Universidade Vigo

SGoF multitesting method under the Bayesian paradigm

Irene Castro-Conde and Jacobo de Uña-Álvarez

Report 13/06

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

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

Universidade Vigo

SGoF multitesting method under the Bayesian paradigm

Irene Castro-Conde and Jacobo de Uña-Álvarez

Report 13/06

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: Universidade Vigo

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

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

SGoF multitesting method under the Bayesian paradigm

Irene Castro-Conde and Jacobo de Uña-Álvarez

December 2, 2013

Abstract

In this paper we consider the Sequential Goodness-of-Fit (SGoF) multitesting procedure under the Bayesian paradigm. For this, it is assumed that the proportion of p-values falling below the significance threshold follows some prior density. Credible intervals and Bayesian preliminary tests for point null hypotheses are combined to define a suitable modification of SGoF method. The performance of Bayesian SGoF is explored through simulations. One of the main conclusions of our research is that the Bayesian viewpoint is suitable to keep the large statistical power of SGoF method even in the presence of strong correlation structures. Application of the method to treatment comparison for acute myocardial infarction and to a microarray study of hereditary breast cancer are included.

1 Introduction

Multiple hypotheses testing is concerned with decision making in situations where a number of null hypotheses are under simultaneous consideration. These may represent a negligible effect of a set of covariates or risk factors in successive univariate regression analysis, the equality of mean response along a number of variables or genes in a two-sample problem, and so on. Often, the 'sampling information' is restricted to a set of n p-values $p_1, ..., p_n$ corresponding to the n nulls at hand, $H_{01}, ..., H_{0n}$ say. In this setting, several methods have been proposed to control for type I errors in a simultaneous way. For example, family-wise error rate (FWER) is defined as the probability of committing one or more than one type I errors along the tests, while the false discovery rate (FDR) is the expected proportion of type I errors among the set of rejected nulls. See Nichols and Hayasaka [1] and Dudoit and van der Laan [2] for a review of FWER- and FDR-controlling procedures, as well as for other error criteria.

Recent research has pointed up that FWER and FDR may be stringent error measures, particularly when the proportion of non-true nulls is small or when the true alternative hypotheses are close to the non-true nulls (weak effects). This means that FWER- and FDR-based methods may be unable to detect even

a single feature in special situations. See for example Carvajal-Rodríguez *et al.* [3] or de Uña-Álvarez [4]. This has motivated the appearance of alternative methods which are more liberal with respect to type I errors, while searching for an improved statistical power (e.g. Storey [5]; Cheng *et al.* [6]; Lehmann and Romano [7]). One of such procedures is the Sequential Goodness-of-Fit (SGoF) method introduced by Carvajal-Rodríguez *et al.* [3].

Given an initial significance threshold γ, SGoF searchs for significance when comparing the observed amount of p-values below γ to the expected amount if all the nulls were true (that is, $n\gamma$). The sequence of p-values is therefore transformed into a sequence $\overrightarrow{X}$ of binary outcomes $X_1 = I(p_1 \leq \gamma), ..., X_n = I(p_n \leq \gamma)$. The intersection or complete null $H_0 = \cap_{i=1}^{n} H_{0i}$ is tested through the usual Z-statistic for the binomial problem

$$Z = Z(\overrightarrow{X}) = \frac{\overline{X}_n - \gamma}{\sqrt{\gamma(1-\gamma)/n}},$$

where $\overline{X}_n$ is the sample mean of the X_i's. Large values of Z reveal that a larger than expected amount of p-values fall below the threshold γ and, consequently, H_0 should be rejected. Note that, in the presence of non-true nulls, the distribution of the p-values will be no longer uniform and its location will be shifted towards zero. Comparison of Z to the $(1-\alpha)$-th quantile of the standard normal, z_α say, leads then to a one-sided (meta)test at level α (binomial quantiles are considered instead if n is small). SGoF procedure declares as non-true nulls the ones corresponding to the $N_n(\alpha)$ smallest p-values, where $N_n(\alpha) = n(\overline{X}_n - \gamma) - \sqrt{n\gamma(1-\gamma)}z_\alpha + 1$ represents the 'excess of significant cases' in the metatest. Such a procedure controls for the FWER only weakly (that is, only under H_0) [3,4]. Besides, if some of the nulls are non-true, it guarantees a 'reasonable FDR' in the sense that the number of false positives (true nulls rejected) is maintained below the number of false negatives (non-true nulls accepted) with probability $1-\alpha$, as long as the null variance $\gamma(1-\gamma)/n$ in the metatest is replaced by its counterpart under the alternative $\overline{X}_n(1-\overline{X}_n)/n$ [8]. The relevant aspect of SGoF multitesting method is that it may allow to detect non-true nulls in difficult situations where other, more standard multitesting methods will fail. Connections of SGoF to the concept of second-level significance testing or higher criticism (cfr. Donoho and Jin [9, 10]) have been discussed [4, 8]).

In this paper, we revisit SGoF method under the Bayesian paradigm. For this, it is assumed that the probability $\theta = P(X_i = 1) = P(p_i \leq \gamma)$ follows a prior density $\pi(\theta)$ supported on the unit interval. Default choice for $\pi(\theta)$ will be the uniform $\pi(\theta) = 1$. The key for the extension of SGoF method to the Bayesian setting is the construction of a $100(1-\alpha)\%$ one-sided credible set for the 'excess of significant cases' $n(\theta - \gamma)$. For a useful summary of the posterior analysis, the posterior probability that the complete null hypothesis $H_0 : \theta = \gamma$ is true will be given. Following Berger and Delampady [11], a prior located at γ is taken instead of the uniform, while default choice for the a priori probabilities of H_0 and H_1 ($\theta \neq \gamma$) is $P_0 = P_1 = 1/2$. Differences of the Bayesian perspective

relative to the (frequentist) original SGoF method are highlighted in Section 2. Simulation studies and illustrations with real medical data are provided in Sections 3 and 4 respectively. A final discussion and the main conclusions of the paper are given in Section 5.

2 Bayesian SGoF

Sequential Goodness-of-Fit (SGoF) multitesting method starts by assuming that, under the complete null, the sequence of available p-values $p_1, ..., p_n$ is a random sample of a uniform distribution on the unit interval, $U(0,1)$. Therefore, two basic assumptions are made: (i) the distributional assumption $p_i \sim U(0,1)$, and (ii) the independence assumption. As a consequence, given an initial significance threshold γ, the transformed sequence of indicators $X_1, ..., X_n$ where $X_i = I(p_i \leq \gamma)$, $i = 1, ..., n$, is a random sample from a Bernoulli population, $Ber(\theta)$, where $\theta = P(X_i = 1) = P(p_i \leq \gamma)$ is an unknown constant parameter. Note that $\theta = \gamma$ under the complete null. If the complete null is false, $p_1, ..., p_n$ is still thought as a random sample but following a non-uniform distribution, typically shifted towards zero. The classical (frequentist) approach to test for $\theta = \gamma$ against to one-sided alternative $\theta > \gamma$ is based on the (frequentist) p-value

$$p_f = P(N(0,1) > Z(\overrightarrow{x})) = 1 - \Phi(Z(\overrightarrow{x})),$$

where $Z(\overrightarrow{x})$ is the actual value of the Z-statistic

$$Z(\overrightarrow{X}) = \frac{\overline{X}_n - \gamma}{\sqrt{\gamma(1-\gamma)/n}}$$

and Φ denotes the cdf of the standard normal, n assumed to be large enough (otherwise the cdf of a binomial distribution is used). At significance level α, the null is rejected if $p_f < \alpha$. Under the alternative, a $100(1-\alpha)\%$ (frequentist) confidence interval for θ is given by

$$I_f = (\overline{X}_n \pm \sqrt{\overline{X}_n(1-\overline{X}_n)/n} z_{\alpha/2})$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. When $\theta = \gamma$ is rejected, SGoF multitesting procedure declares as non-true the null hypotheses with the smallest $N_n(\alpha)$ p-values, where $N_n(\alpha) = n(\overline{X}_n - \gamma) - \sqrt{n\gamma(1-\gamma)} z_\alpha + 1$ represents the 'excess of significant cases' in the (one-sided) metatest; a conservative version of this procedure is given by the corrected amount of rejections $N_n^*(\alpha) = n(\overline{X}_n - \gamma) - \sqrt{n\overline{X}_n(1-\overline{X}_n)} z_\alpha + 1$, which is basically the lower limit of a frequentist $100(1-\alpha)\%$ one-sided confidence interval for $\tau_n(\theta) = n(\theta - \gamma)$. Note that $\tau_n(\theta)$ may be interpreted as the difference between the real and the expected amounts of p-values falling below the threshold γ.

