Estimation non-paramétrique du taux de saut pour des processus de renouvellement marqués non-homogènes

Romain Azaïs, François Dufour & Anne Gégout-Petit

INRIA Equipe-Projet CQFD
Institut de Mathématiques de Bordeaux UMR CNRS 5251
Université de Bordeaux

Résumé. On s'intéresse à l'estimation non-paramétrique du taux de saut et du taux de saut cumulé pour une classe générale de processus de renouvellement marqués non-homogènes, définis sur un espace métrique séparable. L'estimation se fait dans le cadre d'une observation en temps long du processus. Elle est basée sur une généralisation du modèle à intensité multiplicative, introduit par Aalen dans les années soixante-dix. Nous proposons des estimateurs de ces deux fonctions et nous démontrons des résultats de consistance sous des hypothèses portant sur les caractéristiques du processus. Un exemple numérique illustre le bon comportement des estimateurs.

Mots-clés. Processus de renouvellement marqué non-homogène, estimation non-paramétrique, estimation de taux de saut, estimateur de Nelson-Aalen, consistance asymptotique, chaînes de Markov ergodiques.

Abstract. In this work, we deal with the nonparametric estimation of the jump rate and the cumulative rate for a general class of non-homogeneous marked renewal processes, defined on a separable metric space. In our framework, the estimation needs only one observation of the process within a long time. Our approach is based on a generalization of the multiplicative intensity model, introduced by Aalen in the seventies. We provide consistent estimators of these two functions, under some assumptions related only to the primitive data of the process. A numerical example illustrates the good behavior of our estimators.

Keywords. Non-homogeneous marked renewal process, nonparametric estimation, jump rate estimation, Nelson-Aalen estimator, asymptotic consistency, ergodic Markov chains.

Introduction

On s'intéresse ici à l'estimation non-paramétrique du taux de saut pour une classe générale de processus de renouvellement marqués non-homogènes, lorsqu'une seule observation du processus en temps long est disponible. La méthode d'estimation que nous proposons est basée sur une généralisation du célèbre modèle à intensité multiplicative introduit par Aalen dans [1].

On considère une classe de processus de renouvellement marqués non-homogènes, définis sur un sous-ensemble ouvert E d'un espace métrique séparable $(\mathcal{E}, d)$. La dynamique d'un tel processus $(X_t)_{t\geq 0}$ est définie comme suit. On considère $(Z_n)_{n\geq 0}$ une chaîne de Markov sur $(E, \mathcal{B}(E))$ définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbf{P}_x)$, où $Z_0 = x$ presque sûrement, $x \in E$. Il existe donc une fonction ψ et une suite $(\varepsilon_n)_{n\geq 0}$ de variables aléatoires indépendantes et identiquement distribuées telles que,

$$\forall n \geq 1, \quad Z_n = \psi(Z_{n-1}, \varepsilon_{n-1}).$$

Soient $\lambda : E \times \mathbf{R}_+ \to \mathbf{R}_+$ une application mesurable et Φ un flot déterministe sur E. On peut associer à Φ la fonction déterministe d'atteinte de la frontière $t^\star$,

$$\forall \xi \in E, \quad t^\star(\xi) = \inf\{t > 0 \ : \ \Phi(\xi, t) \in \partial E\}.$$

Notons que la fonction $t^\star$ peut prendre la valeur $+\infty$. On construit de manière itérative une suite $(S_n)_{n\geq 1}$ à valeurs dans $(\mathbf{R}_+, \mathcal{B}(\mathbf{R}_+))$. Pour tout $n \geq 1$, la loi de S_n vérifie, pour tout $t \geq 0$,

$$\begin{aligned} \mathbf{P}_{\nu_0}\big(S_n > t | \ \{Z_i : i \geq 0\}, S_1, \ldots, S_{n-1}\big) &= \mathbf{P}_{\nu_0}\big(S_n > t | \ Z_{n-1}\big) \\ &= \exp\Big(- \int_0^t \lambda(Z_{n-1}, s)\mathrm{d}s\Big)\mathbf{1}_{\{0\leq t<t^\star(Z_{n-1})\}}. \end{aligned}$$

