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

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

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.

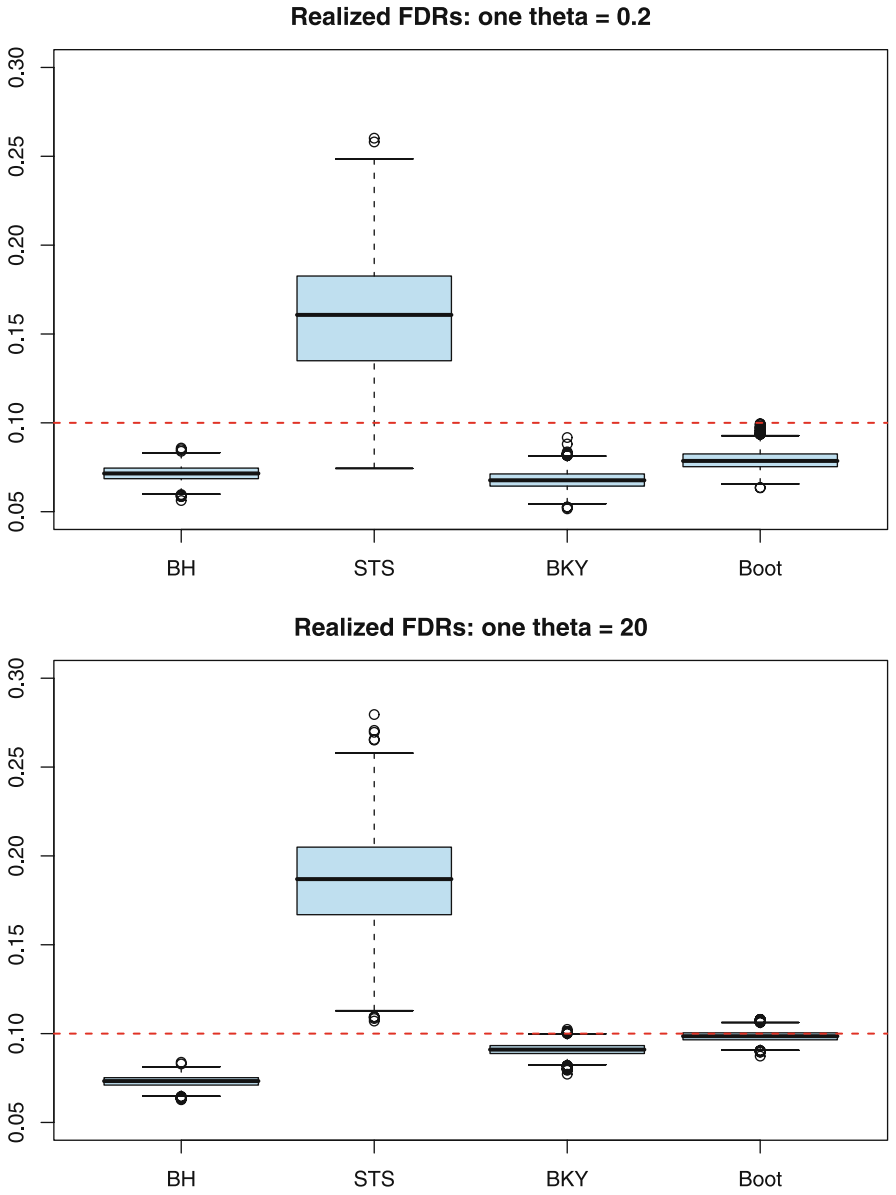

Fig. 1 Boxplots of the simulated FDRs described in Sect. 7.2. The *horizontal dashed lines* indicate the nominal level $\alpha = 0.1$

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

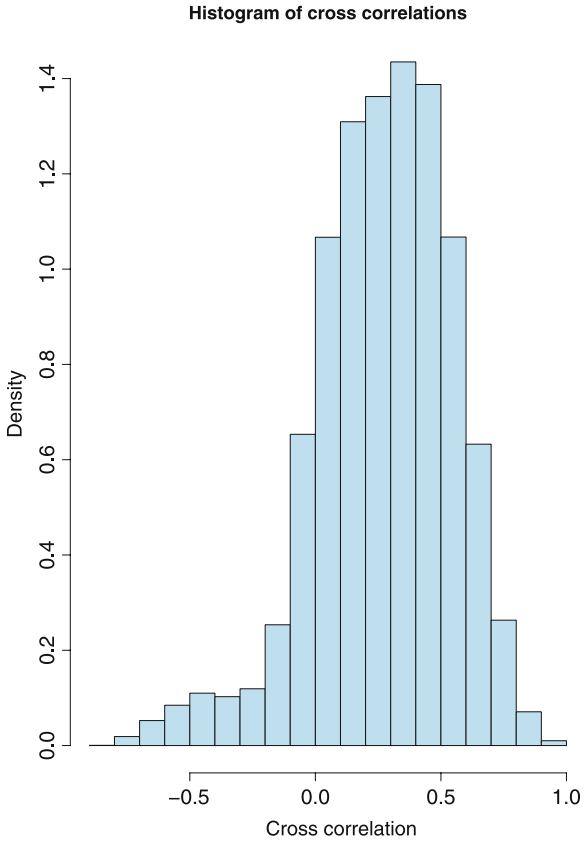

Fig. 2 Histogram of the $208 \cdot 209/2 = 21,736$ cross correlations between the excess returns of the 209 hedge funds. Since it is not true that the majority of these correlations are close to zero, the weak dependence assumption of Storey et al. (2004) is clearly violated

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