Test (2008) 17: 417–442
DOI 10.1007/s11749-008-0126-6

INVITED PAPER

Control of the false discovery rate under dependence using the bootstrap and subsampling

Joseph P. Romano · Azeem M. Shaikh · Michael Wolf

Published online: 30 October 2008

Abstract This paper considers the problem of testing s null hypotheses simultaneously while controlling the *false discovery rate* (FDR). Benjamini and Hochberg (J. R. Stat. Soc. Ser. B 57(1):289–300, 1995) provide a method for controlling the FDR based on p-values for each of the null hypotheses under the assumption that the p-values are independent. Subsequent research has since shown that this procedure is valid under weaker assumptions on the joint distribution of the p-values. Related procedures that are valid under no assumptions on the joint distribution of the p-values have also been developed. None of these procedures, however, incorporate information about the dependence structure of the test statistics. This paper develops methods for control of the FDR under weak assumptions that incorporate such information and, by doing so, are better able to detect false null hypotheses. We illustrate this property via a simulation study and two empirical applications. In particular, the bootstrap method is competitive with methods that require independence if independence holds, but it outperforms these methods under dependence.

This invited paper is discussed in the comments available at:
http://dx.doi.org/10.1007/s11749-008-0127-5, http://dx.doi.org/10-1007/s11749-008-0128-4,
http://dx.doi.org/10.1007/s11749-008-0129-3, http://dx.doi.org/10.1007/s11749-008-0130-x,
http://dx.doi.org/10.1007/s11749-008-0131-9.

J.P. Romano
Departments of Economics and Statistics, Stanford University, Stanford, USA
e-mail: romano@stanford.edu

A.M. Shaikh
Department of Economics, University of Chicago, Chicago, USA
e-mail: amshaikh@uchicago.edu

M. Wolf (✉)
Institute for Empirical Research in Economics, University of Zurich, Bluemlisalpstrasse 10,
8006 Zurich, Switzerland
e-mail: mwolf@iew.uzh.ch

Keywords Bootstrap · Subsampling · False discovery rate · Multiple testing · Stepdown procedure

Mathematics Subject Classification (2000) 62G09 · 62G10 · 62G20 · 62H15

1 Introduction

Consider the problem of testing s null hypotheses simultaneously. A classical approach to dealing with the multiplicity problem is to restrict attention to procedures that control the probability of one or more false rejections, which is called the *familywise error rate* (FWER). When s is large, however, the ability of such procedures to detect false null hypotheses is limited. For this reason, it is often preferred in such situations to relax control of the FWER in exchange for improved ability to detect false null hypotheses.

To this end, several ways of relaxing the FWER have been proposed. Hommel and Hoffman (1988) and Lehmann and Romano (2005a) consider control of the probability of k or more false rejections for some integer $k \geq 1$, which is termed the k-FWER. Obviously, controlling the 1-FWER is the same as controlling the usual FWER. Lehmann and Romano (2005a) also consider control of the *false discovery proportion* (FDP), defined to be the fraction of rejections that are false rejections (with the fraction understood to be 0 in the case of no rejections). Given a user-specified value of γ, control of the FDP means control of the probability that the FDP is greater than γ. Note that when $\gamma = 0$, control of the FDP reduces to control of the usual FWER. Methods for control of the k-FWER and the FDP based on p-values for each null hypothesis are discussed in Lehmann and Romano (2005a), Romano and Shaikh (2006a), and Romano and Shaikh (2006b). These methods are valid under weak or no assumptions on the dependence structure of the p-values, but they do not attempt to incorporate information about the dependence structure of the test statistics. Methods that incorporate such information and are thus better able to detect false null hypotheses are described in Van der Laan et al. (2004), Romano and Wolf (2007), and Romano et al. (2008).

A popular third alternative to control of the FWER is control of the *false discovery rate* (FDR), defined to be the expected value of the FDP. Control of the FDR has been suggested in a wide area of applications, such as educational evaluation (Williams et al. 1999), clinical trials (Mehrotra and Heyse 2004), analysis of microarray data (Drigalenko and Elston 1997, and Reiner et al. 2003), model selection (Abramovich and Benjamini 1996, and Abramovich et al. 2006), and plant breeding (Basford and Tukey 1997). Benjamini and Hochberg (1995) provide a method for controlling the FDR based on p-values for each null hypothesis under the assumption that the p-values are independent. Subsequent research has since shown that this procedure remains valid under weaker assumptions on the joint distribution of the p-values. Related procedures that are valid under no assumptions on the joint distribution of the p-values have also been developed; see Benjamini and Yekutieli (2001). Yet procedures for control of the FDR under weak assumptions that incorporate information about the dependence structure of the test statistics remain unavailable.

This paper seeks to develop methods for control of the FDR that incorporate such information and, by doing so, are better able to detect false null hypotheses.

The remainder of the paper is organized as follows. In Sect. 2 we describe our notation and setup. Section 3 summarizes previous research on methods for control of the FDR. In Sect. 4 we provide some motivation for our methods for control of the FDR. A bootstrap-based method is then developed in Sect. 5. The asymptotic validity of this approach relies upon an exchangeability assumption, but in Sect. 6 we develop a subsampling-based approach whose asymptotic validity does not depend on such an assumption. Section 7 sheds some light on the finite-sample performance of our methods and some previous proposals via simulations. We also provide two empirical applications in Sect. 8 to further compare the various methods. Section 9 concludes.

2 Setup and notation

A formal description of our setup is as follows. Suppose that data $X = (X_1, \ldots, X_n)$ is available from some probability distribution $P \in \Omega$. Note that we make no rigid requirements for Ω; it may be a parametric, semiparametric, or a nonparametric model. A general hypothesis H may be viewed as a subset ω of Ω. In this paper we consider the problem of simultaneously testing null hypotheses $H_i : P \in \omega_i,\ i = 1, \ldots, s$, on the basis of X. The alternative hypotheses are understood to be $H_i' : P \notin \omega_i,\ i = 1, \ldots, s$.

We assume that test statistics $T_{n,i},\ i = 1, \ldots, s$, are available for testing $H_i,\ i = 1, \ldots, s$. Large values of $T_{n,i}$ are understood to indicate evidence against H_i. Note that we may take $T_{n,i} = -\hat{p}_{n,i}$, where $\hat{p}_{n,i}$ is a p-value for H_i. A p-value for H_i may be exact, in which case $\hat{p}_{n,i}$ satisfies

$$P\{\hat{p}_{n,i} \le u\} \le u \quad \text{for any } u \in (0, 1) \text{ and } P \in \omega_i, \tag{1}$$

or asymptotic, in which case

$$\limsup_{n\to\infty} P\{\hat{p}_{n,i} \le u\} \le u \quad \text{for any } u \in (0, 1) \text{ and } P \in \omega_i. \tag{2}$$

In this article, we consider *stepdown* multiple testing procedures. Let

$$T_{n,(1)} \le \cdots \le T_{n,(s)}$$

denote the ordered test statistics (from smallest to largest), and let

$$H_{(1)}, \ldots, H_{(s)}$$

denote the corresponding null hypotheses. Stepdown multiple testing procedures first compare the most significant test statistic, $T_{n,(s)}$, with a suitable critical value c_s. If $T_{n,(s)} < c_s$, then the procedure rejects no null hypotheses; otherwise, the procedure rejects $H_{(s)}$ and then 'steps down' to the second most significant null hypothesis $H_{(s-1)}$. If $T_{n,(s-1)} < c_{s-1}$, then the procedure rejects no further null hypotheses; otherwise, the procedure rejects $H_{(s-1)}$ and then 'steps down' to the third most significant null hypothesis $H_{(s-2)}$. The procedure continues in this fashion until either

one rejects $H_{(1)}$ or one does not reject the null hypothesis under consideration. More succinctly, a stepdown multiple testing procedure rejects

$$H_{(s)}, \ldots, H_{(s-j^*)},$$

where j^* is the largest integer j that satisfies

$$T_{n,(s)} \geq c_s, \ldots, T_{n,(s-j)} \geq c_{s-j};$$

if no such j exits, the procedure does not reject any null hypotheses.

We will construct stepdown multiple testing procedures that control the *false discovery rate* (FDR), which is defined to be the expected value of the *false discovery proportion* (FDP). Denote by $I(P)$ the set of indices corresponding to true null hypotheses; that is,

$$I(P) = \{1 \leq i \leq s : P \in \omega_i\}. \tag{3}$$

For a given multiple testing procedure, let F denote the number of false rejections, and let R denote the total number of rejections; that is,

$$F = \left|\left\{1 \leq i \leq s : H_i \text{ rejected and } i \in I(P)\right\}\right|,$$
$$R = \left|\{1 \leq i \leq s : H_i \text{ rejected}\}\right|.$$

Then, the *false discovery proportion* (FDP) is defined as follows:

$$\text{FDP} = \frac{F}{\max\{R, 1\}}.$$

Using this notation, the FDR is simply $E[\text{FDP}]$. A multiple testing procedure is said to control the FDR at level α if

$$\text{FDR}_P = E_P[\text{FDP}] \leq \alpha \quad \text{for all } P \in \Omega.$$

A multiple testing procedure is said to control the FDR asymptotically at level α if

$$\limsup_{n\to\infty} \text{FDR}_P \leq \alpha \quad \text{for all } P \in \Omega. \tag{4}$$

We will say that a procedure is asymptotically valid if it satisfies (4). Methods that control the FDR can typically only be derived in special circumstances. In this paper, we will instead pursue procedures that are asymptotically valid under weak assumptions.

3 Previous methods for control of the FDR

In this section, we summarize the existing literature on methods for control of the FDR. The first known method proposed for control of the FDR is the stepwise procedure of Benjamini and Hochberg (1995) based on p-values for each null hypothesis. Let

$$\hat{p}_{n,(1)} \leq \cdots \leq \hat{p}_{n,(s)}$$

denote the ordered values of the p-values, and let

$$H_{(1)}, \ldots, H_{(s)}$$

denote the corresponding null hypotheses. Note that in this case the null hypotheses are ordered from most significant to least significant, since small values of $\hat{p}_{n,i}$ are taken to indicate evidence against H_i. For $1 \leq j \leq s$, let

$$\alpha_j = \frac{j}{s}\alpha. \tag{5}$$

Then, the method of Benjamini and Hochberg (1995) rejects null hypotheses $H_{(1)}, \ldots, H_{(j^*)}$, where j^* is the largest j such that

$$\hat{p}_{n,(j)} \leq \alpha_j.$$

Of course, if no such j exists, then the procedure rejects no null hypotheses.

Benjamini and Hochberg (1995) prove that their method controls the FDR at level α if the p-values satisfy (1) and are independent. Benjamini and Yekutieli (2001) show that independence can be replaced by a weaker condition known as positive regression dependency; see their paper for the exact definition. It can also be shown that the method of Benjamini and Hochberg (1995) provides asymptotic control of the FDR at level α if the p-values satisfy (2) instead of (1) and this weaker dependence condition holds.

On the other hand, the method of Benjamini and Hochberg (1995) fails to control the FDR at level α when the p-values only satisfy (1). Benjamini and Yekutieli (2001) show that control of the FDR can be achieved under only (1) if α_j defined in (5) are replaced by

$$\alpha_j = \frac{j}{s}\frac{\alpha}{C_s},$$

where $C_k = \sum_{r=1}^{k} \frac{1}{r}$. Note that $C_s \approx \log(s) + 0.5$, so this method can have much less power than the method of Benjamini and Hochberg (1995). For example, when $s = 1{,}000$, then $C_s = 7.49$. As before, it can be shown that this procedure provides asymptotic control of the FDR at level α if the p-values satisfy (2) instead of (1).

Even when sufficient conditions for the method of Benjamini and Hochberg (1995) to control the FDR hold, it is conservative in the following sense. It can be shown that

$$\mathrm{FDR}_P \leq \frac{s_0}{s}\alpha,$$

where $s_0 = |I(P)|$. So, unless $s_0 = s$, the power of the procedure could be improved by replacing the α_j defined in (5) by

$$\alpha_j = \frac{j}{s_0}\alpha.$$

Of course, s_0 is unknown in practice, but there exist several approaches in the literature to estimate s_0. For example, Storey et al. (2004) suggest the following estimator:

$$\hat{s}_0 = \frac{\#\{\hat{p}_{n,j} > \lambda\} + 1}{1 - \lambda}, \tag{6}$$

where $\lambda \in (0, 1)$ is a user-specified parameter. The reasoning behind this estimator is the following. As long as each test has reasonable power, then most of the "large" p-values should correspond to true null hypotheses. Therefore, one would expect about $s_0(1 - \lambda)$ of the p-values to lie in the interval $(\lambda, 1]$, assuming that the p-values corresponding to the true null hypotheses have approximately a uniform $[0, 1]$ distribution. Adding one in the numerator of (6) is a small-sample adjustment to make the procedure slightly more conservative and to avoid an estimator of zero for s_0. Having estimated s_0, one then applies the procedure of Benjamini and Hochberg (1995) with the α_j defined in (5) replaced by

$$\hat{\alpha}_j = \frac{j}{\hat{s}_0}\alpha.$$

Storey et al. (2004) prove that this adaptive procedure controls the FDR asymptotically whenever the p-values satisfy (2) and a weak dependence condition holds. This condition includes independence, dependence within blocks, and mixing-type situations, but, unlike Benjamini and Yekutieli (2001), it does not allow for arbitrary dependence among the p-values. It excludes, for example, the case in which there is a constant correlation across all p-values. Related work is found in Genovese and Wasserman (2004) and Benjamini and Hochberg (2000).

The adaptive procedure of Storey et al. (2004) can be quite liberal under positive dependence, such as in a scenario with constant positive correlation. For this reason, Benjamini et al. (2006) develop an alternative procedure, which works as follows:

Algorithm 3.1 (BKY Algorithm)

1. Apply the procedure of Benjamini and Hochberg (1995) at nominal level $\alpha^* = \alpha/(1+\alpha)$. Let r be the number of rejected hypotheses. If $r = 0$, then do not reject any hypothesis and stop; if $r = s$, then reject all s hypotheses and stop; otherwise continue.
2. Apply the procedure of Benjamini and Hochberg (1995) with the α_j defined in (5) replaced by $\hat{\alpha}_j = \frac{j}{\hat{s}_0}\alpha^*$, where $\hat{s}_0 = s - r$.

Benjamini et al. (2006) prove that this procedure controls the FDR whenever the p-values satisfy (2) and are independent of each other. They also provide simulations which suggest that this procedure continues to control the FDR under positive dependence.

Benjamini and Liu (1999) provide a stepdown method for control of the FDR based on p-values for each null hypothesis that satisfy (1) and are independent. Sarkar (2002) extends the results of Benjamini and Hochberg (1995), Benjamini and Liu (1999), and Benjamini and Yekutieli (2001) to generalized stepup–stepdown procedures; yet the methods he considers, like those described above, do not incorporate

the information about the dependence structure of the test statistics. In the following sections, we develop multiple testing procedures for asymptotic control of the FDR under weak assumptions that incorporate such information, and, by doing so, are better able to detect false hypotheses. Our procedures build upon the work of Troendle (2000), who suggests a procedure for asymptotic control of the FDR that incorporates information about the dependence structure of the test statistics, but relies upon the restrictive parametric assumption that the joint distribution of the test statistics is given by a symmetric multivariate t-distribution. Yekutieli and Benjamini (1999) also provide a method for asymptotic control of the FDR that exploits information about the dependence structure of the test statistics to improve the ability to detect false null hypotheses, but their analysis requires subset pivotality and that the test statistics corresponding to true null hypotheses are independent of those corresponding to false null hypotheses. Although our analysis will require neither of these restrictive assumptions, the asymptotic validity of our bootstrap approach will rely upon an exchangeability assumption. The subsampling approach we will develop subsequently, however, will not even require this restriction.

4 Motivation for methods

In order to motivate our procedures, first note that for any stepdown procedure based on critical values $c_1, \ldots c_s$, we have that

$$\begin{aligned}\mathrm{FDR}_P &= E_P\left[\frac{F}{\max\{R,1\}}\right] = \sum_{1\le r\le s}\frac{1}{r}E_P[F|R=r]P\{R=r\}\\ &= \sum_{1\le r\le s}\frac{1}{r}E[F|R=r]\\ &\quad\times P\{T_{n,(s)}\ge c_s,\ldots,T_{n,(s-r+1)}\ge c_{s-r+1},T_{n,(s-r)}<c_{s-r}\},\end{aligned}$$

where the event $T_{n,s-r} < c_{s-r}$ is understood to be vacuously true when $r = s$. As before, let $s_0 = |I(P)|$ and assume without loss of generality that $I(P) = \{1, \ldots, s_0\}$. Under weak assumptions, we will show that all false hypotheses will be rejected with probability tending to one. For the time being, assume that this is the case. Let $T_{n,r:t}$ denote the rth largest of the t test statistics $T_{n,1}, \ldots, T_{n,t}$; in particular, when $t = s_0$, $T_{n,r:s_0}$ denotes the rth largest of the test statistics corresponding to the true hypotheses. Then, with probability approaching one, we have that

$$\begin{aligned}\mathrm{FDR}_P &= \sum_{s-s_0+1\le r\le s}\frac{r-s+s_0}{r}\\ &\quad\times P\{T_{n,s_0:s_0}\ge c_{s_0},\ldots,T_{n,s-r+1:s_0}\ge c_{s-r+1},T_{n,s-r:s_0}<c_{s-r}\},\end{aligned}\tag{7}$$

where the event $T_{n,s-r:s_0} < c_{s-r}$ is again understood to be vacuously true when $r = s$.

Our goal is to ensure that (7) is bounded above by α for any P, at least asymptotically. To this end, first consider any P such that $s_0 = |I(P)| = 1$. Then, (7) is

simply

$$\mathrm{FDR}_P = \frac{1}{s} P\{T_{n,1:1} \geq c_1\}. \tag{8}$$

A suitable choice of c_1 is thus the smallest value for which (8) is bounded above by α; that is,

$$c_1 = \inf\left\{x \in \mathbb{R} : \frac{1}{s} P\{T_{n,1:1} \geq x\} \leq \alpha\right\}.$$

Note that if $s\alpha \geq 1$, then c_1 so defined is equal to $-\infty$.

