COMPUTO

ISSN 2824-7795

A fast algorithm to compute a curve of confidence upper bounds for the False Discovery Proportion using a reference family with a forest structure

Guillermo Durand [1] Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay

Date published: 2024-08-12 Last modified: 2024-08-12

Abstract

This paper presents a new algorithm (and an additional trick) that allows to compute fastly an entire curve of post hoc bounds for the False Discovery Proportion when the underlying bound $V^*_{\mathfrak{R}}$ construction is based on a reference family $\mathfrak{R}$ with a forest structure à la Durand et al. (2020). By an entire curve, we mean the values $V^*_{\mathfrak{R}}(S_1), \dots, V^*_{\mathfrak{R}}(S_m)$ computed on a path of increasing selection sets $S_1 \subsetneq \cdots \subsetneq S_m$, $|S_t| = t$. The new algorithm leverages the fact that going from S_t to S_{t+1} is done by adding only one hypothesis.

Keywords: multiple testing, algorithmic, post hoc inference, false discovery proportion, confidence bound

Contents

[1] Corresponding author: guillermo.durand@universite-paris-saclay.fr

1 Introduction

Multiple testing theory is often used for exploratory analysis, like Genome-Wide Association Studies, where multiple features are tested to find promising ones. Classical multiple testing theory like Family-Wise Error Rate (FWER) control or False Discovery Rate (FDR) control (Benjamini and Hochberg, 1995) can be used, but a more recent trend consists in the computation of post hoc bounds, also named post selection bounds or confidence envelopes, for the number of false positives, or, equivalently, for the False Discovery Proportion (FDP). This approach is notably advocated for in the context of exploratory research by (Goeman and Solari, 2011, Section 1).

Mathematically speaking, a confidence upper bound (we prefer to say upper bound instead of envelope for obvious reasons) is a function $\hat{V} : \mathscr{P}(\mathbb{N}_m^*) \to \mathbb{N}_m$, where $\mathbb{N}_m = \{0, \dots, m\}$, $\mathbb{N}_m^* = \{1, \dots, m\}$ and m is the number of hypotheses, such that

$$\forall \alpha \in]0, 1[, \mathbb{P}\left(\forall S \subseteq \mathbb{N}_m^*, |S \cap \mathscr{H}_0| \leq \hat{V}(S)\right) \geq 1 - \alpha. \tag{1}$$

Here, α is a target error rate and $\mathscr{H}_0$ is the set of hypotheses indices that are true null hypotheses. Note that the construction of $\hat{V}$ depends on α and on the random data X and the dependence is omitted to lighten notation and because there is no ambiguity. The meaning of Equation 1 is that $\hat{V}$ provides an upper bound of the number of null hypotheses in S for any selection set $S \subseteq \mathbb{N}_m^*$, which allows the user to perform post hoc selection on their data without breaching the statistical guarantee. Also note that by dividing by $|S| \vee 1$ in Equation 1 we also get a confidence bound for the FDP:

$$\forall \alpha \in]0, 1[, \mathbb{P}\left(\forall S \subseteq \mathbb{N}_m^*, \mathrm{FDP}(S) \leq \frac{\hat{V}(S)}{|S| \vee 1}\right) \geq 1 - \alpha. \tag{2}$$

So post hoc bounds provide ways to construct FDP-controlling sets instead of FDR-controlling sets, which is much more desirable given the nature of the FDR as an expected value. See for example (Bogdan et al., 2015, Figure 4) for a credible example where the FDR is controlled but the FDP has a highly undesirable behavior (either 0 because no discoveries at all are made, either higher than the target level).

The first confidence bounds are found in (Genovese and Wasserman, 2006) and (Meinshausen, 2006), although, in the latter, only for selection sets of the form $\{i \in \mathbb{N}_m : P_i \leq t\}$ where P_i is the p-value associated to the null hypothesis $H_{0,i}$. In (Goeman and Solari, 2011) the authors re-wrote the generic construction of (Genovese and Wasserman, 2006) in terms of closed testing (Marcus et al., 1976), proposed several practical constructions and sparked a new interest in multiple testing procedures based on confidence envelopes. This work was followed by a prolific series of works like (Meijer et al., 2015) and (Vesely et al., 2023). In (Blanchard et al., 2020), the authors introduce the new point of view of references families (see Section 2.2) to construct post hoc bounds, and show the links between this meta-technique and the closed testing one, along with new bounds.

