported too.

Furthermore, the familywise rejection rate (FWRR) was given, defined as the proportion of trials for which one or more than one effect was declared; note that, in the case of the complete null, this is just the FWER or the FDR. But in situations with effects no information about the FDR was reported. Power was not reported either in that simulation study. The reason was in the specific planing for the generation of the outcomes. Indeed, only the indicators $I_{\{u_i \leq \gamma\}}$ were generated, without allowing for the identification of true and non-true nulls.

The contribution of this work is a more intensive simulation study (112 simulated scenarios) where 1000 trials of Monte Carlo are performed in each simulation, and a more extensive results report. Specifically, FDR and power are computed for SGoF, conservative SGoF, BB-SGoF based on four different decisions for the number of blocks (true number of blocks k, k/2, 2k and automatic data-driven choice) and BH method. We also report the proportion of trials for which the number of declared effects was not larger

than the number of effects with p-value below γ. All these calculations were possible due to the fact that in our simulations, we have perfectly identified which p-values correspond to true hypotheses and which with false ones. This makes an important difference with respect to the mentioned paper.

Another relevant difference is that, since we inspire the simulations in the data of Hedenfalk (2-sample tests), we simulate a model which falls out of the scope of the beta-binomial family. In this sense, the provided simulations allows to investigate robustness properties of the BB-SGoF approach.

In this work, `R` code for the implementation of BB-SGoF procedure has been developed. In particular, code for the computation and optimitation of the beta-binomial likelihood along a grid of values for the number of existing blocks of dependent tests is given. This is an important contribution since, so far, the only available implementation of BB-SGoF (cfr. *http://webs.uvigo.es/jacobo/BB-SGoF.htm*) was based on the function `vglm` of the library `VGAM`, for which some problems were detected.

Chapter 2

Simulation Scenario and Simulation Results

2.1 Simulated scenario

In order to further explore the performance of BB-SGoF method, we have carried out the following simulation study where 1000 Monte Carlo simulations were performed.

Having in mind the study of Hedenfalk data, we simulated n=500 or n=1000 2-sample t-tests for comparison of normally distributed 'gene expression levels' in two groups A and B with sizes 7 and 8 respectively. 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 $\rho = 0, 0.1, 0.2$ and 0.8 were taken (note that this ρ refers to the correlation between gaussian outcomes and not to the ρ parameter of the discussed beta-binomial model). With regard to the number of blocks, we considered $k = 10$ or $k = 20$, so we had 50 or 25 tests per block when $n = 500$, and 100 or 50 tests per block when $n = 1000$. For random generation, the function `rmvnorm` of the `R` software was used.

BB-SGoF method with $\gamma = \alpha = 0.05$ was applied under perfect knowledge on the true value of k but also when underestimating $(k/2)$ or overestimating $(2k)$ the true number of blocks. We also applied an automatic (data-driven) choice of k by minimizing the number of effects declared by BB-SGoF along the grid $k = 2, ..., 61$.

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), BB-SGoF guarantees that this proportion (labeled as *Coverage* 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. Computation of these quantities for the original SGoF method for independent tests and its conservative version and for the BH method (with a nominal FDR of 5%) was also included to compare.

2.2 Principal results

Tables 2.1 to 2.9 reported in this section are a sample of the full set of results of the simulations. Due to the large extension of the results, they are restricted to case k=10 and k=20 blocks, n=1000 tests, no effects, 10% of effects or 33% of effects, weak or strong effects ($\mu = 1$ or $\mu = 4$), and within-block correlation $\rho = 0$ (independent setting), $\rho = 0.2$ (moderate correlation), and $\rho = 0.8$ (strong correlation). We collect the remaining Tables in Section 2.3 and Appendix A.

In each table we report the FDR, Power (POW) and the Coverage of seven methods: SGoF, conservative SGoF, BH, BB-SGoF(k), BB-SGoF(k/2), BB-SGoF(2k) 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).

- **Complete null hypothesis**

In first place we are going to 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 hypothesis, all R_n rejected hypothesis are Type I errors and $FDR = FWER$. Obviously, the power in all these situations is 100% since there aren't effects. Moreover,

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.059	1	0.941	0.3142255	0.1835811	0.991	0.1333157	0.3032177	1
Conservative SGoF	0.048	1	0.952	0.3011511	0.1681742	1	0.1260759	0.2851799	1
BH	0.057	1	0.943	0.0491119	0.007868882	1	0.03525682	0.03005536	1
BB-SGoF (k)	0.047	1	0.953	0.2948606	0.1616586	1	0.1246081	0.2807714	1
BB-SGoF (k/2)	0.044	1	0.956	0.2933545	0.159449	1	0.1239398	0.2790337	1
BB-SGoF (2k)	0.044	1	0.956	0.2967956	0.1635291	1	0.1249923	0.2820841	1
Auto BB-SGoF	0.019	1	0.981	0.2664407	0.1356188	1	0.1193362	0.2678764	1

Table 2.1: $n = 1000, \rho = 0, k = 10, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.044	1	0.956	0.3142255	0.1835811	0.991	0.1333157	0.3032177	1
Conservative SGoF	0.033	1	0.967	0.3011511	0.1681742	1	0.1260759	0.2851799	1
BH	0.047	1	0.953	0.0491119	0.007868882	1	0.03525682	0.03005536	1
BB-SGoF (k)	0.035	1	0.965	0.2967956	0.1635291	1	0.1249923	0.2820841	1
BB-SGoF (k/2)	0.036	1	0.964	0.2948606	0.1616586	1	0.1246081	0.2807714	1
BB-SGoF (2k)	0.032	1	0.968	0.2968219	0.1648354	1	0.1252099	0.2829085	1
Auto BB-SGoF	0.01	1	0.99	0.2664407	0.1356188	1	0.1193362	0.2678764	1

Table 2.2: $n = 1000, \rho = 0, k = 20, \mu = 1$.

the coverage coincides to $1 - FDR$ as explained above.

For analizing this case, we have to focus on the first columns ($\Pi_0 = 1$) in Tables from 2.1 to 2.6 where we report the results of no effects in the scenario of weak effects ($\mu = 1$). In the case of strong effects ($\mu = 4$) the results observed are the same because we are considering the global null hypothesis, i.e., we consider that there aren't effects so that parameter is irrelevant.

From Tables 2.1 and 2.2 we see that all the methods respect the nominal FDR of 5% fairly well in the independent setting. For example, SGoF, BH and BB-SGoF(k) report an FDR of 0.059, 0.057 and 0.047, respectively, in Table 2.1 ($\rho = 0, k = 10, \mu = 1$) and 0.044, 0.047, 0.035, respectively, in Table 4.2 ($\rho = 0, k = 20, \mu = 1$). The automatic BB-SGoF reports an FDR below nominal (0.019 for k=10 and 0.01 for k=20), something expected due to its conservativeness.

As correlation grows, original SGoF for independent tests loses control of FWER; for example, when $\rho = 0.2$ and $k = 10$ (Table 2.3), $FDR = 0.166$ and when $\rho = 0.8$ and $k = 10$ (Table 4.5), $FDR = 0.351$, i.e., it is almost 7 times the nominal. The same happens to conservative SGoF. It should be pointed out that this loss of control is

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.166	1	0.834	0.3099808	0.1825662	0.914	0.1351912	0.3028321	0.999
Conservative SGoF	0.145	1	0.855	0.2951096	0.1680559	0.959	0.1277309	0.2849147	1
BH	0.05	1	0.95	0.03895788	0.008349476	1	0.03023891	0.03053031	1
BB-SGoF (k)	0.064	1	0.936	0.2732628	0.1468859	0.992	0.1237546	0.2745344	1
BB-SGoF (k/2)	0.077	1	0.923	0.2751432	0.1474479	0.987	0.1238289	0.2743953	1
BB-SGoF (2k)	0.092	1	0.908	0.2841319	0.1555818	0.988	0.1251061	0.2784617	1
Auto BB-SGoF	0.042	1	0.958	0.2523351	0.1260787	0.994	0.1202209	0.2633641	1

Table 2.3: $n = 1000, \rho = 0.2, k = 10, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.136	1	0.864	0.3192687	0.1821618	0.954	0.1357588	0.30256	1
Conservative SGoF	0.112	1	0.888	0.3062038	0.1668602	0.984	0.1286054	0.284541	1
BH	0.036	1	0.964	0.0387131	0.008084847	1	0.0302904	0.02931974	1
BB-SGoF (k)	0.059	1	0.941	0.2937423	0.1530572	0.996	0.1260919	0.2775567	1
BB-SGoF (k/2)	0.06	1	0.94	0.2921625	0.1530107	0.994	0.1260286	0.2771458	1
BB-SGoF (2k)	0.087	1	0.913	0.2999626	0.1590115	0.995	0.1270276	0.2802654	1
Auto BB-SGoF	0.028	1	0.972	0.2686287	0.1289918	0.998	0.1214661	0.2648467	1

Table 2.4: $n = 1000, \rho = 0.2, k = 20, \mu = 1$.

lesser when $k = 20$ (Tables 2.4 and 2.6). This occurs because more blocks implies less dependence.

On the other hand, BB-SGoF methods adapt well to the correlated settings, this is particularly true for the benchmark method which uses the true k and for the automatic method. When the researcher overestimates the number of blocks, the FDR of BB-SGoF is above the nominal as we can see in the Table 2.5 (FDR=0.118 for $\rho = 0.8$); this is because BB-SGoF decision becomes more liberal as the assumed dependence structure gets weaker. As regards BH method, it respects the nominal FDR regardless the value of ρ, which is expected due to its robustness for dependences. For example, it reports exactly a FDR of 0.05 in Table 2.3 ($\rho = 0.2$), although it is very conservative in the case $\rho = 0.8$ (FDR=0.028).

Summarizing, the results for BB-SGoF are relevant since they suggest FWER control (in the weak sense) even when the simulated model is not beta-binomial.