Having determined c_1, now consider any P such that $s_0 = 2$. Then, (7) is simply

$$\frac{1}{s-1} P\{T_{n,2:2} \geq c_2, T_{n,1:2} < c_1\} + \frac{2}{s} P\{T_{n,2:2} \geq c_2, T_{n,1:2} \geq c_1\}. \tag{9}$$

A suitable choice of c_2 is therefore the smallest value for which (9) is bounded above by α.

In general, having determined $c_1, \ldots, c_{j-1}$, the jth critical value may be determined by considering P such that $s_0 = j$. In this case, (7) is simply

$$\mathrm{FDR}_P = \sum_{s-j+1 \leq r \leq s} \frac{r-s+j}{r} \times P\{T_{n,j:j} \geq c_j, \ldots, T_{n,s-r+1:j} \geq c_{s-r+1}, T_{n,s-r:j} < c_{s-r}\}. \tag{10}$$

An appropriate choice of c_j is thus the smallest value for which (10) is bounded above by α. Note that when $j = s$, (10) simplifies to

$$P\{T_{n,s:s} \geq c_s\},$$

so equivalently

$$c_s = \inf\left\{x \in \mathbb{R} : P\{T_{n,s:s} \geq x\} \leq \alpha\right\}.$$

Of course, the above choice of critical values is infeasible since it depends on the unknown P through the distribution of the test statistics. We therefore focus on feasible constructions of the critical values based on the bootstrap and subsampling.

5 A bootstrap approach

In this section, we specialize our framework to the case in which interest focuses on a parameter vector

$$\theta(P) = \big(\theta_1(P), \ldots, \theta_s(P)\big).$$

The null hypotheses may be one-sided, in which case

$$H_j : \theta_j \leq \theta_{0,j} \quad \text{vs.} \quad H_j' : \theta_j > \theta_{0,j}, \tag{11}$$

or the null hypotheses may be two-sided, in which case

$$H_j : \theta_j = \theta_{0,j} \quad \text{vs.} \quad H_j' : \theta_j \neq \theta_{0,j}. \tag{12}$$

In the next section, however, we will return to more general null hypotheses. Test statistics will be based on an estimate $\hat{\theta}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,s})$ of $\theta(P)$ computed using the data X. We will consider the 'studentized' test statistics

$$T_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j})/\hat{\sigma}_{n,j} \tag{13}$$

for the one-sided case (11) or

$$T_{n,j} = \sqrt{n}|\hat{\theta}_{n,j} - \theta_{0,j}|/\hat{\sigma}_{n,j} \tag{14}$$

for the two-sided case (12). Note that $\hat{\sigma}_{n,j}$ may either be identically equal to 1 or an estimate of the standard deviation of $\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j})$. This is done to keep the notation compact; the latter is preferable from our point of view but may not always be available in practice.

Recall that the construction of critical values in the preceding section was infeasible because of its dependence on the unknown P. For the bootstrap construction, we therefore simply replace the unknown P with a suitable estimate $\hat{P}_n$. To this end, let $X^* = (X_1^*, \ldots, X_n^*)$ be distributed according to $\hat{P}_n$ and denote by $T_{n,j}^*$, $j = 1, \ldots, s$, test statistics computed from X^*. For example, if $T_{n,j}$ is defined by (13) or (14), then

$$T_{n,j}^* = \sqrt{n}\big(\hat{\theta}_{n,j}^* - \theta_j(\hat{P}_n)\big)/\hat{\sigma}_{n,j}^* \tag{15}$$

or

$$T_{n,j}^* = \sqrt{n}\big|\hat{\theta}_{n,j}^* - \theta_j(\hat{P}_n)\big|/\hat{\sigma}_{n,j}^*, \tag{16}$$

respectively, where $\hat{\theta}_{n,j}^*$ is an estimate of θ_j computed from X^* and $\hat{\sigma}_{n,j}^*$ is either identically equal to 1 or an estimate of the standard deviation of $\sqrt{n}(\hat{\theta}_{n,j}^* - \theta_j(\hat{P}_n))$ computed from X^*. For the validity of this approach, we require that the distribution of $T_{n,j}^*$ provides a good approximation to the distribution of $T_{n,j}$ whenever the corresponding null hypothesis H_j is true, but, unlike Westfall and Young (1993), we do not require subset pivotality. The exact choice of $\hat{P}_n$ will, of course, depend on the nature of the data. If the data $X = (X_1, \ldots, X_n)$ are i.i.d., then a suitable choice of $\hat{P}_n$ is the empirical distribution, as in Efron (1979). If, on the other hand, the data constitute a time series, then $\hat{P}_n$ should be estimated using a suitable time series bootstrap method; see Lahiri (2003) for details.

Given a choice of $\hat{P}_n$, define the critical values recursively as follows: having determined $\hat{c}_{n,1}, \ldots, \hat{c}_{n,j-1}$, compute $\hat{c}_{n,j}$ according to the rule

$$\hat{c}_{n,j} = \inf\Bigg\{ c \in \mathbb{R} : \sum_{s-j+1 \le r \le s} \frac{r-s+j}{r} \times \hat{P}_n\{T_{n,j:j}^* \ge c, \ldots, T_{n,s-r+1:j}^* \ge \hat{c}_{n,s-r+1}, T_{n,s-r:j}^* < \hat{c}_{n,s-r}\} \le \alpha \Bigg\}. \tag{17}$$

Remark 1 It is important to be clear about the meaning of the notation $T^*_{n,r:t}$, with $r \leq t$, in (17). By analogy to the "real" world, it should denote the rth smallest of the observations corresponding to the first t true null hypotheses. However, the ordering of the true null hypotheses in the bootstrap world is not $1, 2, \ldots, s$, but it is instead determined by the ordering $H_{(1)}, \ldots, H_{(s)}$ from the real world. So if the permutation $\{k_1, \ldots, k_s\}$ of $\{1, \ldots, s\}$ is defined such that $H_{k_1} = H_{(1)}, \ldots, H_{k_s} = H_{(s)}$, then $T^*_{n,r:t}$ is the rth smallest of the observations $T^*_{n,k_1}, \ldots, T^*_{n,k_t}$.

Remark 2 Note that typically it will not be possible to compute closed form expressions for the probabilities under $\hat{P}_n$ required in (17). In such cases, the required probabilities may instead be computed using simulation to any desired degree of accuracy.

We now provide conditions under which the stepdown procedure with critical values defined by (17) satisfies (4). The following result applies to the case of two-sided null hypotheses, but the one-sided case can be handled using a similar argument. In order to state the result, we will require some further notation. For $K \subseteq \{1, \ldots, s\}$, let $J_{n,K}(P)$ denote the joint distribution of

$$\big(\sqrt{n}\big(\hat{\theta}_{n,j} - \theta_j(P)\big)/\hat{\sigma}_{n,j} : j \in K\big).$$

It will also be useful to define the quantile function corresponding to a c.d.f. $G(\cdot)$ on $\mathbb{R}$ as $G^{-1}(\alpha) = \inf\{x \in \mathbb{R} : G(x) \geq \alpha\}$.

Theorem 1 *Consider the problem of testing the null hypotheses H_i, $i = 1, \ldots, s$, given by (12) using test statistics $T_{n,i}$, $i = 1, \ldots, s$, defined by (14). Suppose that $J_{n,\{1,\ldots,s\}}(P)$ converges weakly to a limit law $J_{\{1,\ldots,s\}}(P)$, so that $J_{n,I(P)}(P)$ converges weakly to a limit law $J_{I(P)}(P)$. Suppose further that $J_{I(P)}(P)$*

(i) *Has continuous one-dimensional marginal distributions*
(ii) *Has connected support, which is denoted by supp$(J_{I(P)}(P))$*
(iii) *Is exchangeable*

Also, assume that

$$\hat{\sigma}_{n,j} \xrightarrow{P} \sigma_j(P),$$

where $\sigma_j(P) > 0$ is nonrandom. Let $\hat{P}_n$ be an estimate of P such that

$$\rho\big(J_{n,\{1,\ldots,s\}}(P), J_{n,\{1,\ldots,s\}}(\hat{P}_n)\big) \xrightarrow{P} 0, \tag{18}$$

where ρ is any metric metrizing weak convergence in $\mathbb{R}^s$.

Then, for the stepdown method with critical values defined by (17),

$$\limsup_{n\to\infty} \mathrm{FDR}_P \leq \alpha.$$

We will make use of the following lemma in our proof of the preceding theorem:

Lemma 1 *Let X be a random vector on $\mathbb{R}^s$ with distribution P. Define $f : \mathbb{R}^s \to \mathbb{R}$ by the rule $f(x) = x_{(k)}$ for some fixed $1 \le k \le s$, where*

$$x_{(1)} \le \cdots \le x_{(s)}.$$

Suppose that (i) *the one-dimensional marginal distributions of P have continuous c.d.f.s and* (ii) *supp(X) is connected. Then, $f(X)$ has a continuous and strictly increasing c.d.f.*

Proof To see that the c.d.f. of $f(X)$ is continuous, simply note that

$$P\big\{f(X) = x\big\} \le \sum_{1 \le i \le s} P\{X_i = x\} = 0,$$

where the final equality follows from assumption (i). To see that the c.d.f. of $f(X)$ is strictly increasing, suppose by way of contradiction that there exists $a < b$ such that $P\{f(X) \in (a, b)\} = 0$, but $P\{f(X) \le a\} > 0$ and $P\{f(X) \ge b\} > 0$. Thus, there exists $x \in \text{supp}(X)$ such that $f(x) \le a$ and $x' \in \text{supp}(X)$ such that $f(x') \ge b$. Consider the set

$$A_{a,b} = \big\{x \in \text{supp}(X) : a < f(x) < b\big\}.$$

By the continuity of $f(x)$ and assumption (ii), $A_{a,b}$ is nonempty. Moreover, again by the continuity of $f(x)$, $A_{a,b}$ must contain an open subset of supp(X) (relative to the topology on supp(X)). It therefore follows by the definition of supp(X) that

$$P\{X \in A_{a,b}\} = P\big\{f(X) \in (a, b)\big\} > 0,$$

which yields the desired contradiction. □

Remark 3 An important special case of Lemma 1 is the case in which X is distributed as a multivariate normal random vector with mean μ and covariance matrix Σ. In this case, assumptions (i)–(ii) of the lemma are implied by the very mild restriction that $\Sigma_{i,i} > 0$ for $1 \le i \le s$. In particular, it is not even necessary to assume that Σ is nonsingular.

Remark 4 Note that even in the case in which $s = 1$, so $f(x) = x$, both assumptions (i) and (ii) in Lemma 1 are necessary to conclude that the distribution of $f(X)$ is continuous and strictly increasing. Therefore, the assumptions used in Lemma 1 seem as weak as possible.

Proof of Theorem 1 Without loss of generality, suppose that $H_1, \ldots, H_{s_0}$ are all true and the remainder false.

In order to illustrate better the main ideas of the proof, we first consider the case in which P is such that the number of true hypotheses is $s_0 = 1$. The initial step in our argument is to show that all false null hypotheses are rejected with probability tending to 1. Since $\theta_j(P) \ne \theta_{0,j}$ for $j \ge 2$, it follows that

$$T_{n,j} = n^{1/2}|\hat{\theta}_{n,j} - \theta_{0,j}|/\hat{\sigma}_{n,j} \xrightarrow{P} \infty$$

for $j \geq 2$. On the other hand, for $j = 1$, we have that

$$T_{n,j} = O_P(1).$$

Therefore, to show that all false hypotheses are rejected with probability tending to one, it suffices to show that the critical values $\hat{c}_{n,j}$ are all uniformly bounded above in probability for $j \geq 2$.

Recall that $\hat{c}_{n,j}$ is defined as follows: having determined $\hat{c}_{n,1}, \ldots, \hat{c}_{n,j-1}$, $\hat{c}_{n,j}$ is the infimum over all $c \in \mathbb{R}$ for which

$$\sum_{s-j+1 \leq r \leq s} \frac{r-s+j}{r} \hat{P}_n\{T^*_{n,j:j} \geq c, \ldots, T^*_{n,s-r+1:j} \geq \hat{c}_{n,s-r+1}, T^*_{n,s-r:j} < \hat{c}_{n,s-r}\} \tag{19}$$

is bounded above by α. Note that (19) can be bounded above by

$$j \hat{P}_n\{T^*_{n,j:j} \geq c\},$$

which can in turn be bounded above by

$$s \hat{P}_n\{T^*_{n,s:s} \geq c\}. \tag{20}$$

It follows that the set of $c \in \mathbb{R}$ for which (20) is bounded above by α is a subset of the set of $c \in \mathbb{R}$ for which (19) is bounded above by α. Therefore, $\hat{c}_{n,j}$ is bounded above by the $1-\alpha/s$ quantile of the (centered) bootstrap distribution of the maximum of all s variables. In order to describe the asymptotic behavior of this bootstrap quantity, let

$$M_n(x, P) = P\left\{ \max_{1 \leq j \leq s} \{n^{1/2} |\hat{\theta}_{n,j} - \theta_j| / \hat{\sigma}_{n,j}\} \leq x \right\},$$

and let $\hat{M}_n(x)$ denote the corresponding bootstrap c.d.f. given by

$$\hat{P}_n\left\{ \max_{1 \leq j \leq s} \{n^{1/2} |\hat{\theta}^*_{n,j} - \theta_j(\hat{P}_n)| / \hat{\sigma}^*_{n,j}\} \leq x \right\}.$$

In this notation, the previously derived bound for $\hat{c}_{n,j}$ may be restated as

$$\hat{c}_{n,j} \leq \hat{M}_n^{-1}\left(1 - \frac{\alpha}{s}\right).$$

By the Continuous Mapping Theorem, $M_n(\cdot, P)$ converges in distribution to a limit distribution $M(\cdot, P)$, and the assumptions imply that this limiting distribution is continuous. Choose $0 < \epsilon < \frac{\alpha}{s}$ so that $M(\cdot, P)$ is strictly increasing at $M^{-1}(1 - \frac{\alpha}{s} + \epsilon, P)$. For such an ϵ,

$$\hat{M}_n^{-1}\left(1 - \frac{\alpha}{s} + \epsilon\right) \xrightarrow{P} M^{-1}\left(1 - \frac{\alpha}{s} + \epsilon, P\right).$$

Therefore, $\hat{c}_{n,j}$ is with probability tending to one less than $M^{-1}(1 - \frac{\alpha}{s} + \epsilon, P)$. The claim that $\hat{c}_{n,j}$ is bounded above in probability is thus verified.

It now follows that, in the case $s_0 = 1$,

$$\mathrm{FDR}_P = \frac{1}{s} P\{T_{n,1} \geq \hat{c}_{n,1}\} + o_P(1).$$

The critical value $\hat{c}_{n,1}$ is the $1 - \alpha s$ quantile of the distribution of $T^*_{n,1}$ under $\hat{P}_n$. If $1 - \alpha s \leq 0$, then $\hat{c}_{n,1}$ is defined to be $-\infty$, in which case,

$$\mathrm{FDR}_P = \frac{1}{s} + o_P(1) \leq \alpha + o_P(1).$$

The desired conclusion thus holds. If, on the other hand, $1 - \alpha s > 0$, then we argue as follows. Note that by assumption (18) and the triangle inequality, we have that

$$\rho\big(J_{\{1\}}(P), J_{n,\{1\}}(\hat{P}_n)\big) \xrightarrow{P} 0.$$

Note further that by Lemma 1, $J_{\{1\}}(\cdot, P)$ is strictly increasing at $J^{-1}_{\{1\}}(1 - s\alpha, P)$. Thus,

$$\hat{c}_{n,1} \xrightarrow{P} J^{-1}_{\{1\}}(1 - s\alpha, P).$$

To establish the desired result, it now suffices to use Slutsky's Theorem.

We now proceed to the general case. First, the same argument as in the case $s_0 = 1$ shows that hypotheses $H_{s_0+1}, \ldots, H_s$ are rejected with probability tending to one. It follows that with probability tending to one, the FDR_P is equal to

$$\sum_{s-s_0+1 \leq r \leq s} \frac{r - s + s_0}{r} \times P\{T_{n,s_0:s_0} \geq \hat{c}_{n,s_0}, \ldots, T_{n,s-r+1:s_0} \geq \hat{c}_{n,s-r+1}, T_{n,s-r:j} < \hat{c}_{n,s-r}\},$$

where the event $T_{n,s-r:j} < \hat{c}_{n,s-r}$ is understood to be vacuously true when $r = s$.

In the definition of the critical values given by (17), recall that $T^*_{n,r:t}$ is defined to be the rth smallest of the bootstrap test statistics among those corresponding to the smallest t original test statistics. Define $T'_{n,r:t}$ to be the rth smallest of the bootstrap test statistics among those corresponding to the first t original test statistics. Define $c'_{n,j}$ to be the critical values defined in the same way as $\hat{c}_{n,j}$ except $T^*_{n,r:t}$ in (17) is replaced with $T'_{n,r:t}$. Recall that we have assumed that null hypotheses $H_1, \ldots, H_{s_0}$ are true and the remainder false. Since the indices of the set of s_0 true hypotheses are identical to the indices corresponding to the smallest s_0 test statistics with probability tending to one, $\hat{c}_{n,j}$ equals $c'_{n,i}$ with probability tending to 1 for $j \leq s_0$. It follows that with probability tending to one, the FDR_P is equal to

$$\sum_{s-s_0+1 \leq r \leq s} \frac{r - s + s_0}{r} \times P\{T_{n,s_0:s_0} \geq c'_{n,s_0}, \ldots, T_{n,s-r+1:s_0} \geq c'_{n,s-r+1}, T_{n,s-r:j} < c'_{n,s-r}\},$$

where, as before, the event $T_{n,s-r:j} < c'_{n,s-r}$ is understood to be vacuously true when $r = s$.

