Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 2108–2124
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

A multivariate version of the Benjamini–Hochberg method

J.A. Ferreira[a,*], S.O. Nyangoma[b]

[a] *Department of Clinical Epidemiology and Biostatistics, VUMC, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands*
[b] *Cancer Research UK Institute for Cancer Studies, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

Received 6 June 2007
Available online 15 February 2008

Abstract

We propose a multivariate method for combining results from independent studies about the same 'large scale' multiple testing problem. The method works asymptotically in the number of hypotheses and consists of applying the Benjamini–Hochberg procedure to the p-values of each study separately by determining the 'individual false discovery rates' which maximize power subject to a restriction on the (global) false discovery rate. We show how to obtain solutions to the associated optimization problem, provide both theoretical and numerical examples, and compare the method with univariate ones.

AMS 2000 subject classifications: 62J15; 62G30; 60F05

Keywords: Multiple testing; Empirical distributions; False discovery rate; Average power

1. Introduction

The development of methods for combining results of independent multiple testing studies has been a subject of interest in recent years, especially in connection with the joint analysis of several microarray experiments (e.g. [13,4]). Typically, a microarray experiment yields gene expression measurements across a very large number of 'genes' under each of a couple of

* Corresponding author.
E-mail addresses: j.ferreira@vumc.nl (J.A. Ferreira), s.o.nyangoma@bham.ac.uk, s.o.nyangoma@amc.uva.nl (S.O. Nyangoma).

0047-259X/$ - see front matter
doi:10.1016/j.jmva.2008.02.013

conditions (e.g. control and treatment), and the main purpose of a statistical analysis of such data is to detect genes whose expression systematically differs under different conditions. Since differences in gene expression are measured by test statistics and formally assessed by means of associated p-values computed under a null hypothesis (e.g. equality of means in the case of the two-sample t-statistic), the analysis of data from a microarray experiment usually reduces to the computation of p-values – one p-value per gene/test statistic/hypothesis – and to the application of the Benjamini–Hochberg [2] method, which declares as 'differentially expressive'/false those genes/hypotheses with p-values below a certain *threshold* or cut-off point. (See, for example, [11].) If we thus regard the result of a multiple testing study as a list of hypotheses rejected on the basis of a sample of p-values, then **a method of combining results from s multiple testing studies** is, by definition, a rule for rejecting hypotheses on the basis of s samples of p-values. In this article we adopt this point of view and introduce a multivariate method for combining the outcomes of different *homologous* multiple testing studies—multiple testing studies that, roughly speaking, address the same set of hypotheses. The method consists of applying a version of the Benjamini–Hochberg procedure to the p-values of each study separately by fixing their thresholds – their *individual false discovery rates* – and using these to construct a list of rejections per study, and then declaring as false those hypotheses which belong to *at least one* of the s lists, in such a way as to maximize *average power* subject to a restriction on the ('global') *false discovery rate* (see Section 2 for the definition of these concepts).

The optimality of our method holds in an asymptotic sense as the number of hypotheses tends to infinity and under rather general conditions (which allow for non-stationarity and dependence in the sequence of p-values); it follows from certain results about an 'adaptive form' of the Benjamini–Hochberg method which we review in Section 2 and require for the exposition of the method in Section 3. Because the determination of the individual false discovery rates amounts to solving numerically an optimization problem with one constraint, part of Section 3 is devoted to the problem of obtaining solutions and to illustrating the workings of the method by means of numerical examples.

A numerical comparison with *univariate* methods is given in Section 4. Univariate methods consist of computing a single p-value or test statistic per hypothesis as a function (e.g. a weighted average) of the homologous p-values or test statistics from the s different studies, and then applying a multiple testing procedure to the resulting sample (e.g. [13,4]). The univariate methods that we consider in Section 4 – the *average method* and *Fisher's method* – are only special cases of the methods of Rhodes et al. [13] and Choi et al. [4]; however, it will be seen that the versions that we chose are fair (and probably even favourable) representatives of the possible variants of these methods.

2. Some basic results

Before considering our problem proper we need to state certain results on the Benjamini–Hochberg method whose proofs may be found in [5,6]. For each n let $Z_1, Z_2, \ldots, Z_n$ be random variables (r.v.'s) with values in $[0, 1]$ representing the p-values associated with n test statistics, n_0 of which are computed under null (or true) hypotheses and the remaining $n_1 = n - n_0$ under alternative (or false) hypotheses. We define γ_n by $n_0 = \gamma_n n$, $n_1 = (1 - \gamma_n)n$, and consider the case where $\gamma_n = n_0/n \to \gamma \in (0, 1)$ as $n \to \infty$. We put $Z_j = X_j$ for $j = 1, \ldots, n_0$, and $Z_{n_0+j} = Y_j$ for $j = 1, \ldots, n_1$. Throughout the paper, F will always denote the distribution function (d.f.) of a standard uniform r.v.

Let $0 \le Z_{1:n} \le \cdots \le Z_{n:n} \le 1$ be the order statistics of the sample of p-values. The Benjamini–Hochberg procedure is this: Fix $q \in [0, 1]$ and reject all hypotheses whose p-values fall below the random threshold $qR_n(q)/n$, where

$$R_n(q) := \max\left\{i : Z_{i:n} \le q\frac{i}{n}\right\}. \tag{2.1}$$

The r.v.'s $R_n(q)$ and $S_n(q) = \sum_{j=1}^{n_0} 1_{\{X_j \le q\frac{R_n(q)}{n}\}}$ give the number of rejected hypotheses and the number of incorrectly rejected hypotheses, respectively. The r.v.'s $\Pi_{1,n}(q) = S_n(q)/(R_n(q) \vee 1)$ and $\Pi_{2,n}(q) = (R_n(q) - S_n(q))/n_1$ are the proportion of incorrectly rejected hypotheses out of all rejected hypotheses and the proportion of correctly rejected hypotheses; their expected values are called **false discovery rate** (f.d.r.) and **average power**, respectively. When they exist, the limits in probability of $\Pi_{1,n}(q)$ and $\Pi_{2,n}(q)$ are referred to as the **asymptotic f.d.r.** and **asymptotic (average) power**. For brevity, and unless confusions may arise, we shall often drop the qualification of 'asymptotic' and simply refer to the r.v.'s and their limits in probability by the same name.

The limits exist and can be computed or estimated under the **assumption** (to be strengthened later on so as to serve our present purpose) that

$$F_{n_0}(u) := \frac{1}{n_0}\sum_{i=1}^{n_0} 1_{\{X_i \le u\}} \to^P F(u) \tag{2.2}$$

and

$$G_{n_1}(u) := \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{\{Y_i \le u\}} \to^P G(u) \tag{2.3}$$

uniformly in u, where $G \ne F$ is some d.f. on $[0, 1]$. From (2.2) and (2.3) it follows that $\sup_u |H_n(u) - H(u)| \to^P 0$, where H_n is the empirical d.f. of $\{Z_1, \ldots, Z_n\}$ and $H = \gamma F + (1 - \gamma)G$; moreover, if G is concave and such that $G'(0+) > (1 - q\gamma)/[q(1 - \gamma)]$ then

$$\frac{R_n(q)}{n} \to^P \rho(q), \tag{2.4}$$

where $q\rho(q)$ is the unique $u > 0$ such that $G(u) = u(1 - q\gamma)/[q(1 - \gamma)]$, and

$$\Pi_{1,n}(q) \to^P q\gamma, \quad \Pi_{2,n}(q) \to^P \rho(q)\frac{(1 - q\gamma)}{(1 - \gamma)}. \tag{2.5}$$

In multiple testing problems it can often be assumed that $G'(0+) = \infty$, in which case the second condition on G is automatically satisfied irrespective of the values of q and γ.

Although for definiteness G **will be assumed concave** throughout the paper, it should be pointed out that $\rho(q)$, and hence asymptotic average power, can be computed as we have just indicated for a much more general G (see Theorem 3.1 and Corollary 3.3 of [6] for the general expressions of $\rho(q)$ and average power).

