ELSEVIER

Journal of Statistical Planning and Inference 82 (1999) 171–196

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics

Daniel Yekutieli *, Yoav Benjamini

Department of Statistics and Operation Research, School of Mathematical Sciences, The Sackler Faculty of Exact Sciences, Tel Aviv University, Israel

Received 1 January 1997; accepted 1 January 1998

Abstract

A new false discovery rate controlling procedure is proposed for multiple hypotheses testing. The procedure makes use of resampling-based p-value adjustment, and is designed to cope with correlated test statistics. Some properties of the proposed procedure are investigated theoretically, and further properties are investigated using a simulation study. According to the results of the simulation study, the new procedure offers false discovery rate control and greater power. The motivation for developing this resampling-based procedure was an actual problem in meteorology, in which almost 2000 hypotheses are tested simultaneously using highly correlated test statistics. When applied to this problem the increase in power was evident. The same procedure can be used in many other large problems of multiple testing, for example multiple endpoints. The procedure is also extended to serve as a general diagnostic tool in model selection.

MSC: 62J15; 62G09; 62G10; 62H11; 62H15

Keywords: Model selection; Multiple comparisons; Meteorology; Multiple endpoints

1. Introduction

The common approach in simultaneous testing of multiple hypotheses is to construct a multiple comparison procedure (MCP) (Hochberg and Tamhane, 1987) that controls the probability of making one or more type I error – the family wise error-rate (FWE). In Benjamini and Hochberg (1995) the authors introduce another measure for the erroneous rejection of a number of true null hypotheses, namely the false discovery rate (FDR). The FDR is the expected proportion of true null hypotheses which are erroneously rejected, out of the total number of hypotheses rejected. When some of the tested hypotheses are in fact false, FDR control is less strict than FWE control, therefore

* Corresponding author. Fax: +972-3-640-9357.
E-mail address: yekutiel@math.tau.ac.il (D. Yekutieli)

0378-3758/99/$ - see front matter
PII: S0378-3758(99)00041-5

FDR-controlling MCPs are potentially more powerful. While there are situations in which FWE control is needed, in other cases FDR control is sufficient.

The correlation map is a data analytic tool used in meteorology, which is a case in point. For example, the correlation between mean January pressure and January precipitation in Israel over some 40 years, is estimated at 1977 grid points over the northern hemisphere, and drawn on a map with the aid of iso-correlation lines (Manes, 1994). On this map, correlation centers are identified, and their orientation and location are analyzed to provide synopticians with the insight needed to construct forecasting schemes. If we treat the correlation map as (partially) a problem of testing independence at 1977 locations we immediately encounter the major difficulty. No control of multiplicity at all, which is the ongoing practice, would result in many spurious correlation centers. But since the multiplicity problem is large, we should be careful about loss of power.

If such centers are identified we can bear a few erroneous ones, as long as they are a small proportion of those identified. If all we face is noise we need full protection. Thus FDR control offers the appropriate mode of protection. Moreover, using data on an even finer grid is highly disadvantageous if we take the traditional approach to multiplicity control, although it is obviously advantageous from a meteorological point of view. A finer grid increases the number of true and non-true null hypotheses approximately by the same proportion. Because the FDR is the *proportion* of true null hypotheses rejected among the rejected, FDR controlling MCPs should approximately retain their power as resolution increases.

The major problem we still face is that the test statistics are highly correlated. So far, all FDR controlling procedures were designed in the realm of independent test statistics. Most were shown to control the FDR even in cases of dependency (Benjamini et al., 1995; Benjamini and Yekutieli, 1997), but they were not designed to make use of the dependency structure in order to gain more power when possible. Here we design new FDR controlling procedures for general dependent test statistics by developing a resampling approach along the line of Westfall and Young (1993) for FWE control. This new procedure is not specific to the meteorological problem, and can be modified to tackle many problems where dependency is suspected to be high, yet of unknown structure. An important example of such a problem is multiple endpoints in clinical trials, reviewed in this issue by Wassmer et al. (1997).

Combining FDR control and resampling is not a straightforward matter. When designing resampling-based p-value adjustments, Westfall and Young relied on the fact that the probability of making any type I error depends on the distribution of the true null hypotheses only, and treating more than necessary as true is always conservative in terms of FWE. This is not generally true for FDR control: the FDR depends on the distribution of *both* the true and false null hypotheses, and failing to reject a false null hypothesis can make the FDR larger.

The approach taken here is as following: we limit ourselves to a family of "generic" MCPs, which rejects an hypothesis if its p-value is less than p. For each p we estimate the FDR of the generic MCP, the *FDR local estimators*. As a "Global" q level FDR

controlling MCP we suggest the most powerful of the generic MCPs whos FDR local estimate is less than q. The resulting MCP is adaptive, since the FDR local estimators are based on the set of observed p-values. It is also a step-up procedure, since the specific generic MCP is chosen in a step-up fashion. Unlike the stepwise resampling procedures suggested by Westfall and Young (1993) and Troendle (1995), in which the entire resampling simulation is repeated in a step-up fashion, the resampling simulation in our proposed method is performed only once on the entire set of hypotheses.

In Section 2 the framework for defining the MCPs is laid, FWE and FDR controlling MCPs, and the relationship between p-value adjustments and local estimators are discussed. In Section 3, the p-value resampling approach is reviewed. In Section 4, two types of FDR local estimators are introduced, a local estimator based on Benjamini and Hochberg's FDR controlling MCP, and two *resampling-based FDR local estimators*. In Section 5, the use of FDR local estimators for inference is presented, and the advantages of the local point of view are discussed, especially using the suggested "local estimates plot". Sections 2, 3 and the beginning of Section 5 (with its references to Section 4) suffice in order to understand the main features of the new procedure and apply it in practice. The results of applying the new MCPs to a correlation map are presented in Section 6, and the use of the p-value adjustment plot is demonstrated. In that example the new MCPs proved to be most powerful, and revealed some new patterns. A simulation study was used to show the global FDR control of the suggested MCP. It was also used to show that the newly proposed procedures are more powerful than the existing MCP. Results of the simulation study are presented in Section 7, proofs of the propositions are given in Section 8.

2. Multiple comparison procedures

The family of m hypotheses which are tested simultaneously includes m_0 true null hypotheses and m_1 false ones. For each hypothesis H_i a test statistic is available, with the corresponding p-value P_i. Denote by $\{H_{01}, \ldots, H_{0m_0}\}$ the hypotheses for which the null hypothesis is true, and by $\{H_{11}, \ldots, H_{1m_1}\}$ the false ones. The corresponding vectors of p-values are $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$. If the test statistics are continuous, $P_{0i} \sim \mathrm{U}[0,1]$. The marginal distribution of each P_{1i} is unknown, but if the tests are unbiased, it is stochastically smaller than the distribution of P_{0i}. Let R denote the number of hypotheses rejected by a MCP, V the number of true null hypotheses rejected erroneously, and S the number of false hypotheses rejected. Of these three random variables, only R can be observed. The set of observed p-values $\boldsymbol{p} = (\boldsymbol{p}_0, \boldsymbol{p}_1)$, and $r(=v+s)$, computed in a multiple hypotheses-testing problem, are *single* realizations of the random variables defined.

In terms of these definitions, the expected number of erroneous rejection, also called the per family error-rate (PFE), is $E_{\boldsymbol{P}} V(p)$. The family wise error-rate (FWE), is the probability of rejecting at least one true null hypotheses, $\mathrm{FWE} = \Pr_{\boldsymbol{P}}\{V \geqslant 1\}$ (and is always smaller than the PFE.) Let H_0^c denote the complete null hypothesis, which is

the intersection of all true null hypotheses (i.e. $m_0 = m$). A MCP offering FWE control under the complete null hypothesis is said to have *weak FWE control*. A MCP offering FWE control for any combination of true and non-true null hypotheses offers *strong FWE control*.

The false discovery rate (FDR) introduced in Benjamini and Hochberg (1995) is a different measure of the type I error rate of a MCP. It is the expected proportion of erroneously rejected hypotheses, among the rejected ones. Let Q denote

$$Q = \begin{cases} V/R & \text{if } R > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then the FDR is defined as $Q_e = E_{\boldsymbol{P}}(Q)$. As shown in Benjamini and Hochberg (1995) the FDR of a MCP is always smaller or equal to the FWE, where equality holds if all null hypotheses are true. Thus control of the FDR implies control of the FWE in the weak sense.

In the same way that an α-level test for a single hypothesis can be defined as "reject whenever the p-value $\leqslant \alpha$" so can a *generic* MCP be defined for simultaneous testing as following: For each $p \in [0,1]$, reject H_i^0 if $p_i \leqslant p$. Let $R(p)$, $V(p)$ and $S(p)$ denote the R, V and S, of the generic MCP. The PFE of such a procedure is $E_{\boldsymbol{P}} V(p) = m_0 \cdot p$. Since rejecting any true null hypotheses by a generic MCP is equivalent to rejecting the hypothesis corresponding to the minimal observed p-value of all true null hypotheses, the FWE of this procedure is $\Pr_{\boldsymbol{P}_0}\{\min_i p_{0i} \leqslant p\}$. Obviously, the larger p in a generic MCP the larger become the two error rates, yet larger is the power as well.

The Bonferroni procedure is the generic MCP with $p = \alpha/m$. For any joint distribution of the p-values, the Bonferroni procedure has FWE $\leqslant \alpha$. Consider now a generic MCP under other distributional assumptions:

Example 2.1 (*Westfall and Young*, 1993, p. 55)**.** If all test statistics corresponding to true null hypotheses are equivalent, all P_{0i}'s are identical, so regardless of m_0, $\min_i P_{0i} \sim P_{01}$. The generic MCP, with $p = \alpha$, would have FWE $= \alpha$.

Thus knowing how $\boldsymbol{P}_0$ is distributed, more powerful MCPs can be constructed. Resampling-based MCPs, introduced in Westfall and Young (1993), use p-value resampling to simulate the distribution of $\boldsymbol{P}_0$ and thereby utilize the dependency structure of the data to construct more powerful MCPs. It is our goal to follow a similar path while controlling the FDR. Alas as we shall see below, this is no trivial task, since the proportion of error Q depends not only on the falsely rejected but also on the number of correctly rejected hypotheses.

