Statistics and Probability Letters 87 (2014) 31–39

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Adaptive linear step-up multiple testing procedure with the bias-reduced estimator

Donggyu Kim*, Chunming Zhang

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA

ARTICLE INFO

Article history:
Received 15 June 2013
Received in revised form 2 January 2014
Accepted 2 January 2014
Available online 9 January 2014

Keywords:
False discovery rate
Multiple testing
Adaptive linear step-up procedure
Storey's estimator

ABSTRACT

This paper suggests two novel adaptive linear step-up procedures to reduce the bias of the estimator of π_0, the proportion of true null hypotheses. Estimators of π_0 are based on the number of p-values less than a threshold which converges to one.

1. Introduction

Since Benjamini and Hochberg (1995) introduced the concept of the false discovery rate (FDR), the FDR has been an important tool in multiple testing. Procedures based on the FDR obtained more power than those based on the family wise error rate (FWER). However, these were not the most powerful test in terms of the false non-discovery rate (FNR) which was discussed in the large sample case by Genovese and Wasswerman (2002). For this reason, Benjamini et al. (2006), Storey (2002) and Nettleton et al. (2006) introduced several adaptive linear step-up procedures.

Benjamini et al. (2006) introduced adaptive linear step-up procedures and established a more powerful test. Since the one stage linear step-up procedure at desired level q was actually a test at level $q\frac{m_0}{m}$, where m is the number of tests and m_0 is the number of true nulls, they tried to estimate the accurate level $q\frac{m_0}{m}$. Specifically, they suggested the two-stage procedure; the first stage for estimating $q\frac{m_0}{m}$ and the second stage for performing hypothesis tests using the one stage linear step-up procedure with the estimator of $q\frac{m_0}{m}$ from the first stage. They proposed the estimator of level $q\frac{m}{m_0}$ as follows:

$$\widehat{q} = q\frac{m}{\widehat{m}_0} \quad (= q/\widehat{\pi}_0),$$

where $\widehat{m}_0$ is the estimator of m_0, and estimated m_0 by using the following relationship

$$m_0 \leq m - (R - V), \tag{1.1}$$

where R is the number of rejections and V is the number of false rejections in the first stage. However, there were two problems when estimating m_0. One is to find a good predictor of V and the other is to obtain the closest estimator to m_0.

For the answer to the first problem, there were two approaches suggested by Benjamini et al. (2006). One is based on the classical rejection method which rejects the null hypotheses if p-value is less than a given critical value, λ. The other is

* Corresponding author. Tel.: +1 608 335 1370.
E-mail address: kimd@stat.wisc.edu (D. Kim).

0167-7152/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2014.01.001

Table 1
2×2 table of the m hypothesis tests classified by whether the null hypotheses are rejected or not and whether the tested null hypotheses are true or not.

	Accept H_0	Reject H_0	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

based on the one stage linear step-up procedure at level, λ. V is approximated as λm_0 by the classical rejection method and $\lambda R \frac{m_0}{m}$ by the one stage linear step-up procedure. Then, based on the classical rejection method, we obtain Storey's estimator (Storey, 2002) as follows:

$$\widehat{m}_0 = \frac{m - R}{1 - \lambda}. \tag{1.2}$$

Meanwhile, we can drive the two-stage linear step-up procedure (Benjamini et al., 2006) with the one stage linear step-up procedure. We will only focus on the classical rejection method in this paper.

The answer to the second problem is related to the derivation of the inequality (1.1). Usually, the set of the rejected hypotheses does not contain all of the alternative hypotheses and it causes the inequality (1.1). In other words, when λ is close to 1, all of the alternative hypotheses would be included in the set of the rejected hypotheses. It implies that the bias of $\widehat{m}_0$ becomes smaller. However, the variance of $\widehat{m}_0$ typically becomes larger. Liang and Nettleton (2012) discussed this problem and showed that their dynamically adaptive procedure (RB20*) was improved in terms of MSE. We suggest procedures that allow λ to go to one with an appropriate convergence rate. The simulation study shows that our novel procedures with appropriate tuning parameters are superior to pre-existing procedures in terms of FDR and power. Furthermore, when the effect size is small, our method is even better than RB20* in terms of MSE.

The rest of the paper is organized as follows. We introduce new adaptive linear step-up procedures and explain the connection with pre-existing procedures in Section 2. Then, we show how these methods control the FDR and behave asymptotically in Section 3. A simulation study is presented in Section 4 to compare our procedures with pre-existing procedures in case where the test statistics are dependent as well as independent. The paper concludes with the discussion in Section 5.

2. Method

2.1. Terminology

Before introducing new procedures, we define several terminologies. Let H_i denote the tested null hypothesis and let $H_i = 0$ (or 1) if the ith tested null hypothesis is true (or false) for $i = 1, \ldots, m$. Let G_1 and G_0 be the set of the alternative hypotheses ($H_i = 1$) and the set of the null hypotheses ($H_i = 0$), respectively. $|G_1| = m_1$ and $|G_0| = m_0$. We can categorize the m number of tests as in Table 1 and then π_0 $(= m_0/m)$ is defined as the proportion of the null hypotheses. Let P_i denote the ith p-value, $i = 1, \ldots, m$, and $P_{(1)} < \cdots < P_{(m)}$ be the ordered p-values.

