Compromis précision-temps de calcul et détection de ruptures

Maxime Brunin, Christophe Biernacki, Alain Celisse

Compromis précision-temps de calcul et détection de ruptures

Maxime BRUNIN, Christophe BIERNACKI & Alain CELISSE

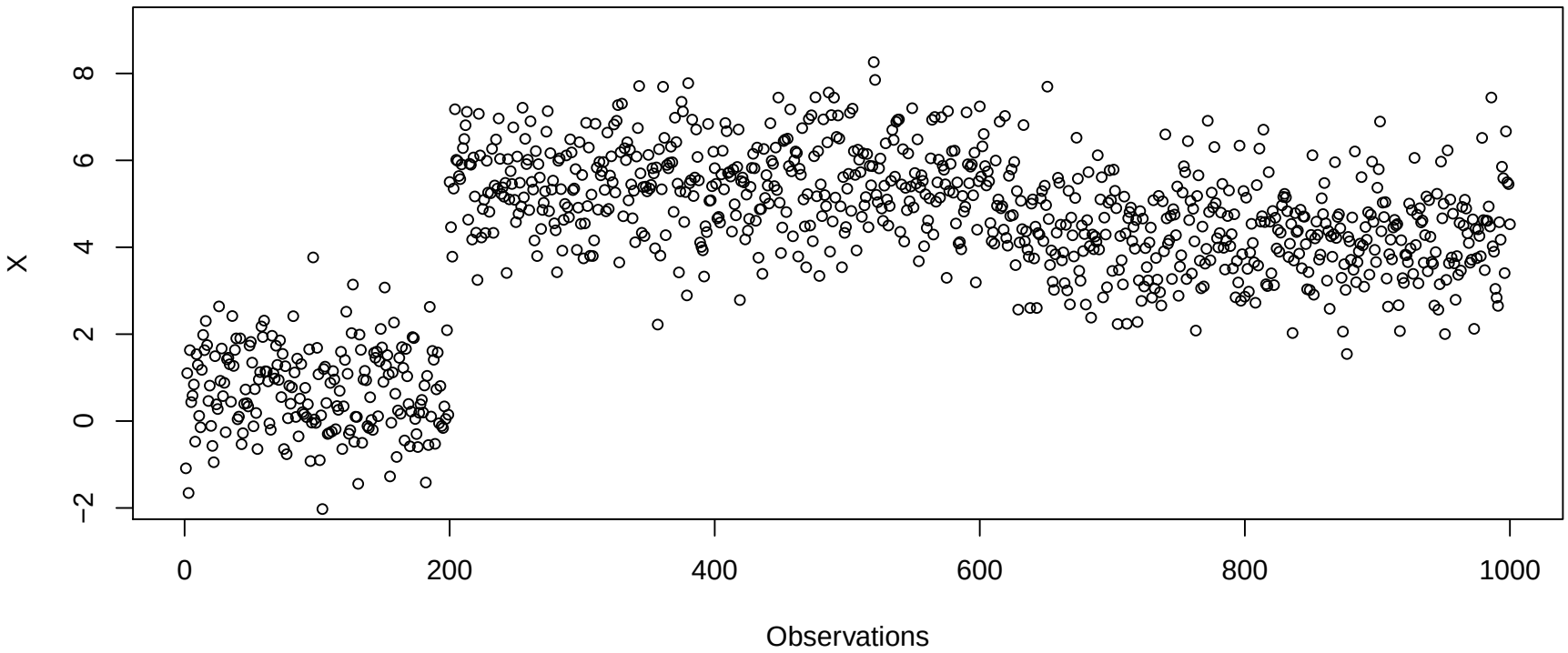
Laboratoire Paul Painlevé, Université de Lille 1
INRIA Lille-Nord Europe, équipe MODAL

29 août 2015

Plan de la présentation

Sommaire

Exemple de détection de ruptures



Détection de ruptures

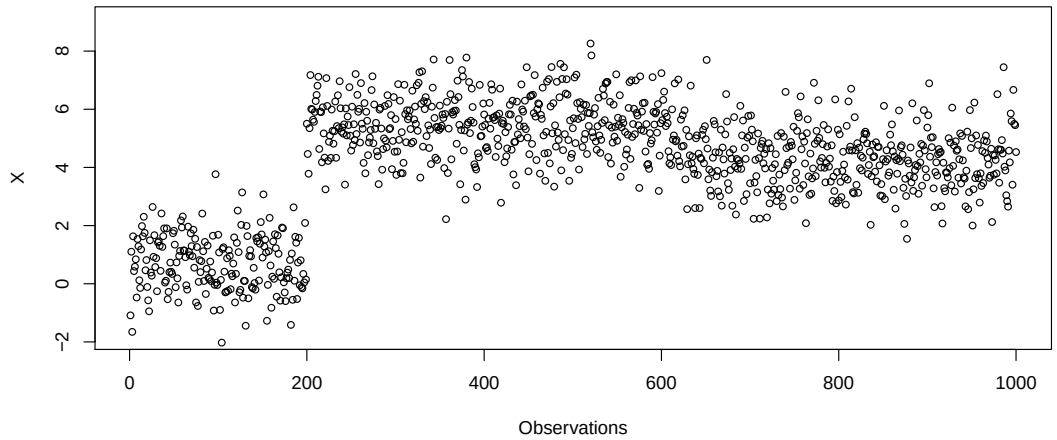
Soit $\{X_i\}_{i \in [\![1,n]\!]}$ un signal présentant des changements dans la distribution à différents instants de ruptures $\{\tau_1^*, \ldots, \tau_{D-1}^*\}$.

