Biometrika (2006), **93,** 3, *pp.* 491–507

Printed in Great Britain

Adaptive linear step-up procedures that control the false discovery rate

BY YOAV BENJAMINI

Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel University, Tel Aviv, Israel

ybenja@tau.ac.il

ABBA M. KRIEGER

Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

krieger@wharton.upenn.edu

AND DANIEL YEKUTIELI

Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel University, Tel Aviv, Israel

yekutiel@tau.ac.il

SUMMARY

The linear step-up multiple testing procedure controls the false discovery rate at the desired level q for independent and positively dependent test statistics. When all null hypotheses are true, and the test statistics are independent and continuous, the bound is sharp. When some of the null hypotheses are not true, the procedure is conservative by a factor which is the proportion m_0/m of the true null hypotheses among the hypotheses. We provide a new two-stage procedure in which the linear step-up procedure is used in stage one to estimate m_0, providing a new level q' which is used in the linear step-up procedure in the second stage. We prove that a general form of the two-stage procedure controls the false discovery rate at the desired level q. This framework enables us to study analytically the properties of other procedures that exist in the literature. A simulation study is presented that shows that two-stage adaptive procedures improve in power over the original procedure, mainly because they provide tighter control of the false discovery rate. We further study the performance of the current suggestions, some variations of the procedures, and previous suggestions, in the case where the test statistics are positively dependent, a case for which the original procedure controls the false discovery rate. In the setting studied here the newly proposed two-stage procedure is the only one that controls the false discovery rate. The procedures are illustrated with two examples of biological importance.

Some key words: False discovery rate; Multiple testing; Two-stage procedure.

1. INTRODUCTION

The traditional concern when testing m hypotheses simultaneously is to control the familywise error rate, the probability of making any false discovery. The restrictiveness of

the familywise error rate criterion leads to multiple testing procedures that are not powerful in the sense that the probability of rejecting null hypotheses that are false must also be small. At the other extreme lies the strategy of ignoring the multiplicity issue altogether, and testing each hypothesis at level α. This is a popular approach which increases the probability of rejecting null hypotheses that are false, but ignores the increased expected number of type I errors.

The false discovery rate criterion was developed by Benjamini & Hochberg (1995) to bridge these two extremes. The false discovery rate is the expectation of the proportion of rejected true null hypotheses among the rejected hypotheses. When the null hypothesis is true for all hypotheses, the false discovery rate and familywise error rate criteria are equivalent. However, when there are some hypotheses for which the null hypotheses are false, a procedure that controls the false discovery rate may reject many more such hypotheses at the expense of a small proportion of erroneous rejections.

The linear step-up procedure, or so-called Benjamini & Hochberg procedure, controls the false discovery rate at a desired level qm_0/m when the test statistics are independent (Benjamini & Hochberg, 1995) or positively dependent (Benjamini & Yekutieli, 2001). Even though this procedure rejects false null hypotheses more frequently than procedures that control the familywise error rate, if we knew m_0 the procedure could be improved by using $q' = qm/m_0$, to achieve precisely the desired level q. In this paper we develop and compare some adaptive false discovery rate controlling procedures that begin by estimating m_0. In § 2, we recall the formal definition of the false discovery rate criterion, the linear step-up procedure, and review the background for the problem at hand.

2. Background

2·1. *The false discovery rate*

Let H_{0i} $(i = 1, \ldots, m)$ be the tested null hypotheses. For $i = 1, \ldots, m_0$ the null hypotheses are true, and for the remaining $m_1 = m - m_0$ the null hypotheses are false. Let V denote the number of true null hypotheses that are erroneously rejected and let R be the total number of hypotheses that are rejected. Now define the proportion of false discoveries by $Q = V/R$ if $R > 0$ and $Q = 0$ if $R = 0$. The false discovery rate is $\text{FDR} = E(Q)$ (Benjamini & Hochberg, 1995).

A few recent papers have illuminated the false discovery rate from different points of view, namely asymptotic, Bayesian, empirical Bayes, as the limit of empirical processes and in the context of penalised model selection; see Efron et al. (2001), Storey (2002), Genovese & Wasserman (2004) and Abramovich et al. (2006). Some of the studies emphasised variants of the false discovery rate, such as its conditional value given that some discovery is made (Storey, 2002), or the distribution of the proportion of false discoveries itself (Genovese & Wasserman, 2004). Procedures were developed for specific settings (Troendle, 1999), and the applicability of existing procedures have been studied; see for example Sarkar (2002).

The linear step-up procedure makes use of the m p-values, $P = (P_1, \ldots, P_m)$. Let $p_{(1)} \leqslant \ldots \leqslant p_{(m)}$ be their ordered observed values.

Definition 1 (*The one-stage linear step-up procedure*)

Step 1. *Let* $k = \max\{i : p_{(i)} \leqslant iq/m\}$.

Step 2. *If such a k exists, reject the k hypotheses associated with* $p_{(1)}, \ldots, p_{(k)}$; *otherwise do not reject any of the hypotheses.*

Benjamini & Hochberg (2000) provide a detailed historical review. For the purpose of practical interpretation and flexibility in use, as well as for comparison with other approaches, the results of the linear step-up procedure can also be reported in terms of the FDR-adjusted p-values. Formally, the FDR-adjusted p-value of $H_{(i)}$ is $p_{(i)}^{\mathrm{LSU}} = \min\{mp_{(j)}/j | j \geqslant i\}$. Thus the linear step-up procedure at level q is equivalent to rejecting all hypotheses whose FDR-adjusted p-value is at most q.

The linear step-up procedure is quite striking in its ability to control the false discovery rate under independence at precisely qm_0/m, regardless of the distributions of the test statistics corresponding to false null hypotheses, when the distributions under the simple null hypotheses are continuous.

Benjamini & Yekutieli (2001) studied the procedure under dependence. For some type of positive dependence they showed that the above remains an upper bound. Even under the most general dependence structure, where the false discovery rate is controlled merely at level $q(1 + 1/2 + 1/3 + \ldots + 1/m)$, it is again conservative by the same factor m_0/m.

2·2. *The role of m_0 in testing*

Knowledge of m_0 can therefore be very useful for improving upon the performance of the FDR controlling procedure. If m_0 were given to us by an 'oracle', the linear step-up procedure with $q' = qm/m_0$ would control the false discovery rate at precisely the desired level q in the independent and continuous case, and would then be more powerful in rejecting hypotheses for which the alternative holds. In a well-defined asymptotic context, Genovese & Wasserman (2002) showed it to be the best possible procedure in that it minimises the expected proportion of the hypotheses for which the alternatives hold among the nonrejected ones, minimising the false nondiscovery rate.

The factor m_0/m plays a role in other settings as well. It is a 'correct' prior for a full Bayesian analysis (Storey, 2002, 2003). Estimating this factor is also an important ingredient in the empirical Bayes approach to multiplicity (Efron et al., 2001).

Even when we are controlling the familywise error rate in the frequentist approach, knowledge of m_0 is useful. Using α/m_0 is a more powerful procedure than the standard Bonferroni, which uses α/m, yet also controls the familywise error rate. Holm's procedure and Hochberg's procedure have been similarly modified by Hochberg & Benjamini (1990) to construct more powerful versions. It is thus interesting to note that estimation of m_0 from the data is needed from the points of view of different schools of thought. Moreover, estimating m_0 becomes easier as more parameters are tested.