The results now stated imply that if G and γ are known then one can compute the asymptotic f.d.r. and average power. However, G and γ are usually unknown in practice; then the Benjamini–Hochberg approach assumes that γ can be anything from 0 to 1, and uses the majorization of the first limit in (2.5) by q to impose a generally conservative bound on the f.d.r. On the other hand, if $\hat{\gamma}_n$ is an estimator of γ then it is advantageous to replace q by $q_n := \delta/\hat{\gamma}_n$,

where $\delta \in (0, \gamma)$, in (2.1); the resulting procedure will be called an *adaptive version of the Benjamini–Hochberg method* and is consistent in the sense that

$$\frac{R_n(q_n)}{n} \to^P \rho\left(\frac{\delta}{\gamma}\right) \tag{2.6}$$

and

$$\Pi_{1,n}(q_n) \to^P \delta, \quad \Pi_{2,n}(q_n) \to^P \rho\left(\frac{\delta}{\gamma}\right)\frac{(1-\delta)}{(1-\gamma)}, \tag{2.7}$$

provided the assumptions stated earlier continue to hold and $\hat{\gamma}_n$ is consistent.

Although the method proposed in this article depends crucially on the availability of consistent estimates of γ and G, we will not consider here the problem of obtaining such estimates. 'Storey-type' estimators of γ such as those proposed by Benjamini and Hehberg [3], Storey [15], Langaas, Lindqvist and Ferkingsland [10] and Ferreira and Zwinderman [5] are not always consistent, sometimes yielding *overestimates* of γ. A consistent and easily implemented estimator of γ which can be used in connection with t-tests and the like has been given by Ferreira and Zwinderman [5]. Another method that is consistent under general conditions has been recently proposed by [9]. If $\hat{\gamma}_n$ is consistent, then $\hat{G}_n = (H_n - \hat{\gamma}_n F)/(1 - \hat{\gamma}_n)$ is obviously consistent for G (under (2.2) and (2.3), as always), so a consistent estimator of $\hat{\gamma}_n$ immediately yields a consistent estimator of G. Apart from the essentially non-parametric estimators of γ and G now mentioned, one useful class of estimators is obtained by fitting the mixture model $H = \gamma F + (1-\gamma)G$, where G is a parametric family of d.f.'s, to the sample of p-values; Allison et al. [1] and Xiang, Edwards and Gadbury [16], for example, following earlier work by Parker and Rothenberg [12], advocate modelling G by a beta d.f. and study the problem of estimating γ and the parameters of G from the sample of p-values. In the sequel we shall assume that in each multiple testing problem considered one can apply the adaptive Benjamini–Hochberg method consistently in the sense that consistent estimators of γ and G are available and (2.6) and (2.7) hold.

For brevity, we shall sometimes refer to the d.f. of the p-values associated with the alternative hypotheses (G, in this section) as the **non-null d.f.**

3. The multivariate method

In order to deal with s independent multiple testing studies we employ the obvious notation X_{ij}, Y_{ij}, Z_{ij}, $F^{(j)}$, $G^{(j)}$, $H^{(j)}$, $q_{n,j}$, ρ_j and δ_j to denote items analogous to those introduced earlier and corresponding to study j ($j = 1, \ldots, s$).

We assume that the s studies are **homologous** in the sense that γ, the proportion of null hypotheses, is fixed and the same in each study, and $F^{(j)} = F$ for all j. However, we assume nothing about $G^{(1)}, \ldots, G^{(s)}$: these d.f.'s may correspond to p-values computed from different test statistics, or from the same statistic with different sample sizes, and hence can be very different from each other.

The natural generalization of (2.2) and (2.3) to the multivariate case is the assumption that they hold uniformly in $u \in \mathbb{R}^s$ with $X_i = (X_{i1}, \ldots, X_{is})$, $Y_i = (Y_{i1}, \ldots, Y_{is})$, F replaced by the product of s standard uniform d.f.'s, $G = G^{(1)} \times \cdots \times G^{(s)}$, and with the inequalities being interpreted pointwise. If each $G^{(j)}$ satisfies the condition on G mentioned in the previous section and if for each multiple testing problem j there is a consistent estimate $\hat{\gamma}_{n,j}$ of γ, then these multivariate versions of (2.2) and (2.3) ensure the validity of the convergence results that follow.

The **multivariate multiple testing procedure** that we propose consists of rejecting those hypotheses in the *union* of the s individual rejection lists, the latter being obtained from s applications of the adapted Benjamini–Hochberg procedure at **individual false discovery rates** $\delta_1, \ldots, \delta_s$ to be determined so as to maximize average power subject to a f.d.r. of δ. Our problem, therefore, is to find the δ_j's that satisfy these two requirements.

The number of rejected hypotheses in the multivariate procedure – the number of hypotheses rejected in at least one of the s applications of the adapted Benjamini–Hochberg procedure – is

$$\mathbf{R}_n = n - \sum_{i=1}^{n} 1_{\left\{Z_{i1} > q_{n,1}\frac{R_{n,1}(q_{n,1})}{n}, \ldots, Z_{is} > q_{n,s}\frac{R_{n,s}(q_{n,s})}{n}\right\}},$$

and the number of incorrect rejections

$$\mathbf{S}_n = n_0 - \sum_{i=1}^{n_0} 1_{\left\{X_{i1} > q_{n,1}\frac{R_{n,1}(q_{n,1})}{n}, \ldots, X_{is} > q_{n,s}\frac{R_{n,s}(q_{n,s})}{n}\right\}}.$$

It is then a consequence of our assumptions that

$$\begin{aligned}
\frac{\mathbf{R}_n}{n} &= 1 - \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{s} 1_{\left\{Z_{ij} > q_{n,j}\frac{R_{n,j}(q_{n,j})}{n}\right\}} \\
&= 1 - \frac{n_0}{n}\frac{1}{n_0}\sum_{i=1}^{n_0}\prod_{j=1}^{s} 1_{\left\{X_{ij} > q_{n,j}\frac{R_{n,j}(q_{n,j})}{n}\right\}} - \frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1}\prod_{j=1}^{s} 1_{\left\{Y_{ij} > q_{n,j}\frac{R_{n,j}(q_{n,j})}{n}\right\}} \\
&\to^P 1 - \gamma\prod_{j=1}^{s}\left\{1 - \frac{\delta_j}{\gamma}\rho_j\left(\frac{\delta_j}{\gamma}\right)\right\} - (1-\gamma)\prod_{j=1}^{s}\left\{1 - G^{(j)}\left(\frac{\delta_j}{\gamma}\rho_j\left(\frac{\delta_j}{\gamma}\right)\right)\right\}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\mathbf{S}_n}{n} &= \frac{n_0}{n} - \frac{n_0}{n}\frac{1}{n_0}\sum_{i=1}^{n_0}\prod_{j=1}^{s} 1_{\left\{X_{ij} > q_{n,j}\frac{R_{n,j}(q_{n,j})}{n}\right\}} \\
&\to^P \gamma\left(1 - \prod_{j=1}^{s}\left\{1 - \frac{\delta_j}{\gamma}\rho_j\left(\frac{\delta_j}{\gamma}\right)\right\}\right),
\end{aligned}$$

and hence that the f.d.r. and average power of our procedure satisfy

$$\mathbf{\Pi}_{1,n} := \frac{\mathbf{S}_n}{\mathbf{R}_n \vee 1} \to^P \frac{\gamma\,(1 - A(d))}{\gamma\,(1 - A(d)) + (1-\gamma)\,(1 - B(d))} \tag{3.1}$$

and

$$\mathbf{\Pi}_{2,n} := \frac{\mathbf{R}_n - \mathbf{S}_n}{n_1} \to^P 1 - B(d), \tag{3.2}$$

where we set $d_j = \delta_j/\gamma$, $d = (d_1, \ldots, d_s) \in [0,1]^s$, and

$$A(d) = \Pi_{j=1}^{s}\left\{1 - d_j\rho_j(d_j)\right\}, \qquad B(d) = \Pi_{j=1}^{s}\left\{1 - G^{(j)}\left(d_j\rho_j(d_j)\right)\right\}.$$

The problem of maximizing asymptotic average power subject to an asymptotic f.d.r. of δ can now be formulated as the problem of finding the values of $d \in [0,1]^s$ which minimize

$$B(d) \quad \text{subject to} \quad \frac{1-A(d)}{1-B(d)} = \frac{\delta(1-\gamma)}{\gamma(1-\delta)}. \tag{3.3}$$

(Since the constraint says $1-A \propto 1-B$, one can alternatively minimize A.)