Dans ce cas, il existe une fonction φ et une suite de variables aléatoires indépendantes et identiquement distribuées $(\delta_n)_{n\geq 0}$, indépendante de la suite $(\varepsilon_n)_{n\geq 0}$, telle que,

$$\forall n \geq 1, \ S_n = \varphi(Z_{n-1}, \delta_{n-1}).$$

On suppose aussi que les deux suites $(\varepsilon_n)_{n\geq 0}$ et $(\delta_n)_{n\geq 0}$ sont indépendantes de Z_0. Le processus de renouvellement marqué $(X_t)_{t\geq 0}$ sous-jacent est défini par,

$$\forall t \geq 0, \ X_t = Z_n \quad \text{si} \quad S_0 + \cdots + S_n \leq t < S_0 + \cdots + S_{n+1},$$

avec la convention $S_0 = 0$. Dans ce cas, les Z_n sont les marques du processus.

Notre objectif ici est de proposer un estimateur non-paramétrique de λ et du taux de saut cumulé, à partir d'une seule observation du processus $(X_t)_{t\geq 0}$. Les résultats que nous présentons sont développés dans notre papier [3].

Dans les années 1970, Aalen s'est intéressé – voir [1], – au modèle à intensité multiplicative. Étant donné un processus de comptage N, ce modèle stipule l'existence d'un processus prévisible Y et d'une fonction déterministe λ – appelée taux de saut en statistique des processus ou taux de risque en analyse de survie, – tels que l'intensité stochastique de N se mette sous la forme $Y\lambda$. Dans ce cadre, Aalen a proposé un estimateur consistant du taux de saut cumulé $\Lambda = \int \lambda$, appelé estimateur de Nelson-Aalen. En 1983, Ramlau-Hansen s'est intéressé dans [8] au lissage par des méthodes à noyau de cet estimateur et a obtenu ainsi une méthode d'estimation directe de λ.

Dans de nombreux problèmes statistiques, on cherche à estimer un taux de saut dépendant à la fois du temps et d'une marque spatiale. Celle-ci peut être vue comme une covariable en analyse de survie, ou comme une marque en statistique des processus. De nombreuses méthodes semi-paramétriques ont été développées lorsque la covariable est à valeurs dans un espace continu. C'est par exemple le cas du modèle de Cox [5] – voir par exemple [2] pour une étude approfondie de ce modèle. De nombreux auteurs se sont intéressés à l'approche non-paramétrique dans différents cadres statistiques [6, 7, 9]. Ces travaux sont à la fois différents et complémentaires des nôtres. Par exemple, les approches de McKeague et Utikal [7], Li et Doss [6] ou Utikal [9] reposent sur la structure euclidienne de l'espace. Dans notre cadre de travail, l'espace d'état est général, et supposé être un espace métrique séparable. Les approches proposées dans la littérature ne sont donc pas appropriées. De plus, dans les papiers mentionnés ci-dessus, les auteurs posent des hypothèses portant à la fois sur des martingales à temps continu, et sur le comportement asymptotique du processus Y. Ces hypothèses peuvent être difficiles à vérifier en pratique, en particulier dans notre problème. De notre côté, nous nous sommes attachés à assurer la consistance de nos estimateurs, sous des hypothèses directement liées aux paramètres définissant le processus.

Résultats

On peut supposer dans un premier temps que le noyau de transition Q ne charge qu'un nombre fini de points. Cela revient à considérer que l'espace d'état du processus est discret. Dans ce cas, nous montrons que le modèle multiplicatif est vérifié pour le processus de comptage

$$N_n(x,t) = \sum_{i=0}^{n-1} \mathbf{1}_{\{Z_i=x\}} \mathbf{1}_{\{S_{i+1}\leq t\}},$$

où x est un point chargé par le noyau Q. On peut alors définir l'estimateur non-paramétrique de Nelson-Aalen de Λ, et se référer à [2] par exemple, pour obtenir des résultats de convergence.

