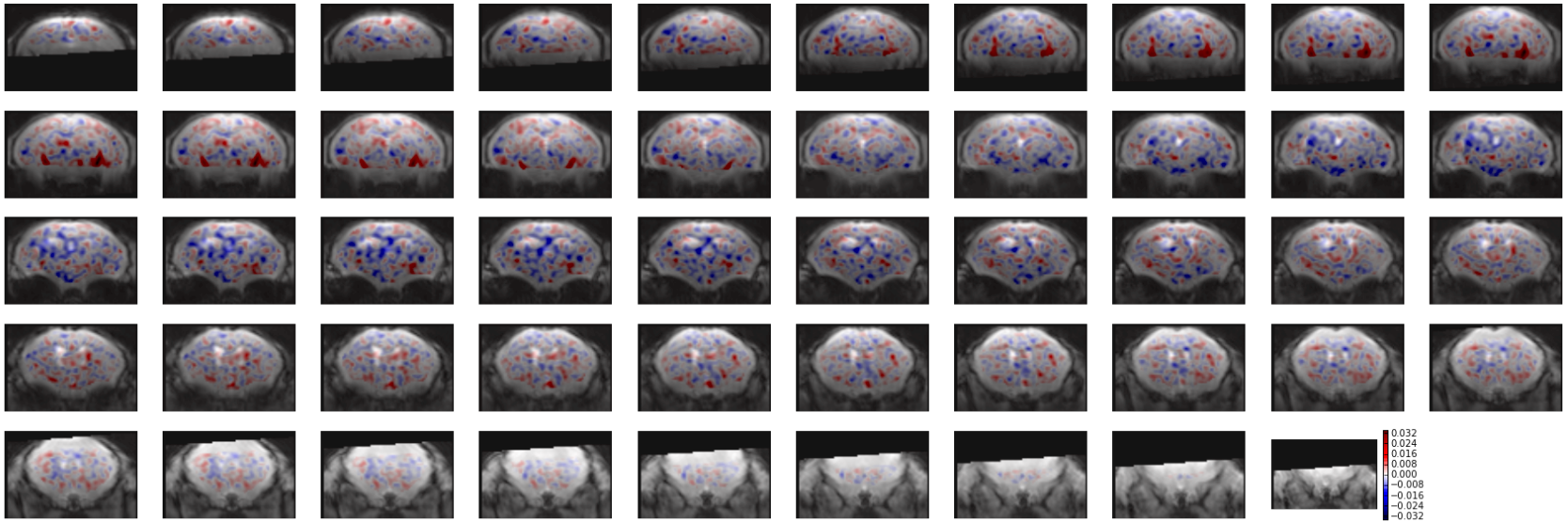
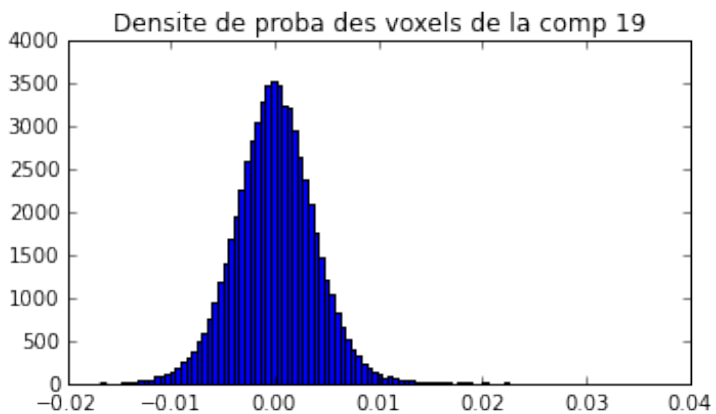
Seuillage en grande dimension : une histoire de p-valeurs, de contrôle global des erreurs et de modélisation précise des données sous $\mathcal{H}_0$