Making use of the fact that $d_j\rho_j(d_j)$ is the unique $u > 0$ such that $G^{(j)}(u) = u\gamma(1-\delta_j)/[\delta_j(1-\gamma)]$, we can also write B as

$$B(d) = \prod_{j=1}^{s} \left\{1 - (1-\gamma d_j)\rho_j(d_j)/(1-\gamma)\right\}.$$

It is then easy to see that each of the points $(\delta/\gamma, 0, \ldots, 0)$, $(0, \delta/\gamma, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, \delta/\gamma)$, is admissible (i.e. satisfies the restriction in (3.3) and belongs to $[0,1]^s$); and since B is bounded and continuous on $[0,1]^s$, it follows that our problem has at least one solution d^* yielding a global minimum (which corresponds to a global maximum of average power). Moreover, setting $d_k = \delta/\gamma$ and $d_j = 0$ for $j \neq k$ we get

$$\rho_k(d_k)\frac{(1-\gamma d_k)}{(1-\gamma)} = 1 - B(d) \leq 1 - B(d^*) \quad \forall k, \tag{3.4}$$

by definition of d^*, which simply means that the power attained by combining the information on the s multiple testing procedures yields at least as much power as that obtained by applying any of the s procedures separately.

Remark. Instead of maximizing power one might be interested in minimizing the asymptotic *false non-discovery rate* of [7], namely the limit in probability of $\mathbf{\Pi}_{3,n} := 1 - (n_0 - \mathbf{S}_n)/[(n - \mathbf{R}_n) \vee 1]$, the proportion of incorrect non-rejections among non-rejections. Because

$$\mathbf{\Pi}_{3,n} \to^P \frac{(1-\gamma)B(d)}{\gamma A(d) + (1-\gamma)B(d)} = \frac{(1-\gamma)B(d)}{\gamma(1-r) + (1-\gamma(1-r))B(d)}$$

under the constraint in (3.3), where we wrote $r = \delta(1-\gamma)/[\gamma(1-\delta)]$, and because the term on the right here increases with $B(d)$, our method yields both maximum power and minimum false non-discovery rate. □

The rest of this section will be concerned with obtaining and interpreting solutions to the optimization problem now formalized and with the analysis of some examples. We shall outline a method for determining the δ_j's – the individual f.d.r.'s – when γ and the $G^{(j)}$'s (and hence the ρ_j's) are known. Recall that in practice these need to be estimated from data; what we propose is to obtain a solution as indicated below with the theoretical items replaced by their estimates or empirical counterparts.

3.1. Obtaining and interpreting solutions

In order to determine the optimal solution d^* it seems best to change variables according to $x_j = d_j\rho_j(d_j)$ and solve the simpler problem of finding the points $x = (x_1, \ldots, x_s) \in [0,1]^s$ which minimize

$$\tilde{B}(x) \quad \text{subject to} \quad \frac{1-\tilde{A}(x)}{1-\tilde{B}(x)} = \frac{\delta(1-\gamma)}{\gamma(1-\delta)}, \tag{3.5}$$

where $\tilde{A}(x) = \Pi_{j=1}^{s}(1 - x_j)$ and $\tilde{B}(x) = \Pi_{j=1}^{s}\left\{1 - G^{(j)}(x_j)\right\}$. Since $d_j\rho_j(d_j)$ is a strictly increasing function of d_j, it is a simple matter to obtain d_j (hence δ_j) from x_j once the latter has been found.

With a view towards solving (3.5) by Lagrange's method, let us observe the properties

$$\frac{D_k\tilde{A}(x)}{\tilde{A}(x)} = -\frac{1}{1 - x_k} \quad \text{and} \quad \frac{D_k\tilde{B}(x)}{\tilde{B}(x)} = -\frac{g^{(k)}(x_k)}{1 - G^{(k)}(x_k)}, \tag{3.6}$$

where as usual D_k denotes differentiation with respect to the k-th variable. The Lagrangian associated with (3.5) is

$$L(x, \lambda) = \tilde{B}(x) + \lambda\left\{\gamma - \delta + \delta(1 - \gamma)\tilde{B}(x) - \gamma(1 - \delta)\tilde{A}(x)\right\},$$

and the $s + 1$ equations satisfied by any extremum are

$$\begin{aligned} &[1 + \lambda\delta(1 - \gamma)]\, D_k\tilde{B}(x) - \lambda\gamma(1 - \delta)D_k\tilde{A}(x) = 0, \quad k = 1, \ldots, s, \\ &\gamma - \delta + \delta(1 - \gamma)\tilde{B}(x) - \gamma(1 - \delta)\tilde{A}(x) = 0. \end{aligned} \tag{3.7}$$

Using (3.6) in the first s equations in (3.7) we can eliminate λ and obtain the $s(s-1)/2$ equations

$$\frac{(1 - x_k)g^{(k)}(x_k)}{1 - G^{(k)}(x_k)} = \frac{(1 - x_l)g^{(l)}(x_l)}{1 - G^{(l)}(x_l)} \quad (k < l). \tag{3.8}$$

Our problem now reduces to solving a system of s equations in s unknowns involving $s - 1$ of the equations in (3.8) and the last equation in (3.7). Although this can be done in a number of ways, it seems particularly convenient to write the system in the form $\psi(x) = x$, where $\psi(x) = (\psi_1, \ldots, \psi_s)$,

$$\psi_k(x) = \frac{1}{C_k}\left\{\frac{1 - G^{(k+1)}(x_{k+1})}{(1 - x_{k+1})g^{(k+1)}(x_{k+1})} - \frac{1 - G^{(k)}(x_k)}{(1 - x_k)g^{(k)}(x_k)}\right\} + x_k$$

for $k = 1, \ldots, s - 1$,

$$\psi_s(x) = \frac{1}{C_s}\left\{\frac{\delta(1 - \gamma)\tilde{A}(x)}{(\gamma - \delta) + \delta(1 - \gamma)\tilde{A}(x)\tilde{C}(x)\left[\frac{1 - G^{(s)}(x_s)}{(1 - x_s)g^{(s)}(x_s)}\right]^s} - \frac{\delta(1 - \gamma)}{\gamma(1 - \delta)}\right\} + x_s,$$

$\tilde{C}(x) = \prod_{j=1}^{s} g^{(j)}(x_j)$ and $C_1, \ldots, C_s$ are positive constants to be chosen, and then use the method of successive approximations to find a fixed point $\xi = \psi(\xi)$. More precisely, given a starting point $x^{(0)}$, we successively obtain $x^{(m+1)}$ from $x^{(m)}$ by $x^{(m+1)} = \psi(x^{(m)})$, $m = 0, 1, \ldots$. As is well known, if $x^{(0)}$ is not too far from ξ and $\max_{u\in\mathbb{R}^s} \|\psi'(x)u\|/\|u\| < 1$ for x near ξ then $\lim_{m\to\infty} x^{(m)} = \xi$; the constants C_k are introduced in order to temper the derivative of ψ. As starting values one may take $x_j^{(0)} = \rho_j(\delta/\gamma)\delta/\gamma$, which typically yield convergence unless the studies have very different or very low power curves.

