Apprentissage pour les modèles d'Aalen et de Cox en grande dimension avec covariables temps-dépendantes

Mokhtar Z. Alaya [1] & Thibault Allart [2] & Agathe Guilloux & [3] & Sarah Lemler [4]

[1] *LSTA, Université Pierre et Marie Curie,*
Boîte 208, Tour 15-16, 2ème étage,
4 place Jussieu, 75252 Paris Cedex 05, France
elmokhtar.alaya@upmc.fr
[2] *LSTA, Conservatoire National des Arts et Métiers & Ubisoft*
thibault.allart@gmail.com
[3] *LSTA, Université Pierre et Marie Curie & École Polytechnique,CNRS UMR 7641*
Boîte 209, Tour 15-16, 2ème étage,
4 place Jussieu, 75252 Paris Cedex 05, France
agathe.guilloux@upmc.fr
[4] *École Centrale Supélec*
Laboratoire de Mathématiques Appliquées aux Systèmes
Grande Voie des Vignes, 92295 Châtenay-Malabry
sarah.lemler@genopole.cnrs.fr

Résumé. Nous considérons le problème d'estimation des coefficients de régression dans les modèles d'Alaen et de Cox quand les coefficients et les covariables (en grande dimension) dépendent du temps. Pour cela, nous utilisons une procédure d'estimation spécifique basée sur la pénalisation par variation totale. Nous donnons des inégalités oracles pour les estimateurs proposés, et nous nous intéressons ensuite à la résolution algorithmique de ces estimateurs par des méthodes proximales en optimisation convexe.

Mots-clés. Régression dynamique, variation-totale, inégalités oracles, opérateur proximal.

Abstract. In a high dimensional covariates setting, we consider the problem of estimating the intensity of a counting process in the time-varying Aalen and Cox models. We introduce a covariate-specific weighted total variation penalization, using data-driven weights that correctly scale the penalization along the observation interval. We prove theoretical guaranties and we present a proximal algorithm to solve the convex studied problems. The practical use and effectiveness of the proposed method are demonstrated by simulation studies and a real data example.

Keywords. Dynamic regression, total-variation, oracles inequalitites, proximal operator.

1 Modèles

Les processus de comptage sont largement utilisés pour décrire l'occurrence d'événements dans des systèmes, par exemple en génomique, biologie, économétrie, etc. (voir [1] pour plus de détails). Dans ces problèmes le but est d'estimer la fonction d'intensité, qui mesure la probabilité instantanée d'un événement.

Pour $i = 1, \dots, n$, nous définissons $N_i(t)$ un processus de comptage marqué par un processus $Y_i(t) \in [0, 1]$ sur un intervalle d'étude, $[0, \tau]$. Nous considérons l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et la filtration $(\mathcal{F}_t)_{t\in[0,\tau]}$ définie par

$$\mathcal{F}_t = \sigma\big(\{N_i(s), Y_i(s), X_i(s),\, 0 \leq s \leq t,\, i = 1, \dots, n\}\big),$$

où $X_i(t) = (X_i^1(t), \dots, X_i^p(t)) \in \mathbb{R}^p$ et le vecteur aléatoire de covariables de l'individu i. À partir de l'échantillon d'apprentissage

$$\mathcal{D}_n = \big\{(X_i(t), Y_i(t), N_i(t)) : t \in [0, \tau], i = 1, \dots, n\big\},$$

nous cherchons à estimer la fonction du risque dans le les modèles suivants :

— Modèle d'Aalen

$$\lambda_\star^{\mathrm{A}}(t, X(t)) = X(t)\beta^\star(t), \tag{1}$$

— Modèle de Cox

$$\lambda_\star^{\mathrm{M}}(t, X(t)) = \exp\big(X(t)\beta^\star(t)\big), \tag{2}$$

avec $\beta^\star$ est une fonction à p variables de $[0, \tau]$ à valeurs dans $\mathbb{R}^p$ que l'on cherche à estimer.