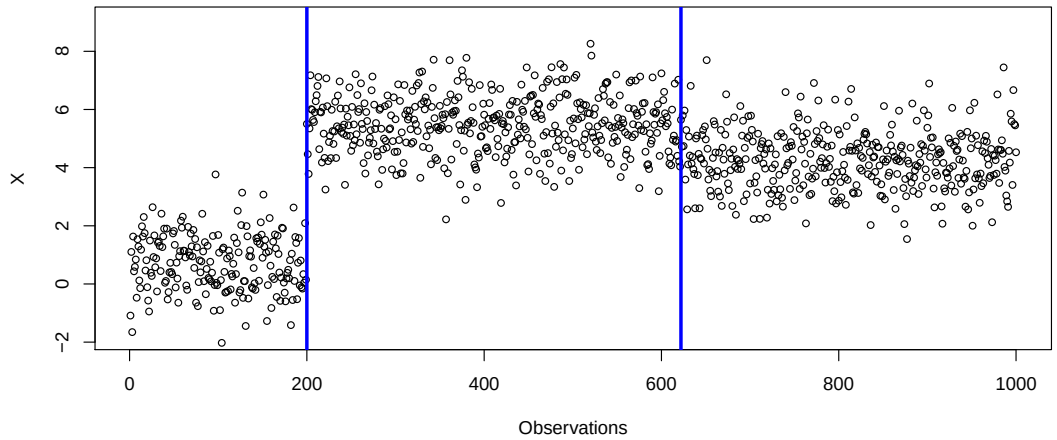
Objectif

Trouver un estimateur de $\{\tau_1^*, \ldots, \tau_{D-1}^*\}$ à l'aide d'un algorithme.

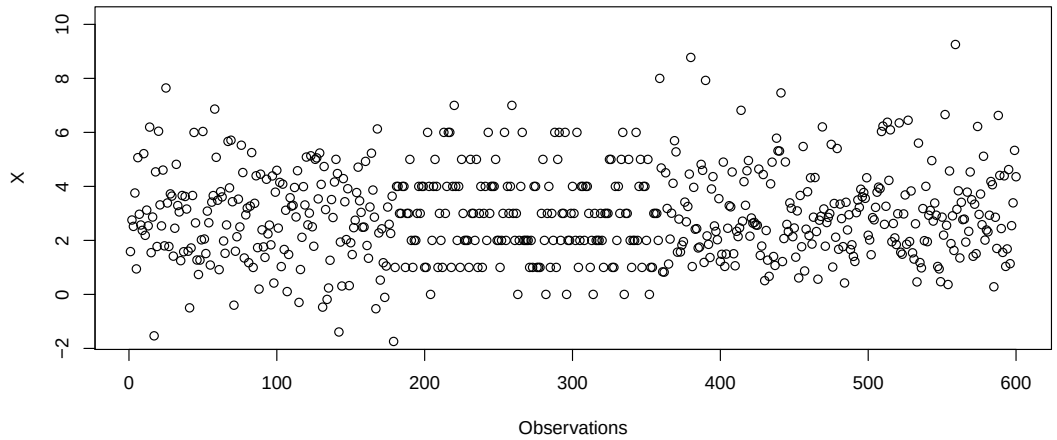
Classiquement, on réalise de la détection de ruptures dans la moyenne.
Je m'intéresse à la détection de ruptures dans la distribution.

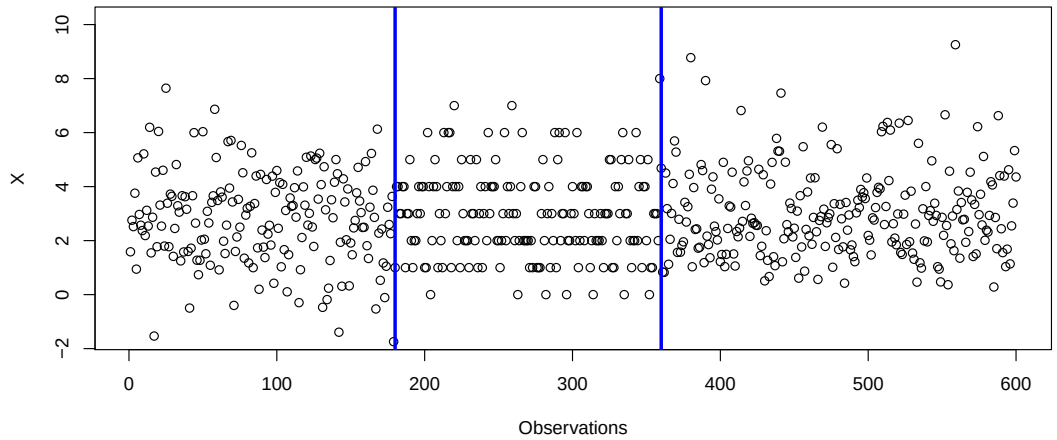
Détection de ruptures dans la moyenne





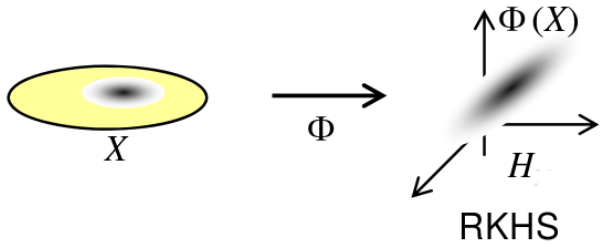
Détection de ruptures dans la distribution





Lien entre détection de ruptures dans la moyenne et détection de ruptures dans la distibution

$\mathcal{H}$ RKHS de noyau $k : \mathcal{X}^2 \mapsto \mathbb{R}$ $(\Phi(x) = k(x, .))$:



On se place dans un RKHS $\mathcal{H}$ de noyau k. On recode le signal initial $\{X_i\}_{i \in [\![1,n]\!]}$ par le signal $\{Y_i\}_{i [\![1,n]\!]} \in \mathcal{H}^n$ défini par $Y_i = k(X_i, .) = k_{X_i}$. Le modèle s'écrit :

$$\forall i \in [\![1, n]\!], \qquad Y_i = \mu_i^* + \epsilon_i.$$

La détection de ruptures dans la distribution est de la détection de ruptures dans la moyenne dans le RKHS $\mathcal{H}$ grâce à la propriété :

$$\forall i \neq j \in [\![1, n]\!]^2, \qquad \mu_i^* \neq \mu_j^* \Rightarrow P_{X_i} \neq P_{X_j}.$$