We shall not seek detailed conditions for convergence and optimality (which seem difficult to obtain and would involve the derivatives of the densities) but note the following points: (i) The presence of the densities in the denominator of ψ_k typically ensures that ψ' can be tempered (by appropriate choice of the constants) if the densities are ∞ at zero. (ii) Obtaining convergence is usually straightforward, though one may sometimes have to try different starting values and to

increase the values of the constants. (iii) The optimality of the solution is usually obvious from the context (see the examples below).

The following two examples illustrate the workings of the method and the kind of improvement in power that it yields.

Example 3.1. We can get an idea of the effect of combining more or less equally powerful studies by considering the case in which $G^{(j)} = G$ for all j. By symmetry, the solution we seek is of the form $x = (\xi, \ldots, \xi)$ for some $\xi \in [0, 1]$; we show that ξ is uniquely determined by the constraint. The constraint can be written as

$$r^{-1} \equiv \frac{\gamma(1-\delta)}{\delta(1-\gamma)} = \frac{1-[1-G(\xi)]^s}{1-(1-\xi)^s} \equiv \frac{\varphi_s(y)}{y}, \quad y = 1-(1-\xi)^s, \tag{3.9}$$

where $\varphi_s(y) = 1 - [1 - G(1 - (1 - y)^{1/s})]^s$ defines a d.f. on $[0, 1]$. Since φ_s is concave (because $y \to [1 - G(1 - (1 - y)^{1/s})]^s$ is a composition of convex functions), we see that Eq. (3.9) has a positive solution $\xi^* = 1 - (1 - y^*)^{1/s}$ – a unique positive solution – if and only if $\varphi_s'(0+) > 1/r$, or equivalently, since $y \downarrow 0 \Leftrightarrow \xi \downarrow 0$, if and only if $G'(0+) > 1/r$. It follows that if $G'(0+) = \infty$ then problem (3.5) always has a solution irrespective of the f.d.r. $\delta \in (0, \gamma)$ chosen.

In order to interpret the workings of the method in this case, observe that (i) φ_s is uniform if and only if G is; (ii) φ_s is the d.f. of $\min\{U_1^*, \ldots, U_s^*\}$, where $U_j^* = 1 - (1 - U_j)^s$ and $(U_1, \ldots, U_s)$ is either a vector of independent standard uniform r.v.'s or a vector of independent r.v.'s with d.f. $G \neq F$. Thus, under the assumption that the s p-values associated with each hypothesis have the same joint distribution as $(U_1, \ldots, U_s)$, our method is asymptotically equivalent to transforming the s samples of p-values by $u \to 1 - (1 - u)^s$, taking the minimum transformed p-value across the s studies for each hypothesis, and applying the Benjamini–Hochberg method to the resulting sample. (In the most interesting situation where the p-values computed under different alternative hypotheses are not identically distributed this interpretation is obviously not valid.) □

Example 3.2. Let $G^{(j)}(x) = x^{\alpha_j}$ for $x \in [0, 1]$ and $\alpha_j \in (0, 1)$, $j = 1, \ldots, s$. Then we have $x_j \equiv d_j \rho_j(d_j) = \left[d_j(1-\gamma)/(1-d_j\gamma)\right]^{1/(1-\alpha_j)}$ for $d_j \in [0, 1]$, and we may first find the solution x of (3.5) and then determine the δ_j's explicitly from $\delta_j = d_j\gamma$ and $d_j = 1/(\gamma + (1-\gamma)x_j^{\alpha_j - 1})$.

Table 1 gives the solutions δ_1, δ_2 in the case $s = 2$ and for several values of γ, δ, α_1 and α_2, and is intended to illustrate the basic workings of the method. For each δ, let $\Pi^{(j)}(\delta)$ represent the power achieved by applying the adaptive Benjamini–Hochberg method to study j with a f.d.r. of δ; we shall call it the **baseline power** (at δ) of study j. In order to express the gain in power achieved by combining the results of the s studies by means of the multivariate method with the same f.d.r., we give, in the last column of the table, the percentage of improvement in power relative to what one would get by using the most powerful of the studies; since in our examples the most powerful study is always the first, this is computed as $[\Pi(\delta) - \Pi^{(1)}(\delta)]/[\Pi^{(1)}(\delta)]$, where $\Pi(\delta)$ is the power of the multivariate procedure.

The application of our method amounts to the determination of the δ_j's which maximize power subject to a f.d.r. of δ. The comparison of the baseline powers $\Pi^{(j)}(\delta)$ with the **individual powers** $\Pi^{(j)}(\delta_j)$ suggests that, at least with this choice of $G^{(j)}$'s, the method tends to work by decreasing the power of the most powerful study and increasing the power of the weaker study (cf. columns 3 and 4, columns 7 and 8), concurrently giving a greater value to δ_2 than to δ_1;

Table 1
Solutions and powers pertaining to the combination of two studies

$\gamma = 0.8, \delta = 0.05$									
α_1	α_2	$\Pi^{(1)}(\delta)$	$\Pi^{(2)}(\delta)$	δ_1	δ_2	$\Pi^{(1)}(\delta_1)$	$\Pi^{(2)}(\delta_2)$	$\Pi(\delta)$	Gain (%)
0.1	0.15	0.6180	0.4657	0.0364	0.0397	0.5956	0.4463	0.7761	25.58
0.1	0.5	0.6180	0.0132	0.0472	0.0898	0.6139	0.0247	0.6234	0.86
0.5	0.55	0.0132	0.0050	0.0486	0.0528	0.0128	0.0054	0.0181	37.42
$\gamma = 0.8, \delta = 0.1$									
α_1	α_2	$\Pi^{(1)}(\delta)$	$\Pi^{(2)}(\delta)$	δ_1	δ_2	$\Pi^{(1)}(\delta_1)$	$\Pi^{(2)}(\delta_2)$	$\Pi(\delta)$	Gain (%)
0.1	0.15	0.6715	0.5313	0.0719	0.0770	0.6451	0.5052	0.8244	22.76
0.1	0.5	0.6715	0.0278	0.0922	0.1532	0.6649	0.0452	0.6800	1.26
0.5	0.55	0.0278	0.0125	0.0968	0.1042	0.0268	0.0132	0.0397	42.81
$\gamma = 0.9, \delta = 0.1$									
α_1	α_2	$\Pi^{(1)}(\delta)$	$\Pi^{(2)}(\delta)$	δ_1	δ_2	$\Pi^{(1)}(\delta_1)$	$\Pi^{(2)}(\delta_2)$	$\Pi(\delta)$	Gain (%)
0.1	0.15	0.6137	0.4605	0.0740	0.0806	0.5916	0.4417	0.7720	25.79
0.1	0.5	0.6137	0.0123	0.0948	0.1740	0.6097	0.0234	0.6188	0.84
0.5	0.55	0.0123	0.0046	0.0973	0.1053	0.0120	0.0050	0.0169	36.96

exceptions to this rule occur when both studies are very powerful (cf. the cases where $\alpha_1 = 0.1$, $\alpha_2 = 0.15$). The gain in power is greater if the two studies have similar power (e.g. $\alpha_1 = 0.1$, $\alpha_2 = 0.15$ compared with $\alpha_1 = 0.1, \alpha_2 = 0.5$); naturally, the weaker the studies are, the greater the gain in power. If the studies are very unbalanced in terms of power (e.g. $\alpha_1 = 0.1, \alpha_2 = 0.5$) then the gain in power is very small; what the method does is then to set δ_1 very close to δ and increase δ_2 substantially, yielding a small improvement relative to what one would get by using the most powerful study alone (which would formally correspond to taking $\delta_1 = \delta$ and $\delta_2 = 0$). These trends change consistently as each one of γ and δ varies with the other held fixed (i.e. as one passes from one panel of the table to the next).

