55376
1992
5

N° d'ordre : 940

THÈSE

présentée à

L'UNIVERSITÉ DES SCIENCES ET
TECHNOLOGIES DE LILLE

pour obtenir

LE TITRE DE DOCTEUR DE L'UNIVERSITÉ

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

UNIVERSITE DES Sc. et Techn. BIBLIOTHEQUE UNIVERSITAIRE LILLE

par

YAZOURH Ouafae

SUR L'UTILISATION DES FONCTIONS ORTHOGONALES DANS LES MODÈLES DE DURÉE.

Soutenue le 6 Juillet 1992 devant la Commission d'Examen :

Président et Rapporteur : J. DELPORTE, Université Catholique de Lille
Rapporteur : D. BOSQ, Université de Paris VI
Examinateurs : C. LEFEVRE, Université Libre de Bruxelles
R. MOCHÉ, Université de Lille
M. DELECROIX, Université de Toulouse I

55376
1992
5

A mes parents,

A mon frère et mes sœurs,

A ma famille et mes amis.

Je tiens à exprimer ma profonde gratitude à :

Monsieur le Professeur Jean DELPORTE qui me fait l'honneur de présider le jury de cette thèse et d'en être rapporteur.

Monsieur le Professeur Michel DELECROIX qui a suivi avec beaucoup d'intérêt l'évolution de ce travail, me prodiguant sans cesse des conseils précieux et bienveillants.

Monsieur le Professeur Denis BOSQ, rapporteur de cette thèse, pour sa lecture de mon travail, pour ses critiques constructives et ses conseils.

Messieurs les Professeurs Claude LEFEVRE et Raymond MOCHÉ, qui ont eu l'extrême obligeance de faire partie du jury de cette thèse.

Je remercie également toutes les personnes qui ont participé à la réalisation matérielle de ce travail, en particulier Madame Arlette LENGAIGNE ainsi que Messieurs Daniel FLIPO et Charles SUQUET qui m'ont aidée à utiliser les techniques du traitement de texte.

PLAN

INTRODUCTION

L'étude des modèles de durée a pour but d'analyser des variables aléatoires réelles, interprétant des phénomènes de durées. Dans la pratique ces variables représentent, la durée passée dans un état donné (durée de chômage, durée d'activité, durée de survie, durée de fonctionnement sans panne, etc...)

Les applications de ces modèles sont assez variées, ils interviennent le plus souvent dans le domaine biomédical, en économie, en théorie de la fiabilité et dans d'autres domaines aussi. La caractéristique fondamentale des modèles de durée, est la présence de "données incomplètes", c'est à dire que certaines observations de l'échantillon ne prennent pas leurs valeurs exactes. On parle alors de modèle censuré. Comme exemple, on peut observer la durée de survie X^0 chez n individus traités pour une certaine maladie, dont m seront décédés par d'autres causes ou perdues de vue aux dates $(C_1, C_2, \ldots, C_m)$. La variable aléatoire observée est en fait

$$X_i = \inf(X_i^0, C_i) = X_i^0 \wedge C_i, \quad i = 1, \ldots, n.$$

Plus précisément, on adoptera le modèle d'EFRON (1967), où on suppose que les X_i^0 sont des variables aléatoires positives, *i.i.d*, de même que les C_i, que les X_i et les C_i sont indépendantes entre elles, et que le statisticien sait si l'observation est censurée ou non (il connaît les valeurs de $\delta_i = \mathbf{1}_{\{X_i^0 \leq C_i\}}, \quad i = 1, \ldots, n$). Les lois des variables aléatoires X^0 de durée, sont caractérisées essentiellement par les fonctions, taux de hasard h et densité f ($h = \frac{f}{1-F}$, où F est la fonction de répartition de X^0). Il est donc naturel que l'analyse des modèles de durée se base sur l'estimation de ces paramètres fonctionnels. Les réponses possibles à ce problème sont les suivantes :

La première est l'approche paramétrique qui a fait l'objet de trés nombreux travaux, dont on peut citer ceux de J. D. Kalbfleish et R. L. Prentice (1980), D. R. Cox et D. Oakes (1984). Dans l'étude de ces modèles paramétriques, on suppose que la loi de la variable X^0, appartient à la classe $\mathcal{P}_0 = \{P_\theta, \ \theta \in \Theta, \ \Theta \subset \mathbf{R}^d\}$, et donc on sait qu'elle dépend d'un certain

nombre de paramètres qu'on peut estimer "facilement" à partir de données observées. Mais l'inconvénient majeur de cette méthode reste la distorsion qui peut exister entre le modèle vrai du phénomène étudié et le modèle retenu.