Sommaire

Notations

- $Y = (Y_1, \ldots, Y_n) \in \mathcal{H}^n$ et $\mu^* = (\mu_1^*, \ldots, \mu_n^*) \in \mathcal{H}^n$.
- $\mathcal{M}$ est l'ensemble des segmentations $m = \{\tau_i\}_{i \in [\![1, D_m - 1]\!]}$.
- Soit $\{\mathcal{S}_m, m \in \mathcal{M}\}$ la famille de modèles avec $\mathcal{S}_m$ l'ensemble des $u = (u_1, \ldots, u_n) \in \mathcal{H}^n$ vérifiant :

$$u_1 = \cdots = u_{\tau_1}, u_{\tau_1+1} = \cdots = u_{\tau_2}, \ldots, u_{\tau_{D_m-1}+1} = \cdots = u_n.$$

- Estimateur du risque empirique : $\hat{\mu}_m = \underset{u \in \mathcal{S}_m}{\arg\min} \|Y - u\|_{\mathcal{H}^n}^2$.

Objectif

Trouver la meilleure segmentation $\hat{m}$ parmi l'ensemble des segmentations candidates $m \in \mathcal{M}$.

Critère pénalisé

$\hat{m}$ est défini par :

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \text{crit}(m).$$

L'objectif est de trouver un critère crit(m) permettant de quantifier la distance entre μ^* et $\hat{\mu}_m$.

Certains critères pénalisés fournissent une inégalité oracle :

$$\text{crit}(m) = \|Y - \hat{\mu}_m\|^2_{\mathcal{H}^n} + \text{pen}(m) \approx \|\hat{\mu}_m - \mu^*\|^2_{\mathcal{H}^n}.$$

$$\|\hat{\mu}_{\hat{m}} - \mu^*\|^2_{\mathcal{H}^n} \leq \inf_{m \in \mathcal{M}} \{\|\hat{\mu}_m - \mu^*\|^2_{\mathcal{H}^n} + pen(m)\} + R_n,$$

R_n est un terme de reste.

Sommaire

Sommaire

Détection de ruptures par programmation dynamique

Cette méthode, basée sur la programmation dynamique, se décompose en deux étapes :

1. Programmation dynamique :

$$\hat{m}_D = \underset{m \in \mathcal{M}(D)}{\arg\min} \ \|Y - \hat{\mu}_m\|^2_{\mathcal{H}^n} .$$

2. Sélection de modèle :

$$\hat{D} = \underset{D \in [\![1, D_{max}]\!]}{\arg\min} \ \{\|Y - \hat{\mu}_{\hat{m}_D}\|^2_{\mathcal{H}^n} + pen(\hat{m}_D)\}.$$

avec :

- $\mathcal{M}(D)$ l'ensemble des segmentations de $\{1, \dots, n\}$ en D segments.
- $pen(m) = \frac{CD_m}{n}\left(\log(\frac{n}{D_m}) + 1\right)$ avec C une constante à calibrer par heuristique de pente.

Compromis précision-temps de calcul

Précision (Arlot, Celisse et Harchaoui 2012)

Sous les hypothèses suivantes,

- $\exists M > 0,\ \sup_{i \in [\![1,n]\!]} \|Y_i\|_{\mathcal{H}}^2 = \sup_{i \in [\![1,n]\!]} k(X_i, X_i) \leq M^2.$
- $\max_{i \in [\![1,n]\!]} v_i \leq v_{max} \quad \left(v_i = E\left[\|\epsilon_i\|_{\mathcal{H}}^2\right]\right).$
- $\exists 0 < c_{min} < +\infty,\ \min_{i \in [\![1,n]\!]} v_i \geq \frac{M^2}{c_{min}} =: v_{min} > 0.$

avec une probabilité supérieure à $1 - \mathrm{e}^{-x}$, si $C_1 \geq L_1 c_{min}^2$, on a le résultat :

$$\frac{1}{n}\|\hat{\mu}_{\hat{m}} - \mu^*\|_{\mathcal{H}^n}^2 \leq 2 \inf_{m \in \mathcal{M}} \{\frac{1}{n}\|\hat{\mu}_m - \mu^*\|_{\mathcal{H}^n}^2 + 2pen(m)\} + \frac{C_1 (\log(4) + x) v_{max}}{n}.$$

