Classification supervisée de Processus de Cox

Gérard Biau[1] & Benoît Cadre [2] & Quentin Paris [3]

[1] *Université Pierre et Marie Curie, Ecole Normale supérieure*
[2] *ENS Rennes, IRMAR*
[3] *CREST, ENSAE-ParisTech*

Résumé. Nous discutons du problème de la classification supervisée de processus de Cox. Les processus de Cox, qui sont en particulier des processus de comptage, ont vocation à modéliser, par exemple, le nombre de visites d'un patient à l'hôpital au cours du temps et leur étude présente de nombreuses applications en Biologie. Notre étude est basée sur un calcul explicite de la règle de Bayes (i.e. la règle de classification optimale). Nous proposons une stratégie de classification de type boosting : la règle de classification proposée minimise un critère empirique convexe régularisé. Nous présentons une inégalité oracle pour évaluer la performance de notre règle de classification. D'autre part, nous montrerons que cette règle de classification converge vers la règle de Bayes à une vitesse qui s'adapte à la régularité inconnue de l'intensité du processus.

Mots-clés. Processus de Cox, classification supervisée, inégalité oracle, regularisation, calcul stochastique.

Abstract. This article addresses the problem of supervised classification of Cox process trajectories, whose random intensity is driven by some exogenous random covariable. The classification task is achieved through a regularized convex empirical risk minimization procedure, and a nonasymptotic oracle inequality is derived. We show that the algorithm provides a Bayes-risk consistent classifier. Furthermore, it is proved that the classifier converges at a rate which adapts to the unknown regularity of the intensity process. Our results are obtained by taking advantage of martingale and stochastic calculus arguments, which are natural in this context and fully exploit the functional nature of the problem.

Keywords. Cox process, supervised classification, oracle inequality, consistency, regularization, stochastic calculus.

1 Introduction

Un exemple de problème pratique qui motive cette étude est le suivant. Supposons qu'un patient, atteint d'une certaine maladie grave, se rend à l'hôpital à une fréquence dépendant de l'évolution de sa maladie. Supposons disposer, d'autre part, d'un certain nombre de patients dont on a déjà observé l'évolution au cours du temps et dont on a pu décider si cette évolution correspondait à une aggravation ou une rémission de la maladie. Peut-on, à l'aide de ces observations, décider si le nouveau patient observé connait lui même une phase d'aggravation ou de rémission de sa maladie ? Un cadre naturel pour formaliser cette étude est celui de la classification supervisée de données modélisées par des processus de comptage. Les processus de Cox, qui sont un type particulier de processus de comptage, sont un outil intéressant pour modéliser, par exemple, l'évolution du nombre de visites d'un patient à l'hôpital au cours du temps. En effet, l'intensité d'un processus de Cox (que l'on peut en première approximation considérer comme un taux de saut instantané) est aléatoire, ce qui autorise une grande fléxibilité du point de vue de la modélisation. Concrètement, un processus $X = (X_t)_{t\in[0,T]}$ est appelé processus de Cox d'intensité $(\lambda(t, Z_t))_{t\in[0,T]}$ si la loi de X sachant tout le processus $(\lambda(t, Z_t))_{t\in[0,T]}$ est celle d'un processus de Poisson d'intensité $(\lambda(t, Z_t))_{t\in[0,T]}$. Ici, le caractère aléatoire de l'intensité est introduit au moyen d'une covariable $Z = (Z_t)_{t\in[0,T]}$ (pour chaque $t \in [0, T]$, Z_t est à valeurs dans $\mathbb{R}^d$) qui peut être considérée, dans le cas de notre exemple, comme l'évolution du dossier médical du patient au cours du temps.

Pour notre étude, nous supposerons qu'il existe deux fonctions inconnues λ_- et $\lambda_+ : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ caractéristiques du processus d'aggravation et de rémission de la maladie. Nos observations $\mathcal{D}_n = \{(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)\}$ sont i.i.d. et de même loi qu'une variable générique (X, Y, Z). Ici, le processus $X = (X_t)_{t\in[0,T]}$ représente l'évolution du nombre de visites à l'hôpital au cours du temps, la covariable $Z = (Z_t)_{t\in[0,T]}$ représente un certain nombre de facteurs dont dépend l'intensité du processus X et Y représente le label ($Y = -1$: aggravation, $Y = +1$: rémission). Nous supposerons que, conditionnellement au fait que $Y = +1$ (resp. $Y = -1$), X est un processus de Cox d'intensité $(\lambda_+(t, Z_t))_{t\in[0,T]}$ (resp. $(\lambda_-(t, Z_t))_{t\in[0,T]}$).

