Estimation fonctionnelle dans les modèles de durée : Méthode des fonctions orthogonales

Ouafae Yazourh

Résumé

On définit un estimateur non paramétrique de la densité de variables aléatoires positives X_i^0 soumises à des censures droites. On démontre la convergence de cet estimateur, construit par la méthode des fonctions orthogonales en divers sens stochastiques, notamment au sens du MISE. On en déduit des résultats de convergence relatifs à un estimateur du taux de hasard des variables X_i^0.

Abstract

Let f and h be the common density and hazard rate of right censored life times. We define new non parametric estimators of f and h, based on the orthogonal functions method. We prove their pointwise and L^2 asymptotic consistencies.

1 Introduction

Le but de ce travail est l'étude de la convergence d'estimateurs de la densité et du taux de hasard de variables aléatoires X_i^0 positives soumises à des censures "droites": il s'agit de l'hypothèse classique des modèles de durées. Plus exactement, nous adopterons dans ce travail le modèle d'Efron (1967).

Received by the editors January 1993
Communicated by M. Hallin
AMS Mathematics Subject Classification : 62G05
Keywords : Données censurées, estimateur non paramétrique du taux de hasard, méthode des fonctions orthogonales, estimateur de Kaplan-Meier, modèle à censures aléatoires.

Appelons X_i^0, $i \geq 1$, les variables aléatoires représentant des durées de vie que l'on veut étudier (temps de chômage, de survie, de non défaillance, etc). Elles sont à valeurs dans $\mathbf{R}^+$, *i.i.d.*; on note f leur densité, supposée continue, F leur fonction de répartition. Le statisticien ne peut alors observer que des variables X_i, $i \geq 1$, où $X_i = \inf\{X_i^0, C_i\}$, les C_i symbolisant les censures apportées aux durées de vie d'intérêt (date d'échantillonnage, etc). On suppose que les C_i forment une suite de variables réelles *i.i.d.* à valeurs dans $\mathbf{R}^+$, de densité g, de fonction de répartition G, et indépendantes des X_i^0. On appellera enfin l la densité des X_i, L leur fonction de répartition $(1 - L = (1 - F)(1 - G))$. Comme le statisticien sait néanmoins si l'observation dont il dispose est censurée ou non, l'échantillon est finalement constitué de n couples (X_i, δ_i) où $\delta_i = I_{\{X_i^0 \leq C_i\}}$ (δ_i est l'indicateur de non censure).

L'estimation non paramétrique de f, du taux de hasard h défini par $h = \frac{f}{1-F}$, a été récemment abordée par de nombreux auteurs parmi lesquels on peut citer, Gneyou (1991) et Grégoire (1991) pour des travaux récents, en renvoyant à Huber et Lecoutre (1990), pour une synthèse des résultats obtenus avant (1988).

Dans la quasi-totalité de ces travaux, on utilise la méthode du noyau pour estimer le pramètre fonctionnel d'intérêt. Kimura (1972) et Hjort (1985) ont cependant suggéré d'utiliser la méthode des fonctions orthogonales, de façon alternative, sans développer les résultats de convergence des estimateurs obtenus. Tout récemment Delecroix et Yazourh (1992) ont étudié le comportement asymptotique d'un estimateur h^* de h déduit directement des observations, sans estimation préalable de la densité f, en montrant que h^* construit par la méthode des fonctions orthogonales obtient sous certaines conditions des performences comparables à celles des estimateurs construits par la méthode du noyau.

Dans ce papier on se propose essentiellement d'estimer de la même façon la densité f des v.a. X_i, en démontrant la convergence de l'estimateur f_n obtenu, localement et dans L^2. On déduira ensuite de f_n un estimateur convergent du taux de hasard h.

Dans la construction de cet estimateur, on se restreint à un intervalle [0,b] pour éviter les difficultés posées par le contrôle de l'écart entre $F_n(t)$ et $F(t)$, pour de grandes valeurs de t. Ceci n'implique pas de restriction pratique, puisqu'on peut choisir b supérieur au plus grand des X_i non censurés qu'on observe et même beaucoup plus grand.

L'intérêt de la méthode des fonctions orthogonales semble résider dans le caractère global de l'estimateur construit. Pour peu que le développement de f par rapport à une base orthonormale de $L^2[0, b]$ converge rapidement, en ramenant l'estimation de f à celle des premiers coefficients de ce développement, on obtiendra, comme nous le verrons une approximation globale de f qui reste satisfaisante même dans les zones où sont apparues peu d'observations, puisqu'on estime les coefficients du développement de f par l'ensemble de l'échantillon.

Le premier paragraphe sera consacré à la définition des divers estimateurs utilisés et à l'énoncé des principaux théorèmes de convergence; les démonstrations seront exposées dans le second paragraphe.

