Bertolino, F., Cabras, S., Castellanos, M. E. & Racugno, W. (2015). Unscaled Bayes factors for multiple hypothesis testing in microarray experiments. *Statistical Methods in Medical Research, 24*(6), pp. 1030-1043.

DOI: 10.1177%2F0962280212437827

Unscaled Bayes factors for multiple hypothesis testing in microarray experiments

Francesco Bertolino,[1] Stefano Cabras,[1] Maria Eugenia Castellanos[2] and Walter Racugno[1]

Abstract

Multiple hypothesis testing collects a series of techniques usually based on *p*-values as a summary of the available evidence from many statistical tests. In hypothesis testing, under a Bayesian perspective, the evidence for a specified hypothesis against an alternative, conditionally on data, is given by the Bayes factor. In this study, we approach multiple hypothesis testing based on both Bayes factors and *p*-values, regarding multiple hypothesis testing as a multiple model selection problem. To obtain the Bayes factors we assume default priors that are typically improper. In this case, the Bayes factor is usually undetermined due to the ratio of prior pseudo-constants. We show that ignoring prior pseudo-constants leads to unscaled Bayes factor which do not invalidate the inferential procedure in multiple hypothesis testing, because they are used within a comparative scheme. In fact, using partial information from the *p*-values, we are able to approximate the sampling null distribution of the unscaled Bayes factor and use it within Efron's multiple testing procedure. The simulation study suggests that under normal sampling model and even with small sample sizes, our approach provides false positive and false negative proportions that are less than other common multiple hypothesis testing approaches based only on *p*-values. The proposed procedure is illustrated in two simulation studies, and the advantages of its use are showed in the analysis of two microarray experiments.

Keywords

false discovery rate, improper priors, local false discovery rate

1 Introduction

The analysis of microarray experiments is a typical statistical problem that involves multiple hypotheses testing (MHT). We perform a large number, say m, of gene expression comparisons across two biological populations by testing a corresponding number of statistical hypothesis using a sample size much smaller than m.

[1]Department of Mathematics and Informatics, University of Cagliari, via Ospedale 72, Cagliari, Italy
[2]Department of Statistics and Operations Research, Rey Juan Carlos University, c/Tulipan sn, Mostoles, Spain

Corresponding author:
Stefano Cabras, Department of Mathematics and Informatics, University of Cagliari, via Ospedale 72, 09124 Cagliari, Italy.
Email: s.cabras@unica.it

Most of the current methods, used in MHT literature, evaluate the joint evidence against m null hypotheses with the objective of guaranteeing an upper bound on the number of false positives (false null rejections or false discoveries) while maintaining also a low number of false negatives (missed null rejections or false non-rejections). Essentially, these procedures guarantee an upper bound to several type I error rates (per comparison, per family, family wise and false discovery rates) along with a reasonable power.

The most commonly known and used MHT methods are based on the evidence provided by suitable test statistics through the corresponding p-values. 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 $U(0, 1)$. This could be a reason why p-values are so popular in MHT. However, outside of the situations where the $U(0, 1)$ can be used, the use of p-values is problematic as shown in Cabras.[1] To this purpose, Efron[2,3] proposed a MHT procedure that estimates the unknown theoretical sampling null distribution of the p-values which differs from $U(0, 1)$. This procedure, later referred as Efron's procedure, plays an important role in this study.

A broad review of MHT literature is beyond the scope of this study and we invite the reader to look at Dudoit et al.[4] and Farcomeni,[5] with the references therein, for specific problematics related to microarray experiments.

The approach to MHT proposed here jointly uses two sources of evidence for each test as done in Perelman et al.,[6] where t-statistics and their corresponding pooled standard deviations are jointly used to detect genes that are differentially expressed (DE). Our study is also located among those that explicitly measure the evidence of the null and the alternative hypothesis as done, for instance, in Moerkerke and Goetghebeur.[7]

Specifically, we look at the MHT as a multiple model selection problem that can be handled with m Bayes factors (BFs in the sequel). These are regarded as a measure of evidence for the model choice that is considered in each test. We concentrate on the use of BF in a default Bayesian setting where the two prior distributions, one for each model under comparison, come from a formal rule. Such priors are typically improper and so the BF, for single hypothesis testing, is undetermined due to the ratio among prior *pseudo*-constants c_1/c_0 (c_0 for the null model, H_0, and c_1 for the alternative, H_1). In order to eliminate the arbitrariness on c_1/c_0, several approaches are considered in literature, such as those in Moreno et al.,[8] Bertolino et al.,[9] Berger and Pericchi[10] and O'Hagan and Forster.[11] Unfortunately, these methods require large samples and their application would be extremely problematic in the analysis of microarray experiments. In fact, m approximations of the full BF would be computationally unfeasible according to the procedures described in Berger and Pericchi[10] and O'Hagan and Forster.[11] It is important to note here that, in the absence of strong prior information for each test, as the case of microarray data analysis, the use of proper 'vague priors' can be problematic because of the well-known Lindley's paradox. Therefore, even the full and well-defined BFs are essentially useless in MHT.

In this article, we ignore c_1/c_0 leading to what we call *unscaled* BFs and we argue that they can be used in MHT with a partial calibration by the corresponding m p-values. Such calibration is only needed in order to approximate their sampling null distribution. The use of the sampling distribution of BFs is not new in literature; recently, it has been used for estimating clusters[12] and Carota[13] uses the sampling distribution of BFs in the context of robustness, while in Pauler et al.,[14] the sampling behaviour of BFs is compared with other frequentist measures. Finally, Sellke et al.[15] analyses the sampling distribution of p-values in connection with BFs in hypothesis testing. Even under simple parametric set-ups, we do not know the sampling null distribution of the unscaled BFs. We estimate it using a data dependent parametric bootstrap resampling scheme, where the parametric model

corresponds to the null model H_0. The partial calibration we propose is needed in order to perform such parametric bootstrap because we estimate the nuisance parameters of the null model using the observations from all tests whose p-values are above a suitable threshold $\widetilde{p}$. We use p-values only as a starting point of our MHT procedure which is mainly based on unscaled BFs. In fact, we cannot appreciate how many of the observed BFs are likely to come from the null model. However, if $\widetilde{p}$ is sufficiently high, we can reasonably assume that the corresponding p-values come from the null model. We fix $\widetilde{p}$ in such a way that, whatever MHT procedure one is willing to use, he/she would never reject hypotheses for which $p > \widetilde{p}$.

