Improving the Performance of the FDR Procedure Using an Estimator for the Number of True Null Hypotheses

Amit Zeisel, Or Zuk, Eytan Domany

W.I.S.

June 15, 2009

Introduction/Motivation

The vast use of high throughput technologies involves testing thousands of hypotheses simultaneously. The field of multiple testing deals with developing methods to determine the level of significance in such a scenario.

Multiple testing

m null hypotheses $H_{0,i}, \forall i = 1, 2, ...m$.

Example: for m variables $H_{0,i} : \mu_i^A = \mu_i^B$. Calculate p-values p_i, and set a threshold for significance. The set of random variables (U, V, T, S), and parameters (m, m_0, m_1) describes this scenario:

"ground truth"	non-rejected hypotheses	rejected hypotheses	total
null hypothesis is true	U	V	m_0
null hypothesis is false	T	S	m_1
total	$m - R$	R	m

Fraction of false discoveries = $\frac{V}{R}$

Control of the FDR=$E(\frac{V}{R^+})$

In 1995 Benjamini and Hochberg proposed a procedure (BH95) to control the FDR for a given set of p-values:

1. Sort and re-label the p-values, $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$.
2. Choose $0 \leq q \leq 1$ the desired FDR level.
3. Define the set of constants $\alpha_i = \frac{i}{m}q, \quad i = 1, 2, , , , m.$
4. Identify $R = max\{i : p_{(i)} \leq \alpha_i\}$.
5. If $R \geq 1$ reject all hypotheses $(i) = 1, 2, , , R$, else no hypothesis is rejected.

BH proved that $FDR \leq \frac{m_0 q}{m} \leq q$.

The bound is not tight, there is room for improvement.

Control vs. estimation, and improved procedures

- Control ($q \Rightarrow R$) - significance is preset at a desired level q. The procedure yields a set of rejected hypotheses with FDR$\leq q$.
- Estimation ($R \Rightarrow q$) - the threshold is preset at a level that yields a desired number of rejections R. The corresponding FDR is estimated.

There were many attempts to produce tighter bounds on the FDR, using an estimator for m_0.

Difficulty: the estimator $\hat{m}_0$ is a fluctuating random variable.

Aim

Produce an improved BH procedure using an estimator $\hat{m}_0$ for m_0, the number of true null hypotheses.

Definitions: monotonic estimator

An estimator for m_0 is a family of functions $\hat{m}_0 \equiv \hat{m}_0^{(m)} : [0,1]^m \rightarrow \mathbb{R}$,

$\hat{m}_0 \equiv \hat{m}_0(p_1, .., p_m)$. $\hat{m}_0$ is a monotonic estimator if it satisfies:

1. $\hat{m}_0^{(m)}(p_1, .., p_i, .., p_m) \geq \hat{m}_0^{(m)}(p_1, .., p_i', .., p_m)$,

   if $p_i \geq p_i'$, $\quad \forall\ i = 1, 2, , , m, \quad m \geq 1$

2. $\hat{m}_0^{(m)}(p_1, .., p_i, .., p_m) \geq \hat{m}_0^{(m-1)}(p_1, .., p_{i-1}, p_{i+1}, .., p_m)$,

   $\forall\ i = 1, 2, , , m, \ \ m \geq 2$

Definitions: modified BH procedure ($m \longrightarrow \hat{m}_0$)

Given m hypotheses of which m_0 are null, let $p_1, .., p_m$ be the respective p-values. The modified BH procedure with estimator $\hat{m}_0$ is:

1. Compute $\hat{m}_0 \equiv \hat{m}_0(p_1, .., p_m)$.
2. Sort and relabel the p-values $p_{(1)} \leq ... \leq p_{(m)}$.
3. Define the set of constants $q_k = \frac{qk}{\hat{m}_0} \quad k = 1, 2..., m$.
4. Let $R = \max\{k : p_{(k)} \leq q_k\}$.
5. If $R \geq 1$ reject $p_{(1)}, .., p_{(R)}$ else don't reject any hypothesis.