3. Adaptive procedures

Adaptive procedures first estimate the number of null hypotheses m_0, and then use this estimate to revise a multiple test procedure. The following adaptive approach is based on the linear step-up procedure.

Definition 2 (*Generalised adaptive linear step-up procedure*)

Step 1. *Compute* $\hat{m}_0$.

Step 2. *If* $\hat{m}_0 = 0$ *reject all hypotheses*; *otherwise*, *test the hypotheses using the linear step-up procedure at level* $qm/\hat{m}_0$.

Schweder & Spjøtvoll (1982) were the first to try to estimate m_0, albeit informally, from the quantile plot of the p-values versus their ranks. This plot will tend to show linear behaviour for the larger p-values which are more likely to correspond to true null

hypotheses. Thus one can inspect the plot and choose the largest k p-values for which the behaviour seems linear, and estimate the slope of the line passing through them. Its reciprocal was used as an estimate of m_0, rejecting the hypotheses corresponding to the $m - m_0$ smallest p-values. Hochberg & Benjamini (1990) formalised the approach and incorporated the estimate into the various procedures that control the familywise error rate.

Benjamini & Hochberg (2000) incorporated their proposed estimator for m_0 into the generalised adaptive linear step-up procedure as follows.

Definition 3 (*The adaptive linear step-up procedure of Benjamini & Hochberg*)
Step 1. *Use the linear step-up procedure at level q, and if no hypothesis is rejected stop; otherwise, proceed.*
Step 2. *Estimate* $m_0(k)$ *by* $(m + 1 - k)/(1 - p_{(k)})$.
Step 3. *Starting with* $k = 2$ *stop when for the first time* $m_0(k) > m_0(k-1)$.
Step 4. *Estimate* $\hat{m}_0 = \min\{m_0(k), m\}$ *rounding up to the next highest integer.*
Step 5. *Use the linear step-up procedure with* $q^* = qm/\hat{m}_0$.

The choice in Step 2 was justified as follows. Let $r(\alpha) = \#\{p_{(i)} \leqslant \alpha\}$. Then $m - r(\alpha)$ is potentially the number of true null hypotheses except that $m_0\alpha$ true null hypotheses are expected to be among the $r(\alpha)$ rejected. Hence, solving $m_0 \backsimeq m - \{r(\alpha) - m_0\alpha\}$ for m_0 yields $m_0 \backsimeq \{m - r(\alpha)\}/(1 - \alpha)$. Use $\alpha = p_{(k)}$ to obtain approximately $m_0(k)$.

The adaptive procedure was shown by simulation to provide tighter control of the false discovery rate than the linear step-up procedure. Not surprisingly, the simulation also showed it to be much more powerful (Benjamini & Hochberg, 2000; Hsueh et al., 2003; Black, 2004). It is of interest to note that this adaptive procedure was the original FDR-controlling procedure suggested by Y. Benjamini and Y. Hochberg in an unpublished Tel Aviv University technical report. Later the authors used the conservative bound of 1 for the m_0/m factor, which enabled the proof of the FDR-controlling property of the linear step-up procedure.

An estimator of the above form evaluated at a single prespecified α quantile of p-values $P_{(i)}$, where $i = \alpha m$, is easier to study. Such an estimator for m_0 is mentioned in passing in Efron et al. (2001) and goes back to earlier versions of Storey (2002), although their interest in the estimator was for different purposes. Using an estimate such as the median of the $\{p_{(i)}\}$, loosely denoted by $p_{(m/2)}$ within the linear step-up procedure, we obtain the following procedure.

Definition 4 (*Median adaptive linear step-up procedure*)
Step 1. *Estimate* m_0 *by* $\hat{m}_0 = (m - m/2)/(1 - p_{(m/2)})$.
Step 2. *Use the linear step-up procedure with* $q^* = qm/\hat{m}_0$.

For estimating m_0, Storey (2002) and Storey & Tibshirani (2003a) recommended using a fixed α, λ in their notation, such as $\alpha = \frac{1}{2}$ in the above. This yields the following procedure.

Definition 5 (*The adaptive linear step-up procedure with Storey's estimator*)
Step 1. *Let* $r(\lambda) = \#\{p_{(i)} \leqslant \lambda\}$.
Step 2. *Estimate* m_0 *by* $\hat{m}_0 = \{m - r(\lambda)\}/(1 - \lambda)$.
Step 3. *Use the linear step-up procedure with* $q^* = qm/\hat{m}_0$.

The above procedure, and the special case with $\lambda = \frac{1}{2}$, are also incorporated in recent versions of SAM software (Storey & Tibshirani, 2003b) even though the p-values are

estimated by resampling. Subsequently, in Storey et al. (2004), the above procedure was modified by replacing $\{m - r(\lambda)\}$ by $\{m + 1 - r(\lambda)\}$ and further requiring that $p_{(i)} \leqslant \lambda$ for a hypothesis to be rejected. These modifications stemmed from theoretical results in their paper, in that they ensure FDR-control in problems where a finite number of hypotheses are tested.

Mosig et al. (2001) suggested a procedure involving an iterated stopping rule which uses the number of p-values falling within some arbitrary cell boundaries over the range $(0, 1)$. The motivation is correct and similar in spirit to the above, but the procedure as published is far from controlling the false discovery rate at the desired level; see § 6.

Other computer-intensive adaptive procedures based on resampling and bootstrapping have also been suggested; see Yekutieli & Benjamini (1999), the resampling-based choice of λ in Storey (2002) and Storey et al. (2004) and the cubic spline fit in Storey & Tibshirani (2003a).

4. The new two-stage procedures

The idea underlying the two-stage procedure is that the value of m_0 can be estimated from the results of the one-stage procedure.

Definition 6 (*The two-stage linear step-up procedure*, TST)

Step 1. *Use the linear step-up procedure at level* $q' = q/(1 + q)$. *Let* r_1 *be the number of rejected hypotheses. If* $r_1 = 0$ *do not reject any hypothesis and stop*; *if* $r_1 = m$ *reject all* m *hypotheses and stop*; *otherwise continue.*

Step 2. *Let* $\hat{m}_0 = (m - r_1)$.

Step 3. *Use the linear step-up procedure with* $q^* = q'm/\hat{m}_0$.

The procedure can be motivated as follows. By definition $m_0 \leqslant m - (R - V)$. The linear step-up procedure used in the first stage ensures that $E(V/R) \leqslant qm_0/m$, so that V is approximately less than or equal to qm_0R/m. Hence, $m_0 \leqslant m - (R - qm_0R/m)$, from which we obtain

$$m_0 \leqslant \frac{m - R}{1 - (R/m)q} \leqslant \frac{m - R}{1 - q} \leqslant (m - R)(1 + q). \tag{1}$$

The right-most bound is the one implicitly used in the above procedure. In § 5 it is proven that this two-stage procedure has a false discovery rate that does not exceed q for independent test statistics.

