Modèle de troncature gauche : Comparaison par simulation sur données indépendantes et dépendantes

Zohra Guessoum [1] & Farida Hamrani [2]

[1] *Lab. MSTD, Faculté de mathématique, USTHB, BP n°32, El Alia, Alger, Algérie,zguessoum@usthb.dz*

[2] *Lab. MSTD, Faculté de mathématique, USTHB, BP n°32, El Alia, Alger, Algérie, fhamrani@usthb.dz*

Résumé. Nous comparons dans ce travail de simulation les performances de l'estimateur à noyau de la fonction de régression pour un modèle tronqué à gauche (RLT), lorsque les données sont indépendantes, α-melangeantes puis associées. Nous rappelons les résultats théoriques obtenus dans les deux premiers cas et nous présentons nos résultats dans le dernier cas c-à-d associé .

Mots-clés. Alpha-mélange, Association, Estimateur à noyau, régression non paramétrique, Troncature aléatoire gauche.

Abstract. Our interest in this work of simulation is to compare the performance of the kernel estimator of the regression function in the random left truncated (RLT) model, when the data are independent, α-mixing and associated. We recall some results for the first and second case and we give our result in the associated case.

Keywords. : Alpha-mixing, Association,Kernel estimator, Nonparametric regression, Random left-truncation model.

1 Introduction

Soit $\{(X_i, Y_i); i = 1, \cdots, N\}$ une suite strictement stationnaire de vecteurs aléatoires définie dans le même espace de probabilté $(\Omega, F, \mathbb{P})$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$, ayant la même loi que (X, Y). Si X et Y ne sont pas independants et X est observable, il est raisonnable de prédire Y à partir des informations apportées par X. Cette relation dite de régression est modélisée par $Y_i = m(X_i) + \epsilon_i; i = 1, \cdots, N$, où $m(.)$ désigne la fonction de régression et $\{\epsilon_i; i = 1, \cdots, N\}$ est une suite d'erreurs indépendante de $\{X_i; i = 1, \cdots, N\}$. Il est connu que le meilleur prédicteur conditionnel de Y sachant X (au sens des moindres carrés) est donné par $m(x) = \mathbb{E}[Y/X = x], x \in \mathbb{R}^d$, qui peut être écrit sous la forme suivante

$$m(x) = \frac{\psi(x)}{v(x)}$$

avec $\psi(x) = \int_{\mathbb{R}} yf(x,y)dy$, $f(.,.)$ est la densié conjointe de (X,Y) et $v(.)$ est la densité marginale de X.

Dans certains échantillons de survie, il arrive que la variable d'intérêt Y ne soit pas complétement observable. Nous nous intéressons ici au modèle de troncature à gauche qui, à l'origine, est apparu en astronomie et ensuite a été étendu à plusieurs domaines. Dans ce modèle, nous n'observons Y_i que si $Y_i \geq T_i$, où $\{T_i, i = 1, \cdots, N\}$ désigne une suite de variables de troncature de même loi que T. Il est clair que nous disposons alors d'un échantillon observé $\{(X_i, Y_i, T_i); i = 1, \cdots, n\}$ de taille $n \leq N$.
En pratique, Le fait de supposer que les données sont toujours indépendantes est peu réaliste, c'est pour cela que depuis quelques années plusieurs auteurs ont concentré leurs études sur des données présentant une certaine forme de dépendance.
Dans ce travail de simulation, nous nous intéressons à l'estimateur à noyau de la fonction de régression pour un modèle tronqué à gauche. Nous rappelons ainsi les résultats existant sur sa convergence dans le cas des données indépendantes (E. Ould Said and M. Lemdani 2006) , $\alpha-$mélangeantes (H. Y. Liang,D. L. Li and Y. C. Qi 2009) et nous présentons nos résultats dans le cas associé.
Une famille finie de variables aléatoires $Y = (Y_1..., Y_N)$ est dite associée si

$$cov(f(Y), g(Y)) \geq 0$$

pour toutes fonctions f et g non décroissante de $\mathbb{R}^N$ dans $\mathbb{R}$ pour lesquelles cette covariance existe. Une famille infinie est dite associée si toute sous famille finie est associée.
Nous effectuons des simulations de l'estimateur dans les differents modèles afin de vérifier sa qualité et de confronter les résultats pratiques à ceux attendus par la théorie.

