Détection de ruptures multiples pour les processus de Poisson

Charlotte Dion-Blanc, E. Lebarbier, Stephane S. Robin

Détection de ruptures multiples pour les processus de Poisson

Charlotte Dion-Blanc [1] & Emilie Lebarbier [2] & Stéphane Robin [1]

[1] *LPSM, Sorbonne Unviersité*
[2] *Modal'X, Université de Nanterre*

Résumé.

La détection des points de rupture a pour but de découvrir les changements de comportement qui se cachent derrière les données de séquences temporelles. Dans ce travail, nous étudions le cas où les données proviennent d'un processus de Poisson hétérogène d'intensité constante par morceaux. Nous présentons une méthodologie de détection, hors ligne, de points de changement multiples, basée sur l'estimateur de contraste minimum. En particulier, nous expliquons comment traiter la nature continue du processus. Enfin, nous sélectionnons le nombre de segments à l'aide d'une procédure de validation croisée, rendu possible ici par une propriété spécifique au processus de Poisson. La méthode proposée est implémentée dans le package R `CptPointProcess`.

Mots-clés. Processus de Poisson, détection de ruptures, cross-validation.

Abstract. Change-point detection aims at discovering behavior changes lying behind time sequences data. In this work, we investigate the case where the data come from an inhomogenous Poisson process, with a piece-wise constant intensity. We present an offline multiple change-point detection methodology based on minimum contrast estimator. In particular we explain how to deal with the continuous nature of the process. Besides, we select the appropriate number of regimes through a cross-validation procedure which is really convenient here due to the nature of the Poisson process. The proposed method is implemented in the `CptPointProcess` R package.

Keywords. Point process, change-point detection, cross-validation.

1 Prsentation du problme et motivation

La détection d'instants de ruptures multiples est l'une des tâches les plus courantes en statistique et en traitement du signal. Le problème peut être formulé de la manière suivante : en considérant un processus observé dans le temps, identifier les instants avant et après lesquels l'intensité du processus est différente. Nous considérons ici un processus de Poisson hétérogène, en supposant que sa fonction d'intensité est constante par morceaux et nous cherchons à trouver les moments où la valeur de l'intensité change.

D'un point de vue statistique, la détection des points de ruptures pose trois problèmes principaux : (i) la localisation des points de changement, (ii) l'estimation des paramètres

régissant le processus entre les changements et (*iii*) l'estimation du nombre de points de rupture. Le problème (*ii*) n'est généralement pas difficile à traiter en utilisant, par exemple, le maximum de vraisemblance, une fois que les points de changement ont été identifiés. La détermination du nombre de points de ruptures (*iii*) est un problème typique de sélection de modèle.

La détermination de la position des changements (*i*) pose ici un problème plus complexe, principalement parce qu'elle fait appel à la minimisation du contraste (disons, la log-vraisemblance négative), qui n'est pas une fonction continue des points de changement. Ce dernier problème peut évidemment être contourné en discrétisant le temps (voir Achcar *et al.*(2008) pour la médecine, Achcar *et al.*(2011) pour la surveillance de la pollution, et Shen et Zhang (2012) pour la génomique), au prix du calcul. L'exploration de l'espace de segmentation continu a été envisagée par West et Ogden (1997) ou Yang et Kuo (2001) pour détecter un point de changement unique dans un processus de Poisson. Yang et Kuo (2001) prouvent que la log-vraisemblance négative est en fait concave entre deux temps d'événements successifs, ce qui permet de résoudre le problème de manière simple et exacte. Cette observation est en fait essentielle et ouvre la voie à la stratégie que nous proposons ici pour la détection de plusieurs points de changement et exacte.

2 Modèle

Considérons un processus de Poisson sur l'intervalle de temps $[0,1]$: $\{N_t\}_{0\leq t\leq 1}$, d'intensité $t \mapsto \lambda(t)$. On note $N_t = N((0,t])$ le nombre d'évènements sur l'intervalle $[0,t]$. Les évènements arrivant sur l'intervalle d'observation sont notés $T_1,\ldots,T_{N_1}$, avec la convention $T_0 = 0$ et $T_{N_1+1} = 1$. On suppose que $N_{T_i} - N_{T_i^-} = 1$ et $T_j - T_j^- = 0$ pour tout j. On notera dans la suite $N_1 = n.$ le nombre de temps de sauts observés sur $[0,1]$.