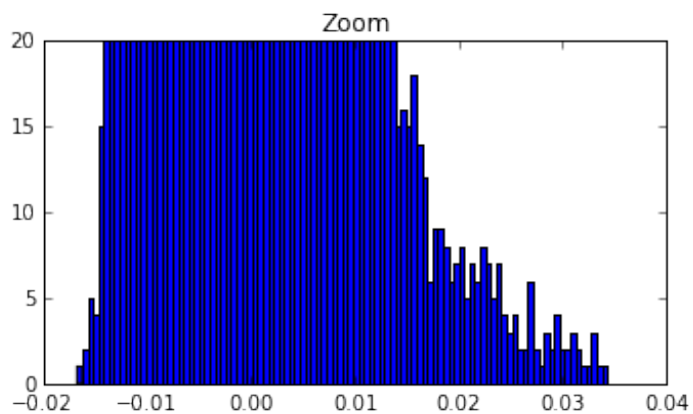
Céline Meillier

Séminaire IMAGeS, 21 juin 2018

Seuillage de cartes en IRM fonctionnelle : composante ICA

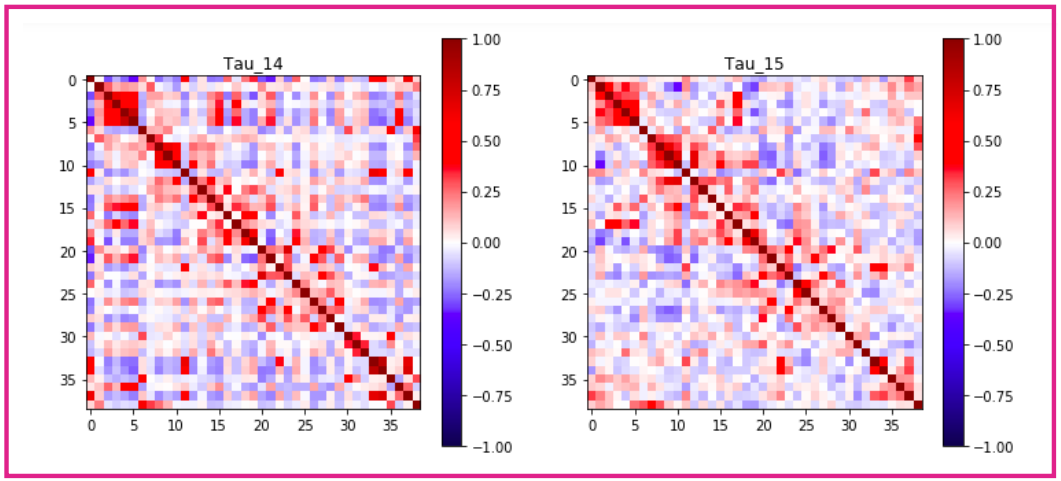


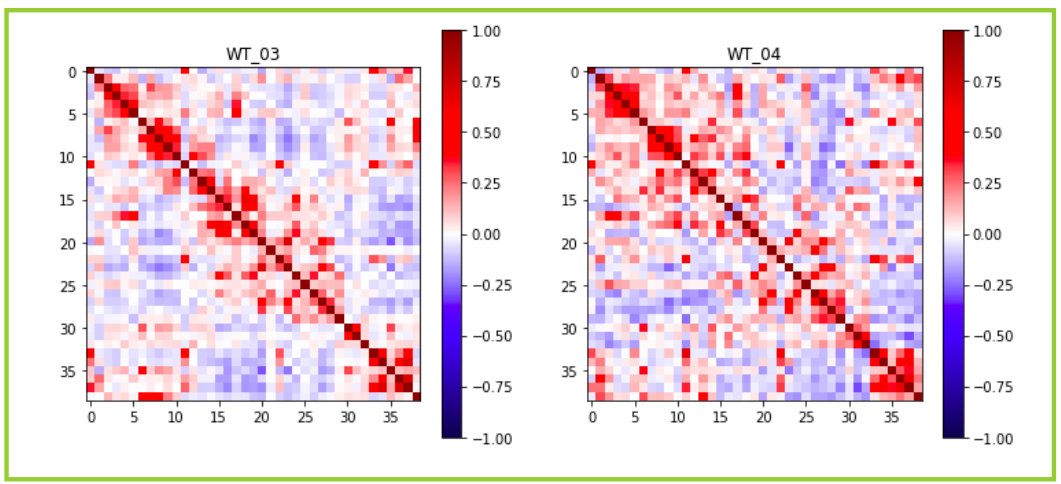




Comparaisons multiples en IRM fonctionnelle : correlations entre ROIs

Souris Alzheimer



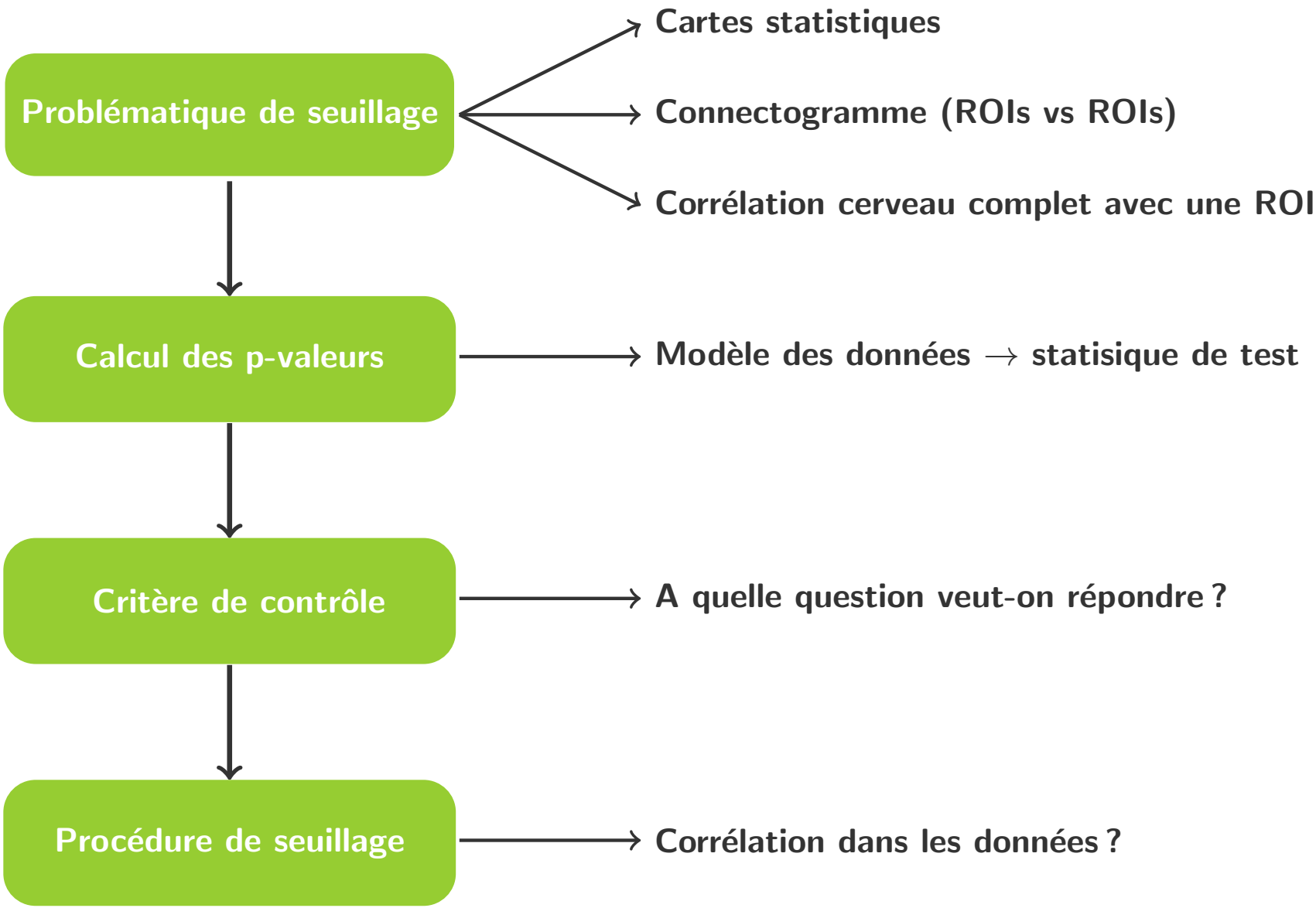


Souris contrôle

Comparaison des corrélations moyennes

Sélection des couples de ROIs significativement différentes ?

Contexte en neuroscience/IRMf



Plan de la présentation

Plan de la présentation

Tests multiples

Définition d'un test :

Soit x une variable à tester et T une statistique de test. La règle de décision associée à T s'écrit :

$$\begin{cases} \mathcal{H}_0 & : \quad T(x) < \eta \quad \text{(absence de signal)} \\ \mathcal{H}_1 & : \quad T(x) \geqslant \eta \quad \text{(présence d'un signal)}, \end{cases}$$

Tests multiples :

Soit $\{x_1, x_2, \cdots, x_n\}$ un ensemble de variables à tester.

On définit pour chaque variable x_i le couple d'hypothèses $\mathcal{H}_0^i / \mathcal{H}_1^i$.

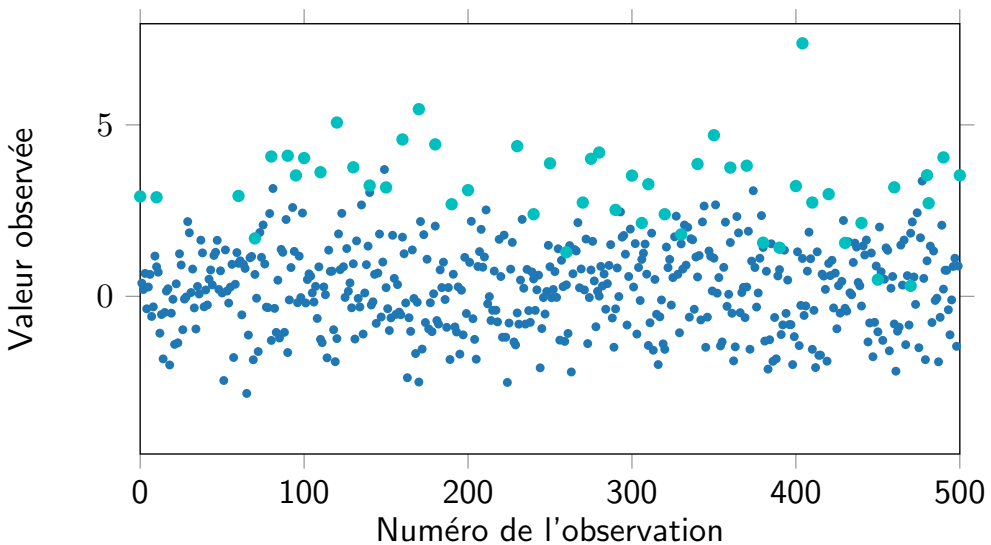
Comment choisir le seuil de décision ?

Un exemple illustratif

Soit $N = 500$ échantillons simulés indépendamment. A chaque échantillon le modèle d'hypothèses binaire suivant est associé :

$$\begin{cases} \mathcal{H}_0^i & : \quad x_i \sim \mathcal{N}(0, \sigma^2) & \text{(bruit seul)} \\ \mathcal{H}_1^i & : \quad x_i \sim \mathcal{N}(\theta, \sigma^2)\ , \quad \theta > 0 & \text{(source + bruit)} \end{cases} .$$

La proportion d'échantillons simulés suivant $\mathcal{H}_0$ est $\pi_0 = 0.9$ ($n_0 = 450$).



Comment choisir le seuil de décision ?

P-valeur : définition et interprétation

Définition de la p-valeur

La p-valeur associée à une observation x_i désigne la probabilité pour $T(x)$ d'être au moins aussi extrême que le résultat $T(x_i)$ du test sur l'observation x_i si l'hypothèse $\mathcal{H}_0$ est vraie.

Pour un test T sur la valeur x_i la p-valeur s'écrit :

$$p_{val}(x_i) = \Pr(T(x) > T(x_i)|\mathcal{H}_0).$$

$\rightarrow$ $\forall x_i$, $p_{val}(x_i) \in [0, 1]$

$\rightarrow$ Sous $\mathcal{H}_0$, $p_{val} \sim \mathcal{U}_{[0,1]}$

Suppose $\mathcal{H}0$ bien spécifiée !

P-valeur : définition et interprétation

Définition de la p-valeur

La p-valeur associée à une observation x_i désigne la probabilité pour $T(x)$ d'être au moins aussi extrême que le résultat $T(x_i)$ du test sur l'observation x_i si l'hypothèse $\mathcal{H}_0$ est vraie.

Pour un test T sur la valeur x_i la p-valeur s'écrit :

$$p_{val}(x_i) = \Pr(T(x) > T(x_i)|\mathcal{H}_0).$$

$\rightarrow \forall x_i,\ p_{val}(x_i) \in [0,1]$

$\rightarrow$ Sous $\mathcal{H}_0$, $p_{val} \sim \mathcal{U}_{[0,1]}$

Suppose $\mathcal{H}0$ bien spécifiée !

