Apprentissage pour des processus de comptage avec covariables

Stéphane Gaiffas, Agathe Guilloux

Apprentissage pour des processus de comptage avec covariables

Stéphane Gaïffas & Agathe Guilloux

Université Pierre et Marie Curie - Paris 6
Laboratoire de Statistique Théorique et Appliquée
175 rue du Chevaleret, 75013 PARIS

1. Abstract. We consider the problem of estimation of the intensity of counting process with covariates. For this model, we introduce an empirical risk, and provide risk bounds for estimators obtained by minimization of this empirical risk. To that end, we derive Bernstein's deviation inequalities and maximal inequalities using the generic chaining mecanism, which was introduced by Talagrand (2005). We give also an oracle inequality for the popular algorithm of aggregation with exponential weights. This provides a way of constructing estimators that are adaptive to the smoothness and to the structure of the intensity. We prove that these estimators are adaptive over anisotropic Besov balls, and that the rates are minimax.

1. Résumé. Nous cherchons à estimer l'intensité d'un processes de comptage avec covariables. Nous proposons un risque empirique naturellement associé à ce modèle, et montrons des bornes pour l'erreur des estimateurs obtenus par minimisation de ce risque empirique. Pour ce faire, nous obtenons des inégalités de Bernstein et des inégalités maximales en utilisant le chaînage générique introduit par Talagrand (2005). Nous proposons également une inégalité d'oracle pour l'algorithme d'agrégation avec poids exponentiels. Cela fournit une stratégie permettant de construire des estimateurs adaptatifs en la régularité de l'intensité, et en sa structure. Nous montrons que ces estimateurs sont adaptatifs sur des espaces de Besov anisotropes, et que les vitesses obtenues sont minimax.

2. Le modèle. Soit N un processus de comptage marqué adapté à une filtration $(\mathcal{F}_t)$ avec compensateur Λ par rapport à $(\mathcal{F}_t)$, de sorte que

$$N - \Lambda = M, \tag{1}$$

où M est une $(\mathcal{F}_t)_{t\geq 0}$-martingale locale. Nous considérons le modèle à intensité multiplicative d'Aalen:

$$\Lambda(t, X) = \int_0^t \alpha_0(t, X)Y(s)ds, \text{ for all } t \geq 0 \tag{2}$$

où X est un vecteur de covariables dans $\mathbb{R}^d$ supposé $\mathcal{F}_0$-mesurable, dont la loi est à support compact, le processus Y est prévisible positif et borné, et α_0 est la fonction d'intensité du processus, supposée déterministe, positive et bornée. Nous cherchons à estimer α_0 à partir de l'observation d'un n-échantillon $(X_i, N^i(t), Y^i(t) : 0 \leq t \leq 1)$ de même loi que $(X, Y(t), N(t) : 0 \leq t \leq 1)$ (on observe les trajectoires à horizon fini) pour $i = 1, \ldots, n$.

Plusieurs exemples, importants en pratiques, sont des cas particuliers du modèle considéré ici. Les trois principaux sont les suivants :

- **Données censurées :** Soient T et C des variables aléatoires réelles, X un vecteur aléatoire de covariables dans $\mathbb{R}^d$, de fonctions de répartitions respectives F_T, G et F_X. On suppose que l'on observe pas T directement, mais que l'on observe n copies i.i.d. (T_i^C, δ_i) de la variable censurée $T^C := T \wedge C$ et de $\delta = I(T \leq C)$ qu'on appelle l'indicateur de censure. Dans ce cas, les processus à considérer (voir e.g. Andersen et al. (1993)) sont donnés par $N^i(t) = I(T_i^C \leq t, \delta_i = 1)$ et $Y^i(t) = I(T_i^C \geq t)$, pour $i = 1, \ldots, n$. On suppose que l'hypothèse dite *faible* en censure est satisfaite : T et C sont indépendantes conditionnellement à X. Dans ce cas, l'intensité α_0 est égale au risque instantané conditionel de T sachant $X = x$, donné par $\alpha(t, x) = f_{T|X}(t, x)/(1 - F_{T|X}(t, x))$, où $f_{T|X}$ est la densité et $F_{T|X}$ est la répartition conditionnelle de T sachant X.