The estimation of BF's null distribution can be considered accurate enough because a large number of observed BFs under the null hypothesis is available. We then compare observed BFs with those expected under its approximated sampling null distribution by means of Efron's procedure. With such procedure, it is possible to estimate the set of true null hypotheses. In particular, the Efron's procedure uses a test statistic Z that quantifies evidence against H_0. It assumes that $z_1, \ldots, z_m$ come from the mixture distribution

$$h(z) = \frac{m_0}{m} h_0(z) + \frac{m - m_0}{m} h_1(z)$$

where $h_0(z)$ and $h_1(z)$ are densities of Z under H_0 and H_1, respectively. In order to set a cut-off on the Z values, the quantity to be controlled is the *local false discovery rate*

$$lfdr(z) \equiv \Pr\{H_0 | Z = z\} = \frac{m_0}{m} \frac{h_0(z)}{h(z)} \leq \frac{h_0(z)}{h(z)}$$

In this article, we assume $h_0(z) = \phi(z)$ (the standard normal density) when $z_1, \ldots, z_m$ are obtained using the parametric bootstrap described in Section 3 and equation (6). Density $h(z)$ is estimated according to the method discussed in Efron[2] (Section 3, equations (3.3) to (3.6)). As in Efron,[2] we assume $m_0/m \geq 0.9$, where m_0 is the unknown number of true nulls. Such assumption on the proportion of DE genes is usual in microarray data analysis.

The *lfdr* is related to the false discovery rate (FDR) controlled by the Benjamini and Hochberg[16] procedure. Efron's[2] procedure also reports an estimation of the expected FDR, *Efdr*, which is a measure of the power of the method, that is, the smaller the *Efdr* the more confident we are in the estimated sets of null and nonnull hypotheses.

The article is organized as follows: in Section 2, we introduce the unscaled BF and illustrate its calculus for a toy example (Section 2.1). We also consider the more realistic problem of assessing the equality of means of two independent heteroscedastic normal populations[17] (Section 2.2). The heteroscedasticity assumption is more realistic than the homoscedasticity in microarray data analysis, because the variance of gene expression levels is often related with their means as argued in Chen et al.,[18] Newton et al.[19] and Perelman et al.[6] Section 3 explains the MHT procedure based on BF and provides the estimated BF's null distribution. Section 4 contains two simulation studies where we illustrate, within the Efron's procedure, the advantages of the use of BF over p-value. Section 5 presents an application to two microarray experiments. Finally, Section 6 contains remarks and a discussion of the benefits that an MHT procedure would have if it were based on the joint combination of the information from p-values and BFs. However, the latter approach is beyond the scope of this article.

2 Default improper priors and unscaled BFs

Let $\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\ \boldsymbol{x}_m)$ be a realization of experiments each with m different features, i.e. m gene expressions. The vector $\boldsymbol{x}_i$ contains n_i replications corresponding to the ith experimental feature, for $i=1,\ldots,\ m$. For the sake of simplicity in the notation, we assume $n_i=n$ with $n\ll m$. In fact, different sample sizes neither modify the definition of the BF nor complicate calculations. Different sample sizes could be of interest in applications with missing data in one or more arrays, but this would just lead to a complication in the adopted notation. The joint sampling distribution of all m features is assumed to be unknown and cannot be accurately estimated from the observed $\boldsymbol{x}$.

We regard the MHT problem as a multiple model selection problem formalized as follows

$$\begin{cases}\mathcal{M}_{i0}: f_{i0}(\boldsymbol{x}_i \mid \theta_{i0}),\ \pi_{i0}(\theta_{i0}),\ \theta_{i0}\in\Theta_{i0}\\ \mathcal{M}_{i1}: f_{i1}(\boldsymbol{x}_i \mid \theta_{i1}),\ \pi_{i1}(\theta_{i1}),\ \theta_{i1}\in\Theta_{i1}\end{cases}\quad i=1,\ldots,m$$

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

$$\begin{cases}\pi_{i0}(\theta_{i0})=\pi_0(\theta_0)\propto c_0 g_0(\theta_0)\\ \pi_{i1}(\theta_{i1})=\pi_1(\theta_1)\propto c_1 g_1(\theta_1)\end{cases}\tag{1}$$

where c_0 and c_1 are the normalizing *pseudo*-constants and $g_0(\theta_0)$, $g_1(\theta_1)$ non-integrable functions. The use of default priors avoids the very difficult task of elicitation on all parameters for all tests.

Prior predictive distributions for null and alternative hypotheses are

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

The BF of $\mathcal{M}_{i1}$ against $\mathcal{M}_{i0}$ is

$$eBF_i=\frac{m_{i1}(\boldsymbol{x}_i)}{m_{i0}(\boldsymbol{x}_i)}=\frac{c_1}{c_0}\cdot\frac{\int_{\theta_1\in\Theta_1} g_1(\theta_1)f_1(\boldsymbol{x}_i\mid\theta_1)\mathrm{d}\theta_1}{\int_{\theta_0\in\Theta_0} g_0(\theta_0)f_0(\boldsymbol{x}_i\mid\theta_0)\mathrm{d}\theta_0}\tag{2}$$

which is unscaled because of the arbitrary ratio c_1/c_0, see for instance O'Hagan and Forster,[11] (para. 7.54). There are several proposals to avoid the arbitrariness of the ratio c_1/c_0 in equation (2), as the fractional and intrinsic BFs.[11] Unfortunately, in our problem, these proposals are unfeasible because of the small sample size. We define the *unscaled* BF as

$$BF_i=\frac{\int_{\theta_1\in\Theta_1} g_1(\theta_1)f_1(\boldsymbol{x}_i\mid\theta_1)\mathrm{d}\theta_1}{\int_{\theta_0\in\Theta_0} g_0(\theta_0)f_0(\boldsymbol{x}_i\mid\theta_0)\mathrm{d}\theta_0}\tag{3}$$

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

$$\frac{\frac{c_1}{c_0}BF_i}{\frac{c_1}{c_0}BF_{i'}}=\frac{BF_i}{BF_{i'}}>1,\ \forall i,i'$$

the evidence in favour of $\mathcal{M}_{i1}$ versus $\mathcal{M}_{i0}$ is larger than that of $\mathcal{M}_{i'1}$ versus $\mathcal{M}_{i'0}$ whatever the ratio $c_1/c_0 > 0$. For numerical convenience we use, instead of the BF_i, the well-known weight of evidence

$$W_i = \log(BF_i)$$

where distances among observed W, $\boldsymbol{w} = \{w_i\}_{i=1}^m$, are invariant with respect to $\log(c_1/c_0)$. This is because $\log(c_1/c_0)$ is the same for each hypothesis. This statement is true in general when prior distributions in equation (1) are the same along all m experiments. A proof of this is reported in the Appendix for Example 1 (next section), indicating also how it could be generalized.