2.2. Fixed linear step-up procedure

For Storey's estimator (1.2) to be unbiased, it needs to satisfy

$$P(G_1 \subset G_R(\lambda)) = 1, \tag{2.1}$$

where $G_R(\lambda) = \{i \in \{1, \ldots, m\} : P_i \leq \lambda\}$. $G_R(\lambda)$ is the set of the rejected hypotheses at a given threshold, λ. When Eq. (2.1) is true, the inequality (1.1) will be an equality almost surely. This indicates that as long as G_R contains G_1, we will be able to achieve the closest estimator to m_0. However, G_R does not usually contain G_1.

To see the case where G_1 is contained in G_R, consider the case of a trivial threshold. When the threshold λ is equal to 1, $G_R(1)$ becomes the whole set and one has $G_1 \subset G_R$. This trivial threshold, however, is inappropriate because we cannot estimate m_0 when $\lambda = 1$. So, we consider $\lambda = 1 - \epsilon$ instead. In other words, we suggest the threshold λ_m such that

$$\lambda_m \nearrow 1 \quad \text{as } m \to \infty. \tag{2.2}$$

The performance of this estimator depends on the convergence rate of λ_m.

To investigate the convergence rate of λ_m, we borrow the model setting, called the random effects (or hierarchical) model, from Efron et al. (2001) and Genovese and Wasswerman (2004). Assume that for $0 \leq \pi_0 \leq 1$:

$(H_i, \Xi_i, P_i)'s$ are independent with H_i, the ith hypothesis, and P_i, the ith p-value;

$H_1, \ldots, H_m \sim \text{Bernoulli}(1 - \pi_0)$;

$\Xi_1, \ldots, \Xi_m \sim L_{\mathbb{F}}$;

$$P_i|H_i = 0,\ \Xi_i = \xi_i \sim \text{Uniform}(0, 1);$$
$$P_i|H_i = 1,\ \Xi_i = \xi_i \sim \xi_i;$$
$$P_i \sim G, \quad \text{with } G(t) = \pi_0 t + (1 - \pi_0)F(t) \text{ and } F(t) = \int \xi(t)dL_{\mathbb{F}},$$

where $\Xi_1, \ldots, \Xi_m$ denote distribution functions and $L_{\mathbb{F}}$ is an arbitrary probability measure over a class of distribution functions $\mathbb{F}$ that is stochastically dominated by the Uniform$(0, 1)$ (i.e. $t \le P(\xi \le t)$ for all t and $t < P(\xi \le t)$ for some t) and G is strictly concave with density $g = G'$.

Under this model,

$$g(t) = \pi_0 + (1 - \pi_0)f(t) \ge \pi_0, \tag{2.3}$$

where $f(t) = F'(t)$. Furthermore, since G is strictly concave,

$$g(1) = \min_{t\in[0,1]} g(t). \tag{2.4}$$

If we let $\pi_g \equiv g(1) \ge \pi_0$, then $\pi_g \ge \pi_0$ by (2.3) and (2.4) and it can be estimated by a kernel function

$$\widehat{\pi}_{g,\,K} = \int_{-\infty}^{\infty} \frac{1}{b} K\left(\frac{1-u}{b}\right) d\widehat{G}(u),$$

where $\widehat{G}$ is an empirical CDF and $b = 1 - \lambda_m$. We set the kernel function $K(t) = \mathrm{I}(|t| \le 1)$, where $\mathrm{I}(\cdot)$ denotes an indicator function. The way $K(\cdot)$ is defined is a little bit different from the traditional definition because we need to deal with the fact that $G(1) = 1$. Then,

$$\widehat{\pi}_{g,\mathrm{I}(|t|\le 1)} = \widehat{\pi}_g = \frac{\widehat{G}(1+b) - \widehat{G}(1-b)}{b} = \frac{1 - \widehat{G}(\lambda_m)}{1 - \lambda_m} = \frac{1 - \#\{P_i \le \lambda_m\}/m}{1 - \lambda_m}. \tag{2.5}$$

Remark 1. The kernel function $K(\cdot)$ to estimate π_g can be defined in different ways.

When λ_m is fixed, this is exactly the same as the well-known Storey's estimator (Storey, 2002) which is based on the estimator of G. However, for fixed m, Storey's estimator does not control the FDR and so, it is modified later in Storey et al. (2004). We also consider the modified version,

$$\widehat{\pi}_g^* = \frac{m - \#\{P_i \le \lambda_m\} + 1}{m(1 - \lambda_m)}.$$

The performance of this estimator depends on the convergence rate of λ_m. We suggest

$$\lambda_m = 1 - c \cdot m^{-1/3}/\log\{\log(m)\},$$

where c is a positive constant number. Using this estimator, we define the new adaptive linear step-up procedure.

Definition 1 (*P-LSU1 (Proposed Linear Step-Up Procedure 1)*). Step 1. Estimate $\widehat{m}_0 = m \cdot \widehat{\pi}_g^*$.

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

We discuss the asymptotic behaviors and the convergence rate of b in Section 3.

2.3. Dynamically adaptive procedure

Similarly to (2.2), consider the following procedure

$$\widehat{\pi}_0 = \frac{m + 1 - (m - \tau_m)}{m\{1 - P_{(m-\tau_m)}\}} = \frac{\tau_m + 1}{m\{1 - P_{(m-\tau_m)}\}},$$

where $\tau_m \to \infty$ and $\tau_m/m \to 0$ as $m \to \infty$. Note that when τ_m is fixed, $\widehat{\pi}_0$ is exactly the same as the quantile adaptive linear step-up procedure in Benjamini et al. (2006). The performance of this estimator relies on the convergence rate of τ_m. We suggest

$$\tau_m = \lfloor m^{\alpha} \rfloor,$$

where $\alpha \in (0, 1)$. Then, the proportion of τ_m to m goes to zero as $m \to \infty$. Based on this estimator, we define another new adaptive linear step-up procedure.

