Contrôle post hoc des faux positifs pour des hypothèses structurées

Guillermo Durand

Travail en commun avec Gilles Blanchard, Pierre Neuvial et Etienne Roquain
Séminaire Statistique, IRMA, Strasbourg

15/03/2021

Table of contents

Multiple testing setting

- Data $X \in (\mathcal{X}, \mathfrak{X})$ with $X \sim P \in \mathcal{P}$ a collection of distributions, P unknown
- m null hypotheses $H_{0,i}$ on P which are subsets of $\mathcal{P}$
- m is large!
- $\mathcal{H}_0 = \{i : P \in H_{0,i}\}$: $i \in \mathcal{H}_0 \Leftrightarrow H_{0,i}$ is true
- m p-values $p_i = p_i(X)$ such that $p_i \succeq \mathcal{U}([0,1])$ if $i \in \mathcal{H}_0$
  - Each p_i provides an α level test : $\mathbb{P}_{P \in H_{0,i}}(p_i \leq \alpha) \leq \alpha$
- Definition: for every subset of hypothese S: $V(S) = |S \cap \mathcal{H}_0|$

Classic MT theory: form a rejection set R with a guarantee on $V(R)$

- $\mathrm{FWER}(R) = \mathbb{P}(V(R) > 0)$
- $\mathrm{FDR}(R) = \mathbb{E}\left[\frac{V(R)}{|R| \vee 1}\right]$

Table of contents

Exploratory analysis in multiple testing

Search interesting hypotheses that will be cautiously investigated after.
Desired properties, as stated by [Goeman and Solari (2011)]:

- Mildness: allows some false positives
- Flexibility: the procedure does not prescribe, but advise
- Post hoc: take decisions on the procedure after seing the data

Example of post hoc decision

GWAS study with 10^6 genetic variants, select the 157 smallest p-values after seing a gap between the 157th and 158th smallest p-values.

Exploratory analysis in multiple testing

[Goeman and Solari (2011)]

This **reverses the traditional roles** of the user and procedure in multiple testing. Rather than, as in FWER-or FDR-based methods, to let the user choose the quality criterion, and to let the procedure return the collection of rejected hypotheses, the **user chooses the collection of rejected hypotheses freely**, and the multiple testing procedure returns the **associated quality criterion**.

FWER is somewhat flexible, FDR is somewhat mild

Post hoc and replication crisis

Replication crisis: many results poorly interpretable and non reproductible

Post hoc done wrong: *p*-hacking

- Pre-selecting variables that seem significant, exclude others from experiment
- Theoretical results no longer hold

Example

- GWAS study with 10^6 genetic variants
- Apply the Bonferroni procedure (FWER control) over the 1000 smallest *p*-values and report the result
- Problem: Bonferroni correction: $\alpha/1000$ instead of $\alpha/10^6$!

Selective inference against replication crisis

Selective inference: methods that account for a post hoc selection step and still provide statistical guarantees.

1. Conditionally to the selection event (e.g. Lasso selected features) [Fithian et al. (2017), Lee et al. (2016), and Tibshirani et al. (2016)]
2. Simultaneously over all possible selection events [Goeman and Solari (2011), Berk et al. (2013), and Blanchard et al. (2020)]

Selective inference against replication crisis

Selective inference: methods that account for a post hoc selection step and still provide statistical guarantees.

1. Conditionally to the selection event (e.g. Lasso selected features) [Fithian et al. (2017), Lee et al. (2016), and Tibshirani et al. (2016)]
2. Simultaneously over all possible selection events [Goeman and Solari (2011), Berk et al. (2013), and Blanchard et al. (2020)] ← The context of this work

Table of contents

Our goal: post hoc inference

Or simultaneous inference

Confidence bounds on any set of selected variables

A confidence bound is a (random) function $\widehat{V}$ such that

$$\mathbb{P}\left(\forall S \subset \mathbb{N}_m, V(S) \leq \widehat{V}(S)\right) \geq 1 - \alpha$$