Following the reference family trail, in (Durand et al., 2020) the authors introduce new reference families with a special set-theoretic constraint that allows an efficient computation of the bound denoted by $V_{\mathfrak{R}}^*$ on a single selection set S. The problem is that one often wants to compute $V_{\mathfrak{R}}^*$ on a whole path of selection sets $(S_t)_{t \in \mathbb{N}_m^*}$, for example the hypotheses attached to the t smallest p-values. Whereas the algorithm provided the aforementioned work (Durand et al., 2020, Algorithm 1) is fast for a single evaluation, it is slow and inefficient to repeatedly call it to compute each $V_{\mathfrak{R}}^*(S_t)$. If the S_t's are nested, and growing by one, that is $S_1 \subsetneq \cdots \subsetneq S_m$ and $|S_t| = t$, there is a way to efficiently compute $\left(V_{\mathfrak{R}}^*(S_t)\right)_{t \in \mathbb{N}_m}$ by leveraging the nested structure.

This is the main contribution of the present paper: a new and fast algorithm computing the curve $\left(V_{\mathfrak{R}}^*(S_t)\right)_{t \in \mathbb{N}_m}$ for a nested path of selection sets, that is presented in Section 3.2. An additional algorithm that can speed up computations both for the single-evaluation algorithm and the new curve-evaluation algorithm is also presented, in Section 3.1. In Section 2.1, all necessary notation and vocabulary is re-introduced, most of it being the same as in (Durand et al., 2020). Finally, a few numerical experiments are presented in Section Section 4 to demonstrate the computation time gain.

2 Notation and reference family methodology

2.1 Multiple testing notation

As is usual in multiple testing theory, we consider a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, a model $\mathscr{P}$on a measurable space $(\mathscr{X}, \mathfrak{X})$, and data that is represented by a random variable $X : (\Omega, \mathscr{A}) \to (\mathscr{X}, \mathfrak{X})$ with $X \sim P \in \mathscr{P}$; that is, the law of X is comprised in the model $\mathscr{P}$.

Then we consider $m \geq 1$ null hypotheses $H_{0,1}, \dots, H_{0,m}$ which formally are submodels, that is subsets of $\mathscr{P}$. The associated alternative hypotheses $H_{1,1}, \dots, H_{1,m}$ are submodels such that $H_{0,i} \cap H_{1,i} = \emptyset$ for all $i \in \mathbb{N}_m^*$. We denote by $\mathscr{H}_0 = \mathscr{H}_0(P)$ (the dependence in P will be dropped when there is no ambiguity) the set of all null hypotheses that are true, that is $\mathscr{H}_0(P) = \{i \in \mathbb{N}_m^* : P \in H_{0,i}\}$. In other words, $H_{0,i}$ is true if and only if $i \in \mathscr{H}_0$. For testing each $H_{0,i}, i \in \mathbb{N}_m^*$, we have at hand a p-value $p_i = p_i(X)$ (the dependence in X will be dropped when there is no ambiguity) which is a random variable with the following property : if $i \in \mathscr{H}_0$, then the law of p_i is super-uniform, which is sometimes denoted $\mathscr{L}(p_i) \succeq \mathscr{U}([0,1])$. This means that in such case, the cumulative distribution function (cdf) of p_i is always smaller than or equal to the cdf of a random variable $U \sim \mathscr{U}([0,1])$:

$$\forall x \in \mathbb{R}, \mathbb{P}\left(p_i \leq x\right) \leq \mathbb{P}\left(U \leq x\right) = 0 \vee (x \wedge 1). \tag{3}$$

For every subset of hypotheses $S \subseteq \mathbb{N}_m^*$, let $V(S) = |S \cap \mathscr{H}_0|$. If we think of S as a selection set of hypotheses deemed significant, $V(S)$ is then the number of false positives (FP) in S. $V(S)$ is our main object of interest and the quantity that we wish to over-estimate with confidence upper bounds (see Equation 1).

Finally let us consider the following toy example, that will be re-used in the remainder of the paper.

Example 2.1 (Gaussian one-sided)**.** In this case we assume that $X = (X_1, \dots, X_m)$ is a Gaussian vector and the null hypotheses refer to the nullity of the means in contrast to their positivity. That is, formally, $(\mathscr{X}, \mathfrak{X}) = (\mathbb{R}^m, \mathscr{B}(\mathbb{R}^m))$, $\mathscr{P} = \{\mathscr{N}(\boldsymbol{\mu}, \Sigma) : \forall j \in \mathbb{N}_m^*, \mu_j \geq 0, \Sigma \text{ positive semidefinite}\}$, for each $i \in \mathbb{N}_m^*$, $H_{0,i} = \{\mathscr{N}(\boldsymbol{\mu}, \Sigma) \in \mathscr{P} : \mu_i = 0\}$ and $H_{1,i} = \{\mathscr{N}(\boldsymbol{\mu}, \Sigma) \in \mathscr{P} : \mu_i > 0\}$. Then we can construct p-values by letting $p_i(X) = \bar{\Phi}(X_i) = 1 - \Phi(X_i)$, where Φ denotes the cdf of $\mathscr{N}(0,1)$ and $\bar{\Phi}$ the associated survival function.