This two-stage procedure uses the number rejected at the first stage to estimate m_0, it controls the false discovery rate and it necessarily increases power. Hence, we may extend this approach using $m - r_2$ at the third stage, and so on. In the multiple-stage linear step-up procedure, the steps of the two-stage linear step-up procedure are repeated as long as more hypotheses are rejected. This procedure can also be expressed in an elegant way using the sequence of constants $ql/(m + 1 - j)$ at each stage. However, using $(1 + q)$ in the denominator does not suffice, since the effective level used is $q^* = qj/(m + 1 - j)$, which may be bigger than q. This suggests inflating the cut-offs to

$$\left\{\frac{q}{1 + qj/(m + 1 - j)}\right\}\frac{l}{m + 1 - j} \tag{2}$$

to obtain the following procedure.

Definition 7 (*The multiple-stage linear step-up procedure*)

Step 1. *Let* $k = \max\{i: \text{for all } j \leqslant i \text{ there exists } l \geqslant j \text{ so that } p_{(l)} \leqslant ql/\{m+1-j(1-q)\}\}$.

Step 2. *If such a k exists, reject the k hypotheses associated with* $p_{(1)}, \ldots, p_{(k)}$; *otherwise reject no hypothesis.*

The last procedure has an interesting internal-consistency property. The number of hypotheses tested minus the number rejected, less the proportion q erroneously rejected, is also the number used in the denominator of the linear step-up procedure as the estimator of m_0. A simpler, more conservative procedure enjoying the same property is to require $l = j$ in the above definition, resulting in a multiple-stage step-down procedure.

5. Analytical results

In this section we derive an expression for the upper bound of the false discovery rate of the generalised adaptive linear step-up procedure for independently distributed test statistics. The distributions of the test statistics may be discrete, in which case the distributions of the p-values under the null hypotheses should be stochastically larger than uniform as usual. The estimator $\hat{m}_0$ of m_0 is assumed to be an increasing function of each of the components of P.

It was shown in Benjamini & Yekutieli (2001) that the false discovery rate of any multiple comparison procedure can be expressed as

$$\begin{aligned}\text{FDR} &= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \text{pr}\{k \text{ hypotheses are rejected one of which is } H_{0i}\} \\ &= m_0 \sum_{k=1}^{m} \frac{1}{k} \text{pr}\{k \text{ hypotheses are rejected one of which is } H_{01}\}.\end{aligned}$$

The second equality follows as the problem is exchangeable in the p-values corresponding to the m_0 true null hypotheses. Let P_{01} be the p-value associated with H_{01}. Note that there must be at least one hypothesis that is null, that is $m_0 \geqslant 1$ because otherwise $\text{FDR} = 0$. Let $P^{(1)}$ be the vector of p-values corresponding to the $m-1$ hypotheses excluding H_{01}. Conditioning on $P^{(1)}$ we can express FDR as

$$\text{FDR} = m_0 E_{P^{(1)}} Q(P^{(1)}), \tag{3}$$

where $Q(P^{(1)})$ is defined by

$$Q(P^{(1)}) = \sum_{k=1}^{m} \frac{1}{k} \text{pr}_{P_{01}|P^{(1)}}\{k \text{ hypotheses are rejected one of which is } H_{01}\}.$$

For each value of $P^{(1)}$, let $r(P_{01})$ denote the number of hypotheses that are rejected, as a function of P_{01}, and let $\iota(P_{01})$ be the indicator that H_{01} is rejected as a function of P_{01}. Then

$$Q(P^{(1)}) = E\left\{\frac{\iota(P_{01})}{r(P_{01})} | P^{(1)}\right\} = \int \frac{\iota(p)}{r(p)} d\mu_{01}(p),$$

where, by the assumed independence, we can take μ_{01} to be the marginal distribution of P_{01}. In the continuous case, μ_{01} is just the uniform distribution on $[0, 1]$ and, in the discrete case, it is necessarily stochastically larger than the uniform.

We make two claims, first that $r(p)$ is a nonincreasing function, and secondly that $\iota(p)$ takes the form $1_{[0,p^*]}$, where $p^* \equiv p^*(P^{(1)})$ satisfies $p^* \leqslant qr(p^*)/\hat{m}_0(p^*)$. If these claims are true, then

$$Q(P^{(1)}) = \int \frac{\iota(p)}{r(p)} d\mu_{01}(p) \leqslant \frac{p^*}{r(p^*)} \leqslant \frac{q}{\hat{m}_0(p^*)}. \tag{4}$$

This follows immediately because $\mathrm{pr}(P_{01} \leqslant p^*) \leqslant p^*$. The bound on FDR follows, in the form

$$\mathrm{FDR} \leqslant qE_{P^{(1)}}\left\{\frac{m_0}{\hat{m}_0(p^*)}\right\}. \tag{5}$$

To prove the claims, note that, for any $P^{(1)}$ and a P_{01} such that $\iota(P_{01}) = 1$, that is H_{01} is rejected, there are exactly $r(P_{01})$ p-values below $qr(P_{01})/\hat{m}_0(P_{01})$, and for any $k > r(P_{01})$ there are strictly fewer than k p-values below $qk/\hat{m}_0(P_{01})$, because otherwise $r(P_{01})$ would be larger by construction. Since $\hat{m}_0(P_{01})$ is increasing in P_{01}, as P_{01} increases all the critical values $qk/\hat{m}_0(P_{01})$ decrease; hence, the number of p-values below each $qk/\hat{m}_0(P_{01})$ cannot increase. It follows that $r(P_{01})$ cannot increase, which proves the first claim. For the second claim, note that $\iota(P_{01}) = 1$ as long as $P_{01} \leqslant qr(P_{01})/\hat{m}_0(P_{01})$.

To prove control of the false discovery rate for a generalised adaptive linear step-up procedure we need to evaluate the right-hand side of (5), which is an expectation over $m-1$ p-values, m_1 of which are generated according to the distribution of alternative hypothesis p-values and $m_0 - 1$ are independent and identically distributed $\mathrm{Un}[0, 1]$. Since $\hat{m}_0$ is stochastically larger as p-values have stochastically larger distributions, the right-hand side of (5) is maximised when the m_1 p-values corresponding to H_{1i} are all zero with probability one. It is interesting to note that, although setting these p-values to 0 maximises the bound in (5), this does not necessarily lead to the maximum value of FDR in (3).

We will derive bounds for the false discovery rates of some of the two-stage procedures discussed earlier. The following is necessary in this regard.

LEMMA 1. *If* $Y \sim \mathrm{Bi}(k-1, p)$ *then* $E\{1/(Y+1)\} < 1/(kp)$.

Proof. Elementary calculations give

$$E\{1/(Y+1)\} = \frac{1}{kp}\{1-(1-p)^k\} \leqslant \frac{1}{kp}. \qquad \square$$

THEOREM 1. *When the test statistics are independent the two-stage procedure controls the false discovery rate at level* q.

