Statistics M.Sc. seminar

Control FDR using bootstrap and subsampling

Dan Navon

The article was written by **Joseph P. Romano, Azeem M. Shaikh & Michael Wolf**

Was published in **October 2008**

Five comments articles were published – two of them are nowadays professors in our department – Ruth Heller and Dani Yekutieli.

We will discuss some of their comments.

Table of contents

FDR - introduction and former methods *

Introduction

* We test **s null hypotheses** simultaneously.
* When s is large, we can't control the possibility to have one mistake or more (**Familywise error rate**), is we want reasonable power.

* Some methods were suggested to control the proportion of the false rejections.

* ***false discovery proportion*** **(FDP)**, defined to be the fraction of rejections that are false rejections

* F – false rejection number ; R – total rejection number

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

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

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

Previous methods

BH procedure

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.

* Let's define

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

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

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

* This method provide asymptotic control of the FDR, under two conditions:

* 1. The p-values are independent.

2. And, under H0: $P\{\hat{p}_{n,i} \leq u\} \leq u \quad \text{for any } u \in (0, 1)$

Or, in some cases: $\limsup_{n\to\infty} P\{\hat{p}_{n,i} \leq u\} \leq u$

* Yekutieli proved that the method is valid also in case of low positive dependency.

* Also he comment that in all possible structure of dependency:

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;

* He had also proved, that under any dependency, we can control FDR if we use the follows:

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

where $C_k = \sum_{r=1}^{k} \frac{1}{r}$. Note that $C_s \approx \log(s) + 0.5$

* But the power in this case is very low.

STS method

* If we know the number of true null hypotheses (S_0), we can improve BH, by using:

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

* But we don't know S_0, so we should estimate it:
* For example, Storey et al. (2004) suggest the following estimator:

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

where $\lambda \in (0, 1)$ is a user-specified parameter.

* **Explanation:** 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 is a small-sample adjustment to make the procedure slightly more conservative, and to avoid an estimator of zero for S_0.

* **How choosing λ?**
* Big and controversial question…

* In the forward simulations the authors have chose λ=**0.5**, because this is the default of the inventor of the method.

* Ruthi Heller, in her comment, argued that the follow is more reliable:

$$\lambda = \frac{\alpha}{1 + \alpha}$$

In the reply, another simulations were presented (according to that choice), but the results weren't better.

Advantages:

* Storey et al. (2004) prove that this adaptive procedure controls the FDR asymptotically whenever a weak dependence condition holds – **independence / dependence within blocks** / mixing-type situations.

Disadvantages:

* 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.

*BKY method

* For this reason, Benjamini et al. (2006) develop an alternative procedure, which works as follows:

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 $\hat{s}_0 = s - r$.

The problem:

* We saw three method, with different demanded conditions .

* But, none of them have used in the **structure of the data**.

The proposed solutions:

* The authors wont to try improving the results, by using the **estimated distribution of the p-values**. And estimate it by bootstrap or subsampling methods.

* **Validity:**
* Benjamini et al. (2006) **prove** that this procedure controls the FDR whenever the p-values are **independent** of each other.

* They also provide **simulations** which suggest that this procedure continues to control the FDR under **positive dependence**.

*Motivation for new methods

* In order to motivate our procedures, first note that for any stepdown procedure based on critical values c1, . . . cs , we have that:

$$\mathrm{FDR}_P = E_P\left[\frac{F}{\max\{R,1\}}\right] = \sum_{1\le r\le s}\frac{1}{r}E_P[F|R=r]P\{R=r\}$$

$$= \sum_{1\le r\le s}\frac{1}{r}E[F|R=r]$$

$$\times P\{T_{n,(s)}\ge c_s,\ldots,T_{n,(s-r+1)}\ge c_{s-r+1},T_{n,(s-r)}<c_{s-r}\}$$

* 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 = s0, $T_{n,r:s0}$ denotes the *r*'th largest of the test statistics corresponding to the true hypotheses.
* Then, with probability approaching one, we can assume that the false H0 were rejected, and we have that:

$$\mathrm{FDR}_P = \sum_{s-s_0+1 \le r \le s} \frac{r-s+s_0}{r} \times P\{T_{n,s_0:s_0} \ge c_{s_0}, \ldots, T_{n,s-r+1:s_0} \ge c_{s-r+1}, T_{n,s-r:s_0} < c_{s-r}\}$$

* Our goal is to ensure that this expression is bounded above by α for any P, at least asymptotically.

