Statistique et analyse des données

G. Collomb
S. Hassani
P. Sarda
Ph. Vieu

Estimation non paramétrique de la fonction de hasard pour des observations dépendantes

Statistique et analyse des données, tome 10, n° 3 (1985), p. 42-49
<http://www.numdam.org/item?id=SAD_1985__10_3_42_0>

Statistique et Analyse des Données
1985 - vol.10 n°3 pp 42-49

ESTIMATION NON PARAMETRIQUE DE LA FONCTION DE HASARD POUR DES OBSERVATIONS DEPENDANTES.

COLLOMB G., HASSANI S., SARDA P., VIEU Ph.

Laboratoire de Statistique et Probabilités
U.A.-C.N.R.S. 745 - Université Paul Sabatier
118, route de Narbonne - 31062 Toulouse Cedex.

Résumé : *Pour une variable aléatoire X à valeurs dans $\mathbb{R}^p$ dont la loi de probabilité dans $\mathbb{R}^p$ admet F pour fonction de répartition et f pour densité, on estime la fonction de hasard h = f/(1-F), à partir de variables aléatoires X_i, i=1,...,n , ayant même loi que X, lorsque le processus $(X_n)_{n\in\mathbb{N}^*}$ est uniformément fortement mélangeant (un tel modèle inclut le cas des "bons" processus markoviens utilisables en théorie de la fiabilité). Nous montrons la convergence uniforme presque complète sur un compact de $\mathbb{R}^p$ de deux estimateurs non paramétriques de h construits directement à partir d'estimateurs de f et F. Au cours des démonstrations nous établissons des résultats analogues concernant l'estimation de f et de F. Enfin nous donnons une application à l'estimation du mode de h.*

Summary : *Let X be a random variable which is valued in $\mathbb{R}^p$, and assume that the distribution of X in $\mathbb{R}^p$ have F for cumulative distribution and f for density. We want to estimate the hazard function defined by h = f/(1-F) from random variables X_i, i=1,...,n , distributed like X, when $(X_n)_{n\in\mathbb{N}^*}$ is ϕ-mixing (such a model includes the cases of "good" Markovian processes which can be used in reliability theory). We introduce two nonparametric estimators of h defined from estimators of f and F.*

We show that both estimators are uniformly completely consistent on a compact set of $\mathbb{R}^p$. We give also an application to the estimation of the mode of h.

Mots clés : *Fonction de hasard, Taux de défaillance, Estimation non paramétrique, Processus ϕ-mélangeant, estimateur à noyau, Estimateur des k points les plus proches, Theorie de la fiabilité.*

Manuscrit reçu le 9 mai 1985
révisé le 28 janvier 1986

1 - INTRODUCTION

Soit X une variable aléatoire à valeurs dans une partie $\mathbb{E}$ non vide et mesurable de $\mathbb{R}^p$ ($p \geq 1$) dont la loi de probabilité dans $\mathbb{R}^p$ admet F (resp. f) pour fonction de répartition (resp. densité). On suppose que f est uniformément continue sur $\mathbb{E}$. La fonction de hasard (appelée aussi taux de hasard, hasard, taux de défaillance, taux de mortalité, fonction d'intensité ... selon les applications envisagées) de X est la fonction définie par

$$h(x) = f(x)/(1-F(x)) \ , \quad \forall\, x \in \mathbb{E} \ , \quad F(x) < 1. \tag{1.1}$$

Le cas où $\mathbb{E} = \mathbb{R}^+$ constitue un important champ d'applications.

Soit $X_1,\ldots,X_n$ une suite de vecteurs aléatoires ayant chacun même loi que X. On suppose que le processus $(X_n)_{\mathbb{N}^*}$ est uniformément fortement mélangeant (ou ϕ-mixing) en ce sens que (Billingsley 1968 p.166) pour tous entiers positifs i et j et pour tout événement A (resp. B) appartenant à la tribu engendrée par $(X_1,\ldots,X_i)$ (resp. par $(X_{i+j}\ ,\ X_{i+j+1},\ldots)$) on a

$$|P(A \cap B) - P(A)\ P(B)| \leq \phi_j\ P(A) \ , \tag{1.2}$$

où $(\phi_n)_{\mathbb{N}^*}$ est une suite réelle positive de limite nulle : une telle condition est satisfaite par $(X_n)_{\mathbb{N}^*}$ dès que ce processus est un "bon" processus markovien - voir remarque 1.1.