- **Processus de Poisson avec covariables :** Soient η^i, pour $i = 1, \ldots, n$, des processus de Cox sur $\mathbb{R}_+$ avec mesure moyenne $\Lambda^i(t) = \int_0^t \alpha(t, X_i)dt$. Dans ce cadre le processus Y est constant p.s. et égal à 1.

- **Processus de Markov :** Soit un n-échantillon d'un processus de Markov inhomogène $P^1, \ldots, P^n$ avec espace d'états fini $\{1, \ldots, k\}$ et soit α_{jl} l'intensité de la transition de l'état j à l'état l. Pour l'individu i avec covariable X_i, $N_{jl}^i(t)$ compte le nombre de transitions directes de l'état j à l avant l'instant t (on peut également ajouter une censure à droite dans ce cadre). Conditionnellement à l'état initial, le processus de comptage N_{jl}^i vérifie le modèle d'Aalen à intensité multiplicative décrit plus haut, avec $N_{jl}^i(t) = \int_0^t \alpha_{jl}(s, X_i)Y_j^i(s)ds + M^i(t)$, où $Y_j^i(t) = I\{P^i(t-) = j\}$, voir Andersen et al. (1993) ou Jacobsen (1982).

4. Un bref aperçu des résultats. Nous considérons des estimateurs basés sur la minimisation du risque empirique suivant :

$$R_n(\alpha) := \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \alpha(t, X_i)^2 Y^i(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^1 \alpha(t, X_i)dN^i(t), \tag{3}$$

qui est la version empirique du risque *théorique*

$$\begin{aligned} R(\alpha) &= \mathsf{E}\Big[\int_0^1 (\alpha(t, X)^2 - 2\alpha(t, X)\alpha_0(t, X))Y(t)dt\Big] \\ &= \|\alpha - \alpha_0\|^2 - \|\alpha_0\|^2, \end{aligned}$$

où $\|\cdot\|$ est la norme pondérée

$$\|\alpha\|^2 := \int_{\mathbb{R}^d}\int_0^1 \alpha(t,x)^2 \mathsf{E}[Y(t)|X=x]dtP_X(dx).$$

Le risque d'excès $R(\alpha) - R(\alpha_0)$ s'écrit donc $R(\alpha) - R(\alpha_0) = \|\alpha - \alpha_0\|^2$. Il est alors naturel de considérer un minimiseur de $R_n(\cdot)$ pour estimer α_0. On fixe donc un ensemble $A \subset L^2 \cap L^\infty$ (appelé souvent *sieve*) et on considère

$$\bar{\alpha}_n \in \operatorname*{argmin}_{\alpha \in A} R_n(\alpha).$$

On propose alors une borne pour $\mathsf{E}^n\|\bar{\alpha}_n - \alpha_0\|^2$. On utilise pour cela la décomposition

$$R_n(\alpha) - R_n(\alpha_0) = \|\alpha - \alpha_0\|_n^2 - \frac{2}{\sqrt{n}} Z_n(\alpha - \alpha_0),$$

où $\|\cdot\|_n$ est la norme empirique

$$\|\alpha\|_n^2 := \frac{1}{n}\sum_{i=1}^n \int_0^1 \alpha(t,X_i)^2 Y^i(t)dt,$$

et Z_n le processus empirique

$$Z_n(\alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^1 \alpha(t,X_i)dM^i(t),$$

où M^i sont les copies indépendantes de l'innovation martingale de (1). On a donc besoin d'une inégalité maximale pour le processus empirique $Z_n(\cdot)$. On montre que l'inégalité suivante (de type Bernstein) est vérifiée :

$$\mathsf{P}^n\Big[Z_n(\alpha) > \delta\sqrt{2x} + \frac{\|\alpha\|_\infty x}{\sqrt{n}},\ \langle Z_n(\alpha)\rangle \le \delta^2\Big] \le \exp(-x), \tag{4}$$