Definition 2 (*P-LSU2 (Proposed Linear Step-Up Procedure 2)*). Step 1. Estimate $\widehat{m}_0 = \frac{\tau_m+1}{1-P_{(m-\tau_m)}}$.

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

3. Analytical result

In this section, we investigate how our proposed procedures control the FDR and behave asymptotically.

3.1. *Controlling the* FDR

For fixed m, the procedures in Definitions 1 and 2 are the same as the modified version of Storey's estimator (Storey et al., 2004) and the quantile adaptive linear step-up procedure (Benjamini et al., 2006), respectively. In light of this, we can derive the following two theorems.

Theorem 1. *If the test statistics are independent, the adaptive linear step-up procedure in Definition 2 controls the false discovery rate at level q.*

Proof. Without loss of generality, let H_1 come from G_0 and let P_1 be the p-value associated with H_1. Define $P^{(1)}$ to be the vector of p-values corresponding to the $(m-1)$ hypotheses excluding H_1. Given $P^{(1)}$, let $\ell(P_1)$ be the indicator that H_1 is rejected using P-LSU2, as a function of P_1. Then, it is enough to show that $E_{P^{(1)}}\{\frac{m_0}{\widehat{m}_0(p^*)}\} \le 1$ by Eq. (5) in Benjamini et al. (2006), where $p^* = p^*(P^{(1)})$ such that $\ell(P_1) = \mathrm{I}(P_1 \le p^*)$. In case of P-LSU2, since $\widehat{m}_0$ does not depend on p^*, $\widehat{m}_0(p^*) = \frac{\tau_m+1}{1-P_{(m-\tau_m)}}$. If $k \le m_1+1$, where $k = m - \tau_m$,

$$E_{P^{(1)}}\left\{\frac{m_0}{\widehat{m}_0(p^*)}\right\} = \frac{m_0}{m-k+1}E_{P^{(1)}}\{1-P_{(k)}\} \le 1.$$

If $k > m_1+1$, at least $k-m_1-1p$-values from $G_0 \setminus \{H_1\}$ are less than or equal to $P_{(k)}$. Let $\widetilde{P}_i$'s be the p-values from $G_0 \setminus \{H_1\}$. Then, $\widetilde{P}_{(k-m_1-1)} \le P_{(k)}$, which implies that

$$\begin{aligned} E_{P^{(1)}}\left\{\frac{m_0}{\widehat{m}_0(p^*)}\right\} &= \frac{m_0}{m-k+1}E_{P^{(1)}}\{1-P_{(k)}\} \\ &\le \frac{m_0}{m-k+1}E_{P^{(1)}}\{1-\widetilde{P}_{(k-m_1-1)}\} \\ &= \frac{m_0}{m-k+1}\left\{1-\frac{k-m_1-1}{m_0}\right\} = 1. \end{aligned}$$

The third equality follows from the fact that $\widetilde{P}_{(k-m_1-1)} \sim \mathrm{Beta}(k-m_1-1, m-k+1)$. ■

Note that the same result as Theorem 1 has been presented in Theorem 2 of Benjamini et al. (2006). However, their proof contained a minor error and here we offer a correction. For example, in their proof, when $k = m_1+1$, the statement does not hold. The following theorem has been showed by Storey et al. (2004, Theorem 3). We prove this again in the light of the approach in Benjamini et al. (2006) as they mentioned in the paper.

Theorem 2. *When the test statistics are independent, the adaptive linear step-up procedure in Definition 1 controls the false discovery rate at level q.*

Proof. We use the same setting as the proof of Theorem 1. Let $\lambda_m = \lambda$.

$$\begin{aligned} E_{P^{(1)}}\left\{\frac{m_0}{\widehat{m}_0(p^*)}\right\} &\le m_0 E_{P^{(1)}}\left\{\frac{1-\lambda}{\#\left\{P^{(1)} > \lambda\right\}+1}\right\} \le m_0\frac{1-\lambda}{E_{P^{(1)}}\#\left\{P^{(1)} > \lambda\right\}+1} \\ &\le m_0\frac{1-\lambda}{E_{P^{(1)}}\#\{\widetilde{P}_i > \lambda\}+1} \le 1. \end{aligned}$$

The second inequality holds by Jensen's inequality. ■

Theorems 1 and 2 show that our proposed procedures control the FDR for any fixed m.

3.2. *Asymptotic results*

From asymptotic theorems for kernel density estimation (see, e.g. Shao, 2003, Chapter 5.1), we can obtain the consistency of $\widehat{\pi}_g$.

Lemma 1. *Let $b \to 0$ and $m \cdot b \to \infty$. If $g(t)$ is left continuous at $t = 1$, then $\widehat{\pi}_g - \pi_g = o_p(1)$.*

Proof. By Taylor's expansion, $g(1) = \frac{G(1)-G(1-b)}{b} + o_p(1)$, which implies that $\widehat{\pi}_g - g(1) = \frac{G(1-b)-\widehat{G}(1-b)}{b} + o_p(1) = O_p(\frac{1}{\sqrt{mb}}) + o_p(1) = o_p(1)$. ■

Combining this result and Lemma 2 below, we can show that the procedure based on $\widehat{\pi}_q$ in (2.5) controls the FDR at level q, asymptotically.

