Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–25
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216628903
smm.sagepub.com
SAGE

A Markov chain representation of the multiple testing problem

Stefano Cabras

Abstract

The problem of multiple hypothesis testing can be represented as a Markov process where a new alternative hypothesis is accepted in accordance with its relative evidence to the currently accepted one. This virtual and not formally observed process provides the most probable set of non null hypotheses given the data; it plays the same role as Markov Chain Monte Carlo in approximating a posterior distribution. To apply this representation and obtain the posterior probabilities over all alternative hypotheses, it is enough to have, for each test, barely defined Bayes Factors, e.g. Bayes Factors obtained up to an unknown constant. Such Bayes Factors may either arise from using default and improper priors or from calibrating *p*-values with respect to their corresponding Bayes Factor lower bound. Both sources of evidence are used to form a Markov transition kernel on the space of hypotheses. The approach leads to easy interpretable results and involves very simple formulas suitable to analyze large datasets as those arising from gene expression data (microarray or RNA-seq experiments).

Keywords

Bayes Factors lower bounds, default Bayes, gene expression, improper priors, RNA-seq

1 Introduction

Multiple hypotheses testing (MHT) consists of a set of statistical techniques in which $m > 1$ statistical tests are stated jointly and the objective is to estimate the partition of the m hypotheses into two sets of sizes m_0 and $m_1 = m - m_0$ regarded as the set of true nulls and true alternatives, respectively. For instance, the analysis of the outcome of gene expression measurements is a typical statistical problem of MHT, where $m > 1$, gene expression levels are compared across two biological populations, by testing a corresponding number of statistical hypotheses with the objective of discovering those genes whose expression level is related with a biological population.

The main input of any MHT procedure is the individual or marginal evidence from each test to obtain some type of joint evidence for all tests. A consistent estimation of the null and alternative sets has the objective of controlling for some type of error rates based on the number of false positives, e.g. false null rejections or false discoveries, while also maintaining a low number of false negatives, e.g. missed null rejections. The typical error rates controlled in MHT procedure

Department of Statistics, Universidad Carlos III de Madrid, Spain; Department of Mathematics and Informatics, Università di Cagliari, Italy

Corresponding author:
Stefano Cabras, Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid/126, 28903 Getafe, Spain.
Email: stefano.cabras@uc3m.es; s.cabras@unica.it

are the False Discovery Rate (FDR) and the False Non-rejection Rate (FNR). These can be broadly defined as the ratio among the number of false discoveries over all discoveries (FDR) and the ratio between the number of missed rejections among all non-rejections (FNR).

The most commonly known and used MHT methods are based on the evidence provided by suitable test statistics through the corresponding *p*-values. The m *p*-values resemble the evidence gathered from observed quantities (i.e. gene expression levels) and the statistical model. In fact, these *p*-values may arise from basic sampling models, such as the test of the equality of the mean in two independent populations, as well as from very complicated ones such as for instance non-parametric models or even from models with a complicated hierarchical structure. It is well known that when the null hypothesis is simple or when the test statistic is ancillary, the theoretical sampling null distribution of the *p*-value is the uniform distribution $U(0,1)$ and the *p*-value is said to be calibrated. The m *p*-values are usually assumed to be calibrated with respect to the $U(0,1)$ distribution and thus MHT inference procedures are reliable with respect to such an assumption. This could be one of the main reasons why *p*-values are so popular in MHT. However, besides the effort in using more complicated and realistic statistical models, the settings in which *p*-values are calibrated are essentially very few and for very simple statistical models. Even asymptotic arguments in favor of such uniformity calibration are unsustainable in light of the fact that, for these kinds of massive experiments such as gene expression studies, replications are usually expensive. Therefore, a part of specific situations, the use of *p*-values in MHT, is problematic as shown in 'Note on multiple testing for composite null hypotheses'–a paper authored by me.[1] In fact, when *p*-values sampling null distribution deviates from the $U(0,1)$, then usual MHT are biased in that no upper bounds on FDR or FNR are guaranteed. To this purpose,[2,3] proposed a MHT procedure (Efron's procedure in the sequel) which estimates, within an empirical Bayes setting, the unknown theoretical sampling null distribution of the *p*-values which may differ from $U(0,1)$.

To work around the difficulties of using *p*-values, barely defined Bayes Factors (BFs in the sequel), $cB_1, \ldots, cB_m$, can be used in MHT, where B_i is the ith BF of the alternative against the null hypothesis in test $i \in \{1, \ldots, m\}$ and c is the unknown indetermination constant assumed to be equal for all tests. In fact, in Bertolino et al.,[4] it is shown that individual fully defined and interpretable BFs are not even necessary in MHT. This avoids the heavy computational techniques required for estimating c to properly calibrate or adjust individual BFs to be individually interpretable. Such types of uncalibrated BFs arise from the fact that BFs in MHT require the elicitation of at least $2m$ priors on models with only an unknown scalar parameter, i.e. one marginal prior for each model under testing. As m is large, the choice of such priors is usually forced to be one, among those that come from formal rules. These types of priors are typically improper and so the BF for single hypothesis testing is undetermined due to the ratio among priors *pseudo*-constants $c = c_1/c_0$, where c_0 and c_1 are the prior normalizing constant for the parameters of null model, H_0, and the alternative one H_1, respectively. Such prior constants are equal for all tests and this comes from the mathematical fact that all $2m$ models for null and alternative hypotheses, share the same parameter spaces and the same priors. If one were interested in eliminating BF arbitrariness due to c, several approaches in the literature could be considered, such as those in the literature.[5–11] Unfortunately, these methods are based on large sampling theory and their application would be extremely problematic in MHT, as the computational effort necessary for BF single calibration would be necessary for each one of the m tests. Another approach that leads to priors from formal rules which are proper, i.e. c_0 and c_1 are known, is that of the conventional prior approach in Bayarri et al.[12] If, for some reason, there were interest in having individual interpretable BFs in linear models, the approach of conventional priors[12] can still be employed in MHT with moderate m (say, $m < 100$) as a substitute for the one considered below.

However, for large m, the approach proposed here may be considered in conjunction with the conventional prior approach, as well as the not very recommended one that says just use "vague proper priors."[12]

The set of undetermined BFs, $cB_1, \ldots, cB_m$ may also arise from calibrating p-values with respect to the infimum of their respective BFs as explained in Sellke et al.[13] and reported here. Let pv_i be the ith p-value for test i, with $pv_i < 1/e$, then the corresponding infimum of the BF for the alternative against the null is

$$cB_i = \begin{cases} [-epv_i \ln(pv_i)]^{-1}, \text{ for } pv_i < 1/e \\ 1, \text{ otherwise} \end{cases}, \tag{1}$$

where c denotes, with an abuse of notation, the calibration constant of the true and unknown BF B_i with respect to its lower bound, i.e. $B_i \geq cB_i$, for which we only know the product cB_i, but not c or B_i, separately. In this case, c is the same for all i because of the following mathematical fact: the infimum of BFs lower bounds are calculated with the same model for all i. In fact, equation (1) comes from assuming that if the test statistic has been appropriately chosen, then its density under the alternative model is decreasing with the p-value. Therefore, considering the class of $Beta(\xi, 1)$ densities for $0 < \xi \leq 1$ for the p-value under the alternative model and the $U(0, 1)$ density under the null one, then the infimum of the BF over $0 < \xi \leq 1$ is (1) for p-values smaller than $1/e$.[13] For all p-values larger than $1/e \approx 0.37$, we assume that the evidence of the alternative against the null model does not differ from that for $pv_i \in [1/e, 1]$. It is worth noting that p-values larger than 0.37 never induce any MHT procedure to reject the corresponding null hypotheses for m, which is also large, as in the case of gene expression measurements (microarray or RNA-seq experiments). The p-value calibration in (1) can also be further improved by considering the sample size in each test with the approach developed in Pérez and Pericchi.[14]

For the sake of comparison, Efron's, Benjamini–Hochberg (BH), and Benjamini–Yekutieli (BY) procedures are considered. The BY procedure is aimed at controlling the mean FDR under weak dependence assumptions among tests[15] with respect to the original BH procedure.[16] The taxonomy of MHT procedures can be divided into two sets: adaptive and non-adaptive procedures. Efron's procedure belongs to the class of adaptive procedures, as the null distribution of p-values is evaluated conditionally on the observed sample of pv_i, $i = 1, \ldots, m$, while the BH and BY procedures belong to the class of non-adaptive MHT methods because they rely on asymptotic results that are unconditional to the observed p-values. By no means, we do mean that such MHT procedures represent the state of the art of MHT, although they certainly are popular choices in MHT. A broad review of the most popular MHT literature is beyond the scope of this work and there are several articles that we invite the reader to consult such as Dudoit et al.[17] and Farcomeni,[18] along with the references therein.

It is worth noting that from a Bayesian perspective, in which the model for the observable data accounts for all m hypotheses, multiple comparisons do not need to be explicitly considered as illustrated in Gelman et al.[19] This is because the effect of the prior on parameters, which represent the m hypotheses, allows posterior parameter distributions to be shrunk in such a way that the multiplicity of tests is trivially accounted for Gelman et al.[19] Alternatively, the MHT may be posed as a model selection problem in a regression setting as in Bayarri et al.,[12] where 2^m models may be competed in estimating the probability of belonging to one of the two biological populations under comparison. In this setting, priors on model parameters and priors on model space become relevant as illustrated in the literature,[12,20,21] where BFs consistency is studied for asymptotic in both sample size and m. However, although the above approach to Objective Bayesian model selection deals with almost analytical solutions for BFs, comparing 2^m models, when m is potentially of the order of say millions, it still implies millions of models in the $\log_2$ scale! For this reason, we here focus on classical approaches with the main aim of post processing the results of studies that have been conducted with separate analyses for each hypothesis under test. Such results may be expressed either in terms of significant evidence from significance testing or conditional to the sample in the case of calibrated p-values or BFs, respectively. The advantage of the approach pursued here is that very complicated analyses can be embedded into one single analysis, where such multiplicity of significance tests (and/or BF evidences) have a Markov chain (MC) representation with the purposes of obtaining the posterior probability of the set of non null hypothesis. The interpretation of this MC representation is not necessary to obtain such a posterior probability, because it comes from a mathematical result, i.e. the definition of the equilibrium distribution for a discrete space and discrete time Markovian process. This way of proceeding also applies when employing Markov Chain Monte Carlo (MCMC) methods to approximate a posterior distribution. In fact, there is not a physical interpretation of the stochastic process used to approximate a posterior, as its equilibrium distribution is interpreted as the posterior distribution. Despite this, we argue that an interpretation of the proposed Markovian process, underlying the estimation of the posterior distribution of non-null hypotheses, can be stated as the process underlying an imaginary or maybe real, but certainly not formalized, discovery process for which there is no explicit reference in the current literature. Specifically, this random process is with regard to a researcher aiming to contrast hypotheses over some real phenomena, e.g. a gene related to a disease. He/she starts from assuming as true a certain hypothesis i and considers a different one j which can better explain the real phenomena. The process of discarding hypothesis i in favor of j is a random Markovian process where the probability of $i \rightarrow j$, α_{ij}, is proportional to the evidence of hypothesis j against i only. The equilibrium distribution of such a process provides the probability of each hypothesis being a true discovery. With such equilibrium distribution, it is possible to set-up a decision rule aimed at estimating the null and alternative sets, possibly controlling FDR and FNR under such an equilibrium distribution.

The rest of the article is organized as follows: Section 2 describes in more details the representation process, transition probabilities, equilibrium distribution, and finally a very simple decision rule. Section 4 illustrates the above theory validating it through theoretical results and a simulation study for a parametric toy example. Section 5 considers a more complex model with an application to cancer risk factor analysis and gene expression measurements in which evidence from BFs and p-values are compared. Section 6 illustrates another more elaborated decision rule that, differently from the previous one, requires a tuning parameter to be fixed beforehand. Conclusions are left to Section 7.