Avec la convention c/o = o pour tout réel c, on définit les deux estimateurs non paramétriques de h suivants

$$h_n(x) = f_n(x)/(1-F_n(x)) \ , \quad \forall\, x \in \mathbb{E} \ , \tag{1.3}$$

et

$$\tilde{h}_n(x) = \tilde{f}_n(x)/(1-F_n(x)) \ , \quad \forall\, x \in \mathbb{E} \ , \tag{1.4}$$

où

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \ , \quad \forall\, x \in \mathbb{E} \ ,$$

et f_n (resp. $\tilde{f}_n$) est l'estimateur à noyau (resp. l'estimateur des k points les plus proches) de f. Ces deux estimateurs, proposés respectivement par Rosenblatt (1956) et Loftsgaarden Quesenberry (1965), sont définis par

$$f_n(x) = f_n^K(x) = (n\, b_n^p)^{-1} \sum_{i=1}^{n} K((x-X_i)/b_n), \quad \forall\, x \in \mathbb{E} \tag{1.5}$$

où $(b_n)_{\mathbb{N}^*}$est une suite réelle strictement positive de limite nulle, et

$$\tilde{f}_n(x) = \tilde{f}_n^K(x) = (n\, R(k_n,x)^p)^{-1} \sum_{i=1}^{n} K((x-X_i)/R(k_n,x)),\ \forall\, x \in \mathbb{E}, \quad (1.6)$$

où $R(k_n,x)$ est la distance entre x et la $k_n^{ième}$ observation la plus proche de x :

$$R(k_n,x) = \inf \{\lambda \in \mathbb{R}^+ : \text{card}\, \{X_i,\ i=1,\ldots,n, |x-X_i| \leq \lambda\} \geq k_n\},$$

la suite $(k_n)_{\mathbb{N}^*}$ étant entière, strictement positive et vérifiant

$$k_n/n \xrightarrow[n\to\infty]{} 0 \ ,$$

la fonction K étant un noyau de $\mathbb{R}^p$, c'est-à-dire une fonction réelle bornée de $L^2(\mathbb{R}^p)$ telle que

$$|z|^p \ K(z) \xrightarrow[|z|\to\infty]{} 0 \ .$$

Nous dirons qu'un noyau K est unitaire lorsque l'on a

$$\int K(z)dz = 1 \ . \quad (1.7)$$

On désigne par C un compact de $\mathbb{E}$ et par $\tilde{C}$ un ε-voisinage compact de C dans $\mathbb{E}$, avec $\tilde{C} = \mathbb{E}$ si $C = \mathbb{E}$. On suppose que f vérifie

$$\exists\, \Gamma < +\infty\ ,\ f(x) \leq \Gamma\ ,\ \forall\, x \in \mathbb{E}, \quad (1.8)$$

$$\exists\, \gamma > 0\ ,\ f(x) \geq \gamma\ ,\ \forall\, x \in \tilde{C}, \quad (1.9)$$

et

$$\exists\, \tau > 0\ ,\ F(x) \leq 1-\tau,\ \forall\, x \in C. \quad (1.10)$$

Nous aurons besoin parfois des hypothèses suivantes sur le noyau K

$$\exists\, M < +\infty,\ \exists\, a > 0\ ,\ \forall\, (z,z') \in \mathbb{R}^{2p}\ \ |K(z)-K(z')| \leq M|z-z'|^a\ , \quad (1.11)$$

$$K(cz) \geq K(z)\ ,\ \forall\, c \in [0,1]\ ,\ \forall\ z \in \mathbb{R}^p \quad (1.12)$$

$$K(z) = 0\ ,\ \forall\, z \in \mathbb{R}^p\ ,\ |z| > 1. \quad (1.13)$$

Nous serons amenés à considérer des noyaux particuliers K_B définis par

$$K_B(z) = \mathbb{1}_B(z)\ ,\ \forall\, z \in \mathbb{R}^p \quad (1.14)$$

où B est un compact de $\mathbb{R}^p$, contenant l'origine et dont la mesure de Lebesgue (ou volume) vaut 1, en remarquant alors que, pour un tel noyau, l'hypothèse (1.12) signifie que B est étoilé par rapport à l'origine, en ce sens que

$$\forall\, z \in B,\ \forall\, c \in [0,1]\ ,\ cz \in B.$$

Les noyaux habituellement utilisés dans les applications sont du type (1.11) ou (1.14) : les résultats que nous établissons concernent ces deux catégories de noyaux.