Lemma 2. *Assume that G is concave and $\widehat{\pi}_0 \xrightarrow{p} \pi_0^*$ for some $\pi_0^* > \pi_0$. If we let $T = \sup_{t\in[0,1]}\{t : \widehat{\pi}_0 t/\widehat{G}(t) \le q\}$, then*

$$E\left\{\mathrm{FDP}(T)\right\} \le q + o(1).$$

Proof. This follows from Theorem 5.2 in Genovese and Wasswerman (2002). ■

If $g(t)$ is smooth enough, we can obtain the asymptotic distribution.

Theorem 3. *If the test statistics are independent, $g(t)$ is left continuously differentiable at $t = 1$ and $mb^3 \to 0$,*

$$\sqrt{mb}(\widehat{\pi}_g - \pi_g) \xrightarrow{d} N(0, g(1)). \tag{3.1}$$

Proof. By Taylor's expansion,

$$\sqrt{mb}\{\widehat{\pi}_g - g(1)\} = \sqrt{m}\left\{\frac{G(1-b) - \widehat{G}(1-b)}{\sqrt{b}}\right\} + O_p(\sqrt{mb^3}).$$

If $\sqrt{m}\{\frac{G(1-b)-\widehat{G}(1-b)}{\sqrt{b}}\} \xrightarrow{d} N(0, g(1))$, (3.1) holds. Now, we check Lindeberg's condition. Let $X_{m,i} = \frac{1}{\sqrt{b}}\{\mathrm{I}\,(P_i \le 1-b) - G(1-b)\}$ and $\sigma_m^2 = \mathrm{Var}(\sum_{i=1}^m X_{m,i}) = m \cdot g(1) + O_p(mb)$. For any $\epsilon > 0$,

$$\begin{aligned}
\frac{1}{\sigma_m^2}\sum_{i=1}^{m} E\left\{X_{m,i}^2 \mathrm{I}\left(X_{m,i}^2 > \epsilon\sigma_m^2\right)\right\} &= \frac{1}{g(1) + O_p(b)} E\left\{X_{m,1}^2 \mathrm{I}\left(X_{m,1}^2 > \epsilon\sigma_m^2\right)\right\} \\
&= \frac{1}{g(1) + O_p(b)} \\
&\quad \times \int_{|\mathrm{I}(x\le -1)-G(1-b)|>\sqrt{\epsilon\sigma_m^2}} \{\mathrm{I}\,(x \le -1) - G(1-b)\}^2 g(bx+1)dx \\
&\to 0 \quad \text{as } m \to \infty \text{ by dominated convergence theorem.}
\end{aligned}$$

Thus, by Lindeberg's central limit theorem, $\sqrt{m}\{\frac{G(1-b)-\widehat{G}(1-b)}{\sqrt{b}}\} \xrightarrow{d} N(0, g(1))$. ■

Remark 2. Even though we choose different kernel functions, if it is appropriate, all of the arguments above will continue to hold.

To obtain the asymptotic normal distribution, Theorem 3 tells us that λ_m should satisfy

$$m(1-\lambda_m)^3 \to 0 \quad \text{and} \quad m(1-\lambda_m) \to \infty.$$

In this point of view, as we mentioned in Section 2, we suggest $\lambda_m = 1 - c \cdot m^{-1/3}/\log\{\log(m)\}$, where c is a positive constant.

4. Simulation study

4.1. Candidate procedures

Simulation study is performed to compare several adaptive linear step-up procedures with the procedures (P-LSU1 and P-LSU2) proposed in this paper. We considered the following procedures:

(a) P-LSU1$^{(c)}$, the proposed linear step-up procedure in Definition 1 with a constant, c;
(b) S-HLF, the adaptive linear step-up procedure proposed by Storey (2002) with $\lambda = 0.5$;
(c) M-S-HLF, the modified version of Storey's estimator in Storey et al. (2004);
(d) TST, the two stage linear step-up procedure by Benjamini et al. (2006);
(e) P-LSU2$^{(\alpha)}$, the proposed linear step-up procedure in Definition 2 with the decaying rate, α;
(f) RB20*, the right boundary procedure proposed by Liang and Nettleton (2012) with a candidate set for λ, $\Lambda = \{0.02, 0.04, \ldots, 0.10, 0.15, 0.2, \ldots, 0.95\}$;
(g) MED, the median adaptive procedure proposed by Benjamini and Hochberg (2000);
(h) ABH, the adaptive linear step-up procedure of Benjamini and Hochberg (2000);
(i) ORC, the linear step-up procedure at level qm/m_0 (ORACLE).

P-LSU2, RB20*, MED, and ABH are dynamically adaptive procedures. All other procedures except ORC are fixed adaptive procedures. All nine procedures control the FDR for fixed m or asymptotically.