- **Weak effects**

The situation with 33% ($\Pi_0 = 0.67$) of weak effects (Tables from 2.1 to 2.6) reveals that SGoF-type strategies are not controlling FDR at any given level. For example,

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.351	1	0.649	0.2252081	0.1528695	0.73	0.1159219	0.2986824	0.832
Conservative SGoF	0.341	1	0.659	0.2157467	0.1426852	0.749	0.1087877	0.2813481	0.886
BH	0.028	1	0.972	0.02606445	0.01244698	0.99	0.02626762	0.04078015	1
BB-SGoF (k)	0.059	1	0.941	0.1372528	0.07316941	0.915	0.08465275	0.2195398	0.992
BB-SGoF (k/2)	0.049	1	0.951	0.1119436	0.06404142	0.96	0.08568549	0.2256876	0.995
BB-SGoF (2k)	0.118	1	0.882	0.1636489	0.09756053	0.856	0.09359753	0.2434332	0.981
Auto BB-SGoF	0.024	1	0.976	0.08273894	0.04767175	0.983	0.07777103	0.2054436	0.999

Table 2.5: $n = 1000, \rho = 0.8, k = 10, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.305	1	0.695	0.2838611	0.1752458	0.715	0.1243796	0.2975322	0.902
Conservative SGoF	0.29	1	0.71	0.2715604	0.1627739	0.757	0.1179747	0.2796748	0.951
BH	0.037	1	0.963	0.03620277	0.01297738	0.998	0.02755297	0.0355607	1
BB-SGoF (k)	0.055	1	0.945	0.2031109	0.107085	0.888	0.1024949	0.2393145	0.997
BB-SGoF (k/2)	0.029	1	0.971	0.1737425	0.08645127	0.948	0.1018557	0.2384888	0.998
BB-SGoF (2k)	0.11	1	0.89	0.2293533	0.127696	0.843	0.1086098	0.256391	0.988
Auto BB-SGoF	0.016	1	0.984	0.1203519	0.05802466	0.976	0.09508935	0.2212269	0.999

Table 2.6: $n = 1000, \rho = 0.8, k = 20, \mu = 1$.

in the independent setting, original SGoF and benchmark BB-SGoF report a FDR of 13.3% and 12.5% respectively (Table 2.1), more than two times the nominal FDR for BH procedure. Results for the dependent setting are of the same order, although for strong correlation ($\rho = 0.8$) these FDRs go down to 11.5% and 8.5% respectively (Table 2.5). However, the proportion of true effects detected by SGoF-type methods is between 5 and 9 times that of BH, the relative performance of SGoF getting better as correlation decreases. At the same time, one may say that BB-SGoF is not detecting 'too many effects' in the sense that, in at least 98.1% of the trials (worst situation, Table 2.5), the number of declared effects is below the number of true effects with p-value below γ. It is not strange that this proportion is just 100% for BH since this method is rejecting only between 3% and 4% of the existing effects. Interestingly, automatic BB-SGoF does not lose much power to respect to its optimal version based on the true number of blocks: its power is 6.4% smaller in the worse situation ($\rho = 0.8$).

With respect to the situation with 10% ($\Pi_0 = 0.9$) of weak effects shown in Tables 2.1-2.6, we see that the only method which respects the FDR al level α is BH. BB-SGoF shows in Table 2.1 a FDR of 0.295 with 10% of weak effects which is more than twice its value with 33% of weak effects (0.125).

On the other hand, the power of automatic BB-SGoF relative to BH was 17 under independence (POW=0.136 and 0.008 respectively, Table 2.1) and above 15 with $\rho = 0.2$ (POW=0.126 and 0.008 respectively, Table 2.3).

- **Strong effects**

Information corresponding to strong effects ($\mu = 4$) is shown in Tables from 2.7 to 2.9.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.0006623815	0.8435305	0.992	0.0001362205	0.9169738	1
Conservative SGoF	0.0002397667	0.7808199	0.999	9.460649e-05	0.8779448	1
BH	0.0443865	0.9998821	0.016	0.03331245	0.9999939	0
BB-SGoF (k)	0.0522245	0.7622275	0.94	8.553288e-05	0.869292	1
BB-SGoF (k/2)	0.0144851	0.7673286	0.98	8.577512e-05	0.8662092	1
BB-SGoF (2k)	0.04587234	0.7711299	0.948	9.204598e-05	0.8716593	1
Auto BB-SGoF	0.0001472093	0.5430974	1	7.653829e-05	0.8532729	1

Table 2.7: $n = 1000, \rho = 0, k = 10, \mu = 4$.

The case with 33% of strong effects allows to see that, in some instances, the FDR of SGoF-type methods may be very small compared to γ (the p-value threshold) or α (the FWER-controlling parameter under the complete null). For example, Tables 2.7, 2.8 and 2.9 indicate that, for the simulated settings, the average proportion of false discoveries of benchmark BB-SGoF lies between 0.07/1000 and 0.4/1000, being even smaller for its automatic version. The reason for this is that, with such strong effects, the non-true nulls report very small p-values, which are clearly separated from those of the true nulls. Still, automatic BB-SGoF is able to detect more than 80% of the existing effects (POW=0.85,0.85, 0.81 in the Tables 2.7-2.9 respectively).

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.004970638	0.8366053	0.935	0.0001785485	0.9172247	0.999
Conservative SGoF	0.001948649	0.7769547	0.977	8.596037e-05	0.8782517	1
BH	0.04512081	0.9998562	0.03	0.03338453	1	0
BB-SGoF (k)	0.0357232	0.727603	0.948	6.714235e-05	0.8673756	1
BB-SGoF (k/2)	0.01609509	0.7460989	0.97	6.747558e-05	0.8650048	1
BB-SGoF (2k)	0.03794572	0.7485775	0.943	8.043789e-05	0.8707239	1
Auto BB-SGoF	0.0002265047	0.4992389	0.998	6.166717e-05	0.8503369	1

Table 2.8: $n = 1000, \rho = 0.2, k = 10, \mu = 4$.

On the other hand, the power of BH procedure is larger than that, according to its higher FDR (0.03); indeed, this power is almost 100% in all the cases. This situation may be regarded as non-optimal however in the sense of the coverage; for example, in the case $\rho = 0.8$, $k = 10$ (Table 2.9), only for 17% of the 1000 Monte Carlo trials the number of effects declared by BH was below the true number of effects with p-value smaller than 0.05, showing an anticonservative performance in this sense (this percentage was even smaller for the other correlation levels and values of k). Also importantly, as for the case with weak effects, the automatic choice of the number of blocks results in a small loss of power (smaller than 2.5% in this case).

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07801752	0.7685382	0.671	0.009759161	0.9046159	0.832
Conservative SGoF	0.06222644	0.7319147	0.719	0.004969988	0.8708095	0.91
BH	0.05045353	0.9998874	0.378	0.03203567	0.9999941	0.17
BB-SGoF (k)	0.02827	0.6296566	0.888	0.0004019391	0.8257775	0.995
BB-SGoF (k/2)	0.01941952	0.5965438	0.946	0.0003115351	0.8249183	0.996
BB-SGoF (2k)	0.0475203	0.6752641	0.818	0.001211916	0.8455773	0.982
Auto BB-SGoF	0.004295736	0.4828479	0.978	9.543074e-05	0.8053235	1

Table 2.9: $n = 1000, \rho = 0.8, k = 10, \mu = 4$.

The case of 10% of strong effects shows that, as in the case of 33% of strong effects, the FDR of SGoF-type methods may be small compared to γ or α although in a lesser degree and excepting SGoF(k) and SGoF(2k), which tend to report a FDR close to 0.05. The lowest FDR and power is reported by Auto BB-SGoF in every case. In particular, Auto BB-SGoF loses about 30% of power compared to its benchmark version.

On the other hand, the performance of BH method is very similar to the case of 33% of strong effects, reporting a power of nearly one and a coverage of 0.016, 0.03 and 0.378 in Tables 2.7, 2.8 and 2.9.

Finally, as a overview of the 'Coverage' of the different methods we have observed that in the case of weak effects ($\mu = 1$), the proportion of trials for which the number of declared effects was not larger than the number of effects with p-value below γ (which should be 95% in SGoF-type methods) is between the 83% and the 100%, increasing with the proportion of effects. See Tables from 2.1 to 2.6. Moreover, in the scenarios where exists strong correlation ($\rho = 0.8$), we see that the 'Coverage' is a bit lower.

On the other hand, when we consider strong effects ($\mu = 4$), we observe a difference. While before the 'Coverage' of BH method was rounding the 100%, in this situation this proportion is zero (Tables 2.7 and 2.8) and a little bit higher in the case of strong correlation (Table 2.9).

2.3 Mention to other scenarios

• Intermediate effects

Important differences were seen when considering intermediate effects ($\mu = 2$) rather than weak ($\mu = 1$) or strong ($\mu = 4$) effects.

When we consider the case of 33% of intermediate effects, shown e.g. in Tables 2.10 and 2.11, we note that BH and BB-SGoF procedures performed similarly in FDR, power and coverage.

But if we consider the 10% of intermediate effects we see that BH and SGoF-type strategies reports very different values of FDR. In this context, Auto BB-SGoF is very conservative and loses much power (around 30%). The other SGoF-type methods do not control the FDR at level $\alpha = 0.05$. On the other hand, BH reported a FDR of 0.045 in Tables 2.10 and 2.11 although it performed similarly to the other procedures in power and coverage.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07914565	0.7245016	0.996	0.03415781	0.8339071	1
Conservative SGoF	0.06500829	0.679775	1	0.02776494	0.8018862	1
BH	0.04490863	0.6099535	1	0.0333315	0.8318117	1
BB-SGoF (k)	0.08239402	0.6576541	0.971	0.02671155	0.7951514	1
BB-SGoF (k/2)	0.06824712	0.6666408	0.988	0.02640069	0.7922354	1
BB-SGoF (2k)	0.09978532	0.6779481	0.954	0.02699543	0.7968584	1
Auto BB-SGoF	0.03706632	0.488855	1	0.02458957	0.7796056	1

Table 2.10: $n = 1000, \rho = 0, k = 10, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.08294895	0.715361	0.93	0.03459625	0.8327661	0.994
Conservative SGoF	0.06871646	0.6714748	0.975	0.02800241	0.8010118	1
BH	0.04524812	0.6018802	0.999	0.03330931	0.830528	0.999
BB-SGoF (k)	0.08232657	0.6301188	0.96	0.02798615	0.773131	0.996
BB-SGoF (k/2)	0.07054939	0.6515584	0.976	0.04691757	0.7974571	0.964
BB-SGoF (2k)	0.09119942	0.6589678	0.956	0.02713911	0.7958391	1
Auto BB-SGoF	0.03386669	0.4438559	1	0.01909913	0.6767205	1

Table 2.11: $n = 1000, \rho = 0.2, k = 10, \mu = 2$.