In order to describe the asymptotic behavior of these critical values, let $(T_1, \ldots, T_{s_0})$ be a random vector with distribution $J_{I(P)}(P)$ and define $T_{r:t}$ to be the rth smallest of $T_1, \ldots, T_t$. Define $c_1, \ldots, c_{s_0}$ recursively as follows: having determined $c_1, \ldots, c_{j-1}$, compute c_j according to the rule

$$c_j = \inf\Bigg\{c \in \mathbb{R} : \sum_{1 \le k \le j} \frac{k}{s - j + k} \times P\{T_{j:s_0} \ge c, \ldots, T_{j-k+1:s_0} \ge c_{j-k+1}, T_{j-k:s_0} < c_{j-k}\} \le \alpha\Bigg\},$$

where, as before, the event $T_{j-k:s_0} < c_{j-k}$ is understood to be vacuously true when $k = j$. We claim for $1 \le j \le s_0$ that

$$c'_{n,j} \xrightarrow{P} c_j. \tag{21}$$

To see this, we argue inductively as follows. Suppose that the result is true for $c'_{n,1}, \ldots, c'_{n,j-1}$. Using assumption (18) and the triangle inequality, we have that

$$\rho\big(J_{\{1,\ldots,j\}}(P), J_{n,\{1,\ldots,j\}}(\hat{P}_n)\big) \xrightarrow{P} 0.$$

Importantly, by the assumption of exchangeability, we have that $J_{\{1,\ldots,j\}}(P) = J_K(P)$ for any $K \subseteq \{1, \ldots, s_0\}$ such that $|K| = j$. Next note that

$$\sum_{1 \le k \le j} P\{T_{j:s_0} \ge c, \ldots, T_{j-k+1:s_0} \ge c_{j-k+1}, T_{j-k:s_0} < c_{j-k}\} = P\{T_{j:s_0} \ge c\}. \tag{22}$$

The right-hand side of (22) is strictly increasing in c by Lemma 1. As a result, at least one of the terms on the left-hand side of (22) is strictly increasing at $c = c_j$. It follows that

$$\sum_{1 \le k \le j} \frac{k}{s - j + k} P\{T_{j:s_0} \ge c, \ldots, T_{j-k+1:s_0} \ge c_{j-k+1}, T_{j-k:s_0} < c_{j-k}\}$$

is strictly increasing at $c = c_j$. The conclusion (21) thus follows. To complete the proof, it now suffices to use Slutsky's Theorem. □

Remark 5 In the definitions of $T^*_{n,j}$ given by (15) or (16) used in our bootstrap method to generate the critical values, one can typically replace $\theta_j(\hat{P}_n)$ by $\hat{\theta}_{n,j}$. Of course, the two are the same under the following conditions: (1) $\hat{\theta}_{n,j}$ is a linear statistic; (2) $\theta_j(P) = \mathrm{E}(\hat{\theta}_{n,j})$; and (3) $\hat{P}_n$ is based on Efron's bootstrap, the circular blocks bootstrap, or the stationary bootstrap in Politis and Romano (1994). Even if conditions (1) and (2) are met, the estimators $\hat{\theta}_{n,j}$ and $\theta_j(\hat{P}_n)$ are not the same if $\hat{P}_n$ is based on the moving blocks bootstrap due to "edge effects." On the other hand, the substitution of $\hat{\theta}_{n,j}$ for $\theta_j(\hat{P}_n)$ does not in general affect the asymptotic validity of the bootstrap approximation, and Theorem 1 continues to hold. Lahiri (1992) discusses this point for the special case of time series data and the sample mean. Still

another possible substitute is $E[\hat{\theta}^*_{n,j}|\hat{P}_n]$, but generally these are all first-order asymptotically equivalent. In the simulations of Sect. 7 and the empirical application of Sect. 8, conditions (1)–(3) always hold, and so we can simply use $\hat{\theta}_{n,j}$ for the centering throughout.

6 A subsampling approach

In this section, we describe a subsampling-based construction of critical values for use in a stepdown procedure that provides asymptotic control of the FDR. Here, we will no longer be assuming that interest focuses on null hypotheses about a parameter vector $\theta(P)$, but we will instead return to considering more general null hypotheses. Moreover, we will no longer require that the limiting joint distribution of the test statistics corresponding to true null hypotheses be exchangeable. Finally, as is usual with arguments based on subsampling, we only require a limiting distribution under the true distribution of the observed data, unlike the bootstrap, which requires (18).

In order to describe our approach, we will use the following notation. For $b < n$, let $N_n = \binom{n}{b}$, and let $T_{n,b,i,j}$ denote the statistic $T_{n,j}$ evaluated at the ith subset of data of size b. Let $T_{n,b,i,r:t}$ denote the tth largest of the test statistics

$$T_{n,b,i,1}, \ldots, T_{n,b,i,t}.$$

Finally, define critical values $\hat{c}_{n,1}, \ldots, \hat{c}_{n,s}$ recursively as follows: having determined $\hat{c}_{n,1}, \ldots, \hat{c}_{n,j-1}$, compute $\hat{c}_{n,j}$ according to the rule

$$\begin{aligned}\hat{c}_{n,j} = \inf\Bigg\{ c \in \mathbb{R} : \frac{1}{N_n} \sum_{1 \le i \le N_n} \sum_{1 \le k \le j} \frac{k}{s-j+k} \\ \times I\{T_{n,b,i,j:s} \ge c, \ldots, T_{n,b,i,j-k+1:s} \\ \ge \hat{c}_{n,j-k+1}, T_{n,b,i,j-k:s} < \hat{c}_{n,j-k}\} \le \alpha \Bigg\},\end{aligned} \tag{23}$$

where the event $T_{n,b,i,j-k:s} < \hat{c}_{n,j-k}$ is understood to be vacuously true when $k = j$. We now provide conditions under which the stepdown procedure with this choice of critical values is asymptotically valid.

Theorem 2 *Suppose that the data $X = (X_1, \ldots X_n)$ is an i.i.d. sequence of random variables with distribution P. Consider testing null hypotheses $H_j : P \in \omega_j$, $j = 1, \ldots, s$, with test statistics $T_{n,j}$, $j = 1, \ldots, s$. Suppose that $J_{n,I(P)}(P)$, the joint distribution of $(T_{n,j} : j \in I(P))$, converges weakly to a limit law $J_{I(P)}(P)$ for which*

(i) *The one-dimensional marginal distributions of $J_{I(P)}(P)$ have continuous c.d.f.s*
(ii) *$\mathrm{supp}(J_{I(P)}(P))$ is connected*

Suppose further that $T_{n,j} = \tau_n t_{n,j}$ and $t_{n,j} \xrightarrow{P} t_j(P)$, where $t_j(P) > 0$ if $P \in \omega_j$ and $t_j(P) = 0$ otherwise. Let $b = b_n < n$ be a nondecreasing sequence of positive integers such that $b/n \to 0$ and $\tau_b/\tau_n \to 0$. Then, the stepdown procedure with critical values

defined by (23) *satisfies*

$$\limsup_{n\to\infty} \mathrm{FDR}_P \le \alpha.$$

Proof We first argue that all false null hypotheses are rejected with probability tending to one. Let $s_0 = |I(P)|$ and, without loss of generality, order the test statistics so that $T_{n,1}, \ldots, T_{n,s_0}$ correspond to the true null hypotheses. Suppose that there is at least one false null hypothesis, for otherwise there is nothing to show, and note that

$$\begin{aligned} &I\{T_{n,b,i,j:s} \ge c, \ldots, T_{n,b,i,j-k+1:s} \ge \hat{c}_{n,j-k+1}, T_{n,b,i,j-k:s} < \hat{c}_{n,j-k}\} \\ &\quad \le I\{T_{n,b,i,j:s} \ge c\}. \end{aligned}$$

Since $\frac{k}{s-j+k} \le 1$, it follows that

$$\hat{c}_{n,j} \le \inf\left\{c \in \mathbb{R} : \frac{1}{N_n} \sum_{1\le i\le N_n} j I\{T_{n,b,i,j:s} \ge c\} \le \alpha\right\},$$

which may in turn be bounded by

$$\begin{aligned} &\inf\left\{c \in \mathbb{R} : \frac{1}{N_n} \sum_{1\le i\le N_n} s I\{T_{n,b,i,s:s} \ge c\} \le \alpha\right\} \\ &\quad = \tau_b \inf\left\{c \in \mathbb{R} : \frac{1}{N_n} \sum_{1\le i\le N_n} I\{t_{n,b,i,s:s} \ge c\} \le \frac{\alpha}{s}\right\}, \end{aligned}$$

where $t_{n,b,i,r:t}$ is defined analogously to $T_{n,b,i,r:t}$. Following the proof of Theorem 2.6.1 in Politis et al. (1999), we have that

$$\inf\left\{c \in \mathbb{R} : \frac{1}{N_n} \sum_{1\le i\le N_n} I\{t_{n,b,i,s:s} \ge c\} \le \frac{\alpha}{s}\right\} \xrightarrow{P} \max_{1\le j\le s} t_j(P) > 0,$$

where the final inequality follows from the assumption that there is at least one false null hypothesis. Now, consider any $T_{n,j}$ corresponding to a false null hypothesis. Since $t_{n,j} \xrightarrow{P} t_j(P) > 0$ and $\tau_b/\tau_n \to 0$, it follows that

$$T_{n,j} = \tau_n t_{n,j} > \tau_b \inf\left\{c \in \mathbb{R} : \frac{1}{N_n} \sum_{1\le i\le N_n} I\{t_{n,b,i,s:s} \ge c\} \le \frac{\alpha}{s}\right\},$$

and thus exceeds all critical values, with probability approaching 1. The desired result is therefore established.

It follows that with probability approaching 1, we have that

$$\begin{aligned} \mathrm{FDR}_P = &\sum_{1\le k\le s_0} \frac{k}{s - s_0 + k} \\ &\times P\{T_{n,s_0:s_0} \ge \hat{c}_{n,s_0}, \ldots, T_{n,s_0-k+1:s_0} \ge \hat{c}_{n,s_0-k+1}, T_{n,s_0-k:s_0} < \hat{c}_{n,s_0-k}\}, \end{aligned}$$

where the event $T_{n,s_0-k:s_0} < \hat{c}_{n,s_0-k}$ is again understood to be vacuously true when $k = s_0$. In order to describe the asymptotic behavior of this expression, let $(T_1, \ldots, T_{s_0})$ be a random vector with distribution $J_{I(P)}(P)$ and define $T_{r:t}$ to be the rth largest of $T_1, \ldots, T_t$. Define $c_1, \ldots, c_{s_0}$ recursively according to the rule

$$c_j = \inf\Bigg\{c \in \mathbb{R} : \sum_{1 \le k \le j} \frac{k}{s-j+k} \times P\{T_{j:s_0} \ge c, \ldots, T_{j-k+1:s_0} \ge c_{j-k+1}, T_{j-k:s_0} < c_{j-k}\} \le \alpha\Bigg\},$$

where, as before, the event $T_{j-k:s_0} < c_{j-k}$ is understood to be vacuously true when $k = j$. By the same argument used in the proof of Theorem 1, we have by Lemma 1 that

$$\sum_{1 \le k \le j} \frac{k}{s-j+k} P\{T_{j:s_0} \ge c, \ldots, T_{j-k+1:s_0} \ge c_{j-k+1}, T_{j-k:s_0} < c_{j-k}\}$$

is continuous and strictly increasing at $c = c_j$. We may therefore argue inductively that for $1 \le j \le s_0$, we have that

$$\hat{c}_{n,j} \xrightarrow{P} c_j.$$

An appeal to Slutsky's theorem completes the argument. □

Remark 6 At the expense of a much more involved argument, it is in fact possible to remove the assumption that $\mathrm{supp}(J_{I(P)}(P))$ is connected. However, we know of no example where this mild assumption fails.

Remark 7 The above approach can be extended to dependent data as well. For example, if the data $X = (X_1, \ldots, X_n)$ form a stationary sequence, we would only consider the $n - b + 1$ subsamples of the form $(X_i, X_{i+1}, \ldots, X_{i+b-1})$. Generalizations for nonstationary time series, random fields, and point processes are further discussed in Politis et al. (1999).

Remark 8 Interestingly, even under the exchangeability assumption and the setup of Sect. 5, where both the bootstrap and subsampling are asymptotically valid, the two procedures are not asymptotically equivalent. To see this, suppose that $s = s_0 = 2$ and that the joint limiting distribution of the test statistics is (T_1, T_2), where $T_i \sim N(0, \sigma_i^2)$, $\sigma_1 = \sigma_2$, and T_1 is independent of T_2. Then, the bootstrap critical value $\hat{c}_{n,1}$ tends in probability to $z_{1-\alpha}$, while the corresponding subsampling critical value tends in probability to the $1 - \alpha$ quantile of $\min\{T_1, T_2\}$, which will be strictly less than $z_{1-\alpha}$.

If the exchangeability assumption fails, i.e., $\sigma_1 \neq \sigma_2$, then the subsampling critical value still tends in probability to the $1 - \alpha$ quantile of $\min\{T_1, T_2\}$. The bootstrap critical value, however, does not even settle down asymptotically. Indeed, in this case, it tends in probability to $z_{1-\alpha}\sigma_1$ with probability $P\{T_1 < T_2\}$ and to $z_{1-\alpha}\sigma_2$ with probability $P\{T_1 \ge T_2\}$.

7 Simulations

Since the proof of the validity of our stepdown procedure relies on asymptotic arguments, it is important to shed some light on finite sample performance via some simulations. Therefore, this section presents a small simulation study in the context of testing population means.

7.1 Comparison of FDR control and power

We generate random vectors $X_1, \ldots, X_n$ from an s-dimensional multivariate normal distribution with mean vector $\theta = (\theta_1, \ldots, \theta_s)$, where $n = 100$ and $s = 50$. The null hypotheses are $H_j : \theta_j \leq 0$, and the alternative hypotheses are $H_j' : \theta_j > 0$. The test statistics are $T_{n,j} = \sqrt{n}\hat{\theta}_{n,j}/\hat{\sigma}_{n,j}$, where

$$\hat{\theta}_{n,j} = \frac{1}{n}\sum_{i=1}^{n} X_{i,j} \quad \text{and} \quad \hat{\sigma}_{n,j}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i,j} - \hat{\theta}_{n,j})^2,$$

that is, we employ the usual t-statistics.

We consider three models for the covariance matrix Σ having (i, j) component $\sigma_{i,j}$. The models share the feature $\sigma_{i,i} = 1$ for all i; so we are left to specify $\sigma_{i,j}$ for $i \neq j$.

- Common correlation: $\sigma_{i,j} = \rho$, where $\rho = 0, 0.5$, or 0.9.
- Power structure: $\sigma_{i,j} = \rho^{|i-j|}$, where $\rho = 0.95$.
- Two-class structure: the variables are grouped in two classes of equal size $s/2$. Within each class, there is a common correlation of $\rho = 0.5$; and across classes, there is a common correlation of $\rho = -0.5$. Formulated mathematically, for $i \neq j$,

$$\sigma_{i,j} = \begin{cases} 0.5 & \text{if both } i, j \in \{1, \ldots, s/2\} \text{ or both } i, j \in \{s/2+1, \ldots, s\}, \\ -0.5 & \text{otherwise.} \end{cases}$$

We consider four scenarios for the mean vector $\theta = (\theta_1, \ldots, \theta_s)$.

- All $\theta_j = 0$.
- Every fifth $\theta_j = 0.2$, and the remaining $\theta_j = 0$, so there are ten $\theta_j = 0.2$.
- Every other $\theta_j = 0.2$, and the remaining $\theta_j = 0$, so there are twenty five $\theta_j = 0.2$.
- All $\theta_j = 0.2$

We include the following FDR controlling procedures in the study.

- (BH) The procedure of Benjamini and Hochberg (1995).
- (STS) The adaptive BH procedure by Storey et al. (2004). Analogously to their simulation study, we use $\lambda = 0.5$ for the estimation of s_0.
- (BKY) The adaptive BH procedure of Benjamini et al. (2006) detailed in Algorithm 3.1. Among all the adaptive procedures employed in the simulations of Benjamini et al. (2006), this is the only one that controls the FDR under positive dependence.

Table 1 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS	BKY	Boot	BH	STS	BKY	Boot	BH	STS	BKY	Boot
All $\theta_j = 0$												
Control	10.0	10.3	9.1	10.0	6.4	16.5	6.0	9.9	4.8	32.8	4.4	9.8
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_j = 0.2$												
Control	7.6	9.5	7.3	7.3	6.4	16.9	7.5	9.3	5.0	26.5	5.8	10.0
Rejected	3.4	3.8	3.4	3.4	3.5	4.2	3.5	4.1	3.7	4.5	3.7	6.0
Twenty five $\theta_j = 0.2$												
Control	5.0	9.5	6.2	6.7	4.3	13.9	7.4	8.9	3.9	18.3	7.1	9.5
Rejected	13.2	17.4	14.5	14.9	12.3	15.1	13.1	14.1	12.6	14.2	12.7	16.6
All $\theta_j = 0.2$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	34.8	49.7	44.9	48.2	31.9	46.9	36.4	39.1	32.1	47.3	32.1	36.4

– (Boot) The bootstrap procedure of Sect. 5. Since the data are i.i.d., we use Efron's (1979) bootstrap with $B = 500$ resamples.

The p-values for use in BH, STS, and BKY are computed as $\hat{p}_{n,j} = 1 - \Psi_{99}(T_{n,j})$, where $\Psi_k(\cdot)$ denotes the c.d.f. of the t-distribution with k degrees of freedom.

We also experimented with the subsampling procedure of Section 6, but the results were not very satisfactory. Apparently, sample sizes larger than $n = 100$ are needed for the subsampling procedure to be employed.