2 Notations, hypothèses et résultats.

2.1 Notations et hypothèses.

On suppose une fois pour toutes que la densité f des v.a. X_i est un élément de $L^2[0,b]$. Alors nous appelons (e_i), $i \geq 1$, une base orthonormale de cet espace $L^2[0,b]$. On peut donc écrire au sens de L^2

$$f = \sum_{i=1}^{\infty} a_i e_i$$

avec

$$a_i = \int_0^b e_i(x) f(x) dx = \int_0^b e_i(x) dF(x).$$

Un estimateur de f construit par la méthode des fonctions orthogonales peut donc s'écrire

$$f_n = \sum_{i=1}^{q(n)} \hat{a}_i e_i$$

où

$$\hat{a}_i = \int_0^b e_i(x) dF_n(x)$$

et $(q(n))_{n\geq 1}$ est une suite d'entiers qui tend vers l'infini. Notons F_n le classique estimateur de Kaplan-Meier (1958) de F. On sait que F_n est la fonction de répartition définie par

$$F_n(t) = \begin{cases} \prod_{i/X_{(i)}\leq t} (\frac{n-i}{n-i+1})^{\delta_{(i)}} & \text{si } t \leq X_{(n)} \\ 1 & \text{si } t > X_{(n)} \end{cases}$$

$(X_{(i)})_{1\leq i\leq n}$, est l'échantillon $(X_i)_{1\leq i\leq n}$ ordonné et $\delta_{(i)}$ représente la valeur de δ pou $X_{(i)}$.

L'estimateur du taux de hasard $h = \frac{f}{1-F}$ sera défini par

$$h_n = \frac{f_n}{1 - F_n + \frac{1}{n}}. \tag{1}$$

Pour énoncer les théorèmes de convergence de f_n et h_n, nous utiliserons les quantités

$$H^i(t) = P(X_j > t, \delta_j = i) \tag{2}$$

et

$$H_n^i(t) = \frac{1}{n} \sum_{j=1}^{n} I_{\{X_j > t, \delta_j = i\}}$$

pour $i = 0, 1$.

Pour toute application m de $\mathbf{R}^+$ dans $\mathbf{R}^+$ supposée continûment différentiable on pose

$$N(m, b) = [\mid m(b) \mid + \int_0^b \mid m'(t) \mid dt]. \tag{3}$$

On peut alors définir à partir de la base $(e_i), i \geq 1$,

$$D_n = \sum_{i=1}^{q(n)} N(e_i, b)^2 = \sum_{i=1}^{q(n)} [\mid e_i(b) \mid + \int_0^b \mid e_i'(t) \mid dt]^2 , \tag{4}$$

$$W_n(x, t) = \sum_{i=1}^{q(n)} e_i(x) e_i(t) \tag{5}$$

pour tout x et t réels, et

$$\bar{f}_n(x) = \sum_{i=1}^{q(n)} a_i e_i(x) = \int_0^b W_n(x, t) dF(t). \tag{6}$$

Nous pouvons maintenant énoncer quelques hypothèses nécessaires à la convergence de f_n et h_n.

$H_0)$: $H^i(b) > 0$, $i = 0, 1$, on travaille sur $[0, b]$, avec b strictement inférieur à la plus grande valeur du support de L.

$H_1)$: Les éléments (e_i), $i \geq 1$, de la base choisie sont des fonctions continûment différentiables et uniformément bornées par un nombre M sur l'intervalle $[0, b]$.

$H_2)$: La densité f est bornée sur tout compact $[0, b]$.

H_1 est liée à la base choisie et vérifiée, nous le verrons, dans le cas des fonctions trigonométriques, par exemple. H_2 est une condition usuelle de la régularité pour la densité f. Nous pouvons alors énoncer les théorèmes de convergence.

2.2 Résultats de convergence de l'estimateur de la densité.

Dans tout ce travail on supposera que b vérifie H_0.

Théorème 1 *On considère $(e_i), i \geq 1$, une base de $L^2[0, b]$ telle que H_1 soit vérifiée.*

Si l'on suppose que $q(n) \longrightarrow \infty$, $\lim_{n \longrightarrow \infty} q(n)/n = 0$ et $\lim_{n \longrightarrow \infty} [D_n (\log n)^4 / n^2] = 0$, alors

$$\lim_{n \longrightarrow \infty} \; p.s. \; E(\| f_n - f \|^2_{L^2[0,b]}) = 0.$$

Si $q(n) \longrightarrow \infty$, de telle sorte que pour tous γ_1 et γ_2 strictement positifs les séries de termes généraux $q(n) \exp\{-\delta_1 n / q(n)\}$ et $n\{\sum_{i=1}^{q(n)} \exp[-\gamma_2 n / \sqrt{q(n)} N(e_i, b)]\}$ convergent, alors

$$\lim_{n \longrightarrow \infty} [\| f_n - f \|^2_{L^2[0,b]}] = 0.$$

Remarque. Les convergences recherchées sont obtenues si "grosso modo" les fonctions e_i ne varient pas trop brutalement sur $[0,b]$ et si $q(n)$ croît lentement vers l'infini: c'est un "vieux" principe de la méthode des fonctions orthogonales qui, en pratique, limite la variance de l'estimateur (mais hélas en augmente le biais!).

Un exemple d'application standard est celui de la base des fonctions trigonométriques de $L^2[0,b]$, $(e_i), i \geq 1$, définie par

$e_1(t) = \frac{1}{\sqrt{b}}$

$e_{2k}(t) = \sqrt{\frac{2}{b}} \sin[\frac{2k\pi}{b}(t - \frac{b}{2})], k \geq 1, t \in [0,b]$

$e_{2k+1}(t) = \sqrt{\frac{2}{b}} \cos[\frac{2k\pi}{b}(t - \frac{b}{2})]$, $k \geq 1$ et $t \in [0,b]$.

On obtient ainsi le corollaire suivant.

