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

DISCUSSION

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

José A. Ferreira · Mark A. van de Wiel

Published online: 30 October 2008

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

DISCUSSION

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

Wenge Guo

Published online: 30 October 2008

1 Introduction

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

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

2 High-dimensional, low sample size data analysis

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

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

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

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

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

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

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

3 Computational problem

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

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

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

where

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

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

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

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

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

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

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

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

4 Some possible extensions

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

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

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

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

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

References

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

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

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

DISCUSSION

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

Sanat K. Sarkar · Ruth Heller

Published online: 30 October 2008

1 Introduction

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

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

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

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

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

2 What is being controlled?

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

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

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

3 Other relevant methods

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

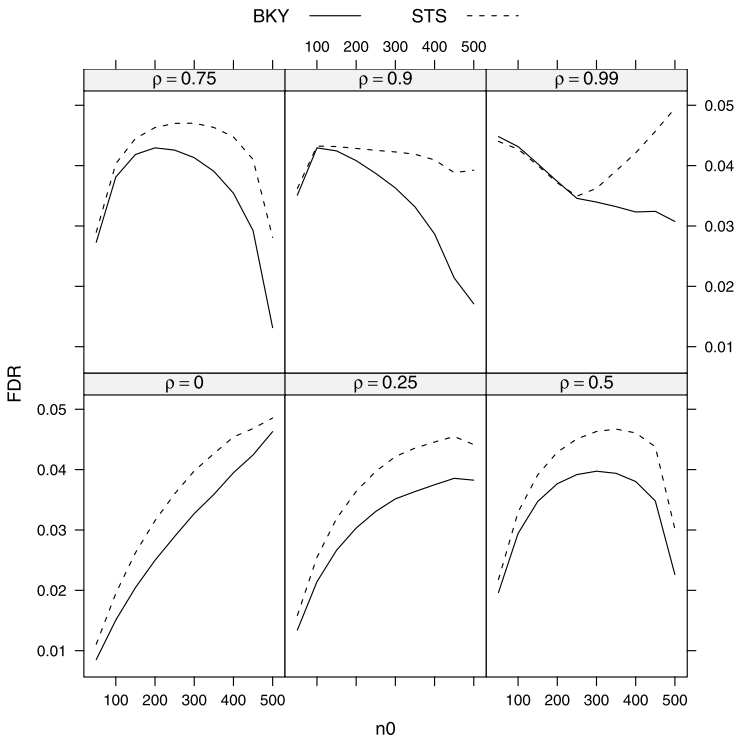

Fig. 1 Comparison of simulated FDR's of the BKY and STS with $\lambda = \alpha/(1+\alpha)$, with $\alpha = 0.05$. Each simulated FDR was based on 20,000 replications, n_0 is the number of true nulls

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

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

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

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

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

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

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

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

4 Final points

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

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

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

References

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

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

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

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

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

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

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

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

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

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

DISCUSSION

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

James F. Troendle

Published online: 30 October 2008

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

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

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

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

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

References

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

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

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

DISCUSSION

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

Daniel Yekutieli

Published online: 30 October 2008

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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