2.2 Post hoc bounds with reference families

2.3 Deterministic regions with a forest structure

3 New algorithms

3.1 Pruning the forest

💡 Tip

Proposition 3.1 (Pruning)**.**

Proof. Content □

Algorithm 1 Pruning of $\mathfrak{R}$

```
procedure PRUNING($\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathscr{K}}$ with $\mathfrak{R}$ complete)
    $\mathscr{K}^{\mathfrak{pr}} \leftarrow \mathscr{K}$
    $H \leftarrow \max_{k\in\mathscr{K}} \phi(k)$                                          ▷ maximum depth
    for $h = H - 1, \ldots, 1$ do
        $\mathscr{K}^h \leftarrow \{k \in \mathscr{K} : \phi(k) = h\}$
        $newVec \leftarrow (0)_{k\in\mathscr{K}^h}$
        for $k \in \mathscr{K}^h$ do
            $Succ_k \leftarrow \{k' \in \mathscr{K}^{h+1} : R_{k'} \subseteq R_k\}$
            if $Succ_k = \varnothing$ then
                $newVec_k \leftarrow \zeta_k$
            else
                if $\zeta_k \geq \sum_{k'\in Succ_k} Vec_{k'}$ then
                    $\mathscr{K}^{\mathfrak{pr}} \leftarrow \mathscr{K}^{\mathfrak{pr}} \setminus \{k\}$
                end if
                $newVec_k \leftarrow \min\left(\zeta_k, \sum_{k'\in Succ_k} Vec_{k'}\right)$
            end if
        end for
        $Vec \leftarrow newVec$
    end for
    return $(\mathscr{K}^{\mathfrak{pr}}, \sum_{k\in\mathscr{K}^1} Vec_k)$
end procedure
```

3.2 Fast algorithm to compute a curve of confidence bounds on a path of selection sets

Theorem 3.1 (Fast curve computation).

Proof. Content □

Corollary 3.1 (Easy implementation).

4 Numerical experiments

5 Conclusion

6 Acknowledments

This work has been supported by ANR-20-IDEES-0002 (PIA), ANR-19-CHIA-0021 (BISCOTTE), ANR-23-CE40-0018 (BACKUP) and ANR-21-CE23-0035 (ASCAI).

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995. ISSN 0035-9246. URL https://www.jstor.org/stable/2346101.

Algorithm 2 Formal computation of $(V^*_{\mathfrak{R}}(S_t))_{0\leq t\leq m}$

```
procedure CURVE($\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathscr{K}}$ with $\mathfrak{R}$ complete, path $(S_t)_{1\leq t\leq m}$ with $S_t = \{i_1, \dots, i_t\}$)
    $\mathscr{P}^0 \leftarrow \{(i,i) : 1 \leq i \leq n\}$                    ▷ the set of all atoms indices
    $\mathscr{K}^-_0 \leftarrow \{k \in \mathscr{K} : \zeta_k = 0\}$
    $\eta^0_k \leftarrow 0$ for all $k \in \mathscr{K}$
    for $t = 1, \dots, m$ do
        if $i_t \in \bigcup_{k\in\mathscr{K}^-_{t-1}} R_k$ then
            $\mathscr{P}^t \leftarrow \mathscr{P}^{t-1}$
            $\mathscr{K}^-_t \leftarrow \mathscr{K}^-_{t-1}$
            $\eta^t_k \leftarrow \eta^{t-1}_k$ for all $k \in \mathscr{K}$
        else
            for $h = 1, \dots, h_{\max}(t)$ do
                $\eta^t_{k^{(t,h)}} \leftarrow \eta^{t-1}_{k^{(t,h)}} + 1$
                if $\eta^t_{k^{(t,h)}} < \zeta_k$ then
                    Pass
                else
                    $h^f_t \leftarrow h.$
                    $\mathscr{P}^t \leftarrow \left(\mathscr{P}^{t-1} \setminus \{k \in \mathscr{P}^{t-1} : R_k \subseteq R_{k^{(t,h^f_t)}}\}\right) \cup \{k^{(t,h^f_t)}\}$
                    $\mathscr{K}^-_t \leftarrow \mathscr{K}^-_{t-1} \cup \{k^{(t,h^f_t)}\}$
                    Break the loop
                end if
            end for
            if the loop has been broken then
                $\eta^t_k \leftarrow \eta^{t-1}_k$ for all $k \in \mathscr{K}$ not visited during the loop, that is all $k \notin \{k^{(t,h)}, 1 \leq h \leq h^f_t\}$
            else
                $\mathscr{P}^t \leftarrow \mathscr{P}^{t-1}$
                $\mathscr{K}^-_t \leftarrow \mathscr{K}^-_{t-1}$
                $\eta^t_k \leftarrow \eta^{t-1}_k$ for all $k \in \mathscr{K}$ not visited during the loop, that is all $k \notin \{k^{(t,h)}, 1 \leq h \leq h_{\max}(t)\}$
            end if
        end if
    end for
    return $\mathscr{P}^t, \eta^t_k$ for all $t = 1, \dots, m$ and $k \in \mathscr{K}$
end procedure
```

Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc confidence bounds on false positives using reference families. *Ann. Statist.*, 48(3):1281–1303, 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1847. URL https://doi.org/10.1214/19-AOS1847.