* To this end, first consider any P such that s0 = 1. Then, FDR is simply:

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

* A suitable choice of c1 is thus the smallest value for the expression above 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*α $\geq$1, then c1 so defined is equal to $-\infty$.

* Having determined c1, now consider any P such that s0 = 2. Then, FDR is simply:

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

* A suitable choice of c2 is therefore the smallest value for which FDR is bounded above α.

* Note that when j = s, FDR simplifies to:

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

* So equivalently,

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

But we don't know P!!

*A bootstrap approach

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

$$\theta(P) = \left(\theta_1(P), \ldots, \theta_s(P)\right)$$

* The null hypotheses may be one-sided, in which case: $H_j : \theta_j \leq \theta_{0,j} \quad \text{vs.} \quad H_j' : \theta_j > \theta_{0,j}$

* or the null hypotheses may be two-sided, in which case $H_j : \theta_j = \theta_{0,j} \quad \text{vs.} \quad H_j' : \theta_j \neq \theta_{0,j}.$

* We will consider the 'studentized' test statistics:

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

* Or:

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

* $\hat{\sigma}_{n,j}$ may either be identically equal to 1 or an estimate of the standard deviation of $\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j})$

* 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 $\hat{P}_n$

*Let's go to the bootstrap world:

$X^* = (X_1^*, \ldots, X_n^*)$ be distributed according to $\hat{P}_n$ and denote by $T_{n,j}^*$, $j = 1, \ldots, s$, test statistics computed from X^*

Then, we get:

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

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

Question: Why we use $\widehat{P_n}$ *instead of* P_0?

* **Answer:** 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 Hj is true.

* **How to choose $\widehat{P_n}$?**
* The exact choice of $\widehat{P_n}$ will, of course, depend on the nature of the data.
* If the data X = (X1, . . . ,Xn) are i.i.d., then a suitable choice of $\widehat{P_n}$ is the **empirical distribution**, as in Efron (1979).
* If, on the other hand, the data constitute a time series, then $\widehat{P_n}$ should be estimated using a suitable **time series bootstrap method**.

Question:

is $\theta\left(\widehat{P_n}\right) = \widehat{\theta_n}$?

* **Answer:** No! The first is the plug-in estimator , by the bootstrap method, and the second is the empirical estimator.
* But, under some conditions, they are equal.
* Importantly, it depends on the **bootstrap method** – it's possible just under Efron's bootstrap, the circular blocks bootstrap, or the stationary bootstrap in Politis and Romano.
* On the other hand, this substitution does not in general affect the **asymptotic validity**.

* Given a choice of $\widehat{P_n}$, define the critical values recursively as follows: having determined $\widehat{c_{n,1}}, \ldots, \widehat{c_{n,j-1}}$ compute $\widehat{c_{n,j}}$ according to the

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

* **Remark:** For each j, we take the j smallest T*, and assume they are correspond to H0.

* **Computational problem:**

* **Wenge Guo:**

"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.**"

* **Alternative procedure:**

* So, he suggests alternative procedure. He's main idea is not calculate $\widehat{P_n}$ by the bootstrap sampling, but using the bootstrap sampling directly.