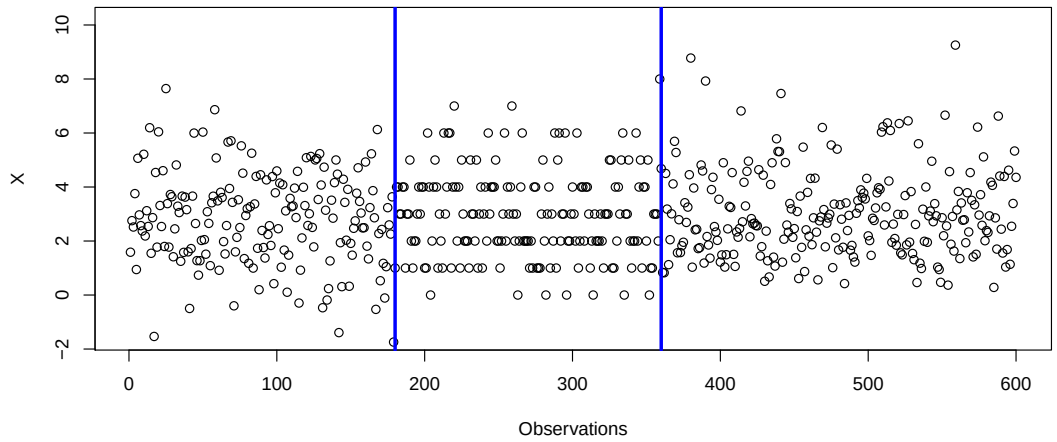
Temps de calcul

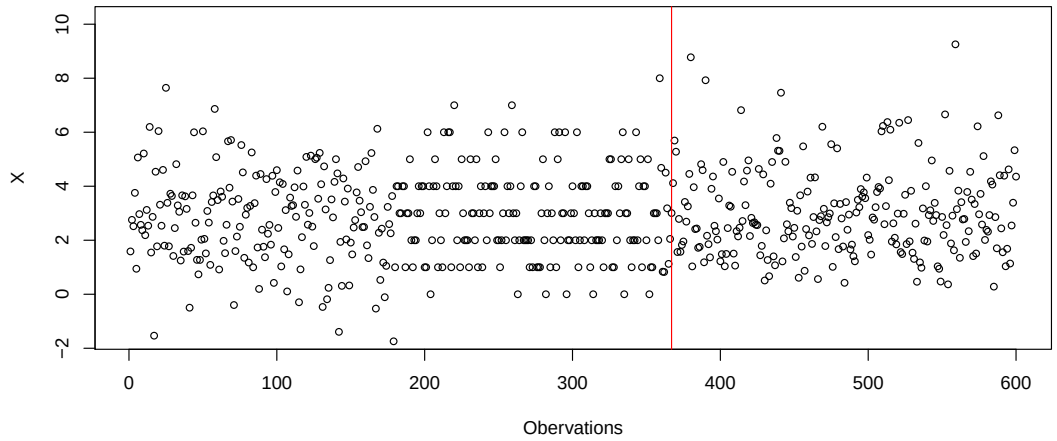
C'est une méthode exacte dont la complexité est de $O\left(D_{max} n^4\right)$ en temps et $O\left(D_{max} n^2\right)$ en espace.