Dans un second temps, nous ne considérons plus que l'espace d'état est discret. Si nous supposons que le noyau Q est diffus, le modèle multiplicatif d'Aalen n'est plus satisfait. En effet, pour tout entier i et pour tout point x de E, la fonction indicatrice $\mathbf{1}_{\{Z_i=x\}}$ vaut presque sûrement 0. Notre approche consiste à introduire une partition finie (A_k) de l'espace d'état. Dans ce cadre, il est naturel de considérer le processus de comptage

$$N_n(A_k,t) = \sum_{i=0}^{n-1} \mathbf{1}_{\{Z_i\in A_k\}} \mathbf{1}_{\{S_{i+1}\leq t\}}.$$

Bien que le modèle multiplicatif ne soit pas vérifié pour ce processus de comptage, l'intensité stochastique de $N_n(A_k,t)$ est presque sûrement équivalente quand n tend vers l'infini au produit $Y_n(A_k,t)l(A_k,t)$, où $l(A_k,t)$ est une fonction *proche* de $\lambda(x,t)$, pour $x \in A_k$, et où

$$Y_n(A_k,t) = \sum_{i=0}^{n-1} \mathbf{1}_{\{Z_i \in A_k\}} \mathbf{1}_{\{S_{i+1} \geq t\}}.$$

Pour $x \in A_k$, il apparaît donc naturel d'estimer $\Lambda(x,t) = \int_0^t \lambda(x,s)\mathrm{d}s$ par

$$\widehat{L}_n(A_k,t) = \int_0^t Y_n(A_k,s)^+ \mathrm{d}N_n(A_k,s),$$

où $Y_n(A_k,t)^+$ est l'inverse généralisé de $Y_n(A_k,t)$. On pose aussi

$$L_n^*(A_k,t) = \int_0^t \mathbf{1}_{\{Y_n(A_k,s)>0\}} l(A_k,s)\mathrm{d}s.$$

Dans les papiers d'Aalen, la différence $\widehat{L}_n(A_k,t) - L_n^*(A_k,t)$ est une martingale à temps continu, alors que dans notre cas, ce n'est pas le cas puisqu'il existe un terme supplémentaire $a_n(t)$ qui tend vers 0 quand n tend vers l'infini. Intuitivement, cela revient à dire que le modèle multiplicatif est asymptotiquement vérifié.

Sous des hypothèses de régularité portant sur les caractéristiques du processus, on montre que $\widehat{L}_n(A,t)$ est un estimateur consistant de $L(A_k,t) = \int_0^t l(A_k,s)\mathrm{d}s$, en utilisant l'inégalité de Lenglart et en contrôlant le comportement asymptotique du terme $a_n(t)$. On en déduit un estimateur consistant de $\Lambda(x,t)$.

Théorème 1 *Soient $\mathcal{K}$ un compact inclus dans E et $\xi \in E$. Pour tous $\varepsilon, \eta > 0$, il existe un entier N et une partition finie $P = (A_k)$ de $\mathcal{K}$ tels que pour tout $n \geq N$ et pour tout t dans un certain intervalle,*

$$\mathbf{P}_\xi\Big(\sup_{x\in\mathcal{K}} \sup_{0\leq s\leq t} \left| \sum_{k=1}^{|P|} \widehat{L}_n(A_k,s)\mathbf{1}_{\{x\in A_k\}} \; - \; \Lambda(x,s) \right| > \eta \Big) < \varepsilon.$$

Dans ce résultat, on suppose qu'il existe une partition finie (A_k) de E, aussi raffinée que l'on veut, telle que la loi invariante de la chaîne de Markov $(Z_n)_{n\geq 0}$ charge l'ensemble A_k, pour tout k. On donne dans [3] un résultat plus général, pour lequel on ne fait pas cette hypothèse. Dans ce cas, il est nécessaire d'estimer la loi invariante de $(Z_n)_{n\geq 0}$. De plus, en lissant cet estimateur par des méthodes à noyau, on obtient un estimateur consistant de λ. On peut trouver dans [3] le résultat correspondant, ainsi que toutes les démonstrations.