The performance criteria are (1) the empirical FDR compared to the nominal level $\alpha = 0.1$; and (2) the empirical power (measured as the average number of false hypotheses rejected). The results are presented in Table 1 (for common correlation) and Table 2 (for power structure and two-class structure). They can be summarized as follows.

– BH, BKY, and Boot provide satisfactory control of the FDR in all scenarios. On the other hand, STS is liberal under positive constant correlation and for the power structure scenario.
– For the five scenarios with ten $\theta_j = 0.2$, BKY is as powerful as BH, while in all other scenarios it is more powerful. In return, for the single scenario with ten $\theta_j = 0.2$ under independence, Boot is as powerful as BKY, while in all other scenarios it is more powerful.
– In the majority of scenarios, the empirical FDR of Boot is closest to the nominal level $\alpha = 0.1$.
– STS is often more powerful than Boot, but some of those comparisons are not meaningful, namely when Boot provides FDR control while STS does not.

Table 2 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is $B = 500$

	Power structure				Two-class structure			
	BH	STS	BKY	Boot	BH	STS	BKY	Boot
All $\theta_j = 0$								
Control	5.4	16.5	4.9	10.2	8.1	7.9	7.5	10.1
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_j = 0.2$								
Control	6.5	17.0	7.4	9.8	6.8	8.0	6.9	8.3
Rejected	3.5	4.2	3.5	4.7	3.2	3.7	3.2	3.6
Twenty five $\theta_j = 0.2$								
Control	4.3	13.9	7.4	9.1	5.0	9.3	6.3	7.4
Rejected	12.3	15.0	13.1	14.8	13.1	17.5	14.3	15.3
All $\theta_j = 0.2$								
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	32.0	47.1	36.0	38.7	35.2	48.8	44.5	47.3

7.2 Robustness of FDR control against random correlations

In the previous subsection, we used three models for the covariance matrix: constant correlation, power structure, and two-class structure. In all cases, BH, BKY, and Boot provided satisfactory control of the FDR in finite samples.

The goal of this subsection is to study whether FDR control is maintained for 'general' covariance matrices. Since it is impossible to employ all possible covariance matrices in a simulation study, our approach is to employ a large, albeit random, 'representative' subset of covariance matrices. To this end, we generate 1,000 random correlation matrices uniformly from the space of positive definite correlation matrices. Joe (2006) recently introduced a new method which accomplishes this. Computationally more efficient variants are provided by Lewandowski et al. (2007), and we use their programming code which Prof. Joe has graciously shared with us.) We then simulate the FDR for each resulting covariance matrix, taking all standard deviations to be equal to one. However, we reduce the dimension from $s = 50$ to $s = 4$ to counter the curse of dimensionality. Note that an s-dimensional correlation matrix lives in a space of dimension $(s-1)s/2$. Since we can only consider a finite number of random correlation matrices, we 'cover' this space more thoroughly when a smaller value of s is chosen. As far as the mean vector is concerned, two scenarios are considered: one $\theta_j = 0.2$ and one $\theta_j = 20$. The latter scenario results in perfect power for all four methods.

The resulting 1,000 simulated FDRs for each method and each mean scenario are displayed via boxplots in Fig. 1. Again, BH, BKY, and Boot provide satisfactory control of the FDR throughout, while STS is generally liberal. In addition, Boot tends to provide FDR control closest to the nominal level $\alpha = 0.1$, followed by BKY and BH.

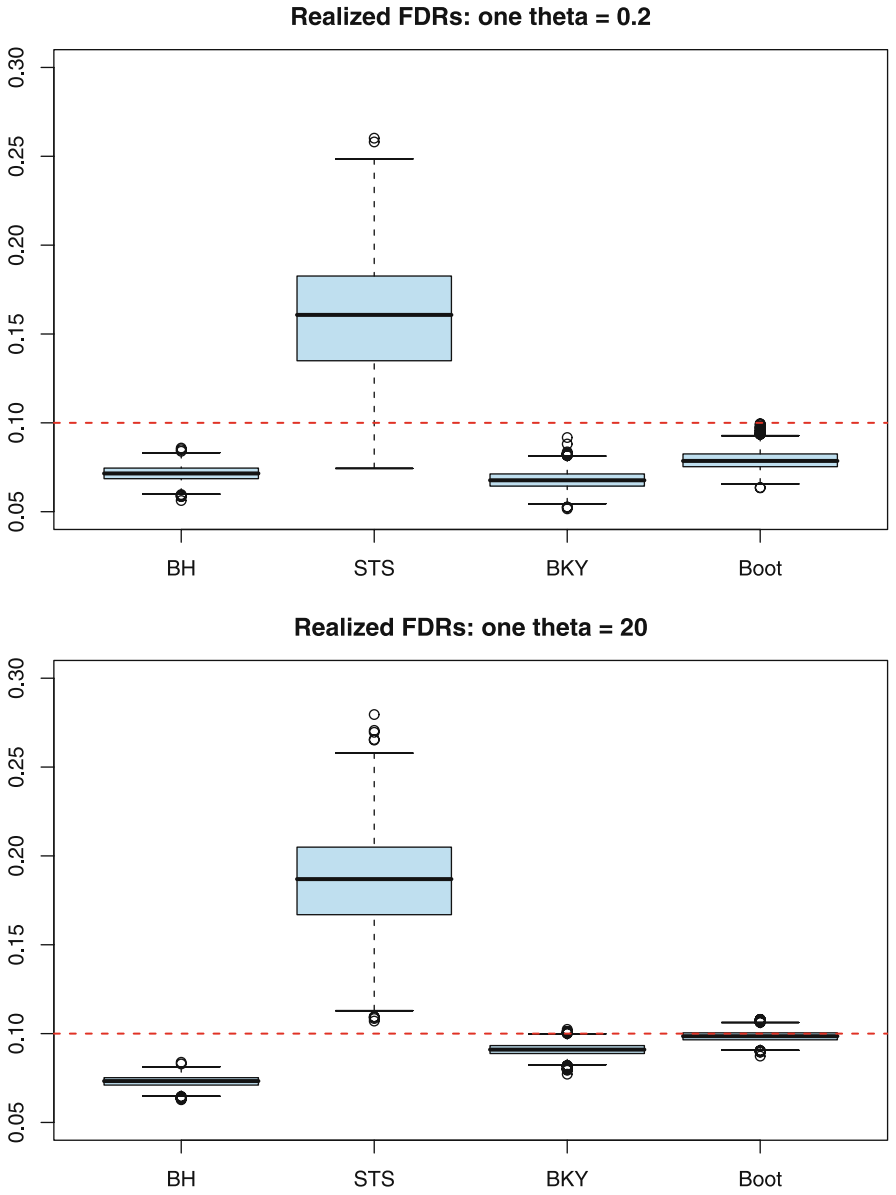

Fig. 1 Boxplots of the simulated FDRs described in Sect. 7.2. The *horizontal dashed lines* indicate the nominal level $\alpha = 0.1$

We also experimented with a larger value of s and different fractions of false null hypotheses. The results (not reported) were qualitatively similar. In particular, we could not find a constellation where any of BH, BKY, or Boot were liberal.

Table 3 Number of outperforming funds identified

Procedure	$\alpha = 0.05$	$\alpha = 0.1$
BH	58	101
STS	173	203
BKY	72	142
Boot	81	129

8 Empirical applications

8.1 Hedge fund evaluation

We revisit the data set of Romano et al. (2008) concerning the evaluation of hedge funds. There are $s = 209$ hedge funds with a return history of $n = 120$ months compared to the risk-free rate as a common benchmark. The parameters of interest are $\theta_j = \mu_j - \mu_B$, where μ_j is the expected return of the jth hedge fund, and μ_B is the expected return of the benchmark. Since the goal is to identify the funds that outperform the benchmark, we are in the one-sided case (11) with $\theta_{0,j} = 0$, for $j = 1, \ldots, s$.

Naturally, the estimator of θ_j is given by

$$\hat{\theta}_{n,j} = \frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \frac{1}{n}\sum_{i=1}^{n} X_{i,B},$$

that is, by the difference of the corresponding sample averages. It is well known that hedge fund returns, unlike mutual fund returns, tend to exhibit non-negligible serial correlations; see, for example, Lo (2002) and Kat (2003). Accordingly, one has to account for this time series nature in order to obtain valid inference. The standard errors for the original data, $\hat{\sigma}_{n,j}$, use a kernel variance estimator based on the prewhitened QS kernel and the corresponding automatic choice of bandwidth of Andrews and Monahan (1992). The bootstrap data are generated using the circular block bootstrap of Politis and Romano (1992), based on $B = 5{,}000$ repetitions. The standard errors in the bootstrap world, $\hat{\sigma}^*_{n,j}$, use the corresponding 'natural' variance estimator; for details, see Götze and Künsch (1996) or Romano and Wolf (2006). The choice of the block sizes for the circular bootstrap is detailed in Romano et al. (2008).

The number of outperforming funds identified by various procedures and for two nominal levels α are presented in Table 3. Both BKY and Boot results in more rejections than BH, with the comparison between BKY and Boot depending on the level. The numbers for STS appear unreasonably high. Apparently, this is due to the fact that the weak dependence (across test statistics) assumption for the application of this method is clearly violated. The median absolute correlation across funds is 0.32; also see Fig. 2.

8.2 Pairwise fitness correlations

We consider Example 6.5 of Westfall and Young (1993), where the pairwise correlations of seven numeric 'fitness' variables, collected from $n = 31$ individuals, are

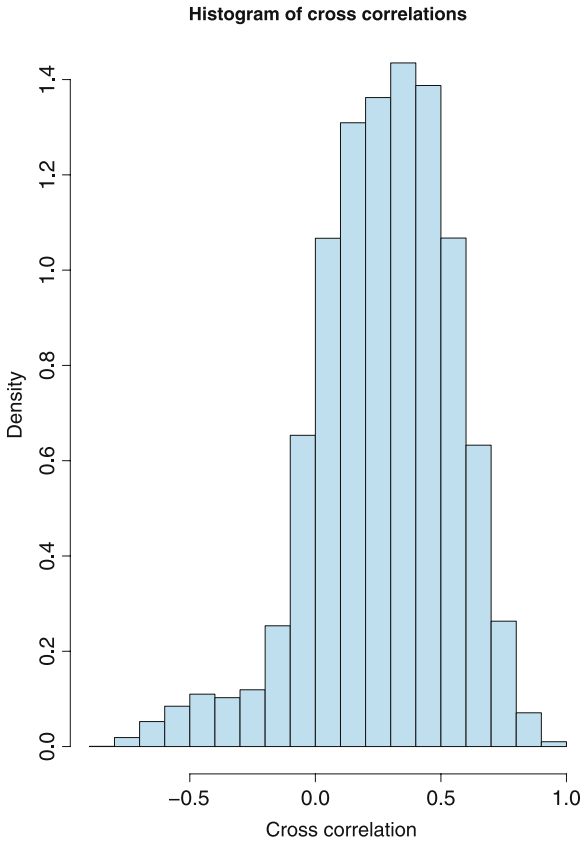

Fig. 2 Histogram of the $208 \cdot 209/2 = 21,736$ cross correlations between the excess returns of the 209 hedge funds. Since it is not true that the majority of these correlations are close to zero, the weak dependence assumption of Storey et al. (2004) is clearly violated

analyzed. Denote the $s = \binom{7}{2} = 21$ pairwise population correlations, ordered in any fashion, by θ_j for $j = 1, \ldots, s$, and let $\hat{\theta}_{n,j}$, $j = 1, \ldots, s$, denote the corresponding Pearson's sample correlations. Since the goal is to identify the nonzero population correlations, we are in the two-sided case (12) with $\theta_{0,j} = 0$ for $j = 1, \ldots, s$.

Westfall and Young (1993) provide two sets of individual p-values: asymptotic p-values based on the assumption of a bivariate normal distribution and bootstrap p-values. As can be seen from their Fig. 6.4, the two are always very close to each other. However, as pointed out by Westfall and Young (1993, p. 194), both sets of p-values are actually for the stronger null hypotheses of independence rather than zero correlation. Obviously, independence and zero correlation are the same thing for multivariate normal data, but we do not wish to make this parametric assumption.

Instead, we use Efron's bootstrap to both compute individual p-values and to carry out our bootstrap FDR procedure. (Of course, the same set of bootstrap resamples is used for both purposes.) The details are as follows. The standard errors for the original data, $\hat{\sigma}_{n,j}$, are obtained using the delta method because, again, we do not want to assume multivariate normality; see Example 11.2.10 of Lehmann and Romano (2005b). This results in test statistics $T_{n,j} = |\hat{\theta}_{n,j}|/\hat{\sigma}_{n,j}$. The bootstrap data are generated using Efron's (1979) bootstrap, based on $B = 5{,}000$ repetitions. The standard

Table 4 Number of nonzero correlations identified

Procedure	$\alpha = 0.05$	$\alpha = 0.1$
BH	2	4
STS	10	20
BYK	2	4
Boot	2	7

errors for the bootstrap data, $\hat{\sigma}^*_{n,j}$, are computed in exactly the same fashion as for the original data. This results in bootstrap statistics $T^*_{n,j} = |\hat{\theta}^*_{n,j} - \hat{\theta}_{n,j}|/\hat{\sigma}^*_{n,j}$. The individual p-values are then derived according to (4.11) of Davison and Hinkley (1997):

$$\hat{p}_{n,j} = \frac{1 + \#\{T^*_{n,j} \geq T_{n,j}\}}{B+1}. \tag{24}$$

The number of nonzero correlations identified by various procedures and for two nominal levels α are presented in Table 4. BKY results in the same number of rejections as BH for both nominal levels. Boot results in the same number of rejections for $\alpha = 0.05$ but yields three additional rejections for $\alpha = 0.1$. The numbers for STS again appear unreasonably high.

An alternative way of testing $H_j : \theta_j = 0$ is to reparametrize θ_j by

$$\vartheta_j = \operatorname{arctanh}(\theta_j) = \frac{1}{2}\log\left(\frac{1+\theta_j}{1-\theta_j}\right).$$

This transformation is known as Fisher's z-transformation, which under normality is variance stabilizing; see Example 11.2.10 of Lehmann and Romano (2005b). Obviously, $\theta_j = 0$ if and only if $\vartheta_j = 0$. The natural estimator of ϑ_j is given by $\hat{\vartheta}_{n,j} = \operatorname{arctanh}(\hat{\theta}_{n,j})$. Using the fact that $\operatorname{arctanh}'(x) = 1/(1-x^2)$, the delta method implies the corresponding standard error $\tilde{\sigma}_{n,j} = \hat{\sigma}_{n,j}/(1-\hat{\theta}^2_{n,j})$. This results in test statistics $T_{n,j} = |\hat{\vartheta}_{n,j}|/\tilde{\sigma}_{n,j}$. Some motivation for bootstrapping the z-transformed sample correlation rather than the 'raw' sample correlation is given in Efron and Tibshirani (1993, Sect. 12.6). Again, the bootstrap data are obtained using Efron's 1979 bootstrap, based on $B = 5,000$ repetitions. The standard errors for the bootstrap data, $\tilde{\sigma}^*_{n,j}$, are computed as $\tilde{\sigma}^*_{n,j} = \hat{\sigma}^*_{n,j}/(1-\hat{\theta}^*_{n,j})^2$. This results in bootstrap statistics $T^*_{n,j} = |\hat{\vartheta}^*_{n,j} - \hat{\vartheta}_{n,j}|/\tilde{\sigma}^*_{n,j}$. The individual p-values are derived as in (24) again.

The number of nonzero correlations identified by various procedures and for two nominal levels α are also presented in Table 4. While making inference for the ϑ_j does not necessarily lead to the same results as making inference for the θ_j, in particular when the sample size n is not large, for this particular data set, none of the numbers of rejections change.

9 Conclusion

In this article, we have developed two methods which provide asymptotic control of the false discovery rate. The first method is based on the bootstrap, and the second is based on subsampling. Asymptotic validity of the bootstrap holds under fairly weak assumptions, but we require an exchangeability assumption for the joint limiting distribution of the test statistics corresponding to true null hypotheses. The method based on subsampling can be justified without such an assumption. However, simulations support the use of the bootstrap method under a wide range of dependence. Even under independence, our bootstrap method is competitive with that of Benjamini et al. (2006) and outperforms it under dependence.

The bootstrap method succeeds in generalizing Troendle (2000) to allow for non-normality. However, it would be useful to also consider an asymptotic framework where the number of hypotheses is large relative to the sample size. Future work will address this.

Acknowledgements We are grateful to Harry Joe for providing R routines to generate random correlation matrices.