To illustrate what happens with more than two studies we collect in Table 2 some results for $s = 3$ in the case $\gamma = 0.8$, $\delta = 0.1$, and for several choices of $\alpha_1, \alpha_2, \alpha_3$. The determination of the δ_j's follows the same principle as before: the method increases power by allowing somewhat greater individual f.d.r.'s in the weaker studies. As the last two columns indicate, the gain in power is greater when combining balanced studies; here, 'Gain1' is calculated as before (as the increase in power relative to the power of the most powerful study), while 'Gain2' gives the increase in power relative to the power achieved by combining the *two* most powerful studies (e.g. those corresponding to $\alpha_1 = 0.1$ and $\alpha_2 = 0.15$ in the second row of numbers, whose power in the combined procedure is given in the middle panel of Table 1) among the three.

In order to analyse the effects of passing from two to three studies we compare Table 2 with the middle panel of Table 1. Naturally, the benefit of adding one extra study is more visible when $s = 2$ (cf. 'Gain' in Table 1 and 'Gain2' in Table 2). The incorporation of a third study is achieved by decreasing the individual f.d.r.'s in the two studies, and hence by decreasing their individual powers (cf. the cases $\alpha_1 = 0.1, \alpha_2 = 0.15$ in Table 1 and $\alpha_1 = 0.1, \alpha_2 = 0.12, \alpha_3 = 0.15$ in Table 2); if the third study is very weak, then δ_1 and δ_2 are set close to their values in the case $s = 2$ (cf. the cases $\alpha_1 = 0.1, \alpha_2 = 0.15$ and $\alpha_1 = 0.1, \alpha_2 = 0.15, \alpha_3 = 0.5$). □

Table 2
Solutions and powers pertaining to the combination of three studies

$\gamma = 0.8, \delta = 0.1$											
α_1	α_2	α_3	δ_1	δ_2	δ_3	$\Pi^{(1)}(\delta_1)$	$\Pi^{(2)}(\delta_2)$	$\Pi^{(3)}(\delta_3)$	$\Pi(\delta)$	Gain1 (%)	Gain2 (%)
0.1	0.12	0.15	0.0559	0.0576	0.0604	0.6262	0.5655	0.4823	0.9159	36.39	7.45
0.1	0.15	0.5	0.0690	0.0741	0.1220	0.6420	0.5014	0.0347	0.8277	23.26	0.40
0.1	0.5	0.55	0.0884	0.1483	0.1581	0.6615	0.0435	0.0238	0.6839	1.84	0.57
0.5	0.52	0.55	0.0957	0.0986	0.1031	0.0265	0.0203	0.0131	0.0586	111.14	23.89

Our last example is somewhat artificial but serves to show on the one hand that Lagrange's method will not always yield the solution, and on the other that in some cases combining studies simply fails to increase power.

Example 3.3. Take $G^{(j)}(x) = 1 - (1-x)^{\beta_j}$ for $x \in [0,1]$ and $\beta_j > 1$. Although $G^{(j)}$ is concave, its right-hand derivative at zero is $\beta_j < \infty$; hence for the basic consistency results (2.6) and (2.7) to hold we need to restrict the choice of δ_j by the condition $\beta_j > \gamma(1-\delta_j)/[\delta_j(1-\gamma)]$ for all j. As pointed out in Section 2, practical problems typically give rise to densities that are more peaked near zero than $g^{(j)} = (G^{(j)})'$, whence the 'artificial'.

Clearly, $g^{(j)}(x) = \beta_j(1-x)^{\beta_j - 1}$; hence $(1-x)g^{(j)}(x)/[1-G^{(j)}(x)] = \beta_j$ for all j and all x, and Eq. (3.8) holds for some (and then for all) x if and only if $\beta_1 = \beta_2 = \cdots = \beta_s$ (a case covered by Example 3.1). (This class of $G^{(j)}$'s is obviously the only one with this property.) In other words, the Lagrange multiplier method fails. Since differentiability properties are not at stake, and since the function $x \to \gamma - \delta + \delta(1-\gamma)\tilde{B}(x) - \gamma(1-\delta)\tilde{A}(x)$ (which corresponds to the constraint) has a non-vanishing derivative, we must conclude that *if the* β_j's *are not all the same* then the optimal solution is a boundary point of the unit cube. In fact, it is easy to see that if all the β_j's are different then the optimal solution must be one of the points $(\delta/\gamma, 0, \ldots, 0), \ldots, (0, \ldots, 0, \delta/\gamma)$, so that the maximum power equals (see (3.4))

$$1 - B(d^*) = \max_{k=1,\ldots,s} \rho_k(\delta/\gamma)\frac{(1-\delta)}{(1-\gamma)},$$

which is to say that the optimal solution consists of using the most powerful study alone. □

4. Comparison with univariate methods

The comparison of our method with univariate methods will be based on a class of models for G of the form

$$G(u; N) = 1 - \Phi\left(\frac{\Phi^{-1}(1-u/2) - \sqrt{N}\mu}{\sqrt{1+N\tau^2}}\right) + \Phi\left(-\frac{\Phi^{-1}(1-u/2) + \sqrt{N}\mu}{\sqrt{1+N\tau^2}}\right),$$

where as usual Φ is the standard normal d.f., Φ^{-1} its inverse, and N, μ and τ^2 are certain parameters. This is one of the 'location models' considered by [5] and arises out of multiple testing problems with t- or z-statistics, for example. In order to explain briefly the rationale behind $G(\cdot; N)$ and the meaning of the parameters (details can be found in the reference just mentioned) let $T_1, \ldots, T_n$ be test statistics such that T_j is a normal r.v. with mean $\theta_j\sqrt{N}$ and variance 1. We think of T_j as some variant of the t-statistic, of θ_j as the 'effect size' and of N

as *sample size*, or more precisely as a factor accounting for sample size (in the case of the two-sample t-statistic, N is the reciprocal of the sum of the inverses of the two sample sizes). Defining the p-value associated with T_j as the probability of observing a result at least as extreme as the observed value of $|T_j|$ in a similar, independent experiment, and assuming that the empirical d.f. of $\{\theta_{n_0+1}, \ldots, \theta_n\}$ (the non-zero effect sizes) converges to that of a normal r.v. with mean μ and variance τ^2 as $n \to \infty$, we get $G(\cdot; N)$ as the asymptotic d.f. of the p-values associated with the false hypotheses. [The empirical d.f. of $\{\theta_{n_0+1}, \ldots, \theta_n\}$ is defined exactly as if the non-zero effect sizes were random. Its convergence may hold in probability (for example) or at every point of the sample space depending on whether the effect sizes are interpreted as random or not; both interpretations are allowed provided the limiting d.f. is non-random.]

To model the outcome of s independent studies we let $\{T_1^{(k)}, \ldots, T_n^{(k)}\}$ $(k = 1, \ldots, s)$ be independent sets of r.v.'s, assume that $T_j^{(k)}$ is normal with mean $\theta_j\sqrt{N_k}$ and variance 1, and also that the empirical d.f. of the p-values associated with the false hypotheses from study k converges to $G^{(N_k)}(\cdot) := G(\cdot; N_k)$. Then, as far as asymptotic results are concerned, the k-th study is characterized by the d.f. $G^{(N_k)}$, which in turn is determined by the 'sample size' N_k and by μ and τ. The fact that μ and τ are the same across the s studies expresses the fundamental assumption that the s studies address the same research question and measure the same thing (the same θ_j's).