On suppose que l'intensité λ du processus est constante par morceaux. Plus précisément, si m est une partition de $[0,1]$ composée de K segments notés $I_k := (\tau_{k-1},\tau_k]$ pour $k = 1,\ldots,K$ avec $0 = \tau_0 < \tau_1 < \ldots < \tau_K = 1$ (les points de ruptures), alors,

$$\lambda(t) := \sum_{k=1}^{K} \lambda_k \mathbf{1}_{I_k}(t) \tag{1}$$

et on note la segmentation $m := \{I_k\}_{1\leq k\leq K}$ et $\tau = (\tau_{\mathbf{1}},\tau_{\mathbf{2}},\ldots,\tau_{\mathbf{K-1}})$ le vecteur des $K-1$ temps de changement. Ce processus est un processus de Poisson homogène par morceaux.

De plus, on note $\Delta\tau_k := \tau_k - \tau_{k-1}$ la longueur de l'intervalle I_k et $\Delta N_k := N((\tau_{k-1},\tau_k]) = N(\tau_k) - N(\tau_{k-1})$ le nombre d'évènements de l'intervalle I_k. La vraisemblance d'une trajectoire $\{N_t\}_{0\leq t\leq 1}$ (que l'on notera N pour raccourcir) est alors

$$p_P(N;\tau,\lambda) = \prod_{k=1}^{K} e^{-\lambda_k\Delta\tau_k}\lambda_k^{\Delta N_k}.$$

On voit facilement que pour une séquence donnée τ, $p_P(N;\tau,\lambda)$ est maximale pour $\widehat{\lambda}_k = \Delta N_k/\Delta\tau_k$, $1 \leq k \leq K$. Le but est donc maintenant de trouver les points de ruptures qui

minimisent le contraste de Poisson suivant :

$$\gamma_P(\tau; N) := -\log p_P(N; \tau, \widehat{\lambda}) = \sum_{k=1}^{K} \Delta N_k \left(1 - \log\left(\frac{\Delta N_k}{\Delta \tau_k}\right)\right). \tag{2}$$

3 Estimation exacte des instants de ruptures

Nous considérons maintenant la détermination de l'ensemble optimal de la localisation des points de changement $\widehat{\tau}$, défini comme le minimiseur du contraste $\gamma_P(\tau; N)$, sur $(0,1)^{K-1}$. Il est important de noter que γ_P n'est ni convexe ni continu par rapport à τ_k.

3.1 Espace de segmentation

Pour un nombre fixé de segments K, on définit l'espace des segmentations comme l'ensemble des partitions de $(0,1)$ en K segments:

$$\mathcal{M}^K := \left\{\tau = (\tau_1, \ldots, \tau_{K-1}) \in (0,1)^{K-1}; 0 = \tau_0 < \tau_1 < \ldots < \tau_K = 1\right\}.$$

Chaque segmentation de $\mathcal{M}^K$ est une séquence de temps de ruptures, de taille $K-1$. Puis on définit l'ensemble des K-uplets d'entiers sommants à n :

$$\Upsilon^{K,n} := \left\{\nu := (\nu_1, \ldots, \nu_K) \in \{0, 1, \ldots, n\}^K, \ \sum_{k=1}^{K} \nu_k = n\right\}.$$

C'est un espace fini. De plus, on ne prend pas en considération les segmentations ν qui auraient deux zéros successifs au moins, on note $\Upsilon_\star^{K,n}$ le nouvel espace qui exclut ces configurations (si $\nu_K = 0$ alors $\nu_{k-1} \neq 0, \nu_{k+1} \neq 0$).

Maintenant, on définit pour $\nu \in \Upsilon_\star^{K,n}$ et une observation $N = \{N_t\}_{0 \leq t \leq 1}$ avec n vnements, le sous espace des segmentations de $\mathcal{M}^K$ dont le nombre d'évènements par segment est donné par ν :

$$\mathcal{M}_\nu^K(N) := \left\{\tau \in \mathcal{M}^K, \forall\ 1 \leq k \leq K :\ \Delta N_k = \nu_k\right\}.$$

Il est importante de noter que cette contrainte impose que pour chaque $1 \leq k \leq K-1$,

$$\tau_k \in \left[T_{\sum_{j=1}^{k} \nu_j}, T_{\sum_{j=1}^{k} \nu_j + 1}\right).$$

On a donc finalement partitionné l'espace des segmentations en $\mathcal{M}_\star^K = \bigcup_{\nu \in \Upsilon_\star^{K,n}} \mathcal{M}_\nu^K(N)$.