2.1. *p-value adjustments*

Instead of describing the Bonferroni procedure as a generic MCP with level α/m, we can adjust each $p_{(i)}$ by multiplying it by m. Then, a hypothesis will be rejected by the Bonferroni procedure if its adjusted p-value is less then α. Generally this is a different

way of phrasing the results of a MCP. The term p-value adjustment (adjusted p-value) is used quite loosely in the literature; see Shafer and Olkin (1983), Heyse and Rom (1988), Westfall and Young (1989), and Dunnet and Tamhane (1992). Wright (1992) presents a nice summary with many example and further references for the p-value adjustment. We have found the need to distinguish between the p-value adjustment and the p-value correction, a difference sometimes masked in single-step procedures.

Definition 2.2. For any MCP $\mathscr{M}$, the FWE p-value adjustment $p_i^{\mathscr{M}}$ of an hypothesis H_i is the size of the strictest level of the test $\mathscr{M}$ still rejecting H_i while controlling the FWE given the observed p-values, i.e.

$$\tilde{p}_i^{\mathscr{M}} = \inf\{\alpha \mid H_i \text{ is rejected at FWE} = \alpha\}.$$

The p-value adjustment can be defined for more complex MCPs as well, as the next example shows.

Example 2.3. Holm's step-down procedure makes use of the ordered p-values $p_{(1)} \leqslant \cdots \leqslant p_{(m)}$. For $i = 1, 2, \ldots, m$ the ordered p-value $p_{(i)}$ is compared to the critical value $\alpha/(m+1-i)$. The null hypotheses corresponding to $p_{(k)}$ is rejected in a α size test if for $i = 1, 2, \ldots, k$, $p_{(i)} \leqslant \alpha/(m+1-i)$. Therefore the p-value adjustment based on Holm's MCP is

$$\tilde{p}_{(k)}^{\text{Holm}} = \max_{i \leqslant k}\{p_{(i)} \cdot (m+1-i)\}. \tag{2}$$

Use of p-value adjustments for multiple-hypotheses testing has the convenience of reporting the results of a single test using p-values, because the level of the test does not have to be determined in advance. If, for example, $\tilde{p}_{(k)}^{\text{Holm}} = 0.0105$, Holm's MCP with $\alpha = 0.01$ would not have rejected $H_{(k)}^0$.

Reporting the p-value adjustments would reveal that if the FWE control is slightly relaxed $H_{(k)}^0$ can be rejected.

In a similar fashion an FDR adjustment is defined.

Definition 2.4. For any MCP $\mathscr{M}$, the *FDR p-value adjustment* of an hypothesis H_i is the size of the strictest level of the test $\mathscr{M}$ still rejecting H_i while controlling the FDR, i.e.

$$\tilde{p}_i^{\mathscr{M}} = \inf\{\alpha \mid H_i \text{ is rejected at FDR} = \alpha\}.$$

By applying the above formulation, the FDR controlling MCP introduced in Benjamini and Hochberg (1995) can be rephrased into a FDR p-value adjustment. Benjamini and Hochberg's MCP is as follows: compare each ordered p-value $p_{(i)}$ with $\alpha \cdot i/m$. Let $k = \max\{i\colon p_{(i)} \leqslant \alpha \cdot i/m\}$, if such k exists reject $H_{(1)}^0, \ldots, H_{(k)}^0$. A q sized MCP rejects an hypotheses $H_{(i)}^0$ if for some k, $k = i, \ldots, m$, $p_{(k)} \cdot m/k \leqslant \alpha$, thus we define the BH FDR p-value adjustment as

$$Q_{\text{adj}}^{\text{BH}}(p_{(i)}) = \min_{i \leqslant k}\{p_{(k)} \cdot m/k\}. \tag{3}$$

2.2. p-value corrections

The FWE correction of a p-value is the adjustment of the generic MCP at this value. Thus the FWE correction of p is defined,

$$\tilde{p} = \Pr_{\boldsymbol{P}_0}\{\min \boldsymbol{P}_0 \leqslant p\}. \tag{4}$$

The FWE correction $\tilde{p}$ can be defined for any $p \in [0,1]$. The FWE correction of 0 is 0, that of 1 is 1, and in between $\tilde{p}$ is increasing in p. Heyse and Rom (1988) refer to such a correction at the observed minimal p-value as the "adjustment". The generic MCP with critical p-value p such that $\tilde{p} = \alpha$, has FWE α by definition. Usually the distribution of $\boldsymbol{P}_0$ is unknown, so the FWE correction cannot be computed, but must be estimated. Let $\tilde{p}_{\text{est}}(\boldsymbol{P})$ be a estimate of the FWE correction, such that for each p and for all $\boldsymbol{P}$, $\tilde{p}_{\text{est}}(\boldsymbol{P}) \geqslant \tilde{p}$. Let $\hat{p}_\alpha = \sup\{p\colon \tilde{p}_{\text{est}} \leqslant \alpha\}$. The MCP based on $\tilde{p}_{\text{est}}$ is: reject H_i^0 if $p_i \leqslant \hat{p}_\alpha$. This MCP is more conservative than the generic MCP with FWE $= \alpha$, thus offers FWE control.

Again, the concept of p-value correction can be extended to FDR control. Let

$$Q(p) = \begin{cases} V(p)/R(p) & \text{if } R(p) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The FDR correction of $p, Q_e(p)$ is defined:

$$Q_e(p) = E_{\boldsymbol{P}} Q(p). \tag{5}$$

For a given p-value p, $Q_e(p)$ is a function of $(S(p), V(p))$. Unlike the FWE p-value correction $Q_e(p)$ is not necessarily increasing in p, although it always satisfies $Q_e(0) = 0$ and $Q_e(1) = m_0/m$. Under the complete null hypothesis $Q_e(p)$ equals the FWE correction $\tilde{p}$.

Using Definition 2.2 the MCPs which are defined in terms of critical values, can be rephrased into a p-value adjustment. As shown, if conservative estimators of the p-value correction are available, a new FWE controlling MCP can be defined based on the set of estimates. Westfall and Young's single-step MCP, which is called a "resampling-based p-value adjustment" is an example of such a procedure. This scheme is discussed in the next section, where it is shown that Westfall and Young's p-value adjustment is a conservative estimator of the FWE correction, and therefore their MCP controls the FWE.

Following the example of Westfall and Young we wish to introduce the resampling approach to the new problem of FDR control. We present a new class of MCPs based on conservative estimators of the FDR correction called FDR local estimators. Two resampling based local estimators are presented, which are conservative and retain good power when true null hypotheses are highly correlated. These two are compared to a local estimator based on the BH MCP which is easy to compute but can be downward biased if $s(p) = 0$ and overly conservative if the test statistics corresponding to the set of true null hypotheses are highly correlated.

3. *p*-value resampling

The construction of powerful MCPs requires knowledge of the distribution of $\boldsymbol{P}_0$. p-value resampling is a method to approximate the distribution of $\boldsymbol{P}_0$ using data gathered in a single experiment. Since the number and identity of the true null hypotheses is not known, p-value resampling is conducted under the complete null hypothesis, i.e. under the assumption that all the hypotheses are in fact true.

The basic setup for p-value resampling: Let $\mathscr{Y}$ denote a data set gathered to test an ensemble of hypotheses.

1. Compute $\boldsymbol{p} = \boldsymbol{P}(\mathscr{Y})$.
2. Model the data according to the complete null hypothesis, with a systematic and a random component, $\mathscr{Y} = \mathscr{Y}(\mathscr{C}_{\mathscr{Y}}, \varepsilon_{\mathscr{Y}})$.
3. Estimate the distribution of the random component $\varepsilon_{\mathscr{Y}}$.
4. Simulate the random component $\varepsilon_{\mathscr{Y}}$, by drawing sample $\varepsilon^*_{\mathscr{Y}}$, with replacement, from the empirical distribution estimated in step 3.
5. Construct a simulated data set $\mathscr{Y}^* = \mathscr{Y}(\mathscr{C}_{\mathscr{Y}}, \varepsilon^*_{\mathscr{Y}})$.
6. Compute $\boldsymbol{p}^* = \boldsymbol{P}(\mathscr{Y}^*)$.
7. Repeat steps 4–6 to get large samples from the simulated distribution $\boldsymbol{P}^*$.

The resampling procedure makes use of these simulated sets of p-values. For the sake of the theoretical discussion we shall further denote the two components of $\boldsymbol{P}^* = (\boldsymbol{P}^*_0, \boldsymbol{P}^*_1)$, according to the subsets of true and non-true null hypotheses. Since resampling is conducted under the complete null hypotheses the distribution of the resample-based p-value vector corresponding to non-true null hypotheses $\boldsymbol{P}^*_1$ is different from its *real* distribution $\boldsymbol{P}_1$. The marginal distribution of all P^*_i is $\mathrm{U}[0,1]$ as is the marginal distribution of all P_{0i}. The property we seek is that the *joint distributions* of $\boldsymbol{P}^*_0$ and $\boldsymbol{P}_0$ are identical. This property called *subset pivotality* is achieved if the distribution of $\boldsymbol{P}_0$ is unaffected by the truth or falsehood of the remaining null hypotheses. Like the assumption of independence it is mostly a property of the design, and is difficult to assess from the data at hand. The formal definition is given in Westfall and Young (1993, p. 42).

Definition 3.1. The distribution of $\boldsymbol{P}$ has the *subset pivotality* property if the joint distribution of the sub-vector $\{P_i;\ i \in K\}$ is identical under the restriction $\bigcap_{i\in K} \mathrm{H}_{0i}$ and $\mathrm{H}^{\mathrm{C}}_0$, for all subsets K of true null hypotheses.

Note that resampling involves sampling empirical distributions, thus subset pivotality can only be achieved asymptotically as sample size of the original data approaches infinity; nevertheless throughout this work we *assume* subset pivotality exists.