For the first part of study, the simulation was repeated 10 000 times. The standard error of the estimated values was the order of 0.002 or less. The target level of the FDR, q, was 0.05. The p-values were generated in the following way. The null

Table 2
Simulation study with independent test statistics. Estimates of the FDR from 8 different procedures are compared for 4 different levels of m and 4 different labels of m_0/m.

m	$m_0/m = 0.25$				$m_0/m = 0.50$				$m_0/m = 0.75$				$m_0/m = 1$			
	64	512	4096	20 000	64	512	4096	20 000	64	512	4096	20 000	64	512	4096	20 000
P-LSU1$^{(1)}$	0.044	0.047	0.047	0.048	0.049	0.049	0.049	0.049	0.051	0.050	0.050	0.050	0.057	0.052	0.048	0.050
P-LSU1$^{(2)}$	0.042	0.045	0.047	0.048	0.048	0.048	0.049	0.049	0.050	0.050	0.050	0.050	0.054	0.051	0.047	0.049
P-LSU1$^{(3)}$	0.040	0.044	0.046	0.047	0.047	0.048	0.049	0.049	0.049	0.049	0.050	0.050	0.054	0.051	0.047	0.049
P-LSU1$^{(4)}$	0.036	0.043	0.046	0.047	0.045	0.047	0.048	0.049	0.048	0.049	0.050	0.050	0.052	0.051	0.047	0.049
P-LSU1$^{(5)}$	0.030	0.042	0.045	0.046	0.042	0.047	0.048	0.049	0.048	0.049	0.049	0.050	0.051	0.050	0.047	0.049
S-HLF	0.044	0.040	0.039	0.039	0.050	0.046	0.046	0.046	0.051	0.049	0.049	0.048	0.055	0.050	0.046	0.049
M-S-HLF	0.040	0.039	0.039	0.039	0.047	0.046	0.046	0.046	0.049	0.049	0.049	0.048	0.054	0.050	0.046	0.049
TST	0.023	0.023	0.022	0.022	0.035	0.034	0.034	0.034	0.042	0.041	0.041	0.041	0.049	0.047	0.044	0.047
P-LSU2$^{(0.1)}$	0.044	0.048	0.049	0.049	0.051	0.051	0.051	0.051	0.056	0.055	0.055	0.054	0.064	0.057	0.058	0.058
P-LSU2$^{(0.2)}$	0.044	0.048	0.048	0.049	0.051	0.050	0.050	0.050	0.056	0.054	0.053	0.052	0.064	0.057	0.054	0.054
P-LSU2$^{(0.4)}$	0.043	0.046	0.047	0.048	0.049	0.049	0.049	0.050	0.052	0.051	0.050	0.050	0.059	0.054	0.050	0.051
P-LSU2$^{(0.5)}$	0.041	0.045	0.047	0.048	0.048	0.049	0.049	0.049	0.051	0.050	0.050	0.050	0.057	0.053	0.048	0.050
P-LSU2$^{(0.6)}$	0.038	0.043	0.045	0.047	0.048	0.048	0.049	0.049	0.050	0.050	0.050	0.050	0.056	0.052	0.048	0.050
P-LSU2$^{(0.8)}$	0.026	0.033	0.038	0.040	0.044	0.046	0.047	0.047	0.049	0.049	0.049	0.049	0.054	0.050	0.047	0.049
P-LSU2$^{(0.9)}$	0.019	0.023	0.027	0.030	0.037	0.041	0.043	0.044	0.048	0.048	0.048	0.049	0.052	0.050	0.047	0.049
RB20*	0.035	0.039	0.043	0.045	0.044	0.044	0.046	0.048	0.048	0.047	0.048	0.049	0.052	0.049	0.046	0.049
MED	0.025	0.024	0.024	0.024	0.044	0.042	0.042	0.042	0.050	0.048	0.048	0.048	0.052	0.050	0.046	0.049
ABH	0.033	0.027	0.024	0.023	0.042	0.037	0.035	0.035	0.047	0.044	0.043	0.043	0.051	0.049	0.046	0.049
ORC	0.050	0.050	0.050	0.050	0.051	0.050	0.050	0.050	0.051	0.050	0.050	0.050	0.050	0.049	0.046	0.049

Table 3
Simulation study with independent test statistics. Estimates of power from 8 different procedures are compared for 4 different levels of m and 3 different labels of m_0/m.