Enfin nous désignons par $(m_n)_{\mathbb{N}^*}$une suite entière qui vérifie conjointement avec la suite $(\phi_n)_{\mathbb{N}^*}$définie en (1.2) l'hypothèse

$$\exists\ A < \infty\ ,\ \forall\ n \in \mathbb{N}^*,\quad n\phi_{m_n}/m_n \leq A\quad ,\ 1 \leq m_n \leq n. \tag{1.15}$$

Remarque 1.1.

On remarque qu'on peut choisir comme suite $(m_n)_{\mathbb{N}^*}$

$$m_n = c[\text{Log } n] + 1\quad ,\quad \forall\ n \in \mathbb{N}^*\ ,$$

dès que le processus $(X_n)_{\mathbb{N}^*}$est markovien et satisfait la condition de Doeblin (Doob 1953, p.209) , le processus $(X_n)_{\mathbb{N}^*}$étant alors géométriquement ϕ-mélangeant en ce sens que

$$\exists\ s < +\infty\ ,\ \exists\ r \in [0,1[\ ,\quad \phi_n \leq s\ r^n\ ,\quad \forall\ n \in \mathbb{N}^*.$$

D'autres exemples de processus stationnaires géométriquement ϕ-mélangeants sont donnés dans Collomb (1985).

De tels modèles de dépendance nous semblent pouvoir convenir dans de nombreux problèmes pratiques où la condition d'indépendance est difficilement acceptable.

2 - LES RESULTATS.

Ces résultats concernent les estimateurs h_n et $\tilde{h}_n$ de la fonction de hasard définis par (1.3)-(1.6) à partir d'un noyau unitaire supposé lipschitzien ou uniforme.

Proposition 1

Lorsque le noyau vérifie (1.7) et soit (1.11), soit (1.14) et (1.12), et lorsque la suite $(b_n)_{\mathbb{N}^*}$vérifie, conjointement avec une suite $(m_n)_{\mathbb{N}^*}$ satisfaisant (1.15), la condition

$$nb_n^p\ /\ (m_n\ \text{Log}\ n) \xrightarrow[n\to\infty]{} \infty\quad , \tag{2.1}$$

alors on a

$$\sup_{x\in C} |h_n(x) - h(x)| \xrightarrow[n\to\infty]{p.co.} 0\ . \tag{2.2}$$

Proposition 2

Lorsque le noyau vérifie (1.7), (1.12), (1.13) et soit (1.11), soit (1.14), et lorsque la suite $(k_n)_{\mathbb{N}^*}$ vérifie, conjointement avec une suite $(m_n)_{\mathbb{N}^*}$ satisfaisant (1.15), la condition

$$k_n/(m_n \text{ Log } n) \xrightarrow[n\to\infty]{} \infty, \tag{2.3}$$

alors on a

$$\sup_{x\in C} |\tilde{h}_n(x) - h(x)| \xrightarrow[n\to\infty]{p.co.} 0 . \tag{2.4}$$

Remarque 2.1.

Pour les processus considérés dans la remarque (1.1) les conditions (2.1) et (2.3) s'écrivent respectivement

$$nb_n^p / (\text{Log } n)^2 \xrightarrow[n\to\infty]{} \infty \quad \text{et} \quad k_n / (\text{Log } n)^2 \xrightarrow[n\to\infty]{} \infty .$$

Remarque 2.2.

Les résultats (2.2) et (2.4) restent valables lorsque le processus $(X_i)_{i\in\mathbb{N}^*}$ est m-dépendant (et donc a fortiori lorsque les v.a. X_i, i=1,2,..., sont indépendantes), la condition (2.1) (resp. (2.3)) devenant alors

$$nb_n^p / \text{Log } n \xrightarrow[n\to\infty]{} \infty, \quad (\text{resp. } k_n / (n \text{ Log } n) \xrightarrow[n\to\infty]{} \infty).$$

Remarque 2.3.

Ces résultats dans le cas de variables dépendantes ou indépendantes améliorent ceux déjà obtenus en estimation non paramétrique de la fonction de hasard (voir par exemple Murthy (1965), Blum et Sursala (1980) pour l'estimateur h_n, ainsi que la revue bibliographique de Hassani et al. (1986), en préparation, pour d'autres estimateurs).

