ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

journal of statistical planning and inference

www.elsevier.com/locate/jspi

False discovery rate control for non-positively regression dependent test statistics

Daniel Yekutieli*

Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Abstract

In this paper we present a modification of the Benjamini and Hochberg false discovery rate controlling procedure for testing non-positive dependent test statistics. The new testing procedure makes use of the same series of linearly increasing critical values. Yet, in the new procedure the set of p-values is divided into subsets of positively dependent p-values, and each subset of p-values is separately sorted and compared to the series of critical values. In the first part of the paper we introduce the new testing methodology, discuss the technical issues needed to apply the new approach, and apply it to data from a genetic experiment.

In the second part of the paper we discuss pairwise comparisons. We introduce FDR controlling procedures for testing pairwise comparisons. We apply these procedures to an example extensively studied in the statistical literature, and to test pairwise comparisons in gene expression data. We also use the new testing procedure to prove that the Simes procedure can, in some cases, be used to test all pairwise comparisons.

The control over the FDR has proven to be a successful alternative to control over the family wise error rate in the analysis of large data sets; the Benjamini and Hochberg procedure has also made the application of the Simes procedure to test the complete null hypothesis unnecessary. Our main message in this paper is that a more conservative approach may be needed for testing non-positively dependent test statistics: apply the Simes procedure to test the complete null hypothesis; if the complete null hypothesis is rejected apply the new testing approach to determine which of the null hypotheses are false. It will probably yield less discoveries, however it ensures control over the FDR.

Keywords: False discovery rate; Pairwise comparisons; Dependent test statistics

1. Introduction

The Benjamini and Hochberg (1995) false discover rate controlling procedure (BH procedure) is known to control the FDR for positively dependent test statistics (Benjamini and Yekutieli, 2001). In this paper we present a modification of the BH procedure for controlling the FDR for non-positive dependent test statistics. The new testing procedure makes use of the series of linearly increasing critical values used in the BH procedure—$\{iq/m\}_{i=1}^{m}$. But while in the BH procedure the entire set of p-values is sorted and compared to the series of critical values, in the new procedure the set of p-values is divided into subsets of positively dependent p-values, and each subset of p-values is separately sorted and compared to the series of critical values.

* Fax: +972 3 640 9357.
E-mail address: yekutiel@post.tau.ac.il.

0378-3758/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.jspi.2007.06.006

The idea of applying the BH procedure to separate subsets was first used in Benjamini and Yekutieli (2005) for testing m, non-positively dependent, two-sided test statistics. The authors suggested to separately apply the BH procedure at level $q/2$ to the two corresponding sets of m positively dependent one-sided test statistics null hypotheses, and showed that the FDR is controlled at level q on all $2m$ one-sided hypotheses. In this paper we will generalize this idea to more than two, not necessarily disjoint, subsets of positively dependent test statistics.

Throughout the paper the vector of p-values are co-monotone transformations of the corresponding test statistics (e.g. one-sided p-values), hence posses equivalent positive dependency properties. For the sake of brevity, we will alternately discuss p-values or test statistics. $\vec{P} = \{P_1, \ldots, P_m\}$ is the vector of p-values corresponding to the tested hypotheses; $P_{(1)} \leqslant \cdots \leqslant P_{(m)}$ are the sorted p-values; m_0 is the number of true null hypotheses ($0 \leqslant m_0 \leqslant m$) and the distribution of each true null hypotheses p-value is stochastically larger than U[0, 1]; we will denote the complete null hypothesis—$\mathrm{H}_{\mathrm{c}}^0$ ($m_0 = m$).

The series of linearly increasing critical values was originally employed (Simes, 1986; Seeger, 1968) to test whether any of the null hypotheses are false:

Definition 1.1. Level α Simes test: if $\exists\ P_{(i)} \leqslant \alpha \cdot i/m$ then reject $\mathrm{H}_{\mathrm{c}}^0$.

It was shown that for independent test statistics (Simes, 1979) and later for positively dependent test statistics (Sarkar and Chang, 1997; Sarkar, 1998) that under $\mathrm{H}_{\mathrm{c}}^0$ the probability that the Simes procedure rejects $\mathrm{H}_{\mathrm{c}}^0$ is less than or equal to α. If, however, $\mathrm{H}_{\mathrm{c}}^0$ is not true ($m_0 < m$) the series of critical values cannot be used to determine which hypotheses are false null hypotheses, while controlling the probability of making at least on type I error at level α.

In their seminal paper, Benjamini and Hochberg (1995) introduced a new measure for type I error in multiple testing—the FDR—employed the series of critical values to test the individual null hypotheses, and showed that the resulting procedure controls the FDR at level $q \cdot m_0/m$ for independently distributed test statistics.

Definition 1.2. The level q BH procedure:

1. Let $k = \max\{i : P_{(i)} \leqslant iq/m\}$.
2. If $\exists k > 0$ then reject the null hypotheses associated with $\vec{R}_{\mathrm{BH}} = \{P_{(i)} : i = 1 \cdots k\}$; otherwise do not reject any of the null hypotheses.

Benjamini and Yekutieli (2001) proved that if the vector of test statistics, $\vec{T}$, is positive regression dependent on the subset of true null hypotheses test statistics $\vec{T}_0$ then the FDR of the level q BH procedure is less than or equal to $q \cdot m_0/m$.

Definition 1.3. $\vec{T}$ is positive regression dependent on $\vec{T}_0$: for any increasing set D, and for each $T_i \in \vec{T}_0$, $\Pr(\vec{T} \in D | T_i = t)$ is non-decreasing in t.

Benjamini and Yekutieli (2001) also presented a general-dependency FDR controlling procedure: applying the BH procedure at level $q/(\sum_{i=1}^{m} 1/i)$ offers FDR control at level q for all joint test statistic distributions. The shortcoming of this testing procedure is that it is considerably less powerful than the BH procedure.