Proof. Recall that, in a two-stage procedure, the first stage is a linear step-up procedure at level $q' \equiv q/(1+q)$, r_1 is the number of hypotheses rejected at stage 1, and $\hat{m}_0$ equals $m - r_1$. Then $\hat{m}_0$ can only be one of two values $\hat{m}_0(0)$ or $\hat{m}_0(1)$. For $P_{01} \leqslant r_1 q'/m$, H_{01} is rejected at both stages of the two-stage procedure, and $\hat{m}_0 = \hat{m}_0(0)$. For $P_{01} > r_1 q'/m$, H_{01} is not rejected at the first stage, and hence $\hat{m}_0 = \hat{m}_0(1)$; however, as long as

$P_{01} \leqslant r(P_{01})q'/\hat{m}_0(1)$, H_{01} is rejected at the second stage, and thus $\hat{m}_0(p^*) = \hat{m}_0(1)$ and, according to (4),

$$Q(P^{(1)}) \leqslant \frac{q'}{\hat{m}_0(1)}. \tag{6}$$

There is just one anomaly. If $\hat{m}_0(1) = m$ then for $P_{01} > r_1 q'/m$ the second stage of the testing procedure is identical to the first stage; thus H_{01} is no longer rejected and $\hat{m}_0(p^*) = \hat{m}_0(0)$. However, note that, as $p^* = r_1 q'/m$ and $r_1 \leqslant r(p^*)$, from the first inequality in (4) we have

$$Q(P^{(1)}) \leqslant \frac{p^*}{r(p^*)} \leqslant \frac{r_1 q'/m}{r_1} = \frac{q'}{m} = \frac{q'}{\hat{m}_0(1)}.$$

Hence inequality (6) is still satisfied.

As $\hat{m}_0(1)$ is stochastically larger than $Y + 1$, where $Y \sim \mathrm{Bi}\{m_0 - 1, 1 - q/(q+1)\}$, if q is replaced with $q' = q/(q+1)$ in Lemma 1, inequality (6) yields

$$\mathrm{FDR} \leqslant m_0 E_{P^{(1)}} Q(P^{(1)}) \leqslant \frac{q}{1+q} E_{P^{(1)}} \frac{m_0}{Y+1} \leqslant \frac{q}{1+q} \frac{m_0}{m_0\{1/(1+q)\}} = q. \qquad \square$$

A proof of the above was given in D. Yekutieli's 2002 Ph.D. thesis from Tel Aviv University. The quantile-based estimator of m_0 is $\hat{m}_0 = (m + 1 - k)/(1 - P_{(k)})$ for an arbitrary k ($1 \leqslant k \leqslant m$). This is the estimator used in the adaptive linear step-up procedure of Benjamini & Hochberg, albeit at a prespecified quantile. Using this estimator in the generalised adaptive linear step-up procedure leads to the following theorem.

Theorem 2. *When the test statistics are independent the quantile adaptive linear step-up procedure controls the false discovery rate at level q.*

Proof. If $k \leqslant m_1$ then according to (5) the false discovery rate of this two-stage procedure is

$$\mathrm{FDR} \leqslant E_{P^{(1)}} \frac{qm_0}{(m+1-k)/(1-P_{(k)})} \leqslant \frac{qm_0}{(m_0+1)} E_{P^{(1)}}(1 - P_{(k)}) \leqslant q.$$

Assume now that $m_1 < k$. We will first compute the expected value of $1/\hat{m}_0(0)$. Since $P_{01} = 0$, in addition to P_{01} there are $k - 1$ p-values less than $P_{(k)}$. The number of true null p-values less than or equal to $P_{(k)}$, not counting P_{01}, is at least $k - m_1 - 1$. Hence the distribution of $P_{(k)}$ is stochastically greater than the $k - m_1 - 1$ ordered p-values out of $m_0 - 1$ p-values that are independent $\mathrm{Un}[0, 1)$. Therefore, with $P_{01} = 0$,

$$E_{P^{(1)}} P_{(k)} \leqslant \frac{(k - m_1 - 1) + 1}{(m_0 - 1) + 1} = \frac{k - m_1}{m_0}.$$

Thus

$$E_{P^{(1)}} \frac{1}{\hat{m}_0(0)} \leqslant \left(1 - \frac{k - m_1}{m_0}\right) \Big/ (m + 1 - k) < 1/m_0.$$

Finally, as $\hat{m}(p^*) \geqslant \hat{m}_0(0)$, returning to (5) we have

$$\mathrm{FDR} \leqslant E_{P^{(1)}} \frac{qm_0}{\hat{m}_0(0)} < q. \qquad \square$$

As an illustration we consider an upper bound for the adaptive linear step-up procedure with Storey's estimator. The definition of $\hat{m}_0$ in Storey (2002) is $\hat{m}_0 = \{m - r(\lambda)\}/(1-\lambda) = \#\{P_i > \lambda\}/(1-\lambda)$. In this case, as P_{01} varies, there are two distinct values of $\hat{m}_0$:

$$\hat{m}_0 = \begin{cases} \#\{P^{(1)} > \lambda\}/(1-\lambda), & \text{if } P_{01} \leqslant \lambda, \\ (\#\{P^{(1)} > \lambda\} + 1)/(1-\lambda), & \text{if } P_{01} > \lambda. \end{cases}$$

When used within a generalised adaptive linear step-up procedure, if $p^* \leqslant \lambda$, then $\hat{m}_0(p^*)$ is stochastically greater than $W/(1-\lambda)$, where W is $\mathrm{Bi}(m_0 - 1, 1-\lambda)$. This causes a technical problem, as $E\{1/\hat{m}_0(p^*)\}$ is infinite because there is a nonzero probability, albeit very small for large m, that W is zero.

In Storey et al. (2004) two modifications were suggested to the original definition of $\hat{m}_0$ when used in testing. First, no hypothesis is rejected with p-value $> \lambda$. Secondly, $\hat{m}_0$ is modified to $(\#\{P_i > \lambda\} + 1)/(1-\lambda)$. With the second modification, (5) and Lemma 1 imply that

$$\mathrm{FDR} \leqslant \frac{qm_0}{m_0(1-\lambda)/(1-\lambda)} = q. \tag{7}$$

Remark 1. Lemma 1 showed that $E\{1/(Y+1)\} = \{1-(1-p)^k\}/kp$, where in our case $k = m_0$ and $p = 1-\lambda$. Substituting this result instead of its bound into (5) yields $\mathrm{FDR} \leqslant (1-\lambda^{m_0})q$, which agrees with the bound in Storey et al. (2004).

Remark 2. For illustration, we computed the bounds on FDR for the original version based on the above analysis in two cases. For $q = 0{\cdot}05$ and $\lambda = 0{\cdot}05$, in which case the condition $m \geqslant \lambda/\{q(1-\lambda)\}$ holds for all m, we obtain the following bounds: when $m = 20$, $\mathrm{FDR} \leqslant 0{\cdot}054$; when $m = 100$, $\mathrm{FDR} \leqslant 0{\cdot}051$; and, when $m = 500$, $\mathrm{FDR} \leqslant 0{\cdot}05$. For $q = 0{\cdot}05$ and $\lambda = 0{\cdot}5$, the results are as follows: when $m = 20$, $\mathrm{FDR} \leqslant 0{\cdot}075$; when $m = 100$, $\mathrm{FDR} \leqslant 0{\cdot}058$; and, when $m = 500$, $\mathrm{FDR} \leqslant 0{\cdot}052$. Thus, in this case, as long as m is in the hundreds the modifications are essential.