In order to derive the theoretical FDR properties of the resampling MCP we shall make use of the additional condition that $S(p)$ and $V(p)$ are independent, which in turn follows if $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ are assumed to be independent for any combination of true and false hypotheses. The conditions of subset pivotality and independence are different:

Usually independence of $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ implies subset pivotality, while the correlation map below is an opposite example where subset pivotality holds yet independence does not. It should be emphasized that the independence condition does not require the independence within $\boldsymbol{P}_0$ or $\boldsymbol{P}_1$.

For each p, denote, $V_0^*(p)=\#\{i \mid P_{0i}^* \leqslant p\}$, $V_1^*(p)=\#\{i \mid P_{1i}^* \leqslant p\}$. Since the identity of the true null hypotheses is not known the only observable variable is $R^*(p) = V_0^*(p)+V_1^*(p)$. The resample-based FWE single-step adjustment (Westfall and Young, 1993) can be viewed as an estimator of the FWE correction, computed by substituting $\boldsymbol{P}_0$ in Definition 2.4 by $\boldsymbol{P}^*$.

$$\tilde{p}^{\mathrm{WF}} = \Pr_{\boldsymbol{P}^*}\{\min \boldsymbol{P}^* \leqslant p\}. \tag{6}$$

Assuming subset pivotality, the resampling-based FWE p-value adjustment defined in Definition 2.4 is greater than the FWE p-value correction:

$$\Pr\{\min \boldsymbol{P}_0 \leqslant p\} \leqslant \Pr\{\min \boldsymbol{P}_0^* \leqslant p \vee \min \boldsymbol{P}_1^* \leqslant p\} = \Pr\{\min \boldsymbol{P}^* \leqslant p\}.$$

4. FDR local estimators

FDR local estimators are estimators of the FDR correction. The first is based on the FDR controlling MCP in Benjamini and Hochberg (1995). The BH FDR local estimator is defined as

$$Q_{\mathrm{est}}^{\mathrm{BH}}(p) = \begin{cases} m \cdot p/r(p) & \text{if } r(p) \geqslant 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Let $R^{-1}(p)$ denote the reciprocal of $R(p)$, taking the value 0 if $R(p)$ is 0. The expected value of the BH local estimator is then $m \cdot p \cdot ER^{-1}(p)$ hence greater than $EV(p) \cdot ER^{-1}(p)$, the FDR correction is $E(V(p) \cdot R^{-1}(p))$ therefore the bias of the BH local estimator as an estimator of the FDR correction is greater than $-\mathrm{Cov}(V(p), R^{-1}(p))$. In general, $V(p)$ and $R(p)$ are positively correlated, making $V(p)$ and $R^{-1}(p)$ negatively correlated and $Q_{\mathrm{est}}^{\mathrm{BH}}(p)$ an upward biased estimator of $Q_{\mathrm{e}}(p)$. But note that if $R(p)$ equals 0 both $R^{-1}(p)$ and $V(p)$ equal their minimal value 0, thus if $R(p)$ is stochastically small (for example extremely small p), $R^{-1}(p)$ and $V(p)$ might become positively correlated.

Let us now examine the BH FDR local estimator when the test statistics corresponding to the true null hypotheses are highly correlated. The most extreme case is the design of Example 2.1, the distribution of $V(p)$ and $S(p)$ are independent and all true null hypotheses are equivalent, furthermore let us assume that $S(p)$ has a single value.

Example 4.1.

$$V(p) = \begin{cases} 0 & \text{with probability } 1-p \\ m_0 & \text{with probability } p \end{cases} \quad \text{and} \quad S(p) \equiv s(p).$$

If $s(p)=0$, $Q_e(p)=p$, while the BH FDR estimator

$$Q_{est}^{BH}(p)=mp^2/m_0=p^2(1+m_0/m_1).$$

If $s(p)>0$, $Q_e(p)=m_0p/(m_0+s)$, and

$$Q_{est}^{BH}(p)=(1-p)\frac{mp}{s}+p\frac{mp}{m_0+s}\geqslant\frac{mp}{m_0+s}.$$

And moreover for small p, $Q_{est}^{BH}(p)\approx mp/s$.

While total dependence is seldom encountered, it is apparent that if $\boldsymbol{P}_0$ is highly intercorrelated, $Q_{est}^{BH}(p)$ might become either downward biased or grossly upward biased, depending on an unobservable random variable $S(p)$.

4.1. Resampling-based FDR local estimators

Using the resampling approach for FDR estimates introduces a new complexity since the estimated correction directly depends on the distribution of $\boldsymbol{P}_1$ through $S(p)$. To solve the problem the *conditional* FDR correction given $S(p)=s(p)$, is estimated instead of the FDR correction. The conditional FDR p-value correction denoted $Q_{V|s}(p)$, is defined as

$$Q_{V|s}(p)=E_{V(p)|S(p)=s(p)}Q(p). \tag{8}$$

Resampling-based estimators mimic expression (8), the expectation is computed using the resample-based distribution R^* instead of the $V|s$, and $s(p)$, the conditioned upon realization of $S(p)$, is replaced by an estimate $\hat{s}(p)$ based on $r(p)$,

$$\hat{Q}_V=E_{R^*}\frac{R^*}{R^*+\hat{s}}.$$

Independence of $V(p)$ and $S(p)$ is needed because p-value resampling approximates the marginal distribution of $V(p)$, but not the conditional distribution of $V(p)$ given $S(p)=s(p)$. Under independence of the two the conditional and marginal distributions of $V(p)$ are identical.

Two resampling-based estimators are introduced differing in their treatment of $s(p)$: point estimator and an upper limit. Obviously, in order to be conservative, we seek a downward biased estimator of $s(p)$.

The first estimator is $r(p)-mp$, which is obviously downward biased given $S(p)=s(p)$,

$$E\{r(p)-mp\}\leqslant s(p)+Ev(p)-m_0p=s(p).$$

We further need $r_\beta^*(p)$, the $1-\beta$ quantile of $R^*(p)$ (we use $\beta=0.05$ in the simulations and in the correlation map example). Using the estimator of $s(p)$ we define the resampling-based FDR local estimator,

$$Q^*(p)=\begin{cases}E_{R^*}\frac{R^*(p)}{R^*(p)+r(p)-p\cdot m} & \text{if } r(p)-r_\beta^*(p)\geqslant p\cdot m,\\ \Pr_{R^*}\{R^*(p)\geqslant 1\} & \text{otherwise.}\end{cases} \tag{9}$$

The second estimator is $r(p) - r^*_\beta(p)$, assuming subset pivotality conditioning on $S(p) = s(p)$:

$$\Pr\{r(p) - r^*_\beta(p) \leqslant s(p)\} = \Pr\{V(p) \leqslant r^*_\beta(p)\} \leqslant \Pr\{R^*(p) \leqslant r^*_\beta(p)\} \geqslant 1 - \beta.$$

The resampling-based $1 - \beta$ FDR upper limit is defined as

$$Q^*_\beta(p) = \sup_{x \in [0,p]} \begin{cases} E_{R^*} \frac{R^*(x)}{R^*(x) + r(x) - r^*_\beta(x)} & \text{if } r(x) - r^*_\beta(x) > 0, \\ \Pr_{R^*}\{R^*(x) \geqslant 1\} & \text{otherwise.} \end{cases} \tag{10}$$

The condition in the definition of each of the FDR estimators is aimed towards the complete null hypotheses, in which $S \equiv 0$ and both resample-based FDR estimators *should* equal the resample-based single-step FWE p-value adjustment.

For small β the $1 - \beta$ quantile of the distribution of R^*, $r^*_\beta(p)$, is generally greater than its expectation $m \cdot p$, so the resampling-based upper limit usually exceeds the point estimator.

The following propositions discuss some of the properties of these estimators and corrections. Throughout this discussion it is assumed that the distributions of $S(p)$ and $V(p)$ are independent, and subset pivotality exists. Proofs of the propositions are given in the appendix.

Proposition 4.2. *If the distribution of $V(p)$ and $S(p)$ are independent, then conditioning on $S(p) = s(p)$, $Q^*_\beta(p)$ exceeds $Q_{V|_s}(p)$ with probability $1 - \beta$.*

Since $Q^*_\beta(p)$ is positive and $V(p)$ and $S(p)$ are independent, for all values of $s(p)$, $E_{V|_s} Q^*_\beta(p) \geqslant (1 - \beta) \cdot Q_{V|_s}(p)$, and therefore,

$$E_{V,S} Q^*_\beta(p) \geqslant (1 - \beta) \cdot Q_e(p).$$

Proposition 4.3. *If the distribution of $V(p)$ and $S(p)$ are independent and $s(p)$ satisfies, $s(p) \geqslant p \cdot m$, then $E_{V(p)|S(p)=s(p)} Q^*(p) \geqslant Q_{V|_s}(p)$.*

Note that the condition of the proposition is on $s(p)$, an unobservable random variable. If the condition is not met, i.e. $s(p) < m \cdot p$, then

Proposition 4.4. *If the distribution of $S(p)$ and $V(p)$ are independent, and $s(p) < p \cdot m$, then conditioning on $S(p) = s(p)$, $Q^*(p)$ exceeds $Q_{V|_s}(p)$ with probability $1 - \beta$.*

As was shown for Q^*_β, for all values of $s(p)$, $E_V Q^*(p) \geqslant (1 - \beta) \cdot Q_{V|_s}$, thus

$$E_{V,S} Q^*(p) \geqslant (1 - \beta) \cdot Q_e(p).$$

Proposition 4.5. *Under the complete null hypothesis,*

$$Q^*(p),\ Q^*_\beta(p) \begin{cases} = Q_e(p) & \text{with probability } \geqslant 1 - \beta, \\ < Q_e(p) & \text{with probability } < \beta. \end{cases}$$

p-value resampling is executed assuming *all* the hypotheses are true null hypotheses, thus the greater the proportion of false null hypotheses the more $R^*(p)$ will exceed $V(p)$. As a result, as m_1 increases the resample-based estimates increase relative to the conditional correction. It seems that the most favorable situation for a given m_0 and m_1, is if the P_{1i}'s are highly correlated and least favorable if the P_{1i}'s are independent, as the following example shows.