Theorem for control (see Benjamini et al 2006 (BKY))

Let $\hat{m}_0 \equiv \hat{m}_0(p_1, .., p_m)$ be a monotonic estimator for m_0. Let $\hat{m}_0^{(1)}(p_1, .., p_m) \equiv \hat{m}_0(p_2, .., p_m)$ be the same estimator, but disregarding the first p-value p_1. Assume that the null p-values are i.i.d. $U[0, 1]$. Then the modified BH procedure satisfies:

$$FDR = E\left[\frac{V}{R^+}\right] \leq m_0 q E\left[\frac{1}{\hat{m}_0^{(1)}}\right] \tag{1}$$

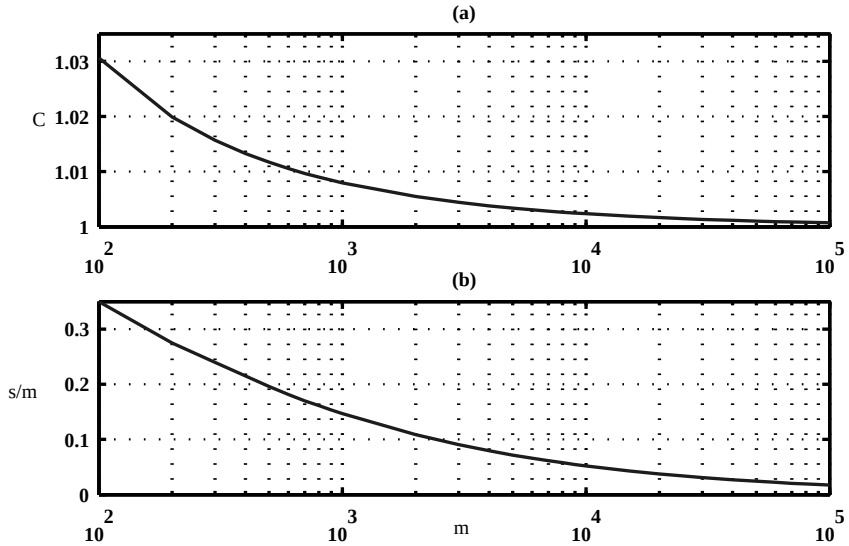
Note: if $E\left[\frac{1}{\hat{m}_0^{(1)}}\right] \leq \frac{1}{m_0}$, then FDR$\leq q$.

Two improved procedures

The two procedures are based on the estimators:

1. IBHsum: $\hat{m}_0 = C(m) \cdot min\left[m, max\left(s(m), 2\sum_{j=1}^{m} p_j\right)\right]$.

   $C(m), s(m)$ are universal correction factors.