6. Simulation study

6·1. *The procedures to be compared*

A simulation study was performed to compare the FDR-control and the power of various adaptive procedures for controlling the familywise error rate and the false discovery rate. The seven procedures that were investigated in some detail can be roughly divided into two types, namely newly suggested adaptive procedures, numbered 1–3 below, and previously suggested adaptive FDR-controlling procedures, numbered 4–7 below. Three other procedures, numbered 8–10, serve as benchmarks for comparing performance.

Procedure 1. The two-stage linear step-up procedure, denoted by TST in the tables; see Definition 6. By Theorem 1, this procedure controls the false discovery rate at level q.

Procedure 2. The modified two-stage procedure, denoted by M-TST. This procedure makes use of q in stage 1, and $q' = q(1+q)$ in stage 2. Even though the proof requires the use of $q/(1+q)$ at both stages, we explore this procedure as well since it is more natural to use q at the first stage.

Procedure 3. The multiple-stage linear step-up procedure, denoted by MST; see Definition 7.

Procedure 4. The adaptive linear step-up procedure of Benjamini & Hochberg (2000), denoted by ABH; see Definition 3.

Procedure 5. The adaptive linear step-up procedure with Storey's estimator, denoted by S-HLF; see Definition 5 with $\lambda = \frac{1}{2}$.

Procedure 6. The adaptive linear step-up procedure as modified in Storey et al. (2004), denoted by M-S-HLF.

Procedure 7. The median adaptive linear step-up procedure, denoted by MED-LSU.

Procedure 8. The adaptive Hochberg procedure of Hochberg & Benjamini (1990), denoted by A-HCH. This adaptive step-up procedure is designed to control the familywise error rate.

Procedure 9. The linear step-up procedure, denoted by LSU; see Definition 1. This nonadaptive procedure controls the false discovery rate at level m_0/m.

Procedure 10. The linear step-up procedure at level qm/m_0, denoted by ORC.

The 'Oracle' procedure, number 10 above, which uses m_0/m to control the false discovery rate at the exact level q, is obviously not implementable in practice as m_0 is unknown. It serves as a benchmark against which other procedures can be compared. It also serves as a variance-reduction method in the simulation study under independence: a large reduction in variance is achieved by comparing the estimated difference in false discovery rates achieved by the procedure in question and that of procedure 10 to zero.

In the first part of the study, the number of tests m was set at $m = 4$, 8, 16, 32, 64, 128, 256 and 512. The fraction of the false null hypotheses was 0%, 25%, 50%, 75% and 100%. The P-values were generated in the following way. First, let, $Z_0, Z_1, \ldots, Z_m$ be independent and identically distributed $N(0, 1)$. Next, let $Y_i = \sqrt{\rho} Z_0 + \sqrt{(1-\rho)} Z_i - \mu_i$, for $i = 1, \ldots, m$, and let $P_i = 1 - \Phi(Y_i)$. We used $\rho = 0$, 0·1, 0·25, 0·5 and 0·75, with $\rho = 0$ corresponding to independence. The values of μ_i are zero for $i = 1, \ldots, m_0$, the m_0 hypotheses that are null. In one case, we let $\mu_i = 5$ for $i = m_0 + 1, \ldots, m$. This leads to $P_i \simeq 0$ for hypotheses that are not null; this is referred to as the 'all at 5' case. In the second case, the value of μ_i was $\mu_i = i$ for $i = 1, 2, 3, 4$. This cycle was repeated to produce the desired m_1 values under H_1. This is referred to as the '1 2 3 4' configuration. The resulting p-values under H_1 are clearly less extreme than those in the first case.

We also tested a few more procedures. The multiple-stage step-down procedure, a variation on Procedure 3, showed performance very similar in terms of FDR-control and was, as expected, slightly less powerful. The advantage of the step-down version is that only the extreme p-values are needed. The procedure of Mosig et al. (2001) did not control the false discovery rate at the desired level, the false discovery rate often exceeding 0·5. However, a slight modification of this procedure is similar to a modification of the ABH procedure and performed similarly. The modified two-stage Procedure 2 is very similar to Procedure 1. It has the advantage that it is run at the first stage at level q. However, it can happen that at the first stage a hypothesis is rejected and at the second stage it is not. This is rare, and occurs only for large m and m_0 close to m.

The simulation results are based on 10 000 replications. The standard error of the estimated false discovery rate is of the order of 0·002 for all the procedures. As mentioned above, the standard error of the performance of a procedure relative to that of the Oracle

Table 1: *Simulation study with independent test statistics. Estimated values of* FDR *for selected values* m_0 *and* m. *The value for the Oracle is set at its expected value* 0·05, *and the others are estimated from the differences. The standard errors are less than* 0·002 *in all cases*

	$m_0/m = 1$			$m_0/m = 0{\cdot}75$			$m_0/m = 0{\cdot}50$		
	$m = 16$	$m = 64$	$m = 256$	$m = 16$	$m = 64$	$m = 256$	$m = 16$	$m = 64$	$m = 256$
TST	0·048	0·048	0·047	0·048	0·048	0·048	0·049	0·049	0·049
M-TST	0·050	0·049	0·047	0·048	0·048	0·048	0·049	0·049	0·049
MST	0·050	0·050	0·050	0·050	0·050	0·050	0·050	0·050	0·050
ABH	0·050	0·050	0·050	0·048	0·049	0·050	0·047	0·049	0·049
S-HLF	0·057	0·051	0·050	0·061	0·052	0·051	0·070	0·053	0·051
M-S-HLF	0·049	0·050	0·050	0·050	0·050	0·050	0·050	0·050	0·050
MED-LSU	0·053	0·050	0·050	0·053	0·051	0·050	0·049	0·049	0·050
A-HCH	0·050	0·050	0·050	0·025	0·008	0·003	0·012	0·001	0·000
LSU	0·050	0·050	0·050	0·038	0·037	0·038	0·025	0·025	0·025
ORC	0·050	0·050	0·050	0·050	0·050	0·050	0·050	0·050	0·050

is even smaller. In Table 1 we used the fact that the expectation of the false discovery rate is exactly 0·05 to estimate the false discovery rate for the other procedures from the difference.

6·2. *Independent test statistics*

The results of the FDR-control under independence for the 10 procedures were higher for the 'all at 5' case than for the '1 2 3 4' case described above. Results are given in Table 1 for some of the configurations. Note that all procedures except S-HLF and MED-LSU control the false discovery rate at levels very close to yet below 0·05 at all configurations. The results for the stated two are above 0·05 for smaller m. The modification in Storey et al. (2004) solves the problem for S-HLF. In both cases the values are very close to the theoretical upper bounds derived in § 5. The fact that the level of false discovery rate approaches 0·05 as m increases supports the theoretical result in Genovese & Wasserman (2004) that, asymptotically in m, it controls the false discovery rate.