2 A MC representation of the discovery process

The aim of this section is to formally describe the MC representation process and how to obtain jump probabilities from uncalibrated BFs and the equilibrium distribution. Finally, a very simple decision rule is described with a more elaborate one in Section 6.

2.1 The discovery process

Let $\mathcal{H} = \left\{H_1^{01}, H_2^{01}, \ldots, H_i^{01}, \ldots, H_m^{01}\right\}$ be the set of tests of cardinality m, always involving two hypotheses in each one: the null and the alternative. The set $\mathcal{H}$ of tests represents m states of nature in which a discovery process may sojourn, in the sense that there is a set of states of nature that implies accepting the corresponding alternative hypotheses as true ones or true discoveries. Parallel to gene expression analysis, we have a set of m genes all compared among them in which only a subset is supposed to be related to the phenotype (e.g. a disease) or the two biological populations under comparison.

The problem indeed is that the analyst is not capable of comparing all hypotheses in $\mathcal{H}$ in a bunch, but is only able to explore the set $\mathcal{H}$ as follows: at time t a certain alternative hypotheses in i is considered as the true one, that is, i is a discovery or the null hypothesis in i is rejected. At time $t+1$, an alternative hypothesis j is picked up at random and considered as the new true one if the probability of the alternative hypothesis in j relative to i, $\alpha_{ij} \in [0,1]$ is large enough. Otherwise, at time $t+1$, i is still the hypothesis declared as the true one. This Markovian process, over the state-space $\mathcal{H}$, is assumed to be repeated infinite times (or by infinite researchers), that is, as $t \to \infty$ and the most visited set of hypotheses is considered to be the most probable set of true alternative hypotheses. This Markovian process plays the same role in approximating the probabilities of true discoveries as MCMC does in approximating a posterior distribution in the usual Bayesian practice. As for the MCMC, there is no interpretation in terms of modeling observable quantities; then the above interpretation, in terms of analyst behavior in assessing discoveries, is not essential as the proposed Markovian process does not directly model any observable quantities. However, we think that such an interpretation provides an intuitive and immediate viability of the resulting probabilities of true discoveries for applied science.

2.2 Jump or transition probabilities

The described Markovian process sustains the idea that the discovery process consists of a random jump process from one explanation of the reality to another. The probability of jumps from explanations are proportional to how relatively likely the new explanation is with respect to that at hand. Specifically, we set the jump probabilities as,

$$\alpha_{ij} = \min\left\{\frac{B_j}{B_i}, 1\right\}, \tag{2}$$

where the unknown constant c simplifies and disappears from the problem. Table 1 reports the transition kernel among hypotheses.

The interpretation of (2) is that the collected and analyzed data provide the relative evidence from one hypothesis to another. This is defined even if it is not calibrated and/or the evidence is

Table 1. Transition kernel among hypotheses.

	Hypotheses				
Hypotheses	1	$\cdots$	j	$\cdots$	m
1	1	$\cdots$	α_{1i}	$\cdots$	α_{1m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	α_{i1}	$\cdots$	$\alpha_{ij} = \min\left\{\frac{B_j}{B_i}, 1\right\}$	$\cdots$	α_{im}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	α_{m1}	$\cdots$	α_{mj}	$\cdots$	1

interpretable in a single test. This sheds new light on the importance of the BFs in MHT even if their own individual scale is not interpretable.

2.3 Equilibrium distribution

Jump probabilities in (2) define a Metropolis algorithm in which the proposal distribution, over the discrete space $\mathcal{H}$, is for the sake of simplicity, the discrete uniform on $\{1, \ldots, m\}$ where the probability of state j to be proposed is $1/m$ and the probability to jump is α_{ij}. In this case, standard theory for MCMC leads us to state that the equilibrium distribution over the discrete state-space $\mathcal{H}$ is given by the following steady-state probabilities:

$$\mathbf{p} = \{p_i\}_i^m, \; p_i = \frac{B_i}{\sum_{i=1}^m B_i}, \text{ for } i = 1, \ldots, m \tag{3}$$

The interpretation of p_i is that the evidence for alternative hypothesis i is directly proportional to its own non-calibrated evidence and inversely proportional to that of all other hypotheses at hand. In such an interpretation, the important argument in MHT comes into play, according to which the evidence of a single hypothesis must account for the evidence of all other hypotheses being tested. Note that while $cB_i/(1 + cB_i)$ cannot be interpreted as the posterior probability of the alternative hypothesis in test i because of the presence of c, formula (3), despite having an intuitive interpretation, it has, at the moment, no statistical justification except under the Markovian process illustrated above.

2.4 A basic decision rule

The posterior probabilities that each of the m alternative hypotheses can be considered as the true ones, $\mathbf{p}$, can be used as the base for a decision rule to estimate the null and alternative sets. Such rules can be more or less complicated depending on the assumption about test dependence or other considerations from a decision theory perspective. Such features of a decision rule are common to all MHT procedures discussed in the literature. Here, we consider a very simple but effective one, leaving the reader to Section 6 for a more elaborate procedure. The reasoning about which hypotheses should be considered as true discoveries is based on an orthodox rule for an Objective Bayesian statistician, that is, on the well known insufficient reason principle for a discrete decision space and applied here conditionally on the given set of m hypotheses. It consists in stating that if there is not sufficient reason to believe that one hypothesis is more probable than another, then $1/m$ prior probability to each one should be assigned before collecting evidence. Therefore, the decision rule consists in rejecting null hypothesis i if

$$p_i > 1/m \tag{4}$$

This also means that alternative hypotheses for which $p_i > 1/m$ should be considered as true discoveries. Such a decision rule has no tuning parameters, given the insufficient reason principle, in contrast to the usual ones in MHT where at least an FDR level must be fixed to have a decision. Of course, rejecting all $p_i > 1/m$ does not guarantee the control of FDR in a finite sample. However, it can be shown that asymptotically, for large sample sizes and a large number of tests, the FDR is negligible (see Appendix 1). This simple rule, while it controls the FDR, does not control the FNR.

To achieve such a goal it is necessary to introduce a tuning parameter as explained in Section 6. Empirical results shown below are compatible with this theoretical result.

3 Illustrative example

To illustrate the method for a possible typical biomedical application, consider the study in García-Arenzana et al.[22] and also discussed in the book of McDonald[23] for illustrating multiple comparisons. This study consists in testing associations of 25 types of diets with mammographic density, which is an important risk factor for breast cancer in Spanish women. The p-values for the association study are published in García-Arenzana et al.[22] and reported in Table 2, which contains the unscaled BFs from (1) and the probabilities $\mathbf{p}$.

Three MHT procedures are considered and at the threshold of 20%, the "Total calories" can always be associated with mammographic density, while the "olive oil" dietary variable only is

Table 2. Results for the association study in García-Arenzana et al.[22] of 25 dietary variables with mammographic density in Spanish women.

Dietary variables	p-values	cB_i	$\mathbf{p}$	Loc. FDR	BH	BY	λ
Total calories	0.001	53.256	0.567	0.027	0.025	0.095	<1
Olive oil	0.008	9.524	0.101	0.289	0.100	0.382	5
Whole milk	0.039	2.908	0.031	0.589	0.210	0.801	8
White meat	0.041	2.809	0.030	0.591	0.210	0.801	12
Proteins	0.042	2.763	0.029	0.592	0.210	0.801	16
Nuts	0.060	2.179	0.023	0.600	0.250	0.954	21
Cereals and pasta	0.074	1.909	0.020	0.605	0.264	1.000	29
White fish	0.205	1.132	0.012	0.730	0.491	1.000	40
Butter	0.212	1.119	0.012	0.742	0.491	1.000	46
Vegetables	0.216	1.111	0.012	0.749	0.491	1.000	53
Skimmed milk	0.222	1.101	0.012	0.760	0.491	1.000	63
Red meat	0.251	1.060	0.011	0.822	0.491	1.000	79
Fruit	0.269	1.042	0.011	0.865	0.491	1.000	99
Eggs	0.275	1.036	0.011	0.880	0.491	1.000	105
Blue fish	0.340	1.003	0.011	1.000	0.533	1.000	138
Legumes	0.341	1.003	0.011	1.000	0.533	1.000	169
Carbohydrates	0.384	1.000	0.011	1.000	0.565	1.000	180
Potatoes	0.569	1.000	0.011	1.000	0.782	1.000	222
Bread	0.594	1.000	0.011	1.000	0.782	1.000	313
Fats	0.696	1.000	0.011	1.000	0.870	1.000	387
Sweets	0.762	1.000	0.011	1.000	0.907	1.000	522
Dairy products	0.940	1.000	0.011	1.000	0.986	1.000	685
Semi-skimmed milk	0.942	1.000	0.011	1.000	0.986	1.000	1147
Total meat	0.975	1.000	0.011	0.303	0.986	1.000	2397
Processed meat	0.986	1.000	0.011	0.122	0.986	1.000	$>10,000$

For each dietary variable, the table contains: p-value for the association study, the corresponding unscaled BFs from (1) and the probabilities $\mathbf{p}$. Three MHT procedures are considered: Efron's procedure for which we reported the value of the corresponding Local FDR (Loc. FDR), the BH and BY procedures for which we reported the corresponding adjusted p-values (columns BH and BY). The last column is the value of the cost parameter λ for the loss function illustrated in Section 6.
FDR: False Discovery Rate; BH: Benjamini–Hochberg; BY: Benjamini–Yekutieli.

associated with the BH procedure. Using the proposed cut-off of $1/m = 1/25 = 0.04$ both total calories and olive oil can be considered as associated with mammographic density according to the proposed approach. The last column relates to the decision rule approach, as illustrated in Section 6.

4 Evidence from BFs

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be a realization of experiments each with m different features, i.e. m genes expression. The vector $\mathbf{x}_i$ contains n replications corresponding to the ith experimental feature, for $i = 1, \ldots, m$. The MHT problem can be formalized as a multiple model selection problem as follows:

$$\begin{cases} H_{i0} : f_{i0}(\mathbf{x}_i|\theta_{i0}),\ \pi_{i0}(\theta_{i0}),\ \theta_{i0} \in \Theta_{i0} \\ H_{i1} : f_{i1}(\mathbf{x}_i|\theta_{i1}),\ \pi_{i1}(\theta_{i1}),\ \theta_{i1} \in \Theta_{i1} \end{cases} i = 1, \ldots, m,$$

where $\pi_{i0}(\theta_{i0})$ and $\pi_{i1}(\theta_{i1})$ are default prior distributions and $\{\Theta_{i0}, \Theta_{i1}\}$ is a partition of $\Theta_i \subset \mathbb{R}^K$, $K \geq 1$. We propose using default and improper priors derived from the same formal rule applied to each $f_{ik}(\cdot|\cdot)$, $k=0, 1$. For the sake of simplicity, we assume that $f_{i0}(\cdot|\cdot)$ and $f_{i1}(\cdot|\cdot)$ are members of the same parametric family for each hypothesis i, namely $f_0(\cdot|\cdot)$ and $f_1(\cdot|\cdot)$, respectively. In this case, we have,

$$\begin{cases} \pi_{i0}(\theta_{i0}) = \pi_0(\theta_0) \propto c_0 g_0(\theta_0) \\ \pi_{i1}(\theta_{i1}) = \pi_1(\theta_1) \propto c_1 g_1(\theta_1) \end{cases}, \tag{5}$$

where c_0 and c_1 are the normalizing *pseudo*-constants as $g_0(\theta_0)$, $g_1(\theta_1)$ could be non integrable functions (i.e. improper priors). Prior predictive distributions for null and alternative hypotheses are assumed to exist and are

$$m_{ik}(\mathbf{x}_i) = \int_{\theta_k \in \Theta_k} f_k(\mathbf{x}_i|\theta_k)\pi_k(\theta_k)\mathrm{d}\theta_k, \quad \text{for } k = 0, 1,\ i = 1, \ldots, m.$$