In Section 2 we will define the new testing approach, address the problem of constructing positively dependent sub-vectors, and apply the new testing procedure to data from a genetic experiment. Section 3 is dedicated to the problem of testing pairwise comparisons. We will present FDR controlling procedures for testing pairwise comparisons. Apply these testing procedures to an example extensively studied in the statistical literature, and to test the pairwise comparisons in the expression level of 7129 genes. Following Yekutieli (2001), we also use the FDR controlling property of the new testing procedure to prove that the Simes procedure can be used to test all pairwise comparisons. In Section 4 we discuss the suggested use of the new testing approach.

2. The separate subsets BH (ssBH) procedure

To apply the ssBH procedure the vector of m p-values, $\vec{P}$, is divided into S sub-vectors, $\vec{P}^s$, for $s = 1 \cdots S$; let m^s denote the number of test statistics in $\vec{P}^s$ and let $\vec{P}_0^s$ denote the p-values corresponding to the true null hypotheses in $\vec{P}^s$.

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

Definition 2.1. *Level q ssBH procedure*:

1. For $s = 1 \cdots S$, apply the BH procedure at level $q \cdot m^s/m$ to test $\vec{P}^s$, and let $\vec{R}^s_{\mathrm{BH}}$ denote the p-values corresponding to the rejected hypotheses.
2. Reject the null hypotheses corresponding to $\vec{R}_{\mathrm{ssBH}} = \bigcup_{s=1}^{S} \vec{R}^s_{\mathrm{BH}}$.

By definition, the ssBH procedure is less powerful than the BH procedure, $\vec{R}_{\mathrm{ssBH}} \subseteq \vec{R}_{\mathrm{BH}}$. However, its power increases if there are $\vec{P}^s$ which include many small p-values, i.e. p-values corresponding to hypotheses rejected by BH procedure. In particular, if $\exists\ \vec{P}^s$ such that $\vec{R}_{\mathrm{BH}} \subseteq \vec{P}^s$ then $\vec{R}_{\mathrm{ssBH}} = \vec{R}_{\mathrm{BH}}$. Thus, it turns out that the BH procedure controls the FDR not only if the set of tested p-values is PRDS, but also if the set of rejected p-values is PRDS.

We are now ready to phrase the main result of this paper.

Proposition 2.2. *If Condition* 2.3 *holds then the level q* ssBH *procedure controls the* FDR *at level* $q \cdot m_0/m$.

Condition 2.3. For each $P_i \in \vec{P}_0$, $\vec{P}(P_i) = \cup\{\vec{P}^s : P_i \in \vec{P}^s\}$ is PRDS on P_i.

The proof of Proposition 2.2 closely resembles the proof of Theorem 1.2 in Benjamini and Yekutieli (2001) and is deferred to the Appendix.

2.1. PRDS random vectors

We will now address the problem of determining whether random vectors of commonly used test statistics are PRDS. In the next section we discuss the verification of Condition 2.3.

2.1.1. Multivariate normal test statistics

$\vec{X} \sim \mathrm{N}(\vec{\mu}, \Sigma)$. The vector of true null hypotheses is $\vec{X}_0 = \{X_i : \mu_i = 0\}$. $\vec{X}$ is PRDS on $\vec{X}_0$ if and only if $\sigma_{i,j} \geqslant 0$ for each $X_i \in \vec{X}_0$ and for each $X_j \in \vec{X}$ (Benjamini and Yekutieli, 2001). The problem is that in many cases the identity of the true null hypotheses is unknown. Thus, a stronger condition may be needed, $\sigma_{k,j} \geqslant 0$ for all X_k and X_j in $\vec{X}$.

2.1.2. Absolute valued multivariate normal

$\vec{Y} = |\vec{X}|$, $\vec{X} \sim \mathrm{N}(\vec{\mu}, \Sigma)$. $\vec{Y}$ is trivially PRDS if $\Sigma \equiv I$. Otherwise, for $0 < m_0 < m$, $\vec{Y}$ is generally not PRDS.

Example 2.4. (X_1, X_0) are bivariate normal with unit variance, $\mu_0 = 0$, $\mu_1 = 2$ and $\rho = 1$ (i.e. $X_1 = X_0 + 2$). $\Pr(|X_1| > 1 \,|X_0| = t) = 1$ for $t < 1$ or $3 < t$, but is $\frac{1}{2}$ for $1 \leqslant t \leqslant 3$.

For $\vec{\mu} = 0$ (i.e. under $\mathrm{H}^0_{\mathrm{c}}$) Karlin and Rinott (1981) proved that $\vec{Y}$ is multivariate total positivity of order 2 (MTP_2) if and only if there exists a diagonal matrix D with elements ± 1 such that the off-diagonal elements of $-D\Sigma^{-1}D$ are all non-negative. MTP_2 is a strong form of positive dependency which implies PRDS on any subset. Thus, absolute valued MVN are PRDS if the sub-vector or true null hypotheses is PRDS on any subset and independently distributed of the sub-vector of false null hypotheses.

Example 2.5. *Search for genetic loci effecting the bronx waltzer mutation*. The study, conducted by Karen P. Steel, consisted of 113 backcross progeny of two inbred mice strains. The data for each mouse included the genotype at 45 genetic loci situated on five chromosomes, and a series of phenotypes—measurements of 16 behavioral traits. A log odds (LOD) score is computed to test for linkage between each genetic marker and each phenotype—in this case a total of 720 tests.

At $q = 0.05$ the BH procedure applied to test all 720 null hypotheses yielded 147 discoveries; while the Benjamini and Yekutieli (2001) general-dependency FDR controlling procedure yielded 27 discoveries.