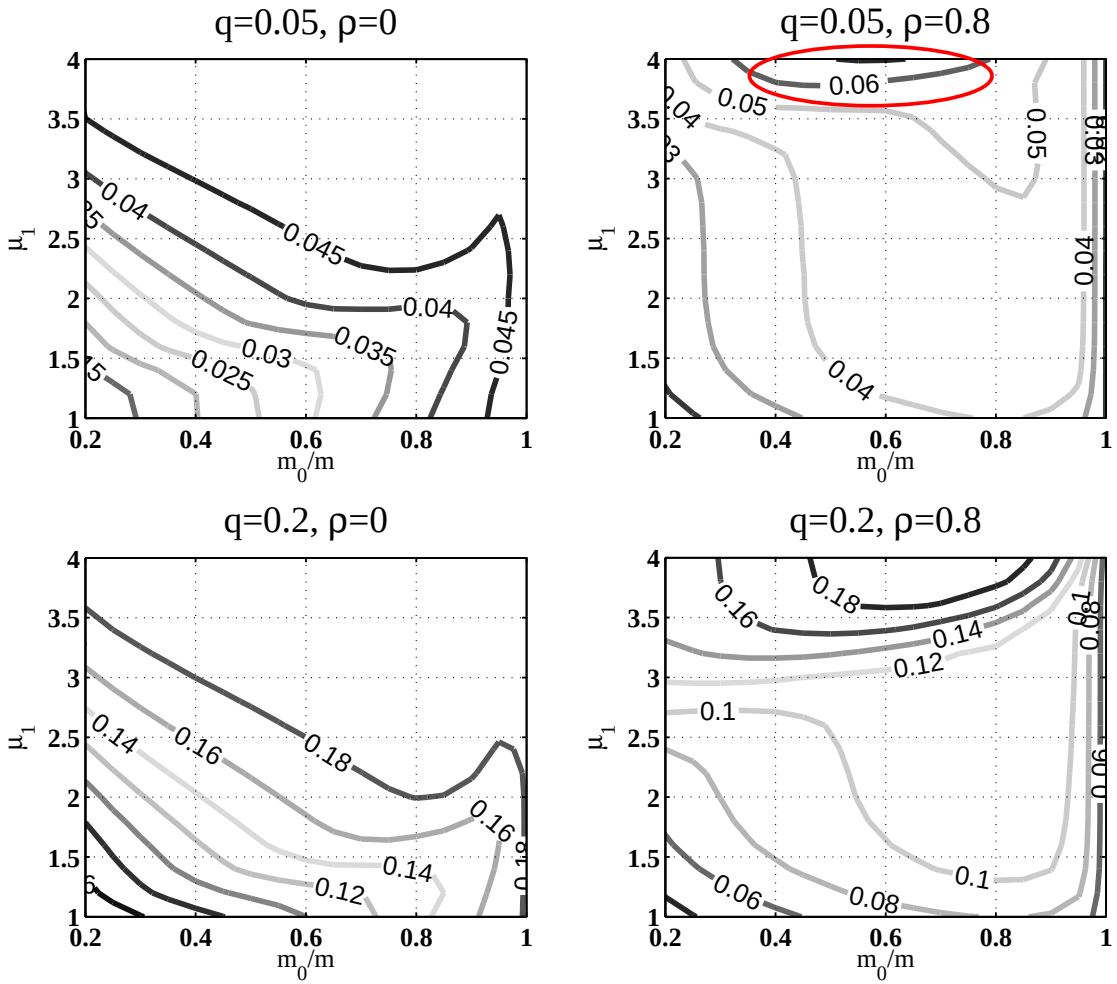


2. IBHlog: $\tilde{m}_0 = 2 - \sum_{i=1}^{m} \log(1 - p_i)$.

Both procedures satisfy $E(V/R^+) \leq q$