Corollaire 1 *Si la base $(e_i), i \geq 1$, est celle des fonctions trigonométriques, la convergence vers zéro de $\| f_n - f \|^2_{L^2[0,b]}$ est assurée*

- *en moyenne si $q(n) \longrightarrow \infty$, $q(n)/n \longrightarrow 0$ et $q(n)^3(\log n)^4/n^2 \longrightarrow 0$.*

- *p.s. si $\forall \gamma$, $\gamma > 0$, $\sum_{n=1}^{\infty} nq(n) \exp\{-\gamma n/q(n)^{\frac{3}{2}}\} < \infty$.*

En particulier si on choisit $q(n)$ égale à la partie entière de $n^{2/3}/(\log n)^{4/3+\alpha}$ avec $\alpha > 0$, alors les deux conditions du corollaire sont vérifiées.

On montrera aussi le résultat suivant, concernant la convergence uniforme presque sûre.

Théorème 2 *Si $\sup_{x\in[0,b]} \mid \bar{f}_n(x) - f(x) \mid \longrightarrow 0$, lorsque $n \longrightarrow \infty$, H_1 est vérifiée, et si pour tout γ_1 et γ_2 strictement positifs les séries de termes généraux $n \sum_{i=1}^{q(n)} \exp\{-\gamma_1 n/q(n)N(e_i,b)\}$ et $q(n)\exp\{-\gamma_2 n/q(n)^2\}$ convergent, alors*

$$\sup_{x\in[0,b]} \mid f_n(x) - f(x) \mid \longrightarrow 0 \ \ p.s., \ \ quand \ \ n \longrightarrow \infty.$$

Corollaire 2 *Dans le cas de la base trigonométrique, si $\sup_{x\in[0,b]} \mid \bar{f}_n(x) - f(x) \mid \longrightarrow 0$ quand $n \longrightarrow \infty$, alors, si, $\forall \gamma > 0$, $\sum_{n=1}^{\infty} nq(n) \exp\{-\gamma n/(q(n))^2\} < \infty$, on a*

$$\sup_{x\in[0,b]} \mid f_n(x) - f(x) \mid \longrightarrow 0 \ \ p.s., \ \ quand \ \ n \longrightarrow \infty.$$

Enfin, on cherchera la limite de J_n définie par

$$J_n = \frac{\sqrt{n}(f_n(x) - E(f_n(x)))}{B_n} \tag{7}$$

où x est un élément fixe de $]0,b[$, $B_n^2 = \int_0^b H_n^2(x,s)\frac{h(s)}{1-L(s)}ds$, et

$H_n(x,s) = S(s)W_n(x,s) + \int_s^b S'(t)W_n(x,t)dt$. S est la fonction de survie définie par $S = 1 - F$, qu'on estime par $S_n = 1 - F_n$.

Alors, en posant $\overline{W}_n : t \longrightarrow W_n(x,t)$, on obtient le théorème suivant.

Théorème 3 *Si* $q(n)/\sqrt{n}B_n = o(1)$ *et* $N(\overline{W}_n, b)(\log n)^2/\sqrt{n}B_n = o(1)$, *alors*

$$J_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Ici encore, un exemple d'application nous est fourni par la base trigonométrique.

Corollaire 3 *Si la base* (e_i), $i \geq 1$, *est celle des fonctions trigonométriques et si* $q(n)^3(\log n)^4/n \longrightarrow 0$, *lorsque* $n \longrightarrow \infty$, *alors*

$$J_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

La démonstration de ce résultat se base essentiellement sur le lemme suivant.

Lemme 1 *Si la base* (e_i), $i \geq 1$, *est celle des fonctions trigonométriques, alors il existe* $A > 0$ *tel que*

$$B_n^2 > q(n)A.$$

2.3 Preuve du lemme 1

On suppose que les (e_i), $i \geq 1$, sont les fonctions trigonométriques, et que $q(n)$ est impair. Dans ce cas un calcul classique donne

$$W_n(x,t) = \sum_{i=1}^{q(n)} e_i(x)e_i(t) = \frac{1}{b}\frac{\sin(q(n)\frac{\pi}{b}(x-t))}{\sin(\frac{\pi}{b}(x-t))}.$$

Dans le cas où $x = t$, on prolonge par continuité le quotient à $q(n)$. On a

$$\begin{aligned} B_n^2 &= \int_0^b \{S(s)W_n(x,s) + \int_s^b S'(t)W_n(x,t)dt\}^2 \frac{h(s)}{(1-L(s))}ds \\ &= \int_0^b \{\frac{1}{b}S(s)\frac{\sin(q(n)\frac{\pi}{b}(x-s))}{\sin(\frac{\pi}{b}(x-s))} \\ &\quad + \frac{1}{b}\int_s^b S'(t)\frac{\sin(q(n)\frac{\pi}{b}(x-t))}{\sin(\frac{\pi}{b}(x-t))}dt\}^2 \frac{h(s)}{(1-L(s))}ds. \end{aligned}$$

En effectuant le changement de variable, $u = q(n)\frac{\pi}{b}(x-s)$, on obtient

$$B_n^2 = q(n)\int_{\mathbf{R}} g_n(u)du$$

où

$$\begin{aligned} g_n(u) &= \frac{1}{\pi b} I_{[q(n)\frac{x-b}{b}\pi, q(n)\frac{x}{b}\pi]}(u)\{S(x - \frac{bu}{q(n)\pi})\frac{\sin u}{q(n)\sin(\frac{u}{q(n)})} \\ &\quad + \frac{1}{q(n)}\int_{x-\frac{bu}{q(n)\pi}}^b S'(t)\frac{\sin(q(n)\frac{\pi}{b}(x-t))}{\sin(\frac{\pi}{b}(x-t))}dt\}^2 \frac{h(x-\frac{bu}{q(n)\pi})}{1-L(x-\frac{bu}{q(n)\pi})}. \end{aligned}$$