LOD score distribution is proportional to absolute valued MVN. The null hypothesis of no linkage is true if there are no QTL affecting the phenotype on the chromosome; if the chromosome contains at least one QTL affecting the phenotype then all the null hypotheses on the chromosome are false. LOD score statistics corresponding to different chromosomes are independent, and Yekutieli (2002) showed that the true null hypotheses on a chromosome corresponding to the

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

same phenotype are PRDS on all subsets. However, according to Example 2.4, when there are QTL on a chromosome affecting some, but not all of the phenotypes, the set of LOD scores corresponding to that chromosome are not PRDS on the subset of true null LOD scores.

The simplest way to ensure FDR control is to divide the 720 test statistics into the 16 disjoint sub-vectors corresponding to each phenotype and apply the ssBH procedure—at $q = 0.05$ this yields 28 discoveries. As LOD score statistics corresponding to markers on separate chromosomes are independent even if they test for linkage with different phenotypes, it is possible to consider 16^5 sub-vectors of 45 LOD scores—LOD scores in each of the five chromosomes can test for linkage with any one of the 16 phenotypes—and still keep Condition 2.3. As testing all 16^5 LOD score sub-vectors is not feasible, we only applied the ssBH to a subset of $S = 80$ sub-vectors. To choose these sub-vectors we examined the results of the BH procedure applied to test all 720 LOD scores, and for each chromosome, recorded the identity of the phenotype with the most discoveries; each of the 80 LOD score sub-vectors consist of the LOD scores for each combination of five chromosomes times 16 phenotypes, and the LOD scores corresponding to the phenotype yielding the most discoveries for the remaining four chromosomes; testing these 80 sub-vectors at level 0.05 the ssBH procedure yielded 41 discoveries.

We would like to point out that in this example multiplicity of phenotypes is unfavorable for the ssBH procedure. The size of PRDS LOD subsets cannot exceed 45 LOD scores, yet the number tests increases with the number of phenotypes. For comparison, if only Phenotypes 1–8 are considered in the analysis than the BH procedure yields 82 discoveries, the general dependency BH procedure only yields 11 discoveries, while the number of ssBH discoveries increases to 43.

The following example illustrates that the joint distribution of pairwise comparisons is not PRDS even if $\vec{\mu} = 0$.

Example 2.6. $X_{i,j} = (Z_i - Z_j)/\sqrt{2}$, for $i \neq j$ and $Z_1 \cdots Z_k$ are iid N(0, 1). The vector of test statistics is $\vec{T} = \{|X_{i,j}| : 1 \leqslant i \neq j \leqslant K\}$. While any pair of test statistics is MTP_2, some of the triplets, for example, the triplet $\{|X_{3,2}|, |X_{3,1}|, |X_{2,1}|\}$, are not PRDS. $D = \{(X_{3,2}, X_{3,1}) : |X_{3,2}| > 1, |X_{3,1}| > 1\}$ is an increasing set in $|X_{3,2}|$ and $|X_{3,1}|$. Conditioning on $X_{2,1} = t$, $X_{3,2}$ is $\mathrm{N}(-t/2, \frac{3}{4})$. As $X_{3,1} = X_{3,2} + X_{2,1}$, for $0 \leqslant t < 2$ the set D is $\{X_{3,2} > 1\} \cup \{X_{3,2} < -1 - t\}$, and for $-2 < t \leqslant 0$ the set D becomes $\{X_{3,2} < -1\} \cup \{X_{3,2} > 1 + t\}$. Thus, for $0 \leqslant t < 2$,

$$\Pr(D \,|\, |X_{2,1}| = t) = 2 \cdot \Phi\left(\frac{-1 - t/2}{\sqrt{3/4}}\right),$$

a decreasing function in t.

2.1.3. Studentized normal test statistics

Benjamini and Yekutieli (2001) proved PRDS dependency for random vectors which are strictly co-monotone, continuously differentiable, transformation of a continuous PRDS random vector and an independent continuous latent variable.

This implies that for $\vec{Y}$ absolute valued MVN PRDS on I_0 and S^2 an independently distributed χ^2_ν, $\vec{T} = \vec{Y}/S$ is PRDS on I_0.

If $\vec{X}$ is PRDS MVN on I_0 and $S^2 \sim \chi^2_\nu$ then $\vec{T} = \vec{X}/S$ is only PRDS on I_0 for either positive or negative values of $\vec{T}$. Nevertheless, Benjamini and Yekutieli (2001) established that this property is sufficient to ensure FDR control of the BH procedure providing that $q < \frac{1}{2}$.

2.2. Verifying Condition 2.3

Following the discussion in the previous section Condition 2.3 may be verified directly. Yet for random transformations of MVN test statistics a simpler condition can be used.

Condition 2.7. For $s = 1 \cdots S$, $\vec{P}^s$ is PRDS on $\vec{P}^s_0$.

Lemma 2.8. *If* $\vec{X}$ *is* MVN *then Condition* 2.7 *implies Condition* 2.3.

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

Proof. Let $\vec{X} \sim \mathrm{N}(\vec{\mu}, \Sigma)$ and $X_i \in \vec{X}_0$. All sub-vectors of $\vec{X}$ are MVN and share the same covariances. To complete the proof, note that for any $X_j \in \vec{X}(X_i)$, $\exists s$ such that X_i and X_j are in $\vec{X}^s$, hence Condition 2.7 implies that $\mathrm{cov}(X_i, X_j) \geqslant 0$. □

Applying the argument used in Section 2.1.3 yields the following corollary.

Corollary 2.9. *Condition* 2.7 *also implies Condition* 2.3 *for continuously differentiable*, *transformation of an* MVN *random vector and an independent continuous latent variable.*

Notice, however, that Condition 2.7 does not imply Condition 2.3 for absolute valued MVN (recall Example 2.6).