*having determined $\widehat{c_{n,1}}, \ldots, \widehat{c_{n,j-1}}$ compute $\widehat{c_{n,j}}$ according to the rule:

$$\mathrm{FDR}_{j,\hat{P}}(c) = \sum_{s-j+1 \le r \le s} \frac{r-s+j}{r} \times \hat{P}\{T_{j:j} \ge c, \ldots, T_{s-r+1:j} \ge \hat{c}_{s-r+1}, T_{s-r:j} < \hat{c}_{s-r}\}$$

$$= \frac{1}{B} \sum_{b=1}^{B} \sum_{s-j+1 \le r \le s} \frac{r-s+j}{r} \times I\left\{T^{*b}_{j:j} \ge c, \ldots, T^{*b}_{s-r+1:j} \ge \hat{c}_{s-r+1}, T^{*b}_{s-r:j} < \hat{c}_{s-r}\right\}$$

* But:

$$I\left\{T^{*b}_{j:j} \geq c, \ldots, T^{*b}_{s-r+1:j} \geq \hat{c}_{s-r+1}, T^{*b}_{s-r:j} < \hat{c}_{s-r}\right\}$$
$$= I\left\{T^{*b}_{j:j} \geq c\right\} \cdots I\left\{T^{*b}_{s-r+1:j} \geq \hat{c}_{s-r+1}\right\} \cdot I\left\{T^{*b}_{s-r:j} < \hat{c}_{s-r}\right\}$$

* For every b = 1, . . . , B, let $r^{*b}{}_j$ denote the total number of rejections when applying a stepdown aprocedure with the critical constants $\widehat{c_i}$, i = 1, . . . , j – 1, to the ordered test statistics $T^{*b}{}_{i:j}$

* and hence:

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

* **Authors reply:**

"We agree that the main drawback of the bootstrap method is its computational burden…

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."

Question: are their claim is right?

*Theorem

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* $\mathrm{supp}(J_{I(P)}(P))$
(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*

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

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

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

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

* There are a lot of assumptions…
* **Exchangeablity:**

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 $|K| = j$.

* It seems like a strong assumption.
* In the subsampling method, we'll can omit this assumption.

* **Another comment:**
* What is the meaning of $\limsup_{n\to\infty} \mathrm{FDR}_P \le \alpha$
* In reply to Heller comment, the authors have cleared it:

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 $\mathrm{FDR}_{n,P} \le \alpha$ for all $n \ge n_0$. 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 $n \ge n_0(\alpha, P)$.

Partial proof

- We'll bring the proof that all false H0 are rejected asymptotically in probability tending to 1:
- 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 s0 = 1.
- Since $\theta_j(P) \neq \theta_{0,j}$ for j≥ 2, it follows that

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

* But this not say anything, if

$$\widehat{c_{n,j}} \to \infty$$

So, we need to proof that $\widehat{c_{n,j}}$ are bounded above.

* Recall that $\widehat{c_{n,j}}$ is defined as follows: having determined $\widehat{c_{n,1}}$,… $\widehat{c_{n,j-1}}$,
* $\widehat{c_{n,j}}$ bring to minimum:

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

* Which is bounded above by: $j\,\hat{P}_n\{T^*_{n,j:j} \ge c\}$
* Which is bounded above by: $s\,\hat{P}_n\{T^*_{n,s:s} \ge c\}$

* Therefore, $\widehat{c_{n,j}}$ is bounded above by the $1-\alpha/s$ quantile of the (centered) bootstrap distribution of the maximum of all s variables.

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

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

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

* In this notation, the previously derived bound for $\widehat{c_{n,j}}$ 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(·,P) is strictly increasing at $M^{-1}(1 - \frac{\alpha}{s} + \epsilon,\ P)$. *For such* 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, $\widehat{c_{n,j}}$ is with probability tending to one less than $M^{-1}(1 - \frac{\alpha}{s} + \epsilon,\ P)$.

* **The claim that $\widehat{c_{n,j}}$ is bounded above in probability is thus verified.**

*The theorem itself is proven much for the assumption that

$$\rho\left(J_{n,\{1,\ldots,s\}}(P), J_{n,\{1,\ldots,s\}}(\hat{P}_n)\right) \xrightarrow{P} 0$$

*We should also pay attention to **Guo comment**:

"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."

*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**.

* 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 i'th subset of data of size b.