Si on considère a réel, tel que $0 < a < 1$, on peut minorer $g_n(u)$ par la fonction suivante :

$$\begin{aligned} g_n^1(u) &= \frac{1}{\pi b} I_{[-a\pi,0]}(u)\{S(x - \frac{bu}{q(n)\pi})\frac{\sin u}{q(n)\sin(\frac{u}{q(n)})} \\ &\quad + \frac{1}{q(n)}\int_{x-\frac{bu}{q(n)\pi}}^{b} S'(t)\frac{\sin(q(n)\frac{\pi}{b}(x-t))}{\sin(\frac{\pi}{b}(x-t))}dt\}^2 \frac{h(x-\frac{bu}{q(n)\pi})}{1-L(x-\frac{bu}{q(n)\pi})}. \end{aligned}$$

On va calculer ensuite $\lim_{n\longrightarrow\infty}\int g_n^1$. Or, on peut écrire la majoration suivante

$$g_n^1 \leq \frac{1}{\pi b} I_{[-a\pi,0]}(u)\{|\ \frac{\sin u}{u}\ |\ \frac{a\pi}{\sin a\pi} + C\frac{q(n)}{q(n)}\int_{\mathbf{R}^+} f(t)dt\}^2 \frac{\sup_{x\leq t\leq x+a} h(t)}{(1-L(x+a))}.$$

(C est une constante positive), car si $-a\pi \leq u \leq 0$ alors, $-a\pi \leq \frac{u}{q(n)} \leq 0$ et $\sin(\frac{u}{q(n)})/\frac{u}{q(n)} \in [\sin(a\pi)/a\pi, 1]$ et $x \leq x - \frac{bu}{q(n)\pi} \leq x + \frac{ab}{q(n)} \leq x + a$.
On utilise aussi pour cette majoration le fait que

$$|\ \frac{\sin(q(n)\frac{\pi}{b}(x-t))}{\sin(\frac{\pi}{b}(x-t))}\ | = b\ |\ \sum_{i=1}^{q(n)} e_i(x)e_i(t)\ | \leq Cq(n).$$

La fonction g_n^1 est donc majorée par une fonction intégrable. Calculons maintenant sa limite. On commence d'abord par calculer

$$\lim_{n\longrightarrow\infty}\int_{x-\frac{bu}{q(n)\pi}}^{b} S'(t)\frac{\sin(q(n)\frac{\pi}{b}(x-t))}{q(n)\sin(\frac{\pi}{b}(x-t))}dt.$$

Or, on a vu que la fonction sous le signe intégrale est majorée en valeur absolue par Cf qui est intégrable. Et en appliquant le théorème de Lebesgue, on trouve

$$\lim_{n\longrightarrow\infty}\int_{x-\frac{bu}{q(n)\pi}}^{b} S'(t)\frac{\sin q(n)\frac{\pi}{b}(x-t)}{q(n)\frac{\pi}{b}(x-t)}\frac{\frac{\pi}{b}(x-t)}{\sin\frac{\pi}{b}(x-t)}dt = 0.$$

Finalement, on obtient

$$\lim_{n\longrightarrow\infty} g_n^1(u) = \frac{1}{\pi b} I_{[-a\pi,0]}(u)(S(x)\frac{\sin u}{u})^2 \frac{h(x)}{1-L(x)}$$

et

$$\int g_n^1(u)du \longrightarrow \frac{1}{\pi}S(x)^2\frac{h(x)}{1-L(x)}\int_{-a\pi}^{0}(\frac{\sin u}{u})^2 du > 0,$$

quand n tend vers l'infini (on utilise le théorème de Lebesgue). Donc, $B_n^2 = q(n)\int_{\mathbf{R}} g_n(u)du > q(n)\int_{\mathbf{R}} g_n^1(u)du$, et par suite

$$\exists A > 0 \quad / \quad B_n^2 > q(n)A.$$

2.4 Application de l'estimation de la densité à celle du taux de hasard.

Un estimateur naturel du taux de hasard h, déjà utilisé dans le cadre de la méthode du noyau par Földes et Réjtö (1981) ainsi que par Los, Mack et Wang (1989), sera $f_n/(1-F_n+\frac{1}{n})$, où f_n est l'estimateur de la densité et F_n est l'estimateur de Kaplan-Meier. On montrera alors que l'estimateur h_n, ainsi construit, vérifie la convergence suivante.