Sommaire

Exemple itération 1

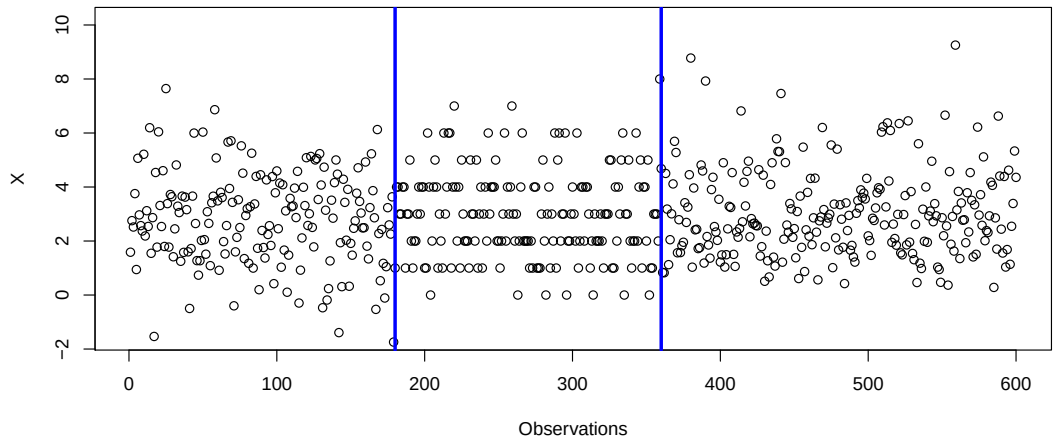
Itération 1

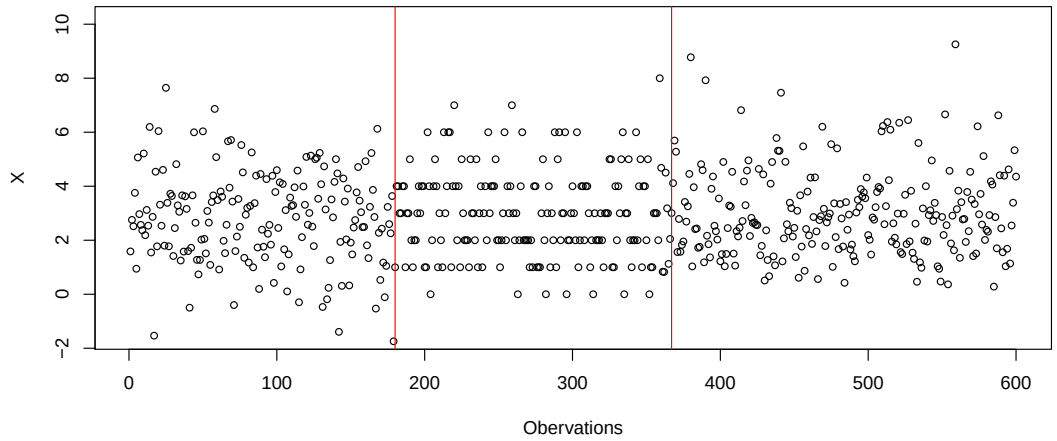




Exemple itération 2

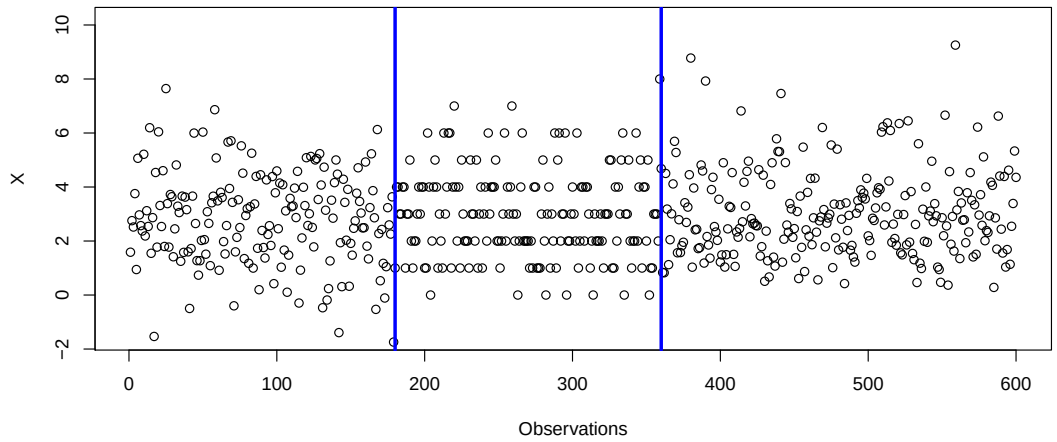
Itération **2**

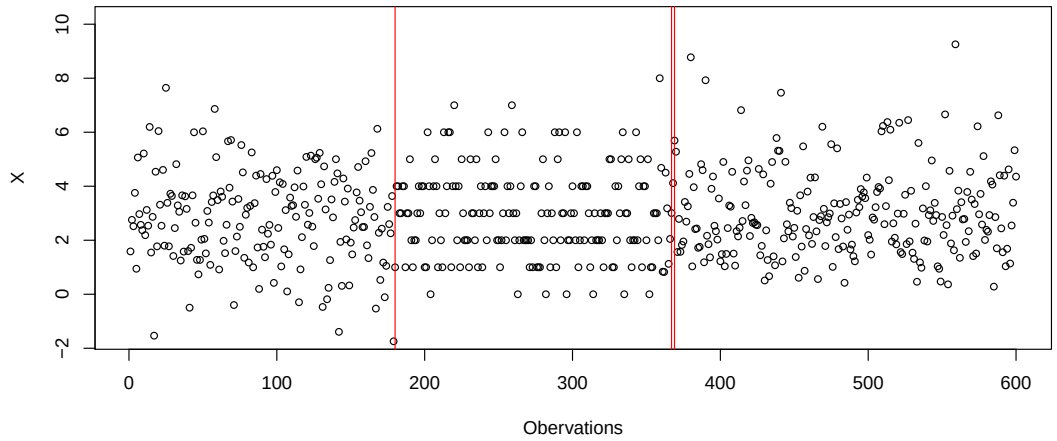




Exemple itération 3

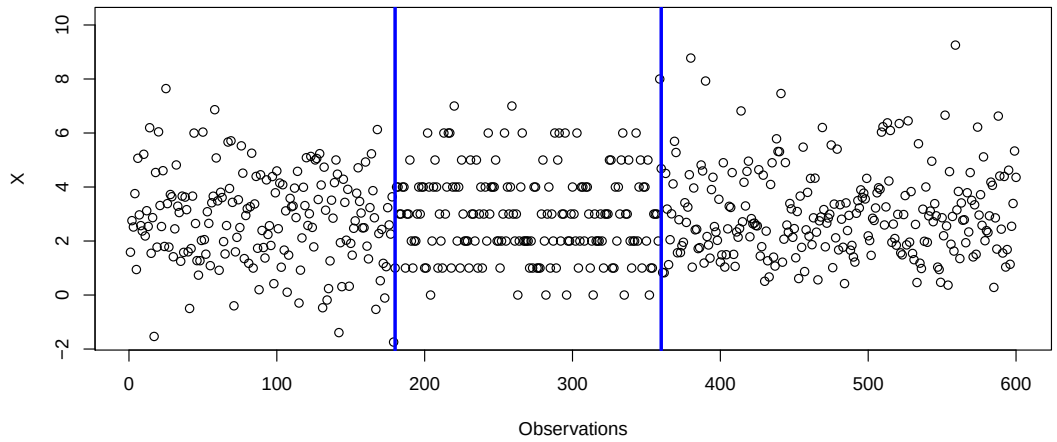
Itération **3**

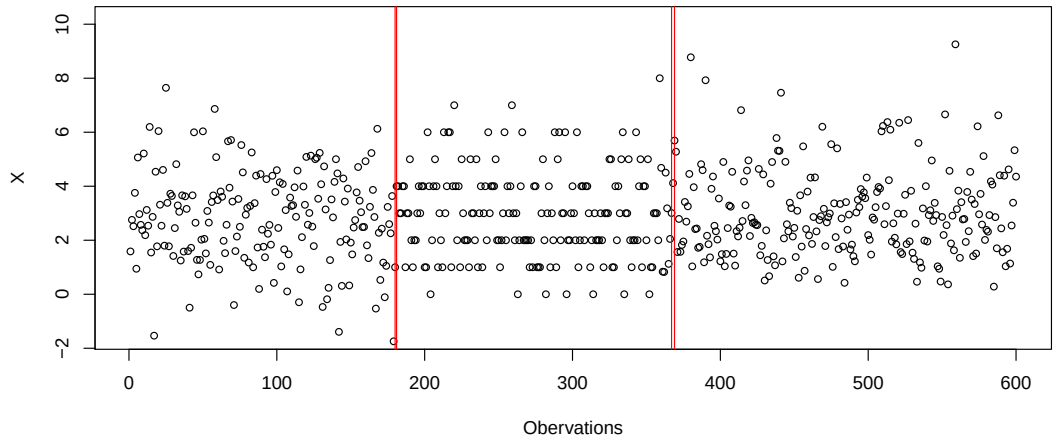




Exemple itération 4

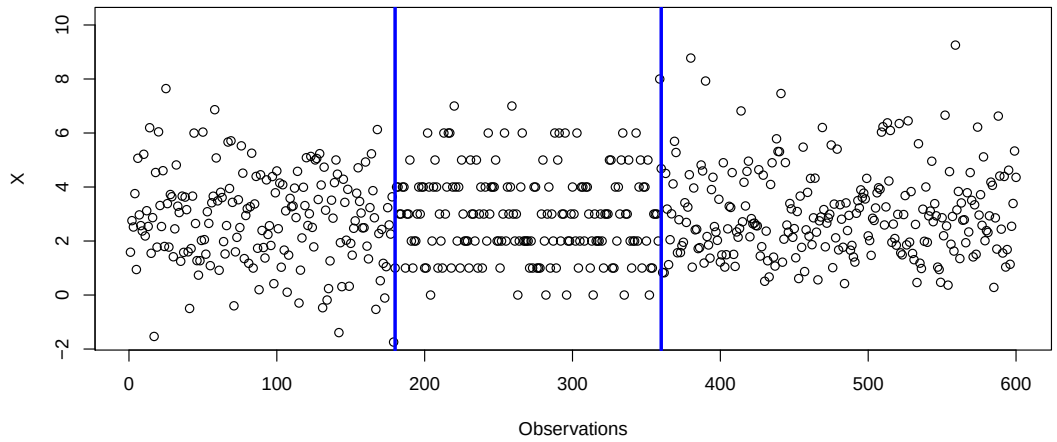
Itération **4**

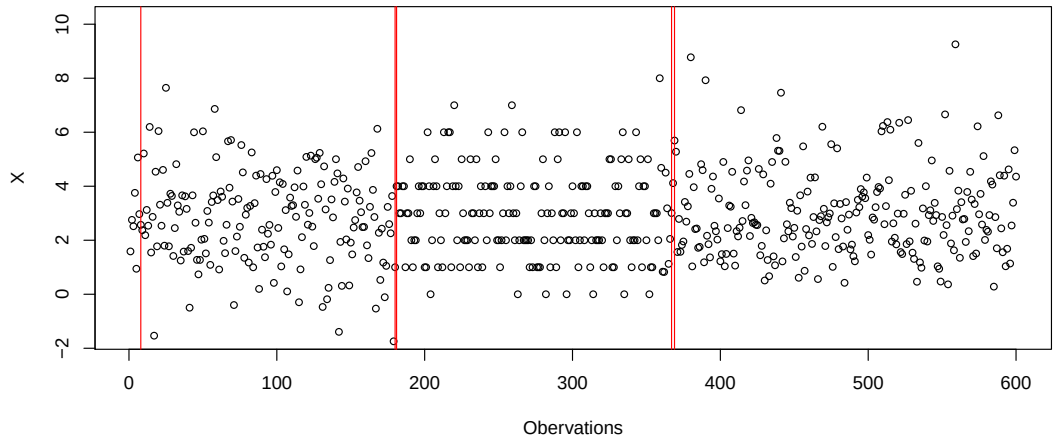