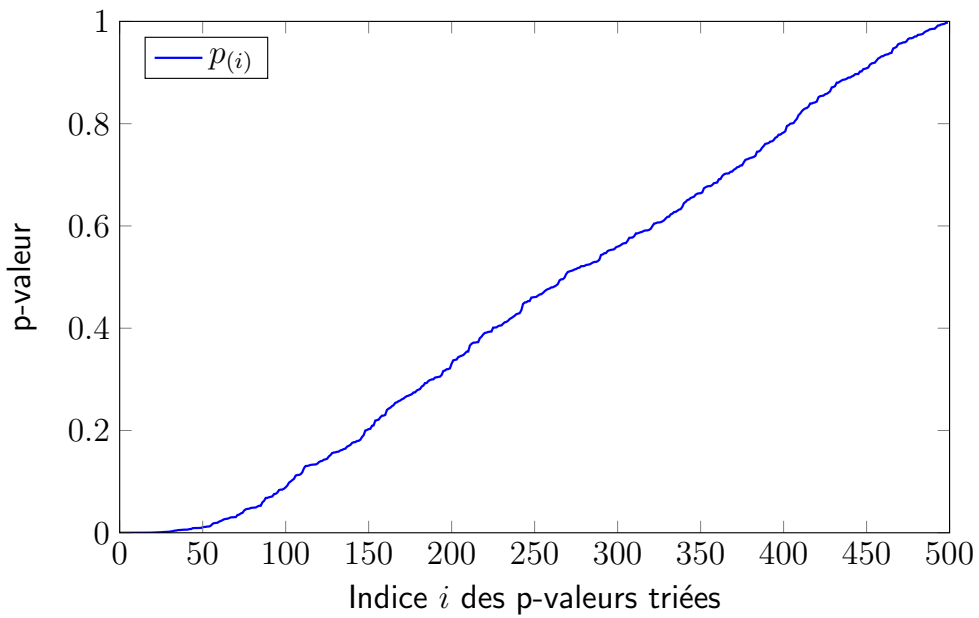
Quantité indépendante de la statistique de T !

On peut comparer la pvaleur issue d'un t-test avec celle issue d'un z-test !

P-valeur : illustration

Soit $N = 500$ échantillons simulés indépendamment. A chaque échantillon le modèle d'hypothèses binaire suivant est associé :

$$\begin{cases} \mathcal{H}_0^i & : \quad x_i \sim \mathcal{N}(0, \sigma^2) & \text{(bruit seul)} \\ \mathcal{H}_1^i & : \quad x_i \sim \mathcal{N}(\theta, \sigma^2)\ , \quad \theta > 0 & \text{(source + bruit)} \end{cases} .$$



Tests multiples et contrôle des erreurs

Répartition des n tests :

Vérité \ Décision	$\hat{\mathcal{H}}_0$	$\hat{\mathcal{H}}_1$	Total
$\mathcal{H}_0$	n_0 - a	a	n_0
$\mathcal{H}_1$	n_1 - b	b	n_1
Total	n - R	R	n

fausses alarmes

détections

Choix du type de contrôle :

p_{FA}	FWER	FDR
$Pr(\hat{\mathcal{H}}_1 \mid \mathcal{H}_0) \leqslant \alpha$	$Pr(a \geqslant 1) \leqslant \alpha$	$E\left[\frac{a}{R}\right] \leqslant q$
$a \nearrow$ si $n \nearrow$	n grand	n grand
Contrôle individuel	Conservatif	Contrôle global

FWER : Family wise error rate

Objectif : Contrôler la probabilité de se tromper au moins une fois en dessous d'un seuil α.

Procédures de seuillage des p-valeurs associées aux n tests :

- $\rightarrow$ Bonferronni
- $\rightarrow$ Holm
- $\rightarrow$ Théorie des champs aléatoires : FWER appliqué aux clusters

A utiliser quand veut à tout prix éviter de faire une fausse alarme parmi tous les tests réalisés

FDR : False discovery rate

Objectif : Contrôler la proportion de fausses découvertes en dessous d'un seuil α.

Procédures de seuillage des p-valeurs associées aux n tests :

- $\rightarrow$ Benjamini-Hochberg (1995) : tests indépendants,
- $\rightarrow$ Benjamini-Yekutieli (2001) : certains types de corrélations entre les tests,
- $\rightarrow$ Théorie des champs aléatoires : FDR topologique (Chumbley et al. 2009)

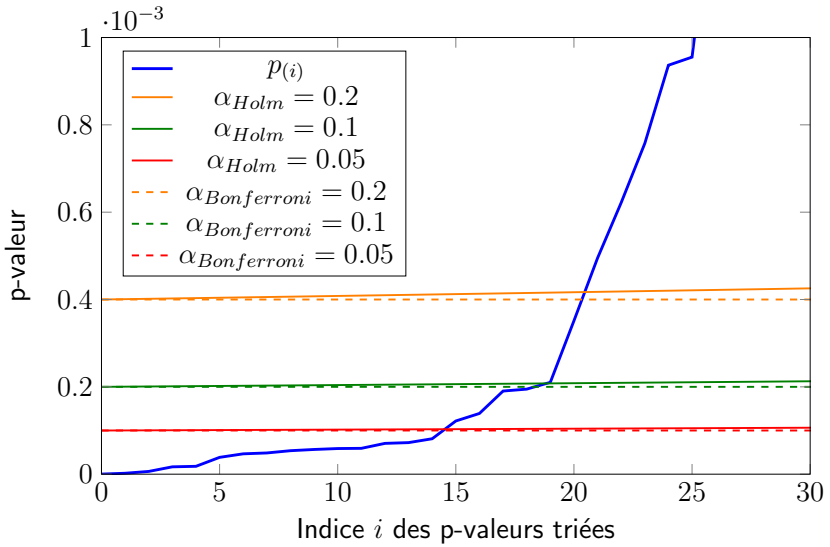
Contrôle moins conservatif $\rightarrow$ on s'autorise une certaine proportion d'erreurs.

$\nearrow$ du nombre b de bonnes détections, au prix d'une $\nearrow$ de la proportion $\frac{a}{R}$ de fausses découvertes

Un exemple illustratif

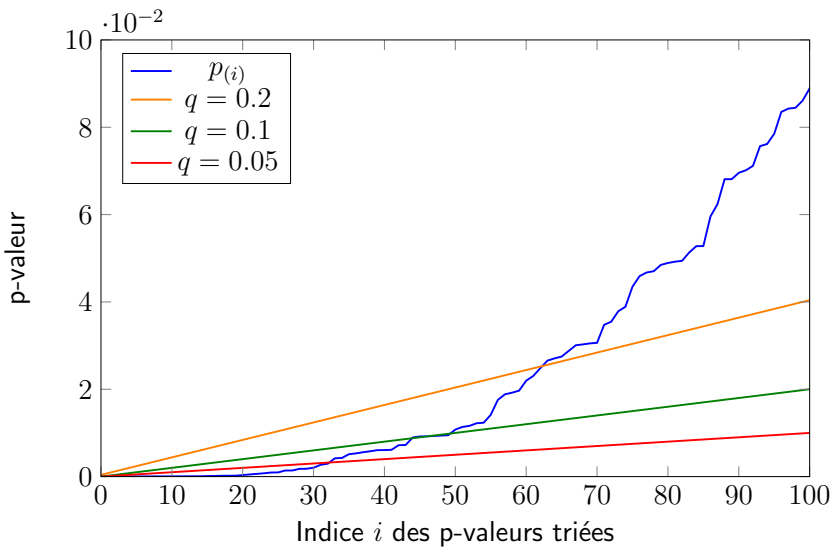
Contrôle du FWER

par Holm et par Bonferroni



Contrôle du FDR

par Benjamini-Hochberg



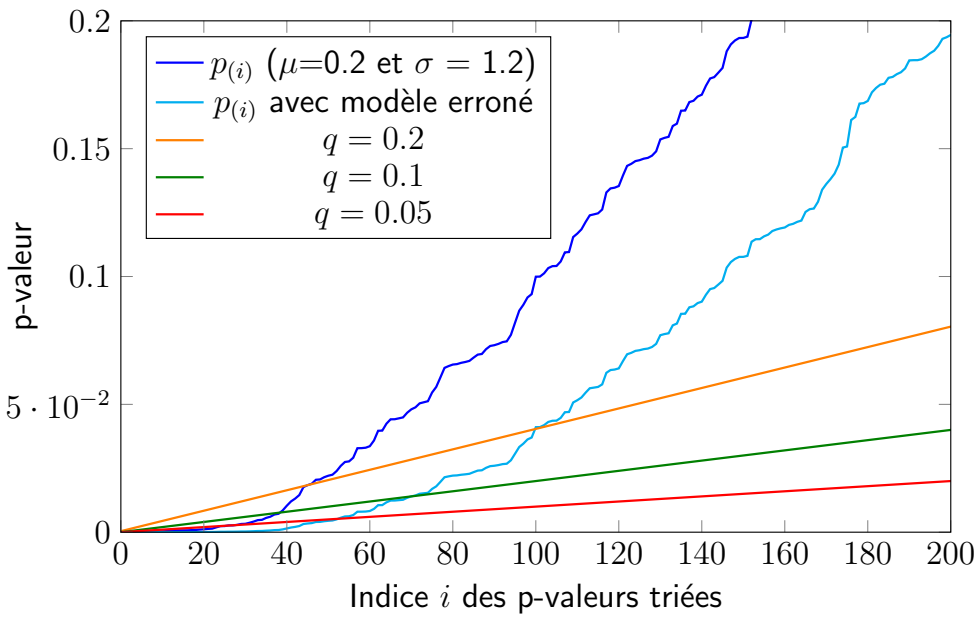
Plan de la présentation

Importance de bien modéliser la loi statistique des tests/des données

Calcul des p-valeurs avec un modèle erroné $\rightarrow$ plus de garantie de contrôler le FDR (ou le critère d'erreur choisi).