3. FDR control for pairwise comparisons

The problem of testing all pairwise comparisons was first discussed in Tukey (1953), and is still among the most challenging multiple comparison problems. Traditionally, the multiplicity problem was addressed by controlling for the family wise error rate (see Hsu, 1999). Williams et al. (1999) note that the power of the BH procedure seems to remain stable as the number of comparisons increase, and recommend use of the BH procedure for testing large pairwise comparison problems.

FDR control of the BH procedure in pairwise comparisons was extensively studied in simulations: Benjamini et al. (1993), Williams et al. (1999), Kesselman et al. (1999), Blair and Hochberg (1995). In all the studies, for all configurations of true and false hypotheses simulated, for balanced and for non-balanced designs, normal and non-normal distributions, the BH procedure controlled the FDR. Furthermore, the configuration of null hypotheses yielding greatest FDR levels, approaching the nominal level q, was $\mathrm{H}^0_{\mathrm{c}}$. There is, however, no theoretical proof for the validity of the BH procedure for testing pairwise comparisons. In this section we will present an ssBH procedure for testing pairwise comparisons. Following Yekutieli (2001), we will use the FDR controlling property of the ssBH procedure to prove FDR control of the BH procedure under $\mathrm{H}^0_{\mathrm{c}}$.

The data for testing pairwise comparisons consists of k independent group means: $\bar{X}_i \sim \mathrm{N}(\mu_i, \sigma_i^2/n_i)$ for $i = 1 \cdots k$. We assume, without loss of generality, that the group means are sorted: $\bar{X}_1 \leqslant \cdots \leqslant \bar{X}_k$. The hypotheses tested are H^0_{ji}: $\mu_j \geqslant \mu_i$ vs. the alternative $\mu_j < \mu_i$, for $j \neq i$. The pairwise comparisons problem is usually expressed as $k \cdot (k-1)/2$ two-sided hypotheses; however, as vectors of two-sided test statistics are not generally PRDS, in order to apply the ssBH procedure we express it as $m = k \cdot (k-1)$ one-sided hypotheses. In either case, applying the BH procedure yields equivalent results: the two-sided null hypothesis $\mu_j = \mu_i$ is rejected iff H^0_{ji} or H^0_{ij} is rejected; furthermore, testing one-sided null hypotheses also offers directional inference (see Benjamini and Yekutieli, 2005). The test statistics can be equal variance two-sample T statistics, or Welch modified two-sample T statistics for non-equal sample variance:

$$T_{ji} = \frac{\bar{X}_i - \bar{X}_j}{S_{ji}},$$

where S_{ji} is an independently distributed standard error estimator. The corresponding p-value is, $\mathbf{P}_{ji} = 1 - F_t(\mathbf{T}_{ji})$, where F_t denotes the corresponding t cdf.

A set of one-sided pairwise comparisons p-value, $\vec{P}^s$, is PRDS only if the two sets of indices $I_+ = \{i : P_{ji} \in \vec{P}^s\}$ and $I_- = \{j : P_{ji} \in \vec{P}^s\}$ are disjoint. As each PRDS subset $\vec{P}^s$ is subset of $\{P_{ji} : i \in I_+, j \notin I_+\}$, there is no need to test all PRDS subsets of p-values but only the subsets of the form

$$\vec{P}^{I_+} = \{P_{ji} : j \notin I_+, i \in I_+\} \quad \text{for } I_+ \subseteq \{1 \cdots k\}. \tag{1}$$

As the test statistics are studentized MVN the validity of the ssBH procedure follows from Corollary 2.9.

Definition 3.1. *The level* q ssBH *procedure for pairwise comparisons*: apply the level q ssBH procedure to all of the pairwise comparison p-value sub-vectors defined in (1).

Example 3.2. To illustrate its use, the ssBH procedure is applied to test the pairwise comparisons in the nitrogen content, between six groups of red clover plants, each inoculated with cultures of Rhizobium bacteria, presented in Erdman (1946).

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

Table 1
Erdman (1946) pairwise comparison p-values

Group means	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$	$\bar{X}_6$
$\bar{X}_1 = 13.26$	–	0.2655	0.01	0.0025	0	0
$\bar{X}_2 = 14.64$	0.7345	–	0.037	0.0115	0	0
$\bar{X}_3 = 18.70$	0.99	0.7105	–	0.2895	0.0115	0
$\bar{X}_4 = 19.92$	0.963	1	0.963	–	0.037	0
$\bar{X}_5 = 23.98$	0.9975	1	1	1	–	0.0175
$\bar{X}_6 = 28.82$	0.9885	0.9885	1	1	0.9825	–

Table 2
The results of the level 0.05 ssBH procedure applied to the Erdman (1946) data pairwise comparisons

k	$\frac{k \cdot 0.05}{30}$	BH	$I_+ = \{3 \cdots 6\}$	$I_+ = \{4 \cdots 6\}$	$I_+ = \{5, 6\}$
1	0.0017	$p_{15} = 0$	$p_{15} = 0$	$p_{15} = 0$	$p_{15} = 0$
2	0.0033	$p_{25} = 0$	$p_{25} = 0$	$p_{25} = 0$	$p_{25} = 0$
3	0.0050	$p_{16} = 0$	$p_{16} = 0$	$p_{16} = 0$	$p_{16} = 0$
4	0.0067	$p_{26} = 0$	$p_{26} = 0$	$p_{26} = 0$	$p_{26} = 0$
5	0.0083	$p_{36} = 0$	$p_{14} = 0.0025$	$p_{36} = 0$	$p_{36} = 0$
6	0.0100	$p_{46} = 0$	$p_{13} = 0.01$	$p_{14} = 0.0025$	$p_{46} = 0$
7	0.0117	$p_{14} = 0.0025$	$p_{24} = 0.0115^*$	$p_{24} = 0.0115$	$p_{35} = 0.0115^*$
8	0.0133	$p_{13} = 0.01$	$p_{23} = 0.0370$	$p_{35} = 0.0115^*$	$p_{45} = 0.0370$
9	0.0150	$p_{24} = 0.0115$		$p_{34} = 0.2895$	
10	0.0167	$p_{35} = 0.0115$			
11	0.0183	$p_{56} = 0.0175^*$			
12	0.0200	$p_{23} = 0.0370$			
13	0.0217	$p_{45} = 0.0370$			
14	0.0233	$p_{12} = 0.2655$			
15	0.0250	$p_{34} = 0.2895$			

