Sélection d'histogrammes dyadiques basés sur des données éventuellement censurées

Nathalie Akakpo, Cécile Durot

Sélection d'histogrammes dyadiques basée sur des données éventuellement censurées

Nathalie Akakpo, Cécile Durot

Université Paris-Sud XI, Laboratoire de mathématiques
Bâtiment 425, 91405 Orsay Cedex (France)
e-mail : nathalie.akakpo@math.u-psud.fr,cécile.durot@math.u-psud.fr

Résumé

Nous étudions une procédure d'estimation basée sur des données éventuellement censurées. Le cadre général que nous considérons permet par exemple de traiter l'estimation de la densité et du taux de survie lorsque les données sont aléatoirement censurées à droite, ainsi que la régression. Nous définissons tout d'abord une collection d'histogrammes construits sur des partitions en intervalles dyadiques, choix inspiré d'un résultat d'approximation dû à DeVore et Yu. Notre procédure consiste à sélectionner le meilleur histogramme parmi cette collection à partir des données, en minimisant un critère de type moindres carrés pénalisé. Notre estimateur vérifie une inégalité de type oracle non-asymptotique ainsi que des propriétés d'adaptativité au sens minimax sur diverses classes de régularité, et notamment sur des classes de régularité inhomogène. En outre, sa complexité algorithmique est seulement linéaire en la taille de l'échantillon.

Mots-clés : *adaptativité, données censurées, sélection de modèles.*

Abstract

We study a procedure dedicated to functional estimation problems based on data that may be censored. The general framework we consider allows for instance to handle density and hazard rate estimation based on randomly right-censored data, or regression. We first define a collection of histograms built on partitions into dyadic intervals, a choice inspired from an approximation result due to DeVore and Yu. Our estimation procedure then consists in selecting the best histogram among that collection from the data, by minimizing a penalized least-squares type criterion. Our estimator satisfies a nonasymptotic oracle-type inequality and enjoys adaptativity properties in the minimax sense over a wide range of smoothness classes, that contain functions of inhomogeneous smoothness. Besides, its computational complexity is only linear in the size of the sample.

Keywords: *adaptivity, censored data, model selection.*

1 Introduction

En analyse des données de survie, de nombreux problèmes d'estimation fonctionnelle sont basés sur des données qui ne peuvent être que partiellement observées. Par exemple, dans une étude clinique consacrée aux temps de survie après une opération, certains patients peuvent survivre au-delà de la fin de l'étude ou décéder d'une cause sans rapport avec l'opération subie. Les temps de survie sont alors souvent considérés comme des variables i.i.d. aléatoirement censurées à droite. Nous présentons une procedure non-paramétrique adaptative permettant d'estimer diverses fonctions, telles que la densité ou le taux de survie, lorsqu'on ne dispose que de ce type de données.

2 Cadre et procédure d'estimation

Nous nous plaçons dans le cadre suivant. Notre but est d'estimer une fonction s à valeurs réelles sur la base d'observations indépendantes $Z_1 = (X_1, Y_1)$, ..., $Z_n = (X_n, Y_n)$, $n \geq 3$, où $X_i \in [0,1]$ peut être considéré comme un instant d'observation et Y_i est à valeurs dans un espace $\mathcal{Y}$ donné. On désigne par $\mathbb{P}_s$ la mesure de probabilité sous-jacente et par $\mathbb{E}_s$ l'espérance associée, et on note $\mathbb{L}_2([0,1])$ l'espace des fonctions de carré intégrable par rapport à la mesure de Lebesgue, muni du produit scalaire et de la norme usuels notés $\langle\,.\,,\,.\,\rangle$ et $\|\,.\,\|$. On suppose que s appartient à un sous-espace donné $\mathcal{S}$ de $\mathbb{L}_2([0,1])$ et que, pour tout $t \in \mathcal{S}$,

