A leave-p-out based estimation of the proportion of null hypotheses in multiple testing problems

ALAIN CELISSE & STÉPHANE ROBIN

UMR 518 AgroParisTech/INRA MIA
SSB Workgroup : Statistics for Systems Biology

Multiple testing

- Test simultaneously a large number m of hypotheses.
- $\pi_0 = m_0/m$ of them are true, but π_0 is unknown.

Goal :

Build a decision rule that make as 'few mistakes' as possible.

False Discovery Rate (Benjamini-Hochberg 95)

$$FDR \;=\; \mathbb{E}\left[\frac{FP}{R}\,\mathbb{1}_{\{R>0\}}\right],$$

where $\begin{cases} FP: \text{number of falsely rejected hypotheses (False Positives)} \\ R: \text{number of Rejections} \end{cases}$.

Benjamini-Hochberg procedure (Decision rule)

- $P_{(1)}, \ldots, P_{(m)}$: ordered p-values,
- Reject hypotheses $H_{(i)},\ 1 \leq i \leq \widehat{k}$, where

$$\widehat{k} = \max\{\, i /\ P_{(i)} \leq i\alpha/m\}.$$

Theorem (BH 95, Storey et al. 04) Applying the BH-procedure under independence assumption,

$$\forall \alpha \in (0,1], \qquad FDR = \pi_0 \alpha \leq \alpha.$$

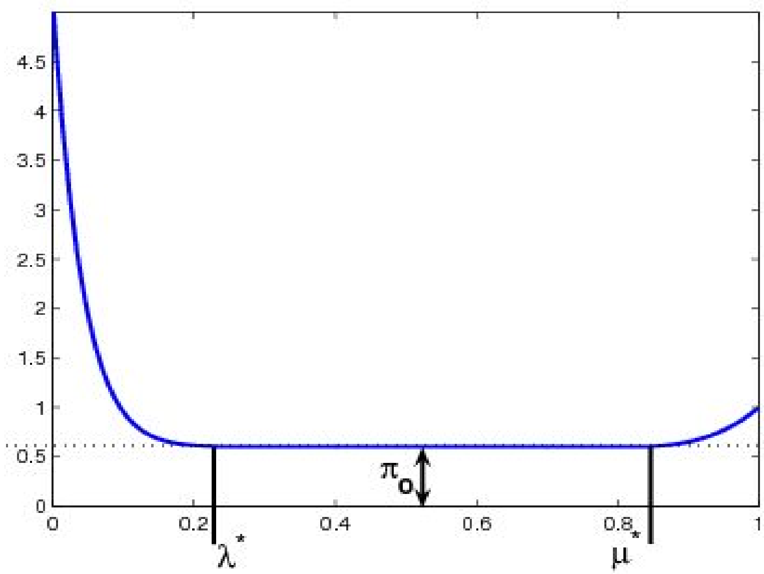
Fact :

Finding accurate conservative $\widehat{\pi}_0$ provides accurate upper-bound of the FDR.

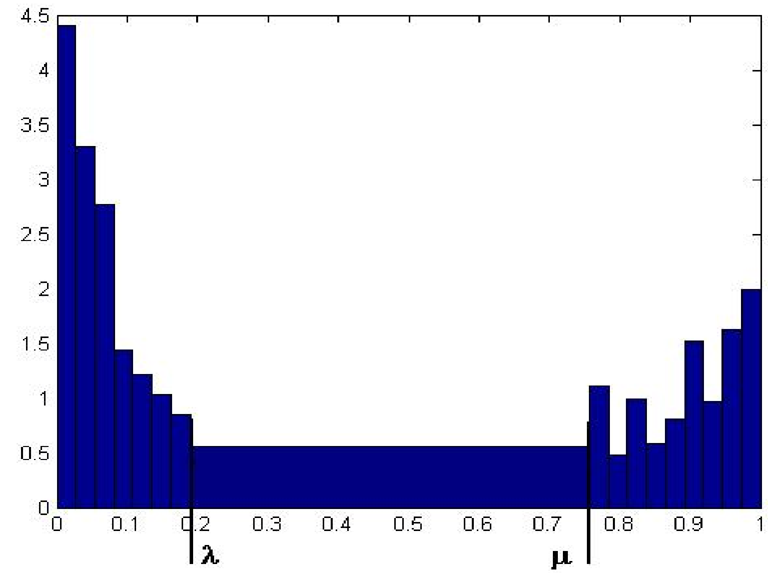
Main assumptions

1. Independence,
2. Mixture model of density : $g = \pi_0 \, \mathbb{1}_{[0,1]} + (1 - \pi_0) f$, where f is unknown,
3. It exists $[\lambda^*, \mu^*] \subset \,]0, 1]$ such that for any $P_i \in [\lambda^*, \mu^*]$, $P_i \sim \mathcal{U}(0, 1)$.