Exemple itération 5

Itération **5**





Principe

A l'itération t, on cherche à minimiser le risque en β pour trouver les instants de ruptures :

$$R(\beta_1^{(t)}, \dots, \beta_n^{(t)}) = R(\beta^{(t)}) = \left\| Y - X^{(t)}\beta^{(t)} \right\|_{\mathcal{H}^n}^2$$
$$= \sum_{i=1}^{n} \left\| Y_i - \sum_{j=1}^{n} X_{i,j}^{(t)} \beta_j^{(t)} \right\|_{\mathcal{H}}^2 ,$$

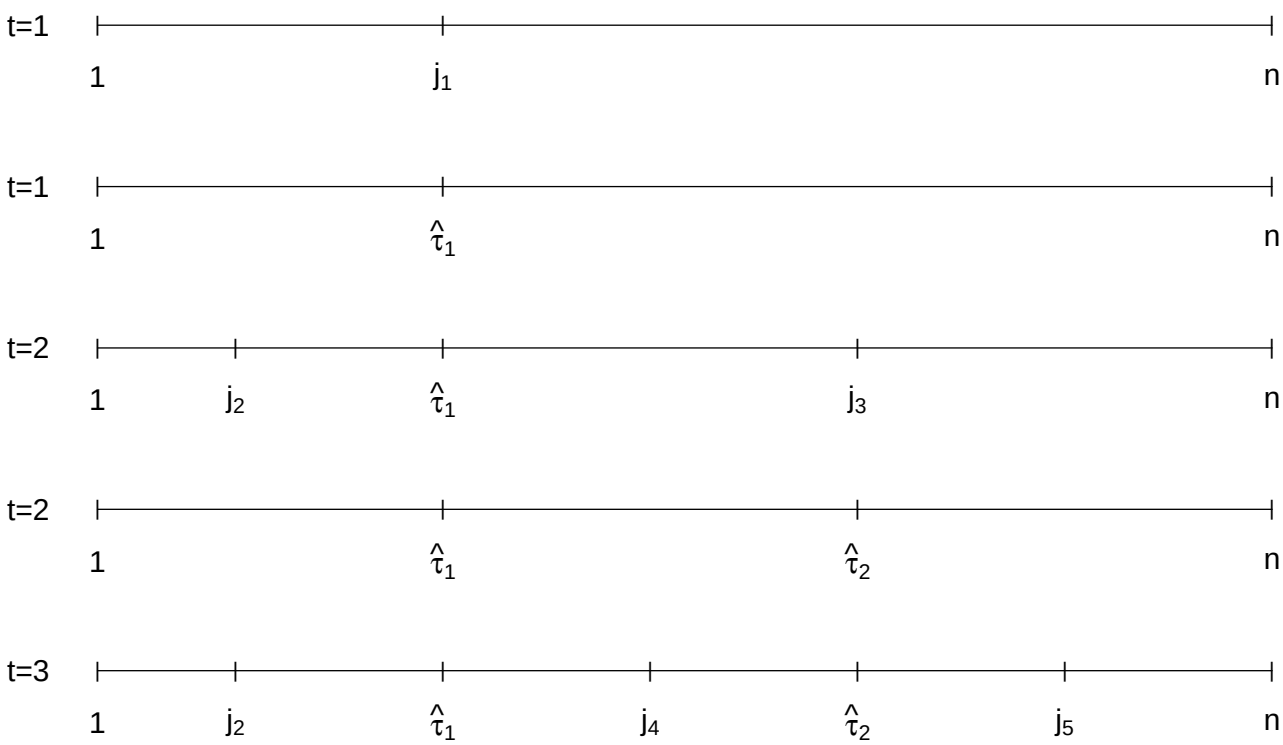
où $X^{(t)}$ est réactualisée à chaque itération.