- **Correlation of $\rho = 0.1$**

When we consider a correlation of $\rho = 0.1$ the simulations reported similar results of the case of correlation of $\rho = 0.2$ as we can see comparing Tables 2.10 and 2.12.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.08049727	0.7199804	0.978	0.03413379	0.8338533	1
Conservative SGoF	0.0658443	0.6759005	1	0.02794623	0.8017551	1
BH	0.04529007	0.6087283	1	0.03309399	0.8315923	1
BB-SGoF (k)	0.1023055	0.6539831	0.945	0.02756254	0.7755336	0.997
BB-SGoF (k/2)	0.07287263	0.6647526	0.983	0.06152693	0.8048206	0.942
BB-SGoF (2k)	0.09407412	0.6672005	0.959	0.0290754	0.7976273	0.997
Auto BB-SGoF	0.0365238	0.4755984	1	0.01973914	0.6918309	1

Table 2.12: $n = 1000, \rho = 0.1, k = 10, \mu = 2$.

- **Influence of the number of test (n)**

We end this section by summarizing the simulations with $n = 500$ tests. The full set of tables are reported in Appendix A. As an example we show two cases.

In first place we show the tables corresponding to the situation of $\rho = 0.1, k = 20$, $\mu = 1$ and $n = 1000, 500$, respectively.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.085	1	0.915	0.3220994	0.1832119	0.981	0.134951	0.3031286	1
Conservative SGoF	0.06	1	0.94	0.3088758	0.1680976	0.994	0.1281242	0.2849963	1
BH	0.045	1	0.955	0.04267951	0.007386369	1	0.03123081	0.02896324	1
BB-SGoF (k)	0.044	1	0.956	0.3010201	0.1606106	0.998	0.1268084	0.2809359	1
BB-SGoF (k/2)	0.051	1	0.949	0.3020164	0.1592527	0.996	0.1264472	0.2801197	1
BB-SGoF (2k)	0.049	1	0.951	0.3049167	0.1638152	0.995	0.1272625	0.2821311	1
Auto BB-SGoF	0.016	1	0.984	0.2733422	0.1347397	0.999	0.1221225	0.267899	1

Table 2.13: $n = 1000, \rho = 0.1, k = 20, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.086	1	0.914	0.29242	0.1689248	0.974	0.131436	0.2963669	1
Conservative SGoF	0.07	1	0.93	0.2681197	0.1401886	0.993	0.120855	0.2682715	1
BH	0.051	1	0.949	0.04890476	0.01333448	1	0.03201893	0.03577015	1
BB-SGoF (k)	0.05	1	0.95	0.2580782	0.1325003	0.995	0.1189151	0.2635621	1
BB-SGoF (k/2)	0.055	1	0.945	0.2557564	0.1306552	0.996	0.1184454	0.2622155	1
BB-SGoF (2k)	0.059	1	0.941	0.26139	0.1347139	0.994	0.1193397	0.2643708	1
Auto BB-SGoF	0.019	1	0.981	0.200694	0.09430312	0.999	0.1116905	0.243856	1

Table 2.14: $n = 500, \rho = 0.1, k = 20, \mu = 1$.

In Tables 2.13 and 2.14 we see that under the complete null the results reported are very similar although the FDR tends to be higher when n=500, contrary what happens to the coverage. In the cases with effects, it occurs that the FDR and power reported by BH is lower when n=1000. On the other hand, SGoF-type methods reported lower FDR, power and coverage with n=500. This fact has an theoretical explanation based in the construction of this kind of methods because an increase in n produces also an increase of the threshold p-value although this result is not so clear when strong correlation is present. These results have already been obtained in previous simulation studies.

In second place, we show Tables 2.15 and 2.16 corresponding to a situation of strong effects.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.04865266	0.80579	0.709	0.003969961	0.9106936	0.901
Conservative SGoF	0.03373382	0.7632626	0.792	0.001484913	0.8742624	0.963
BH	0.04580367	0.9998491	0.255	0.03292385	1	0.07
BB-SGoF (k)	0.01807205	0.6806418	0.888	0.0001706272	0.8426825	0.998
BB-SGoF (k/2)	0.01039443	0.6440988	0.949	0.0001470393	0.8400056	0.998
BB-SGoF (2k)	0.04191248	0.7263315	0.835	0.0004864151	0.8562382	0.989
Auto BB-SGoF	0.001371311	0.505823	0.985	7.679685e-05	0.8194748	1

Table 2.15: $n = 1000, \rho = 0.8, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.03924299	0.7743521	0.76	0.004118932	0.905642	0.915
Conservative SGoF	0.02164147	0.6999143	0.869	0.001066003	0.8503739	0.98
BH	0.04227685	0.9998862	0.415	0.03347978	1	0.135
BB-SGoF (k)	0.007961097	0.6197648	0.936	0.0001887994	0.821963	0.995
BB-SGoF (k/2)	0.003123153	0.5922429	0.972	0.0001799006	0.8205609	0.998
BB-SGoF (2k)	0.0112401	0.6547012	0.917	0.000419386	0.8344119	0.991
Auto BB-SGoF	0.001226188	0.5191446	0.988	7.577354e-05	0.7935625	1

Table 2.16: $n = 500, \rho = 0.8, k = 20, \mu = 4$.

The results reported by SGoF-type strategies in this case are the same to the previous ones. With the 10% and 33% of effects, the FDR and the power reported is always lower when $n = 500$. Moreover, these methods are more conservative when $n = 1000$ in the sense of the coverage.

On the other hand, due to the strong effects, the results reported by BH are similar when n=500 and n=1000. Another fact that we observe on these tables is that, in this scenario, the power of BH is very high, in fact, it is the higher power of all methods. It seems that the nominal level $\alpha = 0.05$ results very high in this context because SGoF-type methods reported a FDR very low compared to 0.05. Moreover, if we focus on the situation of 33% of effects, we see that the coverage of this method is very low, i.e., BH rejected more hypothesis than it should.

2.4 Automatic number of blocks

The number of blocks of dependent tests detected by automatic BB-SGoF was not always close to the true k. In order to illustrate this result we report in the Tables 2.17, 2.18 and 2.19 the number of blocks detected on average and it standard deviations (in brackets), in the case of the true number of blocks is k=10 and n=1000.

In the first table we show the situation of the complete null hypothesis in which case the number of blocks detected was 18.1 (independent setting), 10.4 ($\rho = 0.2$), or 6.9 ($\rho = 0.8$) on average, therefore being decreasing with an increasing correlation. Corresponding standard deviations were 16.7, 12.4, and 10.5, showing a large variability of the selected number of blocks along replicates.

	Π_0=1
ρ=0	18.1(16.7)
ρ=0.2	10.4(12.4)
ρ=0.8	6.9(10.5)

Table 2.17: $n = 1000, k = 10, \mu = 1$.

On the other hand, in the other two tables we see that the average number of blocks detected was decreasing for an increasing proportion of effects although it is not so clear

when $\rho = 0.8$, because in the case of $\mu = 1$ (Table 2.18) this affirmation isn't true.

	Π_0=0.9	Π_0=0.67
ρ=0	13.8(14.7)	9.1(12.7)
ρ=0.2	10.1(11.9)	8.3(10.5)
ρ=0.8	4.6(4.8)	5.2(3.5)

Table 2.18: $n = 1000, k = 10, \mu = 1$.

	Π_0=0.9	Π_0=0.67
ρ=0	16.9(15.6)	5.9 (10.2)
ρ=0.2	15.7 (14.8)	6.3 (10.3)
ρ=0.8	9.2 (11.9)	5.8(7.8)

Table 2.19: $n = 1000, k = 10, \mu = 4$.

Whatever the case, one should keep in mind that the role of automatic BB-SGoF is not to perfectly estimate the number of existing blocks but rather to allow for error control in the multitesting procedure when the value of k is unknown.

2.5 Tarone test

As we said, Tarone (1979) introduced a test for the binomial model $H_0^T : \rho_0 = 0$ against the beta-binomial alternative $H_1^T : \rho_0 > 0$. Here we denote by ρ_0 the correlation between Bernoulli outcomes $I_{\{u_i \leq \gamma\}}$ sharing the same block, which is not the ρ in the simulations (but we have $\rho_0 = 0$ if $\rho = 0$). In Tables 2.20, 2.21 and 2.22 we show the mean and standard deviation of the p-values obtained in our simulations applying this test and the proportion of p-values that have fallen below 0.05 along the 1000 simulations, in the case when the value of k is correctly specified.

	mean p-values	standard deviation p-values	Proportion p-values below 0.05
ρ=0	0.5050367	0.271959	0.067
ρ=0.1	0.3225716	0.2852795	0.257
ρ=0.2	0.1363698	0.2172238	0.596
ρ=0.8	0.00732413	0.06183131	0.979

Table 2.20: $n = 1000, k = 10, \mu = 1, \Pi_0 = 1$.

	mean p-values	standard deviation p-values	Proportion p-values below 0.05
ρ=0	0.5473448	0.2778251	0.052
ρ=0.1	0.4136161	0.2974351	0.167
ρ=0.2	0.2348962	0.2753546	0.413
ρ=0.8	0.023333	0.08792541	0.905

Table 2.21: $n = 1000, k = 10, \mu = 1, \Pi_0 = 0.9$.

	mean p-values	standard deviation p-values	Proportion p-values below 0.05
ρ=0	0.5596636	0.2746229	0.049
ρ=0.1	0.4488361	0.29677	0.123
ρ=0.2	0.2958365	0.2909711	0.3
ρ=0.8	0.003640023	0.03134645	0.986

Table 2.22: $n = 1000, k = 10, \mu = 1, \Pi_0 = 0.67$.

When $\rho = 0$ (null hypothesis of independence) the proportion of p-values below 0.05 is close to 5%, indicating that Tarone's test respects the level well. As ρ departs from zero, this proportion grows (as expected); this power of Tarone's test is decreasing for an increasing proportion of effects when $\rho \leq 0.2$, which means that dependence is more early detected under the complete null. However, the situation is the opposite for $\rho = 0.8$; in the case of strong dependence. The power is maximum with a 33% of effects. More investigation of this issue seems to be needed before reaching general conclusion on this regard.