Théorème 4 *Si* $\sup_{x\in[0,b]} \mid \bar{f}_n(x) - f(x) \mid \longrightarrow 0$ *lorsque* $n \longrightarrow \infty$ *et que les hypothèses* H_1 *et* H_2 *sont vérifiées, alors si pour tout* γ_1 *et* γ_2 *strictement positifs les séries de termes généraux* $n\sum_{i=1}^{q(n)} \exp\{-\gamma_1 n/q(n)N(e_i,b)\}$ *et* $q(n)\exp\{-\gamma_2 n/q(n)^2\}$ *convergent, alors*

$$\sup_{x\in[0,b]} \mid h_n(x) - h(x) \mid \longrightarrow 0 \ \ p.s., \ \ quand \ \ n \longrightarrow \infty.$$

En prenant là aussi, comme exemple, la base trigonométrique pour construire f_n, on obtient

Corollaire 4 *Dans le cas de la base trigonométrique, si* f *vérifie* H_2*, et si* $\sup_{x\in[0,b]} \mid \bar{f}_n(x) - f(x) \mid \longrightarrow 0$ *quand* $n \longrightarrow \infty$*, alors, si,* $\forall\gamma > 0$*,* $\sum_{n=1}^{\infty} nq(n)\exp\{-\gamma n/(q(n))^2\} < \infty$*, on a*

$$\sup_{x\in[0,b]} \mid h_n(x) - h(x) \mid \longrightarrow 0 \ \ p.s., \ \ quand \ \ n \longrightarrow \infty.$$

3 Etude des convergences.

3.1 Principe général.

Pour étudier la convergence de f_n vers f, on utilisera une décomposition de $F_n - F$ sous la forme $\frac{1}{n}\sum_{i=1}^{n} Z_i + R_n$, où les Z_i sont indépendants, et R_n tend asymptotiquement vers zéro. Cette technique a été récemment utilisée par Los, Mack et Wang (1989), dans un autre cadre. Notre décomposition se basera, elle, essentiellement sur les résultats de Reid (1981), permettant de développer $F_n(t) - F(t)$ en fonction des "courbes d'influence" de F_n. Reid a montré exactement que

$$\forall t, \ \ t \geq 0, \quad F_n(t) - F(t) = P_n(t) + R_n(t) \tag{8}$$

où

$$nP_n(t) = \sum_{i=1}^{n} (I_{\{\delta_i=1\}} K_1(t, X_i) + I_{\{\delta_i=0\}} K_2(t, X_i)) \tag{9}$$

avec

$$\begin{cases} K_1(t,s) = S(t)\int_0^{s\wedge t} \frac{h(u)}{1-L(u)}du - S(t)\frac{I_{\{s\leq t\}}}{1-L(s)} \\ \\ K_2(t,s) = S(t)\int_0^{s\wedge t} \frac{h(u)}{1-L(u)}du \end{cases} \tag{10}$$

(K_1 et K_2 sont "les courbes d'influence" de F_n, $s \wedge t = \min(s,t)$).

Le reste R_n est uniformément majoré sur tout intervalle $[0,b]$ de la façon suivante

$$\sup_{t\in[0,b]} \mid R_n(t) \mid \leq C_b\{\sum_{i=0}^{1} \parallel H_n^i - H^i \parallel_\infty^2 + \frac{1}{n}\} \tag{11}$$

où C_b est une constante (variable avec b). Cette majoration est valable dès que l'hypothèse H_0 est vérifiée. (Ce résultat se déduit du lemme 2.1, Mielniczuk(1985)). La décomposition de $F_n(t) - F(t)$ en fonction de $P_n(t)$ et $R_n(t)$ permet d'évaluer asymptotiquement les quantités du type suivant

$$C(m,b) = \int_0^b m(t)dF_n(t) - \int_0^b m(t)dF(t) \tag{12}$$

$$= \int_0^b m(t)d(P_n(t) + R_n(t)).$$

Afin de simplifier les calculs imposés par le développement de ces quantités, on introduit d'abord les variables Z_i définies par

$$Z_i = \int_0^{s\wedge b} S(t)m(t)\frac{h(t)}{1-L(t)}dt + \int_0^b S'(t)m(t)(\int_0^{X_i\wedge t} \frac{h(u)}{1-L(u)}du)dt \tag{13}$$

$$-\delta_i S(X_i)m(X_i)\frac{I_{\{X_i\leq b\}}}{1-L(X_i)} - \delta_i \int_{X_i}^b S'(t)m(t)\frac{1}{1-L(X_i)}dt.$$

Les principales propriétés de ces variables et des quantités $C(m,b)$, sont résumées dans le lemme suivant, qu'on va utiliser dans la plupart des démonstrations des théorèmes déjà énoncés, et qui en constitue en quelque sorte la clé.

Lemme 2 *les Z_i sont des variables aléatoires centrées, de variance égale à*

$$\int_0^b (S(s)m(s) + \int_s^b S'(t)m(t)dt)^2 \frac{h(s)}{1-L(s)}ds,$$

et on peut écrire

$$C(m,b) = \frac{1}{n}\sum_{i=1}^{n} Z_i + R_n^* \tag{14}$$

où

$$\mid R_n^* \mid < N(m,b)\{\sup_{t\in[0,b]} \mid R_n(t) \mid\}. \tag{15}$$

Dès lors on aura

$$\begin{cases} E(C(m,b)) = O\{N(m,b)\frac{(\log n)^2}{n}\} \\ Var(C(m,b)) = \frac{1}{n}Var(Z_i)(1+o(1)). \end{cases} \tag{16}$$

3.2 Démonstration du théorème 1

On commence d'abord par montrer que $\lim_{n\longrightarrow\infty} M_n = 0$, où $M_n = E(\| f_n - f \|^2_{L^2[0,b]})$. Les $(e_i), i \geq 1$, formant une base orthonormale de $L^2[0,b]$, on a

$$M_n = \sum_{i=1}^{q(n)} E[(\hat{a}_i - a_i)^2] + \sum_{i=q(n)+1}^{\infty} a_i^2.$$