On utilise la méthode de descente de coordonnées et la segmentation binaire.

La descente de coordonnées fournit la direction j minimisant $R(\beta^{(t)})$ et son minimiseur $\beta_j \in \mathcal{H}$.

Segmentation binaire

Critère : $R(\beta_1^{(t)}, \ldots, \beta_n^{(t)}) = \left\| Y - X^{(t)}\beta^{(t)} \right\|_{\mathcal{H}^n}^2 = \sum_{i=1}^{n} \left\| Y_i - \sum_{j=1}^{n} X_{i,j}^{(t)} \beta_j^{(t)} \right\|_{\mathcal{H}}^2 .$



Estimateur

A l'itération t, on récupère une segmentation $\hat{m}_D$ en $D = t + 1$ segments :

$$\hat{m}_D = \underset{m \in \mathcal{M}_D(\hat{\tau}_1, \ldots, \hat{\tau}_{D-2})}{\arg\min} \|Y - \hat{\mu}_m\|^2_{\mathcal{H}^n},$$

avec $\mathcal{M}_D(\hat{\tau}_1, \ldots, \hat{\tau}_{D-2})$ est l'espace des segmentations en D segments avec $D - 2$ instants de ruptures calculés à l'itération $t - 1$.

Compromis précision-temps de calcul

Temps de calcul

La segmentation binaire à noyau a une complexité de $O(n^2)$ en temps et $O(n)$ en espace.

Précision

La recherche d'un temps d'arrêt $\hat{D}$ demeure une question.

Algorithme

La complexité de l'algorithme est $O\left(\hat{D}n^2\right)$ en temps et $O\left(\hat{D}n\right)$ en espace.

Une piste

Dans le cas, $\mathcal{H} = \mathbb{R}$, on a le résultat suivant (Fryzlewicz 2014) :

Hypothèses

- $\min_{i \in [\![1, D^*]\!]} |\tau_i^* - \tau_{i-1}^*| \geq \delta_n \geq C_1 n^{\alpha}$ avec $C_1 > 0$ et $\alpha \leq 1$.
- $\min_{i \in [\![1, D^*]\!]} |\mu_{\tau_i^*}^* - \mu_{\tau_{i-1}^*}^*| \geq f_n \geq C_2 n^{-\beta}$ avec $C_2 > 0$ et $\beta \geq 0$.
- $\alpha - \frac{\beta}{2} > \frac{3}{4}$.

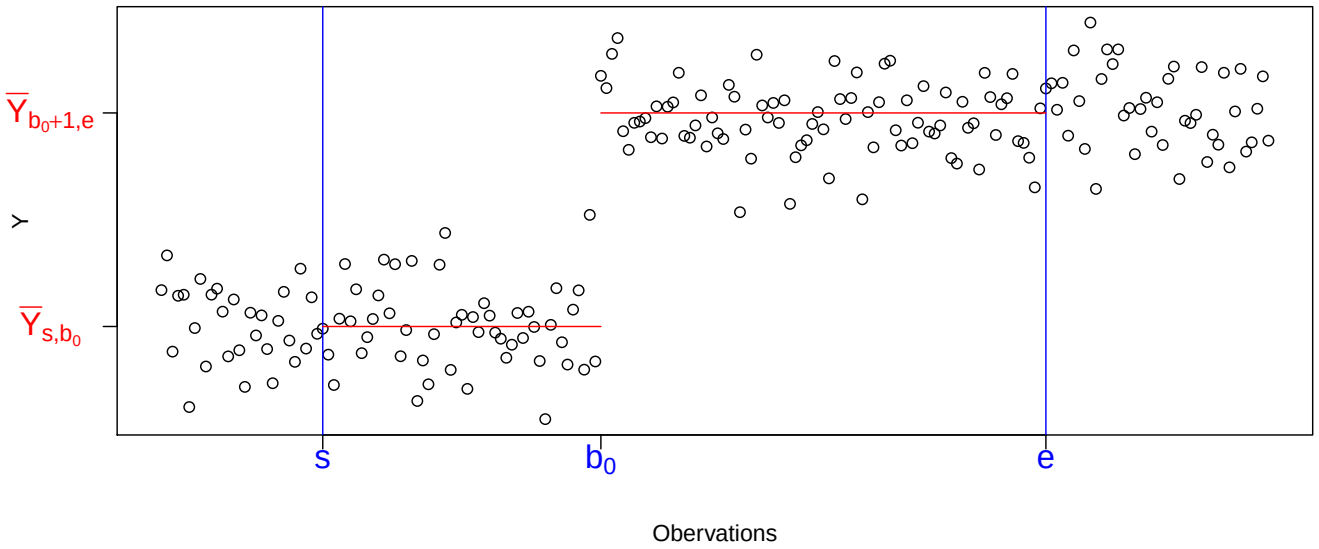
Théorème

Sous des contraintes sur un paramètre de seuil ζ_n, $P(\mathcal{A}_n) \geq 1 - C_3 n^{-1}$,

$$\mathcal{A}_n = \{\hat{D} = D^*, \max_{i \in [\![1, D^* - 1]\!]} |\hat{\tau}_i - \tau_i^*| \leq C \xi_n\},$$

avec $\xi_n = n^2 \delta_n^{-2} f_n^{-2} \log(n)$.