In the Bayesian framework [12], some prior information on the parameter of interest θ is available. Assume that θ has a density $\pi(\theta)$ supported on the

unit interval, which represents the a priori information. In a non-informative setting, $\pi(\theta)$ will be chosen as the uniform density. In general, $\pi(\theta)$ may serve to introduce the researcher's information on the location of the parameter, with a smaller or larger dispersion according to his/her level of uncertainty. The frequentist likelihood of $\overrightarrow{X}$, $f(\overrightarrow{X}|\theta) = \theta^s(1-\theta)^{n-s}$, where $s = n\overline{X}_n$, is then updated to account for the randomness of θ, leading in its turn to the posterior density of the parameter:

$$\pi(\theta|\overrightarrow{x}) = \frac{f(\overrightarrow{x}|\theta)\pi(\theta)}{\int f(\overrightarrow{x}|\theta')\pi(\theta')d\theta'}.$$

A (Bayesian) $100(1-\alpha)\%$ credible interval for θ is then given by

$$I_c = (l_{\alpha/2}(\pi, \overrightarrow{x}), u_{\alpha/2}(\pi, \overrightarrow{x})),$$

where $l_{\alpha/2}(\pi, \overrightarrow{x})$ and $u_{\alpha/2}(\pi, \overrightarrow{x})$ are respectively the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior density $\pi(\theta|\overrightarrow{x})$. More generally, a credible set $A_c = \{\theta : \pi(\theta|\overrightarrow{x}) > k\}$ may be used, where k is chosen to satisfy $P(A_c|\overrightarrow{x}) = 1-\alpha$. Unlike the frequentist interval I_f, which focus in results when 'averaged' along the sampling distribution, the Bayesian counterpart I_c guarantees the nominal coverage $100(1-\alpha)\%$ conditional on observing data of the same 'strength of evidence' as the actual $\overrightarrow{X} = \overrightarrow{x}$. The amount of null hypotheses declared as non-true by a Bayesian analogue of frequentist SGoF is accordingly defined as the lower limit $n(l_\alpha(\pi, \overrightarrow{x}) - \gamma)$ of a $100(1-\alpha)\%$ one-sided credible interval for $\tau_n(\theta) = n(\theta - \gamma)$. We formalize this idea in the following definition.

Definition (Bayesian SGoF). Bayesian SGoF method is defined as the rule which declares as non-true the null hypotheses with the smallest $N_n^b(\alpha)$ p-values, where

$$N_n^b(\alpha) = \max(n(l_\alpha(\pi, \overrightarrow{x}) - \gamma), 0).\square$$

Needless to say, the rule given by Bayesian SGoF must be interpreted under the Bayesian paradigm. By using this method, the researcher is ensuring that, conditionally on $N_n^b(\alpha) > 0$, $100(1-\alpha)\%$ of the times the sample evidence is that of the actual $\overrightarrow{x}$, the difference between the real and expected amounts of p-values below γ ($\tau_n(\theta) = n(\theta - \gamma)$) is at least as large as $N_n^b(\alpha)$. In this sense, Bayesian SGoF is rejecting a 'reasonable' amount of nulls. For illustration, in the next example we consider the situation in which θ follows a beta distribution.

Example (beta prior). Assume $\pi(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$ for some $a, b > 0$, where $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$. That is, θ follows a beta distribution, $\beta(a,b)$. In this case, the X_i's are correlated binary outcomes with mean $E(\theta) = p = a/(a+b)$ and pairwise correlation $\rho = 1/(a+b+1)$ [8]. Note that the non-informative prior $\pi(\theta) = 1$ is just the special case $a = b = 1$. Straightforward calculations give that the posterior distribution of θ is $\beta(a+s, b+n-s)$, where $s = n\overline{X}_n$. A Bayesian point estimate of θ is typically given by using some location parameter; for example, the mean of the posterior distribution is

$$E(\theta|\overrightarrow{x}) = \frac{a+s}{a+b+n}.$$

Note that $E(\theta|\overrightarrow{x}) \approx \overline{X}_n$ as $n \to \infty$, which reflect the well-known predominance of the sampling information on the prior as n grows. We have

$$l_\alpha(\pi, \overrightarrow{x}) = \Psi^{-1}(\alpha; a, b, \overrightarrow{x})$$

where $\Psi(.; a, b, \overrightarrow{x})$ denotes the cdf of a $\beta(a+s, b+n-s)$ random variable. When $a+s$ and $b+n-s$ are large (this is true in particular when n is large), the beta-normal approximation yields

$$l_\alpha(\pi, \overrightarrow{x}) \approx E(\theta|\overrightarrow{x}) - \sqrt{V(\theta|\overrightarrow{x})} z_\alpha$$

where $V(\theta|\overrightarrow{x})$ denotes the posterior variance, namely

$$V(\theta|\overrightarrow{x}) = \frac{(a+s)(b+n-s)}{(a+b+n)^2(a+b+n+1)}.$$

From this it is immediately seen that $n(l_\alpha(\pi, \overrightarrow{x}) - \gamma) \approx N_n^*(\alpha)$ as $n \to \infty$ and, therefore, Bayesian SGoF approaches its frequentist counterpart as the sample size increases. For small to moderate n however, the two approaches will give different answers.□

It is evident from the definition of $N_n^b(\alpha)$ that $N_n^b(\alpha) = 0$ unless $l_\alpha(\pi, \overrightarrow{x}) > \gamma$. In other words, if $l_\alpha(\pi, \overrightarrow{x}) \leq \gamma$, then Bayesian SGoF will accept as true all the null hypotheses under consideration, thus leading to the acceptance of the complete null H_0. Therefore, the location of the null value $\tau_n(\theta) = 0$ relative to the lower bound of the credible interval determines if H_0 is rejected or not. While the relationship between hypothesis testing and confidence sets is well established in the frequentist setting, Bayesian testing of point null hypothesis (like $H_0 : \theta = \gamma$) is not performed on the basis of the construction of credible sets. As quoted by Berger and Delampady [11], "Only by calculating a Bayes factor (or related conditional measure) can one judge how well the data supports a distinguished point θ_0". This makes an important difference between frequentist and Bayesian conceptions of SGoF method, and motivates the introduction of a pre-test (Bayesian) procedure which may complement the information reported by $N_n^b(\alpha)$.

To be precise, consider the (Bayesian) problem of testing $H_0 : \theta = \gamma$ against the alternative $H_1 : \theta \neq \gamma$. Take the usual default prior probabilities for H_0 and H_1, these are $P_0 = P_1 = 1/2$. As prior distribution of θ under the alternative, Berger and Delampady [11] suggested in their Section 3.2.4 the class of conjugate $\pi(\theta)$ with mean γ, among other 'objective' possibilities. In particular, for the binomial distribution, the beta model is a family of conjugate distributions. So take $\pi(\theta) \sim \beta(a, b)$ where $a = (1-\rho)\gamma/\rho$ and $b = (1-\rho)(1-\gamma)/\rho$, ρ to be precise later. In general, the posterior probability that H_0 is true is given by

$$\begin{aligned} P(H_0|\overrightarrow{x}) &= \frac{P_0 f(\overrightarrow{x}|\gamma)}{P_0 f(\overrightarrow{x}|\gamma) + P_1 \int_{\theta \neq \gamma} f(\overrightarrow{x}|\theta)\pi(\theta)d\theta} \\ &= \left[1 + \frac{1-P_0}{P_0} \frac{1}{B(\overrightarrow{x})}\right]^{-1} \end{aligned}$$

where

$$B(\overrightarrow{x}) = \frac{f(\overrightarrow{x}|\gamma)}{\int_{\theta \neq \gamma} f(\overrightarrow{x}|\theta)\pi(\theta)d\theta}$$

is the Bayes factor, which is interpreted as the ratio between the likelihood of the data under the null and the average likelihood of the data under the alternative. For the beta model this becomes

$$B(\overrightarrow{x}) = \gamma^s(1-\gamma)^{n-s}B(a,b)/B(a+s,b+n-s)$$

where n is the number of p-values and s the amount of them falling below γ. Therefore, under the default $P_0 = P_1 = 1/2$, one gets

$$P(H_0|\overrightarrow{x}) = \left[1 + \gamma^{-s}(1-\gamma)^{s-n}B(a+s,b+n-s)/B(a,b)\right]^{-1}.$$

The posterior probability $P(H_0|\overrightarrow{x})$ has been proposed as the suitable way in which evidence against the null should be looked for [11,13]. Its advantages when compared to the classical (frequentist) p-value p_f has been widely discussed (same references). In particular, it has been pointed up that $P(H_0|\overrightarrow{x})$ may be regarded as a frequentist type I error probability, conditional on observing data of the same 'strength of evidence' as the actual $\overrightarrow{x}$. Even when $P(H_0|\overrightarrow{x})$ heavily depends on the value of ρ (indeed, $P(H_0|\overrightarrow{x}) \to 1$ as $\rho \to 1$ regardless the data at hand, something known as Jeffreys's paradox), lower bounds are available and may provide useful guidance in practice. In Table 1, we report for the case $\gamma = 0.05$ and $n = 15$, the lower bounds $\underline{P}(H_0|\overrightarrow{x}) = \inf_\rho P(H_0|\overrightarrow{x})$ depending on the value of s, together with the corresponding frequentist (one-sided) p-values. Even when minimizing the posterior probability of H_0, it becomes clear that the Bayesian perspective may be much more conservative than the classical approach when looking for evidence against the complete null.