For the MED-LSU procedure the overshoot is smaller and it decreases faster. By $m = 64$ it is within simulation-noise level. The source of this problem is the fact that MED-LSU uses $(m - k)$ in the numerator, while the theorem that the false discovery rate is controlled is for the estimator with $(m + 1 - k)$. The ratio of the two decreases as m increases. It is important to emphasise that, in the three procedures S-HLF, M-S-HLF and MED-LSU, unlike in the other two-stage procedures, there is no restriction that $\hat{m}_0 \leqslant m$. It is important not to add such a requirement in the implementation step, because it will harm the FDR-controlling properties. In the other procedures $\hat{m}_0$ cannot exceed m.

Power comparisons are made at the more realistic configuration of '1 2 3 4'; see Table 2. The results are only reported for the procedures that control the false discovery rate. The power of each procedure is divided by the power for the oracle to yield an efficiency-like figure. For example, consider the configuration with $m_0 = 32$ and $m = 64$, where all procedures control the false discovery rate. The regular linear step-up has power of 0·873, and all FDR-controlling adaptive procedures raise the power to within the range of 0·924 to 0·977. It is clearly worth the extra effort to take the second stage in the two-stage procedure when m_0/m is as large as $\frac{1}{2}$. Table 2 shows that M-S-HLF is most powerful in all situations. When m_0/m is small ABH is almost as good. When all hypotheses are

Table 2: *Simulation study with independent test statistics. Power relative to the Oracle procedure for selected values of m_0 and m*

	$m_0/m = 0{\cdot}75$			$m_0/m = 0{\cdot}50$			$m_0/m = 0{\cdot}25$		
	$m = 16$	$m = 64$	$m = 256$	$m = 16$	$m = 64$	$m = 256$	$m = 16$	$m = 64$	$m = 256$
TST	0·956	0·958	0·959	0·917	0·924	0·926	0·862	0·865	0·865
M-TST	0·957	0·958	0·959	0·918	0·925	0·927	0·863	0·866	0·866
MST	0·964	0·967	0·969	0·927	0·935	0·938	0·878	0·888	0·890
ABH	0·968	0·976	0·975	0·946	0·953	0·947	0·906	0·917	0·900
M-S-HLF	0·973	0·989	0·993	0·958	0·977	0·982	0·923	0·949	0·956
LSU	0·945	0·942	0·941	0·874	0·873	0·874	0·781	0·777	0·775

false and m increases MST takes second place. In summary, the real gain seems to be in using a two-stage procedure. Which of these two-stage procedures one uses is of lesser significance.

6·3. *Positively dependent test statistics*

If m_0 is known, the linear step-up procedure controls the false discovery rate even under positive dependence, as expressed in Benjamini & Yekutieli (2001). Furthermore, if m_0 is estimated independently of the p-values used in LSU, if for example it is estimated from a different sample, then, by simply conditioning, we have that FDR $\leqslant qE(m_0/\hat{m}_0)$. Thus if $E(m_0/\hat{m}_0) \leqslant 1$ the two-stage procedure controls the false discovery rate.

To see how the bias and variance of $\hat{m}_0$ affects $E(m_0/\hat{m}_0)$ a straightforward Taylor series expression yields $E(m_0/\hat{m}_0) = 1 - \text{bias}/m_0 + \text{bias}^2/m_0^2 + \text{variance}/m_0^2$. From this we can see that, if the bias is positive, then that helps to meet the condition $E(m_0/\hat{m}_0) \leqslant 1$. If the bias is negligible, then the variance of the estimator plays a key role. For independent tests, the variance of the estimator is of order m_0 so that the variance term goes to zero. For dependent tests, the variance can also be of order m_0^2, in which case the size of the variance can have a large effect.

When the estimate of m_0 is independent of the p-values, even when the p-values themselves are dependent, the above argument led to an inequality, which implied that it is sufficient to show that $E(m_0/\hat{m}_0) \leqslant 1$ in order to obtain control of the false discovery rate. Once the same set of dependent p-values is used in both stages, two issues arise: the inequality on the bound might not hold, and furthermore $E(Q|\hat{m}_0) \leqslant (m/a)(m_0 q/m) = m_0 q/a$ need not hold. It is difficult to study analytically the combined effect that may cause the false discovery rate of the two-stage procedure to be higher than expected, and we therefore resort to a simulation study.

The simulation study allows us to explore the effect of constant positive dependence between the test statistics on the level of the false discovery rate achieved by the adaptive procedures. Figure 1 presents these results for $\rho = 0{\cdot}1$ and $0{\cdot}5$ in comparison to $\rho = 0$. Obviously the Oracle and LSU do control the false discovery rate, as follows from the theory in Benjamini & Yekutieli (2001), even though now at a level that is too low. The two-stage procedures control the false discovery rate well, below but close to the nominal level. All other procedures fail. The MST procedure controls the false discovery rate when the correlation is low, but fails at higher correlations. The MED-LSU procedure does better when m becomes larger. The value for the M-S-HLF procedure is sometimes more than twice the declared false discovery rate.