The BF of H_{i1} against H_{i0} is

$$cB_i = \frac{m_{i1}(\mathbf{x}_i)}{m_{i0}(\mathbf{x}_i)} = \frac{c_1}{c_0} \cdot \frac{\int_{\theta_1 \in \Theta_1} g_1(\theta_1) f_1(\mathbf{x}_i|\theta_1)\mathrm{d}\theta_1}{\int_{\theta_0 \in \Theta_0} g_0(\theta_0) f_0(\mathbf{x}_i|\theta_0)\mathrm{d}\theta_0}, \tag{6}$$

which is unscaled because of the arbitrary ratio $c = c_1/c_0$. We define the *unscaled* BF as,

$$B_i = \frac{\int_{\theta_1 \in \Theta_1} g_1(\theta_1) f_1(\mathbf{x}_i|\theta_1)\mathrm{d}\theta_1}{\int_{\theta_0 \in \Theta_0} g_0(\theta_0) f_0(\mathbf{x}_i|\theta_0)\mathrm{d}\theta_0}. \tag{7}$$

Even if B_i has no interpretation in a single test, it can be used in a comparative approach; in fact, suppose having two tests, i and i', if

$$\frac{cB_i}{cB_{i\prime}} = \frac{B_i}{B_{i\prime}} > 1 \text{ for all } i, i\prime \in \{1, \ldots, m\},$$

the evidence in favor of H_{i1} *versus* H_{i0} is larger than that of H'_{i1} *versus* H'_{i0} whatever c is. To characterize asymptotically the proposed procedure, it is important to state the consistency of p_i as defined in (3). That is, for $n, m \to \infty$, $p_i \to 1$ if i is in the set of true alternatives and $p_i \to 0$, otherwise (see Appendix 1).

4.1 Toy example: Testing zero normal means with unknown variance

We illustrate the proposed method using the following toy example.[4] Let $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, m$ be m independent normal populations with unknown variance σ_i^2. Suppose testing

$$\{H_{0i} : \mu_i = 0 \textit{ versus } H_{1i} : \mu_i \neq 0, \quad \forall \sigma_i^2 > 0\}, i = 1, \ldots, m.$$

Sufficient statistics are $\bar{X}_i = 1/n \sum_{j=1}^n X_{ij}$ and $S_i^2 = 1/n \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$, whose observed values are denoted by $\bar{x}_i$ and s_i^2, respectively.

We assume the usual default and improper priors,

$$\pi_0(\mu_i, \sigma_i^2) = c_0 \sigma_i^{-2} \cdot \mathbf{1}_{\{0\} \times \mathbb{R}^+}(\mu_i, \sigma_i^2),$$
$$\pi_1(\mu_i, \sigma_i^2) = c_1 \sigma_i^{-2} \cdot \mathbf{1}_{\mathbb{R} \times \mathbb{R}^+}(\mu_i, \sigma_i^2),$$

where $\mathbf{1}_A(x)$ is an indicator function for the event $x \in A$. The full BF is $cB_i = c_1/c_0 B_i$, where

$$B_i = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \sqrt{\pi S_i^2} \left(1 + \frac{\bar{X}_i^2}{S_i^2}\right)^{n/2}, \tag{8}$$

is the unscaled BF. The calculation of $\mathbf{p}$ in (3) is immediate.

To have a flavor of the FDR and FNR resulting from the proposed method in this specific parametric toy-example, we consider a study of the following scenarios with, $\sigma_i^2 = 1$ and 1000 simulated datasets in each one:

- $n = 10$, 20, 50, 100;
- $m = 100$, 1000, and 2000;
- $m_0/m = 0.9$, 0.95, and 0.99;
- $\mu_i = \mu_A$ where $\mu_A = 0.5$, 1, 2, and 3.

For each simulated data set, we calculate the FDR and FNR generated by the procedure based on BFs or p-values that in this case is the p-value of the Student's t-test on the mean with unknown variance. Such a p-value is calibrated and its BF lower bound is derived according to (1).

Figure 1 reports the resulting FDR. First of all, we can see that the procedure is consistent for $n \to \infty$ regardless of the proportion of null hypotheses m_0/m and it also improves as the signal μ_A becomes larger. Second, we can appreciate how BFs improves generally over p-values in reporting the evidence of each test. In fact, the FDR using BFs is consistently not larger than those obtained with p-value. This is due to the fact that the BF, explicitly compares two alternative hypothesis in each test, namely the zero mean hypothesis against the alternative. The same cannot be said for the p-value.[4] On the contrary, the use of p-value with this simple rule tends to be less conservative and the FNR is smaller using the calibration of p-values instead of BFs as illustrated in Figure 2.

The fact that the procedure becomes less conservative or more liberal is due to the underlaying Bayesian machinery and in particular to the prior, which is relevant in the inference when the

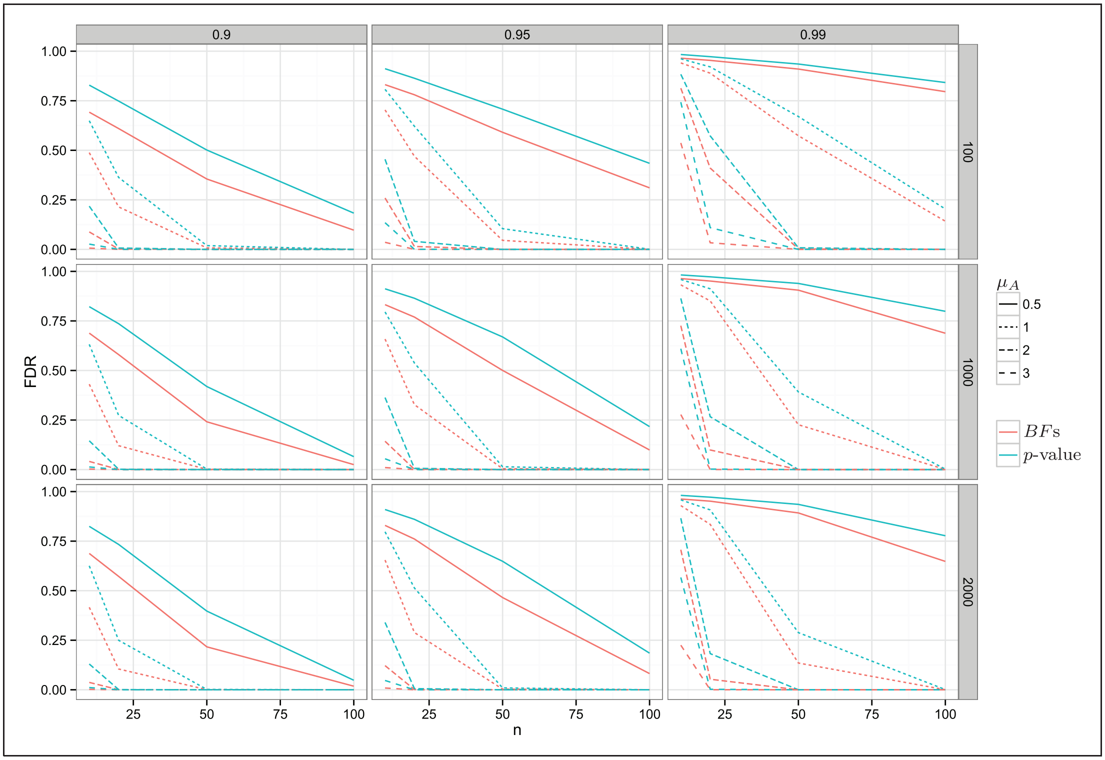

Figure 1. Approximated FDR from simulations. Top scale is the true proportion of null hypotheses m_0/m, right scale is m and bottom scale is the sample size n. Monte Carlo Standard deviation in 1000 simulations is negligible and not reported.

FDR: False Discovery Rate.

evidence is not conclusive. In fact, a priori, it is assumed that at least 1 hypothesis over m is the true alternative, regardless on which one it is. However, it deserves to note that within the Bayesian machinery it is possible to assign different prior probabilities to the each one of the m hypotheses instead of the non-informative $1/m$ and check if a posteriori their probability is increased up to a value that it is judged enough. Essentially the approach here proposed opens many lines of researches in this directions.

Finally, note that for large signals and large n, the procedure consistently estimate the set of true alternatives, as a consequence of BFs consistency. Therefore, both FDR and FNR tends to 0 for large samples and large signals, while in small samples and weak signals things are less clear.

Figures 3 and 4 compare the actual FDR and FNR against that standard procedures, namely BH/BY/Efron and the Bonferroni Procedure, which controls the Family Wise Error Rate (FWER) using the nominal threshold level equal to 5%.

From Figures 3 and 4, it is possible to see that the increment in the sample size n, benefits more the proposed procedure in terms of less FDR and less FNR with respect to the standard ones. This is again the consequence of BFs consistency and hence that of $\mathbf{p}$ consistency. At lower sample sizes, the proposed method, with the simple decision rule, is more liberal as the FDR is larger, but the FNR is lower.

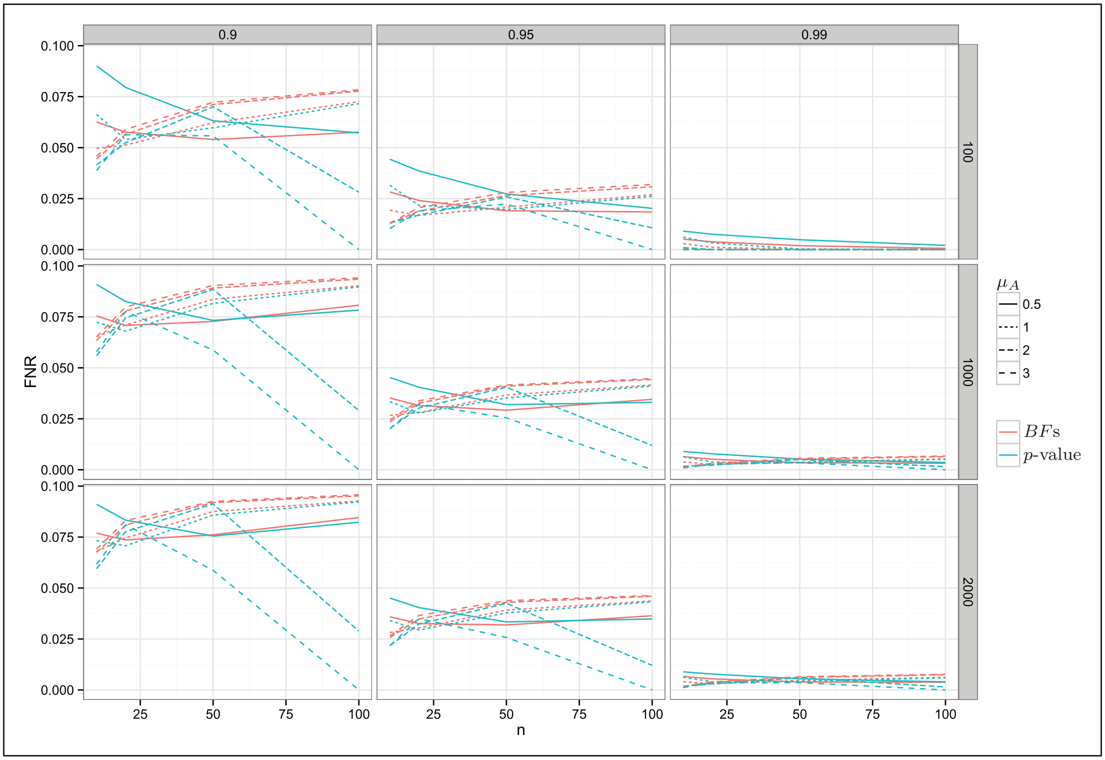

Figure 2. Approximated FNR from simulations. Top scale is the true proportion of null hypotheses m_0/m, right scale is m and bottom scale is the sample size n. Monte Carlo Standard deviation in 1000 simulations is negligible and not reported.