P-value density (g)



Histogram of p-values



Histograms For any partition of $[0,1]$ in D intervals I_k of length $\omega_k = |I_k|$:

$$\widehat{g}_\omega = \sum_{k=1}^{D} \frac{m_k}{m\ \omega_k} \mathbb{1}_{I_k} \quad \left(= \sum_{k=1}^{D} \frac{\sharp\{i /\ P_i \in I_k\}}{m\ \omega_k} \mathbb{1}_{I_k} \right).$$

Minimization of the L^2-risk $\mathcal{G}$: collection of all histograms.

$$g^* = \arg\min_{\widehat{g}\in\mathcal{G}} \underbrace{\left\{ \mathbb{E}_g \left[||g - \widehat{g}||_2^2 \right] - ||g||_2^2 \right\}}_{\stackrel{def}{=} R(\widehat{g})} \qquad \text{(depends on } g\text{)}.$$

Goal : Find an estimator of R : $\widehat{R}$, and then $\widetilde{g}$ such that

$$\widetilde{g} = \arg\min_{\widehat{g}\in\mathcal{G}} \widehat{R}(\widehat{g}).$$

Leave-p-out cross-validation (LPO)

- Cross-validation : a widespread and reliable method to estimate R.
- Usually leave-one-out (LOO) and V-fold are computationally intensive : at each step, you have to compute an estimator and then to assess its performance on remaining data.
- LPO is based on the same idea as LOO, but with p data instead of 1.

In our case :

- We obtain a closed formula for the LPO risk estimator : $\widehat{R}_p$ for any $p \in [\![1, m-1]\!]$.

- This formula is computationally efficient : we do not have to compute any estimator at each step (complexity of the same order as that for reading the data $\mathcal{O}(m)$).

LPO risk estimator $\forall p \in [\![1, m-1]\!]$, and any partition ω,

$$\widehat{R}_p(\omega) = \frac{2m-p}{(m-1)(m-p)} \sum_{k=1}^{D} \frac{m_k}{m\,\omega_k} - \frac{m(m-p+1)}{(m-1)(m-p)} \sum_{k=1}^{D} \frac{1}{\omega_k} \left(\frac{m_k}{m}\right)^2 .$$

Bias of the LPO risk estimator With $\forall k, \quad \alpha_k = \Pr[P_i \in I_k]$,

$$B_p(\omega) = \mathbb{E}_g \left[\widehat{R}_p(\omega) - R\left(\widehat{g}_\omega\right) \right] = \frac{p}{m(m-p)} \sum_{k=1}^{D} \frac{\alpha_k(1-\alpha_k)}{\omega_k} .$$

Remarks :

- Similar expression for the variance.
- Plug-in estimators of bias $\widehat{B}_p$ and variance $\widehat{V}_p$ are obtained replacing α_k by m_k/m in expressions.

Choice of the parameter p

Choose $\widehat{p} \in [\![1, m-1]\!]$ that realizes the best "bias-variance" trade-off according to the MSE criterion ($MSE = B_p^2 + V_p$).

Define for any partition ω

$$\begin{aligned} \widehat{p}(\omega) &= \arg \min_{p \in [\![1, m-1]\!]} \left\{ \widehat{MSE}(p, \omega) \right\}, \\ &= \arg\min_p \left\{ [\widehat{B}_p(\omega)]^2 + \widehat{V}_p(\omega) \right\}. \end{aligned}$$

Final L^2-risk estimator :

$$\forall \omega, \qquad \widehat{R}(\omega) = \widehat{R}_{\widehat{p}(\omega)}(\omega).$$