Modèle théorique : $\mathcal{H}_0 : x_i \sim \mathcal{N}(0, 1)$

Réalité : $\mathcal{H}_0 : x_i \sim \mathcal{N}(0.2, 1.2)$



Importance de bien modéliser la loi statistique des tests/des données

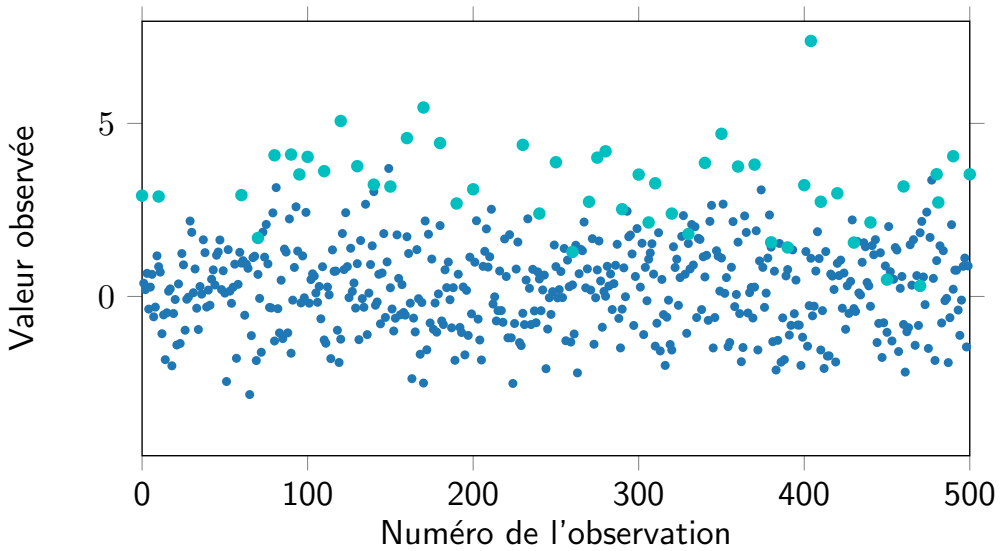
Constats :

- → Les modèles théoriques sont parfois éloignés de la réalité des données.
- → Les données sont corrélées alors qu'elles sont supposées indépendantes.
- → Pour obtenir la loi empirique sous $\mathcal{H}_0$ il faudrait disposer de données de calibration.

Travaux "récents" :

- → Estimation de la loi des données sous $\mathcal{H}_0$ à partir des données (en présence des échantillons $\mathcal{H}_1$).
- → Comparaison de matrices de corrélation (études de groupe ou longitudinale) : test sur les rangs des corrélations plutôt que sur les valeurs → comment définir la statistique du test ?

Le contexte en image

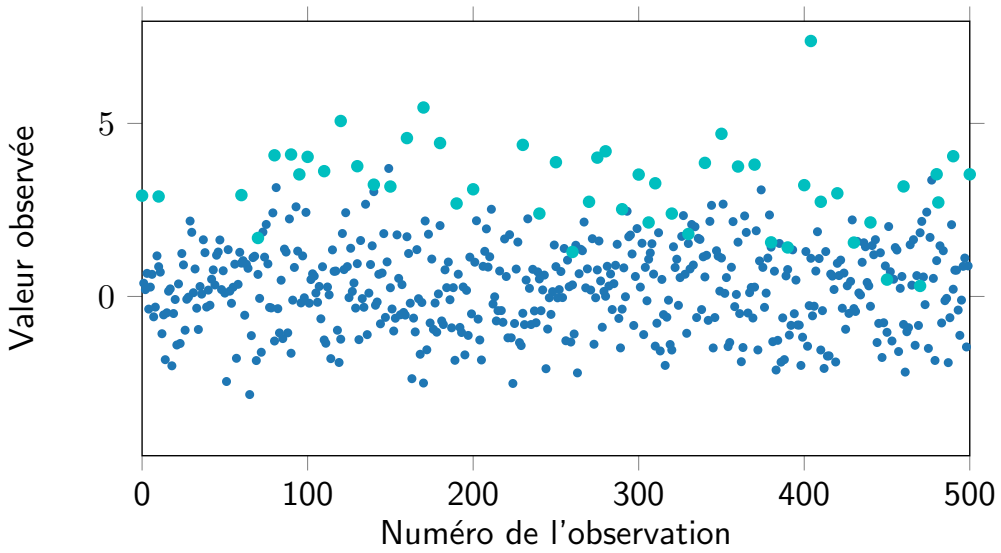


N observations x_i à classer dans $\mathcal{H}_0/\mathcal{H}_1$

$\mathcal{H}_0 : x_i \sim \mathcal{N}(\mu_0, \sigma_0)$
$\mathcal{H}_1 : x_i \sim \mathcal{N}(\mu_0 + \beta, \sigma_0)$

$\mu_0 = 0.2$, $\sigma_0 = 1.2$ et $\beta = 3$

Le contexte en image



N observations x_i à classer dans $\mathcal{H}_0/\mathcal{H}_1$

$\mathcal{H}_0 : x_i \sim \mathcal{N}(\mu_0, \sigma_0)$
$\mathcal{H}_1 : x_i \sim \mathcal{N}(\mu_0 + \beta, \sigma_0)$

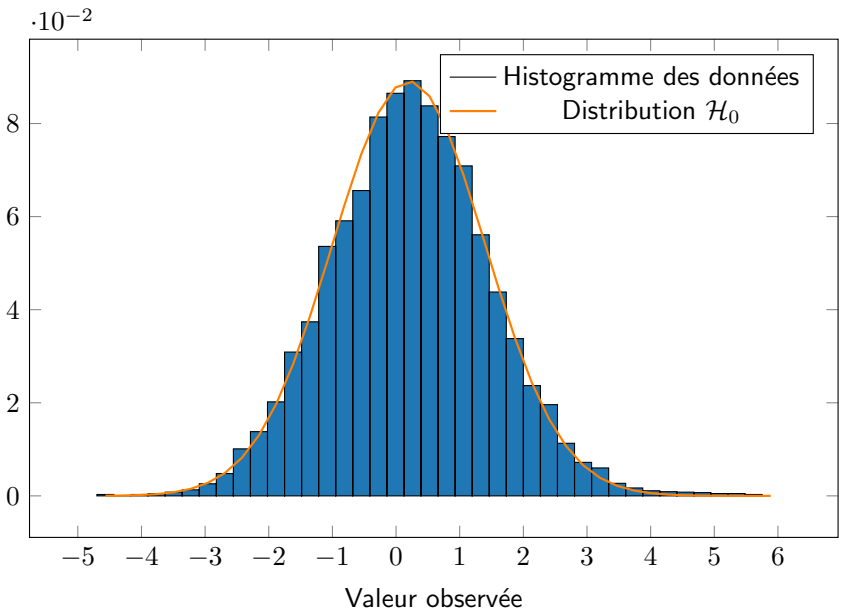
$\mu_0 = 0.2$, $\sigma_0 = 1.2$ et $\beta = 3$

Distribution des données :

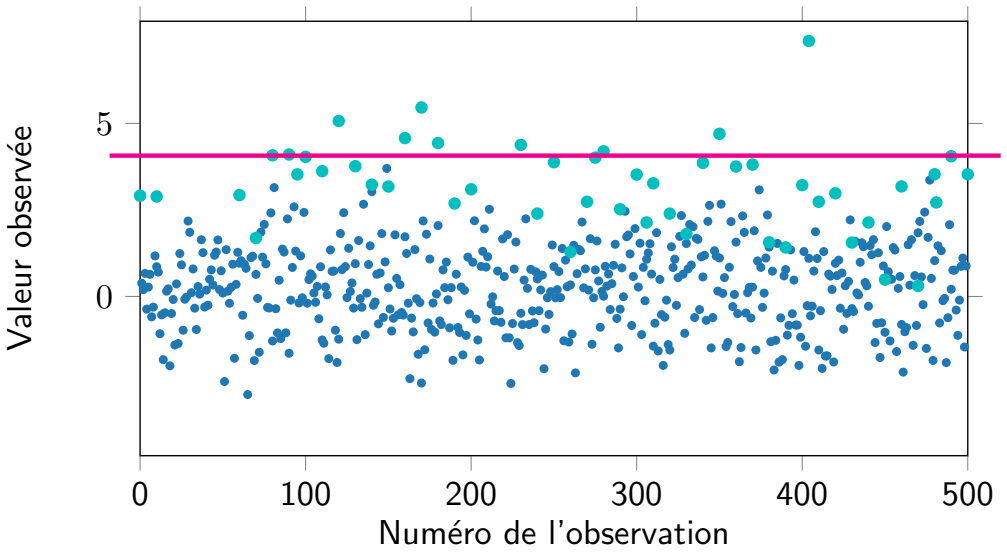
$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$$