Chapter 3

Conclusion and Future research

- **Conclusion**

In this work we have investigated through simulations the performance of BB-SGoF method. Rate of false discoveries (FDR), proportion of detected effects (power), and conservativeness with respect to the true number of effects with p-value smaller than the given threshold have been computed. One conclusion of our research is that BB-SGoF method may control for FWER in the weak sense even when the underlying model is not beta-binomial. 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. As a compromise, the automatic BB-SGoF procedure introduced in de Uña-Álvarez (2012) performs well, with only a small loss of power with respect to the benchmark version when the effects are weak or the proportion of effects is moderate (33%). Another interesting finding was the ability of Tarone's test to detect dependence in practice.

Summarizing, BB-SGoF is a correction of SGoF method with a suitable error control in the presence of dependent tests; its advantages over classical FDR-controlling strategies (e.g. the BH method) remain the same in the dependence scenario as for SGoF in the independent setting, these are: greater power in the case of large number of tests and small to moderate number of weak effects. In such cases application of BB-SGoF is recommended due to its compromise between FDR and power.

- **Future research**

Future lines of research include the study of BB-SGoF in simulated scenarios with blocks of unequal sizes, and the modification of BB-SGoF for dependence situations other than beta-binomial.

Moreover, we aim to develop a *R* package to apply the SGoF and BB-SGoF methods to real data where unadjusted p-values are the only input needed.

Appendix A

Tables

In this appendix we show the remainder of the tables obtained in our simulations. They basically correspond to the case of n=1000 test and intermediate effects ($\mu = 2$), n=1000 test and within-block correlation $\rho = 0.1$, and to the case of n=500 tests.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07923771	1	0.9207623	0.2883931	0.1647738	0.983	0.130887	0.2960884	1
Conservative SGoF	0.05315948	1	0.9468405	0.2671621	0.1358507	0.998	0.1207597	0.2680044	1
BH	0.05717151	1	0.9428285	0.03655794	0.01254583	1	0.03176723	0.03536049	1
BB-SGoF (k)	0.0441324	1	0.9558676	0.2533562	0.1274706	0.998	0.1192404	0.2629619	1
BB-SGoF (k/2)	0.03811434	1	0.9618857	0.2527085	0.1267142	0.998	0.1187473	0.2616227	1
BB-SGoF (2k)	0.04613842	1	0.9538616	0.2577212	0.1293885	0.998	0.1193486	0.2634724	1
Auto BB-SGoF	0.01203611	1	0.9879639	0.1945955	0.08972874	1	0.1133543	0.2428961	1

Table A.1: $n = 500, \rho = 0, k = 10, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.08	1	0.92	0.3038636	0.1660078	0.973	0.1327718	0.2954172	1
Conservative SGoF	0.051	1	0.949	0.2816662	0.1366357	0.994	0.1228241	0.26725	1
BH	0.056	1	0.944	0.04249762	0.01170231	1	0.03405399	0.03543598	1
BB-SGoF (k)	0.043	1	0.957	0.2687651	0.1304969	0.998	0.1216838	0.2628925	1
BB-SGoF (k/2)	0.043	1	0.957	0.2684266	0.1282721	0.996	0.1207352	0.2614755	1
BB-SGoF (2k)	0.041	1	0.959	0.2724224	0.131518	0.996	0.1215773	0.2638062	1
Auto BB-SGoF	0.013	1	0.987	0.2186375	0.09265339	1	0.1139555	0.2419895	1

Table A.2: $n = 500, \rho = 0, k = 20, \mu = 1$.

	Π_0=01			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.171	1	0.829	0.2903807	0.1653884	0.93	0.1324483	0.2937646	0.995
Conservative SGoF	0.142	1	0.858	0.2666627	0.1376886	0.973	0.1225309	0.2656056	1
BH	0.055	1	0.945	0.05172857	0.01401695	1	0.03322214	0.03648515	1
BB-SGoF (k)	0.059	1	0.941	0.2437895	0.1192209	0.989	0.1187629	0.2553919	1
BB-SGoF (k/2)	0.073	1	0.927	0.2432813	0.1205237	0.986	0.1184741	0.25527	1
BB-SGoF (2k)	0.088	1	0.912	0.2536097	0.1262237	0.984	0.1203134	0.2586802	1
Auto BB-SGoF	0.035	1	0.965	0.1961633	0.08922137	0.994	0.1129477	0.236596	1

Table A.3: $n = 500, \rho = 0.2, k = 10, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.129	1	0.871	0.2958333	0.1649297	0.954	0.1309994	0.2959077	0.997
Conservative SGoF	0.092	1	0.908	0.2701832	0.1364936	0.986	0.1202751	0.2678582	1
BH	0.055	1	0.945	0.05015714	0.0130113	1	0.03246352	0.03553716	1
BB-SGoF (k)	0.056	1	0.944	0.2550858	0.1247904	0.992	0.1178665	0.2616067	1
BB-SGoF (k/2)	0.058	1	0.942	0.2498101	0.1246623	0.994	0.117946	0.2610592	1
BB-SGoF (2k)	0.07	1	0.93	0.2608335	0.1297773	0.991	0.118976	0.2634785	1
Auto BB-SGoF	0.017	1	0.983	0.209666	0.0894413	0.998	0.1123566	0.2422092	1

Table A.4: $n = 500, \rho = 0.2, k = 20, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.369	1	0.631	0.240466	0.1594017	0.705	0.1203558	0.2963639	0.841
Conservative SGoF	0.356	1	0.644	0.2231027	0.1393618	0.751	0.1102929	0.2686499	0.919
BH	0.04	1	0.96	0.03411705	0.02087547	0.989	0.02823968	0.04901628	0.997
BB-SGoF (k)	0.067	1	0.933	0.1343833	0.07456597	0.931	0.08655283	0.2173989	0.993
BB-SGoF (k/2)	0.049	1	0.951	0.1173026	0.06614499	0.963	0.08737575	0.2216819	0.994
BB-SGoF (2k)	0.122	1	0.878	0.1651975	0.09616779	0.877	0.09629241	0.2383898	0.981
Auto BB-SGoF	0.029	1	0.971	0.08639863	0.04843845	0.98	0.07920378	0.2005265	0.997

Table A.5: $n = 500, \rho = 0.8, k = 10, \mu = 1$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.3105023	1	0.6860622	0.2534057	0.1573894	0.7565217	0.123459	0.2953726	0.9141165
Conservative SGoF	0.2876712	1	0.7104195	0.2284745	0.1354807	0.8356522	0.1137488	0.2673717	0.9615005
BH	0.03652968	1	0.9607578	0.02983517	0.01634894	0.9973913	0.02745042	0.04033666	1
BB-SGoF (k)	0.05993151	1	0.9424899	0.1616745	0.08443414	0.9426087	0.1005751	0.2334144	0.9960513
BB-SGoF (k/2)	0.04509132	1	0.9566982	0.1351943	0.07372546	0.9669565	0.1003836	0.233322	0.9970385
BB-SGoF (2k)	0.1358447	1	0.8653586	0.1850177	0.1062113	0.9095652	0.1071929	0.2508791	0.988154
Auto BB-SGoF	0.02511416	1	0.9763194	0.09514541	0.05026715	0.9895652	0.09300872	0.2139823	1

Table A.6: $n = 500, \rho = 0.8, k = 20, \mu = 1$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.001245707	0.8084971	0.9768946	0.0001595287	0.9079118	1
Conservative SGoF	0.0001024381	0.7105795	0.9990758	6.323567e-05	0.8491235	1
BH	0.04414288	0.999795	0.1081331	0.03325835	1	0.003
BB-SGoF (k)	8.352882e-05	0.6923637	0.9990758	5.944111e-05	0.8383731	1
BB-SGoF (k/2)	0.0001609868	0.6879245	0.9990758	5.929891e-05	0.8362523	1
BB-SGoF (2k)	8.505639e-05	0.6946836	0.9990758	5.912697e-05	0.8404261	1
Auto BB-SGoF	7.318456e-05	0.6106294	1	5.650203e-05	0.8149006	1

Table A.7: $n = 500, \rho = 0, k = 10, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.005426159	0.8055818	0.936	0.0002347172	0.9071988	0.997
Conservative SGoF	0.001232466	0.711157	0.986	0.0001117649	0.8484888	1
BH	0.04535243	0.9998605	0.122	0.03365939	0.9999884	0.011
BB-SGoF (k)	0.000590578	0.6786624	0.995	8.569257e-05	0.835505	1
BB-SGoF (k/2)	0.0007331328	0.6771473	0.992	8.686684e-05	0.8341268	1
BB-SGoF (2k)	0.0007687861	0.6893755	0.994	9.301219e-05	0.8388986	1
Auto BB-SGoF	0.0004238822	0.605448	0.997	5.128806e-05	0.8112093	1

Table A.8: $n = 500, \rho = 0.2, k = 10, \mu = 4$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.086	1	0.914	0.001354385	0.8084427	0.978	0.0001210263	0.9076068	1
Conservative SGoF	0.054	1	0.946	0.0001132651	0.7106456	0.999	5.053945e-05	0.8488055	1
BH	0.043	1	0.957	0.04449087	0.9997575	0.116	0.03324609	1	0.003
BB-SGoF (k)	0.04	1	0.96	0.0001153365	0.694705	0.999	5.103266e-05	0.8400295	1
BB-SGoF (k/2)	0.044	1	0.956	0.0001132651	0.6919771	0.999	5.136799e-05	0.8376673	1
BB-SGoF (2k)	0.048	1	0.952	9.572863e-05	0.6984297	1	5.129003e-05	0.8410926	1
Auto BB-SGoF	0.017	1	0.983	9.923827e-05	0.6111224	1	4.593761e-05	0.8135864	1