References

Abramovich F, Benjamini Y (1996) Adaptive thresholding of wavelet coefficients. Comput Stat Data Anal 22:351–361

Abramovich F, Benjamini Y, Donoho DL, Johnstone IM (2006) Adapting to unknown sparsity by controlling the false discovery rate. Ann Stat 34(2):584–653

Andrews DWK, Monahan JC (1992) An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 60:953–966

Basford KE, Tukey JW (1997) Graphical profiles as an aid to understanding plant breeding experiments. J Stat Plann Inference 57:93–107

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B 57(1):289–300

Benjamini Y, Hochberg Y (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Stat 25(1):60–83

Benjamini Y, Liu W (1999) A stepdown multiple hypotheses testing procedure that controls the false discovery rate under independence. J Stat Plann Inference 82:163–170

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4):1165–1188

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3):491–507

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Drigalenko EI, Elston RC (1997) False discoveries in genome scanning. Genet Epidemiol 15:779–784

Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7:1–26

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Genovese CR, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Stat 32(3):1035–1061

Götze F, Künsch HR (1996) Second order correctness of the blockwise bootstrap for stationary observations. Ann Stat 24:1914–1933

Hommel G, Hoffman T (1988) Controlled uncertainty. In: Bauer P, Hommel G, Sonnemann E (eds) Multiple hypothesis testing. Springer, Heidelberg, pp 154–161

Joe H (2006) Generating random correlation matrices based on partial correlations. J Multivar Anal 97:2177–2189

Kat HM (2003) 10 things investors should know about hedge funds. AIRC working paper 0015, Cass Business School, City University. Available at http://www.cass.city.ac.uk/airc/papers.html
Lahiri SN (1992) Edgeworth correction by 'moving block' bootstrap for stationary and nonstationary data. In: LePage R, Billard L (eds) Exploring the limits of bootstrap. Wiley, New York, pp 183–214
Lahiri SN (2003) Resampling methods for dependent data. Springer, New York
Lehmann EL, Romano JP (2005a) Generalizations of the familywise error rate. Ann Stat 33(3):1138–1154
Lehmann EL, Romano JP (2005b) Testing statistical hypotheses, 3d edn. Springer, New York
Lewandowski D, Kurowicka D, Joe H (2007) Generating random correlation matrices based on vines and extended Onion method. Preprint, Dept. of Mathematics, Delft University of Technology
Lo AW (2002) The statistics of Sharpe ratios. Financ Anal J 58(4):36–52
Mehrotra DV, Heyse JF (2004) Use of the false discovery rate for evaluating clinical safety data. Stat Methods Med Res 13:227–238
Politis DN, Romano JP (1992) A circular block-resampling procedure for stationary data. In: LePage R, Billard L (eds) Exploring the limits of bootstrap. Wiley, New York, pp 263–270
Politis DN, Romano JP (1994) The stationary bootstrap. J Am Stat Assoc 89:1303–1313
Politis DN, Romano JP, Wolf M (1999) Subsampling. Springer, New York
Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19:368–375
Romano JP, Shaikh AM (2006a) On stepdown control of the false discovery proportion. In: Rojo J (ed) Optimality: the second Erich L Lehmann symposium. IMS lecture notes—monograph series, vol 49, pp 33–50
Romano JP, Shaikh AM (2006b) Stepup procedures for control of generalizations of the familywise error rate. Ann Stat 34(4):1850–1873
Romano JP, Wolf M (2006) Improved nonparametric confidence intervals in time series regressions. J Nonparametr Stat 18(2):199–214
Romano JP, Wolf M (2007) Control of generalized error rates in multiple testing. Ann Stat 35(4):1378–1408
Romano JP, Shaikh AM, Wolf M (2008) Formalized data snooping based on generalized error rates. Econom Theory 24(2):404–447
Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 30(1):239–257
Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66(1):187–205
Troendle JF (2000) Stepwise normal theory test procedures controlling the false discovery rate. J Stat Plann Inference 84(1):139–158
Van der Laan MJ, Dudoit S, Pollard KS (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat Appl Genet Mol Biol 3(1):Article 15. Available at http://www.bepress.com/sagmb/vol3/iss1/art15/
Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for P-value adjustment. Wiley, New York
Williams VSL, Jones LV, Tukey JW (1999) Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. J Educ Behav Stat 24(1):42–69
Yekutieli D, Benjamini Y (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plann Inference 82:171–196

Test (2008) 17: 443–445
DOI 10.1007/s11749-008-0127-5

DISCUSSION

Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling

José A. Ferreira · Mark A. van de Wiel

Published online: 30 October 2008

We congratulate Romano, Shaikh, and Wolf for their interesting work. Our only criticism to the presentation of the article, which is otherwise very readable, concerns Remark 1 on p. 8. This is crucial to understanding the method, because it explains that the estimates of the probabilities under the null are determined by the smaller test statistics, so it should have been made explicit at an earlier stage in Sect. 5. Incidentally, the use of 'rth largest' and 'rth smallest' to denote the rth order statistic on pp. 6 and 8 is confusing.

The assumption that n is large and that the θ_j's are uniformly away from zero ensures that few non-null statistics will be mixed with the null ones and hence that the estimates of the probabilities in (10) are approximately correct. Since the models used in the simulation study conform to this assumption, we guess that the bootstrap method is shown here at its best. We wonder how it will perform under a sequence of alternatives which approach the null in a more continuous fashion, a more plausible scenario in real-life applications.

One interesting aspect of the simulation results presented in Tables 1 and 2 is how well the 'standard' Benjamini–Hochberg method (BH) works in all scenarios of dependence: the FDR is kept below the required 10%, while the power is on average 80% of that of the bootstrap method proposed by the authors. This suggests

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-008-0126-6.

J.A. Ferreira (✉) · M.A. van de Wiel
Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands
e-mail: j.ferreira@vumc.nl

M.A. van de Wiel
e-mail: mark.vdwiel@vumc.nl

M.A. van de Wiel
Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands

that adaptive versions (other than Storey's 2002, referred to as STS) of the method based on better estimates of the proportion of true null hypotheses $\gamma_s := s_0/s$, like the BKY method, might improve its performance. On the other hand, it also calls for an explanation.

If H_s and F_s are the empirical distribution functions of the sample of p-values (more generally: test statistics that tend to take *smaller* values away from the null) and of the sample of p-values corresponding to the s_0 true null hypotheses, then

$$\mathrm{FDP}(x_s) := \gamma_s \frac{F_s(x_s-)}{H_s(x_s-)} \le q \quad \text{whenever } x_s := \sup\bigl\{x : F_s(x) \le q\, H_s(x)\bigr\}$$

and $H_s(x_s-) > 0$. Hence the procedure that rejects all hypotheses with p-values strictly below the random threshold x_s keeps the FDP below q and at the same time is *optimal among all procedures involving no estimates of* γ_s, because taking the supremum above implies that x_s cannot be increased without running the risk of exceeding the required bound on the FDP. Moreover, it is *optimal among all procedures based on the same upper bound* $\bar\gamma_s$ *on* γ_s, because if it is known that $\gamma_s \le \bar\gamma_s$, one can take $q = q'/\bar\gamma_s$ and guarantee $\mathrm{FDP}(x_s) \le q'$. But F_s is unobservable, so the optimal procedure cannot be realized; the BH method attempts to approximate it by replacing x_s by $x_s' := \sup\{x : F(x) \le q\, H_s(x)\}$, where F (typically the uniform distribution function) is an approximation to F_s, the rationale being that if $F_s \approx F$, then $x_s \approx x_s'$ and $\mathrm{FDP}(x_s') \approx \mathrm{FDP}(x_s) \le q$. Since $\mathrm{FDP}(x_s)$ is bounded, the last statement is even stronger than $\mathrm{FDR}(x_s') \approx \mathrm{FDR}(x_s) \le q$.

If s is not too small and the dependence between the p-values is weak ('weak' is a misnomer, since the dependence in question can actually be very strong; see Ferreira and Zwinderman 2003), the approximation of F_s by F is typically good, and the BH method works very well. This observation can of course be illustrated and corroborated by simulation experiments (as, for instance, in Kim and van de Wiel 2008), and it provides us with a justification for using the BH method very generally.

In the situations considered by Romano, Shaikh, and Wolf, where s is relatively small and/or the dependence structure can be as strong as one wishes, it is not so clear how well F_s is approximated by F, and a fortiori how well $\mathrm{FDP}(x_s')$ approximates $\mathrm{FDP}(x_s)$. If one wants to be completely general, there need not even be an obvious candidate for F, but in practice it is usually alright to assume—like the authors do—that all the p-values or statistics generated under the null have (approximately) the same distribution function F, in which case $EF_s = F$ is—irrespective of the dependence structure of the data—really the only candidate to replace F_s. Under such conditions, one would hope that the random variable F_s, despite not approaching a constant limit, does not deviate that much from F, which would explain the success of the BH method. Can the authors comment on how close the empirical distributions of the p-values generated under the null typically are to the uniform distribution in the simulation scenarios they consider?

We were surprised by how bad the STS version of the method does when the data are dependent (as expected, it is close to being optimal under independence). If G_s denotes the empirical distribution function of the p-values computed under the

alternative hypotheses, we have $H_s = \gamma_s F_s + (1 - \gamma_s)G_s$ and

$$H_s(x) \le \gamma_s F_s(x) + (1 - \gamma_s), \quad \text{whence } \gamma_s \le \frac{1 - H_s(x)}{1 - F_s(x)},$$

which suggests taking $\bar{\gamma}_s(x) := (1 - H_s(x))/(1 - F(x))$ as an overestimate of γ_s. The larger x, the tighter the bound on the right is, which suggests taking x as large as possible; if F is uniform, $(1 - H_s(x))/(1 - F(x))$ is, for large x, like the left-derivative of H_s at 1. This motivates the procedure of estimating γ_s by $h_s(1-)$, where h_s is a density estimate constructed from the sample of p-values. The authors used a variant (Storey's) of this overestimate with $x = 0.5$, and it appears (everything else being equal in the case of the BH and BKY methods) from the results of the simulation that $\bar{\gamma}_s(0.5)$ is a serious *underestimate* of γ_s. Since $\bar{\gamma}_s(x) \ge \gamma_s(1 - F_s(x))/(1 - F(x))$, this could be explained by F_s being considerably bigger than F around 0.5. Do the authors think that a different choice of x might improve $\bar{\gamma}_s(x)$ and the performance of the STS method? If not, would it be possible to incorporate—perhaps by means of resampling methods—the dependence between variables into an estimate of γ_s? Intuitively, the fact that "all false hypotheses will be rejected with probability tending to one," implied by the main assumptions, suggests that it should be easy to get a good estimate of γ_s that works well in the scenarios considered by the authors.

The authors perform their simulations of Sect. 7.2 for $s = 4$ in order to cover the space of random correlation matrices "more thoroughly." While we understand that a low-dimensional space is easier to 'fill' than a high-dimensional one, we fail to see why this is relevant to the robustness of multiple testing methods based on the control of the FDR (which are especially designed for testing a substantial number of hypotheses) with respect to random correlations. We wonder whether the situation for $s = 4$ can be extrapolated to the more relevant case of $s \ge 50$, given that the number of correlations increases quadratically in s. Do the authors have any results for larger values of s?

In the Conclusion, the authors touch upon the case $s \gg n$, for which their current asymptotic results are less relevant. Of course, applications in this case are extremely relevant nowadays, and we encourage the authors to consider these. On the other hand, asymptotic results in s by Storey (2003) indicate that under weak dependence the FDR is asymptotically equal to the ratio of the marginal expectations, which obviously does not depend on the dependence structure (and in fact, as pointed out above, even stronger dependencies will not affect this result). Such weak, often local, dependencies are thought to be the most relevant ones in high-dimensional applications to microarrays, mass-spectrometry (proteomics), and functional MRI, so the available FDR algorithms may suffice for these.

References

Ferreira JA, Zwinderman AH (2003) Approximate power and sample size calculations with the Benjamini–Hochberg method. Int J Biostat 2(1):8

Kim KI, van de Wiel MA (2008) Effects of dependence in high-dimensional multiple testing problems. BMC Bioinform 9:114

Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol 64:479–498

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Statist 31:2013–2035

Test (2008) 17: 446–449
DOI 10.1007/s11749-008-0128-4

DISCUSSION

Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling

Wenge Guo

Published online: 30 October 2008

1 Introduction

In this enlightening and stimulating paper, Professors Romano, Shaikh, and Wolf construct two novel resampling-based multiple testing methods using the bootstrap and subsampling techniques and theoretically prove that these methods approximately control the FDR under weak regularity conditions. The theoretical results provide a satisfactory solution to an important and challenging problem in multiple testing on developments of FDR controlling procedures by exploiting unknown dependence among the test statistics using resampling techniques.

In my comments, I address the related statistical and computational issues when applying their bootstrap method to analyze high-dimensional, low sample size data such as microarray data and suggest several possible extensions.

2 High-dimensional, low sample size data analysis

The bootstrap method provides asymptotic control of the FDR when the sample size approaches infinity. Its finite sample performance is evaluated through some simulation studies and analysis of two real data. For the simulated data, the number of hypotheses tested is $s = 50$, and the sample size is $n = 100$. For the real data, one is with $s = 209$ and $n = 120$, and another is with $s = 21$ and $n = 31$. For such simulated and real data, the bootstrap method is competitive with existing methods, such

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-008-0126-6.

W. Guo (✉)
Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709-2233, USA
e-mail: wenge.guo@gmail.com

as Benjamini et al. (2006), under independence and outperforms them under dependence. A common feature of the simulation settings and real data is that s is relatively small and n is relatively large. However, in practice, there are a number of applications where the number of null hypotheses of interest is very large relative to the sample size. For example, in microarray experiments, often there are thousands or tens of thousands of genes, but the sample size is just less than a dozen. A natural question is: Can the bootstrap method be used for analyzing such high-dimensional, low sample size data?

It is often likely for microarray data to contain several extreme outliers. When the bootstrap method is applied to such microarray data, the extreme outliers may appear in some bootstrap samples due to small sample size, resulting in a very large bootstrap statistic. To compute the largest critical value c_s, we take the $(1 - s\alpha)$ quantile of the maximal bootstrap statistics. But, if quite a large fraction of those maximal bootstrap statistics is very large, then the largest critical value will also be very large, which leads to a situation where no hypothesis can be rejected by the stepdown method. Therefore, to make the bootstrap method work well, it is perhaps necessary to perform a preprocessing step to remove these outliers or choose some robust statistics such as the median.

It is also likely that the data sets corresponding to many of the genes in a microarray experiment are skewed. In any bootstrap sample, the maximal bootstrap statistic over a large number of hypotheses is then likely to be quite large, thus resulting in a very large bootstrap critical value to which to compare the largest observed statistic. Since the suggested bootstrap method is a stepdown procedure, it is possible that no hypothesis can be finally rejected at all. Therefore, when applying the bootstrap method to microarray data analysis, it might be necessary to do some transformation to alleviate the skewness of the data or choose some more appropriate test statistics.

With the help of Professor Wolf, I directly applied the bootstrap method in the context of a two-sample t test to a real microarray data (Hedenfalk et al. 2001). Perhaps due to the presence of a few extreme outliers and a large number of skewed data, the bootstrap method could not find any significant gene in this data set.

3 Computational problem

When the bootstrap method is applied to analyzing microarray data, it is a challenge to compute all the critical values. For example, when Professor Wolf applied this method, on my request, to a simulated data set with 4,000 variables, it took him more than 70 hours to do the computations. In the following, we present a possible improvement on the computational method of the critical values.

For a given estimate $\hat{P}$ of the unknown joint distribution P of the underlying test statistics, the critical values, $\hat{c}_i, i = 1, \ldots, s$, are defined recursively as follows: having determined $\hat{c}_1, \ldots, \hat{c}_{j-1}$, compute $\hat{c}_j$ according to the rule

$$\hat{c}_j = \inf\bigl\{c \in \mathbb{R} : \mathrm{FDR}_{j,\hat{P}}(c) \leq \alpha\bigr\},$$

where

$$\begin{aligned}\mathrm{FDR}_{j,\hat{P}}(c) &= \sum_{s-j+1\le r\le s} \frac{r-s+j}{r} \\ &\quad \times \hat{P}\{T_{j:j}\ge c,\ldots,T_{s-r+1:j}\ge \hat{c}_{s-r+1}, T_{s-r:j}<\hat{c}_{s-r}\} \\ &= \frac{1}{B}\sum_{b=1}^{B}\sum_{s-j+1\le r\le s} \frac{r-s+j}{r} \\ &\quad \times I\left\{T^{*b}_{j:j}\ge c,\ldots,T^{*b}_{s-r+1:j}\ge \hat{c}_{s-r+1}, T^{*b}_{s-r:j}<\hat{c}_{s-r}\right\}\end{aligned}$$

is the FDR of the bootstrap method when there are exactly j true null hypotheses under P, and the unknown P is estimated using the empirical distribution $\hat{P}$ of the bootstrap test statistics generated by B bootstrap samples. That is, $\hat{c}_j$ is the α-quantile of $\mathrm{FDR}_{j,\hat{P}}(c)$.

Note that in the above expression of $\mathrm{FDR}_{j,\hat{P}}(c)$,

$$\begin{aligned}&I\left\{T^{*b}_{j:j}\ge c,\ldots,T^{*b}_{s-r+1:j}\ge \hat{c}_{s-r+1}, T^{*b}_{s-r:j}<\hat{c}_{s-r}\right\} \\ &\quad = I\left\{T^{*b}_{j:j}\ge c\right\}\cdots I\left\{T^{*b}_{s-r+1:j}\ge \hat{c}_{s-r+1}\right\}\cdot I\left\{T^{*b}_{s-r:j}<\hat{c}_{s-r}\right\}. \qquad (1)\end{aligned}$$

For every $b=1,\ldots,B$, let r^{*b}_j denote the total number of rejections when applying a stepdown procedure with the critical constants $\hat{c}_i, i=1,\ldots,j-1$, to the ordered test statistics $T^{*b}_{i:j}: i=1,\ldots,j-1$. Then, (1) can be simplified as $I\{T^{*b}_{j:j}\ge c, r=s-r^{*b}_j\}$, and hence $\mathrm{FDR}_{j,\hat{P}}(c)$ can be expressed as

$$\mathrm{FDR}_{j,\hat{P}}(c) = \frac{1}{B}\sum_{b=1}^{B}\frac{j-r^{*b}_j}{s-r^{*b}_j} I\left\{T^{*b}_{j:j}\ge c\right\}. \qquad (2)$$

The expression (2) might be able to greatly simplify computation of the critical values.

Another point we need to be careful about is how the computational precisions of former critical values influence that of the latter. When s is large, the maximum critical value is determined by a large number of former critical values. Even though these former critical values are slightly imprecise, their total effect on the maximum critical values might be huge and thereby greatly changes the final decisions on null hypotheses.

4 Some possible extensions

As we pointed out in Sect. 2, the bootstrap method is sensitive to a few extreme outliers or a large number of skewed data. For such data, it may lead to a very large value for the maximum critical value. Since the bootstrap method is a stepdown procedure, we may fail to detect any false null hypothesis using this method. To overcome the

problems caused by the outliers or skewed data, a possible solution might be to develop stepup procedures that are not sensitive to a few large maximum critical values.