3.2 Estimation par minimisation de contraste

On cherche maintenant $\widehat{\tau}$ ayant K segments tel que:

$$\widehat{\tau} = \arg\min_{\tau \in \mathcal{M}_\star^K} \gamma_P(\tau).$$

La dimension K de la segmentation est fixe dans ce paragraphe. Remarquons que le contraste γ_P donné en équation (2) est additif par segment, c'est-à-dire que l'on peut s'écrire : pour tout $\nu \in \Upsilon_\star^n$ et pour tout $\tau \in \mathcal{M}_\nu(N)$,

$$\gamma_P(\tau) = \sum_{k=1}^{K} C_P(\nu_k, \Delta\tau_k),$$

avec $C_P(\nu_k, \Delta\tau_k) = \nu_k \left(1 - \log\left(\nu_k/\Delta\tau_k\right)\right)$.

De plus, pour tout $\nu \in \Upsilon_\star^n$, $C_P(\nu_k, \Delta\tau_k)$ est une fonction concave en $\Delta\tau_k$. On peut en déduire le résultat suivant.

Proposition 3.1 *Pour tout $\nu \in \Upsilon_\star^n$, le contraste $\tau \to \gamma_P(\tau)$ est concave sur $\mathcal{M}_\nu^K(N)$.*

Ainsi, les points de ruptures sont nécessairement localisés en T_i ou juste avant en T_i^-, c'est ce que dit le théorème suivant.

Théorème 3.2 *Pour $\nu \in \Upsilon_\star^{K,n}$,*

$$\widehat{\tau} = \underset{\tau \in \mathcal{M}_\nu^K(N)}{\arg\min}\, \gamma_P(\tau) \in \{T_{\nu_1}, T^-_{\nu_1+1}\} \times \{T_{\nu_1+\nu_2}, T^-_{\nu_1+\nu_2+1}\} \times \ldots \times \{T_{\nu_1+\ldots\nu_K}, T^-_{\nu_1+\ldots\nu_K+1}\}.$$

Enfin, si $K > 2$, les configurations $(\nu_k, \Delta\tau_k) = (0, 0)$ ne sont pas considérées.

Le problème d'optimisation global est donc réduit à un problème discret sur une grille finie de cardinal $2n$:

$$\{T_1^-, T_1, T_2^-, T_2, \ldots, T_n^-, T_n\},$$

Un algorithme classique de programmation dynamique permet de résoudre ce problème de façon efficace.

3.3 Modification bayésienne du contraste

Il faut noter que pour $K > 2$, la segmentation optimale contient nécessairement des segments de longueur nulle car $C_P(1, 0) = -\infty$. Une telle solution n'est pourtant pas souhaitable.

On propose d'adopter une approche bayésienne et de définir un nouveau contraste basé sur la vraisemblance mais qui n'a pas ce défaut.

On suppose que les paramètres λ_k de l'intensité sont des variables aléatoires i.i.d. de loi Gamma de paramètres $a > 0$ and $b > 0$, notée $p_G(\lambda; a, b)$ Alors, pour une séquence τ donnée, la distribution marginale de N est

$$p_{PG}(N; \tau, a, b) = \int p_P(N; \lambda, \tau)\ p_G(\lambda; a, b) d\lambda = \prod_{k=1}^{K} \frac{b^a\ \Gamma(\Delta N_k + a)}{\Gamma(a)\ (\Delta\tau_k + b)^{\Delta N_k + a}}.$$

Pour $\nu \in \Upsilon_\star^{K,n}$ fixé et pour tout $\tau \in \mathcal{M}_\nu^K(N)$, on propose le contraste Poisson-Gamma suivant

$$\begin{aligned}\gamma_{PG}(\tau) &= -\log p_{PG}(N \mid \tau; a, b) \\ &= \sum_{k=1}^{K} \Big(-a \log b + \log \Gamma(a) + \widetilde{a}_k \log \widetilde{b}_k - \log \Gamma(\widetilde{a}_k) \Big) = \sum_{k=1}^{K} C_{PG}(\nu_k, \Delta\tau_k) \end{aligned} \quad (3)$$

avec $\widetilde{a}_k = \nu_k + a$, $\widetilde{b}_k = \Delta\tau_k + b$. Ce contraste est encore additif par segment et C_{PG} est concave en $\Delta\tau_k$. De plus, $C_{PG}(1, \Delta\tau_k)$ est maintenant borné inférieurement ce qui résout le problème du précédent contraste.