$$\langle t, s\rangle = \mathbb{E}_s\left[\frac{1}{n}\sum_{i=1}^{n} w(Z_i)t(X_i)\right]$$

où $w : [0,1] \times \mathcal{Y} \to \mathbb{R}$ est une fonction mesurable donnée qui dépend éventuellement de paramètres inconnus (par exemple, d'un paramètre de nuisance dû à la censure). Alors, s minimise sur $\mathcal{S}$

$$t \mapsto \mathbb{E}_s\left[\|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} w(Z_i)t(X_i)\right] = \|s-t\|^2 - \|s\|^2,$$

donc nous considérons des estimateurs définis par minimisation du critère

$$\gamma(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{w}(Z_i)t(X_i)$$

sur des espaces donnés, pour un estimateur donné $\hat{w}$ de w. Ce contraste généralise celui introduit par Brunel et Comte [5]. Les espaces sur lesquels nous minimisons γ sont choisis de manière à ce que les estimateurs qui en résultent admettent une expression de type histogramme. Plus précisément, pour une partition m de $[0,1]$, nous notons S_m l'ensemble

des fonctions à valeurs réelles constantes par morceaux sur m et définissons

$$\begin{aligned}\hat{s}_m &= \operatorname*{argmin}_{t \in S_m} \gamma(t) \\ &= \sum_{I \in m} \left(\frac{1}{n|I|} \sum_{i=1}^{n} \hat{w}(Z_i) \mathbb{1}_I(X_i) \right) \mathbb{1}_I,\end{aligned}$$

où $|I|$ désigne la mesure de Lebesgue d'un intervalle I. Puis, nous considérons la famille des $\{\hat{s}_m\}_{m \in \mathcal{M}^\star}$, où $\mathcal{M}^\star$ est composée uniquement de partitions de $[0,1]$ en intervalles dyadiques. Aussi, chaque estimateur $\hat{s}_m$, $m \in \mathcal{M}^\star$, est appelé histogramme dyadique. Il s'agit alors de choisir la meilleure partition dyadique à partir des données. Pour cela, nous considérons la procédure

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}^\star} \big\{ \gamma(\hat{s}_m) + \mathrm{pen}(m) \big\},$$

où $\mathrm{pen} : \mathcal{M}^\star \to \mathbb{R}^+$ est de la forme

$$\mathrm{pen}(m) = \frac{K \dim(S_m)}{n},$$

et nous définissons l'estimateur pénalisé

$$\tilde{s} = \hat{s}_{\hat{m}}.$$

3 Exemples

Le cadre général précédemment décrit présente l'intérêt de couvrir de nombreux problèmes d'estimation. Nous traiterons en particulier des exemples suivants.

Régression. On observe un vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ dont les coordonnnées sont indépendantes et on veut estimer la moyenne $r = \big(r_1, \ldots, r_n\big)$ de $\boldsymbol{Y}$.

Estimation de densité avec données non censurées. On observe des variables aléatoires i.i.d. $X_1, \ldots, X_n$, à valeurs dans $[0,1]$, admettant une densité $s \in \mathbb{L}_2([0,1])$ à estimer.

Estimation de densité avec données censurées. Soient $T_1, \ldots, T_n$ des variables i.i.d. de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^+$, $C_1, \ldots, C_n$ des variables i.i.d. positives et indépendantes des T_i. Les T_i représentent par exemple des temps de survie et les C_i des instants de censures. Ici, on n'observe que

$$X_i = \min(T_i, C_i) \text{ and } Y_i = \mathbb{1}_{T_i \leq C_i}, i = 1, \ldots, n, \tag{3.1}$$

et on veut estimer la restriction s de f à $[0,1]$, en supposant que $s \in \mathbb{L}_2([0,1])$.