Table 1: Lower bounds along ρ for the posterior probability of $H_0 : \theta = \gamma$ for the beta model. Case $\gamma = 0.05, n = 15$.

s	p_f	$\underline{P}(H_0\|\overrightarrow{x})$	ρ
2	.1710	.5000	.0000
3	.0362	.3941	.0652
4	.0055	.1631	.1344
5	.0006	.0344	.1990
6	5.3×10^{-5}	.0044	.2600
9	7.4×10^{-9}	1.2×10^{-6}	.4281

A possible approach to combine a Bayesian pre-test with the computation of $N_n^b(\alpha)$ is to define the pre-test Bayesian SGoF procedure as that rejecting the $N_n^{b*}(\alpha)$ nulls with the smallest p-values, where $N_n^{b*}(\alpha) = I(s \geq s_\alpha)N_n^b(\alpha)$ and $s_\alpha - 1$ is the first value of s when going from n to 0 for which $\underline{P}(H_0|\overrightarrow{x}) \geq \alpha$. In the example of Table 1, for $\alpha = 0.05$ we would have $s_\alpha = 5$. Practical performance of basic Bayesian SGoF ($N_n^b(\alpha)$) and its pre-test version ($N_n^{b*}(\alpha)$) is investigated through simulations in the next Section.

3 Simulation studies

We have designed two simulation studies in order to investigate the performance of Bayesian SGoF. The first simulation concentrates in a model in which (under the alternative) the probability $\theta = P(X_i = 1) = P(p_i \leq \gamma)$ is drawn from the uniform density $\pi(\theta) = 1$ and, therefore, it perfectly fits a Bayesian scenario. The second simulation reproduces the application of a two-sample test along a number of positions (or 'genes'), which results in a sequence of possibly correlated p-values. While the first scenario is suitable for a better understanding of the properties of Bayesian SGoF, the second one allows for the study of the method's performance in a more practical setting.

3.1 Bayesian scenario

For fixed values of γ and n, we simulate data as follows:

Step 1. Draw θ_1 from the uniform density $\pi(\theta) = 1$.

Step 2. Draw independently Y from $Ber(1/2)$, that is, $P(Y = 1) = P(Y = 0) = 1/2$.

Step 3. Compute $\theta = Y\gamma + (1 - Y)\theta_1$.

Step 4. Draw s from a $Bin(n, \theta)$ distribution.

In this model, the complete null $H_0 : \theta = \gamma$ is true with probability $1/2$, while the parameter of interest is uniformly distributed on the unit interval under the alternative. This corresponds to a situation in which the non-informative prior perfectly fits the data when the complete null is violated, while the a priori probability that H_0 is true is $P_0 = 1/2$. In Step 4, the 'data' $\overrightarrow{x}$ are obtained; note that the relevant information for SGoF method is contained in the number of p-values falling below the significance threshold γ, and this can be generated from a binomial distribution. Therefore, in this simulated setting p-values coming from true and non-true nulls are not distinguished; we only know that, when $Y = 1$, all the nulls are true, while some proportion of non-true nulls are present when $Y = 0$. We take $\gamma = 0.05$ and three different sample sizes, $n = 15, 50, 500$. We repeat Steps 1-4 up to get 10,000 simulations.

For $\alpha = 0.05$ we compute the number of rejections provided by the basic Bayesian SGoF and by the pre-test Bayesian SGoF. Therefore, computation of $N_n^b(\alpha) = \max(n(l_\alpha(\pi, \overrightarrow{x}) - \gamma), 0)$ and $N_n^{b*}(\alpha) = I(s \geq s_\alpha)N_n^b(\alpha)$ is done, where $l_\alpha(\pi, \overrightarrow{x})$ is the α-quantile of a $\beta(1 + s, 1 + n - s)$ distribution, and $s_\alpha = 5, 9, 42$ for $n = 15, 50, 500$ respectively. We also compute the posterior probability that the complete null is true, $P(H_0|\overrightarrow{x})$ for each simulation, by using the a priori probabilities $P_0 = P_1 = 1/2$ and the true prior $\pi(\theta) = 1$. For comparison purposes, computation of original (fequentist) SGoF is done too; since the sample size is not always large, here we use the exact formula for the number of rejections, namely $N_n(\alpha) = s - b_{n,\alpha}(\gamma) + 1$, where $b_{n,\alpha}(\gamma) = \inf\{b \in \{0, ..., n\} : P(Bin(n, \gamma) \geq \alpha)\}$ is the $(1 - \alpha)$-quantile of the $Bin(n, \gamma)$ model.

In Table 2 we report, for the three methods and the three sample sizes $n = 15, 50, 500$, the following values. (a) The average number of rejections (Mean)

among the trials with $Y = 0$, that is, under the alternative, and the standard deviation (SD) of the number of rejections. (b) The proportion of times the complete null is rejected among the trials with $Y = 1$, that is, under the complete null; this is just the FWER of each method. (c) The proportion of times the complete null is rejected among the trials with $Y = 0$, thus corresponding to the power (POW) of each method to detect the presence of non-true nulls. And (d) among the trials for which the complete null is correctly rejected (i.e. $Y = 0$ and $N_n^{any}(\alpha) > 0$), the proportion of times the number of rejections is smaller than $\tau_n(\theta) = n(\theta - \gamma)$; this is labelled as COV (from coverage) in Table 2.

Table 2: Results of frequentist SGoF, basic Bayesian SGoF and pre-test Bayesian SGoF along 10,000 Monte Carlo trials (n is the number of tests)

	Mean	SD	FWER	POW	COV
			$n = 15$		
$N_n(\alpha)$	5.64	4.36	.0356	.8117	.7628
$N_n^b(\alpha)$	4.42	3.76	.1718	.8754	.9488
$N_n^{b,*}(\alpha)$	4.29	3.89	.0008	.6761	.9467
			$n = 50$		
$N_n(\alpha)$	20.33	14.46	.0422	.8756	.7966
$N_n^b(\alpha)$	18.24	13.86	.1088	.8988	.9519
$N_n^{b,*}(\alpha)$	18.14	13.98	.0012	.8127	.9520
			$n = 500$		
$N_n(\alpha)$	217.68	142.39	.0459	.9294	.8054
$N_n^b(\alpha)$	211.02	141.84	.0656	.9306	.9505
$N_n^{b,*}(\alpha)$	210.94	141.95	.0010	.9123	.9508

From Table 2 the following features are appreciated. The FWER is controlled at level $\alpha = 0.05$ by frequentist SGoF, while basic Bayesian SGoF is anticonservative and the pre-test Bayesian SGoF is too conservative. That frequentist SGoF controls for FWER under the complete null was expected, since this is one of its well-established properties [3]. The anticonservativeness of basic Bayesian SGoF comes from the fact that it rejects the complete null whenever $l_\alpha(\pi, \overrightarrow{x}) > \gamma$, and no bound is imposed on the probability of this event. However, $P_{H_0}(l_\alpha(\pi, \overrightarrow{x}) > \gamma)$ approaches α as n grows (FWER of $N_n^b(\alpha)$ in Table 2). This can be explained from the beta-normal approximation: one has $l_\alpha(\pi, \overrightarrow{x}) \approx E(\theta|\overrightarrow{x}) - \sqrt{V(\theta|\overrightarrow{x})}z_\alpha$ and the rejection rule becomes

$$\frac{E(\theta|\overrightarrow{x}) - \gamma}{\sqrt{V(\theta|\overrightarrow{x})}} > z_\alpha;$$

since $E(\theta|\overrightarrow{x}) \approx \overline{X}_n$ and $V(\theta|\overrightarrow{x}) \approx \overline{X}_n(1 - \overline{X}_n)/n$ as $n \to \infty$, we conclude that the FWER of Bayesian SGoF will converge to α. Regarding the pre-test Bayesian SGoF, we see in Table 2 that the FWER is very low (about 0.001 for the three sample sizes); this reflects indeed the conservativeness of the frequentist p-value p_f when compared to the (Bayesian) posterior probability of the null

(Table 1). Note that, although condition $P(H_0|\overrightarrow{x}) < \alpha$ would control at level α the FWER along the samples with the same amount of evidence as $\overrightarrow{x}$, the type I error rate becomes much smaller when taking averages along a number of Monte Carlo replicates of a given model. As complementary information, we quote that the mean value of $P(H_0|\overrightarrow{x})$ (computed from the default priors $P_0 = P_1 = 1/2$ and the non-informative $\pi(\theta) = 1$) along the replicates with $Y = 1$ was 0.82 ($n = 15$), 0.87 ($n = 50$) and 0.95 ($n = 500$), with corresponding standard deviations of 0.12, 0.10, and 0.07.