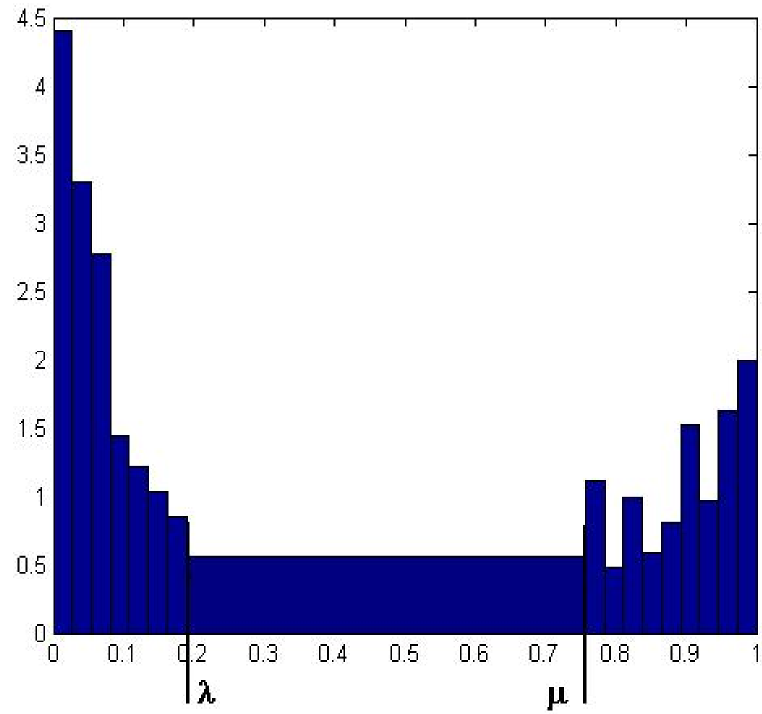
Collection of histograms

For each $N \in \{N_{\min}, \dots, N_{\max}\}$,
consider the regular partition in N intervals.

For every $1 \leq k < \ell \leq N$,
define $\lambda = k/N$ and $\mu = \ell/N$.

The resulting histogram consists in :
(i) k regular columns from 0 to λ of width $1/N$
(ii) a wide large central column from λ to μ,
(iii) $N - \ell$ regular columns of width $1/N$.

$\mathcal{G}$: collection of all these histograms.
$\mathrm{Card}(\mathcal{G}) = N_{\max}\,(N_{\max}^2 - 1)/6\,,$
$(\ N_{\min} = 1)$.



To each partition ω is associated (λ, μ) standing for edges of the widest central column.

Estimation procedure of π_0

Step 1 : $\forall \omega, \qquad \widehat{p}(\omega) = \arg\min_p \widehat{MSE}(p, \omega),$

Step 2 : $\widehat{\omega} = \arg\min_\omega \widehat{R}_{\widehat{p}(\omega)}(\omega),$

Step 3 : $\widehat{\omega} \quad \longrightarrow \quad (\widehat{\lambda}, \widehat{\mu}),$

Step 4 : $\widehat{\pi}_0 = \widehat{\pi}_0(\widehat{\lambda}, \widehat{\mu}) \stackrel{def}{=} \frac{\sharp\left\{i/\ P_i \in \left[\widehat{\lambda}, \widehat{\mu}\right]\right\}}{m(\widehat{\mu} - \widehat{\lambda})}\,.$

Theoretical result

For a given fixed collection of histograms, under independence, we obtain that

$$\widehat{\pi}_0 \quad \xrightarrow[m \to +\infty]{P} \quad \pi_0.$$

$\mathbb{IV}$ Simulations : compact support density f

Storey (2002) with $\lambda = 0.5$

Assumption : For large enough λ, each p-value larger than λ follows $\mathcal{U}(0,1)$.

$$\forall \lambda \in]0,1[, \qquad \widehat{\pi}_0(\lambda) = \frac{\sharp\{i / \ P_i \geq \lambda\}}{m(1-\lambda)} \qquad \text{(SAM : } \lambda = 0.5\text{)}.$$

Simulation design :

– $f(t) = s/\lambda^*(1 - t/\lambda^*)^{s-1} \ \mathbb{1}_{[0,\lambda^*]}(t)$, (density of H_1 p-values)

– $m = 1000$.