- Hence for any selected $\widehat{S}$, $\mathbb{P}\left(V(\widehat{S}) \leq \widehat{V}(\widehat{S})\right) \geq 1 - \alpha$ holds
- Also an FDP bound: $\mathbb{P}\left(\forall S \subset \mathbb{N}_m, \text{FDP}(S) \leq \widehat{V}(S)/|S|\right) \geq 1 - \alpha$
- Originates from [Genovese and Wasserman (2006) and Meinshausen (2006)]
- A guarantee over any selected set instead of a rejected set: advise some $\widehat{S}$ instead of prescribe one R : he MT paradigm is reversed

BNR formalism

[Blanchard et al. (2020)]

Key concept: reference family

- $\mathfrak{R} = (R_k, \zeta_k)$ (random) such that Joint Error Rate (JER) control:

$$\text{JER}(\mathfrak{R}) = \mathbb{P}\left(\exists k, |R_k \cap \mathcal{H}_0| > \zeta_k\right) \leq \alpha$$

- Conversely, $\mathbb{P}\left(\forall k, |R_k \cap \mathcal{H}_0| \leq \zeta_k\right) \geq 1 - \alpha$
- Confidence bound only on the members of $\mathfrak{R}$
- $\Longrightarrow$ Derivation of a global confidence bound by interpolation

BNR formalism

[Blanchard et al. (2020)]

Two different bounds

- $V^*_{\mathfrak{R}}(S) = \max\left\{|S \cap A|, \forall k, |R_k \cap A| \leq \zeta_k\right\}$ optimal but difficult to compute
- $\overline{V}_{\mathfrak{R}}(S) = \min_k \left(\zeta_k + |S \setminus R_k|\right) \wedge |S|$ easy to compute

Main idea: the only information on $\mathcal{H}_0$ is that $\mathcal{H}_0 \in \left\{A, \forall k, |R_k \cap A| \leq \zeta_k\right\}$

$$\begin{aligned}|S \cap A| &= |(S \cap R_k) \cap A| + |(S \setminus R_k) \cap A| \\ &\leq |R_k \cap A| + |S \setminus R_k| \\ &\implies V^*_{\mathfrak{R}}(S) \leq \overline{V}_{\mathfrak{R}}(S)\end{aligned}$$

BNR formalism

A flexible and unified approach

- Compatible with previous works, like the closed testing approach of [Goeman and Solari (2011)]
- BNR approach: $\zeta_k = k - 1$ and find $R_k = \{i : p_i < t_k\}$ such that JER control. Example: $t_k = \alpha k/m$ (Simes inequality)
  - JER control becomes "simultaneous k-FWER control"
- Property: if R_k nested, then $\overline{V}_{\mathfrak{R}} = V^*_{\mathfrak{R}}$

Table of contents

Table of contents

Spatial structure

Informal assumption

The signal is localized in some spatially structured regions, with, possibly, different levels (e.g. active SNPs into genes into chromosomes)

[Meijer et al. (2015)]

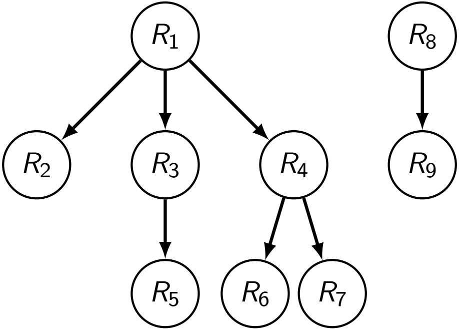
Considering the data at the region level is not only useful because these regions can be **the fundamental units of interest**, but also because these regions can have **an increased signal-to-noise-ratio**.

- Accordingly, find adapted new reference families
- We want $V^*_{\mathfrak{R}}$ to be easy to compute
- Our approach: deterministic R_k's capturing spatial hierarchy, estimate the true nulls inside them (i.e. ζ_k random)
    - opposite of [Blanchard et al. (2020)]

Table of contents

Forest structure

- $\forall k, k' \in \mathcal{K}, \; R_k \cap R_{k'} \in \{R_k, R_{k'}, \varnothing\}$
- Includes nested families or totally disjoint families
- Accommodates to different levels of signal localization through the different depths of the nodes



Forest structure

Important property

Property

There is a partition $(L_n)_{1\leq n\leq N}$ of $\mathbb{N}_m$ (the leaves) such that for each $k \in \mathcal{K}$, there exists some (i,j) with $1 \leq i \leq j \leq N$ and $R_k = L_{i:j}$, where we denote

$$L_{i:j} = \bigcup_{i\leq n\leq j} L_n$$

Identification:

$$\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}} \quad \text{or} \quad \mathfrak{R} = (L_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}$$