Assuming total dependence of all P_{0i}'s and P_{1i}'s respectively, while $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ are independent, then the distribution of $R^*(p)$ is,

$$R^*(p) = \begin{cases} 0 & \text{with probability } (1-p)^2 \\ m_0 & \text{with probability } p(1-p) \\ m_1 & \text{with probability } p(1-p) \\ m & \text{with probability } p^2 \end{cases}$$

Assuming $\hat{s}(p) \equiv s(p)$,

$$\hat{Q}_V = p(1-p)\frac{m_0}{m_0+s} + p(1-p)\frac{m_1}{m_1+s} + p^2\frac{m}{m+s} \approx p\left(\frac{m_0}{m_0+s} + \frac{m_1}{m_1+s}\right).$$

Recall that the conditional correction in this case is $m_0 p/(m_0+s)$ and the BH estimate is mp/s.

Resampling-based FDR estimators are conservative estimators of the conditional FDR correction. Their upward bias decreases as the proportion of true null hypotheses increases, and as the distribution of $\boldsymbol{P}_0$ is more positively correlated. Under the complete null hypotheses, they equal the FDR correction with probability $1-\beta$.

5. Use of local estimates in multiple-hypotheses testing

With an eye towards the user we now outline the multiple testing procedure.

1. Construct a p-value resampling scheme as described in Section 3.
2. Choose the set of p-values for inquiry. If the purpose is testing, it is enough to consider the set of observed p-values $\boldsymbol{p}$. Drawing the FDR local estimates plot, described at the end of this section, might require computing the FDR local estimators on a grid of p-values.
3. For each p-value p, in the set of p-values specified in step 2, compute the resample based distribution $R^*(p)$, using the vectors of resample based p-values $\boldsymbol{P}^*$, generated by the resampling scheme.
4. Find $r^*_\beta(p)$, the $1-\beta$ quantile of $R^*(p)$
5. Using the distribution approximated in step 3, compute either resample based local estimators, which are the resample means of expressions (9) or (10).
6. Let $\hat{Q}$ denote the FDR local estimator computed. Find,

$$k_q = \max_k\{\hat{Q}(p_{(k)}) \leqslant q\},$$

the size q MCP based on the FDR local estimator is: reject $\mathrm{H}^0_{(1)}, \ldots, \mathrm{H}^0_{(k_q)}$.

If the local estimator is computed for the purpose of drawing a local estimates plot, simply plot it alongside other FDR local estimators.

Let us now detail the difficulties, and the properties of the proposed procedure. Recall that FDR local estimators are functions of $r(p)$, conditioning on $S(p) = s(p)$ FDR local estimators are functions of $v(p)$. If the experiment is repeated the cutoff p-value for rejection would be different. The FDR of this MCP is $E_{\boldsymbol{P}} Q(p_{(k_q)}(\boldsymbol{P}))$. Note that MCPs based on FDR local estimators are defined in the same manner as MCPs based on p-value adjustments are defined, but because unlike MCPs based on p-value adjustments the resample based MCPs are not equivalent to FDR controlling MCPs they do not necessarily offer FDR control. Let p_q satisfy, $p_q = \sup\{p \mid Q_{V|_s}(p) \leqslant q\}$.

Recall that $Q_{V|_s}(p)$ is a function of $\boldsymbol{P}_1$, thus p_q is a function of $\boldsymbol{P}_1$. The FDR of this MCP is

$$E_{\boldsymbol{P}} Q(p_q(\boldsymbol{P}_1)) = E_{\boldsymbol{P}_1} E_{\boldsymbol{P}_0|\boldsymbol{P}_1} Q(p_q(\boldsymbol{P}_1)) = E_{\boldsymbol{P}_1} Q_{V|_s}(p_q) \leqslant E_{\boldsymbol{P}_1} q \leqslant q.$$

As discussed in Section 3 given a conservative FWE local estimator, the MCP based on the local estimator controls the FWE. This cannot be applied to FDR control because the FDR local estimators are not uniformally conservative but are conservative in expectation or in probability. Thus, the resulting q sized MCP is not necessarily more conservative than the q sized generic MCP. The following example shows that a MCP can be more conservative than a generic MCP yet have greater FDR. This is possible because, unlike $V(p)$, $Q(p)$ is not increasing in p.

Example 5.1. Two hypotheses are tested, a true null and a false null hypotheses. For each $p \in [0, 1]$ the p-value correction is

$$Q_e(p) = 1 \cdot \Pr\{P_0 \leqslant p, P_1 > p\} + 0.5 \cdot \Pr\{P_0 \leqslant p,\ P_1 \leqslant p\}.$$

Let $\hat{Q}$ denote the FDR local estimator:

$$\hat{Q}(p) = \begin{cases} Q_{\mathrm{e}}(p), & p \leqslant p_0, \\ 1 & p > p_0. \end{cases}$$

Notice that $\hat{Q}(p)$ is a conservative local estimator. The FDR of the generic MCP is by definition $Q_{\mathrm{e}}(p)$, the MCP based on $\hat{Q}$ is equivalent to the generic MCP with one exception: if $p_0 < p_1 \leqslant p$ the generic MCP rejects both null hypotheses, but the $\hat{Q}$ MCP will only reject the true null hypothesis. Hence the FDR of this MCP is

$$Q_{\mathrm{e}}(p) + 0.5 \cdot \Pr\{P_0 \leqslant P_1 \leqslant p\}.$$

Note that in the example a distinction is made between the true and false null hypotheses, and the MCP is designed to produce the maximal FDR. Though $Q(p)$ is not increasing in many examples its expectation $Q_{\mathrm{e}}(p)$ is increasing. Thus in general a more conservative MCP will have less FDR.

According to the following proposition the MCP based on the upper limit FDR estimator is with probability $1 - \beta$ more conservative than the MCP based on the FDR conditional correction.

Proposition 5.2. *Denote,* $p_q^{\mathrm{ul}} = \sup\{p \mid Q_\beta^*(p) \leqslant q\}$, $p_q = \sup\{p \mid Q_{V|_S}(p) \leqslant q\}$. *If the distribution of* $S(p)$ *and* $V(p)$ *are independent then* $\Pr\{p_q^{\mathrm{ul}} > p_q\} \leqslant \beta$.

In the simulation study it is shown that the MCP based on the BH local estimator controls the FDR, and in situations other than the complete null hypothesis, the MCPs based on either of the resampling-based FDR local estimators offer FDR control. Under the complete null hypothesis the FDR slightly exceed the required level. This is consistent with Proposition 4.5, which states that under the complete null hypotheses, $Q_\beta^*(p)$ and $Q^*(p)$ equal $Q_e(p)$ with probability $\geqslant 1-\beta$, but their expected value is less than the FDR correction, and note also that:

Proposition 5.3. *Under the complete null hypothesis the MCP based on* $Q_\beta^*(p)$ *offers* $\beta + q$ *FWE control.*

The problem with MCPs based on FDR local estimates is a *selection bias*. Note that the criterion for choosing the cutoff point is similar to finding the minima of the FDR estimates. Local FDR control is shown for any given p-value, but given the selection process the local FDR estimator at the rejection p-value is downward biased, and may not retain the property of local FDR control. The MCP which uses the resample-based FDR upper limit is less affected by this bias, because its critical value is with probability $1-\beta$ less than the critical value of the MCP based on the conditional FDR correction. Because of this we should not discard procedurewise FDR control as the legitimate goal for an MCP.

5.1. Local estimates plot

The ability to compute conservative estimates of the FDR correction, shifts the emphasis from the properties of the MCP to the property of a specific cutoff decision. This allows the inspection of the implication of the choice of critical value to be made on the error committed. This is best accomplished for all potential decision using the local estimates plot.

The local estimates plot is a multivariate plot of both FDR and FWE local estimates, which presents a complete picture of the expected type-1 error for each value of p. Thus a testing procedure suited to the needs and limitations of the practitioner can then be constructed. An example of use of the local estimates plot is made in the correlation map example where the plot will be described in detail.

6. Applying MCPs to correlation maps

6.1. The problem

The Israeli Meteorological Service has routinely issued seasonal forecasts of precipitation since 1983. These forecasts were constructed by the Seasonal Forecast Research

Group, see Manes et al. (1994). The successful forecasting effort had always involved modeling the association between anomalies in the pressure field over the northern hemisphere and precipitation in Israel.

Within the ongoing forecasting effort, models and methods are ever changing. Recently interest grew in the forecasting of precipitation in individual months. The study reported here is a part of this effort.

The data consist of mean January pressure measured in 1977 points in the northern hemisphere over 39 years from 1950 until 1988, and January precipitation in Israel during that period the correlation is evaluated between each of the 1977 pressure vectors (geopotential height at 500 mb.) and the square root of the precipitation vector. The set of geographic laid correlation coefficients is referred to as the "correlation map". Previously, the analysis of this map involved only their graphical inspection with the aid of isocorrelation lines, and definition of "correlation centers". The configuration, magnitude, and orientation, of the correlation centers, provide synopticians with the insight needed to construct forecasting schemes. It is suspected that some of the structure on the correlation map is the result of noise. We can try to identify the true signal through testing. For each point i, the Pearson correlation coefficient, r_i, is a statistic to test:

- H_0^i: $Z500$ at point i is uncorrelated to the precipitation.
- H_1^i: $Z500$ at point i is correlated to the precipitation.

Testing 1977 hypotheses simultaneously produces a serious multiple hypotheses testing problem. Ignoring this problem and conducting each test at level 0.05 would produce approximately 100 locations with apparent affect on the precipitation in Israel even if no such relationship exists.

