An extension of the Benjamini and Hochberg procedure using data-driven optimal weights with grouped hypothesis

Guillermo Durand

PhD thesis under the supervision of
Etienne Roquain and Pierre Neuvial

29/08/2016

Table of contents

Table of contents

The well-known BH procedure

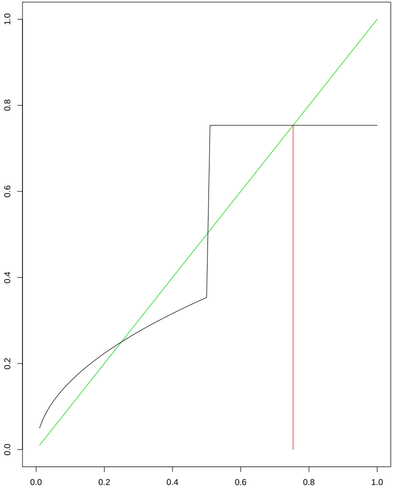
- Order p-values : $p_{(1)} \leq \cdots \leq p_{(m)}$
- Compute $\hat{k} = \max\{k : p_{(k)} \leq \alpha k/m\}$
- Reject all $p_i \leq \alpha \frac{\hat{k}}{m}$
- FDR control at level $\pi_0 \alpha$ when wPRDS

Another formulation

$\frac{\hat{k}}{m} = \max\{u : \widehat{G}(u) \geq u\} := \mathcal{I}\left(\widehat{G}\right)$ where

$$\widehat{G} : u \mapsto m^{-1} \sum_{i=1}^{m} \mathbb{1}_{\{p_i \leq \alpha u\}}, u \in [0,1]$$

An illustration of $\mathcal{I}(F)$



Weighted-BH

With given weights $(w_i)_{1 \leq i \leq m}$ such that $\sum_i w_i = m$ (called a weight vector), form

$$\widehat{G}_w : u \mapsto m^{-1} \sum_{i=1}^{m} \mathbb{1}_{\{p_i \leq \alpha u w_i\}}$$

and reject all $p_i \leq \alpha \hat{u} w_i$ with $\hat{u} = \mathcal{I}\left(\widehat{G}_w\right)$.

Weighted-BH

A generalization : weight functions

From Roquain and Van De Wiel 2009 :

Take a function W such that $(W_i(u))_i$ is a weight vector for all u and

$$\widehat{G}_W : u \mapsto m^{-1} \sum_{i=1}^{m} \mathbb{1}_{\{p_i \leq \alpha u W_i(u)\}}$$

is non-decreasing, then reject all $p_i \leq \alpha \hat{u} W_i(\hat{u})$ with $\hat{u} = \mathcal{I}\left(\widehat{G}_W\right)$.

Weighted-BH

A practical way to compute $\mathcal{I}\left(\widehat{G}_W\right)$

- No need to compute $W(u)$ for each u !

For each $k \in [\![1, m]\!]$, compute the $\frac{p_i}{W_i(\frac{k}{m})}$ and take q_r the r-th smallest. Let $q_0 = 0$.
Then $\mathcal{I}\left(\widehat{G}_W\right) = m^{-1} \max\{k \in [\![0, m]\!] : q_k \leq \alpha \frac{k}{m}\}$.

Optimal weighting

- Unconditional model : each hypothesis is null with proba π_0.
- Consider the procedure R_m^u rejecting p_i if $p_i \leq \alpha u w_i$ for all u.
- Its power is $\mathrm{Pow}_w(u) := (1-\pi_0) m^{-1} \sum_i^m F_i(\alpha u w_i)$ (F_i the c.d.f. under the alternative).
- Maximize it for all u :

Definition of optimal weights :

$$W^*(u) = \underset{(w_i) s.t. \sum_i^m w_i = m}{\mathrm{argmax}} \mathrm{Pow}_w(u)$$

Optimal weighting
Existence and uniqueness

Theorem (Roquain and Van De Wiel 2009)

Assume the following :

- F_i is strictly concave and continuous on $[0,1]$
- F_i has a derivative f_i on $(0,1)$
- $f_i(0^+)$ is constant for all i, same for $f_i(1^-)$
- $\lim_{y \to f_i(0^+)} \frac{f_j^{-1}(y)}{f_i^{-1}(y)}$ exists in $[0,\infty]$ for all i,j

Then we have existence, uniqueness and continuity of W^*, and $u \mapsto uW_i^*(u)$ is non-decreasing.