2.1 Example 1: testing zero normal means with unknown variance

We illustrate the proposed method using the following toy example. 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 to test

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

Sufficient statistics are $\overline{X}_i = 1/n \sum_{j=1}^n X_{ij}$ and $S_i^2 = 1/n \sum_{j=1}^n (X_{ij} - \overline{X}_i)^2$, whose observed values are denoted by $\overline{x}_i$ 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 $eBF_i = c_1/c_0 BF_i$, where

$$BF_i = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \sqrt{\pi S_i^2} \quad 1 + \frac{\overline{X}_i^2}{S_i^2} \Bigg)^{n/2} \tag{4}$$

is the unscaled BF.

2.2 Example 2: testing the equality of normal means

Different from the above example, we consider a parametric test which is more realistic in applications to microarray data. Suppose the usual two-group study, with m features and denote with $\boldsymbol{x}_{m \times n_x}$ the outcome in group X with n_x replications and $\boldsymbol{y}_{m \times n_y}$ the outcome in group Y with n_y replications. Let $X_i \sim N(\mu_{X_i}, \sigma_{X_i}^2)$ and $Y_i \sim N(\mu_{Y_i}, \sigma_{Y_i}^2)$ for $i = 1, 2, \ldots, m$. The set of hypotheses, for $\sigma_{X_i}^2 > 0$, $\sigma_{Y_i}^2 > 0$ unknown, is as follows.

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

With the usual default priors

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

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

$$BF_i = \frac{B(\frac{n_x-1}{2}, \frac{1}{2})B(\frac{n_y-1}{2}, \frac{1}{2})\sqrt{S^2_{X_i}S^2_{Y_i}}}{\int_{\mu_i\in\mathbb{R}}\left(1+(\overline{X}_i-\mu_i)^2/S^2_{X_i}\right)^{-\frac{1}{2}n_x}\left(1+(\overline{Y}_i-\mu_i)^2/S^2_{Y_i}\right)^{-\frac{1}{2}n_y}\mathrm{d}\mu_i} \tag{5}$$

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

3 The empirical null distribution of the *unscaled* BF

In this section, we describe the procedure to approximate the empirical null distribution of the unscaled BF. This distribution is employed to obtain values $z_1, \ldots, z_m$ to be used in the Efron's procedure. First, we associate to each test i the observed w_i and the p-value, p_i. Then, for a fixed threshold $\widetilde{p}$, we consider the set $\mathcal{I}_0 \equiv \{i : p_i > \widetilde{p}\}$ of hypotheses that are assumed to come from the corresponding H_{0i} and that they would have never been rejected from an MHT procedure when m is of the order of thousands. For this reason, in our applications we consider $\widetilde{p}$ to be between 0.1 and 0.2. For all $i \in \mathcal{I}_0$ we resample b times the unscaled BF according to the parametric null models $\mathcal{M}_{0i}$, as specified in Sections 3.1 and 3.2. Let $\widetilde{m}_0$ be the cardinality of $\mathcal{I}_0$, under the assumption that $\widetilde{m}_0 < m_0$; we then have a total of $B = b \cdot \widetilde{m}_0$ bootstrap draws of unscaled BFs under the null hypothesis. Therefore, the corresponding vector $\widetilde{\boldsymbol{w}} = \{\widetilde{w}_1, \widetilde{w}_2, \ldots, \widetilde{w}_B\}$ is assumed to be generated under the null model. The unknown null distribution of W, say $Q_0(w)$, is approximated by the empirical distribution of the bootstrap sample $\widetilde{\boldsymbol{w}}$ and denoted by $\widetilde{Q}_0$. Finally, in order to apply Efron's procedure, we need to transform the observed $\boldsymbol{w}$, into realizations from a normal distribution

$$z_i = \Phi^{-1}\left(\widetilde{Q}_0(w_i)\right), i = 1, 2, \ldots, m \tag{6}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution at point x. In this way, the null distribution of z_i values is supposed to be the standard normal distribution if all hypotheses were true null hypotheses. On the observed z_i, we apply Efron's procedure obtaining the set of null hypotheses that are supposed to be false along with an estimation of *Efdr* over the set of all rejected null hypotheses.

Note that the transformation rule in equation (6) is invariant under location of $\boldsymbol{w}$ and $\widetilde{\boldsymbol{w}}$; so, the choice of $c_1/c_0 = 1$, is not critical and it does not affect the final result. In this way, the unscaled BFs are useful in MHT under Efron's procedure. For the sake of comparison, we also apply Efron's procedure to the p-values. In particular, we use the usual transformation rule of p_i, $i = 1,\ldots, m$, to a normal random variable by means of $\Phi^{-1}(p_i)$. Note that, unlike the unscaled BF, the transformation $\Phi^{-1}(p_i)$ is strictly related to the assumption of uniformity of the p-values under the null model. This assumption is true for Example 1 because the test statistic is ancillary, while for Example 2 it is true only asymptotically.

3.1 Example 1 (continued)

The approximation of the sampling null distribution for the BF_i in equation (4), is obtained by drawing samples, of size n, of $X_i \sim N\left(0, \frac{n}{n-1}s_i^2\right)$ for each $i \in \mathcal{I}_0$. In each draw, we calculate the BF in equation (4) and the corresponding $\widetilde{w}_i$.

The p-value for this test is the usual p-value from the *Student t*-test.