$\Longrightarrow$ leaves represent the thinnest division possible of the structure

Forest structure

Other important property

- Each forest structure can be completed to includes all leaves
- For an added leaf $L_{i:i}$, just state $\zeta_{i,j} = |L_{i:i}|$

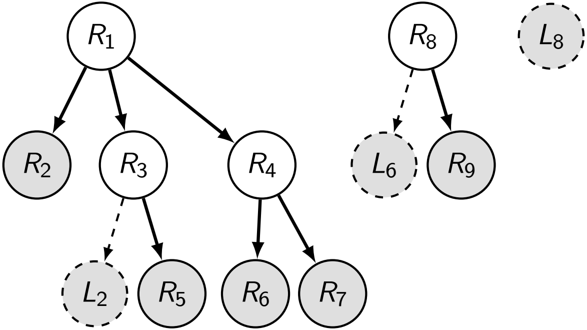


Table of contents

New interpolation bounds

Goal: compute $V^*_{\mathfrak{R}}$ easily with forest structure

Definition

For any $q \leq K = |\mathcal{K}|$,

$$\widetilde{V}^q_{\mathfrak{R}}(S) = \min_{Q \subset \mathcal{K}, |Q| \leq q} \left(\sum_{k \in Q} \zeta_k \wedge |S \cap R_k| + \left| S \setminus \bigcup_{k \in Q} R_k \right| \right),$$

and

$$\widetilde{V}_{\mathfrak{R}}(S) = \widetilde{V}^K_{\mathfrak{R}}(S).$$

Property

$$V^*_{\mathfrak{R}}(S) \leq \widetilde{V}_{\mathfrak{R}}(S) \leq \widetilde{V}^{K-1}_{\mathfrak{R}}(S) \leq \cdots \leq \widetilde{V}^2_{\mathfrak{R}}(S) \leq \widetilde{V}^1_{\mathfrak{R}}(S) = \overline{V}_{\mathfrak{R}}(S)$$

Main results

Compute $V^*_{\mathfrak{R}}$ easily with forest structure

Theorem

$$V^*_{\mathfrak{R}}(S) = \widetilde{V}_{\mathfrak{R}}(S)$$

More precisely,

$$V^*_{\mathfrak{R}}(S) = \widetilde{V}^{\ell}_{\mathfrak{R}}(S),$$

with $\ell =$ number of leaves (without completion).

Proof by construction $\Longrightarrow$ computation algorithm

Corollary

$\ell = 1$ for nested families and BNR property is recovered

Main results

Compute $V^*_{\mathfrak{R}}$ easily with forest structure

Corollary

There is a simple and efficient algorithm to compute $\widetilde{V}_{\mathfrak{R}}$ if $\mathfrak{R}$ is complete ($O(Hm)$ complexity).

Lemma

Completing the family does not change $V^*_{\mathfrak{R}}$ and $\widetilde{V}_{\mathfrak{R}}$.

Corollary

There is a simple algorithm to compute $V^*_{\mathfrak{R}}(S)$ by:

1. Completing the family
2. Travel across the forest from the leaves

Note: all of the above does not depend on the choice of the ζ_k and works for random R_k.