MCPs based on four types of p-value adjustments were used: Westfall and Young's single-step p-value adjustment (WY), BH FDR p-value adjustment (BH), the resampling-based FDR point estimator (RES) and the resampling-based FDR upper limit (UP-RES). The four MCPs were applied twice, with significance levels 0.05 and 0.10. Further analysis was conducted using the newly suggested local estimates plot.

6.1.1. Resampling scheme

Resampling was conducted under the complete null hypothesis, i.e., pressure field is uncorrelated to precipitation in Israel. The pressure field was kept constant over the resampling, the precipitation vector $\boldsymbol{y}^*_{\text{prec}}$ was sampled with replacement from the original precipitation vector. The set of p-values $\boldsymbol{p}^*$ corresponding to the 1977 correlation coefficients between each of the 1977 pressure vectors and $\boldsymbol{y}^*_{\text{prec}}$ is a realization of $\boldsymbol{P}^*$.

The analysis is done while conditioning on the entire pressure matrix. Although conditional inference can yield largely different results than the unconditional one (see Benjamini and Fuchs, 1990), we settled for the conditional inference since the computational effort is considerably smaller. According to a simulation conducted to examine the validity of the conditional inference in this case, conditional inference is similar to the unconditional inference.

6.2. Results

Fig. 1 is the local estimates plot of the 60 most significant p-values. The critical value of r and the number of rejections are listed in Table 1.

All MCPs discover that pressure in grid points near Israel affect precipitation in Israel. Of the 0.05 MCPs only the RES MCP points at an additional correlation region located near Hawaii. Of the 0.1 MCPs the RES, UP-RES and BH MCPs discover the correlation region above Hawaii, but the RES MCP identifies yet an additional correlation center located in northern Italy, as can be seen in Fig. 2 (see Yekutieli (1996) for more details on the example).

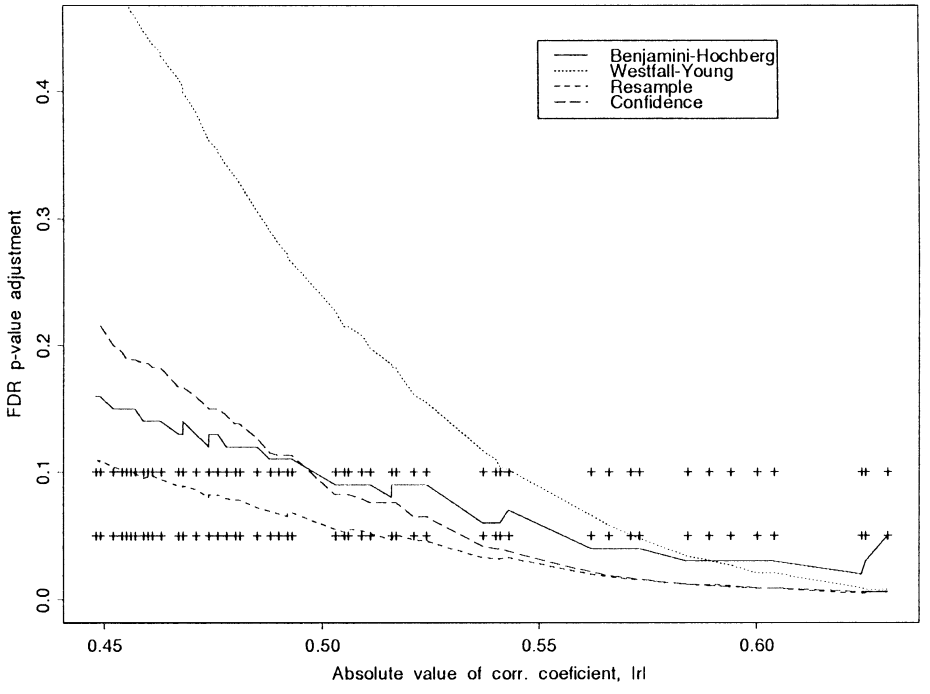

Fig. 1. The X-coordinate is the absolute value of the correlation coefficient. Ordinates are the four p-value estimates, Westfall–Young's ($\tilde{p}^{WF}$), resampling-based estimate (Q^*), resampling-based 1 – 0.05 upper limit ($Q^*_{0.05}$) and Benjamini–Hochberg's FDR estimate (Q^{BH}_{est}). The + signs are ($x = |r_{(k)}|$, $y = 0.05 \vee 0.1$).

Table 1

MCP	0.05 level		0.1 level	
	r	#	r	#
WY	0.573	9	0.562	12
BH	0.562	12	0.502	28
UP-RES	0.537	16	0.503	28
RES	0.516	22	0.457	53

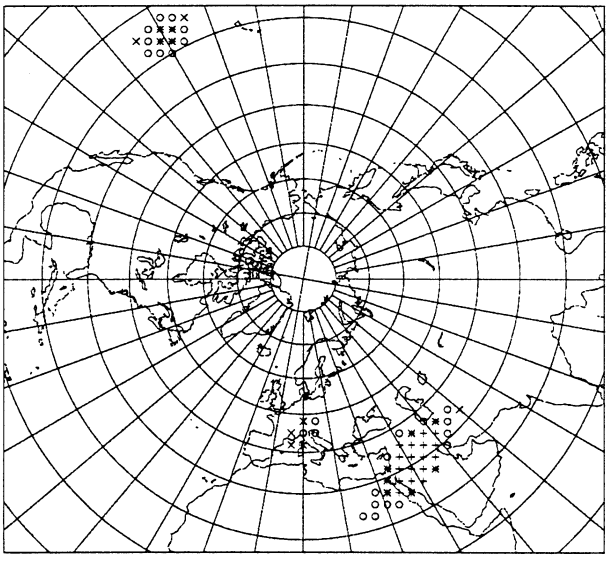

Fig. 2. Location of 60 grid point with maximal $|r|$: plus sign: $|r| \geqslant 0.524$; star: $0.524 > |r| \geqslant 0.5$; circle: $0.5 > |r| \geqslant 0.457$; cross: $0.457 > |r| \geqslant 0.448$.

The correlation map can also be effectively analyzed using the local estimates plot.

The local estimate plot is a scatter plot of the type-I local estimates versus the p-value (or an increasing function of the p-value). In this example, the X-axis of the plot is the absolute value of the correlation coefficient, $|r|$. The MCP based on the FDR local estimate can be constructed by adding the line, $y = q$. The maximal p at which this line crosses the FDR local estimate is $\hat{p}_q$. But the best use of the local estimates plot is as a graphical diagnostics tool.

All local estimates (either FDR or FWE estimates) are on a single scale, the FWE or FDR of the generic MCP. This enables comparison between local estimates for different values of p, or computed using different techniques of estimation. Selection bias a problem of FDR local estimators can be overcome by comparing the resample-based FDR local estimate and upper limit.

The local estimates plot allows the practitioner to decide which hypotheses to reject on a scale relevant to the correlation map setting, in this case the absolute value of the correlation coefficient, while warning him of errors in terms of FDR and FWE. If for example the practitioner decides to reject any hypotheses with $|r|$ greater than 0.5, according to Fig. 1, 28 hypotheses would be rejected, the error in terms of FDR would be approximately 0.05 (RES at 0.5) and at most 0.08 (UP-RES at 0.5), the error in terms of FWE 0.22 (WY at 0.5).

Back to the original purpose it makes sense to combine the pressures in each of the clusters to a single variable, resulting in 2–3 potentially useful variables to join other available forecasting variables in developing the forecasting model.

7. Simulation study

In the previous sections we showed that FDR local estimates are conservative estimators of the FDR correction if the vectors of p-value corresponding to true and non-true null hypotheses are independent. In order to show that the MCPs based on the FDR local estimates control the FDR, we revert to a simulation study.

The performance of the three MCPs is compared in terms of FDR control and power. They are also compared to a fourth MCP based on the FDR p-value correction Q_e (REAL), which can be computed in a simulation study, but obviously cannot be computed when facing a real problem.

The simulation study comprises of two sets of simulations, one in which the ratios of true to false hypotheses varied, and a second set of simulations focusing on the complete null hypothesis.

The performance of the procedures was studied in the following normal shift problem, $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, test $\mathrm{H}_{0i} : \mu_i = 0$ vs. $\mathrm{H}_{1i} : \mu_i < 0$. Each test statistic of H_j is the mean of $n=40$ observations $\bar{y}_{\cdot j}$, and so the p-value is, $p_i = \Pr_{\bar{Y} \sim \mathrm{N}(0,1/n)}(\bar{\mathbf{Y}} \leqslant \bar{y}_{\cdot j})$. The number of hypotheses m was fixed at 40, while five different values of m_0 were used: 0, 20, 30, 35, and 40. The $m_1 = 40 - m_0$ false hypotheses are set at $\mu_j = -(d + j/m_1)/\sqrt{n}$, where $1 \leqslant j \leqslant m_1$. The shift d parameter controls the distance between the true and non-true null hypotheses, $d = 0, 1, 2$. For all i $\mathrm{Var}\, Y_i = 1$. $\mathrm{Cov}(Y_i, Y_j) = 0$ for $1 \leqslant i \leqslant m_0 < j \leqslant m$. $\mathrm{Cov}(Y_i, Y_j) = 0.5$ for $m_0 < i < j \leqslant m$. $\mathrm{Cov}(Y_i, Y_j) = \rho_0$ for $1 \leqslant i < j \leqslant m_0$. Three values of ρ_0 are used: 0, 0.5, 0.941.

Each sample is followed by resamplings and a computational tradeoff has to be made between the number of samples drawn for the simulation and the number of resamplings from each sample.

In the first set of simulations where the proportion of false hypotheses varied, the number of samplings was 200. In the second set of simulations, each simulation consisted of a 1000 samplings. The increased precision, resulting from the larger sample size in the second set of simulations, was needed as estimated FDR was close to the desired level.