où $\langle Z\rangle$ est la variation prévisible de Z et $\|\alpha\|_\infty := \sup_{(t,x)} |\alpha(t,x)|$. Cela permet d'en déduire, en utilisant le chaînage générique de Talagrand (2005), l'inégalité maximale

$$\mathsf{E}^n\big[\sup_{a\in A} Z_n(a)\big] \le L\Big(\gamma_2(A,\|\cdot\|) + \frac{1}{\sqrt{n}}\gamma_1(A,\|\cdot\|_\infty)\Big), \tag{5}$$

où la fonction $\gamma_\nu(A,\|\cdot\|)$ est une mesure de la complexité de A, qui est une alternative (et une amélioration dans certains cas) à l'intégrale de Dudley, puisque l'on a :

$$\gamma_\nu(A,\|\cdot\|) \le L\int_0^{\mathrm{diam}(A,\|\cdot\|)} (\log N(A,\varepsilon,\|\cdot\|))^{1/\nu} \mathrm{d}\varepsilon, \tag{6}$$

où $\log N(A, \varepsilon, \| \cdot \|)$ est l'entropie métrique de A. On peut également utiliser la méthode dite de troncation adaptative, utilisant l'entropie à crochet, due à Ossiander (1987), permet d'enlever le terme en γ_1 de l'inégalité maximale précédente (et qui provient du régime sous exponentiel de l'inégalité de Berstein (4)). On a en effet

$$\mathsf{E}^n \big[\sup_{a \in A} Z_n(a) \big] \leq L \gamma^{[]}(A), \tag{7}$$

où $\gamma^{[]}$ est la version *à crochet* de la fonction γ. Pour montrer une borne pour $\mathsf{E}^n \|\bar{\alpha}_n - \alpha_0\|^2$, il suffit alors d'utiliser la méthode dite de *localisation*, qu'on appelle aussi *peeling* (voir par exemple Massart (2007)): on suppose que pour tout $\delta > \delta_{\min}$ et toutes les boules $A^\delta := \{a \in A : \|a - a_0\|^2 \leq \delta\}$ avec $a_0 \in A$, on a :

$$\mathsf{E}^n \big[\sup_{a \in A^\delta} Z_n(a) \big] \leq L \varphi(\delta),$$

où $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ est une fonction croissante et continue. Cette hypothèse est facilement vérifiée pour des ensembles A de dimension finie, ou des boules d'espaces fonctionel usuel (comme des espaces de Besov), pour lesquelles on peut produire des recouvrement pour la norme $\| \cdot \|_\infty$, et donc majorer les fonctions γ des boules A^δ. Si φ^{-1} est strictement convexe et est telle que $\delta \mapsto \varphi(\delta)/\delta$ est décroissante, on arrive alors a la borne suivante pour l'ERM :

$$\mathsf{E}^n \|\bar{\alpha}_n - \alpha_0\|^2 \leq \inf_{\alpha \in A} \|\alpha - \alpha_0\|^2 + \varphi^{-1*}\Big(\frac{L}{\sqrt{n}}\Big) \vee \delta_{\min},$$

où φ^{-1*} est la conjuguée convexe de l'inverse de φ. Une borne similaire est proposée dans le cas où A est de dimension finie. Ce résultat permet de retrouver, lorsque A est une boule d'un espace fonctionnel (ou lorsque A est de dimension finie) les vitesses de convergence usuelles, après résolution du problème de *biais-variance* (voir Cucker and Smale (2002) par exemple).

Nous utilisons ensuite les outils présentés précédemment pour montrer une inégalité d'oracle pour l'algorithme d'agrégation avec poids exponentiels. Cela nous fournit une stratégie pour construire des estimateurs adaptatifs en la régularité de α_0, et adaptatifs en sa structure. En effet, un avantage de l'agrégation est que la famille d'estimateurs que l'on agrège est presque arbitraire. On peut donc agréger des estimateurs purement non-paramétriques, et des estimateurs qui travaillent sous une hypothèse semi-paramétrique, du type single-index par exemple. Cela permet de *réduire la dimension*, lorsque cela est possible (quand le risque empirique des estimateurs travaillant sous cette hypothèse est faible). Les résultats d'adaptation en la régularité sont établis sur des classes de Besov *anisotropes*, puisque la fonction d'intensité α_0 a typiquement des comportements très différents dans la direction temporelle t et dans la direction x correspondant aux covariables (on peut penser au modèle de Cox par exemple).