Regarding the statistical power to (correctly) reject the complete null, it is seen in Table 2 that Bayesian SGoF is more powerful than frequentist SGoF in all the situations (in agreement with its larger FWER), but both methods become comparable as n grows (87.5% vs. 81.2% for $n = 15$, 93.1% vs. 92.9% for $n = 500$). This is not surprising at all; note (again) that the a priori information on θ is negligible as the sampling information grows. The pre-test Bayesian method exhibits a poor power for $n = 15$ (67.6%); however, its power is remarkably large as n grows despite its low FWER (91.2% for $n = 500$). This is because, when n is large and the complete null is false, the pre-test Bayesian method coincides with the basic Bayesian most of the times (only 77% of the times for $n = 15$, but 90% and 98% of the times for $n = 50$ and $n = 500$ respectively). Here we also mention that the mean value of $P(H_0|\overrightarrow{x})$ (computed once more from the default priors $P_0 = P_1 = 1/2$ and the uniform $\pi(\theta)$) along the replicates with $Y = 0$ was 0.18 ($n = 15$), 0.13 ($n = 50$) and 0.05 ($n = 500$), the standard deviations being 0.31, 0.29, and 0.20 respectively.

Interestingly, it is seen from Table 2 that, despite basic Bayesian detects signal more frequently than frequentist SGoF, the number of effects declared by the classical methods is larger on average. This finding is confirmed in the simulation study performed in Section 4.2. In this sense, one may say that Bayesian viewpoint is more conservative, since it will typically lead to a smaller number of declared features.

The coverages (COV) in Table 2 are defined, as mentioned, as the proportion of times the number of rejections is smaller than $\tau_n(\theta) = n(\theta - \gamma)$ among the trials for which the complete null is correctly rejected. For the basic Bayesian SGoF this is exactly $1 - \alpha$ and, therefore, the figures in Table 2 are roughly of 95%. The pre-test Bayesian SGoF method preserves this property, which means that, when $Y = 0$ and $N_n^b(\alpha) > 0$, the event $s < s_\alpha$ implies that $\tau_n(\theta) > 0$. On the contrary, frequentist SGoF reports coverages systematically below 95% (between 76% and 81% indeed). This is somehow corrected when using its conservative version $N_n^*(\alpha)$, see Section 2, which replaces the term $\gamma(1 - \gamma)$ by $\overline{X}_n(1-\overline{X}_n)$ in the variance (e.g. 93% of coverage for $n = 500$, but below 91% for $n \leq 50$, results not shown); but, in any case, frequentist SGoF is not prepared to cope with the correlation among the X_i's induced by the randomness of θ, so it is not surprising that it behaves in a anticonservative way in the sense of COV. Inspection of the number of rejections (Mean) in Table 2 supports this finding too.

3.2 Two-sample tests scenario

We have designed a simulated scenario similar to the study of Hedenfalk [14], where the mean expression levels of a large number of genes in two different groups A and B of individuals (with sample sizes of 7 and 8) were compared, see Section 4.2. In order to study the influence of the number of null hypotheses in the performance of the multitesting procedures, we considered the cases $n = 10$, $n = 50$, and $n = 500$ tests. Hedenfalk's sample sizes of 7 and 8 were taken for groups A and B respectively. The samples were drawn from n-variate Gaussian populations with different correlation degrees. The 2-sample t-test was applied to test for each null hypothesis of equality of means; the sequence of n p-values is thus coming from the computation of two-sided tails of the Student's t distribution with 13 degrees of freedom. To summarize numerical results, 1000 Monte Carlo trials were performed. The proportion of true nulls (i.e. 'genes

Table 3: Complete null hypothesis: $\Pi_0 = 1$

		$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$
$n = 10$	FDR Binomial SGoF	0.006	0.027	0.056
	FDR Bayesian SGoF	0.082	0.088	0.092
	FDR Bayesian* SGoF	0	0	0.031
	$s \geq s_\alpha$	0	0	0.031
	Posterior	0.8013(0.1234)	0.7947(0.1544)	0.7982(0.2047)
$n = 50$	FDR Binomial SGoF	0.037	0.079	0.117
	FDR Bayesian SGoF	0.1	0.121	0.129
	FDR Bayesian* SGoF	0.001	0.016	0.078
	$s \geq s_\alpha$	0.001	0.016	0.078
	Posterior	0.8788(0.0938)	0.8457(0.1732)	0.7401(0.2398)
$n = 500$	FDR Binomial SGoF	0.038	0.185	0.16
	FDR Bayesian SGoF	0.064	0.192	0.163
	FDR Bayesian* SGoF	0	0.12	0.141
	$s \geq s_\alpha$	0	0.12	0.141
	Posterior	0.9493(0.0622)	0.7079(0.3559)	0.0925(0.2603)

equally expressed') Π_0 was 1 (complete null), 0.9 (10% of effects), 0.7 (30% of effects) or 0.5 (50% 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. We simulated $k = 1$ block of n correlated p-values with correlation levels of $\rho = 0, 0.2$ and 0.8, where $\rho = 0$ means independence and $\rho = 0.8$ indicates strong correlation. For random generation, the function `rmvnorm` of the `R` software [15] was used. For each situation, we computed the FDR, the

power (defined as the proportion of non-true nulls which are rejected, labelled as POW in Tables below), and the coverage (COV), defined here as 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 [8]). Computation of these quantities for the Binomial SGoF method for independent tests and for the basic and the pre-test Bayesian methods are included. We always take $\alpha = \gamma = 0.05$. We also computed the proportion of trials for which $s \geq s_\alpha$ and $N_n^b(\alpha) > 0$ occurred (this proportion is labelled as $s \geq s_\alpha$ in Tables); note that these are the trials for which both the basic Bayesian and the pre-test method reject the complete null. For samples sizes $n = 10, 50$ and 500, $\alpha = \gamma = 0.05$, values of s_α are given by 5,9 and 42, respectively. As complementary information, we calculated the mean and standard deviation of the posterior probability $P(H_0|\overrightarrow{x})$, for de default priors $P_0 = P_1 = 1/2$ and the non-informative $\pi(\theta) = 1$.

Table 4: Proportion of true nulls $\Pi_0 = 0.9$

			$\rho = 0$			$\rho = 0.2$			$\rho = 0.8$		
			FDR	POW	COV	FDR	POW	COV	FDR	POW	COV
$n = 10$	$\mu = 1$	Binomial SGoF	0.0215	0.3632	0.993	0.025	0.3774	0.984	0.0657	0.3881	0.928
		Bayesian SGoF	0.022	0.3613	0.993	0.023	0.3740	0.989	0.0628	0.3845	0.932
		Bayesian* SGoF	0	0.347	1	0.0013	0.002	0.998	0.0207	0.018	0.975
	$\mu = 2$	Binomial SGoF	0.0102	0.4400	0.992	0.014	0.4359	0.984	0.0596	0.4294	0.928
		Bayesian SGoF	0.009	0.4276	0.996	0.0105	0.4241	0.991	0.0555	0.4202	0.935
		Bayesian* SGoF	0.0003	0.001	1	0	0	1	0.029	0.0287	0.966
	$\mu = 4$	Binomial SGoF	0.0053	0.451	0.992	0.0115	0.4497	0.984	0.0554	0.4501	0.928
		Bayesian SGoF	0.0035	0.4373	0.996	0.008	0.4361	0.991	0.0493	0.4411	0.935
		Bayesian* SGoF	0	0.001	1	0.002	0.3615	0.997	0.0253	0.037	0.964
$n = 50$	$\mu = 1$	Binomial SGoF	0.0875	0.0616	0.98	0.112	0.0722	0.937	0.0781	0.0851	0.894
		Bayesian SGoF	0.087	0.0563	0.989	0.1079	0.0663	0.956	0.0697	0.082	0.904
		Bayesian* SGoF	0.0152	0.0102	0.998	0.0297	0.0244	0.972	0.0719	0.069	0.904
	$\mu = 2$	Binomial SGoF	0.0385	0.3502	0.976	0.0618	0.3463	0.933	0.0618	0.2117	0.916
		Bayesian SGoF	0.0266	0.2958	0.994	0.048	0.297	0.966	0.0533	0.1914	0.924
		Bayesian* SGoF	0.0137	0.1499	0.996	0.0081	0.1858	0.968	0.0594	0.1076	0.913
	$\mu = 4$	Binomial SGoF	0.0071	0.4266	0.975	0.0188	0.3918	0.946	0.0581	0.2678	0.897
		Bayesian SGoF	0.0025	0.3214	0.994	0.0096	0.3282	0.971	0.0476	0.2424	0.912
		Bayesian* SGoF	0.0012	0.1799	0.996	0.0096	0.1752	0.971	0.0471	0.1542	0.913
$n = 500$	$\mu = 1$	Binomial SGoF	0.2784	0.1483	0.989	0.2181	0.135	0.874	0.1144	0.1231	0.837
		Bayesian SGoF	0.2646	0.1368	0.997	0.2092	0.1254	0.887	0.1096	0.1191	0.844
		Bayesian* SGoF	0.1946	0.1115	0.997	0.1598	0.1058	0.887	0.1086	0.1121	0.844
	$\mu = 2$	Binomial SGoF	0.0683	0.6744	0.992	0.0789	0.6411	0.857	0.0848	0.4221	0.859
		Bayesian SGoF	0.0529	0.6189	0.999	0.0663	0.5894	0.892	0.0802	0.3893	0.866
		Bayesian* SGoF	0.0529	0.6189	0.999	0.0663	0.5894	0.892	0.0802	0.3796	0.866
	$\mu = 4$	Binomial SGoF	0.0007	0.7838	0.991	0.0302	0.7443	0.861	0.0842	0.4966	0.835
		Bayesian SGoF	0.0001	0.7086	1	0.0211	0.6835	0.905	0.0776	0.4622	0.841
		Bayesian* SGoF	0.0001	0.7083	1	0.0211	0.6832	0.905	0.0776	0.4537	0.841