2 Définition de l' estimateur et présentations des résultats

E. Ould Said and M. Lemdani (2006) ont défini un estimateur de la fonction régression pour un modèle tronqué à gauche par

$$\hat{m}_n(x) = \frac{\hat{\psi}_n(x)}{\hat{v}_n(x)}$$

avec $\hat{\psi}_n(x) = \frac{\alpha_n}{nh_n^d} \sum_{i=1}^{n} \frac{Y_i}{G_n(Y_i)} K_d\left(\frac{x - X_i}{h_n}\right)$ et $\hat{v}_n(x) = \frac{\alpha_n}{nh_n^d} \sum_{i=1}^{n} \frac{1}{G_n(Y_i)} K_d\left(\frac{x - X_i}{h_n}\right)$

- $K_d : \mathbb{R}^d \to \mathbb{R}$ est la fonction à noyau,
- h_n est appelée fenêtre qui tend vers 0 quand $n \to \infty$,
- G_n est l'estimateur de Lynden-bell (1971) de la F.r G de la v.a de troncature T,
- α_n est l'estimateur de $\alpha := \mathbb{P}(Y \geq T)$ proposé par S.He and G. Yang,(1998).

Dans le cas independant et $d = 1$, E. Ould Said and M. Lemdani (2006) ont établi sa convergence uniforme ainsi que sa normalité asymptotique.
Quand les données sont α-melangeantes, H. Y. Liang,D. L. Li and Y. C. Qi (2009) ont donné le taux de convergence uniforme de cet estimateur et H. Y. Liang (2011) a étudié son comportement asymptotique.
Dans le cas où les $(X_i, Y_i)_{i=1,\cdots,N}$ sont associés, nous avons établi la convergence uniforme sur un compact D du même estimateur sous certaines conditions sur la fenêtre, le noyau et des conditions de régularité sur les densités conjointes et marginales. Notre résultat est énoncé dans le théorème suivant dans lequel on supposera aussi une décroissance exponentielle du coefficient de covariance liant la suite $(X_i, Y_i)_{i=1,\cdots,n}$.

Théorème :

$$\sup_{x\in D} |\hat{m}_n(x) - m_n(x)| = O\left\{ \sqrt{\frac{\log n}{n h_n^d}} \vee \left(\frac{\log\log n}{n}\right)^{\theta} \vee h_n \right\} \mathbf{P} \text{ - p.s quand } n \to \infty$$

avec $0 < \theta < \frac{2\gamma}{4\gamma+18+3\kappa}$, κ, γ sont des réels positifs.

Si nous comparons notre vitesse de convergence avec celles obtenues dans les cas indépendant et α-mélangeant, nous voyons apparaître un terme dépendant de θ, qui est dû à l'effet de l'association. Notons que notre vitesse est légérement meilleure que celle obtenue dans le cas indépendant par E. Ould Said and M. Lemdani (2006) grâce à une condition de symétrie sur le noyau, et qu'elle est similaire à celle obtenue par H. Y. Liang,D. L. Li and Y. C. Qi (2009) dans le cas alpha-mélangeant.

3 Simulation

Dans cette partie, nous réalisons des simulations pour étudier la performance de l'estimateur $\hat{m}_n(x)$ de $m(x)$ dans le cas où $d = 1$ sur des échantillons de taille finie. L'estimateur $\hat{m}_n(x)$ dépend du choix du noyau et de la fenêtre h_n. Le choix du noyau influe peu sur la performance de cet estimateur ce qui nous a conduit à privilégier la solution classique du noyau gaussien dans tous les modèles. Le choix de la fenêtre h_n est, en revanche, crucial. Nous choisissons dans les différentes cas les fenêtres optimales suivantes :

- cas indépendant : $h_n = Cn^{-1/5}$ (obtenue par B. W. Silverman (1986) pour les données complètes),
- cas α-mélange et associé : $h_n = C\left(\frac{logn}{n}\right)^{1/3}$(obtenue par E. Liebscher (2002) pour les données complètes α-mélangeantes).

avec C une constante à adapter pour chaque modèle.

On simule dans un premier temps, N valeurs $(X_i, Y_i, T_i)_{i=1,\cdots,N}$ du triplet de variables aléatoires (X, Y, T) avec T indépendante de (X, Y), et $T \sim exp(\lambda)$, λ est adapté de manière à obtenir les différentes valeurs de α . Ensuite on retient les observations $(X_i, Y_i, T_i)_{i=1,\cdots,n}$ vérifiant $Y_i \geq T_i$.
Dans les graphes suivants on remarque que la qualité de l'estimateur est affecté beaucoup plus par n la taille de l'échantillon que par le taux de troncature α.