Définition de $\tilde{Y}^b_{s,e}$



Pour $b \in [\![s, e-1]\!]$,

$$\tilde{Y}^b_{s,e} = \sqrt{\frac{(e-b)(b-s+1)}{e-s+1}} \left(\overline{Y}_{s,b} - \overline{Y}_{b+1,e}\right)$$

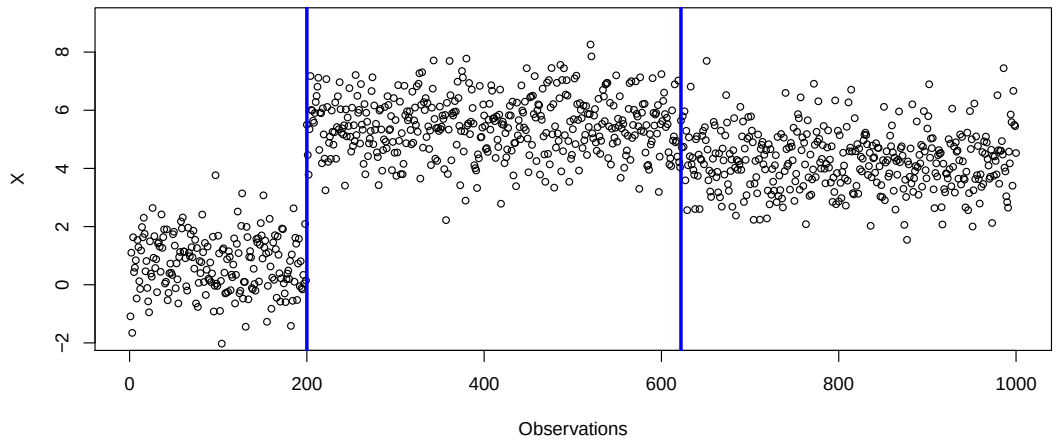
Algorithme (Fryzlewicz 2014)

fonction BinSeg(s,e,ζ_n)
  si e-s<1 alors
    STOP
  sinon
    $b_0 = \underset{b \in [\![s, e-1]\!]}{\arg\max} |\tilde{Y}^b_{s,e}|$
    si $|\tilde{Y}^{b_0}_{s,e}| > \zeta_n$
      ajouter b_0 à l'ensemble des instants de ruptures estimés
      BinSeg(s,b_0,ζ_n)
      BinSeg($b_0 + 1$,e,ζ_n)
    sinon
      STOP
    fin si
  fin si
fin fonction

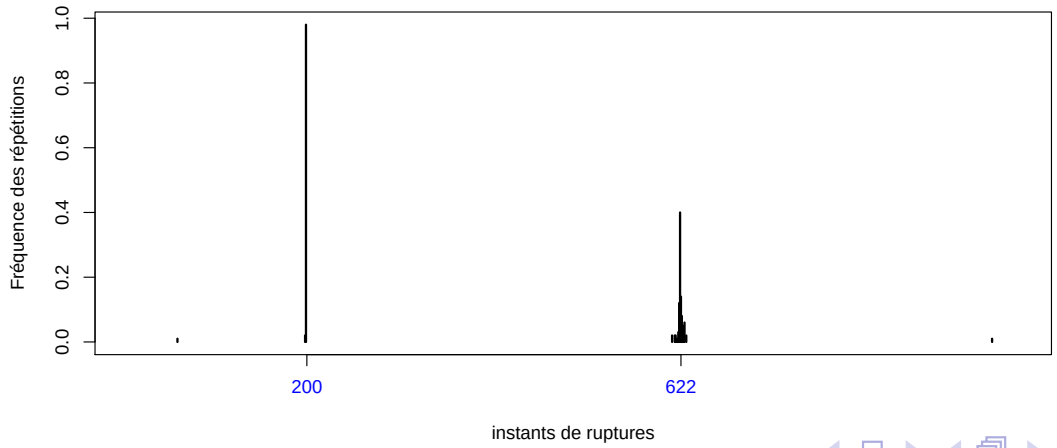
On exécute BinSeg(1,n,ζ_n).

Résultats

Pour 100 répétitions d'un bruit $\forall i \in [\![1, n]\!]\ \epsilon_i \sim \mathcal{N}(0, 1)$ $(n = 1000)$:

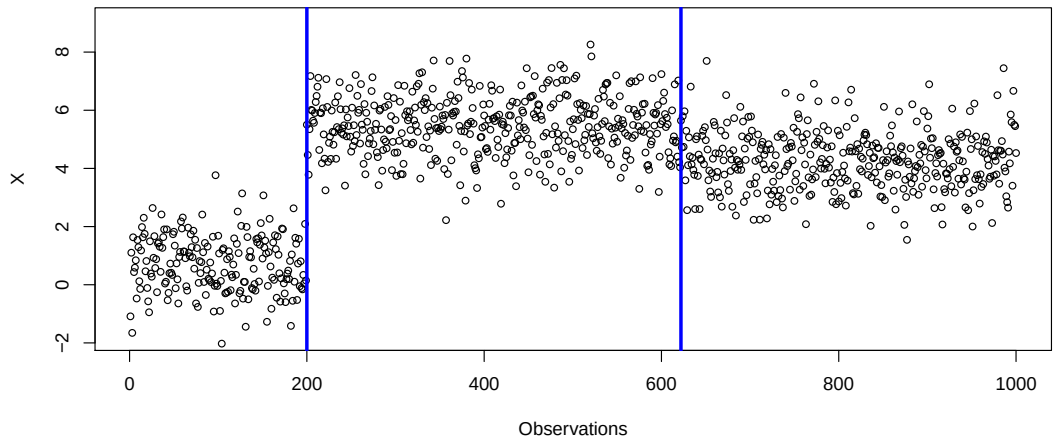


Graphique représentant la fréquence des répétitions en fonction des instants de ruptures



Résultats

Pour 100 répétitions d'un bruit $\forall i \in [\![1, n]\!]\ \epsilon_i \sim \mathcal{N}(0, 1)$ $(n = 1000)$:



Sommaire

Conclusion

Résultats

Algorithme permettant de récupérer des estimateurs des instants de ruptures.

Perspectives

Trouver le temps d'arrêt de l'algorithme.

Bibliographie

Arlot, Sylvain, Alain Celisse et Zaid Harchaoui (2012). *Kernel change-point detection*.

Fryzlewicz, Piotr (2014). « Wild binary segmentation for multiple change-point detection ». In : *The Annals of Statistics* 42.6, p. 2243–2281. issn : 0090-5364, 2168-8966. doi : 10.1214/14-AOS1245.

Questions ?