Remarque 2.4.

Lorsque la fonction h admet sur C un mode unique θ

$$\theta = \arg \max_{x\in C} h(x) ,$$

il découle que sous les hypothèses de la proposition 1 [resp. de la proposition 2] on a

$$\theta_n \xrightarrow[n\to+\infty]{p.\text{¢o.}} \theta \quad , \quad [\text{ resp. } \tilde{\theta}_n \xrightarrow[n\to+\infty]{p.co.} \theta] ,$$

où

$$\theta_n = \arg \max_{x\in C} h_n(x) \quad \text{et } \tilde{\theta}_n(x) = \arg \max_{x\in C} \tilde{h}_n(x).$$

Nous renvoyons à Prakasa Rao (1983, p.273) pour la technique de démonstration de cette remarque.

3 - DEMONSTRATION DES RESULTATS

– En utilisant les conditions (1.8) et (1.10) on obtient en tout point x de $\mathbb{E}$

$$|h_n(x) - h(x)| \leq \frac{|f_n(x)-f(x)|}{\tau-|F_n(x)-F(x)|} + \Gamma\, \tau^{-1}\, |F_n(x) - F(x)| \quad . \tag{3.1}$$

– L'inégalité de Bernstein généralisée à des v.a. ϕ-mélangeantes (Collomb 1984, p.449) donne immédiatement en tout point x de E

$$|F_n(x) - F(x)| \xrightarrow[n\to+\infty]{p.co.} 0 \; . \tag{3.2}$$

En utilisant la croissance des fonctions F et F_n, la relation d'ordre (non totale) sur $\mathbb{R}^p$ étant définie par

$$(u^j)_{1\leq j\leq p} \leq (v^j)_{1\leq j\leq p} \iff \forall \; j=1,\ldots,p \quad u^j \leq v^j \; ,$$

on peut approcher les fonctions F et F_n par leurs valeurs en un nombre fini de points. Cette remarque et le résultat (3.2) amènent

$$\sup_{x\in C} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{p.co.} 0 \; . \tag{3.3}$$

– Il reste à montrer la convergence presque complète de f_n vers f.
Le lemme 3 de Collomb (1984, p.454) donne sous l'hypothèse (2.1)

$$\sup_{x\in C} |f_n^K(x) - E\, f_n^K(x)| \xrightarrow[n\to+\infty]{p.co} 0 \quad , \tag{3.4}$$

lorsque le noyau K est lipschitzien.
Par ailleurs si K_B est un noyau uniforme, par une application du lemme d'Urysohn on peut définir pour tout $\eta \in]0,1[$ deux noyaux k^η et K^η lipschitziens et tels que

$$\left.\begin{aligned} &k^\eta(z) \leq K_B(z) \leq K^\eta(z) \qquad \forall \; z \in \mathbb{R}^p \; , \\ &\|k^\eta - K_B\| \leq \eta \qquad \text{et} \qquad \|K^\eta - K_B\| \leq \eta \quad . \end{aligned}\right\} \tag{3.5}$$

En utilisant alors la définition (1.5) on obtient pour tout x de $\mathbb{R}^p$

$$f_n^{k^\eta}(x) - Ef_n^{k^\eta}(x) + \alpha_n^{k^\eta}(x) \leq f_n^{K_B}(x) - Ef_n^{K_B}(x)$$
$$\leq f_n^{K^\eta}(x) - Ef_n^{K^\eta}(x) + \alpha_n^{K^\eta}(x)$$

où

$$\alpha_n^K(x) = Ef_n^K(x) - Ef_n^{K_B}(x) \text{ pour } K = k^\eta \text{ ou } K^\eta.$$

En utilisant (1.8) et (3.5) on obtient après intégration

$$\alpha_n^K \leq \Gamma \eta ,$$

ce qui permet d'écrire

$$\sup_{x\in C} |f_n^{K_B}(x) - Ef_n^{K_B}(x)| \leq \Gamma\eta + \max_{K=k^\eta, K^\eta} \sup_{x\in C} |f_n^K(x) - Ef_n^K(x)|.$$

En utilisant (3.4) on obtient alors

$$\sup_{x\in C} |f_n^{K_B}(x) - Ef_n^{K_B}(x)| \xrightarrow[n\to\infty]{p.co.} 0. \tag{3.6}$$

- En utilisant la continuité de f sur $\mathbb{E}$ et la condition (1.7) on obtient par un calcul classique (cf. Prakasa Rao, 1983, p.35)

$$\sup_{x\in C} |Ef_n(x) - f(x)| \xrightarrow[n\to\infty]{} 0 . \tag{3.7}$$

Le résultat de la proposition 1 découle alors de (3.1), (3.3), (3.4), (3,7), [resp. de (3.1), (3.3), (3.6), (3.7)] lorsque le noyau est lipschitzien [resp. uniforme].