As seen in Sect. 3, the computation of all critical values for the bootstrap method is a challenging task. To apply the method, we need to go through two steps. We first need to calculate all the critical values and then apply the corresponding stepdown procedure to the observed test statistics. The reason is that the computation starts from the minimum critical value and continues to the larger ones. In practice, it is common that there are only a few false nulls in a large number of null hypotheses of interest. Thus, one natural question is: Could we derive an algorithm which combines computation of every critical value with the corresponding hypothesis testing? For this algorithm, it starts by calculating the maximum critical value and continues up to the critical value for which the corresponding hypothesis is not rejected. Therefore, it is very likely that the whole test will stop in a few earlier steps, and thus we only need to calculate a few of the larger critical values.

The asymptotic control of the suggested methods is proved when the sample size approaches infinity, not the dimension of the data. However, in practice, the data sets with high dimensions and low sample size are becoming more common due to the developments of high throughput technologies. Therefore, it will be interesting and important to develop similar resampling-based methods which can asymptotically control the FDR in theory when the dimensions of the data approach infinity.

Acknowledgements This research is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences [Z01 ES10174-04]. The author thanks Michael Wolf for helpful discussions and for spending a considerable amount of time in computation. The author also thanks Shyamal Peddada, Sanat Sarkar, and Zongli Xu for carefully reading of this manuscript and for their useful comments that greatly improved the presentation.

References

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3):491–507

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J (2001) Gene-expression profiles in hereditary breast cancer. New Eng J Med 344:539–548

Test (2008) 17: 450–455
DOI 10.1007/s11749-008-0129-3

DISCUSSION

Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling

Sanat K. Sarkar · Ruth Heller

Published online: 30 October 2008

1 Introduction

It is a pleasure to congratulate Professors Romano, Shaikh, and Wolf (to be referred to as RSW hereafter) on an interesting and original paper. RSW address an important and challenging issue in multiple testing, to directly incorporate the dependence structure of the p-values while constructing a multiple testing method that provides a control of the false discovery rate (FDR). The dependence among the p-values has often been utilized in an indirect manner, to the extent of just validating that an FDR controlling method developed under the assumption of independent p-values continues to work even when there is a certain form of dependence among the p-values. A more explicit use of the dependence structure should result in a powerful method. The problem is, however, that one has to know the exact distribution of the underlying test statistics, or has to capture it from the data, at least approximately, by means of methods like those relying on resampling techniques. RSW have decided to take the latter approach by appealing to the bootstrap and subsampling methods.

We do like the main idea in the paper, it will provide an impetus for research on developing bootstrap-based multiple testing methods. Nevertheless, we feel that a number of points need to be made to provide a better understanding of the paper and to fill up certain gaps.

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-008-0126-6.

S.K. Sarkar (✉)
Statistics Department, Temple University, Philadelphia, PA 19122, USA
e-mail: sanat@temple.edu

R. Heller
Statistics Department, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: ruheller@wharton.upenn.edu

2 What is being controlled?

Roughly stated, the paper does the following. Given data $(X_1, \ldots, X_n)$ from a distribution $P \in \Omega$, it considers testing of $H_i : P \in \omega_i$ against $H_i' : P \notin \omega_i$, simultaneously for $i = 1, \ldots, s$, based on test statistics $T_{n,i}$, $i = 1, \ldots, s$, which are such that large values of $T_{n,i}$ indicate evidence against the corresponding H_i and all the false null hypotheses are rejected with probability tending to one as $n \to \infty$, and provides both bootstrap and subsampling based algorithms to calculate the critical values $c_{n,1} \leq \cdots \leq c_{n,s}$ of a stepdown test that will guarantee a control of the FDR at α asymptotically as $n \to \infty$ under certain weak assumptions. Let us denote the FDR of this stepdown procedure by FDR_n to properly index it by n since the critical values depend on n. The paper proves that, given s_0, the number of true nulls,

$$\mathrm{FDR}_n \approx E_P\left[\frac{R_{n,0}}{(s - s_0 + R_{n,0}) \vee 1}\right], \tag{1}$$

with probability tending to one as $n \to \infty$, where $R_{n,0}$ is the number of rejections in the stepdown procedure based on any subset of the test statistics corresponding to the s_0 true nulls and the critical values $c_{n,s-s_0+1} \leq \cdots \leq c_{n,s}$. So, the right-hand side in (1) is what is being controlled in the paper, making the proposed stepdown method an asymptotically valid FDR controlling method. For finite n, this equals the FDR in the special case where the non-null test statistics are all larger than the null test statistics. Moreover, it should be noted that by saying that the method in the paper is an asymptotically valid FDR controlling method, in the above sense, it does not necessarily mean that there exists a sufficiently large $n_0 \equiv n_0(\alpha)$ such that $\mathrm{FDR}_n \leq \alpha$ for all $n \geq n_0$.

3 Other relevant methods

RSW have decided to compare their proposed stepdown procedure with three other procedures, the BH and its adaptive versions, the STS and BKY. These three procedures differ from the proposed one in two aspects: (i) they are all stepup procedures, and (ii) they are all marginal procedures (i.e., they do not exploit the joint distribution of the p-values).

Recently, an adaptive stepdown procedure has been given in Gavrilov et al. (2008). While its FDR control has been theoretically established for independent p-values, like in the cases of the BKY and STS, simulations indicate that it can maintain its control even under certain dependence situations. In terms of the p-values, it is based on the following critical values:

$$\alpha_j = \frac{j\alpha}{s - j(1-\alpha) + 1}, \quad j = 1, \ldots, s. \tag{2}$$

Although it is a special case of a multi-stage version of the BKY and has been referred to as a multiple-stage stepdown method in Benjamini et al. (2006), it is actually an

adaptive stepdown analog of the BH considered in Sarkar (2002). To see this, note that with

$$\widehat{\mathrm{FDR}}_\lambda(t) = \frac{\hat{s}_0(\lambda)t}{R(t) \vee 1},$$

where

$$\hat{s}_0(\lambda) = \frac{s - R(\lambda)}{1 - \lambda} \quad \text{and} \quad R(t) = \#\{\hat{p}_{n,j} \le t\},$$

the BH rejects $H_{(1)}, \ldots, H_{(\hat{l}_{SU})}$, where

$$\hat{l}_{SU} = \max\left\{1 \le j \le s : \widehat{\mathrm{FDR}}_{\lambda=0}(\hat{p}_{n,(j)}) \le \alpha\right\} \tag{3}$$

provides the stepup rejection threshold; whereas, the stepdown analog of the BH method rejects $H_{(1)}, \ldots, H_{(\hat{l}_{SD})}$, where

$$\hat{l}_{SD} = \max\left\{1 \le j \le s : \widehat{\mathrm{FDR}}_{\lambda=0}(\hat{p}_{n,(i)}) \le \alpha \ \forall\, i \le j\right\}$$

provides the stepdown rejection threshold. In STS, the FDR is estimated using

$$\widehat{\mathrm{FDR}}^*_\lambda(t) = \frac{\hat{s}^*_0(\lambda)t}{R(t) \vee 1}, \tag{4}$$

that is based on the following slightly different estimate of s_0:

$$\hat{s}^*_0(\lambda) = \frac{s - R(\lambda) + 1}{1 - \lambda}$$

[(6) of the paper], and the stepup rejection threshold in (3) is modified accordingly as

$$\hat{l}^*_{SU} = \max\left\{1 \le j \le s : \widehat{\mathrm{FDR}}^*_\lambda(\hat{p}_{n,(j)}) \le \alpha\right\}$$

for a fixed $\lambda \ne 0$. If we consider modifying the stepdown analog of the BH method using the alternative estimate of the FDR, which is $\widehat{\mathrm{FDR}}^*_t(t)$ [with $\lambda = t$ in (4)], and determining the stepdown rejection threshold based on this estimate, that is,

$$\hat{l}^*_{SD} = \max\left\{1 \le j \le s : \widehat{\mathrm{FDR}}^*_{\hat{p}_{n,(i)}}(\hat{p}_{n,(i)}) \le \alpha \ \forall\, i \le j\right\},$$

we obtain the adaptive stepdown method with the critical values in (2).

A number of other adaptive procedures like the STS and BKY are given in Sarkar (2008). Among these, the following is worth mentioning. Let $R_{SU}(\lambda_1, \ldots, \lambda_s)$ denote the number of rejections in a stepup procedure (in terms of p-values) with the critical values $\lambda_1 \le \cdots \le \lambda_s$. As noted in Sarkar (2008), the BH procedure with its jth critical value replaced by $\hat{\alpha}_j = j\alpha/\hat{s}_0$, where

$$\hat{s}_0 = \frac{s - R_{SU}(\lambda_1, \ldots, \lambda_s) + 1}{1 - \lambda_s} \tag{5}$$

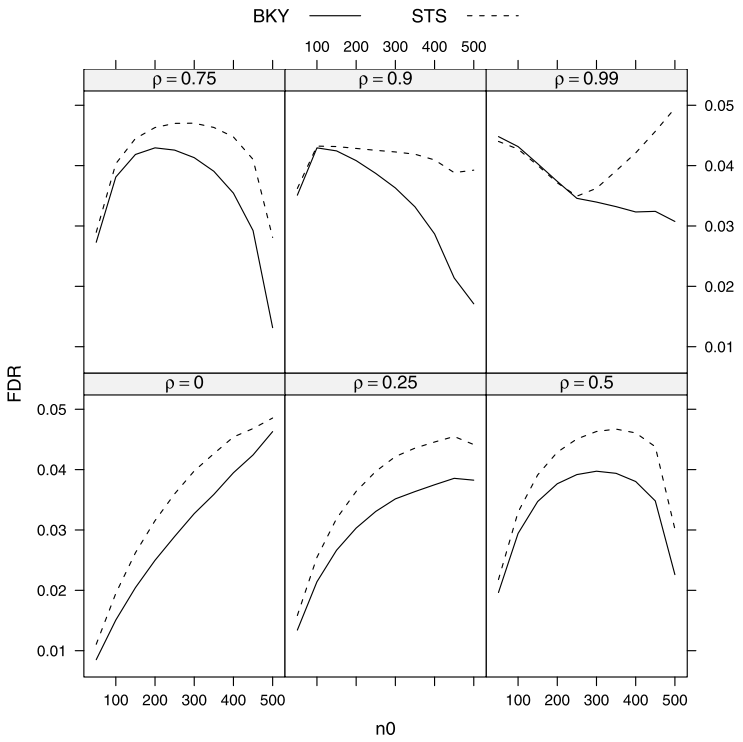

Fig. 1 Comparison of simulated FDR's of the BKY and STS with $\lambda = \alpha/(1+\alpha)$, with $\alpha = 0.05$. Each simulated FDR was based on 20,000 replications, n_0 is the number of true nulls

for any arbitrarily chosen set of critical values $0 < \lambda_1 \leq \cdots \leq \lambda_s < 1$, controls the FDR under the same condition as in the case of the STS or BKY. The STS belongs to this class; it corresponds to the case where $\lambda_j = \lambda$ for any arbitrary λ. Also, the one with $\lambda_j = j\alpha/(1+\alpha)s$ is practically not much different from the BKY.

It should be noted that the rejection threshold chosen to estimate s_0 is much wider in the STS with $\lambda = 0.5$ than in the BKY. This, we suspect, contributes to large variability of the FDR and loss of control over it under dependence of the p-values for the STS with this λ. A smaller λ, we believe, would make the STS more stable in terms of controlling the FDR. A simulation study was conducted to see how the STS compares with the BKY if λ is chosen to be equal to $\alpha/(1+\alpha)$, the same value the BKY chooses as the level of its first stage BH procedure. More specifically, we considered testing whether each of the means of $s = 500$ dependent normal random variables with the same variance 1 and a nonnegative common correlation ρ is 0 or 2 at $\alpha = 0.05$ using both the BKY and STS with $\lambda = .05/1.05$. Figure 1 compares the simulated FDR's of these methods. The STS in this case is seen to have much more favorable performance in terms of the FDR control, even under positive dependence as long as it is not too high.

There exist other procedures that control the FDR by exploiting the joint distribution of the p-values. These procedures include the stepdown procedure of Troendle (2000) under the setting of multivariate normal distribution with a common correlation, as noted in the discussed paper, and the FWER-augmentation procedures towards controlling the FDR suggested in van der Laan et al. (2004) and Pacifico et al. (2004). The FWER-augmentation approach has two stages. At the first stage, an FWER controlling procedure is applied at level α. At the second stage, more discoveries are added to the first stage discoveries while maintaining control of the FDR or of the probability that the FDR is greater than a user-specified value γ. The dependence among the p-values is exploited at the first stage. In Dudoit et al. (2004) the FWER-augmentation procedures of van der Laan et al. (2004) are compared to marginal FDR controlling procedures. Their simulations suggest that there can be substantial power gain in the FWER-augmentation approach due to the incorporation of the joint distribution of the p-values into the procedure.

With a known joint probability distribution P_0 of the test statistics under the null hypotheses, the following stepdown procedure controls the FWER (Pacifico et al. 2004; Dudoit and van der Laan 2008, Chap. 5]):

1. With $t_{n,(1)} \leq \cdots \leq t_{n,(s)}$ being the observed ordered test statistics, let k_j be the hypothesis with the jth smallest test statistic $t_{n,(j)}$.
2. For $r = 1, \ldots, s$, do the following:
   (a) Compute $\hat{p}_{(r)} = P_0\{\max_{j \in V_r} T_{n,j} \geq t_{n,(s-r+1)}\}$.
   (b) If $\hat{p}_{(r)} > \alpha$, stop and reject the $r-1$ hypotheses that correspond to the largest test statistics; if $\hat{p}_{(r)} \leq \alpha$, increase r by 1 and go to Step 2(a).

This stepdown procedure is augmented as follows in Pacifico et al. (2004) to control the FDR at level q (see Dudoit and van der Laan 2008, Chap. 6, for similar procedures):

1. Let $c \in (0, q)$, and let $\alpha = (q - c)/(1 - c)$.
2. Apply the above stepdown procedure at level α. Let R_1 be the number of rejected hypotheses.
3. Let $R_2 = \inf\{r : \frac{r}{r+R_1} \leq q\}$. Reject the R_2 hypotheses corresponding to the largest R_2 test statistics.

Note that the above stepdown FWER controlling procedure is identical to the stepdown FDR procedure of RSW when $r = 1$ but becomes more conservative starting from $r = 2$, and thus will typically reject less hypotheses. In other words, the method suggested by RSW appears to be more powerful than the FWER-augmentation approach. However, the FWER-augmentation approach may control the FDR with finite samples as well as asymptotically (as long as the FWER controlling procedure controls the FWER in the finite sample case).

4 Final points

In sparse settings, where s_0/s is close to 1, the BH procedure is very powerful. It may be interesting to compare the power of the suggested procedure with the BH

procedure in such settings. Example 8.1 in the discussed paper shows that estimating the number of true null hypotheses s_0 and then using this estimate in a marginal procedure (like the BKY) that does not take the joint distribution of the p-values into account may be more powerful than a procedure that takes the joint distribution of the p-values into account without estimating s_0. Maybe, the present method can be improved by incorporating an estimate of s_0?

Acknowledgements The work of Sanat K. Sarkar is supported by the NSF Grant DMS-0603868. We thank Zijiang Yang for doing the numerical calculations.

References

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3):491–507

Gavrilov Y, Benjamini Y, Sarkar SK (2008) An adaptive step-down procedure with proven FDR control. Ann Stat (in press)

Dudoit S, van der Laan MJ (2008) Multiple testing procedures with applications to genomics. Springer series in statistics. Springer, New York

Dudoit S, van der Laan MJ, Birkner MD (2004) Multiple testing procedures for controlling tail probability error rates. Tech report, UC Berkeley division of biostatistics working paper series. Working paper 166. Available in http://www.bepress.com/ucbbiostat/paper166

Pacifico M, Genovese C, Verdinelli I, Wasserman L (2004) False discovery control for random fields. J Am Stat Assoc 99:1002–1014

Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 30(1):239–257

Sarkar SK (2008) On methods controlling the false discovery rate. Unpublished manuscript

Troendle JF (2000) Stepwise normal theory test procedures controlling the false discovery rate. J Stat Plann Inference 84(1):139–158

van der Laan MJ, Dudoit S, Pollard KS (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat Appl Genet Mol Biol 3:15

Test (2008) 17: 456–457
DOI 10.1007/s11749-008-0130-x

DISCUSSION

Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling

James F. Troendle

Published online: 30 October 2008

I would like to commend the authors on a really nice piece of work. It is well written and gives a very general solution to the problem of bootstrap adjustment for multiplicity while controlling the false discovery rate (FDR). At the time that I was working on the normal-theory FDR controlling procedure (Troendle 2000), I had ideas about resampling-based FDR control. However, I have reservations about using FDR-controlling procedures in applications, which led me to discontinue my research on them. The false discovery proportion (FDP) seems like the most natural thing to control when control of the familywise error rate is not needed. In applications there is only one FDP generated, and the bottom line question is "what can you claim about the likelihood of a large FDP with this set of rejected hypotheses?" Even with exact (as opposed to asymptotic) FDR control, the answer is "not much." That is because the FDR is an expected value and says nothing about the tail behavior of the FDP. A simple realistic example give in Korn et al. (2004) showed that a procedure controlling the FDR at 0.1 has an actual FDP ≥ 0.29 with probability 0.1.

One exciting possibility to take from this paper is that the subsampling ideas given in Sect. 6 might be extended to control of the FDP. The fact that the subsampling procedure did not behave well in the simulations for fairly small sample sizes is discouraging, but perhaps that can be overcome. It may take a lot of computation to get satisfactory results because the sample size should be large (for approximately asymptotic behavior to be expected), while the subsample size should also be large yet small relative to the sample size. There are a tremendous number of such subsets

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-008-0126-6.