Optimal weighting
Existence and uniqueness

Proof ideas

Compute an explicit formula using the Lagrange multiplier method :

$$L(\lambda, w) = m^{-1} \sum_{i=1}^{m} F_i(\alpha u w_i) - \lambda \left(\sum_{i=1}^{m} w_g - m \right)$$

gives us

$$W_i^*(u) = \frac{1}{\alpha u} f_i^{-1} \left(\Psi^{-1}(\alpha u) \right)$$

where $\Psi(x) = m^{-1} \sum_{i=1}^{m} f_i^{-1}(x)$.

Optimal weighting

The main problem and the resulting motivation

- The distribution under the alternative F_i needs to be known to compute the W^*.
- Goal : estimate W^* without the knowledge of the alternative and obtain asymptotical results on FDR control and power for the associated weighted-BH procedure.
- Leads to data-driven optimal weighting.

Table of contents

Data-driven optimal weighting

- Assume that the p-values have uniform distribution under the null.

Main idea :

$W^*(u)$ is also the unique maximizer of

$$G_w(u) = \mathbb{E}\left[\widehat{G}_w(u)\right] = \pi_0 m^{-1} \sum_i^m \max(\alpha u w_i, 1) + \mathrm{Pow}_w(u)$$

the mean proportion of rejections done by the procedure R_m^u.
So we can estimate W^* by maximizing G_w's empiric counterpart $\widehat{G}_w$.

Data-driven optimal weighting

Grouping hypotheses

Key assumption

G groups of sizes $(m_g)_{1 \leq g \leq G}$, where p-values have the same distribution.

Examples :

- The Adequate Yearly Progress data set where grouping schools by size avoids a preference for large schools.
- Search for differently expressed genes between individuals with normal copy number or amplified one. Tests are more efficient for genes with ratio normal vs amplified copy numbers near 1.

Data-driven optimal weighting

More technical hypotheses

To obtain asymptotic results on $\widehat{W}^*$, we assume the following :

- p-values are independent.
- The previous hypothesis made by Roquain and Van De Wiel remain, and in addition $f_g(0^+) = \infty \ \forall g$.
- $\frac{m_g}{m} \underset{m\to\infty}{\longrightarrow} \pi_g > 0$.

All the following proofs inspired by Roquain and Van De Wiel 2009, Zhao and Zhang 2014 and Hu, Zhao, and Zhou 2010.

The two main results

Theorem (FDR control)

The False Discovery Proportion converges to $\pi_0\alpha$ almost surely so the FDR too by dominated convergence.

Theorem (optimal power)

Note by $\text{Pow}\,(W)$ the power of a BH procedure using a weight function W. Note by $\mathscr{W}$ the set of all sequences $\left(w^{(m)}\right)$ such that $\sum m_g w_g^{(m)} = m$. Then :

$$\lim_{m\to\infty} \text{Pow}\left(\widehat{W}^*\right) \geq \sup_{\left(w^{(m)}\right)\in\mathscr{W}} \limsup_{m\to\infty} \text{Pow}(w^{(m)}).$$

Some notations

- From now W^* is the asymptotic optimal weight when the F_g are known :

$$W^*(u) = \underset{w:\sum \pi_g w_g = 1}{\operatorname{argmax}} \; G_w^\infty(u)$$
$$= \underset{w:\sum \pi_g w_g = 1}{\operatorname{argmax}} \sum_g \pi_g D_g(\alpha u w_g)$$

with $D_g(\cdot) = \pi_0 \max(\cdot, 1) + (1 - \pi_0) F_g(\cdot)$.

- $P_W^\infty(u) = (1 - \pi_0) \sum_g \pi_g F_g(\alpha u W_g(u))$.
- $\hat{u} = \mathcal{I}\left(\widehat{G}_{\widehat{W^*}}\right)$ and $u^* = \mathcal{I}\left(G_{W^*}^\infty\right)$.

A chain of technical results

A first lemma

$$\sup_{u\in[0,1]} \sup_{w\in(\mathbb{R}^+)^G} \left| \widehat{G}_w(u) - G_w^\infty(u) \right| \xrightarrow{a.s.} 0$$

by Glivenko-Cantelli theorem and $\frac{m_g}{m} \to \pi_g$.