- Le lemme qui suit est une conséquence directe d'un résultat de Moore et Yackell (1977, p.145, théorème 1.1) dont la démonstration ne fait pas intervenir l'hypothèse d'indépendance des X_i, $i\in\mathbb{N}^*$.

Lemme

Si K est un noyau vérifiant (1.12) et (1.13), et B la boule de centre 0 et de volume unité de $\mathbb{R}^p$, toute propriété de convergence (ponctuelle ou uniforme, en probabilité, presque sûre ou presque complète) satisfaite par les estimateurs f_n^K et $f_n^{K_B}$ définis par (1.5) pour une suite $(b_n)_{n\in\mathbb{N}^*}$ reste vérifiée pour l'estimateur $\tilde{f}_n^K$ défini par (1.6) à l'aide du même noyau K pour une suite $k_n \underset{n\to\infty}{\sim} nb_n^p$ [en ce sens que "il suffit d'imposer à $(k_n/n)_{n\in\mathbb{N}^*}$ les mêmes conditions qu'à $(b_n^p)_{n\in\mathbb{N}^*}$ pour que $\tilde{f}_n^K$ jouisse des mêmes propriétés que f_n^K "].

On obtient à partir de (3.4) [resp. (3.6)] , (3.7) et de ce lemme

$$\sup_{x\in C} |\tilde{f}_n(x) - f(x)| \xrightarrow[n\to\infty]{p.co.} 0 , \qquad (3.8)$$

pour un noyau lipschitzien [resp. uniforme]. Par une décomposition analogue à (3.1) on obtient le résultat de la proposition 2 en utilisant (3.3) et (3.8).

Remarque 3.1.

Le résultat (3.2) constitue une amélioration du théorème classique de Glivenko-Cantelli établi pour des observations indépendantes.

De même les résultats de convergence uniforme des estimateurs à noyau et des k-points les plus proches de la densité que nous établissons au cours de cette démonstration améliorent ceux déjà obtenus en estimation non paramétrique de la densité (voir par exemple la revue bibliographique de Bean and Tsakas (1980)).

BIBLIOGRAPHIE

BEAN, S.J. and TSAKAS, C.P. (1980)."Developments in nonparametric density estimation". Bull. Math. Stat., 17, 77-84.

BILLINGSLEY, P. (1968). "Convergence of probability measures". Wiley, New-York.

BLUM, J.R. and SURSALA, V. (1980). "Maximal deviation theory of density and failure rate function estimates based on censored data". Multivariate Analysis, 1, 213-222

COLLOMB, G. (1985). "Nonparametric time series analysis and prediction : uniform almost sure convergence of the window and k.NN autoregression estimates". Statistics 16, 2, 297-307.

COLLOMB, G. (1984). "Propriétés de convergence presque complète du prédicteur à noyau". Zeitschrift. f. Wahrs. v.v. Geb., 66, 441-460.

DOOB, J. (1953). "Stochastic processes". Wiley, New-York.

HASSANI, S., SARDA, P. et VIEU, P. (1986). "Approche non paramétrique en théorie de la fiabilité : revue bibliographique". En préparation.

LOFTSGAARDEN, D.O. and QUESENBERRY, C.D. (1965). "A non parametric estimate of a multivariate density function". A.M.S., 36, 1049-1051.

MOORE, D.S. et YACKELL, J.W. (1977). "Consistency properties of nearest neighbour density functions". Ann. Stat. 5, 143-154.

PRAKASA RAO B.L.S. (1983). "Nonparametric functional estimation". Academic Press. New-York.

MURTHY, V.K. (1965). "Estimation of Jumps, reliability and hazard rate". Ann. Stat. 36, 1032-1040.

ROSENBLATT, M. (1956). "Remarks on some nonparametric estimates of a density function". A.M.S. 27, 642-669.