Resampling scheme is done as following, let $\boldsymbol{i}^* = \{i_1^*, \ldots, i_m^*\}$ denote a sample with replacement from $\{1, \ldots, m\}$. Compute,

$$\bar{y}^*_{\cdot j} = \frac{\sum_{k=1}^{n} y_{i_k^*, j}}{n}, \qquad s^*_{\cdot j} = \sqrt{\frac{\sum_{k=1}^{n} (y_{i_k^*, j} - \bar{y}^*_{\cdot, j})^2}{n-1}}.$$

The statistic is, $t_j^* = (\bar{y}^*_{\cdot j} - \bar{y}_{\cdot, j})/s^*_{\cdot j}/\sqrt{n}$, the p-value, $p_j^* = \Pr_{\boldsymbol{T} \sim t_{n-1}}(\boldsymbol{T} \leqslant t_j^*)$ suggested by Westfall and Young (1993). The number of resamplings in the first set of simulations was varied according to ρ_0: for $\rho_0 = 0$ 400 resamplings, for $\rho_0 = 0.5$, 600 resamplings, and for $\rho_0 = 0.941$, 800 resamplings. In the second set of simulations the number of resamplings was 1000.

Finally in order to apply the REAL MCP, Q_e, has to be computed. A set of 4000 samplings was conducted in order to estimate $Q_e(p)$.

7.1. Computations

The four MCPs were applied at two levels, 0.05 and 0.1. For each sample, each MCP $\mathscr{M}$ and two levels of FDR control, $s_{\mathscr{M}}$ and $v_{\mathscr{M}}$ were computed. The power of an MCP can be described by the simulation mean of $s_{\mathscr{M}}$. The FDR $Q_{\mathscr{M},q}$, of an MCP is the simulation mean of

$$q_{\mathscr{M}} = \begin{cases} \dfrac{v_{\mathscr{M}}}{v_{\mathscr{M}} + s_{\mathscr{M}}} & \text{if } v_{\mathscr{M}} \geqslant 1, \\ 0 & \text{if } v_{\mathscr{M}} = 0. \end{cases}$$

In addition the standard error of both the power and FDR are computed. FDR control can also be shown if the FDR of an MCP is less than the FDR of the REAL MCP. For that purpose the standard error of the difference in FDR is computed.

7.2. FDR control

The primary goal of the simulation study was to determine whether the suggested MCPs offer FDR control.

Fig. 3 is a graphical presentation of the simulation-based FDR values of the four MCPs (at level $q = 0.05$).

In all the plots Q is approximately q. As the percentage of null hypotheses increases FDR values of the BH RES and UP-RES MCPs approach q. For 87.5% and 100% true null hypotheses (rows 3 & 4), C and B exceed q, but by less than a standard error.

The second set of simulations, mentioned before, was conducted to investigate FDR control under the complete null hypothesis. It consisted of three simulations, all under the complete null hypothesis. In each simulation sampling and resampling number was set to 1000. ρ_0 was set to 0, 0.5 and 0.941. Fig. 4 is a graphical summary of the results. From Fig. 4 it seems that under the complete null hypotheses the FDR of the RES MCP exceeds q. In the $q = 0.05$ MCP the FDR of the RES MCP is ≈ 0.06. In the $q = 0.1$ MCP for $\rho_0 = 0$ the FDR is 0.12, but for $\rho_0 = 0.5$ and 0.941 the FDR is slightly less than 0.1. When compared to the REAL MCP, the FDR of the RES MCP exceeds the FDR of the REAL MCP three out of six times by ≈ 0.01. The FDR of the UP-RES MCP exceeds q four out of six times but seem to be less than the FDR of the REAL MCP. We suspect that if sample size was still larger it might have been discovered that the FDR of the UP-RES MCP also exceeds q.

All three MCPs seem to control the FDR when true null hypotheses percentage is less than 100%. Under the complete null hypothesis, the FDR of the RES MCP seems to exceed q by 0.01, the FDR of the UP-RES MCP might also be greater than q, this is also consistent with Proposition 4.5.

7.3. Power of MCPs

S, the number of non-true null hypotheses rejected, is a measure of the power of a MCP.

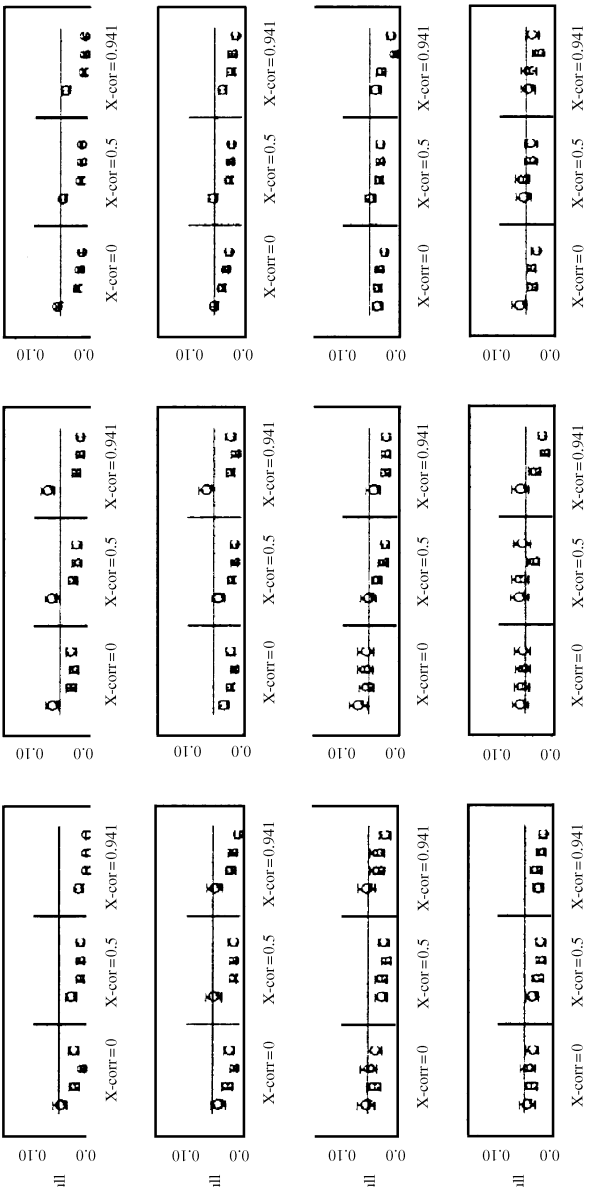

Fig. 3. It consists of 13 subplots and presents the results of the simulations in which the percentage of true null hypotheses was greater than 0. Simulations in each column of plots have the same deviation parameter d, and in each row have the same percentage of true null hypotheses. The simulations within each plot are for the three values of ρ_0. The letters drawn in each subplot are the computed FDR values: Q-the REAL MCP, R-RES MCP, B-BH MCP and C-UP-RES MCP. The two horizontal lines above and below a letter, are estimates of the simulation standard error. The horizontal line is drawn at FDR = 0.05.

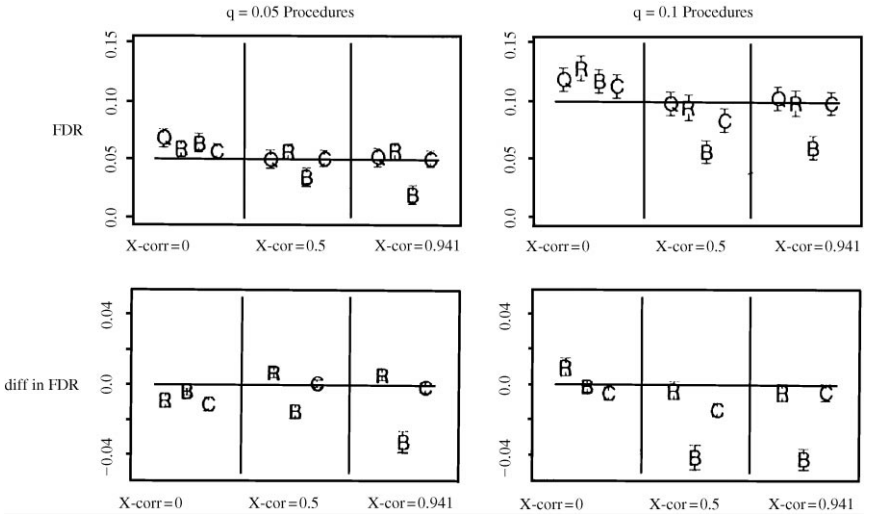

Fig. 4. In the first row of plots the FDR of the 4 MCPs are drawn, the second row is the FDR of the RES BH and con MCPs subtracted by the FDR of the real MCP. In the first column $q = 0.05$, in the second column $q = 0.1$. In each of the plots, the right sub-plot $\rho_0 = 0$, in the middle sub-plot $\rho_0 = 0.5$ and in the left sub-plot $\rho_0 = 0.941$.

Fig. 5 is a graphical summary of the average S values computed in the simulations for $q = 0.05$ MCPs.

Obviously, the efficiency as measured by the average proportion of hypotheses correctly rejected increases as the deviation parameter increases. As percentage of true null hypotheses increases efficiency of all MCPs decreases. The reason for this is that as number of true null hypotheses increases, risk of rejecting true null hypotheses increases thus the p-value adjustment is greater and as a result if a constant FDR rate is maintained the power decreases.

As the percentage of true null hypotheses increases the power of the RES BH and UP-RES MCPs increases relative to the power of the REAL MCP. When computing the Q_e, the basis for the REAL MCP, the number of true null hypotheses m_0 is known. But when computing Q^*, Q^*_β and Q^{BH} it is assumed that all hypotheses are true null hypotheses. In the BH MCP this means replacing m_0 by m. In the RES and UP-RES MCPs, $R^*(p)$ is generated instead of $V^*_0(p)$. As the percentage of true null hypotheses increases m_0 approaches m, and the distribution of $R^*(p)$ approaches distribution of $V(p)$, so the RES BH and UP-RES MCPs perform better relative to the REAL MCP. As X-correlation increases the efficiency of the REAL, RES and UP-RES MCPs increases.