The main technical proposition

Proposition

$$\sup_{u\in[0,1]} \left| \widehat{G}_{\widehat{W}^*}(u) - G^{\infty}_{W^*}(u) \right| \xrightarrow{a.s.} 0$$

or, equivalently,

$$\sup_{u\in[0,1]} \left| G^{\infty}_{\widehat{W}^*}(u) - G^{\infty}_{W^*}(u) \right| \xrightarrow{a.s.} 0.$$

The main technical proposition
Proof ideas

- Play with the triangular inequality and remove the absolute values when able by using the maximality of $\widehat{G}_{\widehat{W}^*}(u)$ and $G^{\infty}_{W^*}(u)$

Problem

They are not maxima on the same sets :
$K^m = \{w : m^{-1} \sum m_g w_g = 1\}$ versus $K^\infty = \{w : \sum \pi_g w_g = 1\}$

The main technical proposition
Proof ideas

- We introduce two shifts $\delta(u) = \sum \pi_g \widehat{W}_g^*(u) - 1$ and $\delta'(u) = \sum \frac{m_g}{m} W_g^*(u) - 1$.
- Then we form shifted weights $\widehat{W}^{\sim}(u) = \widehat{W}^*(u) - \boldsymbol{\delta(u)} \in K^{\infty}$ and $W^{\sim}(u) = W^*(u) - \boldsymbol{\delta'(u)} \in K^m$.

The main technical proposition
Final ideas

- Make appear $\left| G^{\infty}_{\widehat{W}^{\sim}}(u) - G^{\infty}_{W^*}(u) \right| = G^{\infty}_{W^*}(u) - G^{\infty}_{\widehat{W}^{\sim}}(u)$.
- End up with $\sup_u \left| G^{\infty}_{\widehat{W}^*}(u) - G^{\infty}_{W^*}(u) \right| \leq$ $\sup_u \left(\widehat{G}_{W^{\sim}}(u) - \widehat{G}_{\widehat{W}^*}(u) \right) + o_{a.s.}(1)$.
- Use that $\widehat{G}_{W^{\sim}}(u) - \widehat{G}_{\widehat{W}^*}(u) \leq 0$. $\square$

The second important proposition

Proposition

$$\hat{u} \xrightarrow[m\to\infty]{a.s.} u^*$$

from which we deduce $\widehat{G}_{\widehat{W}^*}(\hat{u}) \xrightarrow{a.s.} G^{\infty}_{W^*}(u^*)$ by continuity.

Note $X_m = \sup_{u\in[0,1]} \left| \widehat{G}_{\widehat{W}^*}(u) - G^{\infty}_{W^*}(u) \right| \xrightarrow{a.s.} 0$, take a δ in $(0, u^*)$, note $u^0 = u^* - \delta$ and for all $\delta' \geq \delta$, $u' = u^* + \delta'$.

The second important proposition
Proof

- $s_\delta = \max_{\delta' \geq \delta} \left(G^{\infty}_{W^*}(u') - u' \right) < 0$ because if $s_\delta = 0$ it would contradict u^* maximality.
- $sup_{\delta' \geq \delta} \left(\widehat{G}_{\widehat{W}^*}(u') - u' \right) \leq s_\delta + X_m \to s_\delta < 0$
- So when $m \to \infty$ we must have $\hat{u} < u^* + \delta$.

The second important proposition
Proof

- $G^{\infty}_{W^*}(u^0) \geq G^{\infty}_{w}(u^0)$ with $w = W^*(u^*)$ by maximality.
- $G^{\infty}_{w}(u^0) = \frac{G^{\infty}_{w}(u^0)}{u^0} u^0 > \frac{G^{\infty}_{w}(u^*)}{u^*} u^0 = u^0$ by strict concavity.
- $\widehat{G}_{\widehat{W}^*}(u^0) - u^0 \geq G^{\infty}_{W^*}(u^0) - u^0 - X_m \to G^{\infty}_{W^*}(u^0) - u^0 > 0.$
- So when $m \to \infty$ we must have $\hat{u} > u^* - \delta$. $\square$

Third and last proposition

We have shown that $\widehat{G}_{\widehat{W}^*}(\hat{u}) \xrightarrow{a.s.} u^*$, that is for the denominator of the FDP. Showing that the numerator converges to $\pi_0 \alpha u^*$ is straightforward after this :