Simulation

On considère un processus de renouvellement marqué non-homogène $(X_t)_{t\geq 0}$ défini sur le disque $\mathcal{D} = \{x \in \mathbf{R}^2 : \|x\|_2 \leq 1\}$. Les caractéristiques Q, λ et $t^\star$ sont données pour tout $x = (x_1, x_2) \in \mathcal{D}$, par

- pour tout $A \in \mathcal{B}(\mathbf{R}^2)$, $Q(x, A) = \dfrac{1}{K_x} \displaystyle\int_A \mathbf{1}_{\mathcal{D}}(y) \exp\left(-\frac{1}{8}\|y - x\|_2^2\right) \mathrm{d}y$,
- pour tout $t \geq 0$, $\lambda(x, t) = \dfrac{|x_1| + t}{1 + \|x\|_2}$,
- $t^\star(x) = 2 + \|x\|_2$,

où K_x est la constante de normalisation. On simule une longue trajectoire d'un tel processus : l'observation de 100000 sauts est disponible pour estimer le taux de saut λ. On se focalise sur l'estimation de $\Lambda(x, t)$ et $\lambda(x, t)$, pour $x = (0.2, 0.5)$. On choisit d'approcher $\lambda(x, t)$ par la fonction $l(A_x, t)$, avec $A_x = \{y \in \mathbf{R}^2 : \|y - x\|_2 \leq \varepsilon\}$ et $\varepsilon = 0.2$. La fenêtre de lissage choisie $\beta_n(A_x)$ s'écrit

$$\beta_n(A_x) = \frac{1}{h_n(A_x)^\alpha},$$

où $h_n(A_x)$ est le nombre (aléatoire) de visites dans A_x (environ 4000) et $\alpha = 1/5$. Les résultats sont donnés dans la Figure 1.

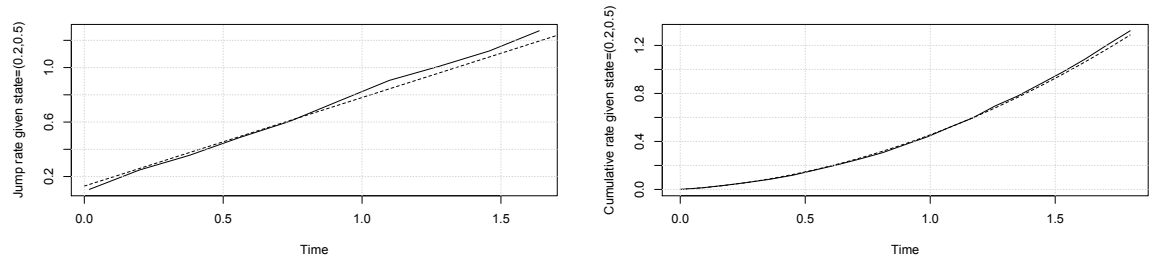

Figure 1: Estimation du taux de saut $\lambda(x, t)$ (à gauche) et du taux de saut cumulé $\Lambda(x, t)$ (à droite) avec $x = (0.2, 0.5)$ et $0 \leq t \leq 1.8$. Les estimateurs sont en lignes pleines, les taux théoriques sont en pointillés.

Conclusion

Nous avons proposé un estimateur du taux de saut et du taux de saut cumulé pour une classe générale de processus de renouvellement marqués non-homogènes. Dans le cadre d'une observation en temps long, nous montrons des résultats de convergence sous des hypothèses d'ergodicité et de régularité des fonctions caractéristiques du processus. De plus, les méthodes que nous proposons sont aussi un point clé de nos travaux [4] sur l'estimation non-paramétrique de la densité conditionnelle des temps inter-sauts pour un processus markovien déterministe par morceaux.

Références

[1] Aalen, O. O. *Statistical inference for a family of counting processes*. ProQuest LLC, Ann Arbor, MI, 1975. Thesis (Ph.D.)–University of California, Berkeley.

[2] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

[3] Azaïs, R., Dufour, F., and Gégout-Petit, A. Nonparametric estimation of the jump rate for non-homogeneous marked renewal processes. *Preprint*.

[4] Azaïs, R., Dufour, F., and Gégout-Petit, A. Nonparametric estimation of the jump rate for piecewise-deterministic Markov processes. *Preprint*.

[5] Cox, D. R. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B 34* (1972), 187–220.

[6] Li, G., and Doss, H. An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist. 23*, 3 (1995), 787–823.

[7] McKeague, I. W., and Utikal, K. J. Inference for a nonlinear counting process regression model. *Ann. Statist. 18*, 3 (1990), 1172–1187.

[8] Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *Ann. Statist. 11*, 2 (1983), 453–466.

[9] Utikal, K. J. Nonparametric inference for Markovian interval processes. *Stochastic Process. Appl. 67*, 1 (1997), 1–23.