Forest algorithm

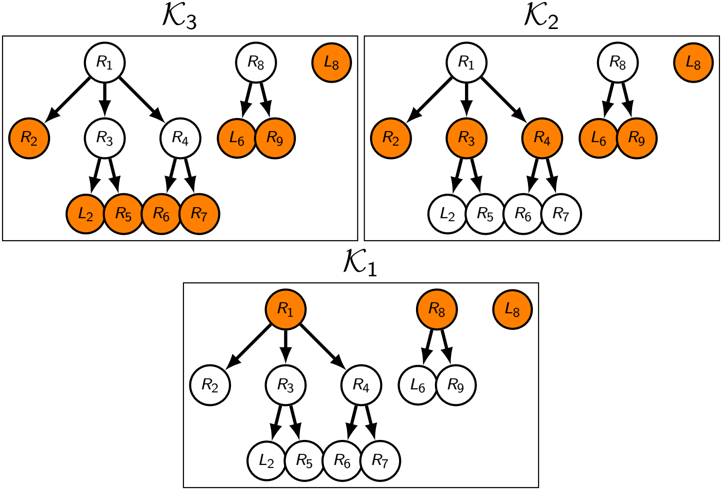


Table of contents

True nulls estimation inside regions

That is, ζ_k computation

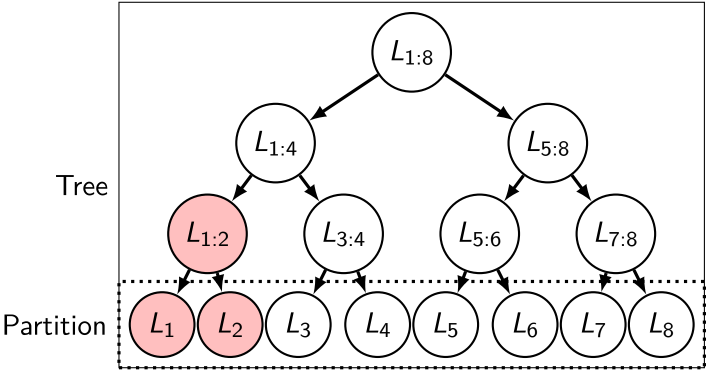
- K deterministic regions, let $C = \sqrt{\frac{1}{2}\log\left(\frac{K}{\alpha}\right)}$
- $\zeta_k = |R_k| \wedge \min_{t\in[0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i\in R_k} \mathbf{1}\{p_i>t\}}{1-t} \right)^{1/2} \right\rfloor^2$
- Comes from carefully handling the DKWM inequality [Dvoretzky et al. (1956) and Massart (1990)]
    - Requires independence!
- Replace $\min_{t\in[0,1)}$ and t above by $\min_{0\leq\ell\leq s}$ and $p_{(\ell)}$ for practical usage $\Longrightarrow$ computation of $(\zeta_k)_k$ is also $O(Hm)$ complex
- α/K instead of α in C: union bound
- Dependence on α (and to K!) only through a log
- $\zeta_k > 0$ (entry cost)

Table of contents

Comparison of 3 bounds

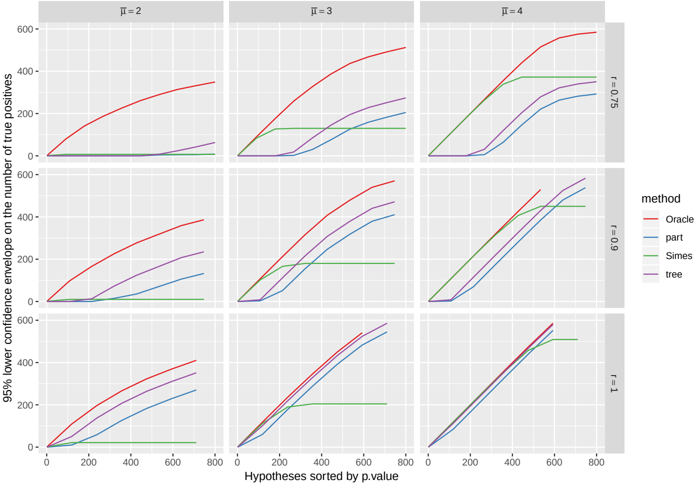
Simes bound of BNR, and 2 new

- V_{tree} and V_{part}, from a complete binary tree or only the partition of leaves
- Signal in adjacent leaves, good performance of V_{tree} expected despite worst K
- Parameters: signal $\bar{\mu}$ and signal proportion in active leaves r