Table 5: Proportion of true nulls $\Pi_0 = 0.7$

			$\rho = 0$			$\rho = 0.2$			$\rho = 0.8$		
			FDR	POW	COV	FDR	POW	COV	FDR	POW	COV
$n = 10$	$\mu = 1$	Binomial SGoF	0.0252	0.1109	0.997	0.0355	0.1144	0.989	0.0382	0.1349	0.943
		Bayesian SGoF	0.0215	0.0965	1	0.034	0.0995	0.995	0.0296	0.1185	0.963
		Bayesian* SGoF	0.0007	0.003	1	0.002	0.0055	0.998	0.0131	0.0088	0.982
	$\mu = 2$	Binomial SGoF	0.0095	0.3869	0.997	0.0183	0.3832	0.989	0.038	0.3643	0.943
		Bayesian SGoF	0.0048	0.2896	1	0.0121	0.29	0.995	0.0277	0.2826	0.967
		Bayesian* SGoF	0.0018	0.131	1	0.0014	0.067	0.999	0.0156	0.0752	0.977
	$\mu = 4$	Binomial SGoF	0.0018	0.4335	0.997	0.005	0.4306	0.989	0.0256	0.412	0.943
		Bayesian SGoF	0.001	0.3149	0.999	0.0025	0.3234	0.996	0.0158	0.3237	0.967
		Bayesian* SGoF	0	0.077	1	0	0.078	1	0.0168	0.107	0.965
$n = 50$	$\mu = 1$	Binomial SGoF	0.1097	0.188	0.989	0.1078	0.1909	0.959	0.0626	0.1951	0.886
		Bayesian SGoF	0.1023	0.1499	0.999	0.0948	0.1548	0.982	0.0507	0.166	0.905
		Bayesian* SGoF	0.0578	0.1057	0.999	0.0598	0.1133	0.982	0.0497	0.1397	0.905
	$\mu = 2$	Binomial SGoF	0.0245	0.696	0.995	0.0271	0.6986	0.972	0.0391	0.6483	0.919
		Bayesian SGoF	0.0147	0.5593	1	0.0156	0.5629	0.998	0.0321	0.5231	0.94
		Bayesian* SGoF	0.0147	0.557	1	0.0151	0.5609	0.998	0.0321	0.5188	0.94
	$\mu = 4$	Binomial SGoF	0.0014	0.7792	0.988	0.0064	0.774	0.959	0.035	0.719	0.896
		Bayesian SGoF	0.0003	0.6223	1	0.0017	0.6234	0.99	0.0262	0.5896	0.929
		Bayesian* SGoF	0.0003	0.6217	1	0.0017	0.6217	0.99	0.0262	0.5868	0.929
$n = 500$	$\mu = 1$	Binomial SGoF	0.1431	0.2798	1	0.1391	0.2811	0.922	0.0557	0.2733	0.877
		Bayesian SGoF	0.1335	0.2568	1	0.1295	0.2586	0.95	0.0504	0.2541	0.89
		Bayesian* SGoF	0.1335	0.2568	1	0.1295	0.2586	0.95	0.0504	0.2521	0.89
	$\mu = 2$	Binomial SGoF	0.0353	0.8105	0.999	0.0397	0.8077	0.923	0.0399	0.7681	0.884
		Bayesian SGoF	0.0269	0.7623	1	0.0306	0.7607	0.963	0.036	0.7197	0.904
		Bayesian* SGoF	0.0269	0.7623	1	0.0306	0.7607	0.963	0.036	0.7197	0.904
	$\mu = 4$	Binomial SGoF	0.0002	0.8953	0.999	0.006	0.8886	0.931	0.0422	0.8313	0.854
		Bayesian SGoF	0.0001	0.8366	1	0.0031	0.8334	0.966	0.0372	0.7836	0.871
		Bayesian* SGoF	0.0001	0.8366	1	0.0031	0.8334	0.966	0.0372	0.7836	0.871

In Table 3 we show the results obtained in the scenario of no effects ($\Pi_0 = 1$). It should be recalled that under the complete null, all rejected hypotheses are Type I errors and therefore FDR collapses to FWER. Obviously, the power in all these situations is 100% since there are no effects. Moreover, the coverage coincides to 1-FDR as explained above. Then, in Table 3 we only report the FDR of the three different methods for every value of n and ρ, together with the proportions of trials with $s \geq s_\alpha$ and a summary (mean and standard deviation) of the posterior probabilities of the complete null.

In first place, from Table 3 we can see that under independence ($\rho = 0$) Binomial SGoF controls the FDR (and thus the FWER), the pre-test Bayesian SGoF is too conservative and the basic Bayesian SGoF reports a FDR greater than the nominal but it converges to 0.05 when the number of tests n grows. This basically mimics results in the previous simulation study (Table 2). The situation changes in the correlated settings; this is because the variance is under-

Table 6: Proportion of true nulls $\Pi_0 = 0.5$

			$\rho = 0$			$\rho = 0.2$			$\rho = 0.8$		
			FDR	POW	COV	FDR	POW	COV	FDR	POW	COV
$n = 10$	$\mu = 1$	Binomial SGoF	0.0265	0.1157	0.997	0.024	0.0942	0.997	0.0141	0.147	0.974
		Bayesian SGoF	0.0282	0.0854	0.998	0.0188	0.0811	0.998	0.0077	0.1075	0.985
		Bayesian* SGoF	0.0036	0.0231	1	0.0042	0.021	0.999	0.0067	0.0359	0.985
	$\mu = 2$	Binomial SGoF	0.0141	0.147	0.974	0.0124	0.5413	1	0.0218	0.5396	0.967
		Bayesian SGoF	0.0077	0.1075	0.985	0.0065	0.3929	1	0.0131	0.3948	0.984
		Bayesian* SGoF	0.0041	0.2567	1	0.0035	0.3538	1	0.0121	0.3128	0.984
	$\mu = 4$	Binomial SGoF	0.001	0.6112	0.997	0.0016	0.6062	0.995	0.0108	0.6013	0.962
		Bayesian SGoF	0	0.4391	1	0.0002	0.4389	0.999	0.0051	0.4441	0.982
		Bayesian* SGoF	0	0.3445	1	0.0002	0.3482	0.999	0.0051	0.4306	0.982
$n = 50$	$\mu = 1$	Binomial SGoF	0.0637	0.2591	0.999	0.0555	0.2616	0.989	0.0323	0.2775	0.917
		Bayesian SGoF	0.0549	0.2046	1	0.0457	0.208	0.997	0.0228	0.2289	0.943
		Bayesian* SGoF	0.0524	0.1956	1	0.0427	0.1972	0.997	0.0228	0.2166	0.943
	$\mu = 2$	Binomial SGoF	0.0162	0.783	0.999	0.0172	0.7777	0.982	0.0287	0.7671	0.918
		Bayesian SGoF	0.0091	0.669	1	0.0105	0.663	0.998	0.0227	0.6557	0.945
		Bayesian* SGoF	0.0091	0.669	1	0.0105	0.663	0.998	0.0227	0.6557	0.945
	$\mu = 4$	Binomial SGoF	0.0002	0.846	0.999	0.0009	0.8497	0.986	0.0157	0.8267	0.926
		Bayesian SGoF	0	0.7234	1	0.0001	0.7281	0.999	0.0106	0.713	0.952
		Bayesian* SGoF	0	0.7234	1	0.0001	0.7281	0.999	0.0106	0.713	0.952
$n = 500$	$\mu = 1$	Binomial SGoF	0.0773	0.3251	1	0.0729	0.3215	0.978	0.0395	0.3243	0.872
		Bayesian SGoF	0.0728	0.3011	1	0.0677	0.2981	0.991	0.0342	0.3046	0.895
		Bayesian* SGoF	0.0728	0.3011	1	0.0677	0.2981	0.991	0.0342	0.3042	0.895
	$\mu = 2$	Binomial SGoF	0.0191	0.847	1	0.0193	0.8475	0.97	0.0275	0.8293	0.905
		Bayesian SGoF	0.015	0.8112	1	0.0151	0.8118	0.995	0.0249	0.793	0.918
		Bayesian* SGoF	0.015	0.8112	1	0.0151	0.8118	0.995	0.0249	0.793	0.918
	$\mu = 4$	Binomial SGoF	0.0001	0.9177	1	0.0013	0.9164	0.976	0.0196	0.8941	0.888
		Bayesian SGoF	0	0.8777	1	0.0006	0.877	0.987	0.017	0.8581	0.906
		Bayesian* SGoF	0	0.8777	1	0.0006	0.877	0.987	0.017	0.8581	0.906

estimated and therefore SGoF-type procedures lose their FWER control. This fact is more clear when n is large. For example, in the case $n = 500$, Binomial SGoF and the pre-test Bayesian SGoF reported a FDR of 0.038 and 0 under independence and 0.16 and 0.141 when $\rho = 0.8$, respectively. We also can see in this table that the pre-test Bayesian method is more conservative than the basic Bayesian one, as expected. For example, for $n = 50$, basic Bayesian SGoF reported a FDR of 0.1,0.121 and 0.129 (depending on ρ), while the pre-test method gave FDR's of 0.001, 0.016 and 0.078, respectively. This happens because the proportion of trials with $s \geq s_\alpha$ is relatively small. These differences between the two procedures decrease (in relative terms) as n and ρ grows, because the probability of the event $s \geq s_\alpha$ increases with n and with ρ. As regards the posterior probability of the complete null, in general, $P(H_0|\overrightarrow{x})$ takes large values, varying between 0.7079 and 0.9493. However, in the special case $n = 500$ and $\rho = 0.8$, the prior information is misleading, giving a posterior probability