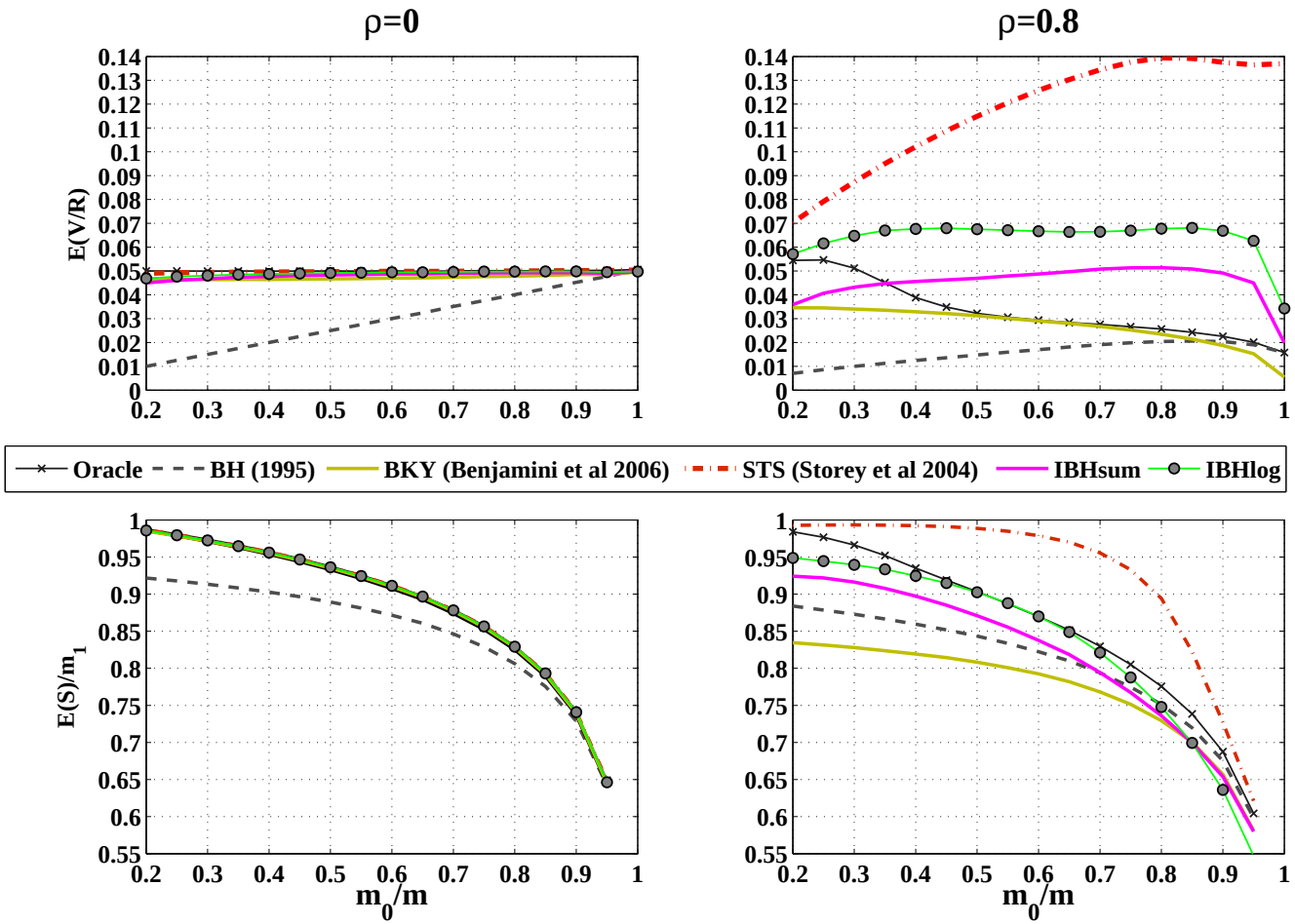
Performance: simulations (IBHsum)

ρ is the correlation between test statistics:

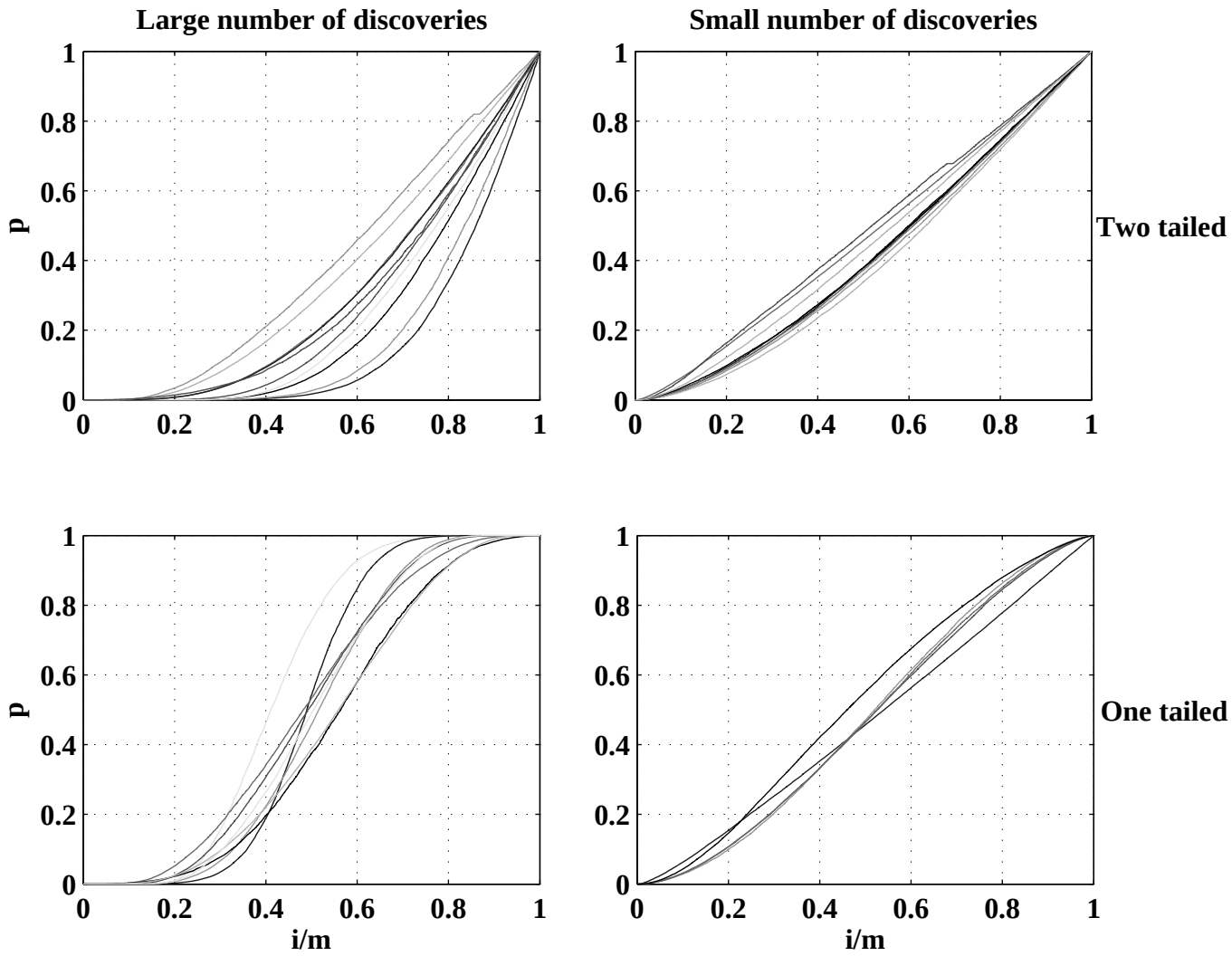


Compare to other methods

Results from simulations, $m = 500, \mu_1 = 3.5$:



Applying to 33 gene expression datasets



q	BKY	STS	IBHsum	IBHlog
a. Two tailed, large number of discoveries (10 studies)				
0.05	1.110 (0.043)	1.239 (0.138)	1.200 (0.110)	1.222 (0.130)
b. Two tailed, small number of discoveries (10 studies)				
0.05	1.003 (0.003)	1.316 (0.197)	1.231 (0.140)	1.291 (0.179)
c. One tailed, large number of discoveries (8 studies)				
0.05	1.049 (0.019)	1.011 (0.033)	1.014 (0.026)	0.108 (0.306)
d. One tailed, small number of discoveries (5 studies)				
0.05	0.998 (0.020)	1.027 (0.052)	1.025 (0.017)	0.882 (0.123)

Summary and conclusions

- We proved a theorem that provides a bound on the FDR for any improved procedure based on an monotonic estimator of m_0.
- We proposed two improved procedures based on the estimators $2\sum_{j=1}^{m} p_j$ and $\sum_{i=1}^{m} \log(1-p_i)$.
- For the case of independent statistics all improved procedures provide similar results: saturation of the bound and more power than BH95.
- We showed by simulations that even in the case of dependent statistics our procedures provide a reliable bound and improved power.
- For real gene expression data, where dependencies are expected, our methods improve, in general, over existing ones.

Theorem for monotonicity

Let $\vec{p} = (p_{1..m})$ be a set of independent p-values. Assume that f, the marginal probability density function of the alternatives, is monotonically non-increasing and differentiable. Let $B^{(i)}$ be two threshold FDR procedures rejecting $R^{(i)}(\vec{p})$ hypotheses and each having $FDR^{(i)}$, $i = 1, 2$. Assume that for any q, $R^{(1)}(\vec{p}) \leq R^{(2)}(\vec{p})$, $\forall \vec{p}$. Then it also holds that $FDR^{(1)} \leq FDR^{(2)}$.

THANKS