Compromis précision-temps de calcul appliqué au problème de détection de ruptures

Maxime Brunin [1] & Christophe Biernacki [2] & Alain Celisse [3]

Inria Lille-Nord Europe & Laboratoire Paul Painlevé, Université Lille 1
[1] *maxime.brunin@inria.fr*
[2] *christophe.biernacki@math.univ-lille1.fr*
[3] *alain.celisse@math.univ-lille1.fr*

Résumé. Le problème de détection de ruptures a pour but de détecter des changements dans la distribution d'observations recueillies au cours du temps entre les instants $1, \dots, T$ dans un contexte *offline*. Ces changements se produisent à certains instants appelés *instants de ruptures*. Notre méthode fournit des estimateurs consistants de ces instants de ruptures obtenus par l'algorithme de segmentation binaire à noyau avec temps d'arrêt (KBS). De plus, notre méthode a une plus faible complexité en temps et en espace que la programmation dynamique à noyau (KDP).

Mots-clés. problème de détection de ruptures, sélection de modèle, temps d'arrêt.

Abstract. The change-point detection problem aims to detect changes in the distribution of observations collected over the time between the instants $1, \dots, T$ in the *offline* context. These changes occur at some instants called *change-points*. Our method provides consistent estimates of the change-points obtained by the Kernel Binary Segmentation algorithm with stopping rule (KBS). Moreover, our method has a lower complexity in time and in space than the Kernel Dynamic Programming (KDP).

Keywords. change-point detection problem, model selection, stopping rule.

1 Introduction

Nous abordons le problème de détection de ruptures, d'observations recueillies au cours du temps pendant les instants $1, \dots, T$, avec l'utilisation des méthodes à noyau dans le cadre du compromis précision-temps de calcul. Ce problème a des applications en génomique dans l'étude des variations du nombre de copies d'ADN dans différents types de cancers. Il est utile aussi pour segmenter le signal sonore d'une émission entre les interviews, bandes d'annonces, musique. Les contributions existantes traitent du cadre de la détection de ruptures dans la moyenne d'un signal réel, présentée par exemple par Lebarbier (2005) et de celui dans la distribution, présentée par exemple par Arlot (2015).

Les méthodes présentées par Lebarbier (2005), Arlot (2015) s'appuient sur des critères pénalisés qui dépendent de constantes optimisées par heuristique de pente, méthode présentée par Massart (2007). Le défaut de l'heuristique de pente est qu'elle nécessite le calcul des meilleurs segmentations en D segments pour $1 \leq D \leq D_{\max}$ (par exemple, par programmation dynamique) qui a une complexité en temps de $O(D_{\max}T^4)$: l'heuristique de pente est donc prohibitive pour T élevé.

Une méthode alternative est la segmentation binaire avec l'utilisation de temps d'arrêt, proposée par Fryzlewicz (2014). Elle permet grâce à son temps d'arrêt $\hat{D}$ d'arrêter d'itérer à l'itération $\hat{D}-1$ (avec $\hat{D} < D_{\max}$) et donc de réduire la complexité en temps : celle-ci est de $O\Big(\hat{D}T^2\Big)$.

Dans ce document, nous présentons : une version noyau de l'algorithme de segmentation binaire avec temps d'arrêt, *Kernel Binary Segmentation* (KBS), permettant de récupérer des estimateurs des instants de ruptures. Puis, nous présentons notre théorème assurant la consistance de ces estimateurs. Enfin, des simulations illustrent la bonne performance de notre méthode.

2 Méthode proposée

2.1 Problème de détection de ruptures dans la distribution

Notations

Le problème de détection de ruptures dans la distribution a pour but de détecter des changements dans la distribution d'un signal $\{X_1, \dots, X_T\}$ à des instants de ruptures inconnus $\{\tau_1^*, \dots, \tau_{D^*-1}^*\}$. La distribution de $\{X_t\}_{t\in[\![1,T]\!]}$ vérifie :

$$P_{X_{\tau_0^*}} = \cdots = P_{X_{\tau_1^*-1}} \neq P_{X_{\tau_1^*}} = \cdots = P_{X_{\tau_2^*-1}} \neq \cdots \neq P_{X_{\tau_{D^*-1}^*}} = \cdots = P_{X_{\tau_{D^*}^*}},$$

avec la convention $\tau_0^* = 1$ et $\tau_{D^*}^* = T$.