FNR: False Non-rejection Rate.

5 Applications

Differently from the above toy example, we consider a another parametric test that is more realistic for applications to, for instance, gene expression measurements data: that is, testing the equality of the mean of two independent normal populations with all parameters unknown. More formally, let m be the number of features and denote with $\mathbf{x}_{m\times n_x}$ the outcome in population X with n_x replications and $\mathbf{y}_{m\times n_y}$ the outcome in population Y with n_y replications. Let $X_i \sim N(\mu_{X_i}, \sigma^2_{X_i})$ and $Y_i \sim N(\mu_{Y_i}, \sigma^2_{Y_i})$ for $i = 1, 2, \ldots, m$. The set of hypotheses, for $\sigma^2_{X_i} > 0$, $\sigma^2_{Y_i} > 0$ unknown, is the following:

$$\{H_{0i} : \mu_{X_i} = \mu_{Y_i} = \mu_i \; versus \; H_{1i} : \mu_{X_i} \neq \mu_{Y_i}, \forall \sigma^2_{X_i} > 0, \forall \sigma^2_{Y_i} > 0\}, \; i = 1, \ldots, m.$$

Under the usual default priors:

$$\pi_0(\mu_i, \sigma^2_{X_i}, \sigma^2_{Y_i}) \propto \sigma^{-2}_{X_i}\sigma^{-2}_{Y_i} \cdot \mathbf{1}_{\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+}(\mu_i, \sigma^2_{X_i}, \sigma^2_{Y_i}),$$
$$\pi_1(\mu_{X_i}, \mu_{Y_i}, \sigma^2_{X_i}, \sigma^2_{Y_i}) \propto \sigma^{-2}_{X_i}\sigma^{-2}_{Y_i} \cdot \mathbf{1}_{\mathbb{R}\times\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+}(\mu_{X_i}, \mu_{Y_i}, \sigma^2_{X_i}, \sigma^2_{Y_i}),$$

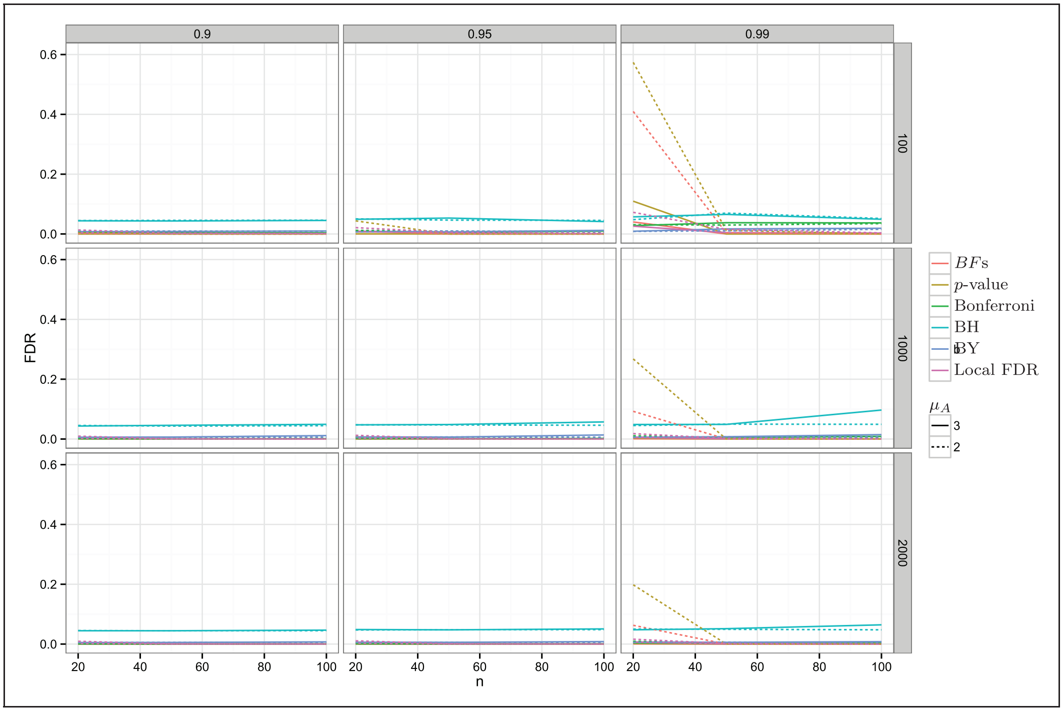

Figure 3. Approximated FDR from simulations for standard procedures and the proposed one. Top scale is the true proportion of null hypotheses m_0/m, right scale is m and bottom scale is the sample size n. Monte Carlo Standard deviation in 1000 simulations is negligible and not reported. The nominal level for all standard procedures is 5%. FDR: False Discovery Rate.

the unscaled BF for H_{1i} against H_{0i} is[4]

$$B_i = \frac{\mathrm{Beta}(\frac{n_x-1}{2},\frac{1}{2})\ \mathrm{Beta}(\frac{n_y-1}{2},\frac{1}{2})\sqrt{S^2_{X_i}S^2_{Y_i}}}{\int_{\mu_i\in\mathbb{R}}\left(1+(\bar{X}_i-\mu_i)^2/S^2_{X_i}\right)^{-\frac{1}{2}n_x}\left(1+(\bar{Y}_i-\mu_i)^2/S^2_{Y_i}\right)^{-\frac{1}{2}n_y}\mathrm{d}\mu_i}, \tag{9}$$

where Beta(a,b) is the beta function evaluated in a, b and $\bar{X}_i, \bar{Y}_i, S^2_{X_i}, S^2_{Y_i}$ are sample means and variances for group X and Y, respectively.

With this model at hand, it is possible to apply the above MHT procedure to the analysis of microarray or RNA-seq experiments. Here, we revisit two old microarray experiments: the first is a calibration experiment where the true differentially expressed (DE) genes are known, whereas the second is a larger study also analyzed, with different approaches, in Singh et al.[24] and Efron.[2] In both cases, we assume that evidence may come either from BFs or from p-values of the usual Student's t-test with the Welch correction for $\sigma^2_{X_i} \neq \sigma^2_{Y_i}$. This is just an asymptotic correction that does not guarantee uniform p-values in finite samples. To be used in the BH/BY and Efron procedures, such p-values are not further calibrated with respect to their BF lower bounds.

The third example is a post analysis of a more recent outcome of an RNA-seq experiment: probabilities $\mathbf{p}$ are calculated from the set of m p-values calibrated with respect to their BF lower bounds according to (1).

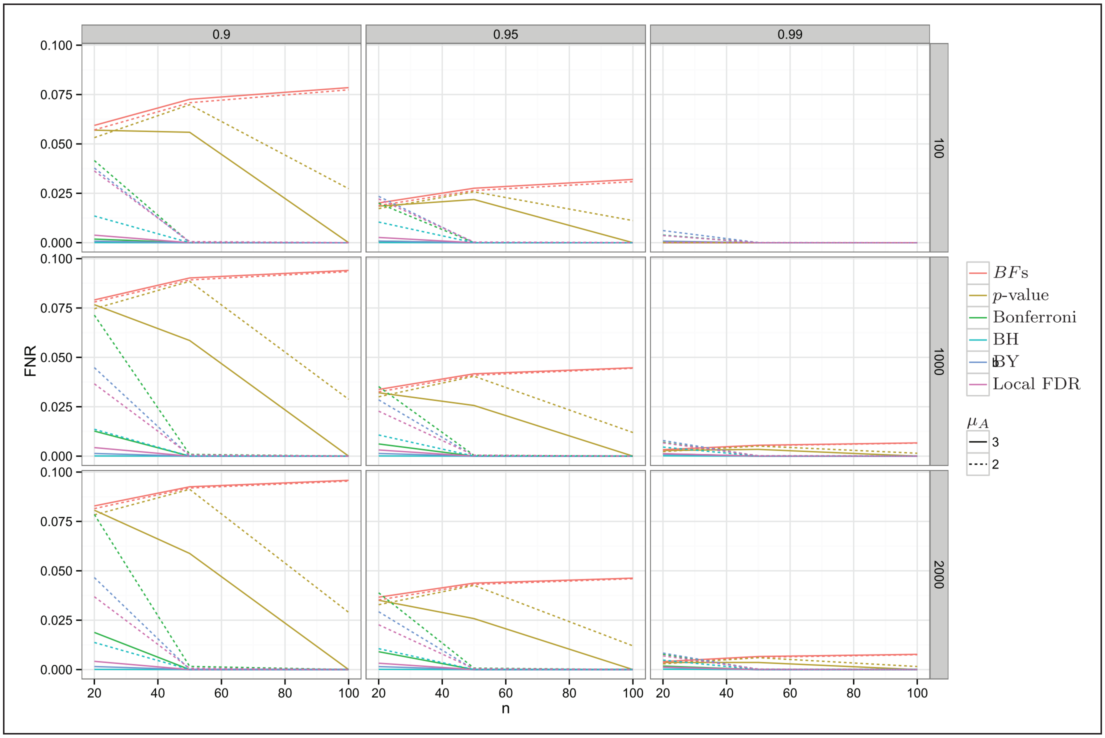

Figure 4. Approximated FNR from simulations for standard procedures and the proposed one. Top scale is the true proportion of null hypotheses m_0/m, right scale is m and bottom scale is the sample size n. Monte Carlo Standard deviation in 1000 simulations is negligible and not reported. The nominal level for all standard procedures is 5%.
FNR: False Non-rejection Rate.

5.1 Microarray: A controlled experiment

We compare results obtained using unscaled BFs and p-values when analyzing gene expression levels of the old Affymetrix HGU95A Latin Square dataset (http://www.affymetrix.com), which has been originally used as the calibration experiment of Affymetrix HGU95A chips. Here, $m = 12626$ and 16 genes have been spiked at controlled levels ranging from 0 to 1024 picoMolar (pM) as shown in Table 3.

A subset of the original 16 replications is considered here to evaluate differences in small samples, specifically the number of replications for X and Y is $n_x = n_y = 12$. In this spiked-in calibration experiment, genes `1597_at` and `38734_at` are the least DE, whereas gene `684_at` is the most DE, because it is absent from population Y and it has the highest concentration in population X. This gene has been eliminated from the analysis as it has a unusually high fold change while that of the not reported genes is between 0.8 and 1.2 fold change. We analyze expression level data obtained from summaries of probe level pairs in the $\log_2$ scale. Such summaries are obtained by pre-processing original microarray measures according to the procedure illustrated in Irizarry et al.[25]

Results of the analysis are reported in Figure 5. For each gene, we reported $\mathbf{p}$ obtained with BF and p-values. Red points indicate genes that are declared as discoveries jointly by Bonferroni/BH/BY/Loc. FDR—Efron procedures. Of course there are still differences among these procedures at 20% of nominal cut-off and with the proposed one. Actual FDR and FNR for all methods are

Table 3. pM concentrations of 16 spiked-in genes in X and Y populations used in this study.