J.F. Troendle (✉)
Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892, USA
e-mail: jt3t@nih.gov

for even moderate sample size, although one would naturally use Monte Carlo here to select only a few, as one does with the bootstrap.

References

Korn EL, Troendle JF, McShane LM, Simon R (2004) Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plann Inference 124:379–398

Troendle JF (2000) Stepwise normal theory test procedures controlling the false discovery rate. J Stat Plann Inference 84(1):139–158

Test (2008) 17: 458–460
DOI 10.1007/s11749-008-0131-9

DISCUSSION

Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling

Daniel Yekutieli

Published online: 30 October 2008

The paper introduces FDR controlling methods that incorporate information about the dependence structure of the test statistics. I congratulate the authors on their fine work—their methods are shown to offer more power than the Benjamini et al. (2006) FDR controlling procedure and still control the FDR for dependent test statistics. However I am bothered by the lack of a theoretical proof for finite sample FDR control. I will comment on this and on two other related points: the Benjamini and Hochberg (1995) procedure does not offer general FDR control yet it controlled the FDR in all the simulations conducted by the authors; in the simulations displayed in Fig. 1 the FDR of the Boot method was very close to $\alpha = 0.1$ for the $\theta = 20$ configuration and much closer to $\alpha \cdot s_0/s$ for $\theta = 0.2$.

Liberalism of the BH procedure The simulations conducted in this paper included studentized multivariate normal test statistics. Working experience and theoretical results (Reiner 2007) suggest that the FDR of BH procedure for this type of test statistics may slightly exceed $\alpha \cdot s_0/s$ but not exceed α. This explains why the BH procedure controlled the FDR in the simulations. An interesting question is how to construct a simulation in which the FDR of the BH procedure exceeds α while the testing methods introduced in this paper control the FDR.

For example, Guo and Rao (2008) construct a joint p-value distribution in which the FDR of the BH procedure reaches its upper bound $(1 + 1/2 + \cdots + 1/s) \cdot \alpha \cdot s_0/s$. To achieve this FDR level the p-values in a random subset of j components are set precisely in the interval $[\alpha \cdot (j-1)/s, \alpha \cdot j/s)$. It is trivial to transform this

This comment refers to the invited paper available at http://dx.doi.org/10.1007/s11749-008-0126-6.

D. Yekutieli (✉)
Department of Statistics and OR, Tel Aviv University, Tel Aviv, Israel
e-mail: yekutiel@post.tau.ac.il

p-value distribution into a multivariate test statistic distribution. However, for the methods described in this paper, each test statistic has to be computed using data consisting of iid samples $X = (X_1, \ldots, X_n)$, and it seems to me very difficult to construct a distribution for the data such that "reasonable" test statistics applied to X will preserve this intricate dependence structure. Furthermore, the joint distribution of "reasonable" test statistics applied to iid samples is asymptotically multivariate normal, thus the FDR of the BH procedure would approach $\alpha \cdot s_0/s$ for sufficiently large n, in any data distribution.

Conservatism of FDR controlling procedures when the non-null tested effects are small In the extreme case that the p-value are marginally $U[0, 1]$, yet $s - s_0$ hypotheses are labeled false null hypotheses, the only effect of increase in s_0 is the occurrence of more false rejections, thus increasing the FDR (and FWER) of any testing procedure; and if the testing procedure is exchangeable, then it is easy to see that the FDR for any value of s_0 is

$$\mathrm{FDR} = \mathrm{FDR}_0 \cdot s_0/s,$$

where FDR_0 is the FDR under the complete null hypothesis, $s_0 = s$. This implies that multiple testing procedures that control the FDR at level α, for all test statistic distributions, will have $\mathrm{FDR} \leq \alpha \cdot s_0/s$ when the non-null tested effects are sufficiently small.

Finite sample FDR control of the new methods Gavrilov et al. (2008) and Benjamini et al. (2006) show that the FDR values of their multiple testing procedures are maximized when the p-values corresponding to false null hypotheses are set to 0; they prove that their testing procedure controls the FDR under this configuration and use this property to prove the validity of their testing approach. Similarly, the methods introduced in this paper are constructed under the assumption that all false null hypotheses are rejected. The authors prove asymptotic FDR control by showing that, as the sample size increases, this occurs with probability tending to one, yet resort to simulations for finite sample FDR control.

Benjamini and Yekutieli (2001) show that the BH procedure is unique in that its FDR level, for independently distributed p-values, is unaffected by the distribution of the p-values corresponding to false null hypotheses: they prove that in step-up multiple testing procedures with a series of constants $\alpha_1 \cdots \alpha_s$ such that α_j/j is increasing in j, when the distribution of false null p-values stochastically decreases, the FDR increases; while in step-up procedures that α_j/j is decreasing in j, the FDR decreases. I think that a similar result for step-down procedures and dependent test statistics is by itself interesting and may also help proving finite sample FDR control of the new methods.

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57(1):289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4):1165–1188

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3):491–507

Gavrilov Y, Benjamini Y, Sarkar SK (2008) An adaptive step-down procedure with proven FDR control under independence. Ann Stat (to appear). http://www.imstat.org/aos/future_papers.html

Guo W, Rao MB (2008) On control of the false discovery rate under no assumption of dependency. J Stat Plann Inference 138:3176–3188

Reiner A (2007) FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. Biom J 49(1):107–126

Test (2008) 17: 461–471
DOI 10.1007/s11749-008-0134-6

DISCUSSION

Rejoinder on: Control of the false discovery rate under dependence using the bootstrap and subsampling

Joseph P. Romano · Azeem M. Shaikh · Michael Wolf

Published online: 9 December 2008

We are extremely appreciative of the insightful comments made by all the responders. The goal of constructing useful multiple testing methods which control the false discovery rate and other measures of error is currently a thriving and important area of research. On the one hand, the bootstrap method presented in the present work seems to work quite well and is supported by some theoretical analysis. On the other hand, many more important practical, computational, and mathematical questions remain, some of which are addressed by the responders and which we touch upon below.

We also appreciate the added references, which help to provide a more thorough discussion of the available methods. Our paper was the development of a particular methodology and was by no means a comprehensive account of the burgeoning FDR literature.

This rejoinder is discussed in the comments available at:
http://dx.doi.org/10.1007/s11749-008-0127-5, http://dx.doi.org/10.1007/s11749-008-0128-4,
http://dx.doi.org/10.1007/s11749-008-0129-3, http://dx.doi.org/10.1007/s11749-008-0130-x,
http://dx.doi.org/10.1007/s11749-008-0131-9.

J.P. Romano
Departments of Economics and Statistics, Stanford University, Stanford, USA
e-mail: romano@stanford.edu

A.M. Shaikh
Department of Economics, University of Chicago, Chicago, USA
e-mail: amshaikh@uchicago.edu

M. Wolf (✉)
Institute for Empirical Research in Economics, University of Zurich, Bluemlisalpstrasse 10,
8006 Zurich, Switzerland
e-mail: mwolf@iew.uzh.ch

1 Reply to José Ferreira and Mark A. van de Wiel

The non-null values $\theta_j = 0.2$ were chosen as an intermediate case between two non-interesting extremes: (i) if θ_j is very large, the corresponding H_j will be rejected with probability (almost) equal to one for all methods, and so there is little distinction in terms of power; (ii) if θ_j is very close to zero, H_j will be rejected with very small probability for all methods, and so, again, there is little distinction in terms of power. By trial and error, the value $\theta_j = 0.2$ was found to be an interesting middle ground. On the other hand, we understand the concern about the performance of our method for a sequence of alternatives which approach the null in a continuous fashion. To shed some light on this issue, we repeated the simulations, restricting attention to the scenario of common correlation, for the values $\theta_j = 0.1$ and 0.01. The results can be found in Tables 1 and 2. The average number of rejections naturally declines with θ_j, but qualitatively the results do not really change very much.

Concerning the empirical distribution of the p-values generated under the null: these p-values were computed using the t_{n-1} distribution for the studentized test statistics. Since under the null, $\theta_j = 0$, as opposed to $\theta_j < 0$, in our simulation set-up, the null test statistics have exactly this t_{n-1} distribution, and so the null p-values have exactly a uniform [0, 1] distribution. We therefore did not feel the need to give some information about the empirical distribution of the null p-values.

We were also quite surprised by how badly the STS version of the BH method does when the data are dependent. The choice of $\lambda = 0.5$ (or what the discussants call $x = 0.5$) may well be partly responsible. However, we would like to point out that we simply used the "default" value of Storey et al. (2004) rather than deliberately choosing a value of λ which makes the STS version look bad. The question of whether a different choice of λ might lead to a better performance is a very good one. This issue is also addressed by S. Sarkar and R. Heller who argue that the choice $\lambda = \alpha/(1+\alpha)$ results in more reliable FDR control under dependence. We redid Table 1 of the paper, replacing STS by STS*, where the latter uses $\lambda = 0.1/1.1$; see

Table 1 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario, and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS	BKY	Boot	BH	STS	BKY	Boot	BH	STS	BKY	Boot
Ten $\theta_j = 0.1$												
Control	8.1	9.7	7.4	7.5	5.6	15.8	5.7	7.7	4.8	27.3	5.3	9.7
Rejected	0.4	0.5	0.4	0.4	0.8	2.1	0.8	1.0	0.9	3.4	0.9	2.2
Twenty five $\theta_j = 0.1$												
Control	5.1	7.6	4.8	4.3	4.7	7.9	4.6	4.4	4.3	11.1	4.8	6.1
Rejected	1.6	2.8	1.5	1.5	1.7	3.5	1.7	1.7	2.6	6.3	2.8	3.5
All $\theta_j = 0.1$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	5.5	23.8	5.6	5.4	6.0	24.2	6.5	6.4	8.0	27.5	9.8	11.9

Table 2 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario, and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS	BKY	Boot	BH	STS	BKY	Boot	BH	STS	BKY	Boot
Ten $\theta_j = 0.01$												
Control	8.1	8.3	7.3	8.1	5.3	13.4	4.9	7.8	4.0	26.6	3.6	7.8
Rejected	0.03	0.03	0.03	0.03	0.19	1.26	0.23	0.30	0.47	3.3	0.45	0.76
Twenty five $\theta_j = 0.01$												
Control	4.7	4.9	4.3	4.7	4.8	5.4	4.4	5.0	3.5	6.8	3.3	5.1
Rejected	0.08	0.09	0.08	0.08	0.10	0.14	0.09	0.10	0.37	1.88	0.39	0.54
All $\theta_j = 0.01$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	0.17	0.20	0.16	0.17	0.20	0.33	0.19	0.20	0.63	3.84	0.69	0.95

Table 3 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario, and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS*	BKY	Boot	BH	STS*	BKY	Boot	BH	STS*	BKY	Boot
All $\theta_j = 0$												
Control	10.0	10.0	9.1	10.0	6.4	8.3	6.0	9.9	4.8	8.5	4.4	9.8
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_j = 0.2$												
Control	7.6	9.3	7.3	7.3	6.4	9.3	7.5	9.3	5.0	8.1	5.8	10.0
Rejected	3.4	3.7	3.4	3.4	3.5	3.7	3.5	4.1	3.7	3.8	3.7	6.0
Twenty five $\theta_j = 0.2$												
Control	5.0	7.8	6.2	6.7	4.3	8.6	7.4	8.9	3.9	8.0	7.1	9.5
Rejected	13.2	16.2	14.5	14.9	12.3	14.3	13.1	14.0	12.6	15.1	12.7	16.6
All $\theta_j = 0.2$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	34.8	46.2	44.9	48.2	31.9	39.9	36.4	39.1	32.1	37.9	32.1	36.4

Table 3. Similar to the simulations carried out by Sarkar and Heller, STS* successfully controls the FDR in all scenarios considered and dominates both BH and BKY in terms of power. Compared to Boot, it is a bit more powerful for $\rho = 0$. Under positive dependence, there is no clear ranking. Depending on the value of $\rho > 0$ and the number of false hypotheses, either method can be more powerful than the other. Of course, SKS* is computationally much less expensive than Boot, which is an im-

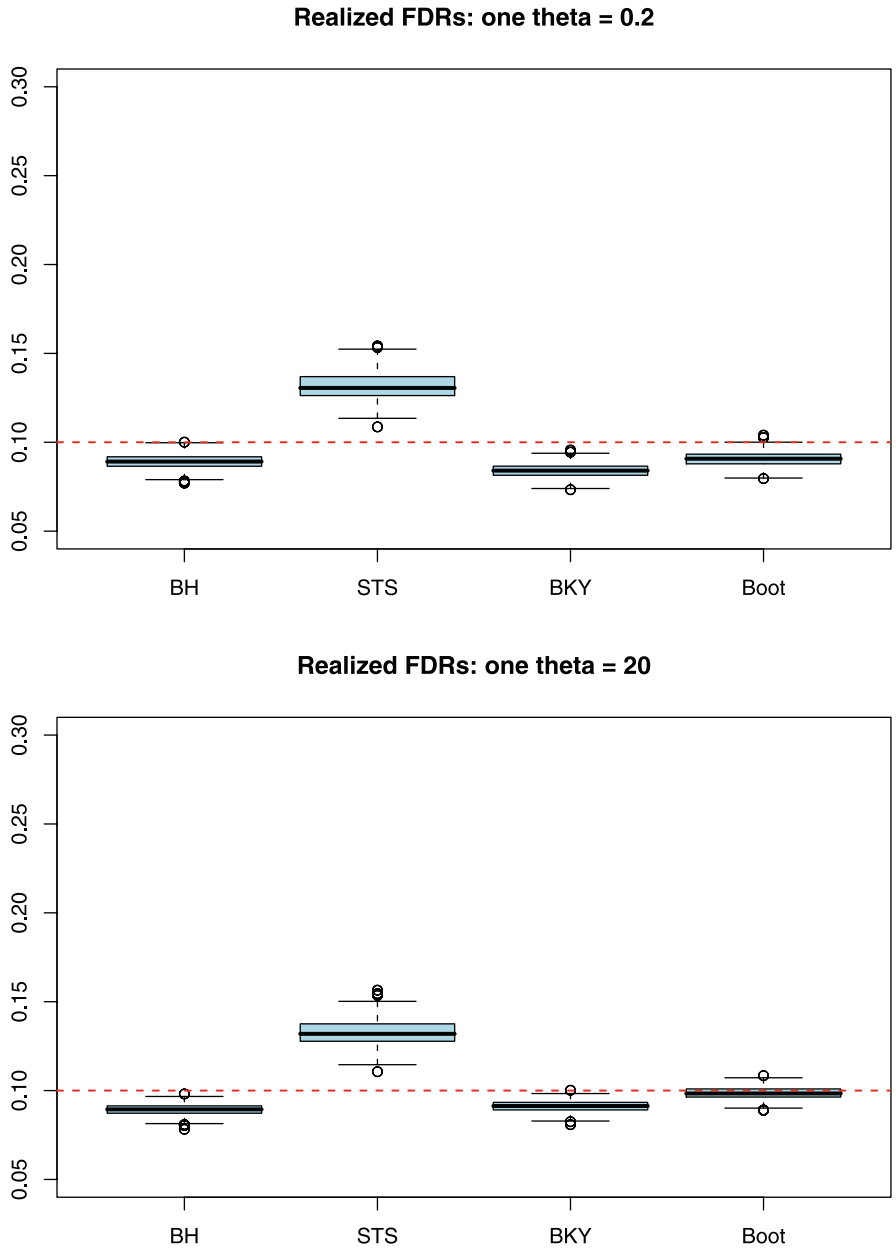

Fig. 1 Boxplots of the simulated FDRs similar to those described in Sect. 7.2, except that we use $s = 10$ instead of $s = 4$ hypotheses now. The horizontal dashed lines indicate the nominal level $\alpha = 0.1$

portant practical advantage, especially when s is very large. There may well be other methods to come up with estimates of s_0 that take the dependence structure into account, say via resampling, but this is beyond the scope of this reply.