Ces modèle incluent plusieurs exemples importants en pratiques : données censurées, processus de Poisson marqué, processus de Markov, voir [11] et [1] pour une liste détaillée. On se contente de rappeler ici le cas des données censurées : soient $T_1, \dots, T_n$ et $C_1, \dots, C_n$ des durées de survie et des durées de censure des n individus considérés. On les consiste en n couples de variables (Z_i, δ_i, X_i) telles que $Z_i = T_i \wedge C_i$ représente la date de l'événement terminal pour le i-ème individu, et $\delta_i = \mathbb{1}(T_i \leq C_i)$ est l'indicateur de la censure. Les processus à considérer sont donnés par $N_i(t) = \mathbb{1}(Z_i \leq t, \delta_i = 1)$ et $Y_i(t) = \mathbb{1}(Z_i \geq t)$, pour tout $i = 1, \dots, n$.

L'estimation non-paramétrique de l'intensité cumulée, en présence de covariables indépendantes de temps, $\int_0^t \lambda(s, x)ds$ a été initié par [3]. Des extensions de ces résulats ont été étudiées dans [7], [16], et [13]. L'estimation semi-paramétrique de $\lambda(t, x)$ a débuté avec [6]. En absence de covariables, la méthode de validation croisée a été suggérée par [9] pour choisir la taille de fenêtre d'un estimateur à noyaux proposé par [19]. D'autres procédures dites adaptatives permettent de construite des estimateurs qui s'adaptent à la régularité de l'intensité. Par exemple, la sélection de modèle étudiée dans [20, 21], [4], et [2] pour l'estimation non-paramétrique de l'intensité d'un processus de Poisson (sur des espace généraux), et par seuillage dans des bases d'ondelettes par [22]. Pour l'estimation paramétrique en grande dimension de (1), [15] considèrent la sélection de covariables, [8] proposent une procédure Lasso avec poids dépendant éventuellement des observations ("data driven") pour estimer (1). [12] et [10] ont étudié le Lasso pour estimer (2).

2 Procédure d'estimation

Pour cela, nous considérons des estimateurs basés sur des histogrammes, voir [17]. Plus précisément, nous disposons d'une L-partition ($L \in \mathbb{N}^*$) de l'intervalle de temps $[0, \tau]$ définie par

$$\varphi_l = \sqrt{\frac{L}{\tau}} \mathbb{1}(I_l) \text{ avec } I_l = \Big(\frac{l-1}{L}\tau, \frac{l}{L}\tau \Big].$$

Pour tout $j = 1, \dots, p$, un candidat pour estimer le j-ème coefficient $\beta_j^\star$ de $\beta^\star$ appartient à l'ensemble de fonctions constantes par morceaux

$$\mathcal{H}_L = \Big\{ \alpha(\cdot) = \sum_{l=1}^{L} \alpha_l \varphi_l(\cdot) : (\alpha_l)_{1 \le l \le L} \in \mathbb{R}_+^L \Big\}.$$

Pour chaque individu i nous lui associons un processus de covariables p-dimensionnel $X_i(t)$, et nous notons par $X_i^j(t)$ le processus associé pour son j-ème covariable. Pour toute fonction à p variables β, estimateur candidat de $\beta^\star$, nous désignons par β_j sa j-ème variable (fonction). Nous définissons l'ensemble d'estimateurs candidats par

$$\Lambda^{\mathrm{A}} = \{ x, t \in [0, \tau] \mapsto \lambda_\beta^{\mathrm{M}}(t, x(t)) = x(t)\beta(t) \mid \forall j \ \beta_j \in \mathcal{H}_L \}$$

pour le modèle d'Aalen et par

$$\Lambda^{\mathrm{M}} = \{ x, t \in [0, \tau] \mapsto \lambda_\beta^{\mathrm{M}}(t, x(t)) = \exp\big(x(t)\beta(t)\big) \mid \forall j \ \beta_j \in \mathcal{H}_L \}$$

pour le modèle de Cox. Pour l'ensemble Λ^{M} ou Λ^{A}, chaque coefficient est une fonction constante par morceaux. β est considérée à la fois comme une fonction à p variables ou comme un vecteur de dimension $p \times L$ défini par

$$\beta = (\beta_{1,\cdot}^\top, \dots, \beta_{p,\cdot}^\top)^\top = (\beta_{1,1}, \dots, \beta_{1,L}, \dots, \beta_{p,1}, \dots, \beta_{p,L})^\top,$$

où $\beta_{j,\cdot}$ appartient à $\mathbb{R}^L$ et $\beta_{j,l}$ est la valeur prise par la j-ème coordonnée dans le l-ème intervalle de notre L-partition $\{I_1, \dots, I_L\}$. Nous considérons la minimisation des fonctionnelles suivantes : les moindres carrées pour le modèle d'Aalen, qui est un critère d'ajustement classique dans ce modèle, voir [20, 21, 14, 8].