Gene	pM X	pM Y	*Gene*	pM X	pM Y
37777_at	512.00	1024.00	36202_at	8.00	16.00
684_at	1024.00	0.00	36085_at	16.00	32.00
1597_at	0.00	0.25	40322_at	32.00	64.00
38734_at	0.25	0.50	407_at	512.00	1024.00
39058_at	0.50	1.00	1091_at	128.00	256.00
36311_at	1.00	2.00	1708_at	256.00	512.00
36889_at	2.00	4.00	33818_at	64.00	128.00
1024_at	4.00	8.00	546_at	8.00	16.00

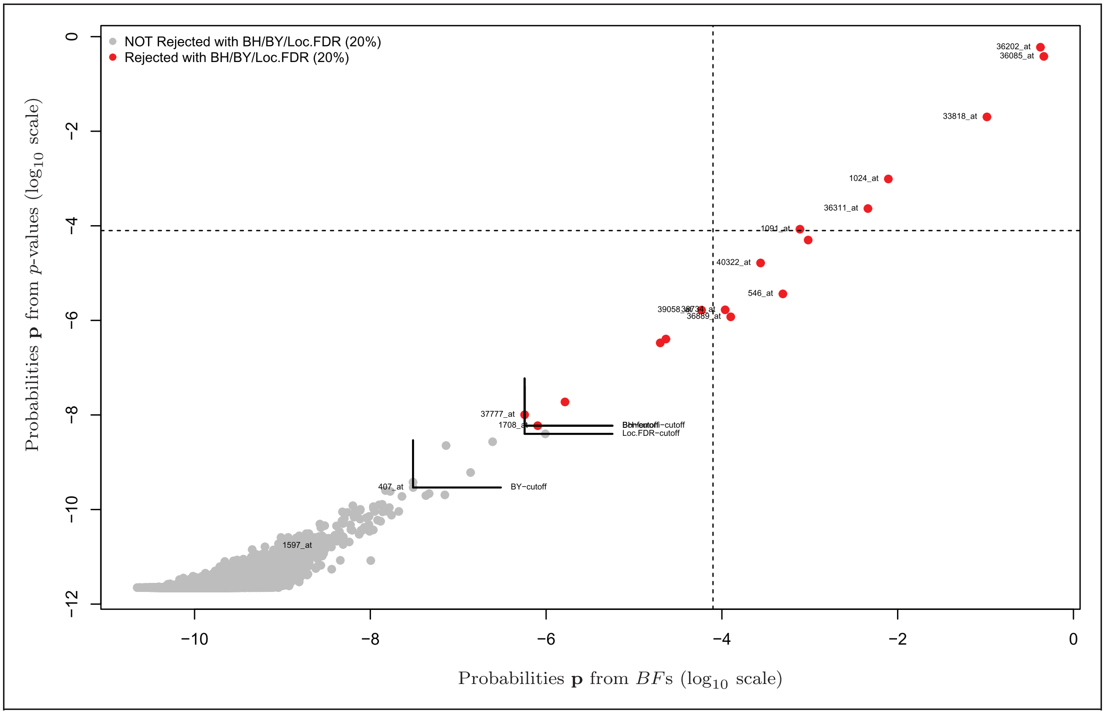

Figure 5. Results from the analysis of a Microarray controlled experiment. Values of the steady-state probabilities for each gene in $\log_{10}$ scale along with the cutoff $\log_{10}(1/m)$ (dashed lines). A total of 17 genes are jointly declared as discovery using all BH/BY/Loc. FDR: Efron procedures with the threshold of 20%. Cut off points are explicitly reported for Efron Local False Discovery Rate BH and BY procedures.

FDR: False Discovery Rate; BH: Benjamini–Hochberg; BY: Benjamini–Yekutieli.

reported in Table 4. The actual FDR for the three procedures is a good deal above the nominal 20%, which just highlight that the interpretation of the nominal 20% is in mean 20%, where the mean is with respect to an infinite resampling. On the contrary, the proposed procedure must be interpreted given the observed sample. There are still a few DE genes that are very difficult for any possible

Table 4. Actual FDR and FNR for the microarray controlled experiment with the considered FDR procedures.

Procedures	Nominal cut-off	FDR	FNR
Efron	20%	$5/18 = 0.28$	$2/12607 = 0.0002$
BH	20%	$4/17 = 0.24$	$2/12608 = 0.0002$
BY	20%	$9/23 = 0.39$	$1/12602 = 0.00008$
Bonferroni	20%	$9/23 = 0.39$	$1/12602 = 0.00008$
$p_i > 1/m$ (on p-value calibration)	none	$0/6 = 0$	$9/12619 = 0.0007$
$p_i > 1/m$ (on BFs)	none	$1/11 = 0.09$	$5/12614 = 0.0004$

BH: Benjamini–Hochberg; BY: Benjamini–Yekutieli.

detection as they have very low probabilities to be visited by the assumed virtual Markovian discovery process.

5.2 Microarray: Prostate data

We compare unscaled BFs with p-values in the analysis of gene expression levels for prostate cancer data.[24] In this study, $m = 6033$ genes with $n_x = 50$ healthy males are compared with $n_y = 52$ prostate cancer cases. Results are shown in Figure 6. The larger sample size of this data set reduces differences between *BF*s and p-values and so the differences between the proposed procedure and Efron procedure or the BY procedure.

In fact, Figure 6 shows that the proposed procedure with $1/m$ cut-off is in between the Efron and BY procedures, with the BH procedure being more conservative. The main message of this example is that for large n, differences among the proposed procedures and the one considered for comparison tend to be small. This is consistent with the simulation study in Section 4.1, namely Figures 3 and 4.

5.3 A RNA-seq experiment: Bovine macrophage response to Mycobacterium bovis infection

In this study, the raw data consists of 3.6 trillion reads of RNA sequences to generate counts of identical sequences that are supposed to represent the abundance of target sequences in the mRNA. Counts are collected for the two biological populations under comparison: bovines infected and non-infected by *Mycobacterium bovis*.[26] Such counts are then used to evaluate the differential gene expressions between said biological populations (see, for instance, Rapaport et al.[27]). After data normalization, there are $m = 11131$ genes that have been compared for differential expression and their corresponding p-values have been calibrated with respect to their BF lower bound (1). The results are compared with Table 1 in Nalpas et al.,[26] which shows the most DE RNA sequences according to their $\log_2$ fold change (*LFC* in the sequel) and also with Table 2 in Nalpas et al.[26] that compares fold-changes in gene expression based on RNA-seq, microarray, and real-time qRT-PCR.

According to our procedure the first two most related sequences are (see Table 5): ENSBTAG00000022396 with probability p_i almost 1 and ENSBTAG00000001725 with a $LFC = 5.67$ that has a raw p-value of 10^{-75}. This latter one has a probability p_i of only 1.1810^{-42},

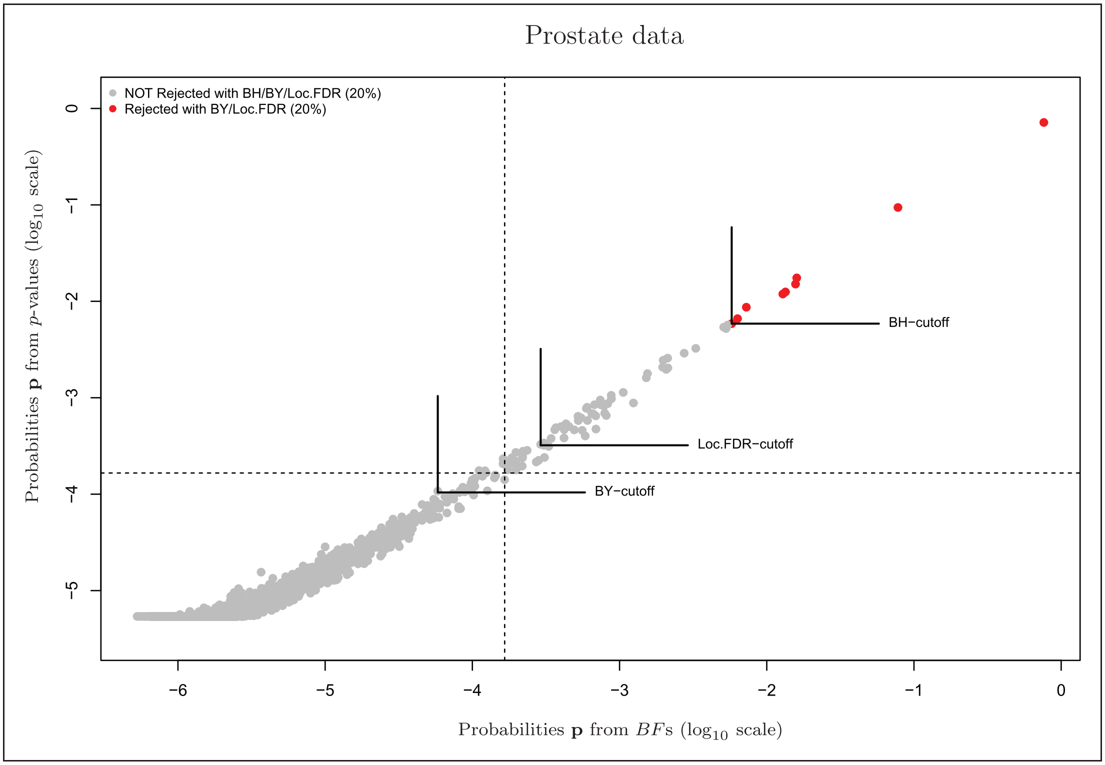

Figure 6. Results from the analysis of Prostate cancer data. Values of the steady-state probabilities for each gene in log$_{10}$ scale along with the cutoff log$_{10}(1/m)$ (dashed lines). Only nine of the most differentiated genes are jointly detected using all procedures. Cut off points are explicitly reported for Efron Local False Discovery Rate BH and BY procedures with the threshold of 20%.

BH: Benjamini–Hochberg; BY: Benjamini–Yekutieli.

with a p-value and/or adjusted p-values which are around 10^{30} times higher than those of ENSBTAG00000022396. This means that, according to our basic procedure only ENSBTAG00000022396, also reported in Table 1 of Nalpas et al.,[26] can be considered the most DE RNA sequence.

Results in Nalpas et al.[26] come from very well established and popular analysis workflow for RNA-seq data, that is, quantification of transcripts using a Python package HTseq followed by identification of DE genes using a R package edgeR. The final result is a sorted list of possible sequences claimed to be DE because of a low enough p-value, according to some MHT procedure[28–30] and large *LFC*. In Nalpas et al.,[26] 2584 genes have been declared as DE, namely those with adjusted p-value with the BH procedure less than 0.05.

There are two differences between the very conservative results obtained here with the simple decision rule and that in Nalpas et al.[26] The first difference is related to Boole's inequality that appears in Proposition 2 of the Appendix 1, which makes the simple rule $p_i > 1/m$ conservative because dependence among all tests is assumed to as strong as possible, which may not be so. Indeed, such a conservativeness can be avoided by introducing a tuning parameter as explained in Section 6 which is usual in MHT approaches implemented in the edgeR package. In fact, if one is

Table 5. RNA-seq experiment: 10 most significant genes.