* Let $T_{n,b,i,r:t}$ denote the t'th largest of the test statistics

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

* Finally, define critical values $\widehat{c_{n,1}}, \ldots, \widehat{c_{n,s}}$ recursively as follows:

* having determined $\widehat{c_{n,1}}, \ldots, \widehat{c_{n,j-1}}$

compute $\widehat{c_{n,j}}$ according to the rule:

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

Theorem 2

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) $\text{supp}(J_{I(P)}(P))$ *is connected*

Then, the stepdown procedure with critical values defined by (23) satisfies

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

Remark: The above approach can be extended to dependent data as well. For example, if the data $X = (X1, \ldots, Xn)$ form a stationary sequence, we would only consider the $n - b + 1$ subsamples of the form $(Xi, Xi+1, \ldots, Xi+b-1)$. Generalizations for nonstationary time series, random fields, and point processes are further discussed in Politis et al. (1999).

* Interestingly, even under the exchangeability assumption, where both the bootstrap and subsampling are asymptotically valid, the two procedures are not asymptotically equivalent.
* To see this, suppose that s = s0 = 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 $\widehat{c_{n,1}}$ 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 $Z_{1-\alpha}$.

*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.

Comparison of FDR control and power

* We generate random vectors X1, . . . , Xn 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 $T_{n,j} = \sqrt{n}\ \hat{\theta}_j / \hat{\sigma}_j$, where

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

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

- **Common correlation:** $\sigma_{i,j} = \rho$, where ρ=0, 0.5, or 0.9.

- **Power structure:** $\sigma_{i,j}$ = ($\rho^{|i-j|}$, where ρ = 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 ρ = 0.5;

and across classes, there is a common correlation of ρ =–0.5.

* **Means:**
* We consider four scenarios for the mean vector $\theta = (\theta_1, \ldots, \theta_s)$
* - All θ_j = 0.
* - Every fifth θ_j = 0.2, and the remaining θ_j = 0, so there are ten θ_j= 0.2.
* - Every other θ_j = 0.2, and the remaining θ_j = 0, so there are twenty five θ_j = 0.2.
* - All θ_j = 0.2

* We'll run five methods: BH, STS, BKY, bootstrap method, and Subsampling method. **The results of the subsampling method were not satisfactory, and omitted from the article!**

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

- 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 θ_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 θ_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 α = 0.1.