Proposition

$$\widehat{W}^*(\hat{u}) \xrightarrow{a.s.} W^*(u^*),$$

or, equivalently,

$$\widehat{W}^{\sim}(\hat{u}) \xrightarrow{a.s.} W^*(u^*).$$

Third and last proposition
Proof ideas

- One can show with the previous results and the triagular inequality that $\left| G^{\infty}_{\widehat{W}^{\sim}(\hat{u})}(u^*) - G^{\infty}_{W^*}(u^*) \right| \xrightarrow{a.s.} 0$.
- By contradiction, if $\widehat{W}^{\sim}(\hat{u}) \overset{a.s.}{\nrightarrow} W^*(u^*)$ then we find a $w^I \neq W^*(u^*)$ maximizing $G^{\infty}_{w}(u^*)$ but $W^*(u^*)$ is unique. $\square$

Optimality in power
Proof ideas

- First, $\mathrm{Pow}\left(\widehat{W}^*\right) = \mathbb{E}\left[\widehat{P}_{\widehat{W}^*}(\hat{u})\right]$ where $\widehat{P}_W(u)$ is m^{-1} times the number of true alternative rejected.
- $\widehat{P}_{\widehat{W}^*}(\hat{u}) \xrightarrow{a.s.} P^{\infty}_{W^*}(u^*)$.
- For each accumulation point for $\mathrm{Pow}(w^{(m)})$ there is an accumulation point w for $w^{(m)}$.
- $\hat{u}^{(m'')} \xrightarrow{a.s.} \mathcal{I}\left(G^{\infty}_w\right)$ and then
- $\widehat{P}_{w^{(m'')}}\left(\hat{u}^{(m'')}\right) \xrightarrow{a.s.} P^{\infty}_w\left(\mathcal{I}\left(G^{\infty}_w\right)\right) \leq P^{\infty}_{W^*}\left(\mathcal{I}\left(G^{\infty}_w\right)\right) \leq P^{\infty}_{W^*}(u^*)$. $\square$

Table of contents

Computation

- Fix a u, form $\tilde{p}_{gi} = \frac{p_{gi}}{\alpha u}$ and order the $\tilde{p}_{gi}$ in each group :

$$\tilde{p}_{g,1} \leq \cdots \leq \tilde{p}_{g,m_g}.$$

  Also note $\tilde{p}_{g,0} = 0$.
- Maximize over $w : \sum m_g w_g = m \iff$ maximize over $w : \sum m_g w_g \leq m$.
- If $\forall g, \tilde{p}_{g,1} > m$, stop and no rejection. If $\exists g, \tilde{p}_{g,1} \leq m$, continue and at least one rejection.

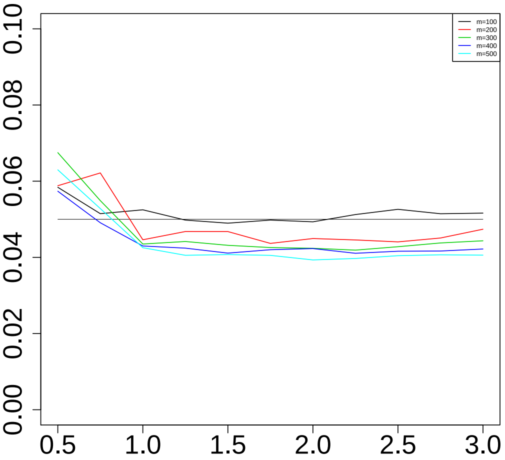
Computation

- Form all G-tuples $\boldsymbol{j} : \sum j_g = 2$ and check if there is one $\boldsymbol{j}$ such that $\sum m_g \tilde{p}_{g,j_g} \leq m$
  - If there is one, at least 2 rejections and continue with G-tuples of sum equal to 3.
  - If not, 1 rejection and use a $w_g = \tilde{p}_{g,\boldsymbol{j_g}}$ with a

    $$\boldsymbol{j} = (0, \dots, 0, \overbrace{1}^{h\text{-th position}}, 0, \dots, 0) \text{ such that } \tilde{p}_{h,1} \leq m.$$
- Reminder : the only values of u that need to be computed are $1, \frac{m-1}{m}, \dots, \frac{1}{m}$.