Table 7: Proportion of trials that $s \geq s_\alpha$ and $N_n^b(\alpha) > 0$ and posterior probabilities corresponding to $\Pi_0 = 0.9, 0.7, 0.5$

			$\mu = 1$			$\mu = 2$			$\mu = 4$		
			$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$
$n = 10$	$\Pi_0 = 0.9$	Posterior	0.722(0.2039)	0.7125(0.2221)	0.7187(0.2602)	0.6034(0.2777)	0.6050(0.2755)	0.6241(0.2895)	0.5917(0.2807)	0.5911(0.281)	0.6127(0.2954)
		$s \geq s_\alpha$	0	0.002	0.033	0.012	0.008	0.043	0.01	0.012	0.045
	$\Pi_0 = 0.7$	Posterior	0.533(0.2956)	0.5272(0.2997)	0.5535(0.3185)	0.2132(0.2663)	0.2105(0.2629)	0.228(0.2778)	0.1844(0.2541)	0.189(0.2523)	0.2097(0.2695)
		$s \geq s_\alpha$	0.01	0.02	0.053	0.196	0.174	0.205	0.232	0.22	0.205
	$\Pi_0 = 0.5$	Posterior	0.3803(0.3086)	0.5435(0.2935)	0.403(0.3322)	0.0454(0.121)	0.0526(0.1377)	0.0497(0.1336)	0.0317(0.1008)	0.039(0.1231)	0.036(0.1123)
		$s \geq s_\alpha$	0.07	0.077	0.106	0.622	0.612	0.619	0.675	0.681	0.665
$n = 50$	$\Pi_0 = 0.9$	Posterior	0.7406(0.2472)	0.7183(0.2789)	0.7462(0.3021)	0.402(0.3355)	0.4174(0.348)	0.5662(0.3437)	0.3548(0.3287)	0.3683(0.3375)	0.526(0.3517)
		$s \geq s_\alpha$	0.032	0.068	0.102	0.288	0.284	0.193	0.285	0.295	0.215
	$\Pi_0 = 0.7$	Posterior	0.2427(0.298)	0.2746(0.3211)	0.4107(0.3911)	0.0028(0.0261)	0.0024(0.0241)	0.0065(0.0487)	0.0007(0.0105)	0.002(0.0235)	0.0032(0.0278)
		$s \geq s_\alpha$	0.459	0.431	0.377	0.996	0.988	0.983	0.998	0.995	0.985
	$\Pi_0 = 0.5$	Posterior	0.0421(0.1309)	0.0574(0.1547)	0.196(0.3236)	0(0)	0(0)	0(0.0004)	0(0)	0(0)	0(0)
		$s \geq s_\alpha$	0.878	0.865	0.675	1	1	0.999	1	1	1
$n = 500$	$\Pi_0 = 0.9$	Posterior	0.2394(0.3064)	0.4082(0.4044)	0.5131(0.4292)	0(0.0006)	0.0007(0.0072)	0.0709(0.1969)	0.0003(0.0081)	0.001(0.0194)	0.0401(0.1355)
		$s \geq s_\alpha$	0.641	0.484	0.245	0.999	0.999	0.894	1	1	0.949
	$\Pi_0 = 0.7$	Posterior	0(0)	0.0001(0.0032)	0.1389(0.3089)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
		$s \geq s_\alpha$	1	1	0.814	1	1	1	1	1	1
	$\Pi_0 = 0.5$	Posterior	0(0)	0(0)	0.0532(0.2043)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
		$s \geq s_\alpha$	1	1	1	0.919	1	1	1	1	1

Table 8: Mean and standard deviation of the interval lengths along the 1000 replicates for the special cases $\Pi_0 = 0.5, 0.9$ and $n = 50$

		$\rho = 0$	$\rho = 0.2$	$\rho = 0.8$
Frequentist	$\mu = 1$	0.1497(0.0367)	0.1494(0.0418)	0.122(0.0669)
$\Pi_0 = 0.9$	$\mu = 2$	0.1881(0.0292)	0.1867(0.0325)	0.1705(0.0453)
	$\mu = 4$	0.1919(0.0291)	0.1905(0.0326)	0.1771(0.0418)
$\Pi_0 = 0.5$	$\mu = 1$	0.2334(0.0219)	0.2325(0.0243)	0.2204(0.052)
	$\mu = 2$	0.2744(0.0038)	0.274(0.0043)	0.2715(0.0124)
	$\mu = 4$	0.274(0.0043)	0.274(0.0046)	0.2708(0.0184)
Bayesian	$\mu = 1$	0.1568(0.0286)	0.1568(0.0329)	0.1387(0.0485)
$\Pi_0 = 0.9$	$\mu = 2$	0.1889(0.0252)	0.1877(0.0279)	0.1746(0.037)
	$\mu = 4$	0.1922(0.0252)	0.1911(0.028)	0.1798(0.0355)
$\Pi_0 = 0.5$	$\mu = 1$	0.2287(0.0195)	0.2279(0.0216)	0.2182(0.0433)
	$\mu = 2$	0.2656(0.0034)	0.2652(0.0039)	0.2629(0.0111)
	$\mu = 4$	0.2652(0.0039)	0.2652(0.0041)	0.2624(0.0156)

as low as 0.09%.

In Tables 4 to 6 we report the FDR, power and coverage obtained by the three methods in the situations with $\Pi_0 = 0.9$,$\Pi_0 = 0.7$ and $\Pi_0 = 0.5$, while in Table 7 we report the corresponding proportions of the event $s \geq s_\alpha$ and summaries of the posterior probability of the complete null.

Tables 4, 5 and 6 reveal that SGoF-type strategies are not controlling FDR under the alternative at any given level, although they can report a very small FDR compared to α when the effects are intermediate to strong ($\mu = 2, 4$) or in the case of $n = 10$ tests. On the other hand, we can see that Bayesian SGoF tends to be more conservative than Binomial SGoF, reporting lower values for FDR and power in all cases. Two particular situations with $n = 10$ are exceptions to this, with the FDR of Bayesian SGoF slightly larger than that of its frequentist analogue.

We have computed $100(1-\alpha)\%$ (frequentist) confidence intervals for $\theta = P(p_i \leq \gamma)$, and $100(1-\alpha)\%$ (bayesian) credible intervals for the same parameter. For illustration, in Table 8 we report the mean and standard deviation of the interval lengths along the 1000 replicates for the special cases $\Pi_0 = 0.5, 0.9$ and $n = 50$. It is seen that the Bayesian intervals are wider than the frequentist intervals for $\Pi_0 = 0.9$, but the opposite occurs for $\Pi_0 = 0.5$. Therefore, one should not always relate the conservative nature of Bayesian SGoF with the chance to get a narrower interval for θ.

Again, as in the case of no effects, the pre-test Bayesian method is more conservative than the basic one, but they get closer to each other in terms of FDR and power when n increases and also when ρ grows (like in Table 3, the probability of having $s \geq s_\alpha$ and $N_n^b(\alpha) > 0$ grows with n and ρ under the

alternative too). On the other hand, when the proportion of effects increases, the differences between the pre-test Bayesian method and the basic one in power vanish. This relates to the fact that the probability of $s \geq s_\alpha$ increases as the proportion of effects grows. The coverage reported by the Bayesian SGoF tends to be larger than the coverage of the Binomial SGoF, and the coverage of the pre-test Bayesian procedure is always the largest. This agrees with the relative conservativeness degree of the various methods. Regarding the posterior probabilities, they decrease as the proportion of effects grows (as expected), reaching the value 0 in many cases; for example, when $n = 500$, $\Pi_0 \geq 0.7$ and $\mu = 2, 4$.

A property claimed to hold for SGoF is that its power increases with the number of tests n [3]. Our simulations show that this feature may fail when considering low sample sizes (from $n = 10$ to $n = 50$), although it is well seen when moving from $n = 50$ to $n = 500$.

4 Real data illustrations

Two real medical datasets are considered in this section for illustration purposes. The first dataset refers to a situation in which the number of tests (n) is small; the tests correspond to a sequence of 15 two-sample comparisons performed on 15 different variables. The second example of application is related to a high-dimensional setting where more than 3,000 tests are performed, corresponding to the comparison of mean gene expression levels in two groups of patients.