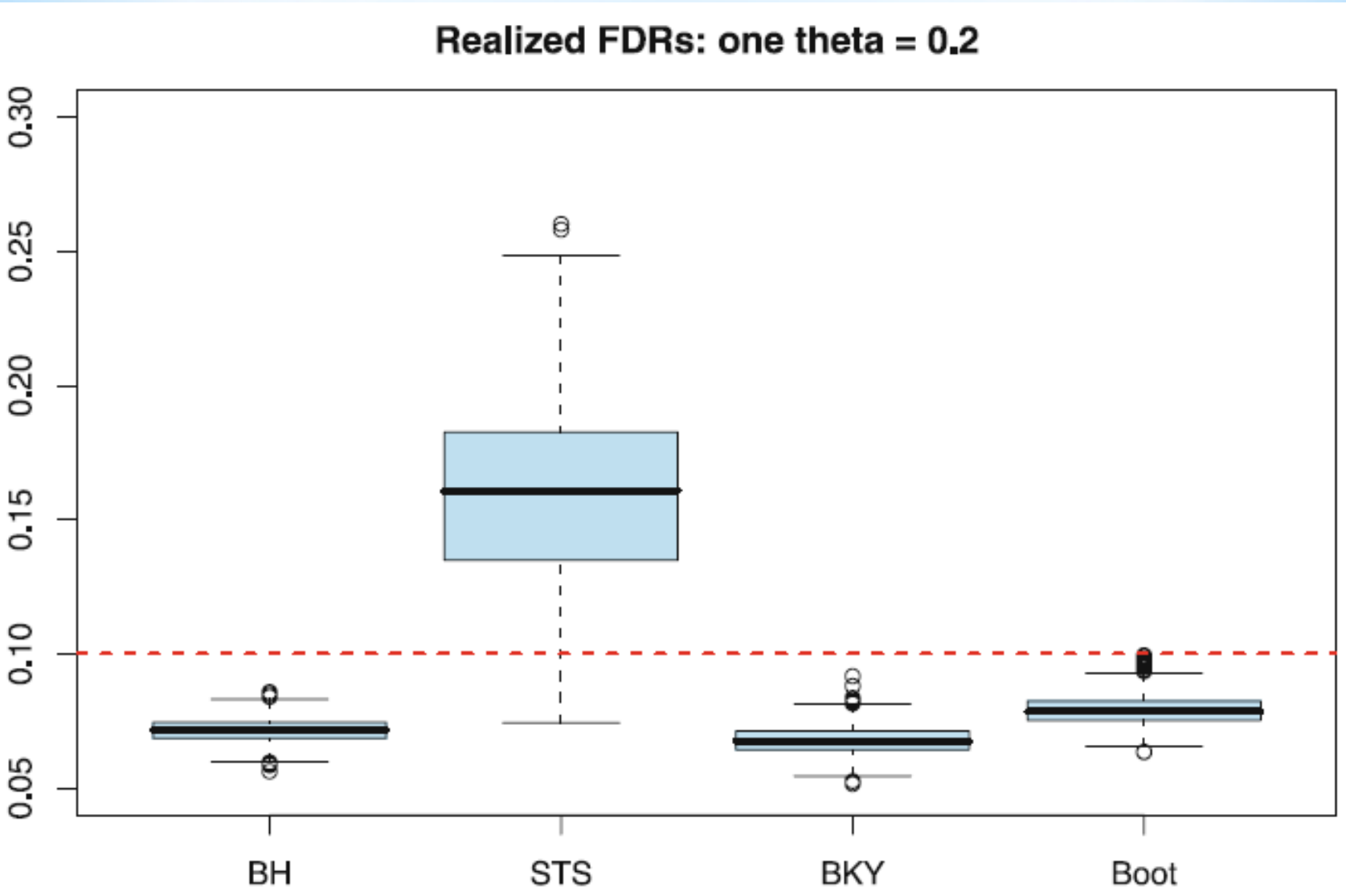
- STS is often more powerful than Boot, but some of those comparisons are not meaningful, namely when Boot provides FDR control while STS does not.

