A discrete Benjamini-Hochberg procedure

Sebastian Döhler
Darmstadt University of Applied Sciences, Department of Mathematics and Science

Introduction

- Benjamini-Hochberg (BH) procedure: Classical method for controlling the false discovery rate for multiple testing problems.
- Originally designed for continuous test statistics.
- For discrete data the BH procedure may be unnecessarily conservative.

'How to derive better FDR procedures in the discrete paradigm remains an urgent but still unresolved problem.'
(Chen and Doerge, 2015)

The False Discovery Rate and the BH procedure

- Assume m hypotheses $H_1, \ldots, H_m \Rightarrow$ p-values $PV_1, \ldots, PV_m$
- For specific decision rule, define
  - $V = \#$ of rejected hypotheses
  - $R = \#$ of falsely rejected hypotheses.

Then the false discovery rate (FDR) is defined by $\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right)$.

The Benjamini-Hochberg procedure

For given p-values $PV_1, \ldots, PV_m$ the Benjamini-Hochberg (BH) step-up procedure with critical values $c_i^{BH} = \frac{i}{m}\alpha$ rejects $H_{(1)}, \ldots, H_{(k)}$ where $k = \max\left\{i | PV_{(i)} \leq \frac{i}{m}\alpha\right\}$ (if $\{\cdots\} \neq \emptyset$)

Distributional assumption

Under H_i^0: $\quad F_i(u) := P(PV_i \leq u) \leq u \qquad (u \in (0,1))$

FDR control for BH procedure

Let $PV_1, \ldots, PV_m$ be independent (or positively dependent) and satisfy the above distributional assumption. Then the BH step-up procedure controls the FDR at level

$$\mathrm{FDR} \leq \frac{\alpha}{m}|I|$$

where $I \subset \{1, \ldots, m\}$ index set of true hypotheses.

Discrete data

- Examples: Counts in clinical studies, next generation sequencing data, ...
- BH procedure valid but conservative!

FDR approaches for discrete data:

- Modify BH 'basis procedure': [3], [4], [5]
- Adaptive approaches: [6], [2], [1]

Main result: A discrete BH procedure

Let $PV_1, \ldots, PV_m$ be independent and satisfy the above distributional assumption and let $\alpha \in (0,1)$. Let $0 \leq c_1 \leq \ldots \leq c_m$ be such that for $k = 1, \ldots, m$

$$\sum_{i=1}^{m}(1+\tau_i) \cdot F_i(c_k) \leq \alpha \cdot k \qquad \text{where} \qquad \tau_i := \frac{F_i(c_m)}{1 - F_i(c_m)}.$$

Then the discrete BH step-up procedure (DBH) based on $c_1, \ldots, c_m$ controls the FDR at level α. More specifically it holds

$$\mathrm{FDR} \leq \max_{1 \leq k \leq m} \frac{1}{k}\sum_{i \in I}(1+\tau_i) \cdot F_i(c_k),$$

where $I \subset \{1, \ldots, m\}$ index set of true hypotheses.

- Proof uses results from [7] for multi-weighted procedures.
- DBH guarantees finite sample FDR-control.
- Uniform improvement over BH procedure (up to τ_i's).
- Simple and transparent approach, no selection of tuning parameters etc. necessary.

Example: Analysis of pharmacovigilance data

- Adverse event data from MHRA
- 2446 drugs in database
- Goal: Investigate association between amnesia and suspected drugs
- For each drug:
  - # reported cases of amnesia
  - # all reported adverse events
  - Fisher's exact test (one-sided)

Results (for $\alpha = 5\%$):

- $\mathrm{FDR}(c^{BH}) \leq 0.0215$
- # rejections: 24 (BH), 27 (DBH)