Or

$$\hat{a}_i - a_i = \int_0^b e_i dF_n - \int_0^b e_i dF = C(e_i, b)$$

et d'après le lemme 1, on a

$$\begin{aligned} E[(\hat{a}_i - a_i)^2] &= \frac{1}{n}\int_0^b \{S(s)e_i(s) + \int_s^b S'(t)e_i(t)dt\}^2 \frac{h(s)}{1-L(s)}ds \\ &\quad + O((N(e_i,b))^2 \frac{(\log n)^4}{n^2}) + O(N(e_i,b)\frac{(\log n)^2}{n}[\frac{1}{n}V(Z_i)]^{\frac{1}{2}}). \end{aligned}$$

L'hypothèse H_1 est supposée vérifiée, donc les $(e_i), i \geq 1$, sont uniformément bornés et par suite, on peut écrire

$$\begin{aligned} \sum_{i=1}^{q(n)} E[(\hat{a}_i - a_i)^2] &= \frac{1}{n}\int_0^b \sum_{i=1}^{q(n)} \{S(s)e_i(s) + \int_s^b S'(t)e_i(t)dt\}^2 \frac{h(s)}{1-L(s)}ds \quad (17) \\ &\quad + O(D_n \frac{(\log n)^4}{n^2}) + O(\sum_{i=1}^{q(n)} N(e_i,b)\frac{(\log n)^2}{n^{\frac{3}{2}}}). \end{aligned}$$

Le premier terme est un $O(q(n)/n)$. S'il tend vers zéro ainsi que le second, le carré du troisième terme qui est un $O(q(n)D_n(\log n)^4/n^3)$ tendra aussi vers zéro. La démonstration de la deuxième partie du théorème, est la même que celle du théorème 1 (cf. Delecroix et Yazourh (1992)), il suffit de remplacer les Z_j^i par

$$\begin{aligned} Z_j^i &= \int_0^{X_i \wedge b} S(t)e_i(t)\frac{h(t)}{1-L(t)}dt + \int_0^b S'(t)e_i(t)(\int_0^{X_i \wedge t} \frac{h(u)}{1-L(u)}du)dt \\ &\quad - S(X_j)e_i(X_j)\frac{1_{\{(\delta_j=1)\cap(X_i\leq b)\}}}{1-L(X_i)} - \int_0^b S'(t)e_i(X_j)\frac{1_{\{(\delta_j=1)\cap(X_i\leq t)\}}}{1-L(X_i)}dt. \end{aligned}$$

3.3 Preuve du corollaire 1

Dans le cas de la base trigonométrique, $\forall i,\ \ i \geq 1,\ \ a_i = O(1/i^2)$ (cf. Sansone (1959) p. 106), donc $\sum_{i=q(n)+1}^{\infty} a_i^2 = O(1/q(n))$ qui tend vers zéro quand n tend vers l'infini, et $N(e_i, b) = O(i)$ ainsi que $D_n = O(q(n)^3)$. En appliquant alors le théorème 1, on obtient les deux convergences souhaitées pour l'estimateur f_n, construit à partir de cette base.

3.4 Preuve du théorème 2

Il suffit de prouver que $\sup_{x\in[0,b]} \mid f_n(x) - f(x) \mid \longrightarrow 0$ *p.s.*, quand $n \longrightarrow \infty$, et donc $\sup_{x\in[0,b]} \mid \sum_{i=1}^{q(n)} (\mid \hat{a}_i - a_i \mid) e_i(x) \mid \longrightarrow 0$ *p.s.*

Pour tout $\varepsilon > 0$, $\sum_n \{\sum_{i=1}^{q(n)} P(\hat{a}_i - a_i) > \varepsilon/q(n)\} < \infty$, cette condition est vérifiée (cf. (14), Delecroix et Yazourh (1992)), si pour tous $\gamma_1 > 0$ et $\gamma_2 > 0$ les séries de termes généraux $n \sum_{i=1}^{q(n)} \exp\{-\gamma_1 n/q(n)N(e_i,b)\}$ et $q(n)\exp\{-\gamma_2 n/q(n)^2\}$ sont convergentes.

3.5 Preuve du corollaire 2

Dans le cas où les $(e_i), i \geq 1$, sont les fonctions de la base trigonométrique, $\forall i \geq 1$, $N(e_i, b) = O(i)$ et donc, $N(e_i, b) = O(q(n))$.

3.6 Preuve du théorème 3

Voir preuve du théorème 2 (cf. Delecroix et Yazourh (1992)); on remplace h par f, h_n par f_n et les $Z_{i,n}$ par

$$\begin{aligned} Z_{i,n} &= \int_0^{X_i \wedge b} S(t)\overline{W}_n(t)\frac{h(t)}{1-L(t)}dt + \int_0^b S'(t)\overline{W}_n(t)(\int_0^{X_i\wedge t} \frac{h(u)}{1-L(u)}du)dt \\ &\quad -S(X_i)\overline{W}_n(X_i)\frac{1_{\{(\delta_j=1)\cap(X_i\leq b)\}}}{1-L(X_i)} - \int_0^b S'(t)\overline{W}_n(t)\frac{1_{\{(\delta_j=1)\cap(X_i\leq t)\}}}{1-L(X_i)}dt. \end{aligned}$$

3.7 Preuve du corollaire 3

D'après le lemme 1, $1/B_n < 1/(q(n)A)^{\frac{1}{2}}$, avec $A > 0$. Les deux hypothèses du théorème 3, sont donc vérifiées si $q(n)/n \longrightarrow 0$ et $q(n)^3(\log n)^4/n \longrightarrow 0$, lorsque $n \longrightarrow \infty$, puisque $N(W_n, b) = O(q(n)^2)$, dans le cas des fonctions trigonométriques.