By the **average method** we mean the procedure of computing the averages of the statistics across studies, namely $\bar{T}_j := \frac{1}{s}\sum_{k=1}^{s} T_j^{(k)}$, $j = 1, \ldots, n$, and applying the Benjamini–Hochberg method to the sample of p-values associated with the tests of $\theta_j = 0$ against $\theta_j \neq 0$. The p-values are easily computed using the fact that $\bar{T}_j$ has a normal distribution with mean $\theta_j \frac{1}{s}\sum_{k=1}^{s}\sqrt{N_k}$ and variance $1/s$, which can also be used to show that the limit of the empirical d.f. of the p-values associated with the false hypotheses is $G^{(N')}(\cdot) := G(\cdot; N')$, with $N' = \sum_{k=1}^{s}\sqrt{N_k/s}$. It follows in particular (by the Cauchy–Schwartz inequality) that:

(i) Under the constraint of a *total sample size* of $N = \sum_{k=1}^{s} N_k$, the average method attains maximum power with $N_j = N/s$ for all j.
(ii) If all the studies are based on the same sample size, then combining them by the average method is equivalent to using a single study with an s-fold increase in sample size.

The second observation shows that there is a certain naturalness to the average method which almost recommends it as the standard for comparison. However, as pointed out by a referee, there is a more general univariate method based on Fisher's method of combining p-values which is equally simple and which for brevity we shall call here **Fisher's method**: it consists of applying the Benjamini–Hochberg to the p-values of the test statistics $\tilde{T}_j := -2\sum_{k=1}^{s}\log Z_{jk}$, $j = 1, \ldots, n$, the p-values being computed using the fact that $\tilde{T}_j$ has a chi-square distribution with $2s$ degrees of freedom if and only if $Z_{j1}, \ldots, Z_{js}$ are standard uniform r.v.'s. Variations of Fisher's method have been proposed and applied to real data by Rhodes et. al. [13], for example, but as far as we know there are no published results which allow us to assess its merits relative to the average method. We conjecture that in the case of t- or z-statistics the average method is uniformly more powerful than Fisher's method, and in the sequel will provide some evidence that this is true in the case of the data generating model described above.

One way of finding out whether one univariate method is uniformly better than another is to compare the non-null d.f.'s associated with them. Specifically, if we show that one non-null d.f. is strictly greater than another non-null d.f. then it will follow that the method associated with the first is more powerful than the method associated with the second *for all choices of* γ

and δ. This is easily seen in the context of Section 2: The asymptotic power for given γ and δ is $G((\delta/\gamma)\rho(\delta/\gamma)) = \rho(\delta/\gamma)(1-\delta)/(1-\gamma)$, and $(\delta/\gamma)\rho(\delta/\gamma)$ is the point where G intersects the line $l(u) = u\gamma(1-\delta)/[\delta(1-\gamma)]$; consequently, if $G > \tilde{G}$, say, G always intersects l at a height greater than $\tilde{G}$ does. This is the most efficient way of finding out which method is best provided one of the methods is indeed uniformly more powerful than the other, since one is not restricted to making comparisons for individual choices of γ and δ, but it is not very useful otherwise.

Now $G(\cdot; 2N)$ is the non-null d.f. associated with the p-values of the average method based on two independent samples of size N. Therefore, in order to show that the average method is better than Fisher's method in combining $s = 2$ studies with the same sample size it suffices to show that $G(\cdot; 2N) \geq \tilde{G}(\cdot; 2N)$, where $\tilde{G}(\cdot; 2N)$ is the (limiting) non-null d.f. associated with Fisher's method. Although we do not have an explicit expression for $\tilde{G}(\cdot; 2N)$ (except in the uninteresting case where all the effect sizes are constant), it is straightforward to compute it approximately using simulation, namely by simulating the θ_j and the $T_j^{(k)}$ according to how we have defined them in the first two paragraphs of this section, and then computing the $\bar{T}_j$'s and the p-values as explained in the third paragraph.

Fig. 1 compares the non-null distributions associated with the average and Fisher methods for $\mu = 0.5$ and a couple of choices of N and τ, and shows that $G(\cdot; 2N) \geq \tilde{G}(\cdot; 2N)$ does indeed hold for these values of the parameters; the fact that it holds for many other values suggests that it holds in general. Apart from this main observation, Fig. 1 also tells us that the superiority of the average method over Fisher's method is more visible with 'large' N and 'large' τ.

The third function plotted in Fig. 1 is $\varphi(x; N) := 1 - [1 - G(1 - (1 - x)^{1/2}; N)]^2$, the analogue of $G(\cdot; 2N)$ and $\tilde{G}(\cdot; 2N)$ in the multivariate method. As shown in Example 3.1, the limiting power of the multivariate method when $s = 2$ corresponds to the height at which $\varphi(\cdot; N)$ intersects l, and so, as before, the inequality $\varphi(\cdot; N) \geq \tilde{G}(\cdot; 2N)$ implied by Fig. 1 suggests that the multivariate method is generally superior to Fisher's method. [$\varphi(\cdot; N)$ is generally not a non-null d.f. as defined at the end of Section 2, as our method is not univariate; it can be interpreted as such only in the less interesting situation described in the last paragraph of Example 3.1, where the multivariate method is asymptotically equivalent to a univariate method.]

In contrast, the comparison between $G(\cdot; 2N)$ and $\varphi(\cdot; N)$ is not straightforward: although $\varphi(u; N) > G(u; 2N)$ for 'moderate or large' u, the opposite is true for x near zero if the sample size or τ is 'small'. This only tells us that the average method is better than the multivariate method for smaller sample sizes and smaller τ for certain δ and γ—those for which the line l has a 'sufficiently large' slope; in order to compare the two methods we need to compute their power for specific choices of γ, δ, and the other parameters.

We shall do this with $s = 2$, $\mu = 0.5, 1$, $\tau = 1$, $\gamma = 0.8$ and $\delta = 0.1$, and over a range of sample sizes N_1, N_2; other choices of the parameters yield qualitatively similar conclusions. Fig. 2 shows the power curves of the two methods as functions of sample size. The left panels refer to the case $N = N_1 = N_2$ and also include the baseline power curve at δ—i.e. the power of the adapted Benjamini–Hochberg method applied to a single study based on a sample of size N. The right panels are intended to illustrate what happens when one study is considerably weaker than the other: they show the power curves of the two methods as functions of the pair of sample sizes N_1, N_2 with $N_2 = 2N_1$; in each case, the ordinate at N_1 is the power obtained by combining a study with sample size N_1 with a study with sample size $2N_1$. In both methods the improvement in power relative to that of the most powerful study is considerable when the two studies are equally powerful, but not impressive when they are too unbalanced. In broad terms, the average method is the more powerful with lower sample sizes and the less powerful