$$\ell_n^{\mathrm{A}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \int_0^\tau \big(\lambda_\beta^{\mathrm{A}}(t, X_i(t))\big)^2 Y_i(t) dt - 2 \int_0^\tau \lambda_\beta^{\mathrm{A}}(t, X_i(t)) dN_i(t) \Big\},$$

et la log-vraisemblance pour le modèle de Cox, voir [14, 12]

$$\ell_n^{\mathrm{M}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \int_0^\tau \log\big(\lambda_\beta^{\mathrm{M}}(t, X_i(t))\big) dN_i(t) - \int_0^\tau Y_i(t) \lambda_\beta^{\mathrm{M}}(t, X_i(t)) dt \Big\}.$$

Nous introduisons la pénalité $(\ell_1 + \ell_1)$-variation totale avec poids défini par

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \sum_{j=1}^{p} \left(\hat{\gamma}_{j,1} |\beta_{j,1}| + \sum_{l=2}^{L} \hat{\gamma}_{j,l} |\beta_{j,l} - \beta_{j,l-1}| \right)$$

pour tout $\beta \in \mathbb{R}^{p \times L}$ avec $\hat{\gamma} = (\hat{\gamma}_{1,\cdot}^\top, \ldots, \hat{\gamma}_{p,\cdot}^\top)^\top$, tel que $\hat{\gamma}_{j,\cdot} \in \mathbb{R}_+^L$ pour tout $j = 1, \ldots, p$, donné par

$$\hat{\gamma}_{j,l} \approx \sqrt{\frac{L \log(pL)}{n} \hat{V}_{j,l}}, \text{ avec } \hat{V}_{j,l} = \frac{1}{n} \sum_{i=1}^{n} \int_{\bigcup_{u=l}^{L} I_u} (X_i^j(t))^2 dN_i(t).$$

Nos estimateurs sont définis par $\hat{\lambda}^{\mathrm{A}} = \lambda^{\mathrm{A}}_{\hat{\beta}^{\mathrm{A}}}$ et $\hat{\lambda}^{\mathrm{M}} = \lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}}$ où

$$\hat{\beta}^{\mathrm{A}} = \mathrm{argmin}_{\beta \in \mathbb{R}^{p \times L}} \left\{ \ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right\}, \tag{3}$$

et

$$\hat{\beta}^{\mathrm{M}} = \mathrm{argmin}_{\beta \in \mathbb{R}^{p \times L}} \left\{ \ell_n^{\mathrm{M}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right\}. \tag{4}$$

3 Résumé des résultats obtenus

Nous présentons ici un résumé de nos principaux résultats pour la procédure $(\ell_1 + \ell_1)$-variation totale avec poids définie dans la Section 2. Nous prouvons des inégalités oracles non-asymptotiques à vitesse lente et rapide vérifiées par $\hat{\lambda}^{\mathrm{A}}$ et $\hat{\lambda}^{\mathrm{M}}$. Pour cela, nous définissons deux fonctions de perte :

— quadratique associée au modèle d'Aalen (1)

$$\|\lambda_\star^{\mathrm{A}} - \lambda^{\mathrm{A}}\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\lambda_\star^{\mathrm{A}}(t, X_i(t)) - \lambda^{\mathrm{A}}(t, X_i(t)) \right)^2 Y_i(t) dt},$$

— divergence de Kullback empirique associée au modèle de Cox (2)

$$\begin{aligned} K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\log \lambda_\star^{\mathrm{M}}(t, X_i(t)) - \log \lambda_\beta^{\mathrm{M}}(t, X_i(t)) \right) \lambda_\star^{\mathrm{M}}(t, X_i(t)) Y_i(t) dt \\ - \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\lambda_\star^{\mathrm{M}}(t, X_i(t)) - \lambda_\beta^{\mathrm{M}}(t, X_i(t)) \right) Y_i(t) dt. \end{aligned}$$

Ci-dessous, Théorèmes 1 et 2 sont des inégalités oracles à vitesse lente.