Noyau symétrique défini positif

On recode les observations initiales $\{X_1, \dots, X_T\}$ par de nouvelles "observations" $\{Y_1, \dots, Y_T\}$. Elles sont définies par $\forall t \in [\![1, T]\!],\ Y_t = k(X_t, .)$, avec k un noyau symétrique défini positif. Par exemple, le noyau Gaussien k_δ est défini par $\forall (x, y) \in \mathbb{R},\ k_\delta(x, y) = \exp\left(-\frac{(x-y)^2}{2\delta^2}\right)$. Le théorème de Moore-Aronszajn assure l'existence d'un RKHS $\mathcal{H}$ (Espace de Hilbert à noyau reproduisant) de noyau k si k est un noyau symétrique défini positif.

Détection de ruptures dans la moyenne dans le RKHS $\mathcal{H}$

On a le modèle de régression :

$$\forall t \in [\![1, T]\!],\ Y_t = f_t + \epsilon_t. \tag{1}$$

Pour des noyaux caractéristiques (par exemple le noyau Gaussien), on a la propriété suivante :

$$P_{X_i} \neq P_{X_j} \Rightarrow f_i \neq f_j.$$

Ainsi, faire de la détection de ruptures dans la distribution est équivalente à faire de la détection de ruptures dans la moyenne dans le RKHS $\mathcal{H}$.

2.2 Kernel Binary Segmentation (KBS)

La segmentation binaire à noyau avec temps d'arrêt est une méthode récursive permettant de récupérer des estimateurs minimisant le risque empirique. On définit tout d'abord quelques notations.

Notations

- $Y = (Y_1, \ldots, Y_T)$.
- Pour $1 \leq s < e \leq T$, $\forall b \in [\![s, e-1]\!]$,

$$\tilde{Y}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} Y_t - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} Y_t,$$

  avec $n = e - s + 1$.
- $m = \{\tau_1, \ldots, \tau_{D_m - 1}\}$ est une segmentation en D_m segments.
- $\hat{\mu}_m$ est la projection orthogonale de Y sur l'espace des fonctions constantes par morceaux avec des discontinuités en chaque τ_i pour $m = \{\tau_i\}_{i \in [\![1, D_m - 1]\!]}$.

On utilisera le lemme suivant, résultat d'un calcul permettant de relier la maximisation de $\left\|\tilde{Y}_{s,e}^b\right\|_{\mathcal{H}}$ à la minimisation du risque empirique.

Lemme 2.1 *Pour $1 \leq s < e \leq T$, un instant de rupture candidat b_0 est défini comme le minimiseur du risque empirique :*

$$b_0 = \underset{b: s \leq b < e}{\arg\min} \ \|Y_s^e - \hat{\mu}_{m_b}\|_{\mathcal{H}^n}^2 = \underset{b: s \leq b < e}{\arg\max} \ \left\|\tilde{Y}_{s,e}^b\right\|_{\mathcal{H}}^2$$

avec $m_b = (s, b, e)$, $n = e - s + 1$ et $Y_s^e = (Y_s, \ldots, Y_e)$.

Algorithme de segmentation binaire à noyau avec temps d'arrêt
Pour plus de clarté, on illustre l'implémentation sur les 2 premières itérations de l'algorithme de segmentation binaire à noyau avec temps d'arrêt (KBS).

Première étape (t=1)

- On cherche j_1 l'instant de ruptures candidat qui minimise le risque empirique $\|Y - \hat{\mu}_{m_1}\|^2_{\mathcal{H}}$ (avec $m_1 = (1, j, T)$) entre $s = 1$ et $e = T$ pour $j \in [\![1, T-1]\!]$. D'après le lemme 2.1, on peut écrire $j_1 = \underset{j:1\leq j<T}{\arg\max} \left\|\tilde{Y}^j_{1,T}\right\|^2_{\mathcal{H}}$.

- Si $\left\|\tilde{Y}^{j_1}_{1,T}\right\|_{\mathcal{H}} > \zeta_T$ (ζ_T est un paramètre de seuil défini dans le théorème 3.1) alors j_1 est un instant de rupture estimé et on note $\hat{\tau}_1 = j_1$.