Estimation du taux de survie avec données censurées. On n'observe toujours que les variables définies en (3.1), avec les mêmes hypothèses sur les T_i et les C_i. Le but est d'estimer la restriction s à $[0,1]$ du taux de survie

$$\lambda = \frac{f}{1-F},$$

où F est la fonction de répartition de T_1, en supposant que $s \in \mathbb{L}_2([0,1])$ et $F(1) < 1$.

4 Performances de la procédure

Etant donnée la forme de la pénalité pen, déterminer $\hat{m}$ revient en fait à résoudre un problème de plus court chemin. De plus, pourvu que la longueur des intervalles d'une partition de $\mathcal{M}^\star$ soit au moins de l'ordre de $1/n$, ce problème peut être résolu avec une complexité algorithmique en $\mathcal{O}(n)$, en utilisant un algorithme similaire à celui décrit dans [1].

D'un point de vue théorique, nous souhaiterions que l'estimateur pénalisé soit presque aussi bon en terme de risque que le meilleur histogramme dyadique de la collection $\{\hat{s}_m\}_{m\in\mathcal{M}^\star}$. En adoptant la même démarche que Birgé et Massart [4], nous obtenons en fait dans le cadre défini en 2, sous des hypothèses assez générales, une inégalité de type oracle non-asymptotique de la forme

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\big] \leq C \inf_{m\in\mathcal{M}^\star}\left\{\|s-s_m\|^2 + \frac{\dim(S_m)}{n}\right\}, \tag{4.2}$$

où C dépend notamment de K, w et s. Les principales hypothèses portent sur

- la longueur minimale 2^{-N} des intervalles d'une partition $m \in \mathcal{M}^\star$;
- les moments exponentiels des variables

$$\sum_{i=1}^{n}\Big(w(Z_i)\mathbb{1}_I(X_i) - \mathbb{E}_s\big[w(Z_i)\mathbb{1}_I(X_i)\big]\Big)$$

  pour les intervalles I de longueur 2^{-N} ;
- la qualité de $\hat{w}$ en tant qu'estimateur de w.

En outre, nous considérons la possibilité d'estimer le réel K apparaissant dans la pénalité tout en conservant une inégalité telle que (4.2).

Grâce à des résultats de théorie de l'approximation prouvés dans DeVore et Yu [6] et Birgé [3], nous démontrons que $\tilde{s}$ atteint la vitesse d'estimation minimax à un facteur près indépendant de n pour diverses classes de fonctions, à variations bornées ou présentant une régularité de type Besov. Les classes de fonctions considérées contiennent notamment des fonctions de régularité inhomogène, c'est-à-dire dont la régularité est mesurée via une norme $\mathbb{L}_p$ avec $p < 2$. A notre connaissance, seuls Baraud et Birgé [2] et Li [7] obtiennent des résultats comparables en considérant des données éventuellement censurées.

Nous illustrerons les performances de l'estimateur $\tilde{s}$ sur les exemples précédemment cités, en explicitant le réel K (éventuellement aléatoire) choisi.

Références

[1] Akakpo N. (2008). Estimating a discrete distribution via histogram selection. En révision pour *ESAIM : P&S*, disponible sur http ://www.math.u-psud.fr/ akakpo/.

[2] Baraud Y., Birgé L. (2008). Estimating the intensity of a random measure by histogram type estimators. To appear in *Probability Theory and Related Fields*.

[3] Birgé L. (2007). Model selection for Poisson processes, in *Asymptotics : Particles, Processes and Inverse Problems*, IMS Lecture Notes Monograph Series, **55**, Institute of Mathematical Statistics.

[4] Birgé L., Massart P. (1997) From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*, Springer New York.

[5] Brunel E., Comte F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhya : The Indian Journal of Statistics*, **67**, 3, 441–475.

[6] DeVore R.A., Yu X.M. (1990). Degree of adaptive appoximation. *Mathematics of computation*, **55**, 192, 625–635.

[7] Li L. (2008). On the block thresholding wavelet estimators with censored data. *Journal of Multivariate Analysis*, **99**, 1518–1543.