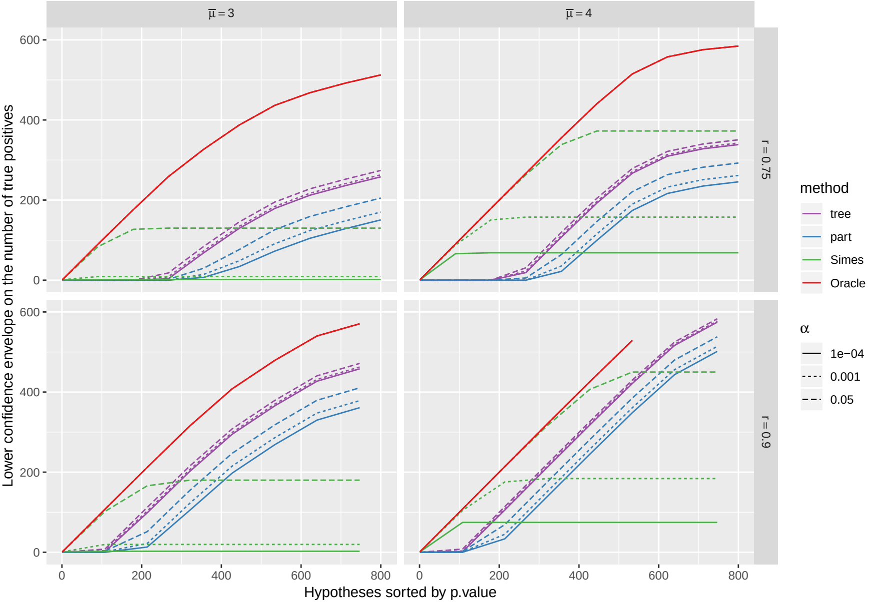
Comparison of 3 bounds

- The choice of S favors the Simes bound of BNR
- But for large r, new bounds better
- V_{tree} better than V_{part} as expected, despite a worst union bound correction



Comparison of 3 bounds

Influence of α



New hybrid bound suggested by the simulations

- $V^{\gamma}_{\text{hybrid}}(\alpha, S) = \min\left(V_{\text{Simes}}((1-\gamma)\alpha, S), V_{\mathrm{tree}}(\gamma\alpha, S)\right)$
- $\gamma = 0.02$: favors Simes, not a problem because V_{tree} is little sensitive to small α

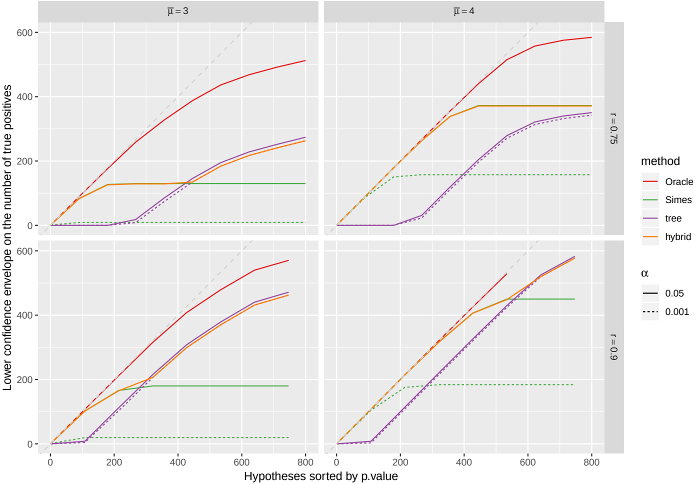


Table of contents

Table of contents

Recap

New confidence bounds that exploit the signal localization to improve on existing bounds, with an acceptable computation time
Limitations:

- DKWM inequality involves independence
- The chosen ζ_k can't reject a whole subset (including individual hypotheses)
- The R_k have to be fixed before seeing the data (not post hoc!)
- The union bound correction chosen may induce conservativeness

Published paper in Scandinavian Journal of Statistics (2020) [Durand et al. (2020)]
Also on arXiv: 1807.01470
R package available on github: sansSouci

Table of contents

Next steps I

- Depart from independence with $\zeta_k(X) = L_k(\alpha/K)$ such that $\mathbb{P}_{X\sim P}\Big(|R_k \cap \mathcal{H}_0(P)| \leq L_k(\lambda)\Big) \leq \lambda$
  - Concentration inequalities for dependent variables?
  - λ-calibration under known dependence or permutation invariance [Hemerik and Goeman (2018) and Blanchard et al. (2020)]
  - Use local tests [Goeman and Solari (2011) and Meijer et al. (2015)], App. B. of my thesis
  - Different L_k at different hierarchical levels [Dobriban (2020)]
- Reduce union bound penalty with some α-recycling (App. B of my thesis)
- Other families combining BNR approach and a deterministic partition
  - $\mathfrak{R} = (R_{k,i_k}, \zeta_{k,i_k})_{\substack{k\in\mathcal{K}\\ 1\leq i_k\leq |R_k|}}, \zeta_{k,i_k} = i_k - 1$
  - The results on forest structures allows the regions to be random
  - A first step toward automatic selection of the forest structure

Next steps II

- Reuse some of those ideas to go back to FWER control (App. B of my thesis)
- Pursue work on closed testing shortcuts for post hoc bounds (App. A.1 of my thesis)
- Applications, real-life favorable cases like neuroimagery [Vesely et al. (2021)]

Bibliography I

Berk, Richard et al. (2013). "Valid post-selection inference". In: *The Annals of Statistics* 41.2, pp. 802–837.