π_0 = prop. de données sous $\mathcal{H}_0$

et $\pi_1 = 1 - \pi_0$



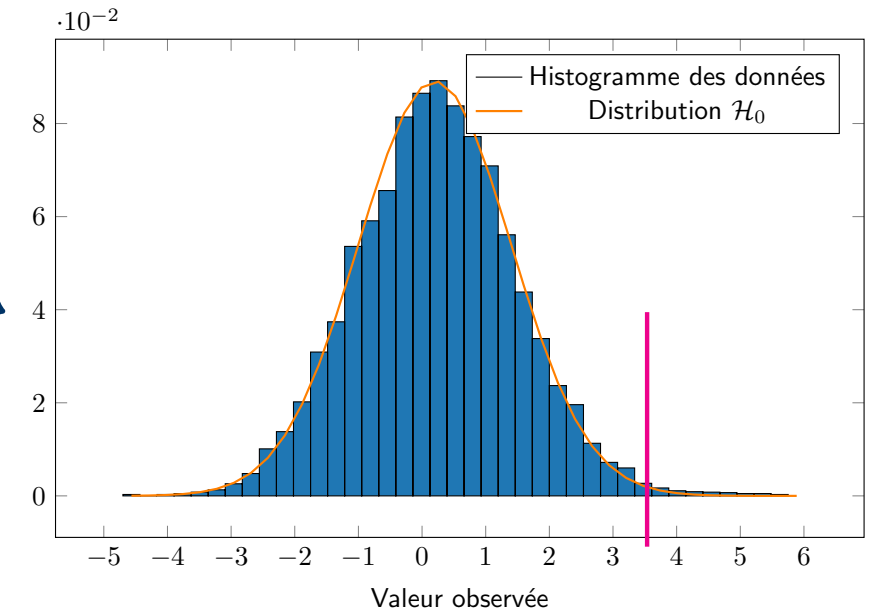
Le contexte en image



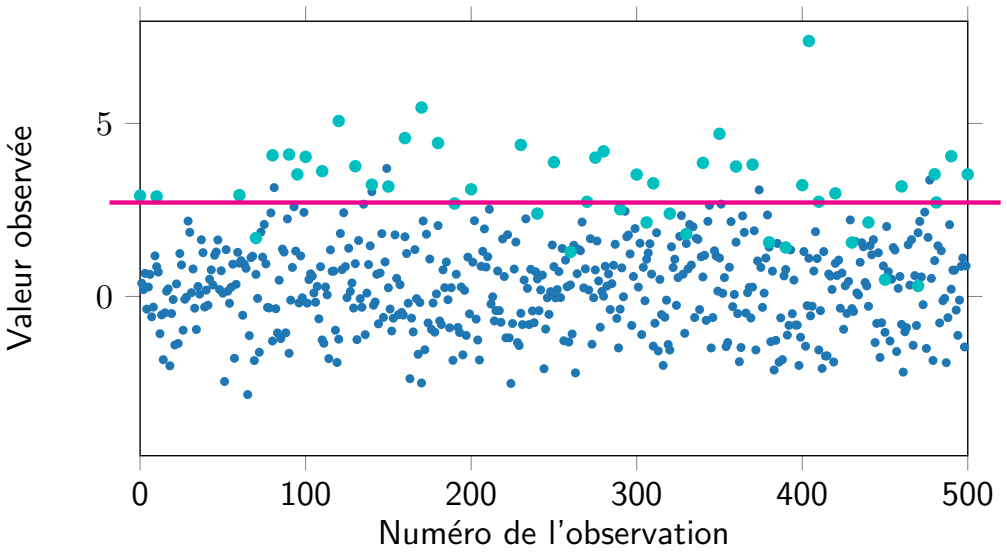
N observations x_i à classer dans $\mathcal{H}_0/\mathcal{H}_1$

$\mathcal{H}_0 : x_i \sim \mathcal{N}(\mu_0, \sigma_0)$
$\mathcal{H}_1 : x_i \sim \mathcal{N}(\mu_0 + \beta, \sigma_0)$

$\mu_0 = 0.2$, $\sigma_0 = 1.2$ et $\beta = 3$



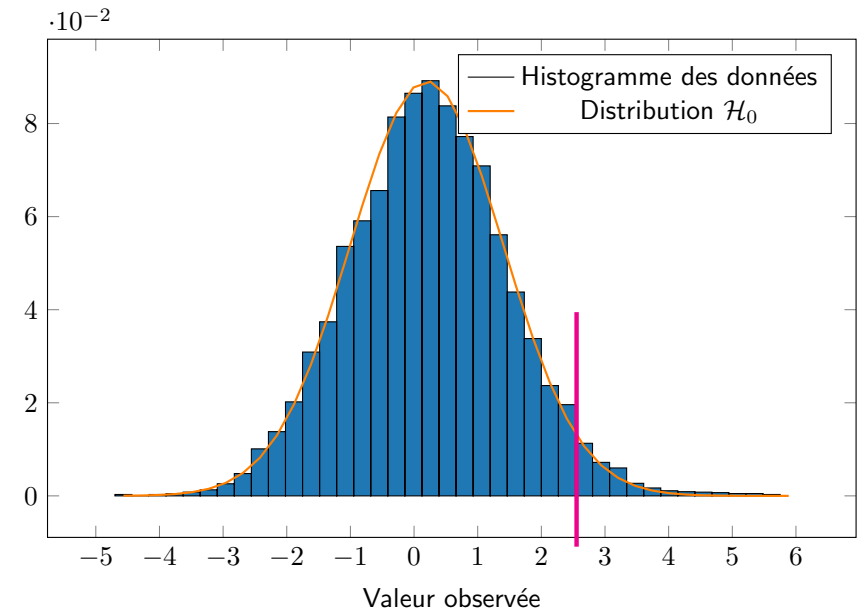
Le contexte en image



N observations x_i à classer dans $\mathcal{H}_0/\mathcal{H}_1$

$\mathcal{H}_0 : x_i \sim \mathcal{N}(\mu_0, \sigma_0)$
$\mathcal{H}_1 : x_i \sim \mathcal{N}(\mu_0 + \beta, \sigma_0)$

$\mu_0 = 0.2$, $\sigma_0 = 1.2$ et $\beta = 3$



Caractérisation de la loi sous $\mathcal{H}_0$

Constat :
Quelle que soit la procédure de seuillage choisie, cela nécessite de connaître la loi des échantillons sous $\mathcal{H}_0$ ou plus généralement la statistique de test sous $\mathcal{H}_0$.

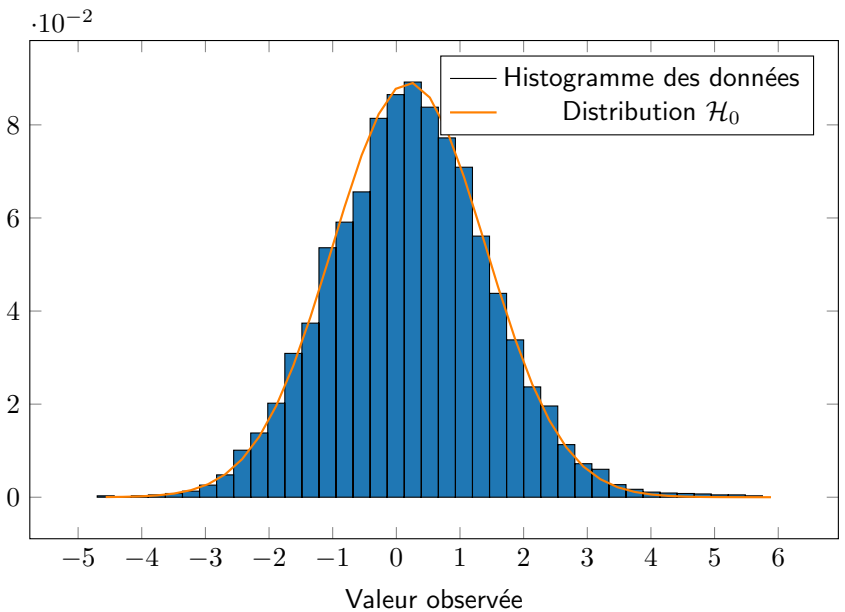
Caractérisation de la loi sous $\mathcal{H}_0$

Constat :
Quelle que soit la procédure de seuillage choisie, cela nécessite de connaître la loi des échantillons sous $\mathcal{H}_0$ ou plus généralement la statistique de test sous $\mathcal{H}_0$.

Comment fait-on ?

$\rightarrow$ utilisation d'un modèle empirique (calibration en l'absence de signal)

$\rightarrow$ utilisation d'un modèle paramétrique pour expliquer les données sous $\mathcal{H}_0$
ex : loi $\mathcal{N}(\mu_0, \sigma_0)$, μ_0 et σ_0 à estimer.



