Apprentissage pour l'intensité d'événements avec points de rupture

ElMokhtar E. Alaya [1], Stéphane Gaïffas [2], Agathe Guilloux [3]

[1] *LSTA, Université Pierre et Marie Curie-Paris VI,*
Boîte 211, Tour 15-25, 2ème étage,
4 place Jussieu, 75252 Paris Cedex 05, France
elmokhtar.alaya@upmc.fr
[2] *CMAP, Ecole Polytechnique,*
Route de Sacaly, 91128 Palaiseau Cedex, France
stephane.gaiffas@cmap.polytechnique.fr
[3] *LSTA, Université Pierre et Marie Curie-Paris VI, Unité INSERM 762 "Instabilité des Microsatellites et Cancers",*
Boîte 209, Tour 15-16, 2ème étage,
4 place Jussieu, 75252 Paris Cedex 05, France
agathe.guilloux@upmc.fr

Résumé. Nous considérons le problème d'estimation de l'intensité d'un processus de comptage, sous une hypothèse de segmentation parcimonieuse. Nous introduisons une procédure d'estimation basée sur la pénalisation par variation-totale avec poids, permettant une calibration fine de la relaxation convexe de l'hypothèse à priori de segmententation parcimonieuse. Nous proposons des inégalités oracles exactes pour cette procédure avec une vitesse rapide de convergence, et nous démontrons la consistance de cette méthode pour la détection des points de rupture. Ces résultats fournissent ainsi une première garantie théorique pour la segmentation basée sur une relaxation convexe au delà du cadre signal + bruit blanc gaussien habituellement considérée.

Mots-clés. Processus de comptage, Points de Rupture, Variation-Totale, Inégalités d'Oracle

Abstract. We consider the problem of learning the inhomogeneous intensity of a counting process, under a sparse segmentation assumption. We introduce a procedure based on a data-driven weighted total-variation penalization, that proposes a sharp tuning for the convex relaxation of the segmentation prior. We prove sharp oracle inequalities for this procedure with fast rates of convergence, and prove consistency of the method for change-points detection, when the number of change-points are both known and unknown. This provides first theoretical guarantees for segmentation with a convex proxy beyond the standard i.i.d signal + white noise setting.

Keywords. Counting processes, Change points, Total-Variation, Oracle inequalities

1 Introduction

Les processus de comptage sont largement utilisés pour décrire l'occurence d'événements dans des systèmes, par exemple en génomique, biologie, économétrie, communications, etc., cf. Anderson et al. (1993). Dans ces problèmes, le but est d'estimer la fonction d'intensité, qui mesure la probabilité instantanée d'un événement. De nombreuses approches ont été proposées dans la litérature. Ramlau-Hansen (1983) propose un estimateur à noyaux, Grégoire (1993) étudie des moindres carrés avec validation croisée. Plus récemment, Reynaud-Bouret (2006) et Baraud and Birgé (2008) ont étudié l'estimation adaptative de l'intensité par séléction de modèle.
Dans ce travail, nous estimons l'intensité $\lambda_0(t)$ d'un processus de comptage $\{N(t), t \in [0,1]\}$ à partir d'un $n-$échantillon de N. Nous travaillons sous l'hypothèse que λ_0 peut être approximée par une fontion constante par morceaux. Nous proposons de traiter ce problème avec un point de vue *segmentation* de signal, qui consiste à découper le signal en segments de durées variables, où l'intensité est à peu près constante. La segmentation a pour but la détermination des temps de changement de régime dans la dynamique du signal.

Plusieurs exemples, importants en pratique, satisfont ce modèle d'intensité avec points de rupture multiples. Prenons le cas de la génomique. L'apparition des technologies de séquençage de nouvelle génération de l'ADN (NGS pour Next Generation Sequencing), permettent la récolte de données haute fréquence, applelées RNA-seq. Ces derniers peuvent être modélisé comme des réplications d'un processus de comptage nonhomogène avec une intensité constante par morceaux, voir Jeremey, Shen and Zhang (2012), où les auteurs adoptent une approche bayésienne pour l'estimation des temps de ruptures.

2 Modèle

Soit $(N_t)_{0\leq t\leq 1}$ un processus de compatage adapté à une filtration $(\mathcal{F}_t)_{0\leq t\leq 1}$, avec un compensateur Λ_0 de sorte que

$$N(t) - \Lambda_0(t) = M(t), \quad 0 \leq t \leq 1,$$

où $(M_t)_{0\leq t\leq 1}$ est une $(\mathcal{F}_t)$-martingale locale. Le modèle est basé sur une écriture de l'intensité $\lambda_0(t)dt = d\Lambda_0(t)$ du processus de comptage sous la forme

$$\lambda_0(t) = \sum_{\ell=1}^{L_0} \beta_{0,\ell} \mathbf{1}_{(\tau_{0,\ell-1}, \tau_{0,\ell}]}(t)$$

pour $0 \leq t \leq 1$, avec la convention $\tau_{0,0} = 0$ et $\tau_{0,L_0} = 1$.

3 Procédure d'estimation

A partir d'un échantillon de taille n, observé entre $t = 0$ et $t = 1$, les paramètres à estimer sont $(\tau_{0,\ell})_{1\leq\ell\leq L_0}$ et $(\beta_{0,\ell})_{1\leq\ell\leq L_0}$. Pour ce faire, nous adoptons une approche qui consiste à ramener la détection de ruptures multiples à un problème de sélection de variables. Nous considérons des estimateurs basés sur la minimisation d'un risque empirique naturellement associé à ce modèle en ajoutant une pénalisation par variation totale pondérée. La variation-totale est simplement la norme ℓ_1 du gradient (discret) du vecteur de paramètres.

$$\hat{\lambda} = \arg\min_{\lambda} \int_0^1 \lambda^2(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^1 \lambda(t)dN_i(t) + \|\nabla\lambda\|_1.$$

Ce problème de minimisation convexe peut se résoudre de façon efficace, avec des algorithmes proximaux. La pénalisation par variation-totale, couplée à une pénalité de type ℓ_1, est nommée *fused lasso* et a été introduite dans Tibshirani et al. (2005). Dans le cadre signal + bruit blanc gaussien, des garanties théoriques sont données par Harchaoui and Lévy-Leduc (2010) . Blekley and Vert (2011) ont proposé le group fused Lasso pour la détection des points de rupture et proposent un algorithme rapide pour le problème convexe correspondant.

Bibliographie

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

[2] Baraud, Y. and Birgé, L. (2008), Estimating the intensity of a random measure by histogram type estimators, *Probab. Theory Related Fields*, 134, 239–284.

[3] Bleakley, K. and Vert, J. P., (2011), The group fused Lasso for multiple change-point detection, *Technical report HAL-00602121.*

[4] Grégoire, G. (1993), Least squares cross-validation for counting process intensities, *Scand. J. Statist*, 20, 343–360.

[5] Harchaoui, Z. and Lévy-Leduc., C. (2010), Multiple change-point estimation with a total variation penalty, *J. Amer. Statist. Assoc.*, 105, 1480–1493.

[6] Jeremy J. Shen and Nancy R. Zhang (2012), Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next- generation DNA sequencing, *Ann. Appl. Stat.*, 6, 476–496.

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

[8] Reynaud-Bouret, P. (2006), Penalized projection estimators of the Aalen multiplicative intensity, *Bernoulli*, 12, 633–661.

[9] Tibshirani, R., Saunders, M., Rosset, R., Zhu, J. and Knight, K. (2005), Sparsity and smoothness via the fused lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67, 91–108.