m	$m_0/m = 0.25$				$m_0/m = 0.50$				$m_0/m = 0.75$			
	64	512	4096	20 000	64	512	4096	20 000	64	512	4096	20 000
P-LSU1$^{(1)}$	0.7656	0.7866	0.7944	0.7977	0.6483	0.6559	0.6596	0.6604	0.5357	0.5343	0.5352	0.5356
P-LSU1$^{(2)}$	0.7685	0.7868	0.7932	0.7964	0.6497	0.6562	0.6592	0.6600	0.5344	0.5347	0.5351	0.5356
P-LSU1$^{(3)}$	0.7627	0.7844	0.7917	0.7952	0.6480	0.6555	0.6587	0.6596	0.5335	0.5347	0.5350	0.5355
P-LSU1$^{(4)}$	0.7521	0.7815	0.7902	0.7941	0.6437	0.6547	0.6582	0.6593	0.5326	0.5344	0.5349	0.5354
P-LSU1$^{(5)}$	0.7320	0.7782	0.7887	0.7931	0.6357	0.6535	0.6578	0.6590	0.5300	0.5341	0.5348	0.5353
S-HLF	0.7778	0.7715	0.7702	0.7702	0.6559	0.6515	0.6513	0.6513	0.5381	0.5339	0.5329	0.5329
M-S-HLF	0.7638	0.7699	0.7700	0.7702	0.6484	0.6506	0.6512	0.6512	0.5336	0.5333	0.5329	0.5329
TST	0.6961	0.6959	0.6957	0.6956	0.6133	0.6127	0.6133	0.6132	0.5170	0.5150	0.5142	0.5144
P-LSU2$^{(0.1)}$	0.7674	0.7768	0.7797	0.7854	0.6497	0.6523	0.6513	0.6541	0.5425	0.5410	0.5411	0.5396
P-LSU2$^{(0.2)}$	0.7674	0.7812	0.7892	0.7928	0.6497	0.6526	0.6542	0.6560	0.5425	0.5396	0.5381	0.5373
P-LSU2$^{(0.4)}$	0.7700	0.7874	0.7944	0.7980	0.6505	0.6558	0.6591	0.6603	0.5391	0.5358	0.5352	0.5354
P-LSU2$^{(0.5)}$	0.7668	0.7859	0.7932	0.7968	0.6505	0.6563	0.6595	0.6604	0.5376	0.5351	0.5352	0.5356
P-LSU2$^{(0.6)}$	0.7598	0.7809	0.7896	0.7937	0.6499	0.6559	0.6590	0.6599	0.5360	0.5349	0.5351	0.5356
P-LSU2$^{(0.8)}$	0.7158	0.7482	0.7651	0.7736	0.6417	0.6500	0.6541	0.6558	0.5336	0.5340	0.5343	0.5347
P-LSU2$^{(0.9)}$	0.6735	0.6981	0.7189	0.7320	0.6226	0.6363	0.6437	0.6471	0.5308	0.5318	0.5324	0.5331
RB20*	0.7474	0.7689	0.7833	0.7901	0.6412	0.6462	0.6530	0.6563	0.5304	0.5297	0.5320	0.5334
MED	0.7071	0.7060	0.7055	0.7055	0.6420	0.6395	0.6395	0.6395	0.5365	0.5327	0.5317	0.5318
ABH	0.7413	0.7188	0.7063	0.7003	0.6343	0.6245	0.6195	0.6167	0.5283	0.5228	0.5198	0.5188
ORC	0.8053	0.8042	0.8037	0.8038	0.6639	0.6624	0.6624	0.6624	0.5398	0.5372	0.5363	0.5363

distributions were identically $N(0, 1)$ and the alternative distributions were $N(\mu_i, 1)$, where $\mu_i = i$ for $i = 1, 2, 3$ and 4. This cycle was repeated to produce the desired m_1 values under H_1. The number of tests, m, was set at 64, 512, 4096 and 20 000 and the fraction of null hypotheses, $\pi_0 (= m_0/m)$, was 25%, 50%, 75% and 100%. We assumed that the test statistics are independent and p-values were calculated as $P_j = 1 - \Phi(Y_j)$ for $j = 1, \ldots, m$, where Y_j's come from the null distribution for $i = 1, \ldots, m_0$ and the alternative distribution for $i = m_0 + 1, \ldots, m$. To investigate sensitivity to the choice of tuning parameters (c and α), for P-LSU1, we chose $c = 1, 2, 3, 4,$ and 5 and the decaying rate, α, was set at 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, and 0.9 on P-LSU2.

4.2. Simulation result for independent tests

Table 2 shows that our two proposed procedures control the FDR well so that their estimates are closer to the target level $q = 0.05$ than others except the case where $\alpha = 0.1, 0.2, 0.8$, and 0.9. The case of $m_0/m = 1$ seems to be an exception, but when m is big enough, they control the FDR well again. Note that when the decaying rate, α, is too small ($\alpha = 0.1, 0.2$), P-LSU2 has slightly bigger FDR in case of $m_0/m = 1$. In Table 3, we compared the power among the procedures. Storey's procedure performs well when $m = 64$ and its power is stable even though the sample size m changes because of the