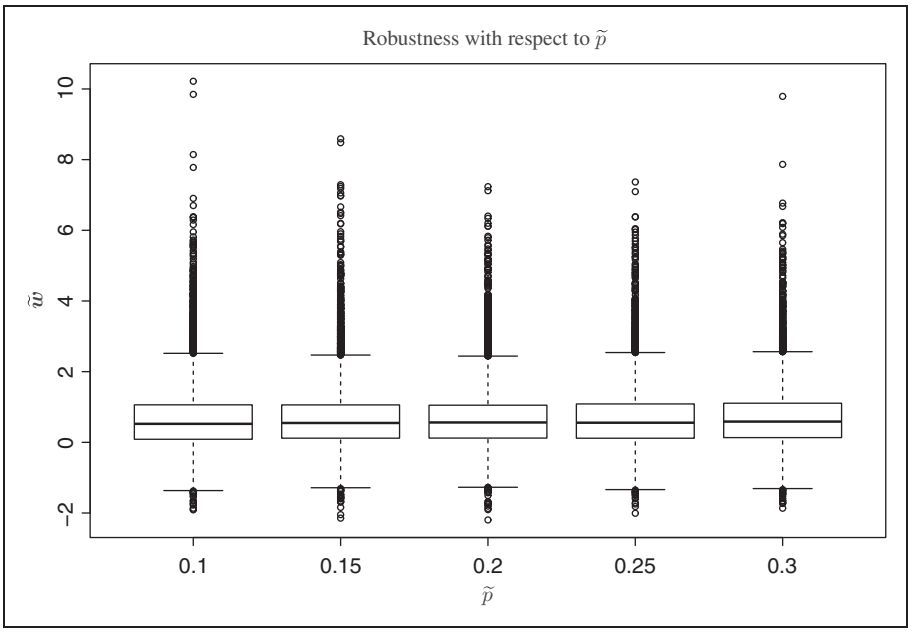

Figure 1. For a simulated dataset corresponding to the toy example, with $m = 5000$, $m_0/m = 0.9$, $\sigma_i = 1$ and $\mu_i = 1 \mid H_{1i}$ boxplots represent the resampling distribution $\widetilde{w}$ for different values of $\widetilde{p}$.

Suppose $m = 5000$, $n = 5$, $m_0/m = 0.9$, $\sigma_i = 1$ and $\mu_i = 1$ as an alternative. Figure 1 shows that the bootstrap approximations of the null distribution of W are fairly robust to $\widetilde{p}$. This result also applies to all parametric set-ups and examples treated in this article. The behaviour of $\left(\widetilde{Q}_0 \mid \widetilde{p}\right)$ is due to the fact that the proportion $m_0/m \approx 1$ in simulated data and also in many real situations.

3.2 Example 2 (continued)

In order to obtain the null sampling distribution for BF_i in equation (5), we consider the parametric bootstrap resampling scheme under the ith null hypothesis for each $i \in \mathcal{I}_0$. This consists of a draw of size n_x of $X_i \sim N(\tilde{\mu}_i, \frac{n_x}{n_x-1} s^2_{x_i})$ and a draw of size n_y of $Y_i \sim N(\tilde{\mu}_i, \frac{n_y}{n_y-1} s^2_{y_i})$, where $\tilde{\mu}_i = \frac{n_x \bar{x}_i + n_y \bar{y}_i}{n_x + n_y}$ is the common mean. For each draw of X_i and Y_i, we calculate the BF in equation (5) and the corresponding $\widetilde{w}$.

In this case, we use p-values from the *Student* t-tests with the Welch correction for $\sigma^2_{X_i} \neq \sigma^2_{Y_i}$. Such correction, however, guarantees only asymptotic uniformity of the p-values under the null model.

4 Simulation study

We describe in this section two simulation studies: the first for the toy example in Section 2.1 and the second for the more general case discussed in Section 2.2. Results cast evidence on the fact that Efron's procedure works better with BFs than p-values alone.

In the simulation study, we consider various scenarios given by the following combinations of sample sizes: $n = 5$, 10; thresholds $\tilde{p} = 0.1, 0.2$; proportions of true nulls: 95% and 99%; means and variances under the alternative hypotheses: (a) $\mu = 2$, $\sigma = 1$, (b) $\mu = 3$, $\sigma = 1$ and (c) $\mu = \sigma = 4$. Note

that combination (3) corresponds to a very weak signal as the variation coefficient equals 1. We compare the use of W against the p-value in 250 simulations each of $m = 5000$ tests, by looking at the the amount of FDR, false non-rejection rate (FNR) and the total error (Tot) given by FDR + FNR. In all simulations, the *local false discovery rate*, *lfdr*, used as threshold in Efron's procedure is 0.2.[2] This means that each rejected null hypothesis has a probability less than 20% to be a false discovery. In this sense, this error rate is local to each test, while the expected error, for the whole set of rejected hypotheses, is estimated by the *Efdr*.

Our first finding was that results are robust with respect to the choice of $\tilde{p}$ and the proportion of true null hypotheses. Therefore, Figure 2 shows only a subset of results that are representative of the whole simulation study. Each boxplot represents the empirical distribution of FDR, FNR and Tot in 250 simulations under the three scenarios and two sample sizes where 95% of $m = 5000$ null hypothesis are true null and $\tilde{p} = 0.1$. The amount of experimental signal is smaller in the first row ($n = 5$) than in the second one ($n = 10$). By columns, the amount of signal is larger in the second column than in the first one. Finally, the signal is very weak in the last column where the variation coefficient is 1. This last situation is relevant in microarray data where the signal grows with noise.