For percentages of true null hypotheses greater than 50%, deviation parameters 0 and sometimes 1, X-correlation 0.5 but especially 0, the RES MCP is more powerful than the REAL MCP. It was shown that in general if $m_0 < m$ then $Q_e(p) \leqslant E_{\boldsymbol{P}} Q^*(p)$,

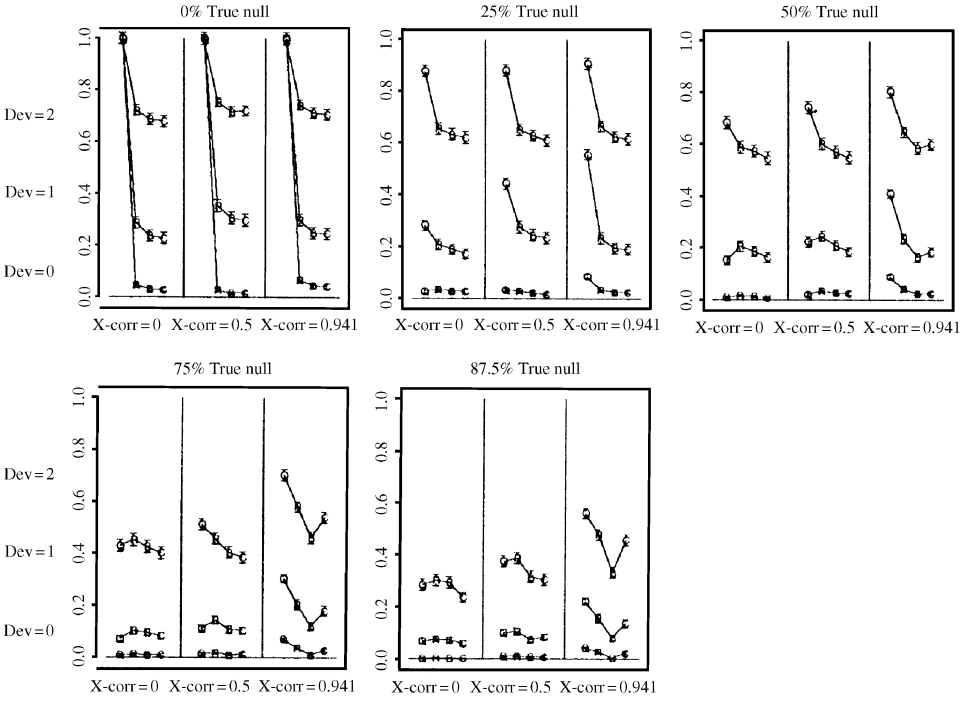

Fig. 5. Consists of 5 plots, a plot for each percentage of true null hypotheses. Each plot is made of sub-plots for each value of ρ_0. The set of letters connected by a line are the proportion of rejected non-true null hypotheses (s/m_1) by each MCP. In the upper set the deviation parameter is 2, in the middle set 1 and then 0.

thus the REAL MCP should be more powerful than the RES and MCP. Yet in cases in which the REAL MCP is weak, the RES MCP seems to be more powerful. A possible explanation is this: as $s(p)$ increase $Q_{V|s}$ decreases, allowing a MCP based on the conditional FDR correction to reject more false hypotheses than the REAL MCP. In situations in which the REAL MCP lacks power only hypotheses with very small p-values are rejected thus the values of $V(p)$ used in the testing are small. Under such conditions the value of $S(p)$ has a substantial affect on the conditional FDR correction, and its deviation from $Q_e(p)$. Recall that the resampling-based estimators estimate the conditional FDR correction. Therefore under such conditions similarity to the conditional MCP overcomes the inherent inferiority due to resampling the entire set of variables and the RES MCP is more powerful than the REAL MCP.

Fig. 6 shows a power comparison between the RES and BH MCPs. The RES MCP is uniformly superior to the BH MCP, especially for small deviation values and large ρ_0. Relative efficiency of RES MCP increases as the deviation parameter decreases, percentage of true null hypotheses and X-correlation increases.

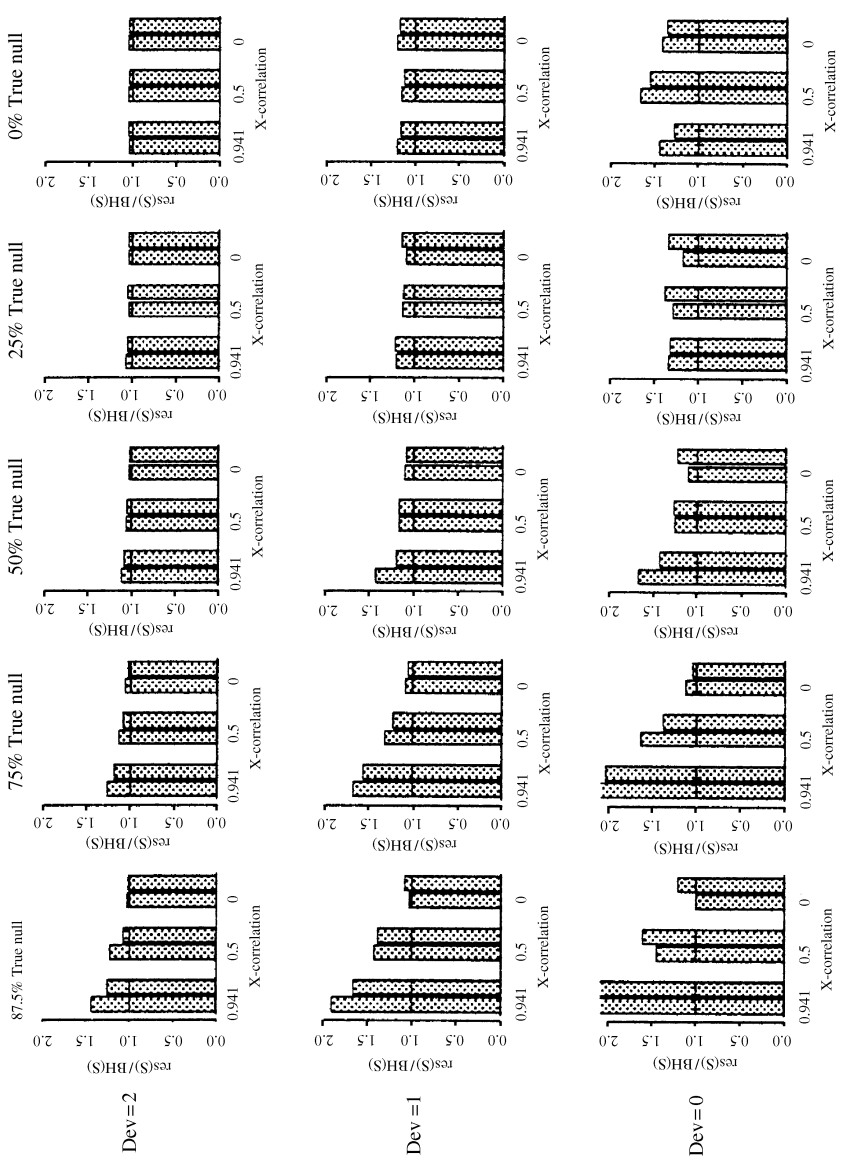

Fig. 6. The height of each bar is the S of the RES MCP divided by the S of the BH MCP. The columns in each figure are arranged according to the percentage of true null hypotheses, rows according to the deviation parameter. Each plot is a bar plot consisting of three pairs of bars, a pair for each value of ρ_0. In each pair of bars the right bar corresponds to the $q = 0.1$ MCP the left bar to the $q = 0.05$ MCP.

8. Proofs of propositions

Proof of Proposition 4.2. As defined, $Q^*_\beta(p) \geqslant E_{R^*} R^*(p)/(R^*(p) + r(p) - r^*_\beta(p))$. Therefore,

$$\Pr\{Q^*_\beta(p) \geqslant Q_V(p)\} \geqslant \Pr\left\{E_{R^*} \frac{R^*}{R^* + r - r^*_\beta} \geqslant E_V \frac{V}{V+s}\right\}$$

(Assuming subset pivotality distribution of $V(p)$ and $V^*_0(p)$ are identical)

$$\begin{aligned}
&\geqslant \Pr\left\{E_{V^*_0} \frac{V^*_0}{V^*_0 + r - r^*_\beta} \geqslant E_{V^*_0} \frac{V^*_0}{V^*_0 + s}\right\} \\
&\geqslant \Pr\{r - r^*_\beta \leqslant s\} = \Pr\{s + v - r^*_\beta \leqslant s\} \\
&= \Pr\{v \leqslant r^*_\beta\} = \Pr\{V^*_0 \leqslant r^*_\beta\} \\
&\geqslant \Pr\{R^*(p) \leqslant r^*_\beta(p)\} \geqslant 1 - \beta. \qquad \square
\end{aligned}$$

Proof of Proposition 4.3. If $S(p)$ and $V(p)$ are independent, the distribution of $V(p)\,|\,S(p) = s(p)$ and $V(p)$ are identical and assuming subset pivotality, the distribution of $V^*_0(p)$ and $V(p)$ are identical thus

$$Q_{V|s}(p) = E_V(p) \frac{V(p)}{s(p) + V_0(p)} = E_V(p)\left\{E_{R^*(p)} \frac{V^*_0(p)}{s(p) + V^*_0(p)}\right\}.$$

Recall that $Q^*(p)$ is defined as

$$Q^*(p) = \begin{cases} E_{R^*} \dfrac{R^*(p)}{R^*(p) + r(p) - p \cdot m} & \text{if } r(p) - r^*_\beta(p) \geqslant m \cdot p, \\ \Pr_{R^*}\{R^*(p) \geqslant 1\} & \text{otherwise.} \end{cases}$$

If $s(p) \geqslant pm$, then $r(p) = s(p) + v(p) \geqslant p \cdot m$, and since $R^*(p) \geqslant V^*_0(p)$,

$$Q^*(p) \geqslant E_{R^*(p)} \frac{R^*(p)}{R^*(p) + r(p) - p \cdot m} \geqslant E_{R^*(p)} \frac{V^*_0(p)}{V^*_0(p) + r(p) - p \cdot m}.$$