$\pi_0 = 0.9$	$\lambda^* = 0.2, \ s = 4$			$\lambda^* = 0.4, \ s = 6$		
Method	Bias	Variance	MSE	Bias	Variance	MSE
LPO	**0.0039**	6.25 10^{-4}	**6.41** 10^{-4}	**0.0056**	7.69 10^{-4}	**8.00** 10^{-4}
LOO	**0.0046**	5.30 10^{-4}	5.52 10^{-4}	**0.0061**	7.29 10^{-4}	7.66 10^{-4}
$\widehat{\pi}_0(0.5)$	-0.0015	9.92 10^{-4}	**9.94** 10^{-4}	0.0024	9.52 10^{-4}	**9.58** 10^{-4}

Conclusions :

– LPO less biased than LOO. MSE of $\widehat{\pi}_0(0.5)$ larger than that of LPO.

– MSE of LPO larger than that of LOO due to the $\widehat{p}$ estimation,

– Even if assumption satisfied, there may be a potential gain in choosing λ.

$\mathbb{IV}$ Simulations : density f on $[0, 1]$

General case with λ^*=1

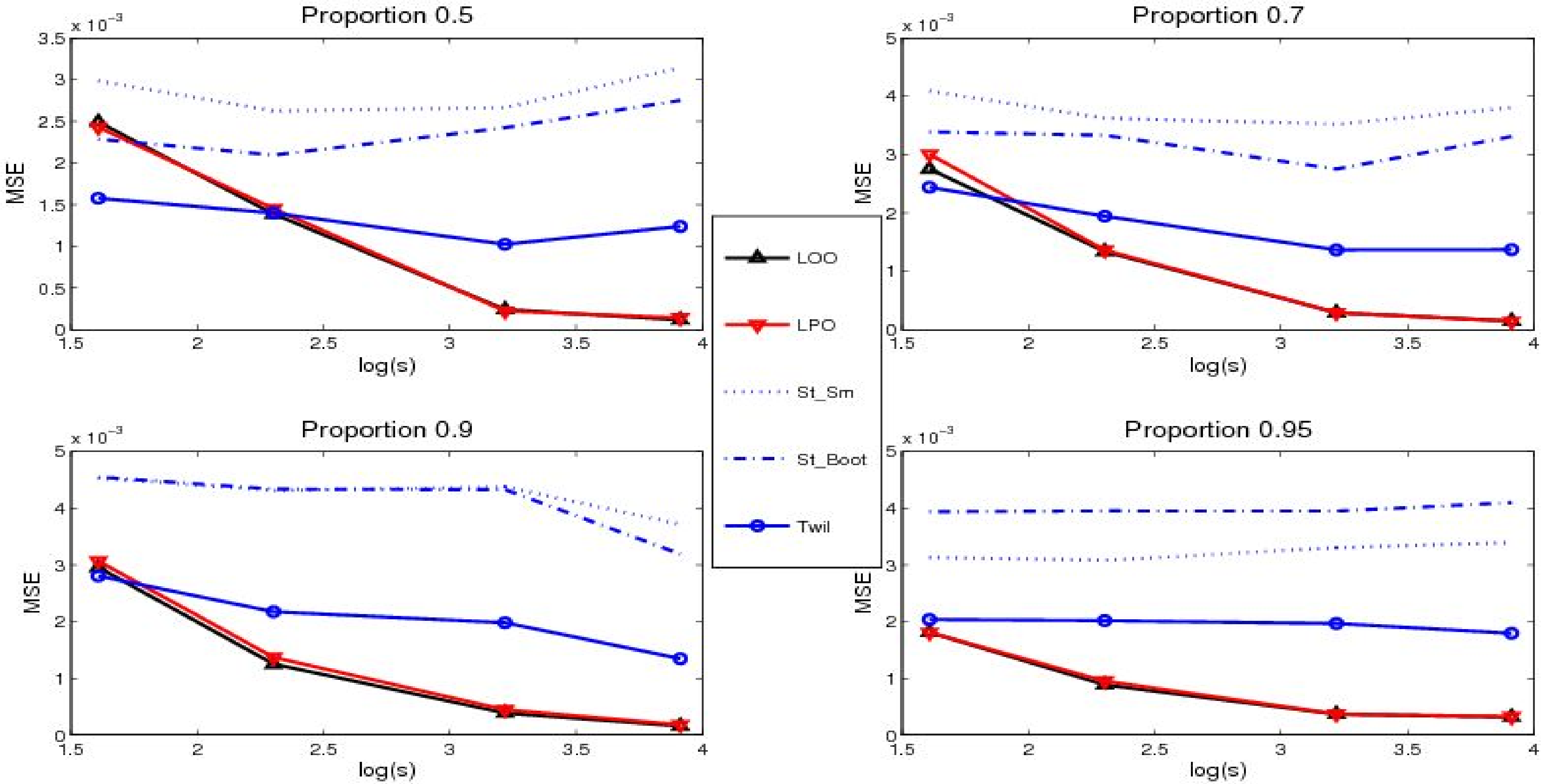
Simulation design :

- $f(t) = s(1-t)^{s-1},\ t \in [0, 1]$, with $s \in \{5, 10, 25, 50\}$,
- $m = 1000$,
- Proportion of true-null hypotheses : 0.5, 0.7, 0.9, 0.95 .