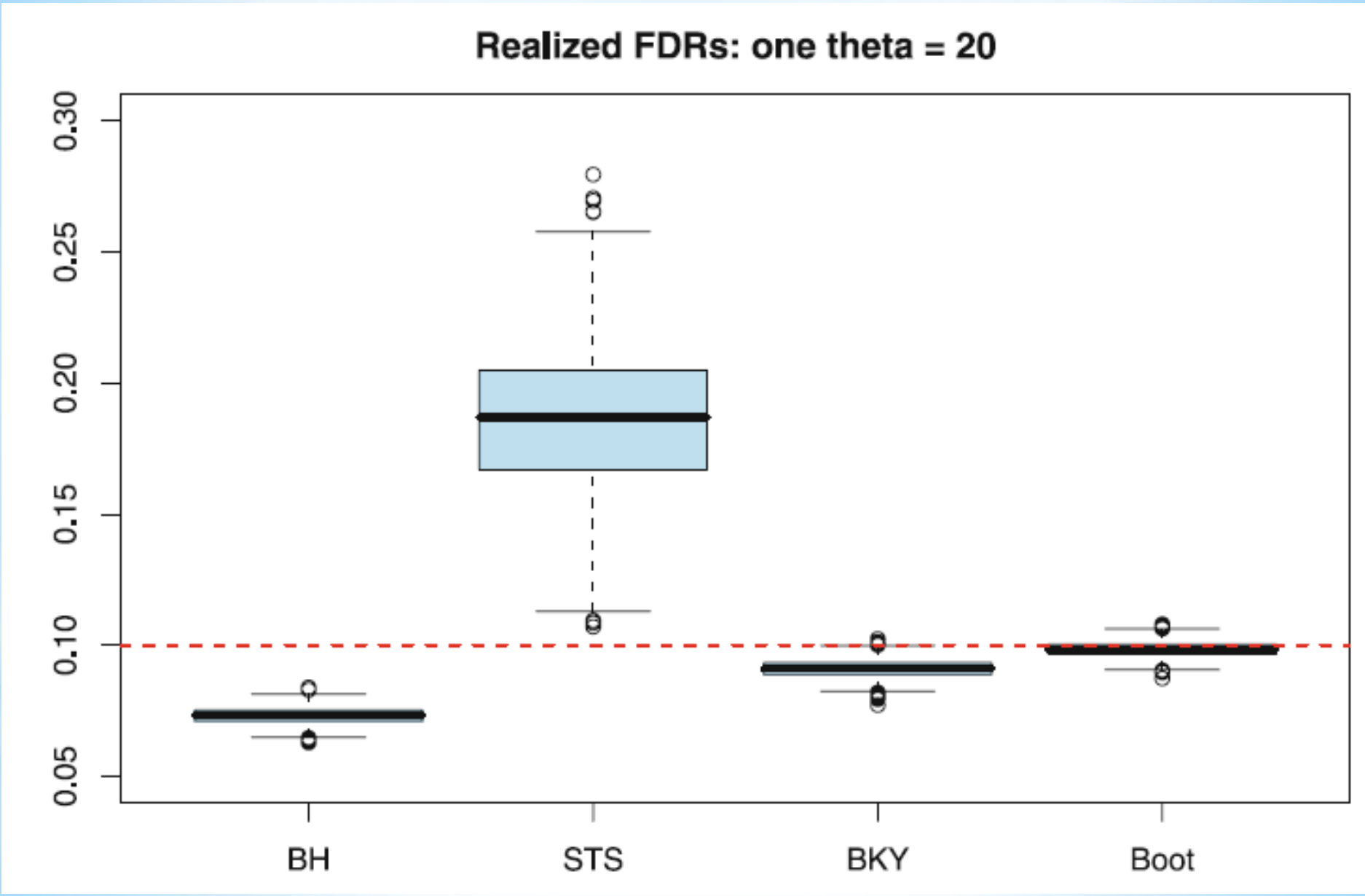
Table 2 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario and the number of bootstrap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

Robustness of FDR control against random correlations

* The goal of this subsection is to study whether FDR control is maintained for ‘general’ covariance matrices.
* We generate 1,000 random correlation matrices uniformly from the space of positive definite correlation matrices. We reduce the dimension from s = 50 to s = 4 to conter the curse of dimensionality.
* As far as the mean vector is concerned, two scenarios are considered: one θ_j = 0.2 and one θ_j = 20. The latter scenario results in **perfect power** for all four methods.


Realized FDRs: one theta = 0.2
0.05
0.10
0.15
0.20
0.25
0.30
BH
STS
BKY
Boot


Realized FDRs: one theta = 20
0.30
0.25
0.20
0.15
0.10
0.05
BH
STS
BKY
Boot

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

Empirical applications

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
* θ_j = μj − μB, where μj is the expected return of the j th hedge fund, and μB is the expected return of the benchmark.

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
* θ_j = μj − μB, where μj is the expected return of the j th hedge fund, and μB is the expected return of the benchmark.

*Naturally, the estimator of θ_j is given by

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

Accordingly, one has to account for this **time series** nature in order to obtain valid inference.

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

Pairwise fitness correlations

- The pairwise correlations of seven numeric ‘fitness’ variables, collected from n = 31 individuals, are analyzed. Denote the $s = \binom{7}{2} = 21$ pairwise population correlations. The analysis is based on B = 5,000 repetitions.
- we use Efron’s bootstrap to both compute individual p-values and to carry out our bootstrap FDR procedure.

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

Another issues to discussion *

* A. Finite sample size.

* B. Big s and low n, or $s \to \infty, as\ n \to \infty$

* C. Outlines

* D. Potential to control FDP?

* E. benefits of stepdown defining of critical values