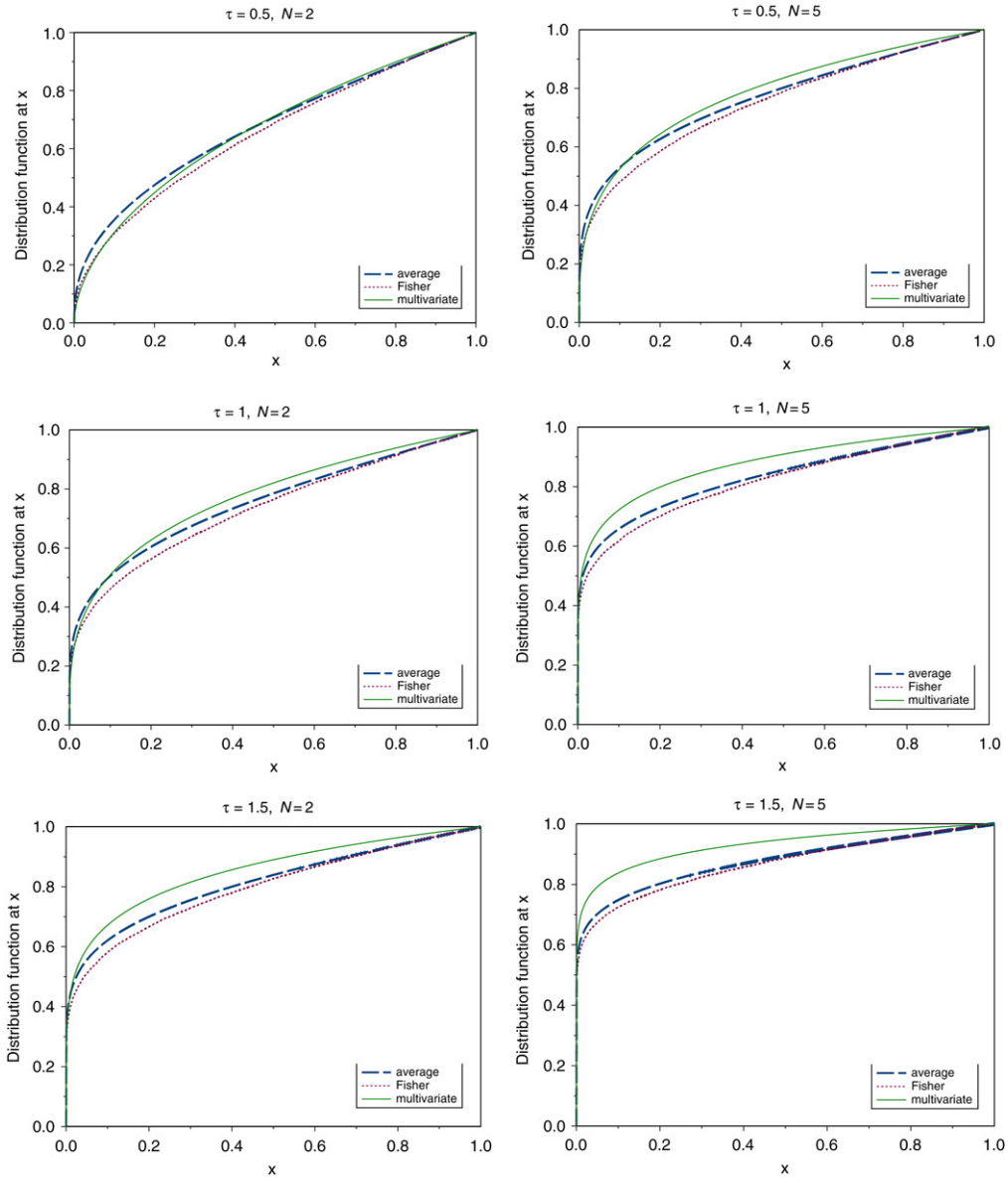

Fig. 1. Comparing the non-null d.f.'s associated with the three different methods when $\mu = 0.5$, $N = 2, 5$ and $\tau = 0.5, 1.0, 1.5$.

with large sample sizes. Taking into account the range of sample sizes and the magnitude of the differences in power shown in Fig. 2, the multivariate method seems overall preferable to the univariate one. However, in applications in which power is rather low one may be better off applying the average method, which is also simpler to implement. In any case, since power can in principle be estimated in both methods, one can always apply the two and keep the results of the most powerful one.

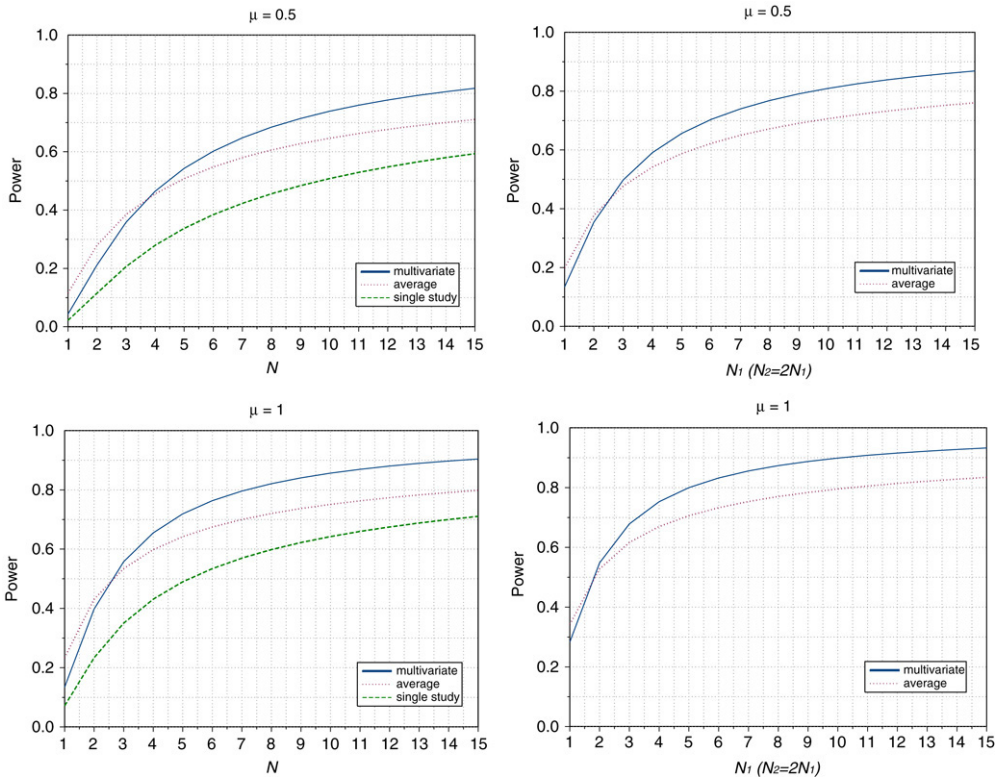

Fig. 2. Comparing the power of the multivariate and average methods for various sample sizes N when $\delta = 0.1$, $\gamma = 0.8$, $\tau^2 = 1$, $\mu = 0.5$ (upper panels) and $\mu = 1$ (lower panels). Left panels: power as a function of $N = N_1 = N_2$; right panels: power as a function of pairs (N_1, N_2) with $N_2 = 2N_1$ (the values in the horizontal axis stand for N_1).

Finally, the comparison of the two methods as developed here provides a partial answer to an interesting question essentially formulated by [8]. Suppose that the result of an experiment consists of $2N$ pairs of measurements ($2N$ on a control and $2N$ on a diseased group, say) on each of a large number of variables and that the interest lies in the detection of mean differences between pairs. In principle, one can use the paired t-test to test for differences across all variables, computing the associated p-values, and then apply the Benjamini–Hochberg method to detect mean differences subject to a given f.d.r. The question is whether it would not be more profitable to split the sample into two samples of size N, computing two sets of t-statistics/p-values, and apply the multivariate procedure to combine the two. In broad terms, and as indicated by the present example, the answer seems to be that it is better to split the samples provided N is not too small; in practice one can determine (by means of estimates of power) the point beyond which splitting is no longer advantageous.

5. Discussion

Whatever the multiple testing procedure, the more hypotheses one rejects the greater the average power attained (though of course power comes at the cost of the false discovery rate). In the multivariate method proposed here, the list of rejections is defined as the union of the lists of rejections obtained from the individual studies. It is the upper bound of the set of rejection

lists: the set which contains all lists, and the smaller set with this property. And, because it is more inclusive than any other set, it can only increase the number of rejections, and hence average power, relative to individual lists. One might think, for example, that it would be better to take the *intersection* of the lists, on the grounds that hypotheses that are rejected in all studies somehow deserve 'more credit' than hypotheses that are rejected only in a couple of studies. However, when combining a weak and a powerful study this line of thinking will force the list of rejections to be a subset of the least inclusive list – the list of the weaker study – which, by definition, cannot yield greater power than that obtained from the weaker study alone. The same holds even if the two studies being combined are equally powerful: one can never be better off by taking the intersection than by using a single study.