Table A.9: $n = 500, \rho = 0, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.003488518	0.807919	0.955	0.0002020055	0.9077372	0.997
Conservative SGoF	0.0005936609	0.7117471	0.995	5.458796e-05	0.8488905	1
BH	0.04464125	0.9999152	0.13	0.03303471	1	0.009
BB-SGoF (k)	0.0002821696	0.6877762	0.998	4.201005e-05	0.8394739	1
BB-SGoF (k/2)	0.0002331002	0.6859521	0.998	4.197309e-05	0.8371173	1
BB-SGoF (2k)	0.0003722998	0.6957992	0.997	4.225438e-05	0.840876	1
Auto BB-SGoF	0.0001447601	0.6092411	0.999	3.713064e-05	0.8108977	1

Table A.10: $n = 500, \rho = 0.2, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.06736602	0.7427713	0.705	0.01006026	0.9010691	0.84
Conservative SGoF	0.04495911	0.6797582	0.784	0.004127951	0.8494872	0.935
BH	0.04433698	0.9998784	0.477	0.0347614	0.9999757	0.233
BB-SGoF (k)	0.01400388	0.5762192	0.914	0.0003824683	0.8073484	0.996
BB-SGoF (k/2)	0.005283893	0.5478367	0.968	0.0005792938	0.8068373	0.993
BB-SGoF (2k)	0.02508536	0.6197454	0.871	0.00116358	0.8250898	0.981
Auto BB-SGoF	0.002920663	0.4818511	0.986	0.0003006911	0.7814984	0.997

Table A.11: $n = 500, \rho = 0.8, k = 10, \mu = 4$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.101	1	0.899	0.2988915	0.1661491	0.975	0.1336018	0.2956407	0.998
Conservative SGoF	0.071	1	0.929	0.2669266	0.1378108	0.994	0.1232211	0.2675328	1
BH	0.051	1	0.949	0.04095476	0.01217435	1	0.02801983	0.03368031	1
BB-SGoF (k)	0.043	1	0.957	0.2526311	0.1267643	0.995	0.1209753	0.2605467	1
BB-SGoF (k/2)	0.043	1	0.957	0.2525723	0.127261	0.997	0.120685	0.2595744	1
BB-SGoF (2k)	0.051	1	0.949	0.2572166	0.1299361	0.996	0.1211933	0.2624426	1
Auto BB-SGoF	0.017	1	0.983	0.1991581	0.09147796	0.999	0.1144367	0.2417406	1

Table A.12: $n = 500, \rho = 0.1, k = 10, \mu = 1$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.001923969	0.8054065	0.974	0.0001955776	0.9083556	0.999
Conservative SGoF	0.0002264906	0.7083698	1	5.739242e-05	0.8493755	1
BH	0.0446109	0.9997978	0.105	0.03335398	1	0.008
BB-SGoF (k)	0.0002141333	0.684281	1	5.135563e-05	0.8385115	1
BB-SGoF (k/2)	0.0002198092	0.6837358	1	5.157976e-05	0.8360728	1
BB-SGoF (2k)	0.0002154179	0.6910005	1	5.128073e-05	0.8406937	1
Auto BB-SGoF	0.0001307692	0.6055045	1	3.038274e-05	0.8142308	1

Table A.13: $n = 500, \rho = 0.1, k = 10, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.001971776	0.8076407	0.976	9.046559e-05	0.9079766	1
Conservative SGoF	0.0003018631	0.7105897	0.997	2.845529e-05	0.8491759	1
BH	0.04464897	0.9998564	0.118	0.03316743	0.9999945	0.01
BB-SGoF (k)	0.0001721084	0.6927552	0.998	2.230361e-05	0.8402856	1
BB-SGoF (k/2)	0.0001525249	0.6889218	0.999	2.237989e-05	0.8380019	1
BB-SGoF (2k)	0.0002281222	0.6971024	0.999	2.239492e-05	0.841421	1
Auto BB-SGoF	6.946254e-05	0.6109991	1	2.300642e-05	0.8133587	1

Table A.14: $n = 500, \rho = 0.1, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.072533	0.6973088	0.977	0.03265808	0.8256552	1
Conservative SGoF	0.0530863	0.623312	0.999	0.02400257	0.7773119	1
BH	0.04404635	0.6067115	1	0.03304869	0.8316294	0.999
BB-SGoF (k)	0.05038741	0.6084955	0.999	0.02263352	0.7675998	1
BB-SGoF (k/2)	0.04921874	0.603523	0.999	0.02243789	0.7655464	1
BB-SGoF (2k)	0.0506564	0.6091134	0.999	0.02295867	0.7696072	1
Auto BB-SGoF	0.03864175	0.5365882	1	0.02025103	0.7439752	1

Table A.15: $n = 500, \rho = 0, k = 10, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.072533	0.6973088	0.977	0.03265808	0.8256552	1
Conservative SGoF	0.0530863	0.623312	0.999	0.02400257	0.7773119	1
BH	0.04404635	0.6067115	1	0.03304869	0.8316294	0.999
BB-SGoF (k)	0.0506564	0.6091134	0.999	0.02295867	0.7696072	1
BB-SGoF (k/2)	0.05038741	0.6084955	0.999	0.02263352	0.7675998	1
BB-SGoF (2k)	0.05082144	0.6133748	0.999	0.02312332	0.7705099	1
Auto BB-SGoF	0.03864175	0.5365882	1	0.02025103	0.7439752	1

Table A.16: $n = 500, \rho = 0, k = 20, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07568323	0.6941068	0.974	0.03301877	0.8242629	1
Conservative SGoF	0.05486157	0.621191	1	0.02448073	0.7758915	1
BH	0.04404874	0.6075871	0.999	0.0332566	0.8309861	0.997
BB-SGoF (k)	0.05012919	0.6031214	1	0.02311982	0.7659528	1
BB-SGoF (k/2)	0.05045137	0.6017372	1	0.02286634	0.7639948	1
BB-SGoF (2k)	0.05160967	0.6074488	1	0.02348001	0.7681017	1
Auto BB-SGoF	0.03816829	0.5369654	1	0.02066399	0.7439237	1

Table A.17: $n = 500, \rho = 0.1, k = 10, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07408192	0.6959091	0.975	0.03265704	0.8258879	1
Conservative SGoF	0.05305046	0.6230924	0.997	0.0245041	0.777106	1
BH	0.04483718	0.6036731	0.999	0.03282497	0.8300028	0.995
BB-SGoF (k)	0.05043844	0.6091867	0.997	0.02348563	0.7693963	1
BB-SGoF (k/2)	0.05013476	0.6062426	0.998	0.02329843	0.7671356	1
BB-SGoF (2k)	0.05100644	0.6128833	0.999	0.02364508	0.7705705	1
Auto BB-SGoF	0.04038673	0.5365966	1	0.02099814	0.7458787	1

Table A.18: $n = 500, \rho = 0.1, k = 20, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07683302	0.6917124	0.935	0.03347812	0.8238487	0.998
Conservative SGoF	0.05674242	0.6189876	0.987	0.02506585	0.7754498	1
BH	0.04487244	0.6099168	1	0.03347651	0.8304555	0.997
BB-SGoF (k)	0.04992505	0.5937152	0.995	0.02338726	0.7649059	1
BB-SGoF (k/2)	0.04985923	0.5937977	0.993	0.02321731	0.7629631	1
BB-SGoF (2k)	0.05201136	0.6025979	0.993	0.02377683	0.7670247	1
Auto BB-SGoF	0.03977856	0.5313044	0.998	0.02085276	0.7411926	1

Table A.19: $n = 500, \rho = 0.2, k = 10, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07763586	0.6949476	0.956	0.03333683	0.8255535	0.997
Conservative SGoF	0.05707085	0.6222882	0.995	0.02524514	0.7766938	1
BH	0.04547261	0.6070965	1	0.0334024	0.8299218	0.998
BB-SGoF (k)	0.05231807	0.6049354	0.998	0.0241714	0.7688535	1
BB-SGoF (k/2)	0.05217206	0.6024414	0.999	0.02394809	0.7665034	1
BB-SGoF (2k)	0.05365607	0.6109247	0.997	0.02429539	0.7699882	1
Auto BB-SGoF	0.04000808	0.5388687	0.999	0.02143568	0.744839	1

Table A.20: $n = 500, \rho = 0.2, k = 20, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.1137246	0.6474072	0.704	0.04093639	0.8194766	0.84
Conservative SGoF	0.09344731	0.5863977	0.779	0.03210974	0.7724307	0.935
BH	0.04234381	0.6110266	0.959	0.03461269	0.8295094	0.91
BB-SGoF (k)	0.05537175	0.4970723	0.918	0.02391017	0.7415504	0.991
BB-SGoF (k/2)	0.04473785	0.4795516	0.959	0.02445863	0.7428574	0.988
BB-SGoF (2k)	0.06931009	0.5334658	0.869	0.02715923	0.7536201	0.978
Auto BB-SGoF	0.03473275	0.4209807	0.977	0.02143062	0.7194113	0.995

Table A.21: $n = 500, \rho = 0.8, k = 10, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.09704707	0.6693581	0.758	0.03626918	0.8233014	0.915
Conservative SGoF	0.07667466	0.6030729	0.861	0.02772658	0.7754364	0.98
BH	0.04186065	0.6090124	0.972	0.03327187	0.831626	0.941
BB-SGoF (k)	0.05352797	0.5354434	0.937	0.02367109	0.7542413	0.995
BB-SGoF (k/2)	0.04624228	0.5160306	0.967	0.02366633	0.7535337	0.995
BB-SGoF (2k)	0.06174804	0.5657011	0.913	0.02539544	0.7632596	0.989
Auto BB-SGoF	0.03662589	0.4553177	0.982	0.02093333	0.7296312	0.997

Table A.22: $n = 500, \rho = 0.8, k = 20, \mu = 2$.