5. Un bref état de l'art sur les processus de comptage. L'estimation non-paramétrique de $\int_0^t \alpha_0(s) ds$ a été initié par Beran (1981). Stute (1986), Dabrowska (1987), McKeague and Utikal (1990) et Li and Doss (1995) ont étendu ces résultats. L'estimation semiparamétrique de α_0 a débuté avec Cox (1972), nous nous référons à Andersen et al. (1993) pour une présentation détaillée de la litérature concernant l'estimation dans ces modèles non-paramétriques. Nous citons également Huang (1999) et Linton et al. (2003) pour des développement récents.

Dans Antoniadis et al. (1999) et Brunel and Comte (2005) sont proposées de nouvelles méthodes d'estimation pour le modèle de régression avec données censurées, comme le seuillage d'ondelettes et la selection de modèle. L'estimation adaptative de la densité conditionnelle en présence de censure a été considérée dans Brunel et al. (2007).

L'estimation nonparamétrique de l'intensité d'un processus de Poisson (sur des espace généraux) par sélection de modèle a été étudiée dans Reynaud-Bouret (2003) et Baraud and Birge (2008), et par seuillage dans des bases d'ondelettes par Reynaud-Bouret and Rivoirard (2008).

Les trois papiers suivants proposent différentes stratégies pour l'estimation de l'intensité d'un processus de comptage sans covariables. Ramlau-Hansen (1983) propose un estimateur à noyaux, Grégoire (1993) étudie des moindres carrés avec cross-validation, et plus récemment, Reynaud-Bouret (2006) a étudié l'estimation adaptative par sélection de modèles. Enfin, nous avons construit dans Comte et al. (2008) des estimateurs adaptatifs de l'intensité α_0 dans le modèle (1), (2) par sélection de modèles.

References

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.

Antoniadis, A., Gregoire, G. and Nason (1999). Density and hazard rate estimation for right censored data using wavelet methods. *JRSS*, **61** 63–84.

Baraud, Y. and Birge, L. (2008). Estimating the intensity of a random measure by histogram type estimators. To appear in Probability Theory and Related Fields, URL http://arxiv.org/abs/math/0608663.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Tech. rep., University of California, Berkeley.

Brunel, E. and Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, **67** 441–475.

Brunel, E., Comte, F. and Lacour, C. (2007). Adaptive estimation of the conditional density in presence of censoring. *Sankhyā*. To appear.

Comte, F., Gaïffas, S. and Guilloux, A. (2008). Adaptive estimation of the conditional intensity of marker-dependent counting processes. Available at http://arxiv.org/abs/0810.4263.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, **34** 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, **39** 1–49 (electronic).

Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Statist.*, **14** 181–197.

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

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.*, **27** 1536–1563.

Jacobsen, M. (1982). *Statistical analysis of counting processes*, vol. 12 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

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

Linton, O. B., Nielsen, J. P. and van de Geer, S. (2003). Estimating multiplicative and additive hazard functions by kernel methods. *Ann. Statist.*, **31** 464–492. Dedicated to the memory of Herbert E. Robbins.

Massart, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

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

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.*, **15** 897–919.

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

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, **126** 103–153.

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

Reynaud-Bouret, P. and Rivoirard, V. (2008). Adaptive thresholding estimation of a poisson intensity with infinite support. Tech. rep., ENS-Paris, Université Paris Sud.

Stute, W. (1986). Conditional empirical processes. *Ann. Statist.*, **14** 638–647.

Talagrand, M. (2005). *The generic chaining*. Springer Monographs in Mathematics, Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.