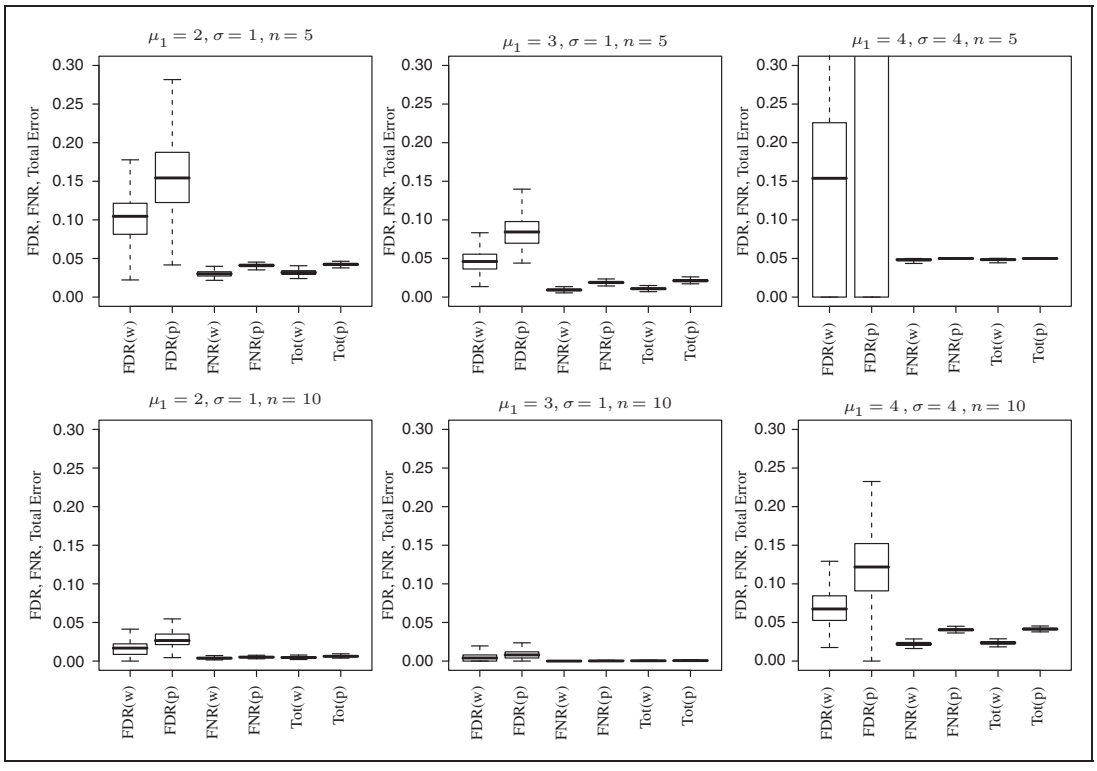

Figure 2. Simulation results for comparison between the weight of evidence (*w*) and the *p*-value (*p*) within Efron's procedure. The comparison is made in terms of false discovery rate (FDR), false non rejection rate (FNR) and total error (Tot).
FDR: false discovery rate; FNR: false non-rejection rate.

In particular, Figure 2 shows that error rates are significantly smaller when using W instead of p-value. This result is persistent in all considered scenarios, even when the error rates are large, as scenario (c).

We propose the above study for the case of two populations, X and Y, with different variances. We assume $Y_i \sim N(0, 1)$ for all i and $\mu_{X_i} = 0$ and $\sigma_{X_i} = 2$ for population X under the null. We consider three scenarios under the alternative: (a) $\mu_{X_i} = 4, \sigma_{X_i} = 2$, (b) $\mu_{X_i} = 3, \sigma_{X_i} = 2$ and (c) $\mu_{X_i} = 4, \sigma_{X_i} = 4$. The rest of the parameters are identical to those used in the above simulation study. Figure 3 provides the simulated error rates in a layout similar to that of Figure 2. Error rates with w are generally smaller than those with p, and thus, the conclusions are compatible to those obtained with the toy example.

5 Application to real datasets

In this section, we consider the application of the proposed procedure to two microarray experiments. The first is a calibration experiment where the true DE genes are known, while the second is a larger study also analysed, with different approaches, in Singh et al.[20] and Efron.[2]

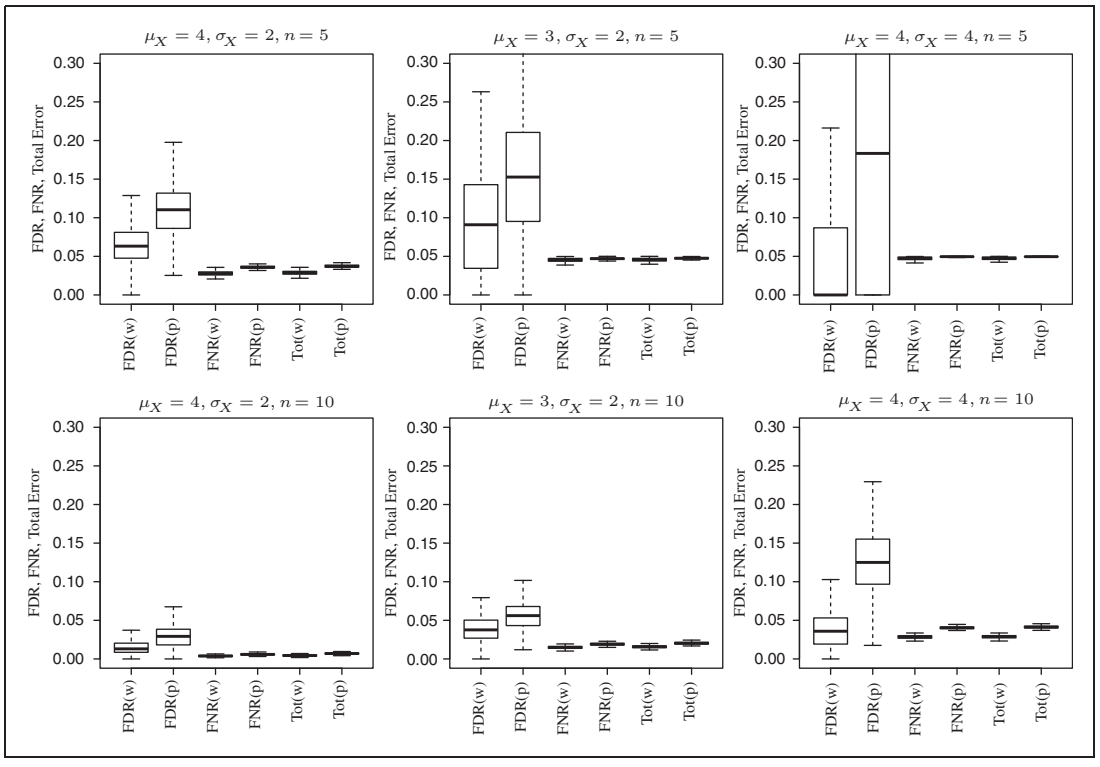

Figure 3. For the case of two populations with different variances, we compare results between the weight of evidence (w) and the p-value (p) within Efron's procedure. The comparison is made in terms of false discovery rate (FDR), false non rejection rate (FNR) and total error (Tot).
FDR: false discovery rate; FNR: false non-rejection rate.