Deuxième étape (t=2)

- On cherche $j_2,\ j_3$ les instants de ruptures candidats minimisant le risque empirique sur $[\![1, \hat{\tau}_1 - 1]\!]$ et $[\![\hat{\tau}_1 + 1, T-1]\!]$ respectivement. On note $s_2 = 1,\ e_2 = \hat{\tau}_1$ et $s_3 = \hat{\tau}_1 + 1,\ e_3 = T$.
  Par exemple pour j_2, d'après le lemme 2.1, $j_2 = \underset{j:1\leq j<\hat{\tau}_1}{\arg\max} \left\|\tilde{Y}^j_{1,\hat{\tau}_1}\right\|^2_{\mathcal{H}}$.

- $\forall l \in [\![2, 3]\!]$ si $\left\|\tilde{Y}^{j_l}_{s_l,e_l}\right\|_{\mathcal{H}} > \zeta_T$, j_l est un instant de ruptures estimé.

Puis, on continue par dichotomie sur les intervalles délimités par les instants de ruptures estimés. L'algorithme s'arrête lorsque plus aucun instant de rupture candidat j vérifie la condition $\left\|\tilde{Y}^j_{s,e}\right\|_{\mathcal{H}} > \zeta_T$.

Remarque 2.1 *Si on utilise KBS avec le noyau linéaire (défini par $\forall x, y \in \mathbb{R},\ k(x, y) = xy$), on réalise de la détection de ruptures dans la moyenne pour un signal réel.*

3 Théorème

Le théorème énoncé dans cette section assure la consistance des estimateurs obtenus par KBS. Il utilise l'inégalité de concentration de Pinelis et Sakhanenko (Boucheron (2013)) pour montrer que les estimateurs des instants de ruptures sont "proches" des vrais instants de ruptures. Nous énoncons d'abord les hypothèses de ce théorème : elles imposent que la distance minimale entre deux vrais instants de ruptures consécutifs ainsi que les sauts des paliers doivent être suffisamment importants.

3.1 Hypothèses

Hypothèses 1 :

- $\forall t \in [\![1, T]\!]$, $\epsilon_t \in \mathcal{H}$ est un processus Gaussien de moyenne nulle et d'opérateur de covariance Σ avec $\mathrm{Tr}(\Sigma) = 1$.

- $\{Y_t\}_{t\in[\![1,T]\!]}$ est bornée.

Hypothèses 2 :

- $\min_{i\in[\![1,D^*]\!]} \left| \tau_i^* - \tau_{i-1}^* \right| \geq C_1 T^{\Theta}$ pour $C_1 > 0$ et $\Theta \leq 1$.
- $\min_{i\in[\![1,D^*-1]\!]} \left\| f_{\tau_i^*} - f_{\tau_i^*-1} \right\|_{\mathcal{H}} \geq C_2 T^{-w}$ pour $C_2 > 0$, $\Theta - \frac{w}{2} > \frac{3}{4}$ et $w \geq 0$.

Limites des hypothèses
Ces hypothèses impliquent notamment que $\min_{i\in[\![1,D^*]\!]} \left| \tau_i^* - \tau_{i-1}^* \right| \geq C_1 T^{3/4}$. Ainsi, l'écart entre deux vrais instants de ruptures doit être supérieur à $C_1 T^{3/4}$ donc les estimateurs des instants de ruptures ne sont pas consistants lorsque l'écart entre deux vrais instants de ruptures est trop petit.

3.2 Théorème

Théorème 3.1 *Soit $\{Y_t\}_{t\in[\![1,T]\!]}$ vérifiant le modèle 1 et supposons que les hypothèses 1 et 2 sont vérifiées. Soit le paramètre de seuil ζ_T vérifiant $\zeta_T = c_1 T^{\theta}$ pour $\theta \in (1-\Theta, \Theta - 1/2 - w)$ si $\Theta \in (3/4, 1)$, ou $\zeta_T \geq c_2(\log(T))^p$ $(p > 1/2)$ et $\zeta_T \leq c_3 T^{\theta}$ $(\theta < 1/2 - w)$ si $\Theta = 1$, quelles que soient les constantes c_1, c_2 et c_3. Alors, il existe des constantes C_3, C_4 telles que $P(\mathcal{A}_T) \geq 1 - C_3 T^{-1}$ où*

$$\mathcal{A}_T = \{\hat{D} = D^*, \max_{i\in[\![1,D^*-1]\!]} |\hat{\tau}_i - \tau_i^*| \leq C_4 \varepsilon_T\},$$

