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: 2024-07-03 Last modified: 2024-07-03

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 $S_1 \subsetneq \dots \subsetneq S_m$, $|S_t| = t$. 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

2 Notation and reference family methodology

2.1 Multiple testing notation

2.2 Post hoc bounds with reference families

2.3 Deterministic regions with a forest structure

3 New algorithms

3.1 Pruning the forest

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 $Succ_k = \emptyset$ **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 forreturn $(\mathscr{K}^{\mathfrak{pr}}, \sum_{k \in \mathscr{K}^1} Vec_k)$
end procedure

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

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).

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

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

References

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.

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
```

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

procedure CURVE($\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathscr{K}}$ with $\mathfrak{R}$ complete, path $(S_t)_{1\le t\le m}$ with $S_t = \{i_1, \dots, i_t\}$)
  $V_0 \leftarrow 0$
  $\mathscr{K}^- \leftarrow \{k \in \mathscr{K} \,:\, \zeta_k = 0\}$
  $\eta_k \leftarrow 0$ for all $k \in \mathscr{K}$
  for $t = 1, \dots, m$ **do**
    if $i_t \in \bigcup_{k\in\mathscr{K}^-} R_k$ **then**
      $V_t \leftarrow V_{t-1}$
    else
      for $h = 1, \dots, h_{\max}(t)$ **do**
        find $k^{(t,h)} \in \mathscr{K}^h$ 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
          $\mathscr{K}^- \leftarrow \mathscr{K}^- \cup \{k^{(t,h)}\}$
          break the loop
        end if
      end for
      $V_t \leftarrow V_{t-1} + 1$
    end if
  end forreturn $(V_t)_{1\le t\le m}$
end procedure

```
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
```