Etat de l'art des méthodes d'estimation robuste

Modification du modèle statistique du bruit :

$\rightarrow$ distribution à queue lourde pour modéliser les incertitudes dûes aux échantillons $\mathcal{H}_1$.

Etat de l'art des méthodes d'estimation robuste

Modification du modèle statistique du bruit :

→ distribution à queue lourde pour modéliser les incertitudes dûes aux échantillons $\mathcal{H}_1$.

Modification de la fonction objectif :

→ M-estimateurs,

→ pénalisation de Huber, etc.

Etat de l'art des méthodes d'estimation robuste

Modification du modèle statistique du bruit :

→ distribution à queue lourde pour modéliser les incertitudes dûes aux échantillons $\mathcal{H}_1$.

Modification de la fonction objectif :

→ M-estimateurs,

→ pénalisation de Huber, etc.

Estimation à partir de statistiques d'ordre ou sur des données tronquées

→ L-estimateurs,

→ σ-clipping : ne converge pas !

→ ajustement du mode central avec une régression de Poisson ([Schwartzman 2008]),

→ ajustement au sens du maximum de vraisemblance à la distribution des données tronquées (MLE, [Efron 2012]).

Procédure de σ-clipping par point fixe (FRONDE)

Collaborations :

Raphael Bacher

Florent Chatelain

Olivier Michel

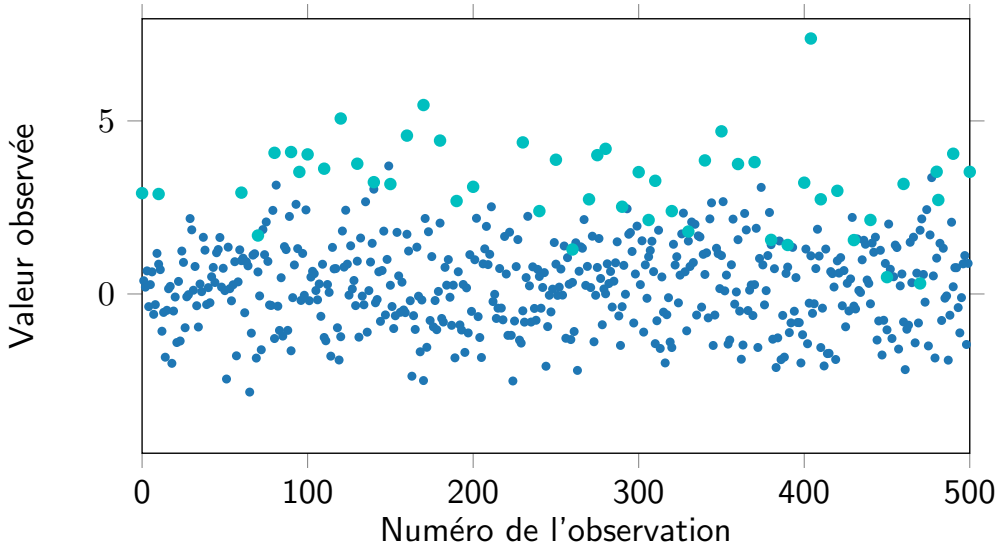
gipsa-lab

Méthode FRONDE : *Fixed-point algorithm for RObust Null Distribution Estimation*

[*Méthode de sigma-clipping par point fixe pour l'estimation de la distribution sous $\mathcal{H}_0$ dans le cadre de tests multiples*, C.Meillier, R. Bacher, F. Chatelain, O. Michel, Gretsi 2017]

Procédure de σ-clipping par point fixe (FRONDE)

Cadre :



N observations x_i à classer dans $\mathcal{H}_0/\mathcal{H}_1$

$\mathcal{H}_0 : x_i \sim \mathcal{N}(\mu_0, \sigma_0)$
$\mathcal{H}_1 : x_i \sim \mathcal{N}(\mu_0 + \beta, \sigma_0)$

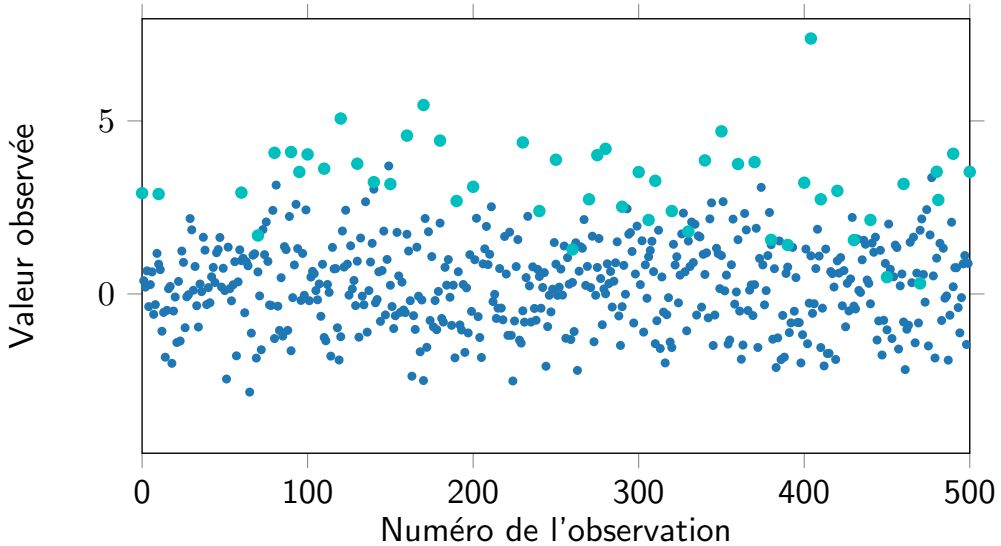
avec $\beta > 0$

Fonction de répartition des données observées

$$F(x) = \pi_0 F_0(x) + \pi_1 F_1(x) \tag{1}$$

Procédure de σ-clipping par point fixe (FRONDE)

Cadre :



N observations x_i à classer dans $\mathcal{H}_0/\mathcal{H}_1$

$\mathcal{H}_0 : x_i \sim \mathcal{N}(\mu_0, \sigma_0)$
$\mathcal{H}_1 : x_i \sim \mathcal{N}(\mu_0 + \beta, \sigma_0)$

avec $\beta > 0$

Fonction de répartition des données observées

$$F(x) = \pi_0 F_0(x) + \pi_1 F_1(x) \tag{1}$$

Objectif : estimer F_0 à partir de F (empirique).

$\rightarrow$ F_0 paramétrique : $F_0(x) = \Phi(\frac{x-\mu_0}{\sigma_0})$

Troncature des données

Principe du σ-clipping

- $\rightarrow$ troncature des données à $\pm\kappa\sigma_0$ autour de μ_0,
- $\rightarrow$ écarter les échantillons aberrants ($\mathcal{H}_1$) afin de ne pas biaiser l'estimation de μ_0 et σ_0

Troncature des données

Principe du σ-clipping

- $\rightarrow$ troncature des données à $\pm\kappa\sigma_0$ autour de μ_0,
- $\rightarrow$ écarter les échantillons aberrants ($\mathcal{H}_1$) afin de ne pas biaiser l'estimation de μ_0 et σ_0

Troncature des données dans le cadre de tests à grande échelle

- $\rightarrow$ estimation très robuste : biais faible,
- $\rightarrow$ conservation d'un nombre suffisamment grand de données : faible variance de l'estimateur.

Troncature des données

Fonction de répartition des données après troncature

$$F_t(x) = \frac{F(x) - F(l)}{F(r) - F(l)} \tag{2}$$

avec dans notre cas une troncature symétrique : $l = \mu_0 - \kappa\sigma_0$ et $r = \mu_0 + \kappa\sigma_0$.

Troncature des données

Fonction de répartition des données après troncature

$$F_t(x) = \frac{F(x) - F(l)}{F(r) - F(l)} \tag{2}$$

avec dans notre cas une troncature symétrique : $l = \mu_0 - \kappa\sigma_0$ et $r = \mu_0 + \kappa\sigma_0$.

Postulat (P1) : La probabilité d'observer un échantillon distribué selon $\mathcal{H}_1$ est nulle sur le domaine de troncature : $F_1(r) = F_1(l) = 0$

Troncature des données

Fonction de répartition des données après troncature

$$F_t(x) = \frac{F(x) - F(l)}{F(r) - F(l)} \tag{2}$$

avec dans notre cas une troncature symétrique : $l = \mu_0 - \kappa\sigma_0$ et $r = \mu_0 + \kappa\sigma_0$.