Rank	Gene ID	Related protein	*LFC*	*p*-value	Bonferroni	BH	BY	Loc. FDR	p	ln(λ)
1	ENSBTAG00000022396	Serum amyloid A	7	−117	−113	−113	−112	−109	0	0
2	ENSBTAG00000001725	Chemokine	6	−75	−71	−71	−70	−69	−42	107
3	ENSBTAG00000013167	Sialic acid	5	−70	−66	−67	−66	−65	−47	111
4	ENSBTAG00000008612	Complement c1	5	−68	−64	−64	−63	−63	−49	116
5	ENSBTAG00000016061	Radical S-adenosyl	5	−65	−61	−62	−61	−60	−52	120
6	ENSBTAG00000034954	Beta-defensin 5	6	−65	−61	−62	−61	−60	−52	121
7	ENSBTAG00000038639	Chemokine	6	−65	−61	−62	−61	−60	−52	122
8	ENSBTAG00000005603	Chemokine	7	−64	−60	−61	−60	−59	−53	124
9	ENSBTAG00000020602	Indoleamine 2	5	−64	−59	−60	−59	−59	−53	125
10	ENSBTAG00000018119	Acyloxyacyl hydrolase	5	−63	−59	−60	−59	−58	−54	127

Columns report for each gene, from left to right: the gene rank according to **p** and λ (see Section 6); gene identification and related protein; *LFC*, raw *p*-value and adjusted ones according to: Bonferroni, BH, BY; the corresponding Local FDR. Last two columns report **p** and ln(λ). All values are in Log$_{10}$ scale, except *LFC* and ln(λ) which are in ln$_2$ and natural ln scales, respectively.
LFC: Log$_2$ Fold Change; BH: Benjamini–Hochberg; BY: Benjamini–Yekutieli.

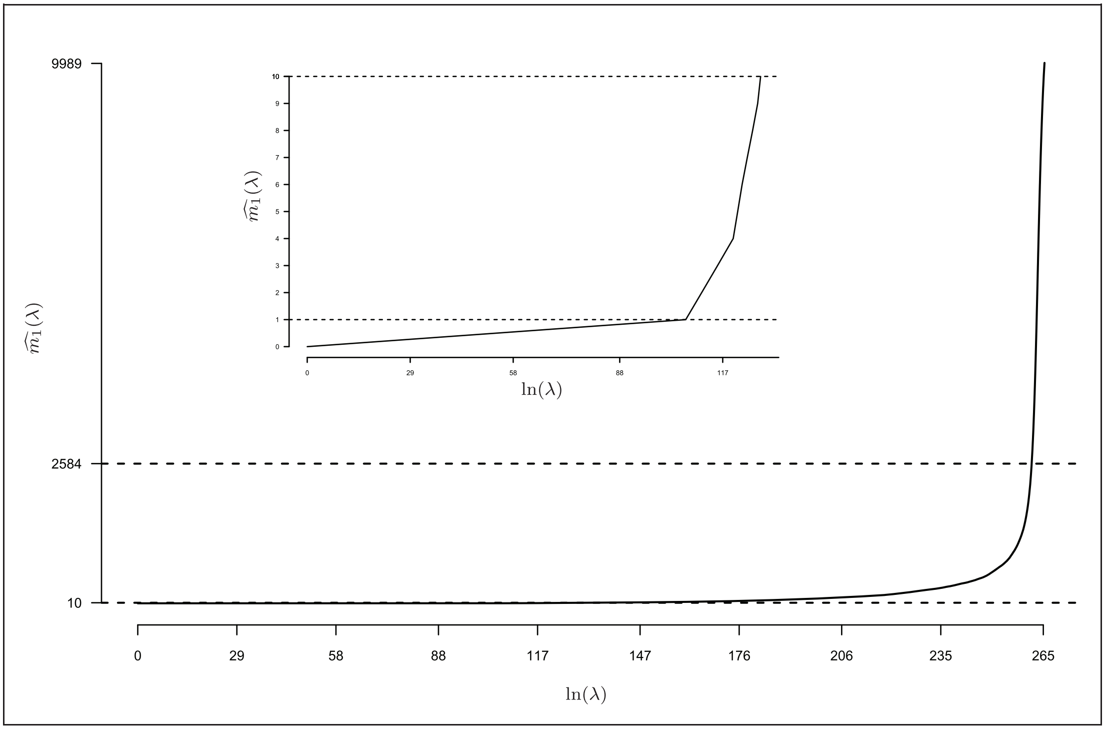

Figure 7. RNA-seq experiment: function $\widehat{m_1}(\lambda)$. The cutoff horizontal lines represent the first 10 genes of Table 1 in Nalpas et al.,[26] the 2584 genes claimed to be differentially expressed according to the BH procedure at the nominal level of 5%. The λ values are reported in natural log scale.
BH: Benjamini–Hochberg.

willing to consider larger set of possible DE genes, Section 6 would allow this by providing also relative cost of a missed rejection with respect to a false discovery that compete to such a set of possible DE genes. This cost opportunity is represented by the λ parameter, which is reported in Figure 7 as the argument of the function of the number of genes declared as DE, $\widehat{m_1}(\lambda)$.

We can see that λ is almost constant for the first 10 genes, suggesting that there could be signal there or even slightly more genes, while 2584 genes appears to be excessive as the derivative of $\widehat{m_1}(\lambda)$ is almost maximum.

To further illustrate the outcome of this experiment and the role of λ, next Figure 8 reports *LFC* and unadjusted *p*-value.

We can see that for low values of λ, signal from genes DE is quiet strong and it decreases with λ, that is either *LFC* decreases or *p*-values increase.

Looking then at the list of the 10 most DE genes in Table 5, it is interesting to note that apart from the first gene, we have few genes related to the chemokine protein, which has also been mentioned in Table 2 in Nalpas et al.,[26] where a comparison of fold-changes in gene expression based on RNA-seq, microarray, and real-time qRT-PCR was performed.

Finally, it is worth to discuss the second difference between the approach here proposed and that in Nalpas et al.[26] Such a difference is more on the basis of the involved fundamentals of statistics. Essentially, the `edgeR` implements, for large samples, the use of *p*-values *alone* to assess the significance of a hypothesis. This is a very popular statistical practice although a questionable one, to the point that some journals started banning *p*-values.[31] However, this practice, in the proposed procedure, is not repudiated, but embraced under the posterior probability principle (PPP) instead of the significance principle.[14] Under the PPP, *p*-values have been calibrated with respect to the minimum probability of the hypothesis of non differential expression (1) given all observed evidence.

6 Remarks: Another decision rule

This section is devoted to illustrating another decision rule based on the vector of steady-state probabilities $\mathbf{p}$. Assume that $\mathcal{H}_1$ and $\mathcal{H}_0$ are the sets of alternative hypotheses declared as true (m_1 in total, with $0 \geq m_1 \geq m$) and false (m–m_1), respectively (e.g. $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_0$ and $\mathcal{H}_1 \cap \mathcal{H}_0 = \emptyset$). There can be different ways of partitioning $\mathcal{H}$, but let us assume that it is defined upon the ordered vector of steady-state probabilities $1 \geq p_{i_1} \geq p_{i_2} \geq \ldots \geq p_{i_{m_1}} \geq \ldots \geq p_{i_m} \geq 0$ and thus on a given number of rejected null hypotheses m_1, that is $\mathcal{H}_1 = \left\{i : i \in i_1, \ldots, i_{m_1}\right\}$ and $\mathcal{H}_0 = \mathcal{H} \setminus \mathcal{H}_1 = \left\{i : i \notin i_1, \ldots, i_{m_1}\right\}$.

The FDR is the expected proportion of false discoveries in $\mathcal{H}_1$, that is,

$$FDR(\mathcal{H}_1) = \frac{E_{\mathbf{p}}(\text{False Discoveries in } \mathcal{H}_1)}{\text{Card}(\mathcal{H}_1)} = \frac{1}{m_1} \sum_{i \in \mathcal{H}_1} (1 - p_i), \tag{10}$$

for $m_1 > 0$ and $FDR(\mathcal{H}_1) = 0$ for $m_1 = 0$. Analogously, the FNR is the expected proportion of missed discoveries in $\mathcal{H}_0$, that is,

$$FNR(\mathcal{H}_0) = \frac{E_{\mathbf{p}}(\text{Missed Discoveries in } \mathcal{H}_0)}{\text{Card}(\mathcal{H}_0)} = \frac{1}{m - m_1} \sum_{i \in \mathcal{H}_0} p_i, \tag{11}$$

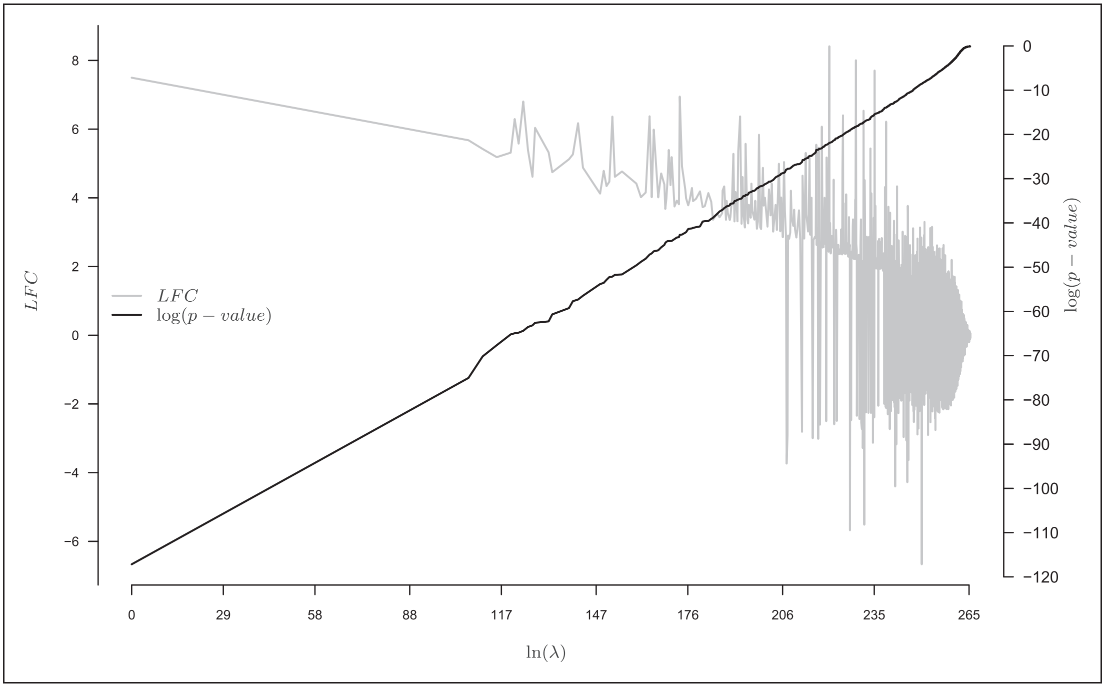

Figure 8. RNA-seq experiment: *LFC* and unadjusted *p*-values as functions of λ.
LFC: Log_2 Fold Change.

A suitable way to partition $\mathcal{H}$ is by fixing m_1 to minimize the following loss function[32]:

$$L_\lambda(m_1) = FDR(\mathcal{H}_1) + \lambda FNR(\mathcal{H}_0),$$

where $\lambda > 0$ is a user specified constant which expresses the relative cost of a missed rejection with respect to a false discovery. It can also be interpreted as the cost-complexity parameter in classification as also λ represents the cost that we have to pay to complicate the explanation of the reality with an increasing set of possible causes labeled as discoveries. For λ fixed, m_1 can be estimated, conditionally on λ using,

$$\widehat{m_1} = \arg\min L_1(m_1) \tag{12}$$

It is worth noting that λ, in this decision rule, is the only tuning parameter and the function $\widehat{m_1}(\lambda)$ is very informative to choose a cutoff value for λ that leads to the optimal decision in the sense that minimizes L_λ. The arguments regarding the choice of such a cutoff are the same as those related to the choice of an optimal cost-complexity parameter in classification.

To have an idea of the behavior of the function $\widehat{m_1}(\lambda)$, we consider the toy example illustrated in Section 4.1 with $m = 100$, $m_1 = 10$, $n = 10$, $\sigma_i^2 = 1$ and $\mu_1, \ldots, \mu_{m_1} = \mu_A$. We analyze four simulated scenarios where $\mu_A = 0$, 1, 2 or 3 in each scenario. These have different signal intensities as in the first there is no signal and in the last one the mean for the alternative hypotheses is three standard deviations from that under the null.