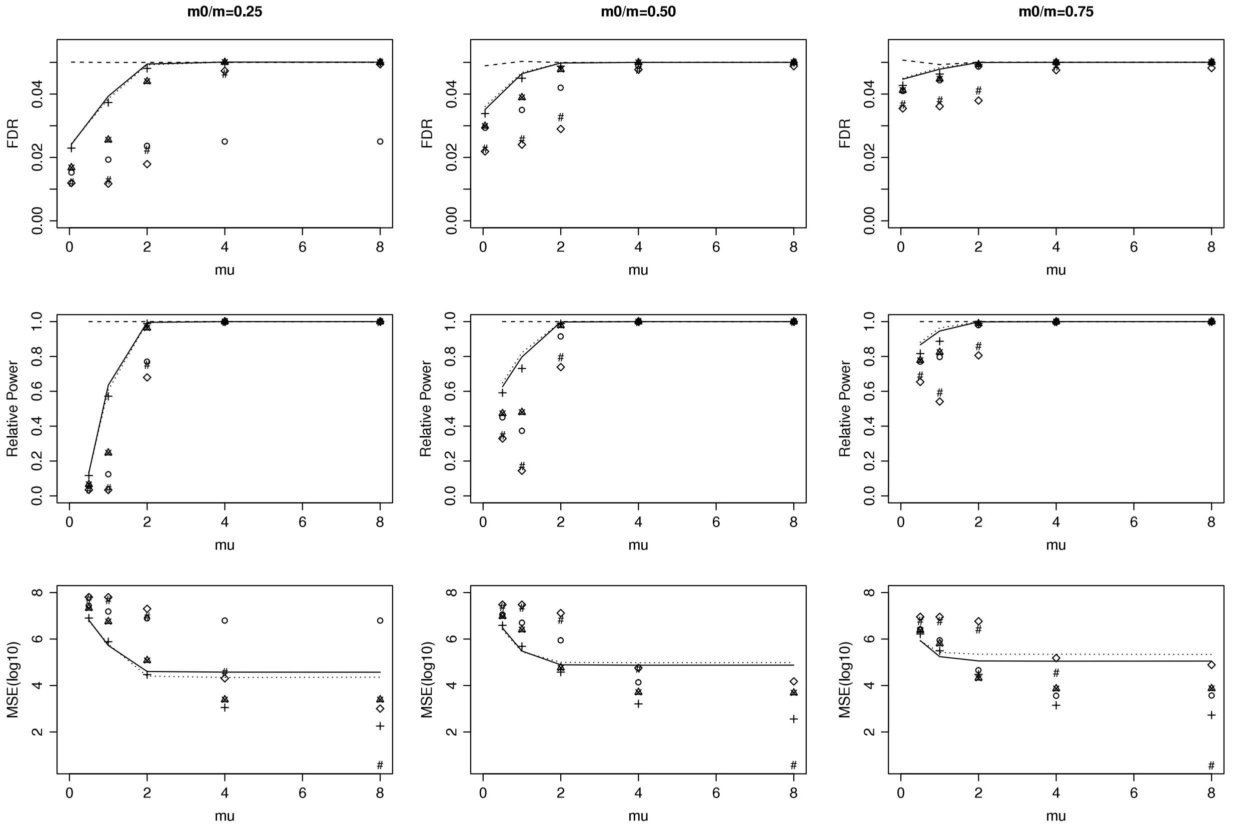

Fig. 1. Under the independence assumption, 9 procedures are compared in terms of FDR, relative power and MSE of m_0. Solid line: P-LSU1; dotted line: P-LSU2; dashed line: ORC; $\times$: S-HLF; $\triangle$: M-S-HLF; $\diamond$: TST; $+$: RB20*; $\circ$: MED; #: ABH.

fixed choice of λ. The proposed procedures perform better for large m. When $m = 20\,000$, the performance of P-LSU1 with $c = 1, 2, 3, 4$ and P-LSU2 with $\alpha = 0.4, 0.5, 0.6$ are better than that of any other procedures. This is a predictable result because when m is small, the asymptotic variance of the estimator of m_0 is expected to be big, but when m is big enough, the affect of variance is small.

When $m = 64$, the choice of tuning parameter, c, influences on the performance of P-LSU1. For example, when $c = 4$ or 5, the power and FDR of P-LSU1 are relatively worse. However, the performance of P-LSU1 becomes robust to the choice of the tuning parameter, when m is big enough. For P-LSU2, when the decaying rate, α, is too small or too big ($\alpha = 0.1, 0.2, 0.8, 0.9$), the performance of P-LSU2 is not good. On the other hand, when we choose α between 0.4 and 0.6, P-LSU2 is robust to the choice of α and works well. Based on this sensitivity study, we suggest to choose c between 1 and 3, and α between 0.4 and 0.6.

4.3. *Simulation study for comparing* MSE

4.3.1. *Simulation result for independent tests*

To investigate MSE, another simulation study was carried out with the same setting as the one conducted by Liang and Nettleton (2012). Specifically, the simulation was repeated 10 000 times and the number of tests, m, was 10 000. The fraction of null hypotheses, m_0/m, was set to be 25%, 50% and 75%. The null distribution and the alternative distribution were $N(0, 1)$ and $N(\mu, 0)$, respectively. For the alternative distribution, μ was set at $\mu = 0.5, 1, 2, 4$ and 8 to cover the range of all possible effect sizes. When the effect size is small ($\mu = 0.5$), the alternative p-values are well mixed with the null p-values. On the other hand, when effect size is large ($\mu = 8$), the alternative p-values are well-separated from the null p-values. The tuning parameters were set at $c = 3$ and $\alpha = 0.6$.

In Fig. 1, MSE of P-LSU1 and P-LSU2 is smaller than those of any others when the possible effect size is small ($\mu = 0.5$ or 1). This is because when the effect size is small, the bias influences on the results a lot. In terms of FDR and power, P-LSU1 and P-LSU2 are similar and are the most powerful and closest to the desired level 0.05 over all 5 different levels of μ. This shows that even though P-LSU1 and P-LSU2 are not the best in terms of MSE for the large effect size, these procedures are superior in terms of power.