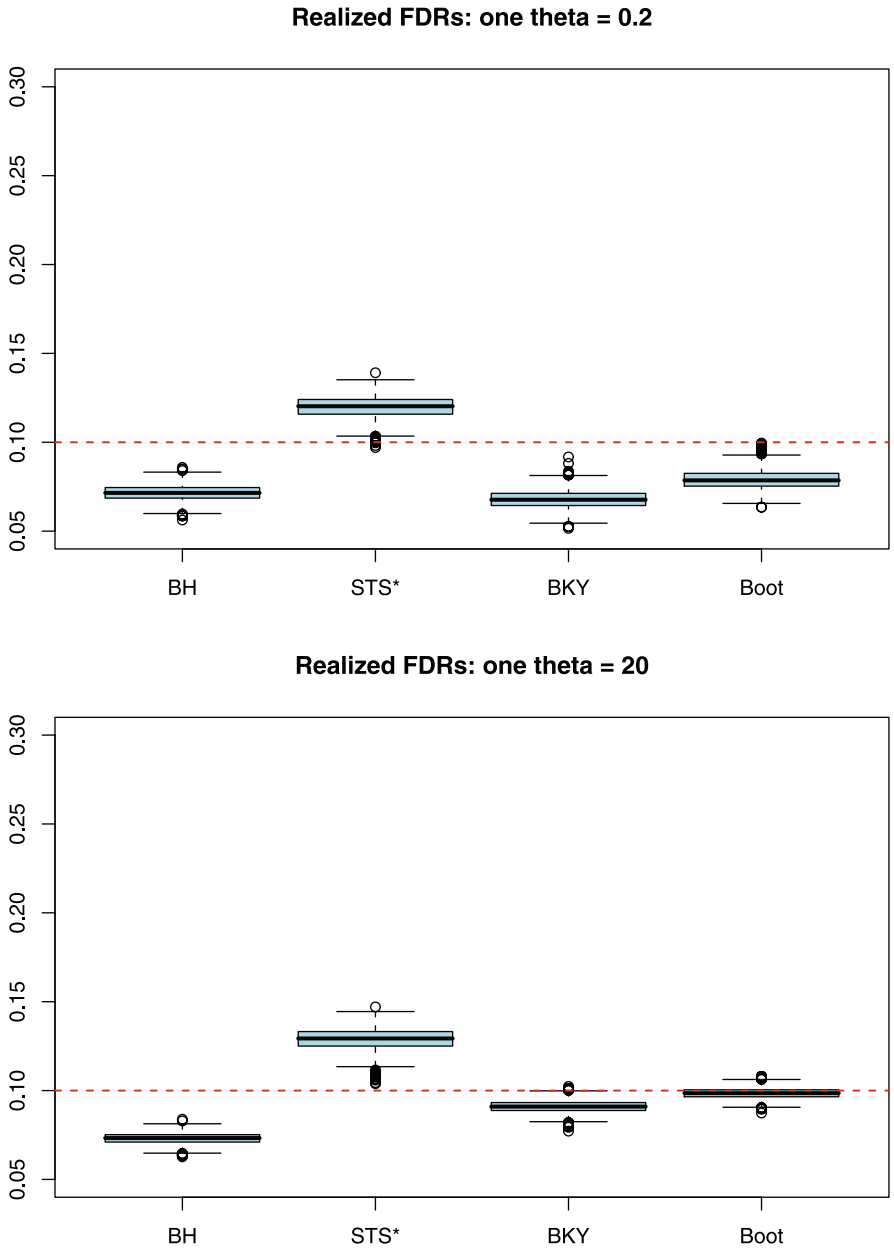

Fig. 2 Boxplots of the simulated FDRs similar to those described in Sect. 7.2, except that STS is replaced by STS* now. The horizontal dashed lines indicate the nominal level $\alpha = 0.1$

Concerning the simulations in Sect. 7.2, there were actually two reasons for the choice $s = 4$. On the one hand, we wanted to cover the space of random correlation matrices "more thoroughly." On the other hand, something like $s = 50$ is computationally infeasible. Unfortunately, the computational burden of our method is a draw-

back. While simulating a single scenario with $s = 50$ is no problem, doing it 1,000 times over (for 1,000 different correlation matrices) would take weeks. However, we were able to at least redo the exercise for the larger value $s = 10$; see Fig. 1. In terms of the FDR control of our bootstrap method, the results do not change qualitatively compared to $s = 4$. So there is reason to hope that they would continue to hold for $s = 50$, say. However, note that there is generally a reduced variation in the boxplots, especially for STS. This indicates that indeed, we cover the space of random correlation matrices "less thoroughly" for $s = 10$. For example, while all the realizations for STS lie above 0.1, we know that for some correlation matrices, the FDR is actually successfully controlled (e.g., for the identity matrix). On the other hand, while all the realizations lie below 0.2, we know that for some correlation matrices, the FDR is actually higher than that (e.g., for the constant correlation matrix with correlation close to one). So in some sense the plot for $s = 4$ is indeed more informative.

In view of the above discussion, we repeated the exercise, keeping $s = 4$ but replacing STS by STS*; see Fig. 2. It is seen that STS* is more conservative than STS but still fails to generally control the FDR. Therefore, if one wishes to use a method based on the marginal p-values and is ignorant about the underlying dependence structure, it might be safer to use BKY rather than STS*.

Finally, we agree with you that Remark 1 could be clearer, and we wish we had the possibility of reviews before the final version. In any case, for the benefit of new readers, the values of both $T_{n,r:t}$ and $T^*_{n,r:t}$ are always meant to be ordered so that they are nondecreasing as r increases.

2 Reply to Wenge Guo

2.1 High-dimensional, low sample size data analysis

We agree that for many applications, these days the number of hypotheses, s, is very large, while the number of data points, n, is very small (at least in comparison). Our bootstrap procedure was not designed for such applications. At this point, the justification of our methods is based on the assumption that $n \to \infty$ while s remains fixed. Also, mild assumptions are imposed on the data generating mechanism P, from which it follows that all false null hypotheses will be rejected with probability tending to one. Arguably, such assumptions are problematic when $s = 2{,}000$ and $n = 10$, for example, which might be considered a "typical" combination for microarray data.

Contamination with outliers, which is quite common for microarray data, is a severe problem for our procedure, at least when non-robust test statistics are used, such as the usual t-statistic. However, the problem lies more with these outliers *not* appearing in bootstrap resamples. Take the case of a single small sample that is "well behaved," apart from a solitary, very large outlier. The t-statistic, for testing the null hypothesis that the population mean is zero, will be close to one in absolute value (as the outlier gets larger and larger in absolute value). Whenever the outlier does not appear in the bootstrap sample, the bootstrap t-statistic—centered by the sample mean of the original data rather than zero—will be large in absolute value, and this happens with probability $(1 - \frac{1}{n})^n \approx 1/e \approx 0.38$. So the bootstrap test, applied to this single

sample, will not reject the null at any customary significance level, just like the usual t-test. Now consider a multiple testing set-up. Our bootstrap method is a stepdown procedure and the "first" critical value (that is, the critical value used to compare the largest of the test statistics) is the $1-\alpha$ quantile of the sampling distribution of the largest bootstrap t-statistic $T^*_{n,(s)}$. Even a single data set with a very large outlier, out of all s individual data sets, can dominate the sampling distribution of this maximum, leading to large critical value. As a result, not even a single hypothesis might get rejected. It is plausible that stepup methods are more robust in this sense. Unfortunately, no bootstrap stepup methods have been suggested in the literature at all so far, not even for the more traditional FWER. This appears, therefore, an important topic for future research.

On the other hand, the fact that stepup procedures based on individual p-values are more robust, in their ability to make rejections at all, to very large outliers in individual samples, does not necessarily mean that they will lead to reliable inference, at least when based on non-robust individual test statistics such as the usual t-statistic. It might be worthwhile to explore suitable robust test statistics as an alternative.

2.2 Computational problem

We agree that the main drawback of the bootstrap method is its computational burden. We are grateful for the suggestions to improve matters. However, consider expression (2). As pointed out, for the bth bootstrap data set, one has basically to compute the number of rejections determined by the critical constants $\hat{c}_i$, $i = 1, \ldots, j-1$, and the ordered test statistics $T^{*b}_{i:j}$, $i = j-1$. For a given value of c, this number, denoted by r^{*b}_j, together with $T^{*b}_{j:j}$, determines the contribution of the bth bootstrap sample to the expression $\mathrm{FDR}_{j,\hat{P}}(c)$. Actually, our software implementation is really comparable in computational complexity to this suggestion. So, unfortunately, things could not be sped up significantly along these lines.

The number of bootstrap repetitions, B, is not all that crucial in successfully controlling the FDR. Note that in our simulations we only used $B = 200$. On the other hand, consider two researchers applying the method to the same data set, both using the same value of B but a different random number generator (or a different seed value). It may well happen that, due to the randomness of the critical values which are computed sequentially, the two researchers might obtain quite different results in terms of the rejected hypotheses. It is therefore indeed desirable to pick B as large as possible, given the computational resources.

2.3 Some possible extensions

We agree that bootstrap stepup methods should be less sensitive to a few extreme outliers or a large number of skewed data sets, as typical with microarray data. However, to the best of our knowledge, no such methods have been developed yet in the multiple testing literature, even for the presumably simpler problem of controlling the FWER (at least not in the nonparametric setting under weak conditions). This remains an exciting field for future research.

As pointed out, the computation of the critical values progresses from the "bottom up" rather than "top down." The latter would save much time in case the number of false hypotheses is relatively small. Unfortunately, we have not yet been able to come up with a "top down" method.

At this point, if the number of hypotheses is very large compared to the sample size, we would not be comfortable with applying the bootstrap method. In such applications, it is probably safer to use methods based on the marginal p-values. But as much effort as possible should be made to ensure that the distribution of the null p-values is as close as possible to the uniform [0,1] distribution in finite samples. Using the usual t-test to compute individual p-values in the presence of extreme outliers or skewed data, combined with small sample sizes, does not appear prudent, yet it seems quite common in practice.

It would be very desirable to develop bootstrap methods that provide error rate control (whether FWER, FDP, or FDR) under more general asymptotics where the number of hypotheses is allowed to tend to infinity together with the sample size. This appears a very challenging task, but we hope to make some progress here in future research.

3 Reply to James F. Troendle

We fully agree that for many, if not most, applications, it would be preferable to control the FDP rather than the FDR. As pointed out, by controlling an expected value, one cannot really say anything of much use about the realized FDP for a given data set. (Of course, one can apply Markov's inequality to get some crude information; see (34) of Lehmann and Romano (2005a). In this sense, it is indeed unfortunate to see that many researches use FDR controlling methods and then interpret their results as if they had actually controlled the FDP instead.

However, control of the FDR is widespread, while control of the FDP is still used comparatively rarely. We hope that this will change over time. In the meantime, and also for those applications where control of the FDR might actually be preferred, we tried to develop a resampling method to account for the unknown dependence structure in order to improve power or the ability to detect false null hypotheses.

Notably, in our own research, we have worked on resampling methods for FDP control first; see Romano and Wolf (2007) and Romano et al. (2008). In the latter paper, inspired by the example in Korn et al. (2004), we also addressed the tail behavior of the realized FDP under FDR control. It was seen that, especially under strong dependence, high values of the FDP can become very likely, even though the FDR is perfectly controlled.

We also agree that there is potential for the subsampling method when the sample size is much larger than one considered in our simulation study, that is, $n = 100$. It is interesting that, even in testing problems involving mean-like parameters and statistics, the asymptotic behavior of the bootstrap and subsampling method are quite distinct in the behavior of critical values. Usually, their first-order asymptotic behavior is the same, but not in the setting of the present paper. It is also frustrating that we could not justify the bootstrap without the exchangeability assumption, even though

this assumption is not needed for subsampling. Future research will be dedicated to these issues.

4 Reply to Sanat K. Sarkar and Ruth Heller

In the setting of our paper, weak assumptions are imposed on the mechanism generating the data, denoted by P, with the number of data points n asymptotically tending to ∞ while the number of tests s remains fixed. It is a consequence of these assumptions (rather than a basic assumption) that all false null hypotheses are rejected with probability tending to one. As Sarkar and Heller point out, the false discovery rate, which is indeed both a function of n and P, now denoted $\mathrm{FDR}_{n,P}$, behaves asymptotically like their expression (1).

In order to interpret our asymptotic results, let us be clear. As pointed out, our results do not imply that there exists a sufficiently large $n_0 = n_0(\alpha)$ such that $\mathrm{FDR}_{n,P} \leq \alpha$ for all $n \geq n_0$. The actual statement is that, for any $\epsilon > 0$, there exists a sufficiently large $n_0 = n_0(\alpha, P)$ such that $\mathrm{FDR}_{n,P} < \alpha + \epsilon$ for all $n \geq n_0(\alpha, P)$. Notice that $n_0(\alpha, P)$ depends on the unknown P; that is, our asymptotic analysis is pointwise in P. Uniform asymptotic convergence over a broad class $\mathbf{P}$ of P would demand that n_0 not depend on $P \in \mathbf{P}$. The distinction between pointwise and uniform convergence in the case of single testing is discussed in Sect. 11.1 of Lehmann and Romano (2005b). Since P is unknown, the stronger uniform convergence results are generally more desirable, though they require additional arguments and sometimes do not hold (for example, as a consequence of the Bahadur–Savage result). Although we did not prove the stronger uniform convergence result in this paper, for the special case where the test statistics are studentized sample means like those considered in the simulations, we expect our results to hold uniformly over a broad class $\mathbf{P}$. In the single testing case, one restriction is that the underlying family of distributions have a uniformly bounded $2 + \delta$ moment, and a weaker condition is given in (11.77) in Theorem 11.4.4 of Lehmann and Romano (2005b). A multivariate extension of that theorem that is relevant for the multiple testing situation studied here is given in Lemma 3.1 of Romano and Shaikh (2008).

A certain limitation of our theoretical analysis is the assumption that n gets large while s remains fixed. We should mention that some literature has considered the large s situation; see, for example, Genovese and Wasserman (2004), Storey et al. (2004), and Efron (2008). However, note that, in some ways, the problem of large s is made easier by stronger assumptions and by the ability to average out errors over many tests. For instance, with the commonly used mixture model, the tests cannot be that different from one another in that their average behavior must settle down, so that, for example, the density of the distribution of test statistics corresponding to false null hypotheses is the same for all such test statistics and can therefore be estimated by usual techniques. Our goal here was to see what can be accomplished in a more general setting which allows for a great deal of heterogeneity (in the sense that the limiting covariance matrix of the test statistics is quite general), but with s fixed.

Sarkar and Heller present an interesting derivation of the stepdown procedure of Gavrilov et al. (2008) as an adaptive stepdown analog of the Benjamini–Hochberg

procedure. The procedure is adaptive in that it modifies the BH procedure by incorporating an estimate of the number of true null hypotheses s_0. Interestingly, the resulting stepdown critical constants, given by (2) in the discussion of Sarkar and Heller, are nonrandom, even though the motivation was based on incorporating a data-dependent estimate of s_0.

We appreciate the discussion of the choice of $\lambda = 0.5$. We also redid some of our simulations, using your suggestion of $\alpha/(1+\alpha)$; see our above rejoinder to Ferreira and van de Wiel.

Sarkar and Heller summarize the use of augmentation methods suggested by Pacifico. et al. (2004) and Dudoit and van der Laan (2008). Our experience with these methods is that they are not as powerful as other resampling methods we have considered, at least in the context of other error rates; see the comparisons in Romano and Wolf (2007). While augmentation is a general approach that exploits the relationship between the familywise error rate and a given generally weaker measure of error control, it appears that the idea behind augmentation is too crude in that the construction does not really make full use of the given measure of error control desired. Nor does it take into account the dependence structure in the problem, outside the first stage where control of the familywise error rate is used. Indeed, after the first stage, a given number of additional hypotheses are rejected at the second stage, and this number only depends on the number of rejections at the first stage and not, for example, on the dependence structure of the remaining test statistics to be tested.

Finally, it would be interesting to improve the procedure, perhaps by incorporating an estimate of s_0. An alternative but similar approach might first apply some kind of thresholding (say by a familywise error rate controlling procedure) to reduce the number of hypotheses under consideration.

5 Reply to Daniel Yekutieli

Of course, we wish we could propose a method with finite sample validity which implicitly or explicitly accounts for the dependence structure in the problem. Unfortunately, even in single testing, this is usually too much to hope for in nonparametric problems, but we believe that resampling methods can still be quite useful and reliable with sufficiently informative data. Of course, we point out the obvious fact that, in order for the BH procedure, or any other procedure which claims finite sample control based on marginal p-values, to truly exhibit finite sample control, the p-values must be exact in the sense of (1) in the paper. Of course, this requirement is almost never satisfied in practice, as p-values often rely on either asymptotic or resampling approximations.

Apparently, it is indeed quite challenging to construct a reasonable scenario where the Benjamini–Hochberg (BH) method fails to control the FDR. However, suppose we are in a situation where the exact sampling distribution of the test statistics is multivariate normal with a known covariance matrix Σ, which corresponds to an asymptotic approximation of the problem studied here. In the case $s = 2$ with both null hypotheses true and with negative correlation between the test statistics, control of the BH method reduces to the validity of Simes inequality. In this case, it is known

to fail; see, for example, Samuel-Cahn (1996) for a counterexample in the one-sided case. To the best of our knowledge, it is not known in general whether the BH method ever fails in the two-sided case, even if the covariance matrix exhibits extreme negative dependence. The statement that the FDR of the BH method approaches $\alpha s_0/s$ for large n and any P seems unsubstantiated, unless one has further knowledge of the limiting covariance matrix Σ. The validity of the BH method for multivariate normal test statistics in the two-sided case is interesting and deserves further thought. Certainly, a highlight of our work is that no assumptions are required on the limiting covariance matrix, in either the one- or two-sided cases.

Yekutieli's argument for the conservatism of FDR controlling procedures when the non-null tested effects are small is nice. The problem is essentially reduced to the study of control of the FDR under the complete null hypothesis when all null hypotheses are true. However, the argument does assume exchangeability, and one must know that the given method controls the FDR under the complete null. Of course, the BH method may not do so in general, and one is left with deciding which method is most appropriate.

To be clear, we do not assume that all false null hypotheses are rejected with probability tending to one; rather, it is a proven consequence of very basic assumptions concerning the limiting behavior of the test statistics under the fixed known data generating mechanism P. A more complete asymptotic framework would consider uniformity with respect to P, as well as s getting large (as discussed above in the response to Sarkar and Heller).

References

Dudoit S, van der Laan MJ (2008) Multiple testing procedures with applications to genomics. Springer series in statistics. Springer, New York

Efron B (2008) Microarrays, empirical Bayes and the two-groups model. Stat Sci 23:1035–1061

Gavrilov Y, Benjamini Y, Sarkar SK (2008) An adaptive step-down procedure with proven FDR control. Ann Stat (in press)

Genovese CR, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Stat 32(3):1035–1061

Korn EL, Troendle JF, McShane LM, Simon R (2004) Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plann Inference 124:379–398

Lehmann EL, Romano JP (2005a) Generalizations of the familywise error rate. Ann Stat 33(3):1138–1154

Lehmann EL, Romano JP (2005b) Testing statistical hypotheses, 3rd edn. Springer, New York

Pacifico. M, Genovese. C, Verdinelli. I, Wasserman. L (2004) False discovery control for random fields. J Am Stat Assoc 99:1002–1014

Romano JP, Shaikh AM (2008) Inference for identifiable parameters in partially identified econometric models. J Stat Plann Inference 138:2786–2807

Romano JP, Wolf M (2007) Control of generalized error rates in multiple testing. Ann Stat 35(4):1378–1408

Romano JP, Shaikh AM, Wolf M (2008) Formalized data snooping based on generalized error rates. Econom Theory 24(2):404–447

Samuel-Cahn E (1996) Is the Simes improved Bonferroni procedure conservative? Biometrika 83:928–933

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66(1):187–205