Enfin, en remarquant que la loi conditionnelle de $(\lambda \mid N; \tau)$ est une loi gamma de paramètres $\widetilde{a}_k$ et $\widetilde{b}_k$, la moyenne a posteriori est

$$\mathbb{E}[\lambda_k \mid N, \tau] = \widetilde{a}_k / \widetilde{b}_k. \quad (4)$$

On peut alors estimer les paramètres d'intensité par $\widehat{\lambda}_k = \widetilde{a}_k / \widetilde{b}_k$.

4 Algorithme de détection de ruptures multiples

Nous proposons maintenant une procédure pour sélection K, le nombre de segments de la partition. Rappelons pour cela une propriété important des processus de Poisson.

Proposition 4.1 *Soit N un processus de Poisson de fonction d'intensité $\lambda(t)$. En échantillonnant chaque temps d'évènement de N avec probabilité f donne un processus de Poisson hétérogène N^A d'intensité $\lambda^A(t) = f\lambda(t)$. Les évènements restants forment un processus de Poisson N^B d'intensité $\lambda^B(t) = (1-f)\lambda(t)$, indépendant de N^A.*

Cette proposition a deux conséquences. Tout d'abord, la fonction d'intensité d'intérêt est constante par morceaux, donc λ^A et λ^B aussi, avec les mêmes instants de ruptures que λ. Ensuite, quelque soit la forme de λ, le ratio des intensités de N^A et N^B est constant égal à $\lambda^B(t)/\lambda^A(t) \equiv (1-f)/f$. Ce constat conduit à utiliser la procédure de cross-validation suivante : (i) échantillonner N avec une probabilité f pour former le processus d'apprentissage N^L et celui de test N^T; (ii) pour une collection de valeurs de K, estimer $(\widehat{\tau}^{K,L}, \widehat{\lambda}^{K,L})$ en utilisant N^L, (iii) évaluer le contraste sur le processus test N^T avec les paramètres $(\widehat{\tau}^{K,L}, \frac{1-f}{f}\widehat{\lambda}^{K,L})$.

Voici l'algorithme complet pour M répétitions.

Algorithme 4.2

Input: *une réalisation du processus N.*

Cross-validation: *pour $m = 1$ à M*

1. *créer l'échantillon d'apprentissage $N^{m,L}$ à partir de N avec probabilité f, et former l'échantillon test $N^{m,T}$ avec les temps restants,*

2. *pour* $K = 1$ *à* $K_{\max}$,

- *segmenter l'échantillon d'apprentissage* $N^{m,L}$ *en utilisant* γ_{PG} (3) *pour obtenir*

$$\widehat{\tau}^{m,K,L} = \underset{\nu \in \Upsilon_\star^{K,n(m,L)}}{\arg\min} \underset{\tau \in \mathcal{M}_\nu^K(N^{m,L})}{\min} \gamma_{PG}(\tau),$$

  où $n(m, L)$ *est le nombre d'évènement de* N^L

- *calculer les estimateurs* $\widehat{\lambda}^{m,K,L}$ *en utilisant* (4)

$$\widehat{\lambda}_k^{m,K,L} = \frac{a^{m,L} + \Delta N_k^{m,L}}{b^{m,L} + \Delta \hat{\tau}_k^{m,K,L}} \quad \textit{pour tout } k = 1, \dots K$$

- *calculer*

$$\gamma^{m,K,T} = -\log p_P\left(N^{m,K,T}; \widehat{\tau}^{m,K,L}, \frac{1-f}{f}\widehat{\lambda}^{m,K,L}\right).$$