Postulat (P1) : La probabilité d'observer un échantillon distribué selon $\mathcal{H}_1$ est nulle sur le domaine de troncature : $F_1(r) = F_1(l) = 0$

Expression des quartiles $q_{i,t}$ de la loi tronquée (d'après P1)

$$\underbrace{\frac{i}{4} = F_t(q_{i,t})}_{\text{données tronquées}} = \frac{F_0(q_{i,t}) - F_0(l)}{F_0(r) - F_0(l)} = \underbrace{\frac{F_0(q_{i,t}) - F_0(l)}{1 - 2F_0(l)}}_{\text{loi paramétrique } F_0 \text{ non tronquée}}, \tag{3}$$

Equations du point fixe

On a les équations suivantes :

$$\mu_0 = q_{2,t} \text{ et } \sigma_0 = \frac{q_{3,t} - q_{1,t}}{\lambda_\kappa} \tag{4}$$

avec

- $\rightarrow$ $\lambda_\kappa = 2\Phi^{-1}\left(\frac{1}{2}\left(\frac{1}{2} + \Phi(\kappa)\right)\right)$ qui est une constante
- $\rightarrow$ $q_{1,t}$, $q_{2,t}$ et $q_{3,t}$ qui dépendent de μ_0, σ_0 et Φ

Ce qui donne des équations du point fixe (μ_0, σ_0) :

$$\mu_0 = g_1(\mu_0, \sigma_0) = q_{2,t}, \tag{5}$$

$$\sigma_0 = g_2(\mu_0, \sigma_0) = (q_{3,t} - q_{1,t})/\lambda_\kappa, \tag{6}$$

Algorithme

Entrées : ensemble des données $I = \{x_1, \ldots, x_N\}$, facteur de troncature κ

$\mu_0^0 \leftarrow$ médiane empirique(I)
$\sigma_0^0 \leftarrow$ écart-type empirique(I)
$k \leftarrow 0$

while (μ_0^k, σ_0^k) n'a pas convergé **do**
$k \leftarrow k + 1$
$I_k \leftarrow \{x_i \ : \ |x_i - \mu_0^{k-1}| \leqslant \kappa \sigma_0^{k-1}\}$
$\mu_0^k \leftarrow$ médiane(I_k)
$q_{1,t} \leftarrow$ quartile$(I_k, 25\%)$
$q_{3,t} \leftarrow$ quartile$(I_k, 75\%)$
$\sigma_0^k \leftarrow (q_{3,t} - q_{1,t})/\lambda_\kappa$.

Sorties : μ_0^k, σ_0^k

Convergence de l'algorithme

→ difficile d'établir théoriquement les conditions de convergence dans le cas où μ_0 et σ_0 sont inconnus,

→ si σ_0 fixée, $(\mu_0^k)_k$ = suite monotone et bornée, donc convergente,

→ par construction μ_0 et σ_0 peuvent prendre un nombre fini de valeurs, et en pratique il arrive parfois qu'on observe un cycle d'ordre 1 pour (μ_0^k,σ_0^k),

→ dans ce cas : σ_0 est fixé à la valeur la plus élevée du cycle (valeur la + conservative au sens du test d'hypothèse), et μ_0 est ensuite mis à jour seul.

Consistance des estimateurs

$\rightarrow$ $\bar{F}$ converge uniformément vers F quand N$\nearrow$,

$\rightarrow$ dans le cas gaussien, sous (P1), $g(\mu_0, \sigma_0) = (g_1(\mu_0, \sigma_0), g_2(\mu_0, \sigma_0))$ est contractante au voisinage du point fixe,

$\rightarrow$ l'approximation empirique de $g(\mu_0, \sigma_0)$ converge vers (μ_0, σ_0).

Données synthétiques

Données proposées dans [Schwartzman 2008] :

- → 1000 cubes de taille $64 \times 64 \times 64$ contenant un champ aléatoire gaussien,
- → données i.i.d. $\mathcal{N}(0, 1)$ convoluées par un noyau gaussien de $\sigma = 1.5$,
- → décalage et mise à l'échelle du champ gaussien afin que $\mu_0 = 0.2$ et $\sigma_0 = 1.2$,
- → ajout d'un signal constant $\beta = 3$ dans un sous cube de taille $T \times T \times T$ $(\mathcal{H}_1)$
- → $\pi_0 = 1 - \frac{T^3}{64^3}$

Données synthétiques

Données proposées dans [Schwartzman 2008] :

- → 1000 cubes de taille $64 \times 64 \times 64$ contenant un champ aléatoire gaussien,
- → données i.i.d. $\mathcal{N}(0, 1)$ convoluées par un noyau gaussien de $\sigma = 1.5$,
- → décalage et mise à l'échelle du champ gaussien afin que $\mu_0 = 0.2$ et $\sigma_0 = 1.2$,
- → ajout d'un signal constant $\beta = 3$ dans un sous cube de taille $T \times T \times T$ $(\mathcal{H}_1)$
- → $\pi_0 = 1 - \frac{T^3}{64^3}$

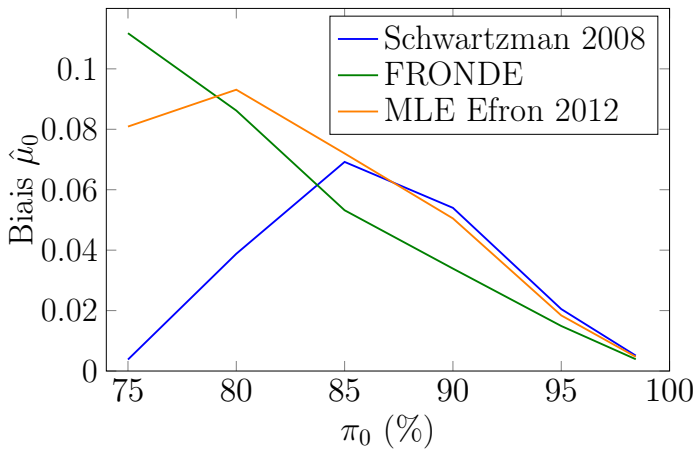
Comparaison des performances :

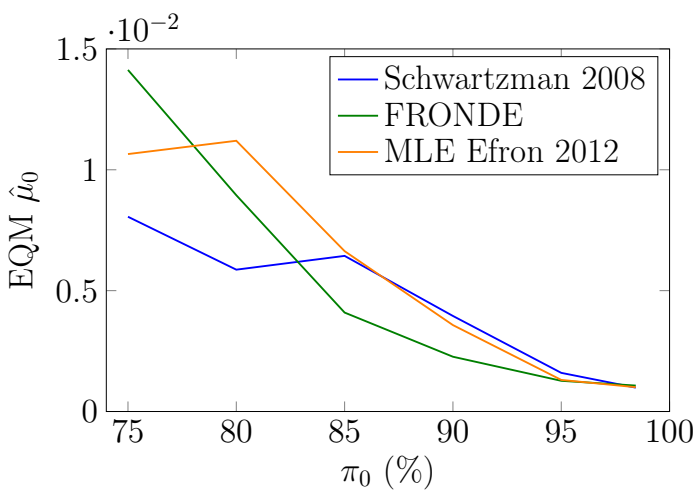
- → σ-clipping par point fixe (FRONDE),
- → ajustement du mode central avec une régression de Poisson ([Schwartzman 2008]),
- → ajustement au sens du maximum de vraisemblance à la distribution des données tronquées (MLE, [Efron 2012]).

Influence de la proportion π_0

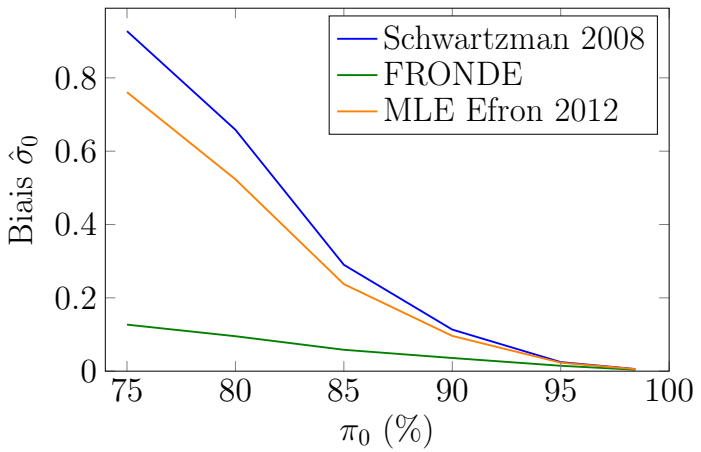
Proportion de données conservées dans la troncature : 80%