2 Stratégie de classification

Notre objectif est ici de construire une règle de classification $g_n = g_n(., \mathcal{D}_n) : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$ dont la probabilité d'erreur

$$L(g_n) = \mathbb{P}\left(Y \neq g_n(X, Z)|\mathcal{D}_n\right),$$

est la plus faible possible. Notre stratégie, de type boosting, consiste à construire la règle g_n de la manière suivante. Etant donnée une classe $\mathcal{F}$ de fonctions $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ et la

fonction convexe $\phi(u) = \ln_2(1 + e^u)$, on pose

$$f_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(-Y_i f(X_i, Z_i)).$$

Finalement, on pose

$$g_n = \text{signe}(f_n) = \mathbf{1}\{f_n \geq 0\} - \mathbf{1}\{f_n < 0\}.$$

La performance de cette stratégie dépend évidemment du choix de la fonction ϕ ainsi que de la classe $\mathcal{F}$ choisie. Le choix particulier de la fonction ϕ dans notre cas est motivé par un lien particulier qui existe entre la formulation de notre problème est le principe du maximum de vraisemblance. Le choix de la classe $\mathcal{F}$ que nous proposons est motivé par le résultat préliminaire suivant qui donne explicitement la règle de Bayes.

Theorem 2.1 *La règle de Bayes $g^\star$, qui réalise le minimum*

$$L(g^\star) = \inf_g L(g),$$

où l'inf est pris sur l'ensemble des applications mesurables $g : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$, est telle que

$$g^\star(X, Z) = \textit{signe}\left(\xi - \ln \frac{p_-}{p_+}\right),$$

avec

$$\xi = \int_0^T (\lambda_- - \lambda_+)(s, Z_s)\mathrm{d}s + \int_0^T \ln \frac{\lambda_+}{\lambda_-}(s, Z_s)\mathrm{d}X_s.$$

Ce résultat permet en particulier un choix judicieux de la classe $\mathcal{F}$. En effet, disposant d'une certaine base de fonctions $(\varphi_j)_{j \geq 1}$ dans laquelle les fonctions $\lambda_- - \lambda_+$ et $\ln \frac{\lambda_+}{\lambda_-}$ admettent des développements, il est raisonnable de vouloir approcher la règle de Bayes par une fonction du type signe(f) où f appartient à la classe

$$\mathcal{F}_B = \left\{ f = \sum_{j=1}^{B} [a_j \Phi_j + b_j \Psi_j] + c : \max\left(\sum_{j=1}^{B} |a_j|, \sum_{j=1}^{B} |b_j|, |c| \right) \leq B \right\},$$

où

$$\Phi_j(x, z) = \int_0^T \varphi_j(s, z_s)\mathrm{d}s, \qquad \Psi_j(x, z) = \int_0^T \varphi_j(s, z_s)\mathrm{d}x_s.$$

Ici, le paramètre entier B joue le rôle d'un paramètre de régularisation. Plus B est grand, plus la classe $\mathcal{F}_B$ a de bonnes propriétés d'approximation et plus la procédure d'estimation est difficile. Pour choisir à partir des données le paramètre B de manière à effectuer ce compromis "biais-variance", nous procédons de la manière suivante. Nous introduisons

une suite croissante $(B_k)_{k\geq 1}$ de paramètres entiers de régularisation et définissons pour chaque valeur de $k \geq 1$,

$$\hat{f}_k \in \arg\min_{f\in\mathcal{F}_{B_k}} \frac{1}{n}\sum_{i=1}^{n} \phi(-Y_i f(X_i, Z_i)).$$

Disposant maintenant d'un terme de pénalité pen : $\mathbb{N}_+ \to \mathbb{R}_+$ qui reste à choisir, on pose finalement

$$\hat{f} = \hat{f}_{\hat{k}},$$

où

$$\hat{k} \in \arg\min_{k\geq 1}\{A_n(\hat{f}_k) + \text{pen}(k)\}.$$

Le résultat suivant permet d'évaluer la performance de cette procédure. Ici, on note $A(f) = \mathbb{E}\phi(-Yf(X,Z))$ et $f^\star$ l'unique application mesurable réalisant le minimum de $A(f)$ lorsque f parcourt l'ensemble des application mesurables $\mathcal{X}\times\mathcal{Z} \to \mathbb{R}$.