Mał gorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015. ISSN 1932-6157,1941-7330. doi: 10.1214/15-AOAS842. URL https://doi.org/10.1214/15-AOAS842.

Guillermo Durand, Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc false positive control for structured hypotheses. *Scand. J. Stat.*, 47(4):1114–1148, 2020. ISSN 0303-6898. doi: 10.1111/sjos.12453. URL https://doi.org/10.1111/sjos.12453.

Christopher R. Genovese and Larry Wasserman. Exceedance control of the false discovery proportion.

Algorithm 3 Implementation of $(V^*_{\mathfrak{R}}(S_t))_{0\leq t\leq m}$

```
procedure CURVE($\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ with $\mathfrak{R}$ complete, path $(S_t)_{1\leq t\leq m}$ with $S_t = \{i_1, \dots, i_t\}$)
    $V_0 \leftarrow 0$
    $\mathcal{K}^- \leftarrow \{k \in \mathcal{K} \,:\, \zeta_k = 0\}$
    $\eta_k \leftarrow 0$ for all $k \in \mathcal{K}$
    for $t = 1, \dots, m$ do
        if $i_t \in \bigcup_{k\in\mathcal{K}^-} R_k$ then
            $V_t \leftarrow V_{t-1}$
        else
            for $h = 1, \dots, h_{\max}(t)$ do
                find $k^{(t,h)} \in \mathcal{K}^h$ such that $i_t \in R_{k^{(t,h)}}$
                $\eta_{k^{(t,h)}} \leftarrow \eta_{k^{(t,h)}} + 1$
                if $\eta_{k^{(t,h)}} < \zeta_k$ then
                    pass
                else
                    $\mathcal{K}^- \leftarrow \mathcal{K}^- \cup \{k^{(t,h)}\}$
                    break the loop
                end if
            end for
            $V_t \leftarrow V_{t-1} + 1$
        end if
    end for
    return $(V_t)_{1\leq t\leq m}$
end procedure
```

J. Amer. Statist. Assoc., 101(476):1408–1417, 2006. ISSN 0162-1459. doi: 10.1198/016214506000000339. URL https://doi.org/10.1198/016214506000000339.

Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statist. Sci.*, 26(4):584–597, 2011. ISSN 0883-4237. doi: 10.1214/11-STS356. URL https://doi.org/10.1214/11-STS356.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.655. URL https://doi.org/10.1093/biomet/63.3.655.

Rosa J. Meijer, Thijmen J. P. Krebs, and Jelle J. Goeman. A region-based multiple testing method for hypotheses ordered in space or time. *Stat. Appl. Genet. Mol. Biol.*, 14(1):1–19, 2015. ISSN 2194-6302. doi: 10.1515/sagmb-2013-0075. URL https://doi.org/10.1515/sagmb-2013-0075.

Nicolai Meinshausen. False discovery control for multiple tests of association under general dependence. *Scand. J. Statist.*, 33(2):227–237, 2006. ISSN 0303-6898. doi: 10.1111/j.1467-9469.2005.00488.x. URL https://doi.org/10.1111/j.1467-9469.2005.00488.x.

Anna Vesely, Livio Finos, and Jelle J. Goeman. Permutation-based true discovery guarantee by sum tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 85(3):664–683, 2023. ISSN 1369-7412,1467-9868. doi: 10.1093/jrsssb/qkad019. URL https://doi.org/10.1093/jrsssb/qkad019.

Session information

```
R version 4.4.0 (2024-04-24)
Platform: x86_64-pc-linux-gnu
```

```
Running under: Ubuntu 22.04.4 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so;  LAPACK version 3.10.0

locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
 [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
[10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

time zone: UTC
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base

loaded via a namespace (and not attached):
 [1] compiler_4.4.0    fastmap_1.1.1     cli_3.6.2         htmltools_0.5.8.1
 [5] tools_4.4.0       yaml_2.3.8        rmarkdown_2.26    knitr_1.46
 [9] jsonlite_1.8.8    xfun_0.43         digest_0.6.35     rlang_1.1.3
[13] renv_1.0.7        evaluate_0.23
```