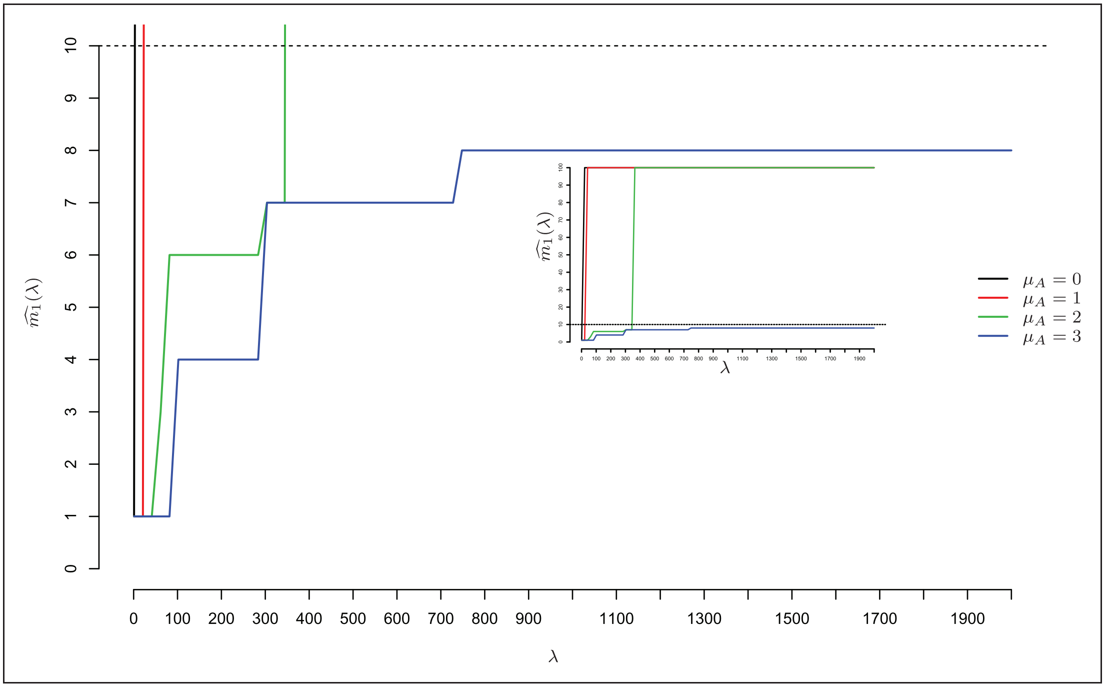

Figure 9. Function $\widehat{m_1}(\lambda)$ for testing zero normal means with unknown variance. We have $m = 100$ hypotheses where $m_1 = 10$ hypotheses come from the alternative model. The four samples are of size $n = 10$, $\sigma_i^2 = 1$ and for value of the signals $\mu_1, \ldots, \mu_{m1} = \mu_A$, where $\mu_A = 0,1,2$, or 3.

The evolution of $\widehat{m_1}(\lambda)$ for the above scenarios is reported in Figure 9, while the actual and estimated FDR and FNR are reported in Figure 10.

Even if these are four examples of MHT, Figure 9 shows quite well the comparative behavior of $\widehat{m_1}(\lambda)$ functions at increasing signal levels that illustrate a general behavior. In fact, when there is no signal, (i.e. $\mu_A = 0$) all hypotheses are true nulls; then they can be considered to explain the phenomena at hand and there is no price that can be paid for missing a discovery. As long as the signal grows, the price of complicating the explanation by considering more hypotheses jointly becomes higher. For example, with $\mu_A = 3$, the data do not practically support more than eight true alternative hypothesis as for $\lambda > 800$ we can pay more, but such payments do not correspond to any new hypothesis that can be considered as true discovery. Of course, for $\lambda \to \infty$ all m hypotheses can be considered as true alternatives, but the price with respect to false discoveries may be unaffordable. Essentially, wide flat regions of $\widehat{m_1}(\lambda)$, indicate possible cut-off values for λ and hence the size of the first $\widehat{m_1}$ most probable hypotheses to be considered as true discoveries. For example, for $\mu_A = 3$, it is clear that a reasonable range to be considered is $\lambda \in (800, 2000)$, that is, the first eight most probable hypotheses should be jointly rejected with a realized FDR of 0% and a FNR of 2% (see Figure 10). At $\mu_A = 2$, we would consider no more than six or seven hypotheses ($\lambda \in (100, 400)$) as at larger values of λ the data do not support any reasonable small set of true alternative hypotheses. Finally, it is clear that when there is no signal, no hypotheses should be really considered as true alternatives because at $\lambda \to 0$ data suggest that all m hypotheses should be rejected.

Note that the cut-offs induced by $\widehat{m_1}(\lambda)$ are data dependent and this is a relevant difference from the usual MHT procedures in which cut-offs are fixed beforehand.

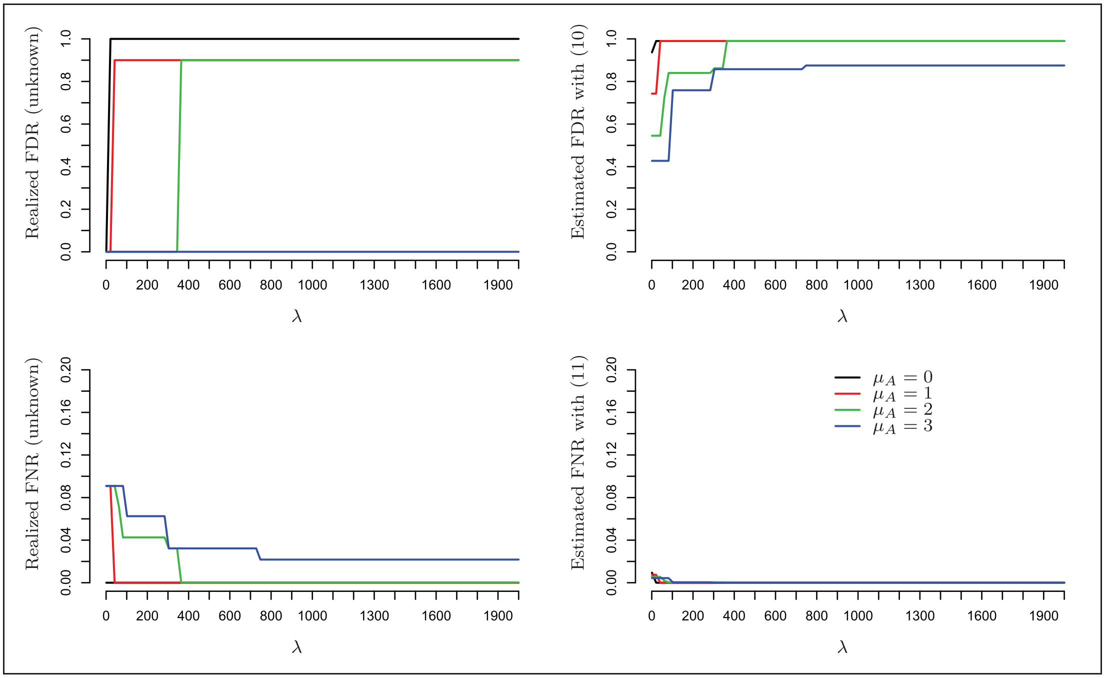

Figure 10. FDR and FNR realized (always unknown) and estimated conditionally on λ with equations (10) and (11) for testing zero normal means with unknown variance. We have $m = 100$ hypotheses where $m_1 = 10$ hypotheses come from the alternative model. The four samples are of size $n = 10$, $\sigma_i^2 = 1$ and for value of the signals $\mu_1, \ldots, \mu_{m1} = \mu_A$, where $\mu_A = 0,1,2$, or 3. The small figure reports the same but with the full vertical scale.
FDR: False Discovery Rate; FNR: False Non-rejection Rate.

Finally, we return to the last column of Table 2 that reports the values of λ for the 25 dietary variables. We can appreciate how reasoning in terms of loss function $L_\lambda(m_1)$ leads to similar conclusions, which is that if the cost of a false discovery is no larger than that of a missed rejection then only "total calories" should be considered as related to the mammographic density. To also consider the "olive oil" dietary variable, we should assume that costs for missed rejection are 5 times larger than those for a false discovery. The behavior of the function $\widehat{m_1}(\lambda)$ for the 25 dietary variables is reported in Figure 11.

The detail for the first 15 most evident dietary variables does not suggests any cut-off as the increment in $\widehat{m_1}(\lambda)$ is almost constant across values of λ between 1 and 62. After that, there seems to be stationary steps, but the only strong signal seems to be between all dietary variables up to "Total meat" and "Processed meat." Of course, if the analyst is willing to consider as related to mammographic density all dietary variables except "Processed meat," then it means that he/she is also willing to consider missed discoveries to be 2397 times more expensive than false discoveries. Such a value of λ, depending on the application, may be too high.

7 Conclusions

The decision rule discussed above is just an example of a possible more sophisticated MHT procedure based on the vector of true discovery probabilities $\mathbf{p}$. However, the key point of our

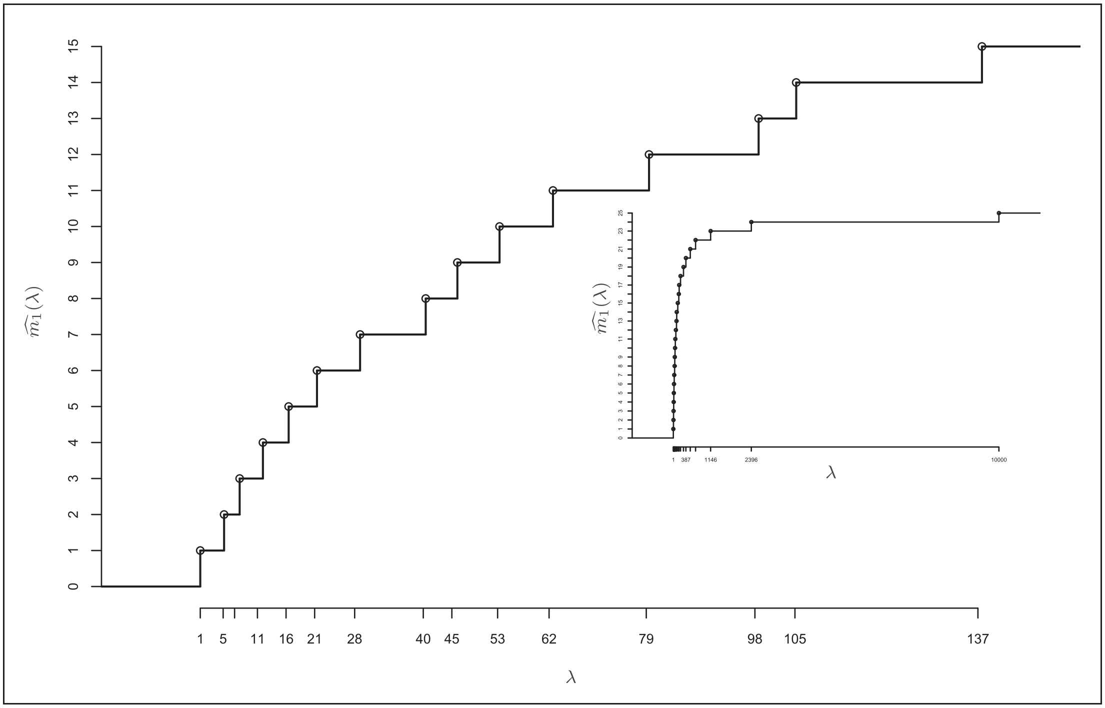

Figure 11. Function $\widehat{m_1}(\lambda)$ for the 25 dietary variables related to mammographic density in Spanish women.[22]

exposition is the Markovian representation of the MHT problem that can be applied to virtually any source of evidence for the single test and is suitable for very large data bases, as the involved formulas are very simple. The strong point of this approach is the straightforward interpretation of **p** as probabilities that hypotheses can be considered as true discoveries given the collected data and the involved statistical models used to analyze them. This is not trivial with current MHT procedures as the interpretation of, say, adjusted p-values requires deep knowledge of statistical reasoning for an applied scientist. Finally, it is important to stress that the proposed Markovian process is to MHT as MCMC is to posterior approximation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author has been funded by Ministerio de Ciencia e Innovación grant MTM2013-42323, ECO2012-38442, RYC-2012-11455, and ACOMP/2015/202, by Ministero dell'Istruzione, dell'Univesità e della Ricerca of Italy, and Regione Autonoma della Sardegna CRP-59903.