5.1 Microarray-controlled experiment

We compare the results obtained using unscaled BF and p-values when analysing gene expression levels of Affymetrix HGU95A Latin square dataset (`http://www.affymetrix.com`). Here, $m = 12626$, and 16 genes have been spiked at controlled levels ranging from 0 to 1024 pM, as presented in Table 1.

The number of replications for X and Y are $n_x = n_y = 5$. In this case, genes `1597_at` and `38734_at` are the less DE, while gene `684_at` is the most DE because it is absent from population Y and it has the highest concentration in population X. We analyse expression level data obtained from summaries of probe level pairs in the $\log_2$ scale according to the procedure in Irizarry.[21]

Figure 4 (top) shows the results of the analysis with the procedure here proposed, while the bottom the results using the p-values alone in Efron and Benjamini and Hochberg[16] procedures. Histograms refer to observed w (top) and observed P (bottom), while continuous density is $\widetilde{Q}_0$ approximated with $b = 5$ and $\tilde{p} = 0.2$. As expected, this density fits quite well with most of the observed w.

The main discrepancy between results obtained with W and P is that genes of Table 1 are mostly separated in the set $\boldsymbol{w}$ rather than in that of the p-values. This may be ascribed to the fact that we are evaluating the evidence of the null model and the alternative model. In fact, p-value does not provide an evaluation of the evidence for an alternative model.

This larger separation improves the power of Efron's procedure. In fact, using W, we have 12 discoveries of which 3 are false, while using P, we have 3 discoveries and none of them are false. Using the Benjamini and Hochberg procedure, we have 8 discoveries with 2 of them false. Looking at the number of false negatives, we may see that with W it decreases to 7 against 13 for Efron's procedure based on p-values and 10 for the Benjamini and Hochberg procedure. Finally, the reported estimate of $Efdr$ is 20% for W against 28% for P, leading the analyst to be more confident in the set of genes discovered with the proposed procedure.

5.2 Prostate data

We compare unscaled BFs with p-values in the analysis of gene expression levels for prostate cancer data.[20] 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 5. The larger sample size reduces differences between W and P with Efron's procedure as also suggested by the simulation study. In fact, at $lfdr = 0.2$, we have 57 and 53 discoveries with W and P, respectively, with all the 53 discoveries included into the 57 ones. Using the Benjamini and Hochberg procedure at the same level of 20%, we obtain 103 discoveries. Despite the fact that the results of Efron's procedure almost agree using W or P, we may

Table 1. pM concentrations of 16 spiked-in genes in X and Y populations used in this study.

Gene	X (pM)	Y (pM)	Gene	X (pM)	Y (pM)
`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

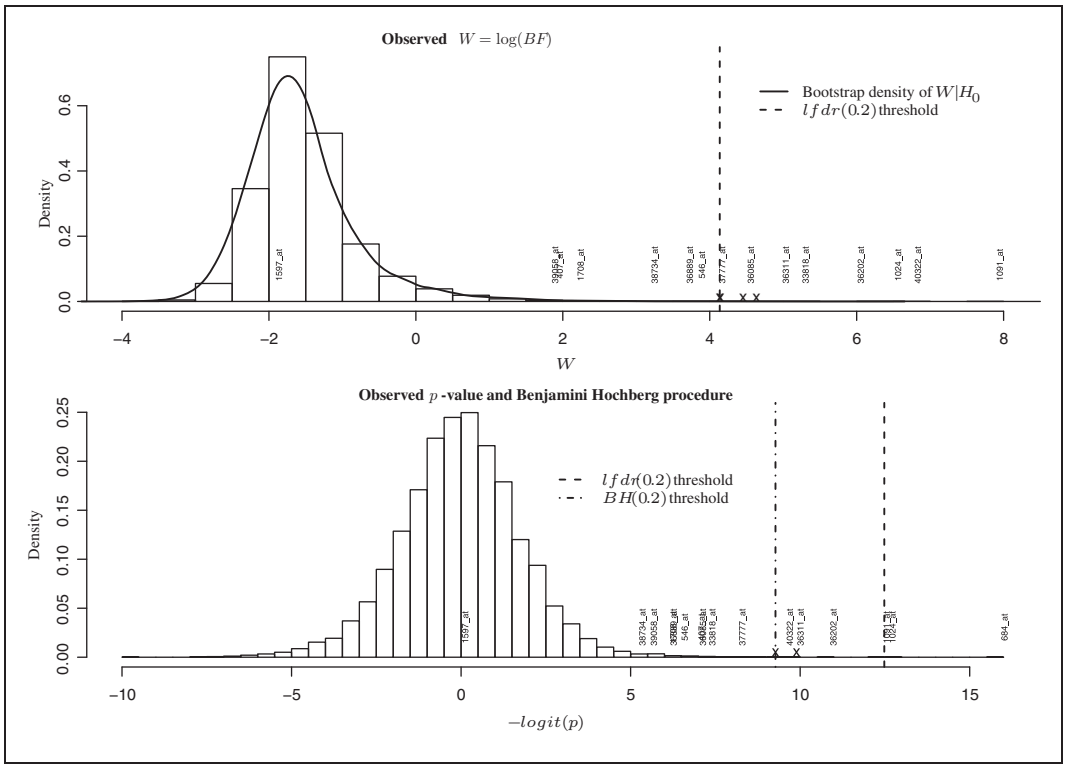

Figure 4. Comparison among *BF* and *p*-value for a controlled microarray experiment. Points indicated with x are false discoveries. In the top figure, gene `684_at` is out of scale with $w \simeq 15$.

still see that the observed w declared as discoveries are further away from the bulk of the null density than the corresponding observed P points. This is also reflected in the reported estimation of *Efdr* because we have 61% for W against 63% for P. We are then slightly more confident in using Efron's procedure with unscaled BFs rather than the same procedure with *p*-values.