First simulations : $\alpha = 0.05$, 70% null hypothesis

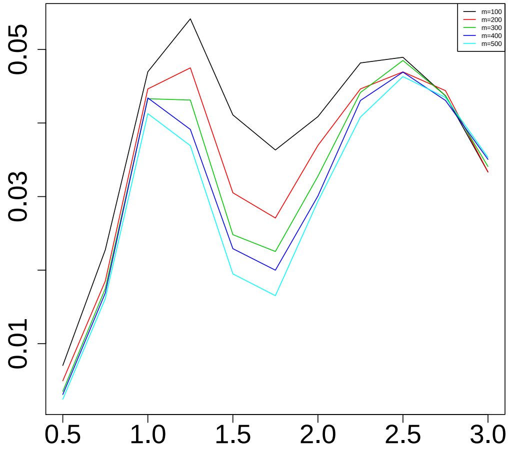
FDR plot, 2 groups, $\pi_1 = \pi_2 = 0.5$



- x axis : the $\bar{\mu}$ parameter. $\mu_1 = \bar{\mu}$ and $\mu_2 = 2\bar{\mu}$.
- y axis : the FDR.

First simulations : $\alpha = 0.05$, 70% null hypothesis

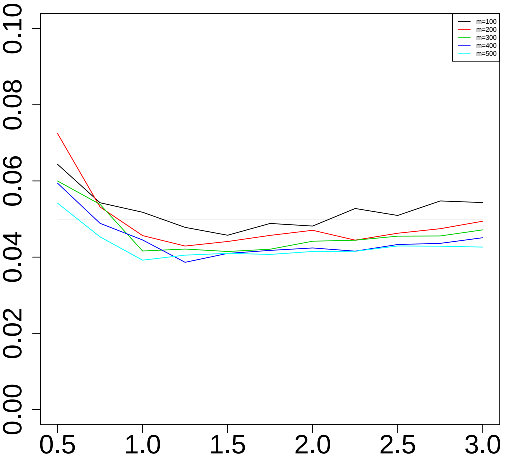
Relative power plot, 2 groups, $\pi_1 = \pi_2 = 0.5$



- x axis : the $\bar{\mu}$ parameter. $\mu_1 = \bar{\mu}$ and $\mu_2 = 2\bar{\mu}$.
- y axis : the difference in power between our procedure and the BH procedure.

First simulations : $\alpha = 0.05$, 70% null hypothesis

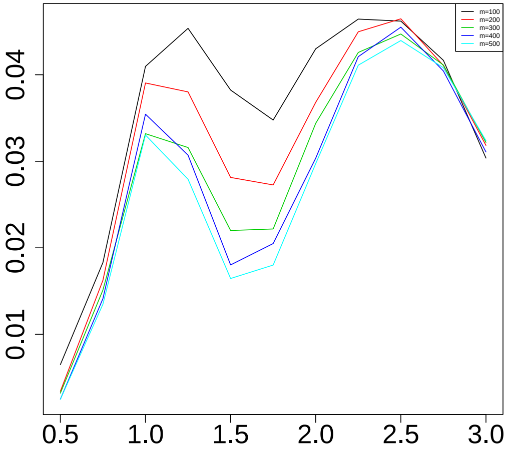
FDR plot, 2 groups, $\pi_1 = 0.3, \pi_2 = 0.7$



- x axis : the $\bar{\mu}$ parameter. $\mu_1 = \bar{\mu}$ and $\mu_2 = 2\bar{\mu}$.
- y axis : the FDR.

First simulations : $\alpha = 0.05$, 70% null hypothesis

Relative power plot, 2 groups, $\pi_1 = 0.3, \pi_2 = 0.7$



- x axis : the $\bar{\mu}$ parameter. $\mu_1 = \bar{\mu}$ and $\mu_2 = 2\bar{\mu}$.
- y axis : the difference in power between our procedure and the BH procedure.

Some perspectives

- Estimate π_0 to control the FDR at level α instead of $\alpha\pi_0$.
- A different π_0 in each group.
- Use wPRDS instead of independence.
- Optimize the computation.

The end

Thank you for your attention !

Bibliography

Hu, James X., Hongyu Zhao, and Harrison H. Zhou (2010). "False discovery rate control with groups". In: *Journal of the American Statistical Association* 105.491.

Roquain, Etienne and Mark A. Van De Wiel (2009). "Optimal weighting for false discovery rate control". In: *Electronic Journal of Statistics* 3, pp. 678–711.

Zhao, Haibing and Jiajia Zhang (2014). "Weighted p-value procedures for controlling FDR of grouped hypotheses". In: *Journal of Statistical Planning and Inference* 151, pp. 90–106.