4.1 Neuhaus data

Neuhaus *et al.* [16] investigated in a randomized multicenter clinical trial with 421 patients the effects of two different treatments for acute myocardial infarction: improved thrombolysis (rt-PA), which has been reported to yield higher patency rates than those individuals with standard regimens of thrombolytic treatment, and anisoylated plasminogen streptokinase activator (APSAC). Four families of hypotheses were identified; we focus on the study of cardiac and other events after the start of thrombolitic treatment ($n = 15$ hypotheses related to reinfarction, recurrent ischemia, blood preasure decrease, bleeding, allergic reaction, cardiogenic shock and in-hospital death). In the original paper, there is no attention to the problem of the multiplicity of tests.

Benjamini and Hochberg [17] analyzed this set of p-values with a FDR-based strategy. When controlling the FDR at 5%, Benjamini-Hochberg procedure, they were able to reject the 4 nulls with the smallest p-values, thus identifying significant improvements of rt-PA when compared to APSAC for allergic reaction, two different aspects of bleeding (bleeding pucture site and bleeding overall), and mortality. Classical SGoF procedure with $\gamma = \alpha = 0.05$ was applied to this dataset by Castro-Conde and de Uña-Álvarez. SGoF method provided 7 rejections (9 out of the 15 p-values fell below γ), which makes also significant another aspect of bleeding (bleeding transfusion), cardiogenic shock,

and a blood pressure decrease. A 95% confidence interval for $\theta = P(p_i \leq \gamma)$ is given by $(0.3333, 0.8667)$. The classical frequentist (one-sided) p-value for the complete null hypothesis is $p_f = 7.42 \times 10^{-9}$.

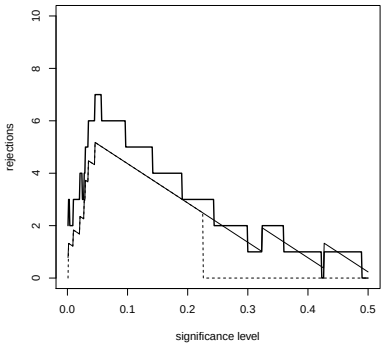
When using the non-informative prior $\pi(\theta) = 1$, Bayesian SGoF reports $N_n^b(0.05) = 5.1152$ rejections, which is more conservative than original SGoF. One explanation for this is that, under the alternative, Bayesian SGoF is implicitly assuming a marginal correlation of $\rho = 1/3$ between each pair of indicators $(X_i, X_j) = (I(p_i \leq \gamma), I(p_j \leq \gamma))$, which are taken as independent by original SGoF. Correlation results indeed in an extra variance [18, 8]. More specifically, the estimated standard error of the frequentist $\widehat{\theta} = \overline{X}_n$ is 0.1265 under independence. On its turn, the standard deviation of the posterior distribution of θ is $\sqrt{V(\theta|\overrightarrow{x})} = 0.1760$, larger than 0.1265 in any case. Another explanation for $N_n^b(0.05) < N_n(0.05)$ is that the non-informative prior is located at 0.5, while the frequentist estimation of θ reports a larger value (0.6). This has, however, a second-order influence as n grows (see Section 4.2).

The mean of the posterior distribution is 0.5882, close to the frequentist 0.6. A 95% Bayesian credible interval for θ is given by $I_c = (0.3543, 0.8025)$, suggesting evidence against the complete null $H_0 : \theta = 0.05$. Indeed, the posterior probability of H_0 based on the default a priori probabilities $P_0 = P_1 = 1/2$ and the non-informative prior $\pi(\theta) = 1$ is $P(H_0|\overrightarrow{x}) = 1.15 \times 10^{-7}$. On the other hand, one may apply the pre-test Bayesian SGoF, by taking $\pi(\theta)$ to be a beta model located at the null. As indicated in Section 2, for $P_0 = P_1 = 1/2$, $n = 15, s = 9$ and $\gamma = 0.05$, the minimum value of $P(H_0|\overrightarrow{x})$ is 1.2×10^{-6} (corresponding to a pairwise correlation of $\rho = 0.4281$, see Table 1), which is below $\alpha = 0.05$. Only values of ρ smaller than 0.006361 or greater than 0.999989 lead to $P(H_0|\overrightarrow{x}) \geq 0.05$, but no one would consider the corresponding priors $\pi(\theta)$ as a reasonable choice for the pre-test. Summaryzing, even when biasing the Bayesian analysis towards H_0, there is no reason to accept that all the null hypotheses under consideration true.

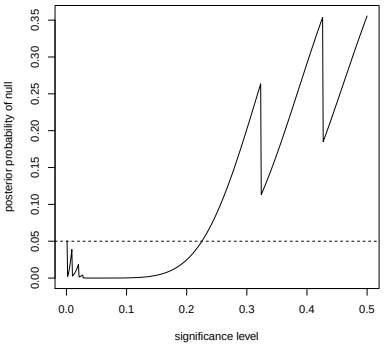
To illustrate how the pre-test may influence the results, we consider now the more stringent case $\gamma = 0.001$. Then, only 2 p-values fall below γ, which reports a one-sided frequentist p-value of $p_f = 0.0001$ for the complete null. In this case, the lower bound for $P(H_0|\overrightarrow{x})$ is $\underline{P}(H_0|\overrightarrow{x}) = 0.0504 > 0.05$ and, therefore, no evidence against H_0 is obtained at level 0.05 from a Bayesian viewpoint. With $\gamma = 0.001$ and $\alpha = 0.05$, frequentist SGoF rejects the nulls with the smallest 2 p-values, while Bayesian SGoF without the pre-test would reject 1 null ($N_n^b(\alpha) = 0.7822$).

It is interesting to look at the relative results provide by frequentist and Bayesian SGoF procedures when letting the significance level γ vary. To this end, in Figure 1, left, we depict the values of $N_n(\alpha)$ and $N_n^b(\alpha)$ (again for $\alpha = 0.05$) when γ changes on a grid of values from 0.001 to 0.5 with step 0.001. Both curves are roughly increasing-decreasing, corresponding to the fact that SGoF criterion is based on a comparison between the observed cdf of the p-values and the uniform cdf, which is the expected one under the complete null; the distance between these cumulative curves along γ is maximum at some

Figure 1: Neuhaus data.



(a) Number of rejected nulls along the significance threshold γ: frequentist SGoF (thick solid line), basic Bayesian SGoF (thin solid line), and pre-test Bayesian SGoF (dashed line).



(b) Solid line: Lower bound for the posterior probability that the complete null is true depending on the significance threshold γ(solid line). Dashed line: line y=0.05

central point γ. On the other hand, from Figure 1, left, it is seen that Bayesian SGoF is more conservative than its frequentist counterpart along the several significance levels. Indeed, when rounding the number of rejections to the closest integer, $N_n^b \leq N_n$ happened for the 500 values of γ, $N_n^b \leq N_n - 1$ for 220, and $N_n^b \leq N_n - 2$ for 33 cases. For comparison purposes, results of pre-test Bayesian SGoF are reported in Figure 1 too. It is seen that $N_n^b = N_n^{b*}$ but for the case $\gamma = 0.001$ and for large thresholds (namely $\gamma > 0.225$), where the Bayesian evidence against the complete null vanishes. Figure 1, right, displays the curve $\gamma \mapsto \underline{P}(H_0|\overrightarrow{x})$ which is used by the pre-test method to update the rejection rule given by basic Bayesian SGoF.

4.2 Hedenfalk data

Hedenfalk et al. [14] performed a microarray study of hereditary breast cancer. 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, a comparison of means was performed through a suitable two-sample test; the sizes of the groups were 7 and 8 subjects respectively. Following previous analysis of these data [19], 56 genes were eliminated because they had one or more measurements above 20. This left $n = 3,170$ genes. Application of Benjamini-Hochberg FDR method (at 5% level) reported 157 significant cases [4].