In Table 1 we list the 30 p-value computed to test the pairwise comparisons. In Table 2 we list the results of the BH and ssBH procedures: in column 2 we list the series of 15 constants for testing the 15 hypotheses corresponding to p-values which are less than $\frac{1}{2}$; in column 3 we list the 15 ordered p-values used in the BH procedure. Comparing the p-values in column 3 to the critical values in column 2 yields 11 discoveries. In columns 4–6 we present the sorted p-values tested in the three sub-vectors corresponding to: $I_+ = \{3 \cdots 6\}$; $I_+ = \{4 \cdots 6\}$; $I_+ = \{5, 6\}$. These three sub-vectors yielded 10 discoveries. The only discovery missed was the comparison of group means 5 and 6.

For $k = 5$, the ssBH procedure involves testing 59 additional p-value sub-vectors. However, notice that H^0_{56} is only rejected if it is the 11th sorted p-value: according to Table 2, its p-value is smaller than the 11th critical value, yet greater than the 10th critical value; as the 10 p-values smaller than P_{56} include the negatively correlated P_{15} and P_{25}–P_{56} cannot be the 11th sorted p-value in any of the sub-vectors tested in the ssBH procedure. Hence H_{56} cannot be rejected by the ssBH procedure.

Procedure 3.1 includes $2^k - 2$ sub-vectors, many of which are redundant—if $I_+ = \{1 \cdots l\}$ then all of the p-values in $\vec{P}^{I_+}$ are greater then $\frac{1}{2}$; Yekutieli (2001) suggested a shorthand version of the ssBH procedure:

Definition 3.3. *The abridged level q* ssBH *procedure for pairwise comparisons*: apply the level q ssBH procedure to the sub-vectors $\vec{P}^{I_l} = \{P_{ji} : j < l, l \leqslant i\}$ where $l = 2 \cdots k$.

Procedure 3.1 may, in some cases, yield more discoveries than Procedure 3.3. Yet, as testing all $2^k - 2$ subsets is only feasible for small k, for large k we recommend using Procedure 3.3, and only if it produces considerably less discoveries than the BH procedure testing additional sub-vectors should be considered.

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

3.1. The validity of Simes procedure for testing all pairwise comparisons

In this section we assume that the pairwise comparisons test statistics share a common distribution F_T. Let $c_l = F_T^{-1}(1 - q \cdot l/m)$, and define w_k as the minimal number of pairwise comparisons within 3 disjoint sets of sub-groups:

$$w_k = \min_{k_1+k_2+k_3=k} \binom{k_1}{2} + \binom{k_2}{2} + \binom{k_3}{2}. \tag{2}$$

Lemma 3.4. *For q and k such that*: (1) $3 \cdot c_{m/2} > c_1$ *and* (2) $2 \cdot c_{m/2-w_k} > c_1$, *if the level q* BH *procedure rejects at least one hypothesis then Procedure* 3.3 *at level q will also reject at least one hypothesis.*

Procedure 3.3 controls the FDR, i.e. under H_c^0 the probability that Procedure 3.3 at level q rejects at least one hypotheses is less than or equal to q. Thus, Lemma 3.4 implies that under H_c^0 the probability that the BH rejects at least one hypothesis, and thus H_c^0, is also less than or equal to q:

Corollary 3.5. *Under the conditions of Lemma* 3.4, *the Simes procedure is valid for testing all pairwise comparisons.*

The following table list maximal values of k for which the two conditions in Lemma 3.4, given q and F_T, are kept:

q	t_{20}	t_{50}	t_{100}	t_{200}	N(0, 1)
0.1	13	19	23	25	27
0.05	22	39	50	58	68
0.01	55	164	285	398	587

Proof. It is assumed that at least one null hypothesis is rejected by the level q BH procedure, i.e. $r_{\mathrm{BH}} = |R_{\mathrm{BH}}| > 0$. Let $I^+ = \{i : P_{ji} \in \vec{R}_{\mathrm{BH}}\}$, $I^- = \{j : P_{ji} \in \vec{R}_{\mathrm{BH}}\}$. Now denote $I_2 = I^+ \cap I^-$, $I_3 = I^+ - I^-$ and $I_1 = I^- - I^+$. I_1 and I_3 are not empty. If $r_{\mathrm{BH}} > 0$ then the hypothesis corresponding to the minimal p-value, P_{1k}, must be rejected, hence $1 \in I^-$ and $k \in I^+$. Furthermore, $1 \notin I_2$ and $k \notin I_2$, thus $1 \in I_1$ and $k \in I_3$. We will now show that for each configuration of I_2 at least one hypothesis is rejected in Procedure 3.3.

Case 1: *I_2 is empty*. If I_2 is empty then R_{BH} is a subset of $\{P_{ji} : j \in I_1\}$—one of the sub-vectors tested in Procedure 3.3, in which case $\vec{R}_{\mathrm{ssBH}} = \vec{R}_{\mathrm{BH}}$.

Case 2: $\exists\ j, i \in I_2$ such that the $P_{ji} \in \vec{R}_{\mathrm{BH}}$. As half of the m hypotheses correspond to negative test statistics (hence cannot be rejected) all of the rejected p-values must be smaller than $q/2$, and the corresponding test statistics are all greater than $c_{m/2}$. Thus, $T_{ji} \geqslant c_{m/2}$, and since j and i are in I_2: $T_{1j} \geqslant c_{m/2}$ and $T_{ik} \geqslant c_{m/2}$. According to the first condition:

$$T_{1k} = T_{1i} + T_{ij} + T_{jk} \geqslant 3 \cdot c_{m/2} > c_1,$$

therefore $P_{1k} \leqslant q/m$ and $P_{1k} \in \vec{R}_{\mathrm{ssBH}}$.