3.8 Preuve du théorème 4

Considérons

$$h - h_n = \frac{f}{1-F} - \frac{f_n}{1-F_n+\frac{1}{n}} = f\frac{F - F_n + \frac{1}{n}}{(1-F)(1-F_n+\frac{1}{n})} + \frac{f-f_n}{1-F_n+\frac{1}{n}}.$$

Le premier terme se majore par

$$\frac{1}{(1-F(b))(1-F_n(b)+\frac{1}{n})}\{\sup_{t\in[0,b]} \mid f(t) \mid (\sup_{t\in[0,b]} \mid F(t) - F_n(t) \mid + \frac{1}{n})\}.$$

b vérifie l'hypothèse H_0, donc $F(b) < 1$, et comme f vérifie H_2, on a

$$\sup_{t\in[0,b]} \mid h(t) - h_n(t) \mid = O(\frac{\sup_{t\in[0,b]} \mid F(t) - F_n(t) \mid + \frac{1}{n} + \sup_{t\in[0,b]} \mid f(t) - f_n(t) \mid}{1 - F_n(b) + \frac{1}{n}}).$$

D'autre part, $F(b) < 1$: $\exists \varepsilon > 0$ / $F(b) \leq 1 - \varepsilon$, et $F_n \longrightarrow F$ $p.s.$, quand $n \longrightarrow \infty$, donc $\forall \varepsilon' > 0, \exists \eta > 0 / \forall n \geq \eta; 1 - F_n > 1 - F - \varepsilon'$ $p.s.$ En choisissant, $\varepsilon' = \varepsilon/2$, on a, $1 - F_n(b) + \frac{1}{n} \geq \varepsilon/2$ $p.s.$, c.à.d. $1/(1 - F_n(b) + \frac{1}{n}) \leq 2/\varepsilon$ $p.s.$ On en déduit finalement que

$$\sup_{t\in[0,b]} \mid h(t) - h_n(t) \mid = O\{ \sup_{t\in[0,b]} \mid F(t) - F_n(t) \mid + \frac{1}{n} + \sup_{t\in[0,b]} \mid f(t) - f_n(t) \mid\}.$$

Donc il suffit que les hypothèses du théorème 2, soient vérifiées pour que $\sup_{t\in[0,b]} \mid f(t) - f_n(t) \mid$ tende vers zéro presque sûrement. Et comme b vérifie H_0, $\sup_{t\in[0,b]} \mid F(t) - F_n(t) \mid$ tend vers zéro presque sûrement. On obtient donc la convergence souhaitée.

3.9 Preuve du corollaire 4

Voir preuve du corollaire 2.

4 Commentaires.

En utilisant les convergences uniformes p.s., ponctuelle et L^2, l'estimateur f_n peut être comparé à l'estimateur f_n^* construit par la méthode du noyau et étudié par Földes et al (1981), Mielniczuk (1983) et Los, Mack et Wang (1989).

Dans le cas de l'estimateur f_n, on a vu que la convergence uniforme $p.s.$, dépend surtout de la série qui doit être uniformément convergente, alors que pour f_n^*, Fôldes et Al (1981) ont obtenu cette convergence, en imposant plus de conditions sur les distributions des X_i et C_i.

Mielniczuk (1985) et Los, Mack et Wang (1989), ont étudié les convergences ponctuelles des estimateurs à noyau de la densité. Ils ont montré que le MSE asymptotique optimal est en $n^{-4/5}$. Les résultats obtenus permettent d'exprimer la variance asymptotique de f_n^* sous la forme

$$V(f_n^*(x)) = \frac{A_1}{nh_n} \frac{f(x)}{1 - G(x)} + O(\frac{1}{nh_n})$$

avec $h_n \longrightarrow 0$ et $nh_n \longrightarrow \infty$ (h_n est la "fenêtre" et A_1 est une constante qui dépend du "noyau"). Dans le cas de l'estimateur f_n, si $(e_i), i \geq 1$, est la base trigonométrique, on montre en remplaçant dans le lemme 2 $m(.)$ par $W_n(x, .)$ que $V(f_n(x)) \sim \frac{q(n)^2}{n} \frac{f(x)}{1-G(x)} A_2$. Il suffit donc de choisir $h_n = q(n)^{-2}$ pour avoir la même vitesse de convergence pour les deux variances.

Si en plus la densité f est telle que $\sum_{q(n)+1}^{\infty} a_i e_i(x)$ converge assez rapidement vers zéro (par exemple en $q(n)^{-4}$, ce qui correspond au biais usuel que donne le noyau avec $h_n = q(n)^{-2}$) alors le MSE correspondant à f_n est en $n^{-4/5}$, donc f_n est aussi performant que f_n^* de ce point de vue.

