Test
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

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

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

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