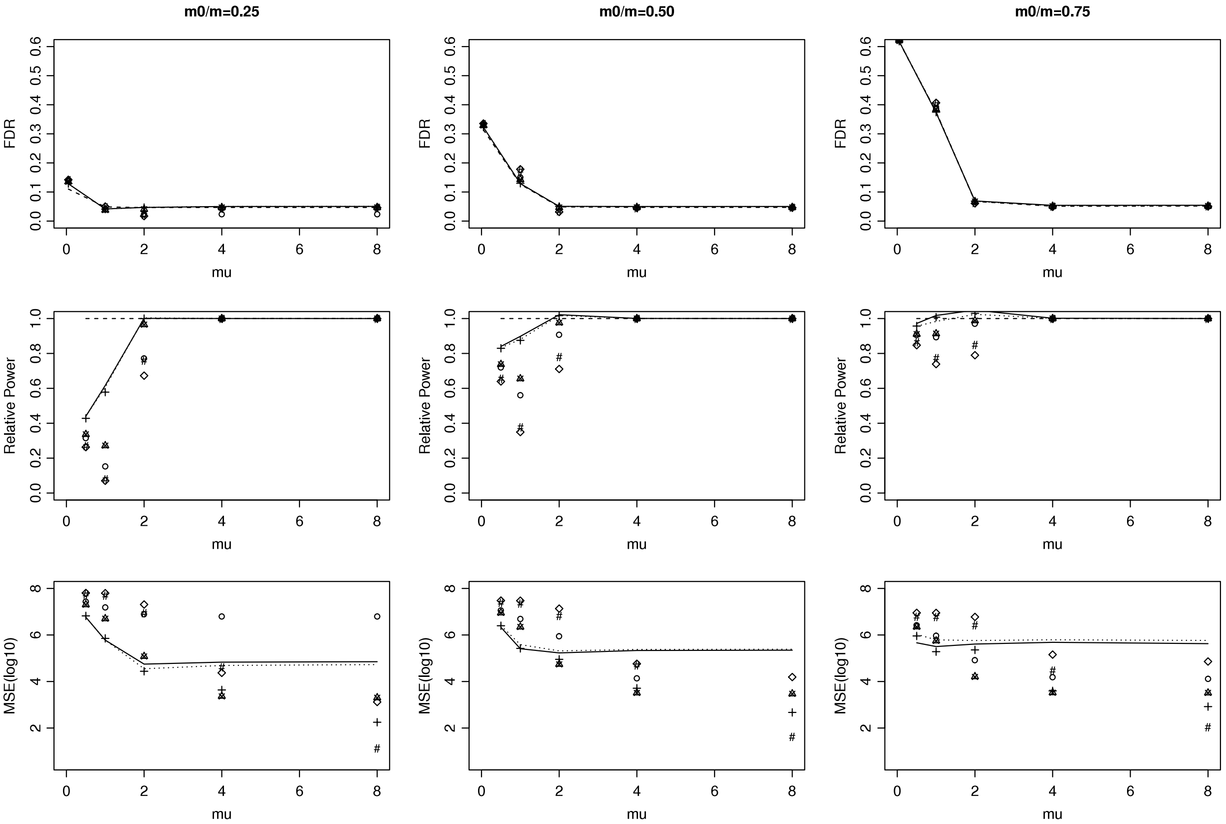

Fig. 2. Under the dependence assumption, 9 procedures are compared in terms of FDR, relative power and MSE of m_0. Solid line: P-LSU1; dotted line: P-LSU2; dashed line: ORC; ×: S-HLF; △: M-S-HLF; ◊: TST; +: RB20*; ○: MED; #: ABH.

4.3.2. Simulation result for dependent test

To examine the case where the test statistics are correlated, we borrow the setting from Liang and Nettleton (2012). The test statistics have block auto-regressive order 1 correlation structure with correlation of $\rho^{|i-j|}$ for the ith element and jth element within any block and block size of 50. The correlation coefficient ρ was set at -0.9. The tuning parameters were set at $c = 3$ and $\alpha = 0.6$.

In Fig. 2, since there exists a correlation between the test statistics, when the effect size is small ($\mu = 0.5$ or 1), all procedures, even including ORC, do not control the FDR very well. The performance of P-LSU1 and P-LSU2 are similar. They are the most powerful test among all procedures and have the smallest MSE for the small effect size ($\mu = 0.5$).

We conclude this section with a comment on the usefulness of our proposed models. Practically, we usually consider two-sided tests so that the effective size is not big enough as in case of $\mu = 0.5$ or 1 and the correlation structure is unknown. In this point of view, P-LSU1 and P-LSU2 are recommended in the real data analysis.

5. Discussion

When the effect size is big enough, the performance of existing adaptive procedures are good and we can find an appropriate threshold λ by dynamically adaptive procedures such as RB20*. However, in practice, we do not know whether the effect size is big or not. Also, since the two-sided tests are usually preferred by practitioners and the test statistics are commonly correlated, it is very likely that the null p-values are mixed with the alternative p-values. Therefore, P-LSU1 and P-LSU2 are recommended to use when analyzing real data.

To use our proposed procedures, we need to determine an appropriate c for P-LSU1 and α for P-LSU2. In the paper, we study the sensitivity of the choice of c and α. We find that when c is between 1 and 3 and α is between 0.4 and 0.6, the two proposed procedures are robust to the choice of the tuning parameters.

Acknowledgments

The authors would like to thank Ju Hee Cho for her helpful suggestions. The authors also appreciate the Editor, Professor H. Koul, and an anonymous referee for their careful reading of this paper and valuable comments. Zhang's research is supported by the NSF grants DMS-1106586 and DMS-1308872.

References

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

Benjamini, Y., Hochberg, Y., 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. 25, 60–63.

Benjamini, Y., Krieger, A., Yekutieli, D., 2006. Adaptive linear stepup procedures that control the false discovery rate. Biometrika 93, 491–507.

Efron, B., Tibshirani, R., Storey, J., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96, 1151–1160.

Genovese, C.R., Wasswerman, L., 2002. Operating characteristics and extensions of the FDR procedure. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 499–518.

Genovese, C.R., Wasswerman, L., 2004. A stochastic process approach to false discovery control. Ann. Statist. 32, 1035–1061.

Liang, K., Nettleton, D., 2012. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. J. R. Stat. Soc. Ser. B Stat. Methodol. 74, 163–182.

Nettleton, D., Hwang, J., Caldo, R., Wise, R., 2006. Estimating the number of true null hypotheses from a histogram of p values. J. Agric. Biol. Environ. Stat. 11, 337–356.

Shao, J., 2003. Mathematical Statistics, second ed. Springer, New York.

Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 479–498.

Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 66, 187–205.