Théorème 1. *Pour $x > 0$ fixé, l'estiamteur $\hat{\lambda}^{\mathrm{A}}$ défini dans (3), vérifie avec une probabilité supérieure à $1 - C_{\mathrm{A}} e^{-x}$, (avec $C_{\mathrm{A}} > 0$)*

$$\|\lambda_\star^{\mathrm{A}} - \hat{\lambda}^{\mathrm{A}}\|_n^2 \leq \inf_{\beta \in \mathbb{R}^{p \times L}} \left(\|\lambda_\star^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2 + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right). \tag{5}$$

Théorème 2. *Pour $x > 0$ fixé, l'estimateur $\hat{\lambda}^{\mathrm{M}}$ defini dans (4), vérifie avec une probabilité supérieur à $1 - C_{\mathrm{M}}\mathrm{e}^{-x}$ ($C_{\mathrm{M}} > 0$),*

$$K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) \leq \inf_{\beta \in \mathbb{R}^{p\times L}} \Big(K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 2||\beta||_{\mathrm{gTV},\hat{\gamma}} \Big). \tag{6}$$

Ces théorèmes admettent l'interprétation suivante : les risques théoriques de nos estimateurs sont bornés par les meilleurs risques atteignables sur l'ensemble des modèles basés sur les histogrammes (Λ^{A} et Λ^{M}), plus un terme satisfaisant

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \asymp \|\beta\|_{\mathrm{gTV}} \max_{j=1,\dots,p} \max_{l=1,\dots,L} \sqrt{\frac{L \log pL}{n} \hat{V}_{j,\ell}}. \tag{7}$$

pour tout $\beta \in \mathbb{R}^{p\times L}$. La norme $\|\cdot\|_{\mathrm{gTV}}$ représente la $(\ell_1 + \ell_1)$-variation totale sans poids (*i.e.* $\hat{\gamma}_{j,l} = 1$). Le terme dominant dans (7) est d'ordre $\|\beta\|_{\mathrm{gTV}} \big(L \log(pL)/n\big)^{1/2}$ (à un facteur de $\log\log$ près).

La résolution algorithmique des problèmes (3) et (4) s'est fait par l'implémentation d'un algorithme proximal de descente du gradient stochastique, voir [18, 5, 23]. Nous illustrons nos méthodes sur des données simulées et réelles en les comparons avec les procédures étudiées dans [14].

Bibliographie

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

[2] Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2) :239–284, 2009.

[3] Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, 1981.

[4] L. Birgé. *Model selection for Poisson processes*, volume Volume 55 of *Lecture Notes–Monograph Series*, pages 32–64. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

[5] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[7] D. M. Dabrowska. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, 14(3) :181–197, 1987.

[8] S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the Lasso. *Electron. J. Stat.*, 6 :522–546, 2012.

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

[10] J. Huang, T. Sun, Z. Ying, Y. Yu, and Zhang C.-H. Oracle inequalities for the lasso in the cox model. *Ann. Statist.*, 41(3) :1142–1165, 2013.

[11] A.F. Karr. *Point processes and their statistical inference*, volume 7. CRC press, 1991.

[12] S. Lemler. Oracle inequalities for the lasso in the high-dimensional multiplicative aalen intensity model. *Les Annales de l'Institut Henri Poincaré, arXiv preprint*, 2013.

[13] G. Li and H. Doss. An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, 23(3) :787–823, 06 1995.

[14] T. Martinussen and T. H Scheike. *Dynamic regression models for survival data*. Springer Science &amp ; Business Media, 2007.

[15] T. Martinussen and T. H. Scheike. Covariate selection for the semiparametric additive risk model. *Scand. J. Statist.*, 36(4) :602–619, 2009.

[16] I. W. McKeague and K. J. Utikal. Inference for a nonlinear counting process regression model. *Ann. Statist.*, 18(3) :1172–1187, 09 1990.

[17] S. A. Murphy and P. K. Sen. Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1) :153–180, 1991.

[18] N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3) :127–239, 2014.

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

[20] P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, 126(1) :103–153, 2003.

[21] P. Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4) :633–661, 2006.

[22] P. Reynaud-Bouret and V. Rivoirard. Adaptive thresholding estimation of a poisson intensity with infinite support. *arXiv preprint arXiv :0801.3157*, 2008.

[23] L. Rosasco, S. Villa, and B. C. Vu. Learning with stochastic proximal gradient.