	Π_0=1			Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.117	1	0.883	0.3156009	0.1844194	0.968	0.1346371	0.3028142	1
Conservative SGoF	0.099	1	0.901	0.3019483	0.1690155	0.994	0.1281644	0.2845311	1
BH	0.038	1	0.962	0.03394455	0.008303924	1	0.03192522	0.02999156	1
BB-SGoF (k)	0.07	1	0.93	0.2922072	0.1586136	0.998	0.1255315	0.2785209	1
BB-SGoF (k/2)	0.074	1	0.926	0.2908984	0.157168	0.997	0.1251351	0.2775521	1
BB-SGoF (2k)	0.073	1	0.927	0.296652	0.1624913	0.998	0.12631	0.2806594	1
Auto BB-SGoF	0.037	1	0.963	0.2687569	0.1351199	0.999	0.1210977	0.2670855	1

Table A.23: $n = 1000, \rho = 0.1, k = 10, \mu = 1$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.0006740034	0.8401602	0.987	0.0001362205	0.9169738	1
Conservative SGoF	0.0002311201	0.7776983	1	9.460649e-05	0.8779448	1
BH	0.04548345	0.9998823	0.004	0.03331245	0.9999939	0
BB-SGoF (k)	0.03695506	0.7684464	0.959	9.204598e-05	0.8716593	1
BB-SGoF (k/2)	0.05406239	0.7543145	0.938	8.553288e-05	0.869292	1
BB-SGoF (2k)	0.09967354	0.7971993	0.884	8.531566e-05	0.8728143	1
Auto BB-SGoF	9.252428e-05	0.5581372	1	7.653829e-05	0.8532729	1

Table A.24: $n = 1000, \rho = 0, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.002403633	0.8381385	0.964	0.0001665559	0.9171046	1
Conservative SGoF	0.0007487605	0.7770165	0.991	8.217097e-05	0.8781257	1
BH	0.04502026	0.9998799	0.018	0.03362871	1	0
BB-SGoF (k)	0.03448514	0.7516695	0.956	7.301997e-05	0.8705118	1
BB-SGoF (k/2)	0.0429189	0.7333313	0.944	7.308826e-05	0.8684568	1
BB-SGoF (2k)	0.07640079	0.7857539	0.907	6.965155e-05	0.8723347	1
Auto BB-SGoF	0.0001049291	0.484406	1	6.038407e-05	0.8512987	1

Table A.25: $n = 1000, \rho = 0.2, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.001086717	0.8381086	0.985	0.0001314909	0.9172807	1
Conservative SGoF	0.0003371053	0.7760309	0.999	7.048533e-05	0.8782766	1
BH	0.04533316	0.9998879	0.014	0.03314732	1	0
BB-SGoF (k)	0.04020032	0.7345943	0.952	6.144447e-05	0.8692124	1
BB-SGoF (k/2)	0.01422176	0.760218	0.98	5.462311e-05	0.8663461	1
BB-SGoF (2k)	0.05867043	0.7666504	0.934	6.801737e-05	0.8716981	1
Auto BB-SGoF	0.00012231	0.5337324	1	5.20479e-05	0.8530562	1

Table A.26: $n = 1000, \rho = 0.1, k = 10, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.0009190963	0.8393794	0.984	0.0001338674	0.9175348	1
Conservative SGoF	0.0002703556	0.7770134	0.999	9.05142e-05	0.8785189	1
BH	0.04566781	0.9998923	0.012	0.03358453	0.999997	0
BB-SGoF (k)	0.05212026	0.7627921	0.941	8.101023e-05	0.8719207	1
BB-SGoF (k/2)	0.04155303	0.7456478	0.95	8.125841e-05	0.8696619	1
BB-SGoF (2k)	0.08513919	0.7916971	0.896	8.089809e-05	0.8732959	1
Auto BB-SGoF	0.0001165149	0.5313321	1	7.217647e-05	0.8528029	1

Table A.27: $n = 1000, \rho = 0.1, k = 20, \mu = 4$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07508295	0.721779	0.99	0.03415781	0.8339071	1
Conservative SGoF	0.06109634	0.6770049	1	0.02776494	0.8018862	1
BH	0.04363838	0.6102663	1	0.0333315	0.8318117	1
BB-SGoF (k)	0.1078664	0.6817703	0.94	0.02699543	0.7968584	1
BB-SGoF (k/2)	0.08927069	0.6602341	0.958	0.02671155	0.7951514	1
BB-SGoF (2k)	0.1432331	0.7041613	0.896	0.02718156	0.7979617	1
Auto BB-SGoF	0.03559753	0.4886356	1	0.02458957	0.7796056	1

Table A.28: $n = 1000, \rho = 0, k = 20, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.07936596	0.7212782	0.988	0.03450013	0.8337944	1
Conservative SGoF	0.06503996	0.6768046	0.998	0.0280016	0.8018842	1
BH	0.04604996	0.6086079	1	0.03348249	0.8316665	0.999
BB-SGoF (k)	0.1004214	0.6728464	0.951	0.02705642	0.7967117	1
BB-SGoF (k/2)	0.09869188	0.6596282	0.948	0.02684758	0.7950654	1
BB-SGoF (2k)	0.1622113	0.711834	0.876	0.02733005	0.7978248	1
Auto BB-SGoF	0.03797591	0.4847879	1	0.02473473	0.7802142	1

Table A.29: $n = 1000, \rho = 0.1, k = 20, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.08062639	0.719712	0.955	0.03453465	0.8331109	1
Conservative SGoF	0.06608729	0.6757069	0.992	0.02843441	0.800965	1
BH	0.04539546	0.6078756	1	0.03359248	0.8315152	0.999
BB-SGoF (k)	0.0964806	0.6616635	0.953	0.02827552	0.7966276	0.998
BB-SGoF (k/2)	0.08837635	0.6413935	0.954	0.02730659	0.7743909	0.997
BB-SGoF (2k)	0.1418842	0.6976114	0.897	0.02947609	0.7981732	0.98
Auto BB-SGoF	0.03244395	0.4437876	0.999	0.01898726	0.6707074	1

Table A.30: $n = 1000, \rho = 0.2, k = 20, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.131014	0.6758194	0.648	0.04220296	0.8246939	0.825
Conservative SGoF	0.116342	0.640568	0.695	0.0357867	0.7937275	0.894
BH	0.04905769	0.6003912	0.952	0.03406133	0.8272614	0.915
BB-SGoF (k)	0.07609181	0.5479928	0.883	0.02685931	0.7467612	0.984
BB-SGoF (k/2)	0.0631698	0.5220823	0.93	0.04309167	0.7655717	0.95
BB-SGoF (2k)	0.09564965	0.5868202	0.818	0.03156706	0.7754603	0.958
Auto BB-SGoF	0.0456457	0.4248208	0.96	0.0197259	0.6312541	0.993

Table A.31: $n = 1000, \rho = 0.8, k = 10, \mu = 2$.

	Π_0=0.9			Π_0=0.67		
	FDR	POW	Coverage	FDR	POW	Coverage
SGoF	0.1114072	0.6961378	0.717	0.03658664	0.8296586	0.901
Conservative SGoF	0.09676163	0.6572573	0.784	0.03050042	0.7979467	0.96
BH	0.04737694	0.6060968	0.972	0.03298176	0.8322969	0.955
BB-SGoF (k)	0.07171496	0.5873553	0.894	0.02530539	0.775195	0.998
BB-SGoF (k/2)	0.06904177	0.5650304	0.929	0.02523273	0.7745534	0.996
BB-SGoF (2k)	0.08993389	0.6226097	0.843	0.02744011	0.7846601	0.987
Auto BB-SGoF	0.04220414	0.4413102	0.969	0.02276628	0.7568643	0.998

Table A.32: $n = 1000, \rho = 0.8, k = 20, \mu = 2$.