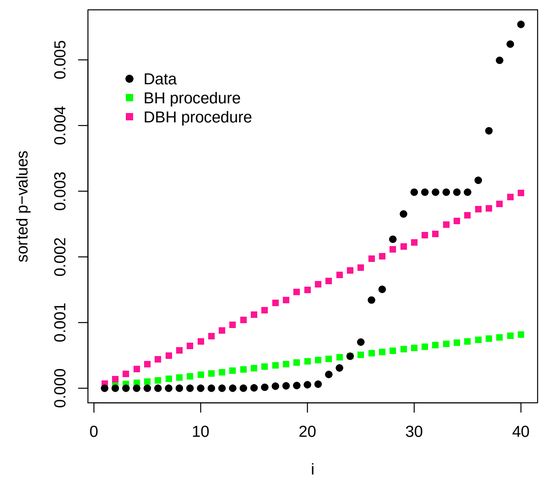

Figure 1: Rejection curves for the BH and DBH procedures.

Simulation study

- Clinical study with 2 treatment groups ($N = 25$ subjects per group)
- $m = 2000$ independent binary responses ('adverse events')
- For $i = 1, \ldots, m$: p_{1i}, p_{2i} success probabilities
- Hypotheses: For $i = 1, \ldots, m$

$$H_i^0 : p_{1i} = p_{2i} \qquad \text{vs.} \qquad H_i^1 : p_{1i} \neq p_{2i}$$

- Fisher's exact test (two-sided).

For $m = m_1 + m_2 + m_3$ generate data by

- $Bernoulli(0.01)$ at m_1 positions (both groups)
- $Bernoulli(0.10)$ at m_2 positions (both groups)
- $Bernoulli(0.10)$ at m_3 positions (group 1), $Bernoulli(0.40)$ at m_3 positions (group 2)
- Proportion of alternatives: $m_3/m = 10\%, 30\%, 60\%$ (small, intermediate, large)
- For each simulation evaluate rejections for BH and DBH
- Power: (average) proportion of rejections among m_3 alternatives.

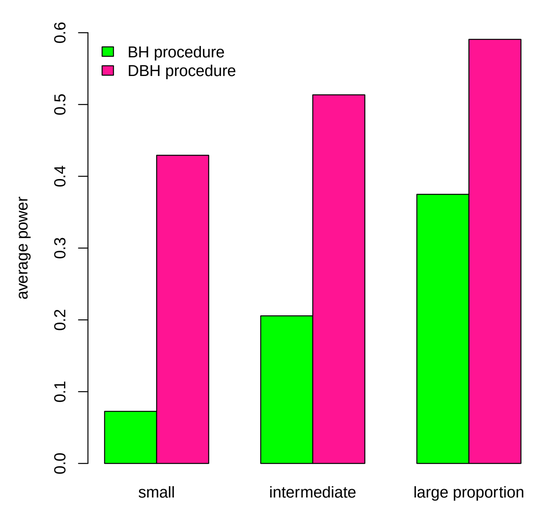

Figure 2: Power for the BH and DBH procedures.

Conclusions

- Discrete BH procedure can offer considerable improvement.
- Degree of improvement

'... depends upon the specific characteristics of the discrete distributions. Larger gains are possible when the number of tests is large and where many variables are sparse.' (Westfall and Wolfinger, 1997)

Acknowledgement Thanks to Etienne Roquain (Paris 6 University) for helpful discussions.

References

[1] X. Chen and R. Doerge. A weighted fdr procedure under discrete and heterogeneous null distributions. *arXiv:1502.00973*, 2015.

[2] Thorsten Dickhaus, Klaus Straßburger, and Daniel Schunk. How to analyze many contingency tables simultaneously in genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 11(4), 2012.

[3] P. G. Gilbert. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society. Series C*, 54(1):143–158, 2005.

[4] Ruth Heller and Hadas Gur. False discovery rate controlling procedures for discrete tests. *arxiv:1112.4627v2*, 2012.

[5] Joseph F. Heyse. A false discovery rate procedure for categorical data. In *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pages 43–58. World Scientific, 2011.

[6] Stan Pounds and Cheng Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987, 2006.

[7] Etienne Roquain and Mark A. van de Wiel. Optimal weighting for false discovery rate control. *Electron. J. Statist.*, 3:678–711, 2009.