Blanchard, Gilles, Pierre Neuvial, and Etienne Roquain (2020). "Post hoc confidence bounds on false positives using reference families". In: *The Annals of Statistics* 48.3, pp. 1281 –1303. DOI: `10.1214/19-AOS1847`. URL: `https://doi.org/10.1214/19-AOS1847`.

Dobriban, Edgar (2020). "Fast closed testing for exchangeable local tests". In: *Biometrika* 107.3, pp. 761–768.

Durand, Guillermo et al. (2020). "Post hoc false positive control for structured hypotheses". In: *Scandinavian journal of Statistics* 47.4, pp. 1114–1148.

Dvoretzky, Aryeh, Jack Kiefer, and Jacob Wolfowitz (1956). "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator". In: *The Annals of Mathematical Statistics*, pp. 642–669.

Fithian, William, Dennis Sun, and Jonathan Taylor (2017). "Optimal inference after model selection". In: *arXiv preprint arXiv:1410.2597*.

Genovese, Christopher R and Larry Wasserman (2006). "Exceedance control of the false discovery proportion". In: *Journal of the American Statistical Association* 101.476, pp. 1408–1417.

Goeman, Jelle J and Aldo Solari (2011). "Multiple testing for exploratory research". In: *Statistical Science*, pp. 584–597.

Bibliography II

Hemerik, Jesse and Jelle J Goeman (2018). "False discovery proportion estimation by permutations: confidence for significance analysis of microarrays". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1, pp. 137–155.

Lee, Jason D et al. (2016). "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3, pp. 907–927.

Marcus, Ruth, Peritz Eric, and K Ruben Gabriel (1976). "On closed testing procedures with special reference to ordered analysis of variance". In: *Biometrika* 63.3, pp. 655–660.

Massart, Pascal (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *The Annals of Probability*, pp. 1269–1283.

Meijer, Rosa J, Thijmen JP Krebs, and Jelle J Goeman (2015). "A region-based multiple testing method for hypotheses ordered in space or time". In: *Statistical Applications in Genetics and Molecular Biology* 14.1, pp. 1–19.

Meinshausen, Nicolai (2006). "False discovery control for multiple tests of association under general dependence". In: *Scandinavian Journal of Statistics* 33.2, pp. 227–237.

Tibshirani, Ryan J et al. (2016). "Exact post-selection inference for sequential regression procedures". In: *Journal of the American Statistical Association* 111.514, pp. 600–620.

Vesely, Anna, Livio Finos, and Jelle J Goeman (2021). "Permutation-based true discovery guarantee by sum tests". In: *arXiv preprint arXiv:2102.11759.*

Classical theory

Family-Wise Error Rate (FWER)

- $\mathrm{FWER}(R) = \mathbb{P}(V(R) > 0)$
- Bonferroni method: reject all $p_i \leq \frac{\alpha}{m}$ (union bound)
- Variant: k-$\mathrm{FWER}(R) = \mathbb{P}(V(R) \geq k)$
  - Choice of k? Often post hoc!

False Discovery Rate (FDR)

- $\mathrm{FDR}(R) = \mathbb{E}\left[\frac{V(R)}{|R| \vee 1}\right]$
- Benjamini-Hochberg method for positive dependence
  - Reject all $p_i \leq \frac{\alpha \hat{k}}{m}$
  - $\hat{k} = \max\{k : p_{(k)} \leq \frac{\alpha k}{m}\}$, $p_{(1)} \leq \cdots \leq p_{(m)}$

Closed testing for post hoc inference

Designed for FWER control [Marcus et al. (1976)]