Comparison of different methods :

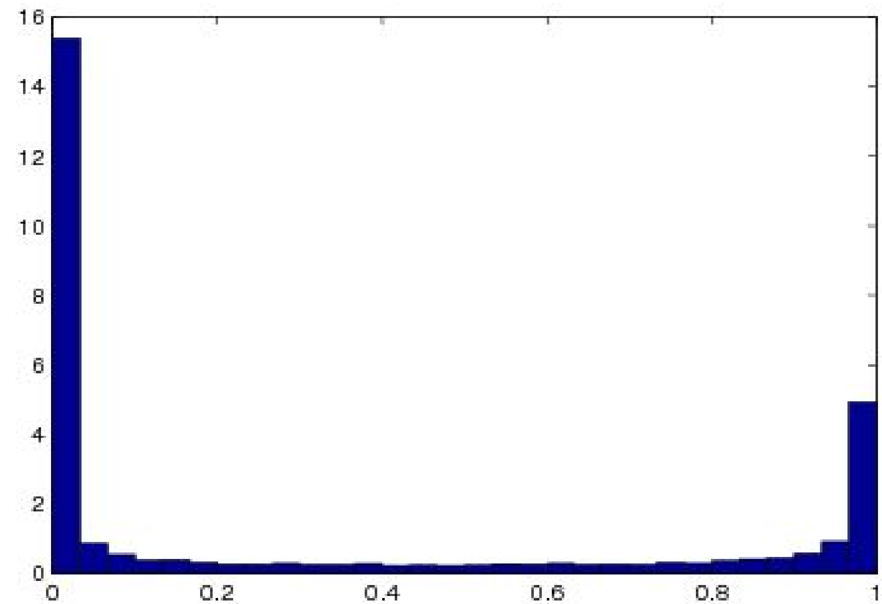
1. LPO : proposed estimator of π_0 based on leave-p-out,
2. LOO : LPO with $p = 1$,
3. $Bootstrap$: Storey (2002), based on bootstrap and MSE,
4. $Smoother$: Storey et al.(2003), relying on spline adjustment,
5. $Twilight$: Scheid et al.(2004), based on both minimization of a penalized criterion and bootstrap.

$\mathbb{IV}$ Simulations : density on $[0, 1]$



U-shape density of real data

Histogram of pooled p-values (Pounds et al.(2005))



Pounds et al.(2005) observed a U-shape on real data, for Affymetrix present-absent p-values.

It appears in one-sided tests when non tested alternative is true.

Simulation design : (Test of $\mu = 0$ against $\mu > 0$.)

- $m = 1000$,
- Data simulated $\sim \pi_0 \mathcal{N}(0, 0.75) + \frac{1-\pi_0}{2} \mathcal{N}(\mu, 0.75) + \frac{1-\pi_0}{2} \mathcal{N}(-\mu, 0.75)$,
- $\mu \in \{1, 1.5\}$.

Comparison in the U-shape case

MSE :

π_0	0.25	0.5	0.7	0.8	0.9
LPO	**0.0068**	**0.0057**	**0.0047**	**0.0044**	**0.0024**
LOO	**0.0071**	**0.0078**	**0.0066**	**0.0057**	**0.0028**
Smoother	0.56	0.25	0.09	0.04	0.0098
Bootstrap	0.187	0.084	0.03	0.01	0.0032
Twilight	0.536	0.226	0.08	0.03	0.0066

Conclusions :

- LPO has lower MSE than LOO,
- The gap between LPO/LOO and other methods decreases as π_0 grows, but still in favor of LPO/LOO.

Conclusion :

- Our estimator of π_0 relies on a LPO risk estimator,
- It is not computation-time consuming,
- This estimator seems to outperform other tested methods in the general framework,
- LPO estimator is still reliable even in the case of U-shape density, where other methods highly overestimate π_0.

Thank you!