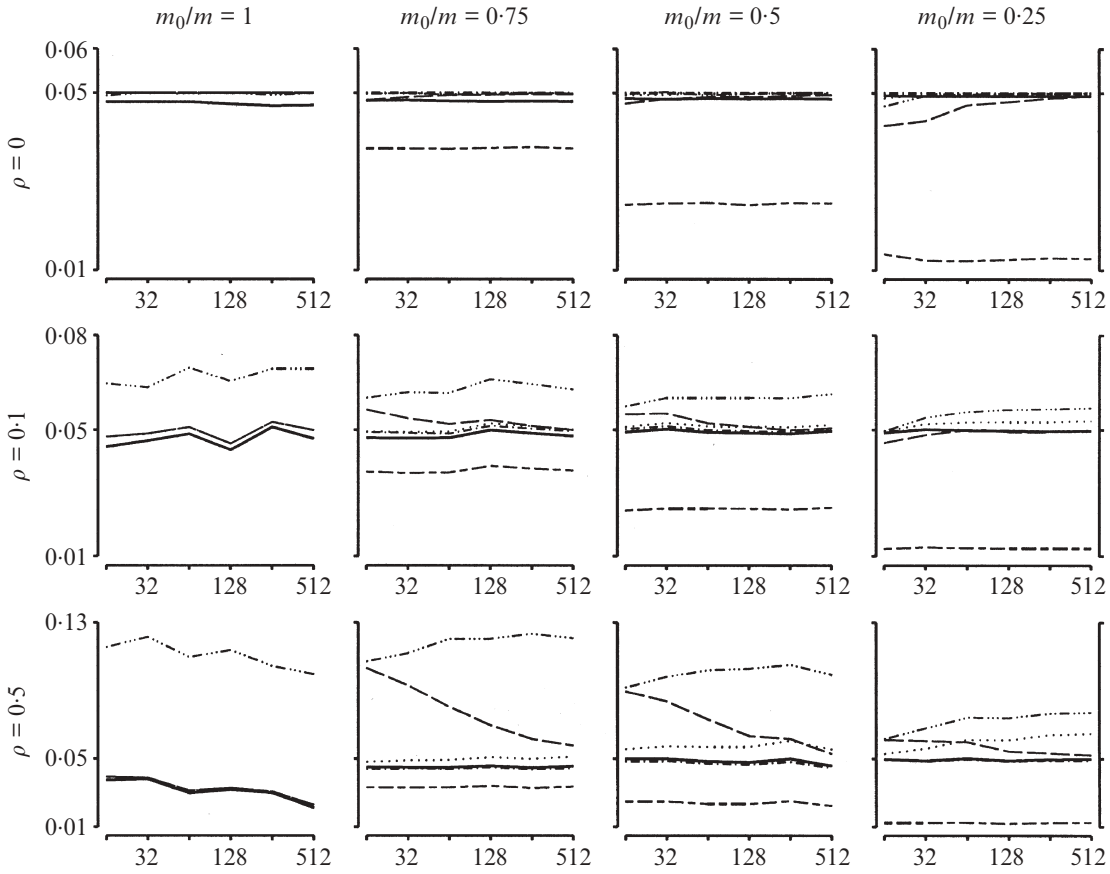

Fig. 1: Simulation study. Estimated FDR values for $m = 16, \ldots, 512$ and $\rho = 0$, 0·1, and 0·5. Results for procedure TST, solid line; MST, dotted line; ORC, dotted-dashed; ABH, dashed; M-S-HLF, dashed triple-dotted; LSU, short-dash long-dash.

Since many applications of the false discovery rate and theoretical results involve a large number of tests, simulations were also conducted for $m = 5000$, 10 000 and 15 000. The p-values were generated in the same way except that the values of μ were chosen to be $5i/(m - m_0)$ for $i = 1, \ldots, m - m_0$. In addition, the values 0, 0·01, 0·05 and 0·10 were used for the fraction of false null hypotheses. The results that appear in Fig. 2 indicate that it does not seem that the false discovery rate level gets closer to 0·05 as m increases for the M-S-HLF procedure.

How can this difference be explained? Figure 3 presents the distribution of the estimators of m_0 that are used in the TST procedure, ABH and M-S-HLF procedures, both under independence and under $\rho = 0{\cdot}5$. The figure shows that variability for M-S-HLF relative to that for either TST or ABH is only about two in the independence case. In the dependent case, however, this ratio increases to about ten; note that the biases are comparable. This results in M-S-HLF overshooting the nominal 0·05 false discovery rate level even for large m, stabilising at a level of 0·08, as is evident from Fig. 2. Note from Fig. 3 that more than a quarter of m_0 estimates obtained by M-S-HLF are above the maximal possible value of $m = 64$. Thus the deviation from the desired false discovery rate will be even greater if in practice $\min(\hat{m}_0, m)$ is used instead of $\hat{m}_0$.

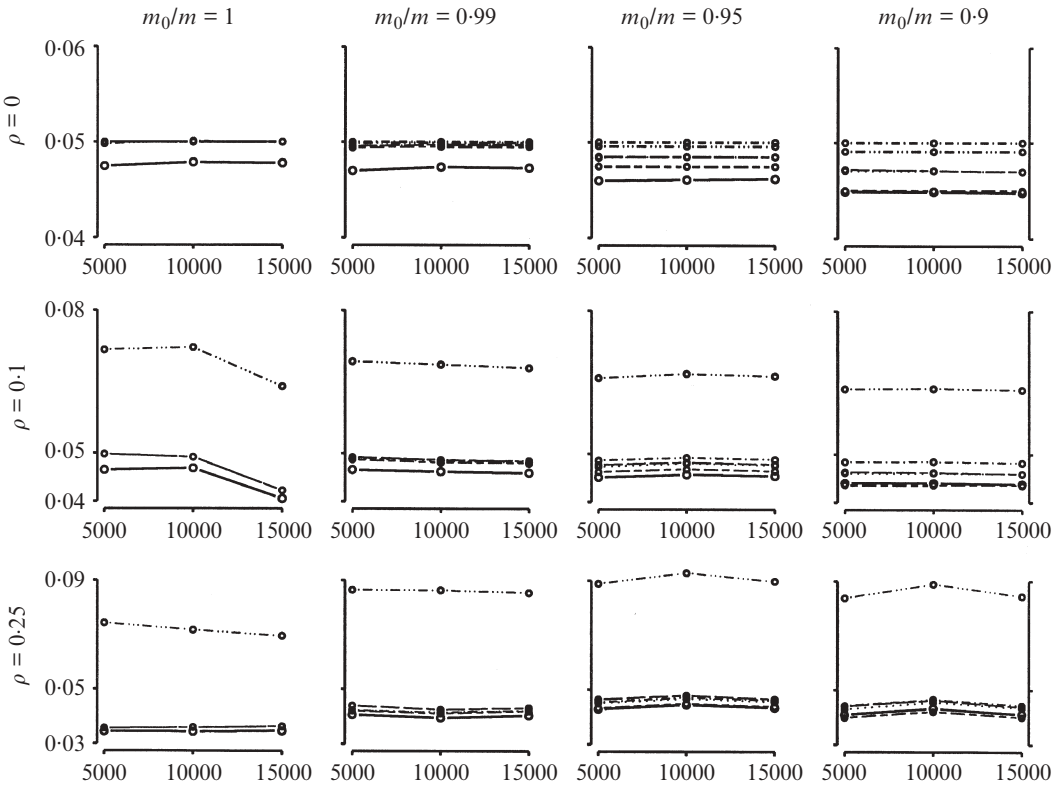

Fig. 2: Simulation study. Estimated FDR values for $m = 5000$, 10 000 and 15 000 and $\rho = 0$, 0·1 and 0·25. results for procedure TST, solid line; MST, dotted line; ORC, dotted-dashed; ABH, dashed; M-S-HLF, dashed triple-dotted; LSU, short-dash long-dash.

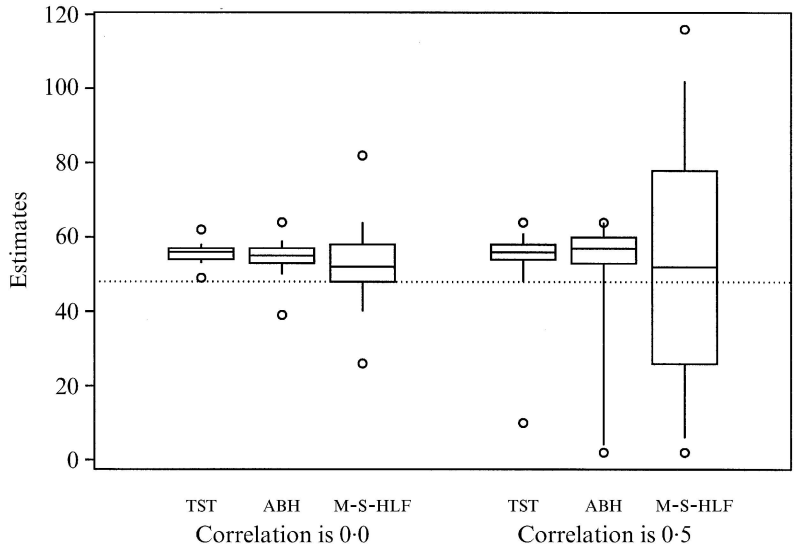

Fig. 3: The simulated distribution of the estimators $\hat{m}_0$ used in the TST, ABH and M-S-HLF adaptive procedures for estimating the number of true hypotheses with independent and positively correlated statistics for the case of $m_0 = 48$ and $m = 64$. Each box displays the median and quartiles as usual. The whiskers extend to the 5% and the 95% quantiles. The circles are located at the extremes, i.e. the 0·01% and 99·99% percentiles.