Moyenne *pour* $K = 1$ *à* $K_{\max}$, *calculer le contraste moyen*

$$\overline{\gamma}^{K,T} = \frac{1}{M}\sum_{m=1}^{M} \gamma^{m,K,T}.$$

Sélection *sélectionner* K *comme*

$$\widehat{K} = \underset{K}{\arg\min} \overline{\gamma}^{K,T}.$$

5 Exemple

Nous considérons ici les éruptions du volcan Kilauea à Hawaii, présentées par Ho et Badhuri (2017). Le jeu de données est constitué des dates d'éruptions enregistrées de l'année 1750 à l'année 1983. Sur cette périodes $n = 63$ éruptions ont été observées. Les données originales ne rapportent que le nombre d'éruptions par an (allant de 0 à 4). Les temps continus ont été restaurés en associant à chaque éruption une date uniformément distribuée, dans l'année correspondante.

Le résultat de l'algorithme est présenté en figure 1. On représente à gauche l'évolution du contraste en fonction de la valeur de K et la valeur sélectionnée (détectée par la barre verticale), et à droite la segmentation finale sur le processus avec $\widehat{K} = 4$ segments.

Bibliographie

Achcar, J. et Martinez, E. et Ruffino-Netto, A. et Paulino, C. et Soares,P. *A statistical model investigating the prevalence of tuberculosis in new york city using counting processes with two change-points.* Epidemiology & Infection, 136(12):1599–1605, 2008.

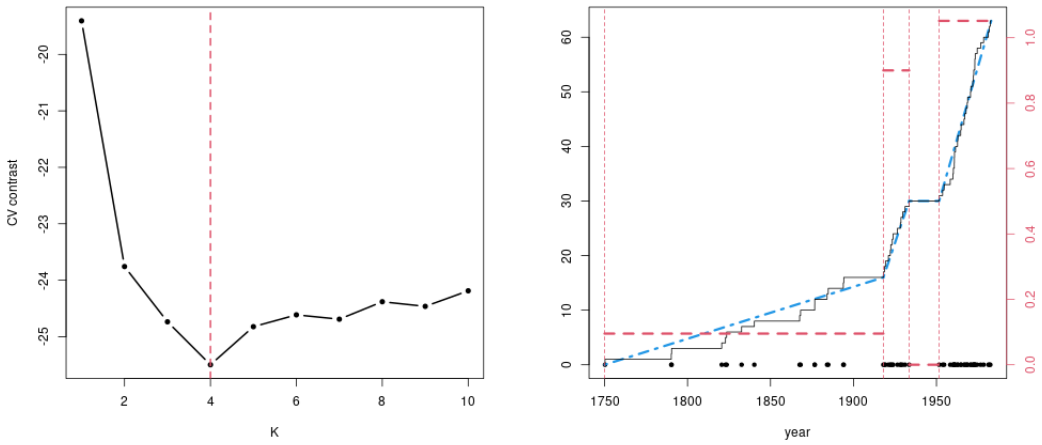

Figure 1: Eruptions du volcan Kilauea. Gauche : sélection du nombre de segments K: line noire = critère $\overline{\gamma}^{K,T}$, ligne verticale pointillée = $\widehat{K}$. Droite : les points noirs sont les temps d'éruption, les lignes pointillées rouges verticales sont $\widehat{\tau}_k$, les lignes bleues représentent l'intensité estimée cumulée, et les lignes rouges pointillées horizontales sont les $\widehat{\lambda}_k$.

Achcar,J et Rodrigues, E. et Tzintzun, G. *Using non-homogeneous poisson models with multiple change-points to estimate the number of ozone exceedances in mexico city.* Environmetrics 22(1):1–12, 2011.

Ho, C. et Bhaduri, M. *A quantitative insight into the dependence dynamics of the Kilauea and Mauna Loa volcanoes, Hawaii.* Mathematical Geosciences 49(7):893–911, 2017.

Shen, J. et Zhang, N. *Change-point model on nonhomogeneous poisson processes with application in copy number profiling by next-generation dna sequencing.* The Annals of Applied Statistics, 6(2):476–496, 2012.

West, W et and Ogden, T. *Continuous-time estimation of a change-point in a poisson process.* Journal of Statistical Computation and Simulation, 56(4):293–302, 1997.

Young, T. et Kuo, L *Bayesian binary segmentation procedure for a poisson process with multiple changepoints.* Journal of Computational and Graphical Statistics, 10(4):772–785, 2001.