Case 3: $\forall\ j, i \in I_2\ P_{ji} \notin \vec{R}_{\mathrm{BH}}$. Per definition, all hypotheses comparing group means within I_1 and within I_3 are also not rejected. Therefore, number of hypotheses not rejected by the BH procedure, $m - r_{\mathrm{BH}}$, exceeds the sum of the $m/2$ negative contrasts plus the number of contrasts comparing group means within I_1, I_2 and I_3. Thus, $m - r_{\mathrm{BH}} \geqslant m/2 + w_k$, or equivalently $r_{\mathrm{BH}} \leqslant m/2 - w_k$. This means that any rejected hypothesis corresponds to a test statistic greater than $c_{m/2-w_k}$. Let i belong to I_2, then both H_{1i}^0 and H_{ik}^0 are rejected. According to the second condition:

$$T_{1k} = T_{1i} + T_{ik} \geqslant 2 \cdot c_{m/2-w_k} > c_1.$$

Again $P_{1k} \leqslant q/m$ and $P_{1k} \in \vec{R}_{\mathrm{ssBH}}$. □

3.2. Analysis of microarray data

In this example we apply the ssBH procedure to data from a gene expression study in breast tumors (West et al., 2001). Gene expression levels were measured in 49 breast tumor mRNA samples using Affymetrix high-density

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

Table 3
The number of pairwise comparison differences discovered in the analysis of the West et al. (2001) gene expression data

p-value vector	BH $q = 0.05$	ssBH $q = 0.05$	BH $q = 0.025$	General dependency
$\vec{P}_{12}$	22	9	11	3
$\vec{P}_{13}$	3	2	2	0
$\vec{P}_{14}$	20	5	11	2
$\vec{P}_{23}$	8	6	2	0
$\vec{P}_{24}$	4	1	2	0
$\vec{P}_{34}$	1	0	0	0
$\vec{P}_{21}$	34	24	17	2
$\vec{P}_{31}$	0	0	0	0
$\vec{P}_{41}$	43	31	24	5
$\vec{P}_{32}$	3	0	0	0
$\vec{P}_{42}$	0	0	0	0
$\vec{P}_{43}$	21	15	8	2
$\vec{P}$	159	93	77	14

oligonucleotide chips containing 7129 human probe sequences (HuGeneFL chips). Two outcomes were measured for each tumor sample: estrogen receptor status—ER^-/ER^+; presence of affected lymph nodes—LN^-/LN^+.

The gene expression data was analyzed in R (R 2.0.1—Copyright, 2004, the R Development Core Team). The data was transformed—vsn R package (Huber et al., 2002), and then standardized—for each array the mean was set to 0 and the MAD was set to 1.

The 49 arrays were divided into four groups according to the lymph node and estrogen receptor status. Group 1: (LN^-, ER^-), $n_1 = 13$ arrays; Group 2: (LN^-, ER^+), $n_2 = 12$ arrays; Group 3: (LN^+, ER^-), $n_3 = 11$ arrays; Group 4: (LN^+, ER^+), $n_4 = 13$ arrays; For $i = 1 \cdots 4$, $\bar{X}_i$ is the 7129-component vector of mean standardized expression levels in Group i. It is assumed $\bar{X}_1 \cdots \bar{X}_4$ are independent MVN random vectors, with marginal distributions: $\bar{X}_{gi} \sim \mathrm{N}(\mu_{gi}, \sigma_g^2/n_i)$, where $g = 1 \cdots 7129$. In this example, we will consider testing, for each gene, all pairwise comparisons. The test statistics computed are:

$$T_{gji} = \frac{\bar{X}_{gi} - \bar{X}_{gj}}{S_g\sqrt{1/n_i + 1/n_j}},$$

where S_g is the pooled estimator of σ_g.

The only source of correlation between the expression level of a non-differentially expressed gene and any other gene is correlated measurement error. Reiner et al. (2003) argue that gene expression measurement errors are positively correlated. Thus, each vector of t-statistics, $\vec{T}_{ji} = \{T_{gji}\}_{g=1}^{7129}$, is PRDS on the subset of the components corresponding to non-differentially expressed genes $\vec{T}^0_{ji} = \{T_{gji} : \mu_{gj} = \mu_{gi}\}$. Employing this argument, if $j' \neq i$ and $i' \neq j$ then the vector $\vec{T}_{j'i'}$ is also PRDS on $\vec{T}^0_{ji}$. Let $P_{gji} = 1 - F_t(T_{gji})$, where F_t is the 45 degrees of freedom t cdf. Then Condition 2.7 holds for each:

$$\vec{P}_{I_+} = \{P_{gij} : g = 1 \cdots 7129,\ i \in I_+,\ j \notin I_+\} \quad \text{where } I_+ \subseteq \{1 \cdots 4\}. \tag{3}$$

As all $\vec{P}_{I_+}$ are MVN studentized normal, Condition 2.3 follows. Thus, the ssBH procedure applied to the p-value sub-vectors defined in (3) controls the FDR.

In Table 3 we present the results of the statistical analysis. At first we applied the BH procedure at $q = 0.05$ to test all 85, 548 pairwise comparisons. The results are listed in column 2 of Table 3. The BH procedure yielded 159 discoveries. The greatest number of differentially expressed genes was discovered in the Group 1 (ER−, LN−) and Group 4 (ER+, LN+) comparison: 43 genes were found to have higher expression levels in Group 4 than in Group 1 (row 9); 20 genes were found to have lower expression levels in Group 4 than in Group 1 (row 3).