To give an example where this can be shown explicitly, consider the situation of Example 3.1 (combining s identical studies) with the model from Example 3.2: $G(x) = G^{(j)}(x) = x^{\alpha}$ for $x \in [0, 1]$, $\alpha \in (0, 1)$ and all j. As in Section 3, one can show that the f.d.r. and average power of *the multivariate procedure based on the intersection of the s lists* are $\gamma \bar{A}(d)/[\gamma \bar{A}(d) + (1-\gamma)\bar{B}(d)]$ and $\bar{B}(d)$, where

$$\bar{A}(d) = \prod_{j=1}^{s} \frac{\delta_j(1-\gamma)}{\gamma(1-\delta_j)} \quad \text{and} \quad \bar{B}(d) = \prod_{j=1}^{s} G^{(j)}\left(\frac{\delta_j}{\gamma}\rho_j\left(\frac{\delta_j}{\gamma}\right)\right).$$

Thus, with our choice of G, maximizing power subject to a f.d.r. of δ amounts to finding δ^* that satisfies $\{\delta^*(1-\gamma)/[\gamma(1-\delta^*)]\}^s = \delta(1-\gamma)/[\gamma(1-\delta)]$ and maximizes $G\left((\delta^*/\gamma)\rho\left(\delta^*/\gamma\right)\right)^s$, and it can be seen from the expression of $(\delta^*/\gamma)\rho\left(\delta^*/\gamma\right)$ given in Example 3.2 that the optimal average power is $(\delta(1-\gamma)/[\gamma(1-\delta)])^{\alpha/(1-\alpha)}$, which does not depend on s. In particular (take $s = 1$), the power of the multivariate method based on the intersection is equal to the power provided by a single study, and hence, from what we have said around (3.4), inferior to the power of the multivariate method that we propose when $s \geq 2$ (the methods are obviously identical when $s = 1$). In fact, although the average power of our method cannot be computed explicitly in this case, it can be shown – and seen from the numerical results of Example 3.2 – that it does increase with s.

As an alternative to taking the intersection of the rejection lists one might also consider defining the list of rejections as the list of hypotheses rejected in at least k of the studies. However, if $k = 1$ this is exactly our approach, if $k = s$ it is the approach based on the intersection, and if $1 < k < s$ it amounts to a more or less stringent form of intersection or to a more or less restricted form of union; so it is obvious that taking the union of the lists still leads to a more inclusive rejection list and hence, as far as average power is concerned, to the best result.

If we define (somewhat loosely) a *multivariate* method of multiple testing as a procedure based on the determination of individual false discovery rates and on the construction of a rejection list from individual rejection lists, it is apparent from the above discussion that our method is 'the right' multivariate Benjamini–Hochberg method. However, it does not follow that the multivariate method is superior to *univariate* methods—methods that reduce the multivariate problem to a univariate multiple testing problem by computing p-values or test statistics as a function of the p-values or test statistics from the individual studies. In Section 4 we compared the multivariate method with the simple and conceptually sound *average method*, which in turn was found to be uniformly more powerful than *Fisher's method*. We have seen that, under the particular numerical setting that we chose, both the average and multivariate methods are powerful, that the former tends to be somewhat more powerful than the latter with relatively weak studies, and that the contrary is true with more powerful studies. As already implied in the

Introduction, the average method (and Fisher's method as well) has been assessed here at its best. Indeed, we have assumed that the variances of the test statistics are known and identical, while in practice one might have to replace the $\bar{T}_j$'s by weighted averages, with weights to be determined (using Exercise 3 on p. 12 of [14], for example). Thus our numerical examples suggest that the multivariate method is a powerful method of combining different multiple testing studies.

On the other hand (point (ii) of Section 4) in the case of s 'replicate' studies the average method amounts to the univariate method applied to the sample obtained by pooling the 'raw observations' (those that enter in the calculation of the $T_j^{(k)}$'s) from the s studies. Consequently, in that case the average method is 'playing at home', being a univariate method applied to a genuine univariate problem, and our method is attempting to beat it, transforming the univariate problem into a multivariate one by splitting the sample into s identical samples; thus the fact that the latter often beats the former if the splitting is properly done (if s is not 'too large', as discussed at the end of Section 4) shows that *the multivariate Benjamini–Hochberg method can improve upon (univariate) Benjamini–Hochberg method*.

Irrespective of their relative performance, an obvious advantage of the average method over ours is its somewhat greater computational simplicity; an advantage of ours is generality: it applies to many more situations, not just to those in which test statistics can be averaged across studies in a meaningful way (specifically, in a way that accounts for the sample sizes, variances, and other study-specific characteristics), and, as mentioned at the beginning of Section 3, it does not even require the test statistics to be the same across studies.

Finally, we recall that our fundamental requirement that the studies be homologous implies that the proportion of true hypotheses is the same across studies. The latter condition can be checked by comparing the estimates of γ obtained in different studies; the method for obtaining confidence intervals for γ proposed by Lai [9], for example, may be useful in this connection. Of course, a certain amount of agreement between such estimates is expected if the studies are approximately homologous and if the assumptions behind the Benjamini–Hochberg method are tenable; if no agreement is found, then there is little point in proceeding with the method proposed here (with any method, really, since then one must look for the causes of the disagreement). An application of this and other methods to several real data sets from molecular biology is currently under way and will be published elsewhere.

Acknowledgments

The first author was funded in part by 'Nederlandse Hartstichting' (the Dutch Heart Foundation) through the grant no. 2003B215. The second author was funded in part by NBIC (Netherlands Bioinformatics Centre) through the Biorange programme.

References

[1] D. Allison, G. Gadbury, M. Heo, J. Fernandez, C. Lee, T. Prolla, R. Weindruch, A mixture model approach for the analysis of microarray gene expression data, Comput. Statist. Data Anal. (39) (2002) 1–20.

[2] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. Roy. Statist. Soc. Ser. B (57) (1995) 289–300.

[3] Y. Benjamini, Y. Hochberg, On the adaptive control of the false discovery rate in multiple testing with independent statistics, J. Educ. Behav. Statist. (2000) 25.

[4] J.-K. Choi, U. Yu, S. Kim, O. Yoo, Combining multiple microarray studies and modeling interstudy variation, Bioinformatics 19 (Suppl. 1) (2003) i84–i90.

[5] J. Ferreira, A. Zwinderman, Approximate power and sample size calculations with the Benjamini–Hochberg method, Int. J. Biostatistics 2 (1) (2006). Available online at Article 8: http://www.bepress.com/ijb/vol2/iss1/8.

[6] J. Ferreira, A. Zwinderman, On the Benjamini-Hochberg method, Ann. Statist. 34 (4) (2006).
[7] C. Genovese, L. Wasserman, Operating characteristics and extensions of the false discovery rate procedure, J. Roy. Statist. Soc. Ser. B (64) (2002) 499–517.
[8] R. Gentleman, M. Ruschaupt, W. Huber, On the synthesis of microarray experiments, Bioconductor project working papers, 2005. http://www.bepress.com/bioconductor/paper8.
[9] Y. Lai, A moment-based method for estimating the proportion of true null hypotheses and its application to microarray gene expression data, Biostatistics 8 (4) (2007) 744–755.
[10] M. Langaas, B. Lindqvist, E. Ferkingstad, Estimating the proportion of true null hypotheses, with application to DNA microarray data, J. Roy. Statist. Soc. Ser. B 67 (4) (2005) 555–572.
[11] G. McLachlan, K.-A. Do, C. Ambroise, Analysing Microarray Gene Expression Data, Wiley, 2004.
[12] R. Parker, R. Rothenberg, Identifying important results from multiple statistical tests, Stat. Med. 17 (1988) 1031–1043.
[13] D. Rhodes, T. Barrette, M. Rubin, D. Ghosh, A. Chinnaiyan, Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, Cancer Res. (62) (2002) 4427–4433.
[14] G. Seber, A. Lee, Linear Regression Analysis, 2nd ed., Wiley, 2003.
[15] J. Storey, A direct approach to false discovery rates, J. Roy. Statist. Soc. Ser. B (64) (2002) 479–498.
[16] Q. Xiang, J. Edwards, G. Gadbury, Interval estimation in a finite mixture model: Modeling p-values in multiple testing applications, Comput. Statist. Data Anal. (2006).