Concernant la convergence $L^2[0,b]$ dans le cas de la base trigonométrique par exemple, on montre d'après (17), en choisissant $q(n)$ impair de la forme : $q(n) = 2q'(n)+1$, que

$$\begin{aligned}\sum_{i=1}^{q(n)} E[(\hat{a}_i - a_i)^2] &= \frac{2}{nb}\sum_{i=1}^{q'(n)}\int_0^b \{(S^2(s) + S(s)S(b) \\ &+ (\int_s^b S'(t)\sin(\frac{2i\pi}{b}(t-\frac{b}{2}))dt)^2 \\ &+ (\int_s^b S'(t)\cos(\frac{2i\pi}{b}(t-\frac{b}{2}))dt)^2\}^2 \frac{h(s)}{1-L(s)}ds \\ &+ O(q'(n)^3\frac{(\log n)^4}{n^2}) + O(q'(n)^2\frac{(\log n)^4}{n^{\frac{3}{2}}}).\end{aligned}$$

Le premier terme est un $O(q'(n)/n)$ et les deux derniers sont des $o(q'(n)/n)$, pourvu que l'on choisisse $q'(n)$ tel que $q'(n)(\log n)^2/\sqrt{n}$ tende vers zéro. Donc le $MISE$ asymptotique s'écrit

$$AMISE = O(\frac{q'(n)}{n}) + O(\frac{1}{q'(n)})$$

puisque le $MISE$ est égale à $\sum_{i=1}^{q(n)} E[(\hat{a}_i - a_i)^2] + \sum_{i=q(n)+1}^{\infty} a_i^2$ et d'après Sansone (1959) p. 106, on montre que $\forall k,\ a_{2k}^2 + a_{2k+1}^2 = O(1/k^2)$, donc $\sum_{i=q(n)+1}^{\infty} a_i^2 = O(1/q'(n))$.

Finalement, on a le $MISE$ asymptotique qui équivaut à un $O(1/q'(n))$, sous la contrainte imposée à $q'(n)$. L'écart optimal est donc plus grand que dans le cas de l'estimateur à noyau et dépend étroitement du reste de la série $\sum_{i=q(n)+1}^{\infty} a_i^2$, et de la base $(e_i), i \geq 1$ choisie.

L'estimateur du taux de hasard proposé ici est construit de la même manière que les deux estimateurs h_1^* et h_2^* étudiés par Földes et al (1981) et Los, Mack et Wang (1989) et construits à partir des estimateurs à noyau de la densité. Dans les deux méthodes on remarque que les résultats de convergence dépendent de ceux des estimateurs de la densité, donc à priori, pour comparer les performances de ces estimateurs, il suffit de comparer celles des estimateurs de la densité.

En conclusion, on peut dire que l'efficacité de f_n et de h_n est liée au choix de $q(n)$ et de la base qui devrait être " la plus adaptée possible" à la densité f, problème qui nécessite une étude détaillée accompagnée par des simulations.

References

[1] Delecroix M. et Yazourh O. (1992). Estimation non paramétrique du taux de hasard en présence de censures droites: méthode des fonctions orthogonales, *Statistique et Analyse des Données*, 16, 39-62.

[2] Droesbeke J. J. , Fichet B. et Tassi Ph. (1989), Eds, *Analyse statistique des durées de vie*. Economica, Paris.

[3] Efron B. (1967). The two sample problem with censored data, *Proc. 5th. Berkeley Symp. vol.* 4, 831-853.

[4] Földes A. , Rejtö L. and Winter B. B. (1981). Strong consistency properties of non parametric estimators for randomly censored data II, Estimation of density and failure rate, *Period. Math. Hungar.* 12, 15-29.

[5] Gneyou K. E. (1991). Inférence statistique non paramétrique pour l'analyse du taux de panne en fiabilité, thèse soutenue à l'Université de Paris VI.

[6] Grégoire G. (1991). Choix de la taille de la fenêtre pour l'estimation de l'intensité d'un processus ponctuel, application aux données de durée de vie censurées. Communication présentée aux XXIIIèmes Journées de Statistique.

[7] Huber C. et Lecoutre J. P. (1990). Estimation fonctionnelle dans les modèles de durée. In Droesbeke J. J. , Fichet B. et Tassi Ph. , Eds, *Analyse statistique des durées de vie*. Economica, Paris.

[8] Hjort N. L. (1985). Discussion contribution to Andersen and Borgan's review article, *Scand. J. Statist.* 12, 141-150.

[9] Kaplan E. and Meier P. (1958). Non parametric estimation from incomplete observations, *J.A.S.A.* 53, 457-481.

[10] Kimura D. K. (1972). Fourier series methods for censored data, thèse soutenue à Washington University Seattle.

[11] Los H., Mack Y. P. and Wang J. L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator, *Prob. Th. and rel. Fields*, 80, 461-473.

[12] Mielniczuk J. (1985). Properties of some kernel estimators and of the adapted loftsgarden-Quesenberry estimator of a density function for censored data, *Periodica Math. Hungarica,* 16, 69-80.

[13] Reid N. (1981). Influence functions for censored data, *Annals of Statist.* 9, 78-92.

[14] Sansone G.(1959). Orthogonal functions, *Pure and Applied Mathematics* IX, Interscience Publishers, New York.

O. Yazourh
Laboratoire de Statistique et Probabilités
U.F.R. de Mathématiques
Université des Sciences et Technologies de Lille
F-59655 VILLENEUVE D'ASCQ CEDEX