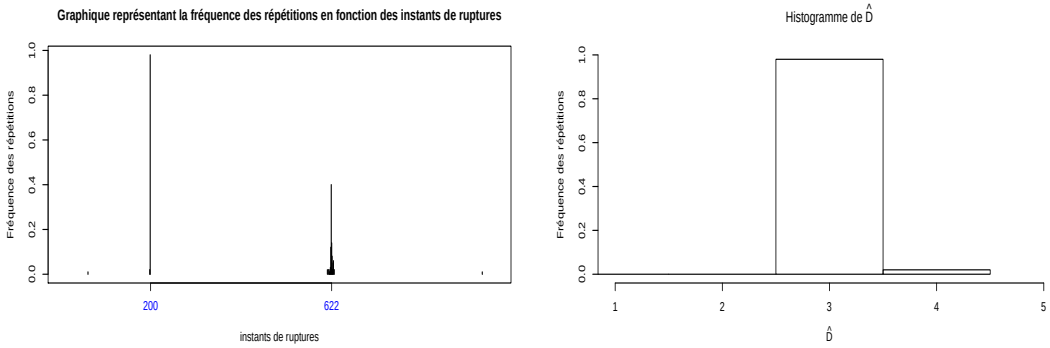
avec $\varepsilon_T = \left(\frac{T\sqrt{\log(T)}}{\left(\min_{i\in[\![1,\dots,D^*]\!]} |\tau_i^* - \tau_{i-1}^*|\right)\left(\min_{i\in[\![1,\dots,D^*-1]\!]} \left\| f_{\tau_i^*} - f_{\tau_i^*-1} \right\|_{\mathcal{H}}\right)} \right)^2$

Limites du théorème
Le théorème 3.1 fournit une expression du paramètre de seuil ζ_T qui doit être calculé pour utiliser KBS. Néanmoins, celui-ci dépend des paramètres Θ et w qui sont inconnus car ils sont liés à la fonction de régression.

4 Simulations

On se place dans le cadre de la détection de ruptures dans la moyenne d'un signal réel $\{X_t\}_{t\in[\![1,T]\!]}$ de longueur $T = 1000$ (on utilise KBS avec le noyau linéaire). La vraie segmentation m^* a deux instants de ruptures $\tau_1^* = 200$ et $\tau_2^* = 622$, et $f_1 = 0.7$, $f_{\tau_1^*} = 5.4$ et $f_{\tau_2^*} = 4.2$. On effectue $m = 100$ répétitions d'un bruit de distribution $\mathcal{N}(0,1)$. On a construit la fonction de régression $\{f_t\}_{t\in[\![1,T]\!]}$ de telle sorte qu'elle vérifie les hypothèses 1 et 2. On évalue la bonne performance de KBS sur ces répétitions.

La figure 1(a) montre que la fréquence des instants de ruptures estimés est moins étalée pour $\tau_1^* = 200$ que pour $\tau_2^* = 622$: cela est due à un plus fort saut de palier en τ_1^* qu'en τ_2^*. La figure 1(b) montre une fréquence proche de 1 pour $\hat{D} = D^* = 3$, ce qui confirme que l'évènement $\hat{D} = D^*$ se réalise avec grande probabilité.



(a) Fréquence des instants de ruptures estimés par KBS.

(b) Fréquence des nombre de segments estimés par KBS ($D^* = 3$).

5 Conclusion

Pour répondre au probème de détection de ruptures dans la distribution, l'algorithme KBS fournit des estimateurs des instants de ruptures dont nous avons prouvé la consistance sous conditions. Un avantage de notre méthode KBS (complexité en temps $O\left(\hat{D}T^2\right)$) est qu'elle est moins coûteuse en temps de calcul que la méthode basée sur la programmation dynamique (complexité en temps $O(D_{\max}T^4)$). L'estimation du paramètre de seuil ζ_T dont dépend KBS reste à explorer.

Bibliographie

[1] Arlot S., Celisse A. et Harchaoui Z. (2015), A kernel change-point algorithm via model selection, *Journal of Machine Learning Research*.
[2] Birgé L., Massart P. (2007). Minimal penalties for Gaussian model selection, *Probab. Th. Rel. Fields*, 138, 33-73.
[3] Boucheron S., Lugosi G., Massart P. (2013), Concentration Inequalities A Nonasymptotic Theory of Independence, *Oxford University Press*.
[4] Fryzlewicz P. (2014), Wild binary segmentation for multiple change-point detection, The *Annals of Statistics*, 42, 6, 2243-2281.
[5] Lebarbier E. (2005), Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection, *Signal Processing*, 85, 4, 717-736.