Appendix B

R code

In this second appendix we provide the code used in the R software to get the simulations. This includes the code for the computation and optimitation of the beta-binomial likelihood, which is a novelty of this work.

```
########################################
#Introduction of the data:
########################################

M=1200 #no of trials
ro=0.2  #the correlation inside the blocks
s=1000 #number of tests
s0=900 #the number of true nulls
pio=s0/s  #the proportion of true nulls
ss1=7;ss2=8  #the sample sizes group 1, group 2
mu1=0  #mean in group 1
mu2=2 #mean in group 2 for effects, 1/3 of cases
mu22=-2   #mean in group 2 for effects, 2/3 of cases
ki=20   #the number of blocks
j0=1  #independent blocks
n1i=s%/%ki   #block size, j=1,...,ki
sigma=matrix(0,s,s)  #the correlation matrix
```

```
##############################
#correlation only in effects:
#s1=s-s0
#mymatrix <- function(s1, l, ro)
#{
#k<-s1/l
#sigma1<-matrix(0,nrow=s1,ncol=s1)
#for (i in seq(1, s1, k))
#{
#j <- i +k -1
#sigma1[i:j,i:j]<-diag(1-ro,k)+matrix(ro,k,k)
#}
#sigma1
#}
#sigma[(s0+1):s,(s0+1):s]<-mymatrix(8,4,0.3)
#sigma[1:s0,1:s0]=diag(1,s0)
##############################
#correlation only in true nulls:
#mymatrix <- function(s0, l, ro)
#{
#k<-s0/l
#sigma1<-matrix(0,nrow=s0,ncol=s0)
#for (i in seq(1, s0, k))
#{
#j <- i +k -1
#sigma1[i:j,i:j]<-diag(1-ro,k)+matrix(ro,k,k)
#}
#sigma1
#}
#sigma[1:s0,1:s0]<-mymatrix(12,1,0.3)
#sigma[-(1:s0),-(1:s0)]=diag(1,s-s0)
##############################
#correlation in both:
```

```
#l=n de bloques
mymatrix <- function(s, l, ro)
{
k<-s/l
sigma<-matrix(0,nrow=s,ncol=s)
for (i in seq(1, s, k))
{
j <- i +k -1
sigma[i:j,i:j]<-diag(1-ro,k)+matrix(ro,k,k)
}
sigma
}
sigma<-mymatrix(s,ki,ro)
############################
nblocks2=rep(ki/2,M)
nblocks3=rep(2*ki,M)
t12=matrix(nrow=M,ncol=s) #statistics
x=matrix(nrow=M,ncol=s)  #the p-values of the t-test
g=0.05  #the gamma value of SGoF
abh=0.05  #FDR of BH
a=0.05  #the alpha controlling FWER for SGoF (weak sense)
Fng=vector(length=M)   #empirical distribution of the p-values in g
ka=qbinom(a,s,g,lower.tail=F) #automatic choice  of  ka (we reject when k>=ka)
x=matrix(nrow=M,ncol=s)#matrix of p-values
k=vector(length=M)#number of p-values below gamma
n=vector(length=M)  #effects declared by SGoF(  k-ka+1 smallest p-values )
n1=vector(length=M)
power=rep(1,M)
v=vector(length=M) #times that the rank of a p-value of a true null is less than
or equal to  n, entering to effects declared by SGoF (false positive)

#aprox normal SGoF:
nz=vector(length=M)
```

```
n1z=vector(length=M)
powerz=rep(1,M)
vz=vector(length=M)

# BH
nbh=vector(length=M)
vbh=vector(length=M)
powerbh=rep(1,M)

# conservative SGoF
Nindep0=vector(length=M) #effects declared
vc=vector(length=M)
powerc=rep(1,M)

#SGOF BB-SGOF Benchmark
Ndepbench=vector(length=M)  #effects declared
vb=vector(length=M)
powerb=rep(1,M)

#BB-SGOF underestimation
nblocks2=vector(length=M)  #no. of blocks
Ndepauto2=vector(length=M) #effects declared
vu=vector(length=M)
poweru=rep(1,M)

#BB-SGOF overestimation
nblocks3=vector(length=M) #no. of blocks
Ndepauto3=vector(length=M) #effects declared
vo=vector(length=M)
powero=rep(1,M)

#automtico1
nblocks51=vector(length=M) #no. of blocks
```

```
Ndepauto51=vector(length=M) #effects declared
va1=vector(length=M)
powera1=rep(1,M)

##################################
#Simulation
##################################

dm=matrix(nrow=M,ncol=s)
se0g=vector(length=M)

for (j in 1:M) {

dm[j,]=rbinom(s,1,pio) #indicator of true/non-true null,1=true null
d=dm[j,]
s0=sum(d)  #n0. of true nulls
s1a=(s-s0)%/%3
s1b=s-s0-s1a

library(mvtnorm)
xx1=rmvnorm(ss1, rep(mu1,s),  sigma,method="chol")
per<-sample(c(rep(mu1,s0),rep(mu2,s1a),rep(mu22,s1b)))
xx2=rmvnorm(ss2,per,sigma,method="chol")
per=per*as.numeric(per>=0)-per*as.numeric(per<0)

ii<-which(per==mu2)
jj<-which(per==0)
per[ii]=0
per[jj]=1

for (k in 1:s){
t12[j,k]=t.test(xx1[,k],xx2[,k],var.equal=T)$statistic
}
```

```
x[j,]=2*(1-pt(abs(t12[j,]),df=ss1+ss2-2))
dm[j,]=per
h=x[j,]
Fng[j]=ecdf(h)(g)
se0g[j]=sqrt(ecdf(h)(g)*(1-ecdf(h)(g))/s)  #variance of Fng under H0: Fng(g)=g
}
Zvalue=matrix(nrow=60,ncol=M)
pvalue=matrix(nrow=60,ncol=M)
low=matrix(nrow=60,ncol=M)
tarone=vector(length=M)
sebench=vector(length=M)
prob=matrix(nrow=60,ncol=M)
ro=matrix(nrow=60,ncol=M)
vp=matrix(nrow=60,ncol=M)
vb1=matrix(nrow=60,ncol=M)
logver=matrix(nrow=60,ncol=M)

for (j in 1:M) {

w=as.numeric(x[j,]<=g)
p=mean(w)

for (l in j0:6) {

k=l+1  #dependent blocks
A=vector(length=k)  #sj
n11=s%/%k
for (i in 1:k){A[i]=sum(w[((i-1)*n11+1):(i*n11)])}
A[k]=sum(w[((k-1)*n11+1):s])
B=c(rep(n11,k-1),length(w[((k-1)*n11+1):s]))   #nj

l1=vector(length=(length(A)))
```

```
L1<-function(pe,rho){
for(i in 1:length(A)){
if(A[i]==0){l1[i]=0} else {l1[i]<-sum(log(pe+(-rho/(rho-1))*(0:(A[i]-1))))}
}
return(sum(l1))}

l2=vector(length=(length(B-A)))
L2<-function(pe,rho){
for(i in 1:length(B-A)){
if((B[i]-A[i])==0){l2[i]=0} else

{l2[i]<-sum(log(1-pe+(-rho/(rho-1))*(0:(B[i]-A[i]-1))))}
}
return(sum(l2))}

l3=vector(length=(length(B)))
L3<-function(pe,rho){
for(i in 1:length(B)){
if(B[i]==0){l3[i]=0} else {l3[i]<-sum(log(1+(-rho/(rho-1))*(0:(B[i]-1))))}
}
return(sum(l3))}

L<-function(pe,rho){
L<-L1(pe,rho)+L2(pe,rho)-L3(pe,rho)
return(L)
}

l111=vector(length=(length(A)))
L111<-function(pe,rho){
for(i in 1:length(A)){
if(A[i]==0){l111[i]=0} else

{l111[i]<-sum((-1)/((pe+(-rho/(rho-1))*(0:(A[i]-1)))^2))}
```

```
}
return(sum(l111))}

l112=vector(length=(length(B-A)))
L112<-function(pe,rho){
for(i in 1:length(B-A)){
if((B[i]-A[i])==0){l112[i]=0} else

 {l112[i]<-sum((-1)/((1-pe+(-rho/(rho-1))*(0:(B[i]-A[i]-1)))^2))}
}
return(sum(l112))}

der11<-function(pe,rho){
der11<-L111(pe,rho)+L112(pe,rho)
return(der11)}

l121=vector(length=(length(A)))
L121<-function(pe,rho){
for(i in 1:length(A)){
if(A[i]==0){l121[i]=0} else

 {l121[i]<-sum((-(0:(A[i]-1)))/((pe+(-rho/(rho-1))*(0:(A[i]-1)))^2))}
}
return(sum(l121))}

l122=vector(length=(length(B-A)))
L122<-function(pe,rho){
for(i in 1:length(B-A)){
if((B[i]-A[i])==0){l122[i]=0} else

 {l122[i]<-sum((0:(B[i]-A[i]-1))/((1-pe+(-rho/(rho-1))*(0:(B[i]-A[i]-1)))^2))}
}
return(sum(l122))}
```

```
der12<-function(pe,rho){
der12<-L121(pe,rho)+L122(pe,rho)
return(der12)}

l221=vector(length=(length(A)))
L221<-function(pe,rho){
for(i in 1:length(A)){
if(A[i]==0){l221[i]=0} else

{l221[i]<-sum((-((0:(A[i]-1))^2))/((pe+(-rho/(rho-1))*(0:(A[i]-1)))^2))}
}
return(sum(l221))}

l222=vector(length=(length(B-A)))
L222<-function(pe,rho){
for(i in 1:length(B-A)){
if((B[i]-A[i])==0){l222[i]=0} else

{l222[i]<-sum((-((0:(B[i]-A[i]-1))^2))/((1-pe+(-rho/(rho-1))*(0:(B[i]-A[i]-1)))^2))}
}
return(sum(l222))}

l223=vector(length=(length(B)))
L223<-function(pe,rho){
for(i in 1:length(B)){
if((B[i])==0){l223[i]=0} else

{l223[i]<-sum(((0:(B[i]-1))^2)/((1+(-rho/(rho-1))*(0:(B[i]-1)))^2))}
}
return(sum(l223))}

der22<-function(pe,rho){
```

```
der22<-L221(pe,rho)+L222(pe,rho)+L223(pe,rho)
return(der22)}

MM<-function(pe,rho){
matrix(c(der11(pe,rho),der12(pe,rho),der12(pe,rho),der22(pe,rho)),2,2)
}

LL<-function(pe,rho){
LL<-sum(log(choose(B[1:k],A[1:k])))+L1(pe,rho)+L2(pe,rho)-L3(pe,rho)
return(LL)
}

Max<- function(x) L(x[1], x[2])
opt<-optim(c(0.01,0.01), Max,lower = 0.001, upper = 0.999,

method = "L-BFGS-B",control=list(fnscale=-1))

pmodel<-opt$par[1]
rho<-opt$par[2]
nI<-(-1)*MM(pmodel,rho)
varp=(solve(nI))[1,1]
beta1=log(pmodel/(1-pmodel))
g2=(exp(2*beta1))/((1+exp(beta1))^4)
varb1=varp/g2
se=sqrt(varb1)
prob[l,j]=pmodel
ro[l,j]=rho
vp[l,j]=varp
vb1[l,j]=varb1

##############################
#Tarone statistic:
S=sum((A-p*B)^2)/(p*(1-p))
```

```
Zvalue[l,j]=(S-sum(B))/sqrt(2*sum(B*(B-1)))
pvalue[l,j]=1-pnorm(Zvalue[l,j])
########################################

low[l,j]=log(pmodel/(1-pmodel))-se*qnorm(.95)
}

tarone[j]=pvalue[ki-1,j]
sebench[j]=(log(prob[ki-1,j]/(1-prob[ki-1,j]))-low[ki-1,j])/qnorm(.95)
}
############################
########################
gg=vector(length=M)
for (i in 1:M){
gg[i]=sum(vb1[,i]<=0)
}
ii=which(gg>0)#trials con errores

yy=which(vb1[ki-1,]<0)
ll<-M-length(yy)
r=ro[ki-1,-yy][-(1001:ll)]
mean(r)
sd(r)
p=prob[ki-1,-yy][-(1001:ll)]
mean(p)
sd(p)

yy=which(vb1[2*ki-1,]<0)
ll<-M-length(yy)
r=ro[2*ki-1,-yy][-(1001:ll)]
mean(r)
sd(r)
p=prob[2*ki-1,-yy][-(1001:ll)]
```

```
mean(p)
sd(p)

yy=which(vb1[ki/2-1,]<0)
ll<-M-length(yy)
r=ro[ki/2-1,-yy][-(1001:ll)]
mean(r)
sd(r)
p=prob[ki/2-1,-yy][-(1001:ll)]
mean(p)
sd(p)

###############end simulation##############################

#########################################################################
#FDR and POWER:
#########################################################################
N1=vector(length=M)#p-values of false nulls <=g
I1=vector(length=M)
I2=vector(length=M)
I3=vector(length=M)
Ic=vector(length=M)
Ib=vector(length=M)
Io=vector(length=M)
Iu=vector(length=M)
Ia1=vector(length=M)
Ia1=vector(length=M)

for (j in 1:M) {

tlow=exp(low[,j])/(1+exp(low[,j]))-.05
d=dm[j,]
#d=rep(1,s)  #activate this when there are no effects
```

```
r=rank(x[j,])
k[j]=length(x[j,][x[j,]<=g])

#SGoF
n[j]=max(k[j]-ka+1,0)
v[j]=length(r[r<=n[j]&d==1])
n1[j]=max(k[j]-ka+1,1,na.rm = T)
if (sum(d)<s) power[j]=(n[j]-v[j])/(s-sum(d))

#normal SGoF
nz[j]=round(max(k[j]-s*g-qnorm(1-a)*sqrt(s*Fng[j]*(1-Fng[j]))+1,0,na.rm = T))
vz[j]=length(r[r<=nz[j]&d==1])
n1z[j]=round(max(k[j]-s*g-qnorm(1-a)*sqrt(s*Fng[j]*(1-Fng[j]))+1,1))
if (sum(d)<s) powerz[j]=(nz[j]-vz[j])/(s-sum(d))

#bh
#nbh[j]=round(max(c(r[x[j,]<=(r/sum(d))*abh],0),na.rm = T) ) #BH ptimo
#activamos el BH original
nbh[j]=round(max(c(r[x[j,]<=(r/s)*abh],0),na.rm = T) )
vbh[j]=length(r[r<=nbh[j]&d==1])
if (sum(d)<s) powerbh[j]=(nbh[j]-vbh[j])/(s-sum(d))

#conservative SGoF
w=s*Fng[j]
Nindep0[j]=round(max(sum(w)-s*.05-sqrt(0.05*(1-0.05)*s)*qnorm(.95)+1,0,na.rm = T))
vc[j]=length(r[r<=Nindep0[j]&d==1])
if (sum(d)<s) powerc[j]=(Nindep0[j]-vc[j])/(s-sum(d))

#BB-SGoF Benchmark
Ndepbench[j]=round(max(s*tlow[ki-1],0,na.rm = T))
vb[j]=length(r[r<=Ndepbench[j]&d==1])
if (sum(d)<s) powerb[j]=(Ndepbench[j]-vb[j])/(s-sum(d))
```

```
#BB-SGoF underestimation
nblocks2[j]=ki/2
Ndepauto2[j]=round(max(s*tlow[nblocks2[j]-1],0,na.rm = T))
vu[j]=length(r[r<=Ndepauto2[j]&d==1])
if (sum(d)<s) poweru[j]=(Ndepauto2[j]-vu[j])/(s-sum(d))

#BB-SGoF overestimation
nblocks3[j]=2*ki
Ndepauto3[j]=round(max(s*tlow[nblocks3[j]-1],0,na.rm = T))
vo[j]=length(r[r<=Ndepauto3[j]&d==1])
if (sum(d)<s) powero[j]=(Ndepauto3[j]-vo[j])/(s-sum(d))

#automatic

Mini=min(tlow,na.rm = T)
Aux2=min(which(tlow==Mini))
nblocks51[j]=Aux2
Ndepauto51[j]=round(max(s*tlow[nblocks51[j]],0,na.rm = T))
va1[j]=length(r[r<=Ndepauto51[j]&d==1])
if (sum(d)<s) powera1[j]=(Ndepauto51[j]-va1[j])/(s-sum(d))
}

for (j in 1:M){
#coverage
N1[j]<-sum(x[j,]<=g&dm[j,]==0)
I1[j]<-sum(n[j]<=N1[j])
I2[j]<-sum(nz[j]<=N1[j])
I3[j]<-sum(nbh[j]<=N1[j])
Ic[j]<-sum(Nindep0[j]<=N1[j])
Ib[j]<-sum(Ndepbench[j]<=N1[j])
Iu[j]<-sum(Ndepauto2[j]<=N1[j])
Io[j]<-sum(Ndepauto3[j]<=N1[j])
Ia1[j]<-sum(Ndepauto51[j]<=N1[j])
```

```
}

ll<-M-length(ii)

fdr=v[-ii][-(1001:ll)]/n1[-ii][-(1001:ll)]
afdr=mean(fdr)
apower=mean(power[-ii][-(1001:ll)])

fdrz=vz[-ii][-(1001:ll)]/n1z[-ii][-(1001:ll)]
afdrz=mean(fdrz)
apowerz=mean(powerz[-ii][-(1001:ll)])

fdrbh=rep(0,M)
fdrbh[nbh>0]=vbh[nbh>0]/nbh[nbh>0]
afdrbh=mean(fdrbh[-ii][-(1001:ll)])
apowerbh=mean(powerbh[-ii][-(1001:ll)])

fdrc=rep(0,M)
fdrc[Nindep0>0]=vc[Nindep0>0]/Nindep0[Nindep0>0]
afdrc=mean(fdrc[-ii][-(1001:ll)])
apowerc=mean(powerc[-ii][-(1001:ll)])

fdrb=rep(0,M)
fdrb[Ndepbench>0]=vb[Ndepbench>0]/Ndepbench[Ndepbench>0]
afdrb=mean(fdrb[-ii][-(1001:ll)])
apowerb=mean(powerb[-ii][-(1001:ll)])

fdru=rep(0,M)
fdru[Ndepauto2>0]=vu[Ndepauto2>0]/Ndepauto2[Ndepauto2>0]
afdru=mean(fdru[-ii][-(1001:ll)])
apoweru=mean(poweru[-ii][-(1001:ll)])

fdro=rep(0,M)
```

```
fdro[Ndepauto3>0]=vo[Ndepauto3>0]/Ndepauto3[Ndepauto3>0]
afdro=mean(fdro[-ii][-(1001:ll)])
apowero=mean(powero[-ii][-(1001:ll)])

fdra1=rep(0,M)
fdra1[Ndepauto51>0]=va1[Ndepauto51>0]/Ndepauto51[Ndepauto51>0]
afdra1=mean(fdra1[-ii][-(1001:ll)])
apowera1=mean(powera1[-ii][-(1001:ll)])


afdr;afdrz;afdrbh;afdrc;afdrb;afdru;afdro;afdra1;

apower;apowerz;apowerbh;apowerc;apowerb;apoweru;apowero;apowera1;

mean(I1[-ii][-(1001:ll)]);mean(I2[-ii][-(1001:ll)]);mean(I3[-ii][-(1001:ll)]);

mean(Ic[-ii][-(1001:ll)]);mean(Ib[-ii][-(1001:ll)]);mean(Iu[-ii][-(1001:ll)]);

mean(Io[-ii][-(1001:ll)]);mean(Ia1[-ii][-(1001:ll)]);mean(N1[-ii][-(1001:ll)])

###parameters of the simulation:
M  #trials
s;s0;pio  #tests, true nulls (random, last trial), average proportion of true nulls
mu2;mu22;mean(Fng)   #alternative means and mean proportion of p-values <=gamma
ki   #number of blocks
mean(nblocks51)
sd(nblocks51)
mean(nblocks52)
sd(nblocks52)
#plot(colMeans(dm))   #average proportion of true nulls for each test
#plot(rowMeans(dm))  #proportion of true nulls for each trial
M-length(ii)
sum(ii<=1000)
```