6 Conclusions

'Simulation study' and applications to real datasets suggest that unscaled *BF* in MHT is a valuable tool. Under the considered normal sampling models, the weight of evidence, W, is in general more powerful than P, and the simulation study shows that W produces less false discoveries than the *p*-value. This result is particularly emphasized in small samples, provided a correct specification of the sampling model is given. In fact, the null distribution of the BFs relies on correct model specification. In particular, the calculus of $z_1,\ldots, z_m$ is made under independent parametric bootstrap resampling. The missed inclusion of correlation between genes and the presence of random effects could lead to a null distribution, $h_0(z)$, which may differ from the standard normal used here (see Efron[2] and Chapter 6 of Efron[22]). When independence and precise null effects cannot be assumed, it is possible to make use of the empirical null distribution proposed in Efron.[2]

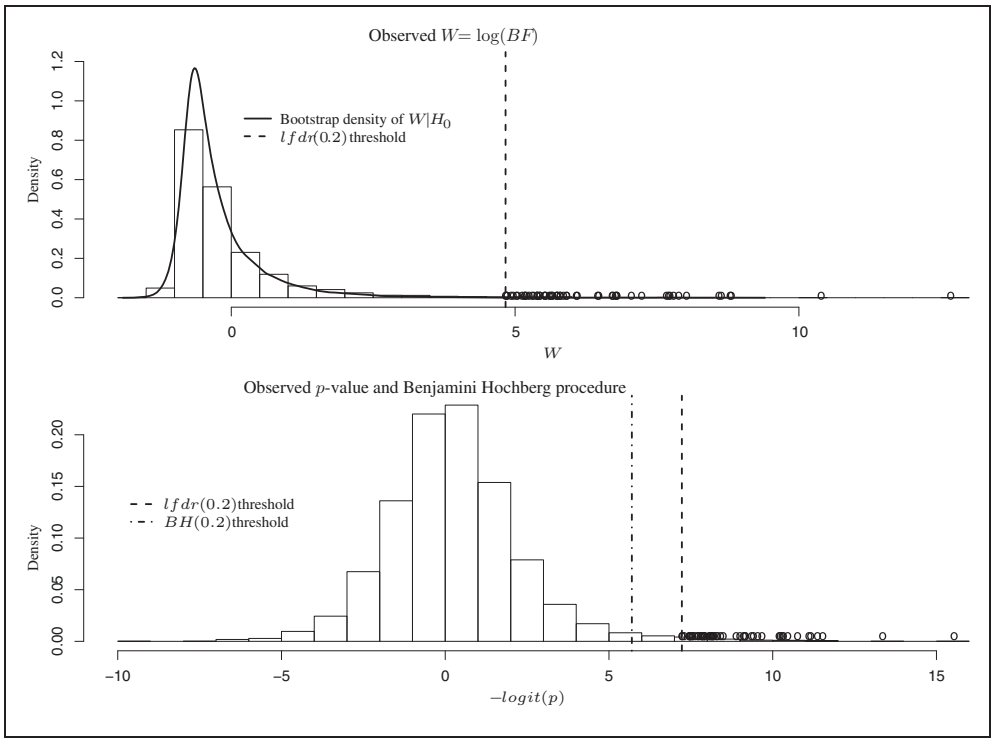

Figure 5. Comparison among unscaled *BF*s and *p*-values for prostate cancer data. Points labelled by 'o' are discoveries. Using the threshold of *lfdr* = 0.2, we have 57 and 53 with *W* and *P*, respectively, while using Benjamini and Hochberg procedure, we have 103 discoveries.

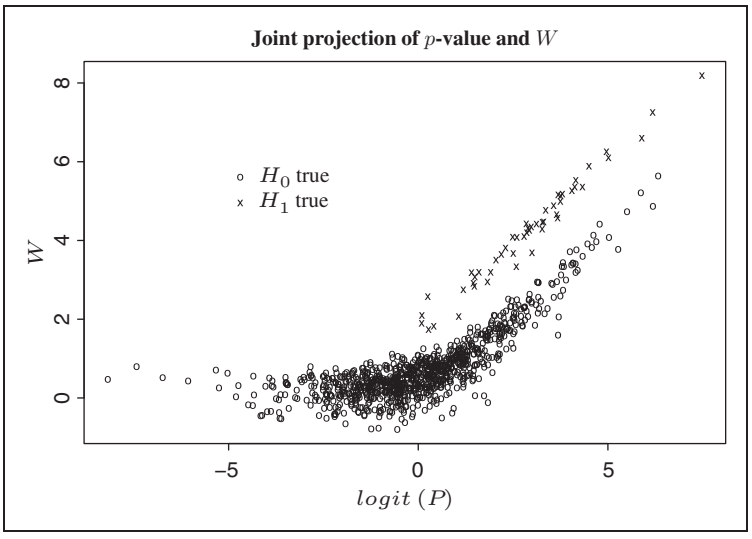

Figure 6. Joint projection of the logit of the *p*-value and weight of evidence *W* for the toy example with $m = 1000$ where 950 tests are true nulls: $X \sim N(0, 1)$ and 50 comes from the alternative $X \sim N(4, 3)$ and $n = 5$.

However, we found that the combined use of W and P would improve MHT procedures. To support this statement, we showed, in Figure 6, the joint values of W and the logit of P for the toy example 1 with $n = 5$, $m = 1000$ tests in which $m_0 = 950$ are true nulls, $X \sim N(0, 1)$, while 50 comes from the alternative $X \sim N(4, 3)$. We may see that in this bivariate space, the cloud of points coming from the alternative hypotheses may clearly be isolated using both dimensions rather than only one. A similar approach is not new in literature,[23] but the combination of such well-known measures of evidence tailed for hypothesis testing has never been considered. The main difficulty for this approach is finding the joint null distribution of W and P. Such null distribution could be embedded in an extension of Efron's procedure applied to the bivariate space induced by W and P as that proposed in Ploner et al.[23]

Funding

Authors F. Bertolino, S. Cabras and W. Racugno were partially supported by the Italian Ministry of Education, University and Research. M.E. Castellanos was partially supported by the Spanish Ministry of Science and Technology, under grant MTM2010-19528, CAM of Spain under grant S2009/esp-1594 and the visiting professor program of Regione Autonoma della Sardegna of Italy.