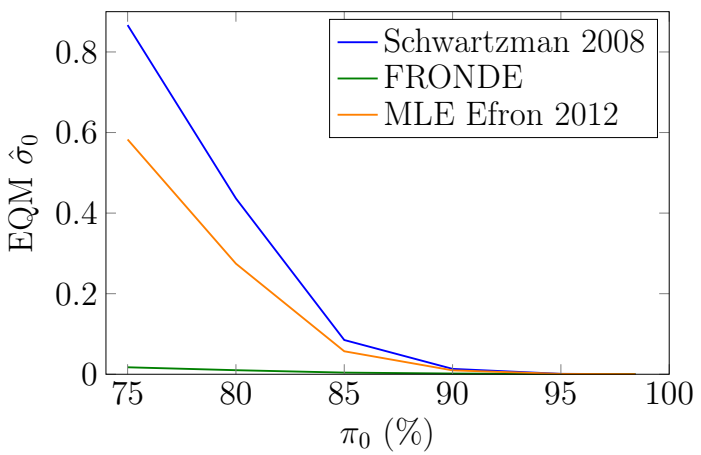
Estimateurs de la moyenne μ_0





Estimateurs de l'écart-type σ_0

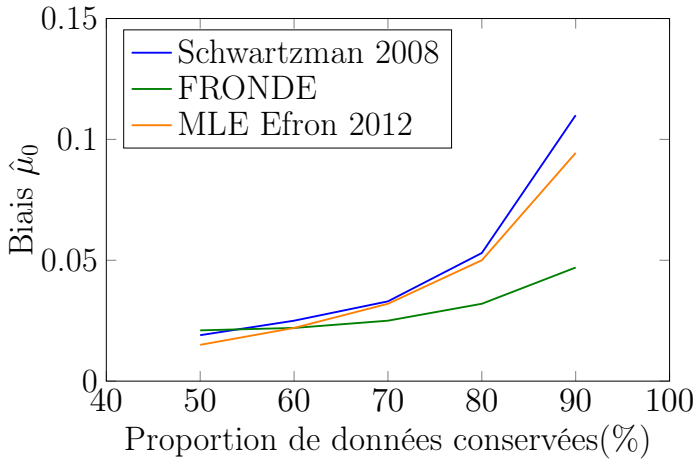


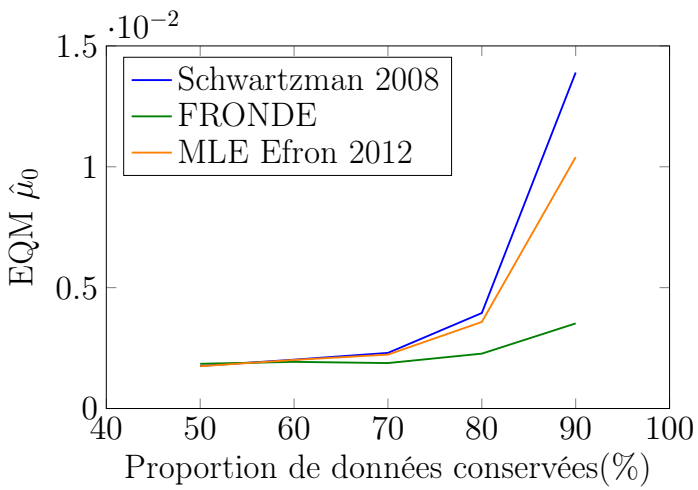


Influence de la proportion de données conservées dans la troncature

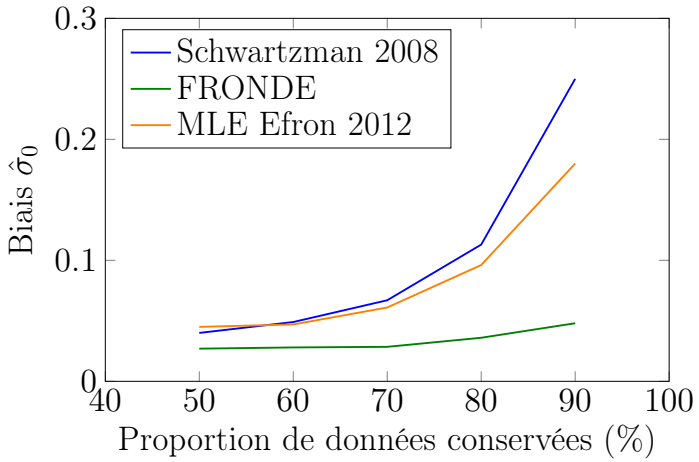
Proportion de données sous $\mathcal{H}_0$: $\pi_0 = 90\%$

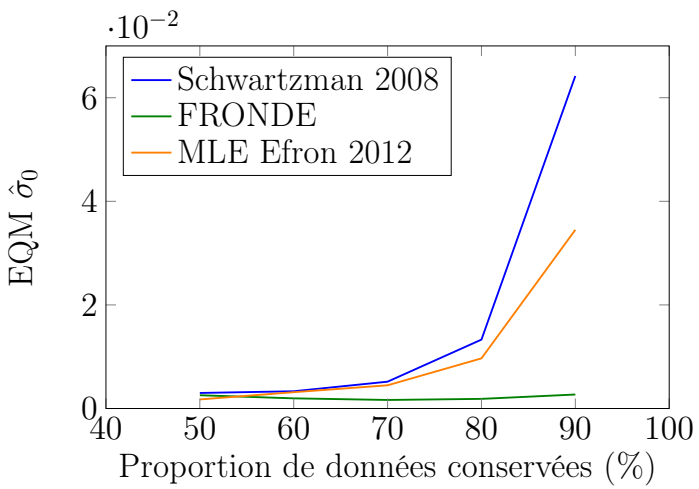
Estimateurs de la moyenne μ_0





Estimateurs de l'écart-type σ_0





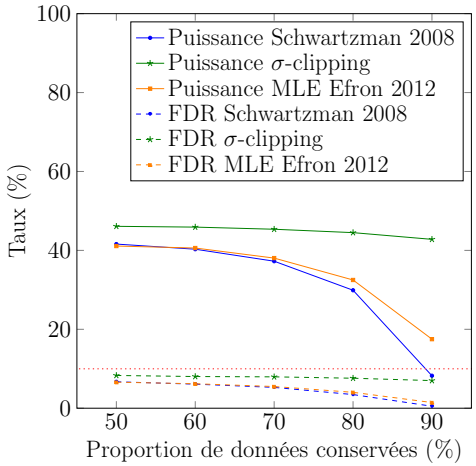
Plan de la présentation

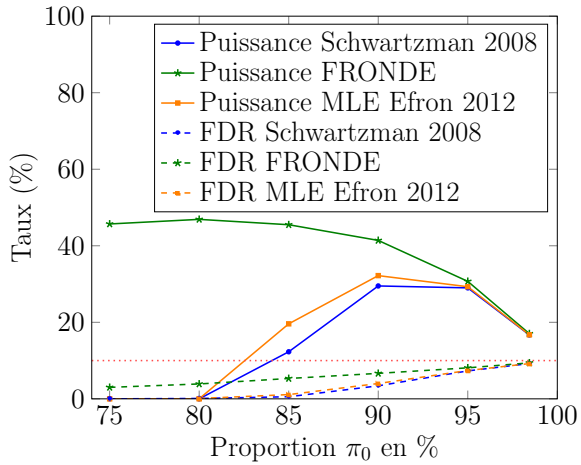
Quelques mots pour la fin...

Récapitulatif : pourquoi cette méthode ?

→ Calcul des p-valeurs : nécessité d'avoir une estimation précise et robuste de $\mathcal{H}_0$

→ Exemple du contrôle du FDR :



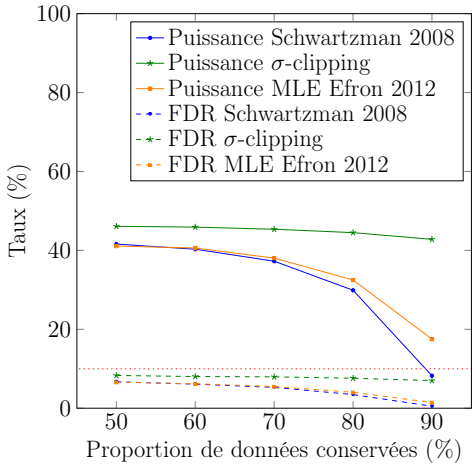


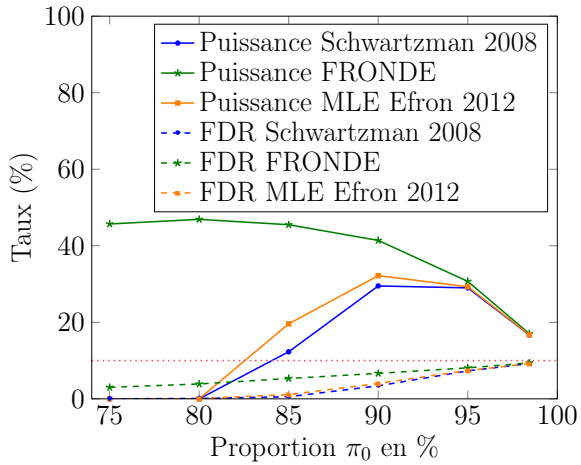
Quelques mots pour la fin...

Récapitulatif : pourquoi cette méthode ?

→ Calcul des p-valeurs : nécessité d'avoir une estimation précise et robuste de $\mathcal{H}_0$

→ Exemple du contrôle du FDR :





Code disponible :

→ implémentation python : https ://github.com/raphbacher/fronde-py

→ implémentation matlab : https ://github.com/raphbacher/fronde-matlab