Acknowledgements

Financial support from the Grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science and Innovation is acknowledged. This work was also supported by Agrupamento INBIOMED, 2012/273, from DXPCTSUG-FEDER 'Unha maneira de facer Europa'.

Bibliography

[1] 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), Vol. 57, No. 1, 289-300.

[2] Benjamini Y and Yekutieli D (2001). *The Control of the False Discovery Rate in multiple testing under dependence.* The Annals of Statistics, Vol. 29, No. 4, 1165-1188.

[3] Carvajal-Rodríguez A, De Uña-Álvarez J and Rolán-Álvarez E (2009). *A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests.* BMC Bioinformatics 10:209.

[4] Castro Conde I, de Uña-Álvarez J (2013) *Performance of Beta-Binomial SGoF multitesting method for dependent gene expression levels: a simulation study.* Proceedings of BIOINFORMATICS 2013 International Conference on Bioinformatics Models, Methods and Algorithms (Pedro Fernandes, Jordi Solé-Casals, Ana Fred and Hugo Gamboa Eds.), SciTePress, to appear.

[5] Chatfield C and Goodhart GJ (1970). *The beta-binomial model for consumer purchasing behaviour.* Appl. Statist., 19, 240-50.

[6] De Uña-Álvarez J (2011). *On the statistical properties of SGoF multitesting method.* Statistical Applications in Genetics and Molecular Biology Vol. 10, Iss. 1, Article 18.

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

[8] Dudoit S and Van der Laan MJ (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

[9] Griffiths DA (1973). *Maximum likelihood estimation for the beta-binomial distribution and a application to the household distribution of the total number of cases of a disease*. Biometrics, 29, 637-48.

[10] 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.

[11] Hedenfalk data. Library *qvalue* of the software *R*.

[12] Hochberg Y (1988). *A Sharper Bonferroni Procedure for Multiple Tests of Significance*. Biometrika 75 (4): 800-802.

[13] Hommel G (1988). *A stagewise rejective multiple test procedure based on a modified Bonferroni test*. Biometrika 75 (2): 383.

[14] Johnson NL, Kotz S and Balakrishnan N (1970). *Distributions in statistics, continuous univariate distributions-2*. Houghton Mifflin, Boston.

[15] Kemp CD and Kemp AW (1956). *The analysis of point quadrat data*. Aust. J. Bot. 4, 167-74.

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

[17] Owen A (2005). *Variance of the number of flase discoveries*. Journal of the Royal Statistical Society Series B-Statistical Methodology, Vol. 67, 411-426.

[18] Pham TV, Hoffmann S, Kubbutat MH, Jimnez CR et al. (2010). *Workflow comparison for label-free, quantitative secretome proteomics for cancer biomarker discovery: method evaluation, differential analysis and verification in serum*. J. Proteome Res 9, 1913-1922.

[19] R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[20] Simes RJ (1986). *An improved Bonferroni procedure for multiple tests of significance*. Biometrika 73, 751-754.

[21] Storey JD (2003). *The positive False Discovery Rate: a Bayesian interpretation and the q-value*. The Annals of Statistics, Vol. 31, No. 6, 2013-2035.

[22] Storey JD and Tibshirani R (2003). *Statistical significance for genomewide studies*.www.pnas.org.

[23] Williams DA (1975). *The analysis of binary responses form toxicological experiments involving reproduction and teratogenicity*. Biometrics, 31, 949-52.