References

1. Cabras S. A note on multiple testing for composite null hypotheses. *J Stat Plan Inference* 2010; **140**: 659–666.
2. Efron B. Microarrays, Empirical bayes and the two-groups model. *Stat Sci* 2008; **23**(1): 1–22.
3. Efron B. Size, power and false discovery rates. *Ann Stat* 2007; **35**(4): 1351–1377.
4. Dudoit S, Shaffer J and Boldrick J. Multiple hypothesis testing in microarray experiments. *Stat Sci* 2003; **18**: 71–103.
5. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat Meth Med Res* 2008; **17**: 347–388.
6. Perelman E, Ploner A, Calza S, et al. Detecting differential expression in microarray data: comparison of optimal procedures. *BMC Bioinf* 2007; **8**(28): 1–10.
7. Moerkerke B and Goetghebeur E. Selecting "Significant" differentially expressed genes from the combined perspective of the null and the alternative. *J Comput Biol* 2006; **13**(9): 1513–1531.
8. Moreno E, Bertolino F and Racugno W. An intrinsic limit procedure for model selection and hypotheses testing. *J Am Stat Assoc* 1998; **93**: 1451–1460.
9. Bertolino F, Moreno E and Racugno W. Bayesian model selection approach to analysis of variance under heteroscedasticity. *Statistician* 2000; **49**: 503–517.
10. Berger JO and Pericchi LR. The intrinsic bayes factor for model selection and prediction. *J Am Stat Assoc* 1996; **91**: 109–122.
11. O'Hagan A and Forster J. *Kendall's advanced theory of statistics: Bayesian inference*. Vol. 2, New York: Hafner, 2004.
12. Fuentes C and Casella G. Testing for the existence of clusters. *SORT* 2009; **33**(2): 115–146.
13. Carota C. Local robustness of Bayes factors for nonparametric alternatives. *Lect Notes Monogr Ser* 1996; **29**: 283–291.
14. Pauler DK, Wakefield JC and Kass RE. Bayes factors and approximations for variance component models. *J Am Stat Assoc* 1999; **94**(448): 1242–1253.
15. Sellke T, Bayarri MJ and Berger JO. Calibration of p-values for testing precise null hypotheses. *Am Stat* 2001; **55**(1): 62–71.
16. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; **57**: 289–300.
17. Moreno E, Bertolino F and Racugno W. Default Bayesian analysis of the Behrens-Fisher problem. *J Stat Plan Inference* 1999; **81**: 323–333.
18. Chen Y, Dougherty ER and Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt* 1997; **2**: 364–374.
19. Newton MA, Kendziorski CM, Richmond CS, et al. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 2002; **8**: 37–52.
20. Singh D, Febbo P, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002; **1**: 302–309.
21. Irizarry RA, Bolstad BM, Collin F, et al. Summaries of affymetrix genechip probe level data. *Nucleic Acid Res* 2003; **31**: 4–15.
22. Efron B. *Large-scale inference: empirical Bayes methods for estimation, testing and prediction*. IMS Monographs. New York: Cambridge University Press, 2010.
23. Ploner A, Calza S, Gusnanto A, et al. Multidimensional local false discovery rate for microarray studies. *Bioinformatics* 2006; **22**(5): 556–565.

Appendix

In this section, we justify that the ratio of *pseudo*-constants c_1/c_0 in equation (4) does not depend on data. For any hypothesis i, we consider the following priors

$$\pi_0^M(\sigma_i^2) = c_0^M \sigma_i^{-2} \cdot \mathbf{1}_{D_M}(\sigma_i^2),\ c_0^M = (2\log(M))^{-1} \tag{7}$$

$$\pi_1^M(\mu_i, \sigma_i^2) = c_1^M \sigma_i^{-2} \cdot \mathbf{1}_{A_M \times D_M}(\mu_i, \sigma_i^2),\ c_1^M = (4M\log(M))^{-1} \tag{8}$$

where $D_M = (M^{-1}, M)$ and $A_M = (-M, M)$. For instance, the ratio of the two constants is

$$\frac{c_1^M}{c_0^M} = \frac{1}{2M} = O(M^{-1}).$$

For a value of M greater enough such that

$$\int_{\mathbb{R}^+ \setminus D_M} \pi_0^M(\sigma_i^2) f_0(\overline{x}_i, s_i^2 \mid \sigma_i^2) d\sigma_i^2 < \epsilon$$

$$\int_{(\mathbb{R} \setminus A_M) \times (\mathbb{R}^+ \setminus D_M)} \pi_1^M(\mu_i, \sigma_i^2) f_1(\overline{x}_i, s_i^2 \mid \mu_i, \sigma_i^2) d\sigma_i^2 d\mu_i < \epsilon$$

it is trivial to obtain the following result for the eBF_i^M defined using priors in equations (7) and (8):

$$eBF_i^M = \frac{m_1^M(\overline{x}_i, s_i^2)}{m_0^M(\overline{x}_i, s_i^2)} = \delta_i(M) \cdot BF_i$$

where $\delta_i(M) = O(M^{-1})$ and

$$BF_i = \frac{m_1(\overline{x}_i, s_i^2)}{m_0(\overline{x}_i, s_i^2)} = \frac{\int_{\mathbb{R} \times \mathbb{R}^+} \sigma_i^{-2} f_1(\overline{x}_i, s_i^2 \mid \mu_i, \sigma_i^2) \mathrm{d}\sigma_i^2 \mathrm{d}\mu_i}{\int_{\mathbb{R}^+} \sigma_i^{-2} f_0(\overline{x}_i, s_i^2 \mid \sigma_i^2) \mathrm{d}\sigma_i^2}$$

is the *unscaled* BF.

It is very important to note that for any $M \in \mathbb{R}^+$ and for every i, $i' = 1, 2, \ldots, m$, we have that $\delta_i(M)/\delta_{i'}(M) = 1$. That is, the undetermined constant c_1/c_0 does not depend on test index. Note that this proof is generalizable to cases in which normalizing constants c_0^M and c_1^M do not depend on test index i, which is true if priors in equations (7) and (8) have the same functional form and support for all tests.