In expectation on the distribution of $V(p)$,

$$E_{V(p)} Q^*(p) \geqslant E_{V(p)}\left[E_{R^*(p)} \frac{V^*_0(p)}{V^*_0(p) + V(p) + s(p) - p \cdot m}\right].$$

Thus to prove the proposition it is sufficient to show that (dropping the "p")

$$E_V E_{R^*}\left[\frac{V^*_0}{V^*_0 + V + s - p \cdot m} - \frac{V^*_0}{s + V^*_0}\right] \geqslant 0.$$

$$\begin{aligned}
&E_V E_{R^*}\left[\frac{V^*_0}{V^*_0 + V + s - p \cdot m} - \frac{V^*_0}{s + V^*_0}\right] \\
&\quad = E_V E_{R^*}\left[\frac{sV^*_0 + (V^*_0)^2 - (V^*_0)^2 - VV^*_0 - sV^*_0 + pmV^*_0}{(V^*_0 + V + s - pm)(s + V^*_0)}\right] \\
&\quad = E_V E_{R^*}\left[\frac{V^*_0(pm - V)}{(s + V^*_0)(s + V + V^*_0 - p \cdot m)}\right]
\end{aligned}$$

$$=E_V\left[(pm-V)E_{R^*}\left[\frac{V_0^*}{(s+V_0^*)(s+V+V_0^*-p\cdot m)}\right]\right]$$

(denote, $P_v=P(V=v)$, $\varphi(V)=E_{V_0^*,V_1^*}V_0^*/(s+V_0^*)(s+V+V_0^*-p\cdot m)$, since $s\geqslant p\cdot m\varphi$ is positive)

$$=E_V(p\cdot m-V)\varphi(V)=\sum_{v=0}^{m_0}(p\cdot m-v)P_v\varphi(v)$$

$$=\sum_{v=0}^{[p\cdot m]}(p\cdot m-v)P_v\varphi(v)+\sum_{v=[p\cdot m]+1}^{m_0}(p\cdot m-v)P_v\varphi(v)$$

(because the left summation consists of positive expressions, right summation of negative expressions, and $\varphi(0)\geqslant\cdots\geqslant\varphi(m_0)$)

$$\geqslant\sum_{v=0}^{[p\cdot m]}(p\cdot m-v)P_v\varphi([p\cdot m])+\sum_{v=[p\cdot m]+1}^{m_0}(p\cdot m-v)P_v\varphi([p\cdot m]).$$

$$=\varphi([p\cdot m])\left\{\sum_{v=0}^{m_0}P_v\cdot p\cdot m-\sum_{v=0}^{m_0}P_v v\right\}=\varphi([p\cdot m])\cdot\{p\cdot m-E_V V\}$$

$$=\varphi([p\cdot m])\cdot p\cdot(m-m_0)\geqslant 0. \qquad \square$$

Proof of Proposition 4.4. If $S(p)$ and $V(p)$ are independent, the distribution of $V(p)\,|\,S(p)=s(p)$ and $V(p)$ are identical; therefore,

$$\begin{aligned}\Pr_{R|s=s}\{r(p)-r_\beta^*(p)<p\cdot m\}&=\Pr_V\{V(p)+s(p)-r_\beta^*(p)<p\cdot m\}\\&\geqslant\Pr_V\{V(p)-r_\beta^*(p)\leqslant 0\}\\&=\Pr_V\{V(p)\leqslant r_\beta^*(p)\}\end{aligned}$$

(assuming subset pivotality)

$$=\Pr_{R^*}\{V_0^*(p)\leqslant r_\beta^*(p)\}\geqslant\Pr\{R^*(p)\leqslant r_\beta^*(p)\}\geqslant 1-\beta.$$

Recall that $Q^*(p)$ was defined as

$$Q^*(p)=\begin{cases}E_{R^*}\frac{R^*(p)}{R^*(p)+r(p)-p\cdot m} & \text{if } r(p)-r_\beta^*(p)\geqslant p\cdot m,\\ \Pr_{R^*}\{R^*(p)\geqslant 1\} & \text{otherwise.}\end{cases}$$

Therefore under the conditions of the proposition, $Q^*(p)=\Pr_{R^*}(R^*(p)\geqslant 1)$ with probability $1-\beta$. And because, $\Pr_{R^*(p)}\{R^*(p)\geqslant 1\}\geqslant Q_V(p)$,

$$\Pr_{V(p)}\{Q^*(p)\geqslant Q_V(p)\}\geqslant 1-\beta. \qquad \square$$

Proof of Proposition 4.5. Under the complete null hypothesis the FWE equals the FDR and assuming subset pivotality the distribution of $R^*(p)$ and $R(p)$ are identical, thus

$$\Pr\{r(p) - r^*_\beta \leqslant 0\} = \Pr\{R^*(p) \leqslant r^*_\beta\} \geqslant 1 - \beta$$

and

$$\Pr\{R^*(p) \geqslant 1\} = Q_e(p).$$

To complete the proof recall that, $Q^*(p)$ and $Q^*_\beta(p)$ equal $\Pr\{R^*(p) \geqslant 1\}$, if $r(p) - r^*_\beta < pm$ and $r(p) - r^*_\beta \leqslant 0$ accordingly. □

Proof of Proposition 5.2. As defined $Q^*_\beta(p^{ul}_q)$ and $Q_{V|s}(p_q)$ equal q. Since $Q^*_\beta(p)$ is increasing in p, if $p^{ul}_q > p_q$ then $Q_{V|s}(p_q) = Q^*_\beta(p^{ul}_q) \geqslant Q^*_\beta(p_q)$. Thus, using the result proven in Proposition 8.4,

$$\Pr\{p^{ul}_q > p_q\} \leqslant \Pr\{Q^*_\beta(p^{ul}_q) > Q_{V|s}(p^{ul}_q)\} \leqslant \beta. \qquad \square$$

Proof of Proposition 5.3. Denote $p^{ul}_q = \sup_p\{Q^*_\beta(p) \leqslant q\}$, p^{ul}_q is a function of $\boldsymbol{P}$, the FWE of the MCP based on Q^*_β is $\Pr_{\boldsymbol{P}}\{V(p^{ul}_q(\boldsymbol{P})) \geqslant 1\}$. Denote $p_q = \sup_p\{\Pr(V(p) \geqslant 1) \leqslant q\}$, recall that under the complete null hypotheses $S \equiv 0$, thus the conditional FDR correction equals the FWE correction, since $V(p)$ is increasing in p, $p^{ul}_q \leqslant p_q$ implies

$$\Pr\{V(p^{ul}_q) \geqslant 1\} \leqslant \Pr\{V(p_q) \geqslant 1\} = q,$$

and according to Proposition 7.6,

$$\Pr_{\boldsymbol{P}}\{p^{ul}_q(\boldsymbol{P}) \geqslant p_q\} \leqslant \beta$$

therefore,

$$\begin{aligned}
\Pr_{\boldsymbol{P}}\{V(p^{ul}_q(\boldsymbol{P})) \geqslant 1\} &= \Pr_{\boldsymbol{P}}\{p^{ul}_q(\boldsymbol{P}) \leqslant p_q \wedge p^{ul}_q(\boldsymbol{P}) \geqslant 1\} \\
&\quad + \Pr_{\boldsymbol{P}}\{p^{ul}_q(\boldsymbol{P}) > p_q \wedge V(p^{ul}_q(\boldsymbol{P})) \geqslant 1\} \\
&\leqslant \Pr_{\boldsymbol{P}}\{V(p^{ul}_q) \geqslant 1 | p^{ul}_q \leqslant p_q\} + \Pr_{\boldsymbol{P}}\{p^{ul}_q(\boldsymbol{P}) > p_q\} \\
&\leqslant q + \beta. \qquad \square
\end{aligned}$$

Acknowledgements

We are thankful to Dr. Manes from the Israeli Meteorological Service (IMS) and to Prof. Alpert from Tel-Aviv University for introducing us to the meteorological problem, and making the data accessible. We are also thankful to the many comments of one of the referees which have helped us improve the style of the presentation.

References

Benjamini, Y., Fuchs, C., 1990. Conditional versus unconditional analysis in some regression models. Comm. Statist. Theory Methods 19 (12), 4731–4756.

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

Benjamini, Y., Hochberg, Y., Kling, Y., 1995. False discovery rate controlling procedures for pairwise comparisons. Dept. of Statistics and Operations Research Tech 95-2, Tel Aviv University.

Benjamini, Y., Yekutieli, D., 1997. The control of the false discovery rate in multiple testing under positive dependency. Dept. of Statistics and Operations Research RS-SOR-97-04, Tel Aviv University.

Dunnet, C.W., Tamhane, A.C., 1992. A step-up multiple test procedure. J. Amer. Statist. Assoc. 87, 162–170.

Heyse, J.F., Rom, D., 1988. Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. Biomet. J. 30, 883–896.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. Wiley, New York.

Manes, A. (Ed.), 1994. Seasonal forecasting of precipitation in Israel. Research Report No. 1, Israeli Meteorological Service, Beit Dagan, April 1994.

Shafer, G., Olkin, I., 1983. Adjusting p-values to account for selection over dichotomies. J. Amer. Statist. Assoc. 78, 674–678.

Troendle, J.F., 1995. A stepwise resampling method of multiple hypothesis testing. J. Amer. Statist. Assoc. 90, 370–378.

Wassmer, G., Reitmer, P., Kieser, M., Lehmacher, W., 1997. Procedures for testing multiple endpoints in clinical trials: an overview. J. Statist. Plann. Inference 82, 69–81 (this issue).

Westfall, P.H., Young, S.S., 1989. p-value adjustment for multiple tests in multivariate binomial models. J. Amer. Statist. Assoc. 84, 780–786.

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing. Wiley, New York.

Wright, P., 1992. Adjusted p-values for simultaneous inference. Biometrics 48, 1005–1013.

Yekutieli, D., 1996. Resampling-based FDR controlling multiple hypotheses testing. M.Sc. Dissertation, Department of Statistics and Operation Research, Tel Aviv University, Tel Aviv Israel.