- Form $H_{0,I} = \bigcap_{i \in I} H_{0,i}$ all intersection hypotheses
- Have a collection of α level local test $\phi_{\boldsymbol{I}}$
- Examples:
  - Bonferroni test $\phi_{\boldsymbol{I}} = 1$ if $\exists i \in I : p_i \leq \alpha/|I|$
  - Simes test $\phi_I = 1$ if $\exists i \in I : p_{(i:I)} \leq \alpha i/|I|$ (under PRDS)
- Test $H_{0,I}$ only if all $H_{0,J}$, $J \supseteq I$, are rejected
- Reject the individual hypotheses $H_{0,i}$ such that $H_{0,\{i\}}$ has been rejected that way
- Then FWER(Closed testing) $\leq \alpha$

Closed testing for post hoc inference

[Goeman and Solari (2011)]

Main idea

The closed testing provides more information than just the individual rejects:

- Let $\mathcal{X}$ the set of all I such that we rejected $H_{0,I}$
- Simultaneous guarantee over all $H_{0,I}$, $I \in \mathcal{X}$:

$$\mathbb{P}\left(\forall I \in \mathcal{X}, H_{0,I} \text{ is false}\right) \geq 1 - \alpha$$

Confidence bound derivation:

- $V_{\mathrm{GS}}(S) = \max_{\substack{I \subseteq S \\ I \notin \mathcal{X}}} |I|$ is a confidence bound because

$$\begin{aligned} \exists S, |S \cap \mathcal{H}_0| > V_{\mathrm{GS}}(S) &\implies \exists S, S \cap \mathcal{H}_0 \in \mathcal{X} \\ &\implies \exists I \in \mathcal{X}, H_{0,I} \text{ is true} \end{aligned}$$

- $V_{\mathrm{GS}}(S) = V^*_{\mathfrak{R}}(S)$ with $\mathfrak{R} = (I, |I| - 1)_{I \in \mathcal{X}}$

DKWM use

- Let $S \subset \mathbb{N}_m$
- $N_t(S) = \sum_{i \in S} \mathbf{1}\{p_i(X) > t\}$
- $v = |S \cap \mathcal{H}_0|$

$$v \leq \min_{t \in [0,1)} \left(\frac{\sqrt{\log(1/\lambda)/2}}{2(1-t)} + \left\{ \frac{\log(1/\lambda)/2}{4(1-t)^2} + \frac{N_t(S)}{1-t} \right\}^{1/2} \right)^2$$

comes from

$$v^{-1} \sum_{i=1}^{v} \mathbf{1}\{U_i > t\} - (1-t) \geq -\sqrt{\log(1/\lambda)/(2v)}, \ \forall t \in [0,1],$$

with probability at least $1-\lambda$ ($U_1, \ldots, U_v$ i.i.d. uniform, $N_t(S)$ dominates $\sum_{i=1}^{v} \mathbf{1}\{U_i > t\}$ by independence)

- $S = R_k$ and $\lambda = \alpha/K$ (union bound)

Forest algorithm

Computation of $V^*_{\mathfrak{R}}(S)$

```
Data: $\mathfrak{R} = (L_{i:j}, \zeta_{i,j})_{(i,j)\in\mathcal{K}}$ and $S \subset \mathbb{N}_m$.
Result: $V^*_{\mathfrak{R}}(S)$.
$\mathfrak{R} \longleftarrow \mathfrak{R}^{\oplus}$; $\mathcal{K} \longleftarrow \mathcal{K}^{\oplus}$ (completion);
$H \longleftarrow \max_{k\in\mathcal{K}} \phi(k)$ (max depth);
$V \longleftarrow (\zeta_k \wedge |S \cap R_k|)_{k\in\mathcal{K}^H}$;
for $h \in \{H-1, \dots, 1\}$ do
    $newV \longleftarrow (0)_{k\in\mathcal{K}^h}$;
    for $k \in \mathcal{K}^h$ do
        $Succ_k \longleftarrow \{k' \in \mathcal{K}^{h+1} : R_{k'} \subset R_k\}$;
        $newV_k \longleftarrow \min\left(\zeta_k \wedge |S \cap R_k|, \sum_{k'\in Succ_k} V_{k'}\right)$;
    end
    $V \longleftarrow newV$;
end
return $\sum_{k\in\mathcal{K}^1} V_k$.
```