Theorem 2.2 *Supposons que* $\sum_{k\geq 1} B_k^{-\alpha} \leq 1$ *pour un certain* $\alpha > 0$. *Alors, pour tout* $\delta > 0$, *si*

$$\text{pen}(k) \geq C\left[\frac{B_k^4 \ln n}{n} + \frac{\alpha \ln B_k + \delta}{n}\right],$$

où $C > 0$ *désigne une constante universelle, on obtient*

$$A(\hat{f}_n) - A(f^\star) \leq 2\inf_{k\geq 1}\left\{\inf_{f\in\mathcal{F}_{B_k}} (A(f) - A(f^\star)) + \text{pen}(k)\right\},$$

avec probabilité $1 - e^{-\delta}$.

Le résultat suivant prouve que, sous réserve que les fonctions λ_- et λ_+ soit suffisamment régulières, on peut contrôler le terme de biais qui apparaît dans le résultat précédent.

Theorem 2.3 *Supposons sue* $(\varphi_j)_{j\geq 1}$ *soit une base orthonormale de* $\mathbb{L}^2([0,T]\times[0,1]^d)$ *et que les fonctions* $\lambda_- - \lambda_+$ *et* $\ln\frac{\lambda_+}{\lambda_-}$ *appartiennent à la classe*

$$\mathcal{W}(\beta, M) = \left\{\sum_{j\geq 1} a_j\varphi_j : \sum_{j\geq 1} j^{2\beta}a_j^2 \leq M\right\},$$

pour un certain $\beta \geq 1$ *et un certain* $M > 0$. *Alors, il existe une constante* $C' > 0$ *dépendant explicitement des constantes du problème et telle que pour tout* $B \geq 1$ *entier,*

$$\inf_{f\in\mathcal{F}_B} (A(f) - A(f^\star)) \leq \frac{C'}{B^\beta}.$$

Une conséquence de ces deux résultats est que

$$A(\hat{f}_n) - A(f^\star) = O\left(\frac{\ln n}{n}\right)^{\frac{\beta}{\beta+8}},$$

aver grande probabilité. Finalement, les méthodes récentes qui relient le risque convéxifié $A(f)$ à l'erreur de classification $L(\text{signe}(f))$ permettent d'établir que la règle $g_n = \text{signe}(\hat{f})$ vérifie

$$L(g_n) - L(g^\star) = O\left(\frac{\ln n}{n}\right)^{\frac{\beta}{2\beta+16}},$$

aver grande probabilité.

Pour conclure, nous avons proposé une stratégie de classification dans le contexte de données fonctionnelles en ayant recours à une procédure de minimisation convexe. Pour notre analyse, nous avons utilisé à un certain nombre de résultats issus du calcul stochastique et de la théorie des martingales qui exploitent la nature fonctionnelle du problème. Pour une lecture complémentaire de nos travaux (Biau, Cadre, and Paris, 2013), nous renvoyons le lecteur aux travaux importants et connexes que sont Cox (1955, 1972, 1975); Andersen and Gill (1982); O'Sullivan (1993); Bouzas, Valderrama, Aguilera, and Ruiz-Fuentes (2006); Illian, Benson, Crawford, and Staines (2006); Zhu, Song, and Taylor (2011); Cadre (2012); Denis (2012); Hansen, Reynaud-Bouret, and Rivoirard (2014).

References

P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

G. Biau, B. Cadre, and Q. Paris. Cox process learning. *hal:00820838*, 2013.

P.R. Bouzas, M.J. Valderrama, A.M. Aguilera, and N.R. Ruiz-Fuentes. Modelling the mean of a doubly stochastically poisson process by functional data analysis. *Computational Statistics & Data Analysis*, 50:2655–2667, 2006.

B. Cadre. *Supervised classification of diffusion paths*. Preprint, Ecole Normale Supérieure de Cachan, Antenne de Bretagne, Bruz, 2012.

D.R. Cox. Some statistical methods connected with series of events. Series b. *Journal of the Royal Statistical Society*, 17:129–164, 1955.

D.R. Cox. Regression modes and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

D.R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

C. Denis. *Classification in postural style based on stochastic process modeling.* hal-00653316, 2012.

N.R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, page (to appear), 2014.

J. Illian, E. Benson, J. Crawford, and H. Staines. Principal component analysis for spatial point processes - assessing the appropriateness of the approach in an ecological context. *Lecture Notes in Statistics*, 185:135–150, 2006.

F. O'Sullivan. Nonparametric estimation in the Cox model. *The Annals of Statistics*, 21: 124–145, 1993.

B. Zhu, P.X.-K. Song, and J.M.G. Taylor. Stochastic functional data analysis: a diffusion model-based approach. *Biometrics*, 67:1295–1304, 2011.