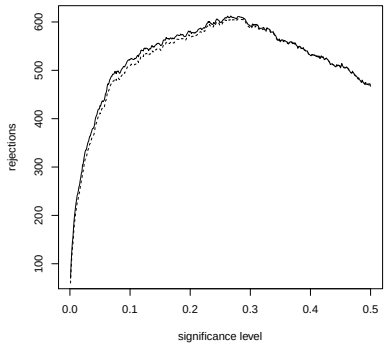
For Hedenfalk data, $s = 606$ p-values (19.12% of the 3170) fell below the significance threshold $\gamma = 0.05$. This resulted in a highly significative frequentist p-

value in the binomial test ($p_f = 0$). A 95% confidence interval for $\theta = P(p_i \leq \gamma)$ is given by $I_f = (0.1773, 0.2047)$. When taking $\alpha = 0.05$, original (frequentist) SGoF ($N_n(\alpha)$) declared 428 genes as differentially expressed; this amount decreased to 412 when applying its conservative version $N_n^*(\alpha)$ [8]. Application of basic Bayesian SGoF with non-informative (uniform) prior resulted also in 412 nulls rejected. These findings are in agreement with the relationship between Bayesian SGoF and conservative frequentist SGoF for large sample sizes discussed before (see the Example in Section 2). To understand why Bayesian SGoF quotes a smaller number of effects compared to $N_n(\alpha)$ note that, by using the beta-normal approximation $l_\alpha(\pi, \overrightarrow{x}) \approx E(\theta|\overrightarrow{x}) - \sqrt{V(\theta|\overrightarrow{x})} z_\alpha$ and since $E(\theta|\overrightarrow{x}) = (1 + n\overline{X}_n)/(2+n)$ and $V(\theta|\overrightarrow{x}) \approx \overline{X}_n(1-\overline{X}_n)/n$ as $n \to \infty$, we have ($l_\alpha(\pi, \overrightarrow{x}) > \gamma$ provided)

$$\begin{aligned} N_n(\alpha) - N_n^b(\alpha) &\approx n(\overline{X}_n - E(\theta|\overrightarrow{x})) + \sqrt{n}\left[\sqrt{nV(\theta|\overrightarrow{x})} - \sqrt{\gamma(1-\gamma)}\right] z_\alpha + 1 \\ &\approx n\frac{2\overline{X}_n - 1}{2+n} + \sqrt{n}\left[\sqrt{\overline{X}_n(1-\overline{X}_n)} - \sqrt{\gamma(1-\gamma)}\right] z_\alpha, \end{aligned}$$

the second term being dominant and positive as long as $\gamma < \overline{X}_n < 1 - \gamma$ (as it happens in this case). Therefore, variance is the main responsible for the different results.

Figure 2: Number of rejected nulls along the significance threshold γ: frequentist SGoF (solid line) and Bayesian SGoF (dashed line). Hedenfalk data.



The mean of the posterior distribution of θ is 0.1912, while a 95% (Bayesian) credible interval for θ is $I_c = (0.1779, 0.2052)$. The posterior probability of $H_0 : \theta = \gamma$ for uniform $\pi(\theta)$ and default priors $P_0 = P_1 = 1/2$ is $P(H_0|\overrightarrow{x}) = 8.50 \times 10^{-173}$. A lower bound for $P(H_0|\overrightarrow{x})$ based on the very default priors and a model for $\pi(\theta)$ located at $\theta = \gamma$ is $\underline{P}(H_0|\overrightarrow{x}) = 1.28 \times 10^{-172}$ (attained for a pairwise correlation between the indicators $X_i = I(p_i \leq \gamma)$ of $\rho = 0.1144$). Moreover, it happens $P(H_0|\overrightarrow{x}) < 0.05$ along $\rho \in \left(10^{-6}, 1 - 10^{-6}\right)$. This strongly

suggests that at least one gene is differentially expressed; the existence of true effects in Hedenfalk data has been object of some discussion in recent research, due to the existing correlation among the tests [20].

To illustrate the influence of the significance threshold γ, in Figure 2 we report the values of $N_n(\alpha)$ and $N_n^b(\alpha)$ for $\alpha = 0.05$ when γ changes on a grid of values from 0.001 to 0.5 with step 0.001. The values for $N_n^{b*}(\alpha)$ are not displayed in this case since the pre-test had no influence on the results. Like for Neuhaus data, both curves are roughly increasing-decreasing; on the other hand, they are apparently close to each other. However, when summaryzing the results for $\gamma \leq 0.2$, we find that the differences $N_n(\alpha) - N_n^b(\alpha)$ are 12.15 on average; this amount changes to 14.34 and 14.79 when considering the results corresponding to $\gamma \leq 0.1$ and $\gamma \leq 0.05$ respectively. For $\gamma \leq 0.4$ the differences were all positive. Therefore, Bayesian SGoF reports results more conservative than those of its frequentist counterpart, and differences between frequentist and Bayesian SGoF criteria are more clearly seen for small significance levels.

5 Discussion and main conclusions

In this paper, Sequential Goodness-of-Fit (SGoF) multitesting procedure has been considered under the Bayesian paradigm. This has two important consequences in the method's application and interpretation of results. First, since SGoF involves a pre-testing of a point null hypothesis ('the proportion θ of p-values falling below threshold γ is γ'), the differences between Bayesian and frequentist viewpoints in such setting play a role. To be brief, Bayesian hypotheses testing for point nulls is based on the conditional probability that the null is true, given the sampling information; and this is a conservative criterion when compared to frequentist p-values (as those used by classical SGoF). Indeed, frequentist p-values are seen as a wrong way to measure significance in Bayesian inference [11]. In practice, this implies that SGoF method relying on a Bayesian pre-test will accept the absence of features in situations when classical SGoF detects signal. Second, when the complete null of no effects is rejected, Bayesian SGoF proceeds by constructing a credible interval for the 'excess of significant cases' when counting p-values below threshold γ, $\tau_n(\theta) = n(\theta - \gamma)$; this interval is directly obtained from the posterior distribution of θ. The analogue of this in the classical frequentist setting is a (frequentist) confidence interval. Again, in practice this results in that Bayesian SGoF declares a smaller amount of features when compared to its frequentist counterpart. These relative properties of Bayesian and frequentist versions of SGoF have been investigated by simulations and real data applications.

Regarding the interpretation of the results, it should be noted that Bayesian inference is based on a conditional analysis. That is, results are valid for all the situations with the same strength of evidence as the actual data $\overrightarrow{x}$. For example, when a pre-test is performed and the researcher rejects the complete null H_0 whenever $P(H_0|\overrightarrow{x}) < \alpha$, then it is guaranteed a type I error rate of α for samples with the very amount of evidence of $\overrightarrow{x}$. Similarly, credible intervals for

$\tau_n(\theta)$ should only be interpreted conditionally on $\overrightarrow{x}$. Unlike Bayesian inference, classical (frequentist) methods aim to ensure error bounds when averaging the results over all the possible samples, but often they say no much when the interest is restricted to the actual sampling information.

The provided simulations and real data applications have demonstrated that, when the number of tests under consideration (n) is large, Bayesian procedure mimics the conservative version of classical SGoF, $N_n^*(\alpha)$, at least when summaryzing its results along a number of Monte Carlo trials. This is not surprising, since the prior information in which Bayesian inference is based on is less and less relevant as the sample size (n) grows. Still, there is an important drawback behind the application of classical SGoF ($N_n(\alpha)$ or $N_n^*(\alpha)$): the underlying assumption of independence among the tests. This assumption is often violated in practice. Even if one does not like very much a Bayesian approach, it is true that the randomness of θ in the Bayesian setting induces (and hence allows for) a pairwise correlation between the indicators $X_i = I(p_i \leq \gamma)$. In practice, a Bayesian pre-test for $H_0 : \theta = \gamma$ will allow for correlated indicators under the alternative $H_1 : \theta \sim \pi(\theta)$; the researcher may then include a guess for the correlation degree in the prior density $\pi(\theta)$, the natural location of $\pi(\theta)$ being the null value $\theta = \gamma$ otherwise. This flexibility to cope with the correlated setting is not shared by frequentist SGoF; although some extensions of classical SGoF for dependent tests have been proposed [8], a number of practical issues are still unsolved. Bayesian SGoF preserves the pleasant properties of classical SGoF (e.g. large statistical power) while permitting the existence of dependences. In particular, the number of significant features selected by Bayesian SGoF for Neuhaus data (or for Hedenfalk data), see Section 4, is larger than that of FDR-based methods.

The usual criticism against the application of Bayesian methods is their dependence on the prior information. However, it should be mentioned that a Bayesian analysis may be quite objective when based on default priors (such as a prior probability of 1/2 for the complete null) and non-informative prior densities for θ ($\pi(\theta) = 1$). For the Bayesian pre-test, objective choices for $\pi(\theta)$ have been largely discussed in the literature (e.g. Berger and Delampady [11]), so these concerns may be reasonably solved. Summaryzing, the Bayesian perspective over SGoF method have a number of advantages and no visible inconvenient, and it seems to be a promising way to look for significance in the setting of multiple comparisons.

Acknowledgement

This work was supported by Grant MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science and Education. Support from INBIOMED project -FEDER 'Unha maneira de facer Europa' is also acknowledged.

References

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

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

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

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

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

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

[7] Lehmann EL, Romano JP. Generalizations of the familywise error rate. *Annals of Statistics* 2005; **33**: 1138-1154.

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

[9] Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* 2004; **32**: 962-994.

[10] Donoho D, Jin J. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Phil. Trans. R. Soc. A* 2009; **367**: 4449-4470.

[11] Berger J, Delampady M. Testing precise hypotheses. *Statistical Science* 1987; **2**: 317-352.

[12] Berger J. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag: New York, 1985.

[13] Berger J, Sellke T. Testing a point null hypothesis: the irreconcilability of p-values and evidence. *JASA* 1987; **82**: 112-122.

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

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

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

[17] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995; **57**: 289-300.

[18] Owen A Variance of the number of false discoveries. *Journal of the Royal Statistical Society Series B* 2005; **67**: 411-426.

[19] Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of National Academy of Science* 2003; **100**: 9440-9445.

[20] Efron B. Correlation and large-scale simultaneous significance testing. *JASA* 2007; **102**: 93-103.