For the ssBH procedure we constructed the 14 ($=2^4 - 2$) p-value vectors $\vec{P}_{I_+}$ defined in (3), corresponding to: $I_+ = \{4\}$; $I_+ = \{3\}$; $I_+ = \{3, 4\}$; $I_+ = \{2\}$; $I_+ = \{2, 4\}$; $I_+ = \{2, 3\}$; $I_+ = \{2, 3, 4\}$; $I_+ = \{1\}$; $I_+ = \{1, 4\}$; $I_+ = \{1, 3\}$;

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

$I_+ = \{1, 3, 4\}$; $I_+ = \{1, 2\}$; $I_+ = \{1, 2, 4\}$; $I_+ = \{1, 2, 3\}$. This yielded a total of 93 discoveries, see column 3 of Table 3.

The ssBH procedure was less powerful than the BH procedure at level 0.05; but was, generally, more powerful than the BH procedure at level 0.025 (see column 4). It yielded considerably more discoveries in $\vec{P}_{41}$, $\vec{P}_{21}$, $\vec{P}_{43}$ and $\vec{P}_{23}$ that were tested simultaneously in $I_+ = \{1, 3\}$; but since $\vec{P}_{14}$ cannot be tested alongside the many pairwise comparison discoveries in $\vec{P}_{41}$, $\vec{P}_{42}$, and $\vec{P}_{21}$ the ssBH procedure yielded considerably less discoveries than the level 0.025 BH procedure in $\vec{P}_{14}$.

Finally, in column 5 we present the results of the level 0.05 general-dependency FDR controlling procedure, which is simply the BH procedure at level $0.05/11.934 = 0.0042$. This procedure only yielded 14 discoveries.

4. Discussion

Control over the FDR has proven to be a successful alternative to control over the FWE in the analysis of large data sets. The use of the BH procedure also filled a gap in the statistical inference. Before its introduction a statistician would apply the Simes procedure and reject H_c^0—thus be able to tell the client that not all the null hypotheses tested are true null hypotheses, yet, in some cases, he would also have to inform the client that he could not tell him which of the null hypotheses are false.

Our offering in this paper is a testing procedure which employs the same set of constants used in the BH and Simes procedures, is more complicated to apply and makes less discoveries than the BH procedure, but controls the FDR for non-PRDS test statistics and is more powerful than the Benjamini and Yekutieli (2001) general-dependency procedure.

We have shown that in many cases even though the p-value distribution is not generally PRDS, it is still PRDS under H_c^0 (e.g absolute valued MVN); furthermore, there are cases in which the test statistics are not PRDS even under H_c^0, yet the Simes procedure is still valid (pairwise comparisons); we have also shown for pairwise comparisons, under quite general conditions, that if the Simes procedure rejects H_c^0 then at least one hypothesis is rejected by the ssBH procedure.

We therefore recommend to first apply the Simes procedure to test H_c^0 at level $q = 0.05$; if H_c^0 is rejected we suggest using the level 0.05 ssBH procedure to determine which null hypotheses are false; it will yield less discoveries than the level 0.05 BH procedure, however, it ensures that the FDR will not exceed 0.05.

Appendix

Proof of Proposition 2.2. We will assume that $H_1 \cdots H_{m_0}$ is the set of true null hypotheses. For each value of value of $\vec{P}$ and $i \in \{1 \cdots m_0\}$ we set $P_i = 0$ and apply the ssBH procedure. Let $s_{\max}$ denote the index of the sub-vector $\vec{P}^s$ yielding the maximal number of rejections of all the sub-vectors such that $P_i \in \vec{P}^s$, and denote the number of rejections $r_{\max} = |\vec{R}_{\mathrm{BH}}^{s_{\max}}|$.

Leaving the remaining components of $\vec{P}$ unchanged we gradually increase P_i. Notice that as long as $P_i \leqslant q \cdot r_{\max}/m$ then $r_{\max}$ hypotheses are rejected in the BH test of $\vec{P}^s$ among them H_i. However, if $P_i > q \cdot r_{\max}/m$ then H_i is no longer rejected by the ssBH procedure at all.

Benjamini and Yekutieli (2001) express the FDR of any testing procedure:

$$\mathrm{FDR} = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \Pr(k \text{ null hypotheses are rejected including } \mathrm{H}_i). \tag{4}$$

Notice that $r_{\max}$ is less than or equal to the total number of hypotheses rejected by the ssBH procedure, $r_{\max} \leqslant |\vec{R}_{\mathrm{ssBH}}|$. Furthermore, it is only determined by $\vec{P}^{(i)} = \vec{P}(P_i) - \{P_i\}$. Therefore, if we define $C_k^{(i)} = \{\vec{P}^{(i)} : r_{\max} = k\}$ we get the following upper bound for the FDR of the ssBH procedure:

$$\mathrm{FDR} \leqslant \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \Pr\{P_i \leqslant kq/m \cap C_k^{(i)}\}. \tag{5}$$

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

Define $D_k^{(i)} = \bigcup_{j \leqslant k} C_j^{(i)}$ for $k = 1 \cdots m$. $D_k^{(i)}$ is an increasing set in $\vec{P}^{(i)}$. To see this, set $P_i = 0$ and increase any of the components of $\vec{P}^{(i)}$. For each $\vec{P}^s$ such that $P_i \in \vec{P}^s$ the number of rejections will remain unchanged or decrease, thus $r_{\max}$ will either remain the same or decrease, leaving us in $D_k^{(i)}$.