7. Examples

Example 1: *Multiple endpoints analysis.* Multiple endpoints analysis in clinical trials is one of the most commonly encountered multiplicity problems in medical research. This example on multiple endpoints illustrates the various procedures and shows the increased number of rejections when m_0 is estimated and accounted for as in our procedure. Since the data represent multiple measurements on the same individual, an individual's innate level can be viewed as latent. The assumption of constant positive dependence is arguably plausible, at least approximately. For the specific multiple problem described in detail in Benjamini & Hochberg (1995), the significance of the treatment effect on each of the 15 endpoints is given by the ordered $p_{(i)}$'s: 0·0001, 0·0004, 0·0019, 0·0095, 0·0201, 0·0278, 0·0298, 0·0344, 0·0459, 0·3240, 0·4262, 0·5719, 0·6528, 0·7590 and 1·000. Four hypotheses were rejected using the single linear step-up procedure at level 0·05. Four were also rejected at the first stage of the two-stage procedure run at level 0·05/1·05. At the second stage the linear step-up procedure is used at level $(0{\cdot}05/1{\cdot}05) \times 15/(15-4) = 0{\cdot}06494$, resulting in the rejection of the eight hypotheses whose p-values are less than or equal to 0·0344.

The multiple-stage procedure continues with the linear step-up procedure at level 0·0893, which is obtained from $0{\cdot}0515/\{15+1-8(1-0{\cdot}05)\}$. In this case, the ninth hypothesis with p-value of $0{\cdot}0459 \leqslant 0{\cdot}0893 \times \frac{9}{15}$ is also rejected. Interestingly all hypotheses with p-values less than 0·05 were rejected, as if no correction had been made.

Another interesting observation is that all adaptive procedures considered here rejected either 8 or 9 hypotheses; the procedures ABH, MST and MED-LSU resulted in 9 rejections. However, since positive dependence may be present among the measured endpoints, taking the more conservative finding of the two-stage procedure is recommended.

Example 2: *Quantitative trait loci analysis using false discovery rate.* Genetic researchers considered control of the false discovery rate in this important biological area (Weller et al., 1998), and Mosig et al. (2001) pioneered the use of adaptive procedures in the context of quantitative trait loci analysis of milk production in cattle, although without considering analysis of this particular example (Mosig et al., 2001). However, they did not consider the theoretical properties of their procedure. The purpose of their analysis is to identify regions on the chromosomes containing genes which affect the level of some quantitative property of the milk that a cow produces, such as volume, fat content or protein content. This kind of analysis is based on testing the statistical linkage between genetic markers on the chromosomes and the quantity of interest. Since molecular genetic markers can now be identified on a very dense map, the issue of multiple testing and its increased type I error probabilty is of fundamental concern. Lander & Kruglyak (1995) discuss this issue and set out guidelines that emphasise the genome-wise control of the familywise error rate, and Benjamini & Yekutieli (2005) established the appropriateness of the linear step-up procedure for this purpose. Their result relies on the positive regression dependence structure within chromosomes, inherent in the genetic problem.

Using the linear step-up procedure on single-site data, Mosig et al. (2001) identified 34 markers, out of a total number of 138, to be significant at the 0·05 level of false discovery rate. Using their original adaptive two-stage procedure they identified 8 additional significant markers. However, as was shown here, this procedure need not control the false discovery rate.

With the new two-stage procedure, the same 34 markers were rejected at the first stage at the 0·05/1·05 level. At the second stage the linear step-up procedure is used at level

$q^* = \frac{0{\cdot}05}{1{\cdot}05} \times \frac{138}{104} = 0{\cdot}063$. This increase identified a total of 37 markers, that is more than 34, but fewer than the 42 found by Mosig et al. (2001).

8. Discussion

Benjamini & Hochberg (2000) state, regarding adaptive procedures, that 'in cases where most of the hypotheses are far from being true there is hardly any penalty due to the simultaneous testing of many hypotheses'. As carefully analysed and explained by Black (2004), introducing the adaptive component into the linear step-up procedure is also the reason for the power advantage of the direct approach to false discovery rate of Storey et al. (2004). It is demonstrated here that the differences in power among the various adaptive procedures are much smaller than the differences among all adaptive procedures and the linear step-up procedure. Adaptive procedures that control familywise error have even less power. Of course these advantages are not realised when m_0/m is close to 1.

The results of the simulation study raise interesting issues. Some procedures are more sensitive to the nature of the correlation structure, equal positive correlations that we impose in our simulation study, and other approaches, most notably our new procedure, seem to perform well even in the correlated case. In practice, tests tend to be correlated. Understanding how different procedures that perform equally well for independent tests behave in correlated environments that reflect important applications is critical, but remains to be investigated.

Acknowledgement

This research was partly supported by the Focal Initiatives in Research Science and Technology foundation of the Israeli Academy of Sciences and Humanities, and by a grant from the U.S. National Institutes of Health. We would like to thank the referees for their helpful suggestions.

References

Abramovich, F., Benjamini, Y., Donoho, D. & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653.

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

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Behav. Educ. Statist.* **25**, 60–83.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

Benjamini, Y. & Yekutieli, D. (2005). The false discovery rate approach to quantitative trait loci analysis. *Genetics* **171**, 783–9.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *J. R. Statist. Soc.* B **66**, 297–304.

Efron, B., Tibshirani, R. J., Storey, J. D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.* **96**, 1151–60.

Genovese, C. & Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B **64**, 499–517.

Genovese, C. & Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–61.

Hochberg, Y. & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statist. Med.* **9**, 811–8.

Hsueh, H., Chen, J. J. & Kodell, R. L. (2003). Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J. Biopharm. Statist.* **13**, 675–89.

Lander, E. S. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–7.

Mosig, M. O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M. & Friedmann, A. A. (2001). Whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* **157**, 1683–98.

Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 239–57.

Schweder, T. & Spjøtvoll, E. (1982). Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64**, 479–98.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Statist.* **31**, 2013–35.

Storey, J. D. & Tibshirani, R. (2003a). Statistical significance for genome-wide studies. *Proc. Nat. Acad. Sci.* **100**, 9440–5.

Storey, J. D. & Tibshirani, R. (2003b). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software*, Ed. G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger, pp. 272–90. New York: Springer.

Storey, J. D., Taylor, J. E. & Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unifield approach. *J. R. Statist. Soc.* B **66**, 187–205.

Troendle, J. (1999). A permutational step-up method for testing multiple outcomes. *J. Statist. Plan. Infer.* **84**, 139–58.

Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A. & Ron, M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**, 1699–706.

Yekutieli, D. & Benjamini, Y. (1999). Resampling based false discovery rate controlling procedure for dependent test statistics. *J. Statist. Plan. Infer.* **82**, 171–96.

[*Received September* 2003. *Revised February* 2006]