Cas indépendant : Modèle $\begin{cases} X_i \sim N(0,1), \\ Y_i = 2X_i + 1 + \epsilon_i, \epsilon_i \sim N(0.2, 1). \end{cases}$

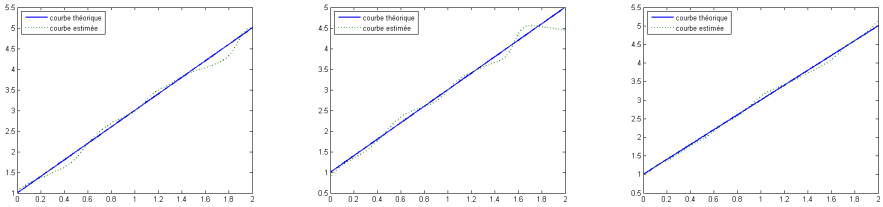

FIG. 1 – $m(x) = 2x + 1$ avec $\alpha \approx 70\%$ et $n = 50, 100, 300$, respectivement.

Cas $\alpha-$melange : Modèle $\begin{cases} X_i = 0.1X_{i-1} + \epsilon_{1i}, \epsilon_{1i} \sim N(0,1), \\ Y_i = 2X_i + 1 + \epsilon_{2i}, \epsilon_{2i} \sim N(0.2, 1). \end{cases}$

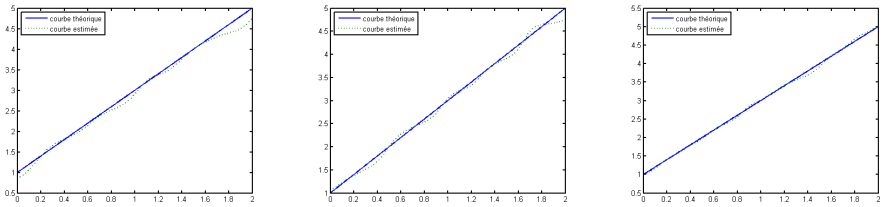

FIG. 2 – $m(x) = 2x + 1$ avec $\alpha \approx 70\%$ et $n = 50, 100, 300$, respectivement.

Cas associé : Modèle $\begin{cases} X_i = \exp(W_{i-1}/2)\exp(W_{i-2}/2), W_i \text{ sont N+1 v.a iid} \sim N(0,1), \\ Y_i = 2X_i + 1 + \epsilon_i, \epsilon_i \sim N(0.2, 1). \end{cases}$

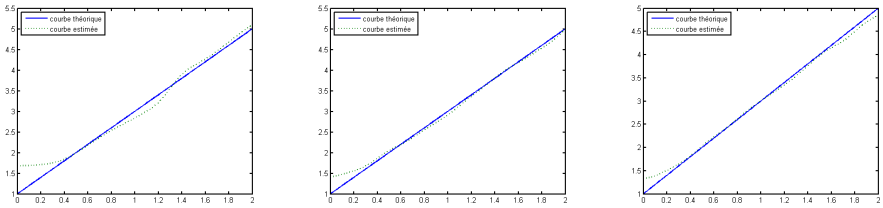

FIG. 3 – $m(x) = 2x + 1$ avec $\alpha \approx 70\%$ et $n = 50, 100, 300$, respectivement.

Bibliographie

[1] He,S., Yang,G.(1998). Estimation of the truncation probability in the random truncation model. The Annals of Statistics. Vol 26 : 1011-1027.
[2] Liang,H.Y.,Li,D. L.,Qi, Y.C. (2009).Strong convergence in nonparametric regression with truncated dependent data. J.Multivariate Anal. Vol 100 :162-174.
[3] Liang,H.Y.(2011). Asymptotic normality for regression function estimate under truncation and α-mixing conditions. C. Statistics-Theory and Methods. Vol 40 :1999-2021.
[4] Liebscher,E. (2002). Kernel density and hazard rate estimation for censored data under α-mixing condition. Ann.Inst.Statist.Math. Vol 34 :19-28.
[5] Lynden-Bell, D., (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. Monthly Notices Royal Astronomy Society Vol 155 : 95-118.
[6] Ould Saïd,E. Lemdani,M. (2006). Asymptotic properties of a nonparametric regression function estimator with randomly truncated data.Ann.Inst.Statist.Math. Vol 58 :357-378.