For brevity, let $q_l = k \cdot q/m$. We now shall make use of the PRDS property, which states that for $p \leqslant p'$

$$\Pr(D \mid P_i = p) \leqslant \Pr(D \mid P_i = p'). \tag{6}$$

Following Lehmann (1966) it is easy to see that for $j \leqslant l$ since $q_j \leqslant q_l$:

$$\Pr(D \mid P_i \leqslant q_j) \leqslant \Pr(D \mid P_i \leqslant q_l), \tag{7}$$

for any non-decreasing set D, or equivalently,

$$\frac{\Pr(\{P_i \leqslant q_k\} \cap D_k^{(i)})}{\Pr(P_i \leqslant q_k)} \leqslant \frac{\Pr(\{P_i \leqslant q_{k+1}\} \cap D_k^{(i)})}{\Pr(P_i \leqslant q_{k+1})}. \tag{8}$$

As $D_{j+1}^{(i)} = D_j^{(i)} \cup C_{j+1}^{(i)}$ for $k = 1 \cdots m-1$, expression (8) yields

$$\begin{aligned} &\frac{\Pr(\{P_i \leqslant q_k\} \cap D_k^{(i)})}{\Pr(P_i \leqslant q_k)} + \frac{\Pr(\{P_i \leqslant q_{k+1}\} \cap C_{k+1}^{(i)})}{\Pr(P_i \leqslant q_{k+1})} \\ &\quad \leqslant \frac{\Pr(\{P_i \leqslant q_{k+1}\} \cap D_k^{(i)})}{\Pr(P_i \leqslant q_{k+1})} + \frac{\Pr(\{P_i \leqslant q_{k+1}\} \cap C_{k+1}^{(i)})}{\Pr(P_i \leqslant q_{k+1})} \\ &\quad = \frac{\Pr(\{P_i \leqslant q_{k+1}\} \cap D_{k+1}^{(i)})}{\Pr(P_i \leqslant q_{k+1})}. \end{aligned} \tag{9}$$

Staring with $C_1 = D_1$, we repeatedly use the above inequality for $k = 1 \cdots m-1$, to fold the sum on the left into a single expression,

$$\sum_{k=1}^{m} \frac{\Pr(\{P_i \leqslant q_k\} \cap C_k^{(i)})}{\Pr(P_i \leqslant q_k)} \leqslant \frac{\Pr(\{P_i \leqslant q_m\} \cap D_m^{(i)})}{\Pr(P_i \leqslant q_m)} = 1, \tag{10}$$

where the last equality follows because $D_m^{(i)}$ is the entire space.

Going back to expression (5), as $\Pr(P_i \leqslant q_k) \leqslant k \cdot q/m$,

$$\begin{aligned} \mathrm{FDR} &\leqslant \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \Pr(\{P_i \leqslant q_k\} \cap C_k^{(i)}) \\ &\leqslant \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \cdot \frac{k \cdot q}{m} \cdot \frac{\Pr(\{P_i \leqslant q_k\} \cap C_k^{(i)})}{\Pr(P_i \leqslant q_k)}. \end{aligned}$$

Finally, invoking (10) yields,

$$\mathrm{FDR} \leqslant \frac{q}{m} \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{\Pr(\{P_i \leqslant q_k\} \cap C_k^{(i)})}{\Pr(P_i \leqslant q_k)} \leqslant \frac{m_0}{m} q.$$

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B 57, 289–300.

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

Benjamini, Y., Yekutieli, D., 2005. False discovery rate-adjusted multiple confidence intervals for selected parameters. J. Amer. Statist. Assoc. 100, 71.

Benjamini, Y., Hochberg, Y., Kling, Y., 1993. False discovery rate control in pairwise comparisons. Research Paper 93-02, Department of Statistics and O.R., Tel Aviv University.

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006

Blair, C.R., Hochberg, Y., 1995. Improved bonferroni for testing overall and pairwise homogeneity hypotheses. J. Statist. Comput. Simul. 51, 281–289.

Erdman, L.W., 1946. Studies to determine if antibiosis occurs among Rhizobia: between *Rhizobium meliloti* and *Rhizobium trifoli*. J. Amer. Soc. Agron. 38, 251–258.

Hsu, J.C., 1999. Multiple Comparisons. Chapman & Hall, CRC, London, Boca Raton, FL.

Huber, W., et al., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18, S96–S104.

Kesselman, H.J., Cribbie, R., Holland, B., 1999. The pairwise multiple comparison multiplicity problem. An alternative approach to familywise and comparisonwise type I error control. Psychol. Methods 4 (1), 58–69.

Lehmann, E.L., 1966. Some concepts of dependence. Ann. Math. Statist. 37, 1137–1153.

Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19, 368–375.

Sarkar, S.K., 1998. Some probability inequalities for ordered MTP_2 random variables: a proof of Simes' conjecture. Ann. Statist. 26 (2), 494–504.

Sarkar, S.K., Chang, C.K., 1997. The Simes method for multiple hypotheses testing with positively dependent test statistics. J. Amer. Statist. Assoc. 92, 1601–1608.

Seeger, 1968. A note on a method for the analysis of significances en mass. Technometrics 10, 586–593.

Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

Tukey, J.W., 1953. The problem of multiple comparisons. Mimeographed monograph, appears in full, in: Braun, H. (Ed.), Collected work of J.W. Tukey, vol. VII. 1994, Chapman & Hall Inc., NY.

West, M., et al., 2001. Predicting the clinical status of human breast cancer using gene expression profile. Proc. Nat. Acad. Sci. 98, 11462–11467.

Williams, V.S.L., Jones, L.V., Tukey, J.W., 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. J. Educ. Behav. Statist. 24 (1), 42–69.

Yekutieli, D., 2001. Elkund-Seeger-Simes is conservative for testing all pairwise comparisons. Research Paper 99-07, Department of Statistics and O.R., Tel Aviv University.

Yekutieli, D., 2002. Theoretical results needed for applying the False Discovery Rate in statistical problems. Ph.D. Thesis, Department of Statistics and O.R., Tel Aviv University, Tel Aviv.

Please cite this article as: Yekutieli, D., False discovery rate control for non-positively regression dependent test statistics. J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.06.006