References

1. Cabras S. A note on multiple testing for composite null hypotheses. *J Stat Plain Infer* 2010; **140**: 659–666.
2. Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat Sci* 2008; **23**: 1–22.
3. Efron B. Size, power and false discovery rates. *Ann Stat* 2007; **35**: 1351–1377.
4. Bertolino F, Cabras S, Castellanos ME, et al. Unscaled Bayes factors for multiple hypothesis testing in microarray experiments. *Stat Methods Med Res* 2015; **24**: 1030–1043.
5. Moreno E, Bertolino F and Racugno W. An intrinsic limit procedure for model selection and hypotheses testing. *J Am Stat Assoc* 1998; **93**: 1451–1360.
6. Bertolino F, Moreno E and Racugno W. Bayesian model selection approach to analysis of variance under heteroscedasticity. *J Roy Stat Soc D-Sta* 2000; **49**: 503–517.
7. Berger JO and Pericchi LR. The intrinsic Bayes factor for model selection and prediction. *J Am Stat Assoc* 1996; **91**: 109–122.
8. O'Hagan A and Forster J. *Kendall's advanced theory of statistics: Bayesian inference*. Vol. 2, New York: Hafner, 2004.
9. Berger JO and Pericchi LR. Objective Bayesian methods for model selection: Introduction and comparison. In: Lahiri P (ed.) *Model selection*. Vol. 38, Beachwood, OH: Institute of Mathematical Statistics, 2001, pp.135–207http://www.jstor.org/stable/4356165 (accessed 12 January 2016).
10. Berger JO and Pericchi LR. Training samples in objective Bayesian model selection. *Ann Stat* 2004; **32**: 841–869.
11. Moreno E and Girón FJ. Comparison of Bayesian objective procedures for variable selection in linear regression. *Test* 2008; **3**: 472–492.
12. Bayarri M, Berger J, Forte A, et al. Criteria for Bayesian model choice with application to variable selection. *Ann Stat* 2012; **40**: 1550–1577.
13. Sellke T, Bayarri MJ and Berger JO. Calibration of p-values for testing precise null hypotheses. *Am Stat* 2001; **55**: 62–71.
14. Pérez ME and Pericchi LR. Changing statistical significance with the amount of information: The adaptive α significance level. *Stat Probabil Lett* 2014; **85**: 20–24.
15. Benjamini Y and Yekutieli D. The control of the False Discovery Rate in multiple testing under dependence. *Ann Stat* 2001; **29**: 1165–1188.
16. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B* 1995; **57**: 289–300.
17. Dudoit S, Shaffer J and Boldrick J. Multiple hypothesis testing in microarray experiments. *Stat Sci* 2003; **18**: 71–103.
18. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat Methods Med Res* 2008; **17**: 347–388.
19. Gelman A, Hill J and Yajima M. Why we (usually) don't have to worry about multiple comparisons. *J Res Educ Eff* 2012; **5**: 189–211.
20. Casella G, Girón FJ, Martínez ML, et al. Consistency of Bayesian procedures for variable selection. *Ann Stat* 2009; **37**: 1207–1228.
21. Moreno E, Girón FJ and Casella G. Consistency of objective Bayes factors as the model dimension grows. *Ann Stat* 2010; **38**: 1937–1952.
22. García-Arenzana N, Navarrete-Muñoz EM, Lope V, et al. Calorie intake, olive oil consumption and mammographic density among Spanish women. *Int J Cancer* 2014; **134**: 1916–1925.
23. McDonald JH. *Handbook of biological statistics*, 3rd ed. Baltimore, MD: Sparky House Publishing, 2014.
24. Singh D, Febbo P, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002; **1**: 302–309.
25. Irizarry RA, Bolstad BM, Collin F, et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003; **31**: 4–15.
26. Nalpas NC, Park SDE, Magee DA, et al. Whole-transcriptome, high-throughput RNA sequence analysis of the bovine macrophage response to Mycobacterium bovis infection in vitro. *BMC Genomics* 2013; **14**: 230.
27. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013; **14**: R95.
28. McCarthy DJ, Chen Y and Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 2012; **40**: 4288–4297.
29. Robinson MD and Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 2007; **23**: 2881–2887.
30. Robinson MD and Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 2008; **9**: 321–332.
31. Trafimow D and Marks M. Editorial board EOV. *Basic Appl Soc Psychol* 2014; **36**: 1–2.
32. Genovese C and Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. *J Roy Stat Soc B* 2002; **64**: 499–518.

Appendix I

In this appendix, we show that the proposed procedure is asymptotically consistent for sample size n and the number of tests m in the sense that FDR is negligible for large enough n and m.

First, assume that the original set $\mathcal{H}$ is partitioned into $\mathcal{H}_1$ and $\mathcal{H}_0$ which are the sets of true nulls and alternative hypotheses, respectively (e.g. $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_0$ and $\mathcal{H}_1 \cap \mathcal{H}_0 = \emptyset$) of cardinality $m_1 = m - m_0$ and $m_{0,}$ respectively, with $m_1 \ll m_0$. The following Proposition assures asymptotic consistency of p_i with respect to the m, n under $\mathcal{H}_0$ and $\mathcal{H}_1$.

Proposition 1: For $m, n \to \infty$ and $i \in \mathcal{H}_0$, then $p_i \to 0$, while if $i \in \mathcal{H}_1$, then $p_i \to 1$.

Proof. The proof is based on the well known BFs consistency (see e.g. Berger and Pericchi[9]), for $n \to \infty$, that is if $i \in \mathcal{H}_0$ then $cB_i = \frac{m_{i1}(\mathbf{x}_i)}{m_{i0}(\mathbf{x}_i)} \to 0$, while if $i \in \mathcal{H}_1$ then $cB_i = \frac{m_{i1}(\mathbf{x}_i)}{m_{i0}(\mathbf{x}_i)} \to \infty$ Therefore, by (3) $p_i = \frac{B_i}{B_i + \sum_{j \in \mathcal{H} \setminus i}^{m} B_i} \to 0$ for $m_0 < m$. For $m_0 = m$ and because of the fact that evidence is evaluated under the same model $f(\cdot)$ and/or same calibration scale (e.g., calibration formula (1)) if $i \in \mathcal{H}_0$ $p_i \to 1/m$ and finally for $m \to \infty$ $p_i \to 0$. If $i \in \mathcal{H}_1$ by (3) $p_i = \frac{B_i}{B_i + \sum_{j \in \mathcal{H} \setminus i}^{m} B_i} \to 1$ for $m_1 = 1$ while for $m_1 > 1$ for the same argument as above, that the evidence is evaluated under the same model $f(\cdot)$ and/or same calibration scale $p_i \to (m - m_0)/m$ and finally for $m \to \infty$ $p_i \to 1$ under $i \in \mathcal{H}_1$. □

The simple rule of declaring as discoveries all $p_i > 1/m$ controls the FWER and hence the FDR.[16] To show this, let $\pi(p_i)$ be the proposed prior of p_i and $\pi(p_i|Data)$ be the posterior distribution of p_i conditioned to the observed evidence (either from BFs or p-values.). Under the insufficient reason principle $\pi(p_i) = 1/m$, while the consequence of Proposition 1 implies that $\Pr_{\pi(p_i|Data)}(p_i > 1/m) \to 0$ for $i \in \mathcal{H}_0$ and $\Pr_{\pi(p_i|Data)}(p_i > 1/m) \to 1$ for $i \in \mathcal{H}_1$ when $m, n \to \infty$.

The following corollary, necessary to prove the next propositions, clarifies that under the set of null hypotheses and a posteriori the probability of having $p_i > 1/m$ is asymptotically negligible.

Corollary 1: For $m, n \to \infty$ and for any set of $i \in \mathcal{H}_0$ then $\Pr_{\pi(p_i|Data)}(\cap_{i \in \mathcal{H}_0}\{p_i > 1/m\}) \to 0$.

Proof. In fact, $\Pr_{\pi(p_i|Data)}(\cap_{i \in \mathcal{H}_0}\{p_i > 1/m\}) \leq \sum_{i \in \mathcal{H}_0} \Pr_{\pi(p_i|Data)}\{p_i > 1/m\} \to 0$ as the upper bound tends to 0 because of Proposition 1, which refers to tests under $\mathcal{H}_0$ and $\mathcal{H}_1$, where $\sum_{i \in \mathcal{H}_0} \Pr_{\pi(p_i|Data)}\{p_i > 1/m\} = 1 - \sum_{i \in \mathcal{H}_1} \Pr_{\pi(p_i|Data)}\{p_i > 1/m\}$, with $\sum_{i \in \mathcal{H}_1} \Pr_{\pi(p_i|Data)}$ $\{p_i > 1/m\} \to 1$. □

Proposition 2: For $m, n \to \infty$ the simple rule in (4) controls the FWER a priori and a posteriori.

Proof. By definition we have

$$\begin{aligned} FWER &= \Pr_G\left(\cup_{\mathcal{H}_0}\{p_i > 1/m\}\right) \\ &\leq \sum_{\mathcal{H}_0} \Pr_G(p_i > 1/m), \text{ by Boole Inequality,} \end{aligned}$$

where G is the probability measure on the random variables p_i. If $G = \pi(p_i)$ then $FWER = 0$ for every m and n and the a priori control is assured. For *a posterior* control, once evidence has been collected then $G = \pi(p_i|Data)$ and by Proposition 1 $FWER \to 0$ for $m, n \to \infty$ because probabilities are calculated under the null set $\mathcal{H}_0$. □

The following Proposition shows that it is even possible to be more explicit for the control of FDR, although that of the FWER would be enough to bound the FDR.

Proposition 3: For $m, n \to \infty$ the simple rule in (4) controls the FDR.

Proof. By definition of FDR in Benjamini and Hochberg[16] with expectation calculated under $\pi(p_i|Data)$ we have, using Proposition 1

$$\begin{aligned} FDR &= E\left[\frac{\#FalseDiscoveries}{\#Alldiscoveries}\right] \\ &= E\left[\frac{\sum_{\mathcal{H}_0} \mathbb{I}\{p_i > 1/m\}}{\sum_{\mathcal{H}} \mathbb{I}\{p_i > 1/m\}}\right] \\ &= E\left(\sum_{\mathcal{H}_0} \mathbb{I}\{p_i > 1/m\} | \sum_{\mathcal{H}} \mathbb{I}\{p_i > 1/m\} = r\right)/r, \ \text{see (16, pag. 292)} \\ &= \sum \Pr_{\mathcal{H}_0}\left(p_i > 1/m | \sum_{\mathcal{H}} \mathbb{I}\{p_i > 1/m\} = r\right)/r \\ &\to 0, \ \text{for any r} > 0 \text{ and n, m} \to \infty. \end{aligned}$$

□