La deuxième réponse est l'approche non paramétrique qui est susceptible de vaincre cette difficulté, puisqu'on suppose que la loi de X^0 est quelconque. Cette méthode a été aussi, utilisée par de nombreux auteurs, comme A. Földes, L. Rejtö et B. B. Winter (1981), A. Tanner et Wing Hung Wong (1983), J. Mielniczuk (1985), K. E. Gneyou (1991) et G. Gregoire (1991). Dans la quasi totalité de ces travaux, on s'intéressait à l'estimation du taux de hasard h et de la densité f, qui se base en premier lieu sur l'estimation de la fonction de répartition F. Vu la présence de censures dans les modèles de durées étudiés, deux méthodes sont donc possibles : la première consiste à construire un estimateur de F à partir de l'échantillon réduit et dans ce cas, non seulement on aura une perte d'information mais aussi, les procédures statistiques usuelles seront considérablement compliquées. La deuxième démarche, est celle qui a été choisie par E. Kaplan et P. Meier (1958), ils ont défini l'estimateur produit limite de la fonction de répartition à partir de toutes les observations de l'échantillon, et ont montré l'intérêt de cet estimateur par rapport à celui qu'on pourrait construir à partir de l'échantillon réduit, c'est à dire sans tenir compte des observations censurées. Cet estimateur produit limite, est effectivement la base de très nombreux travaux concernant les modèles de durée. On peut citer ceux des auteurs ci-dessus, qui pour la plupart ont construit des estimateurs du taux de hasard h et de la densité f par la méthode du noyau. On a vu que cette approche non paramétrique présente l'avantage de ne présupposer aucune forme particulière du taux de hasard et de la densité, mais en contre partie, il est nécessaire de disposer d'un nombre important d'observations, puisqu'il s'agit d'estimation de paramètres fonctionnels appartenant à un espace de dimension infinie.

Une troisième approche qu'on peut situer entre les deux précédentes, consiste en la modélisation semi-paramétrique, si la loi vraie de la v.a. X^0, appartient à la classe de loi définie par:

$$\mathcal{P} = \{\mathbf{1}_{\theta,e};\ \theta \in \Theta,\ e \in E\}$$

où Θ est l'espace des paramètres et G une collection de fonctions. Ce type de modèles a été introduit dès 1978 par J. Kalbfleisch, et en suite utilisé par J. Begun, W. Hall et J. Wellner (1983) et bien d'autres.

Dans ce travail on propose la méthode des fonctions orthogonales, qu'on peut situer dans le cadre d'une approche semi-paramétrique, pour estimer la densité et le taux de hasard. Après une introduction de cette méthode et une synthèse des résultats déjà obtenus, dans le premier chapitre, on présente dans le deuxième, une étude d'estimation du taux de hasard par projection dans un espace L^2. Enfin, dans le troisième chapitre, on estime la densité par la même méthode et on en déduit un deuxième estimateur du taux de hasard.

CHAPITRE I

Estimateur de Kaplan-Meier et synthèse des résultats de convergence

1 Estimateur de Kaplan-Meier de la fonction de survie

1.1 Définition

Selon la modélisation d'Efron (1967), on considère deux suites de *v.a.* , *i.i.d.* , positives et mutuellement indépendantes. Les X_i^0, $i \geq 1$, représentent les durées de vie étudiées, et les C_i, $i \geq 1$, symbolisent les censures (droites) éventuellement apportées aux X_i^0. On appelle f la densité des X_i^0, supposée continue et F leur fonction de répartition. Les mêmes fonctions correspondantes aux C_i seront notées g et G. On a vu que pour ce modèle, l'échantillon observé par le statisticien, est l'ensemble des n couples (X_i, δ_i), $1 \leq i \leq n$, où $X_i = \inf(X_i^0, C_i)$ et $\delta_i = \mathbf{1}_{\{X_i^0 \leq C_i\}}$. On appellera l la densité des *v.a.* X_i et L leur fonction de répartition vérifiant, $1 - L = (1 - F)(1 - G)$.

L'estimateur de la fonction de survie $S = 1 - F$, introduit par E. Kaplan et P. Meier (1958), est défini par:

$$\hat{S}_n(t) = \begin{cases} \prod_{i/X_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} & \text{si } t \leq X_{(n)} \\ 0 & \text{si } t > X_{(n)} \end{cases}$$

$(X_{(i)})_{1 \leq i \leq n}$, est l'échantillon $(X_i)_{1 \leq i \leq n}$ ordonné et $\delta_{(i)}$ représente la valeur de δ pour $X_{(i)}$. Un estimateur naturel de la fonction de répartition F, sera donc $\hat{F}_n = 1 - \hat{S}_n$.

L'idée de construction de cet estimateur produit limite est la suivante: la fonction de survie étant définie par $S(t) = P(X^0 > t)$, peut s'écrire aussi sous forme d'un produit de probabilités, tel que:

$$P(X^0 > t) = \prod_{i/X_{(i)} \leq t} P(X^0 \geq X_{(i)} / X^0 \geq X_{(i-1)})$$

avec $X_{(0)} = 0$.

Chaque probabilité conditionnelle est estimée ensuite à partir du rapport du nombre d'observations après $X_{(i)}$ par le nombre d'observations après $X_{(i-1)}$. C'est à dire on estime $P(X^0 \geq X_{(i)} / X^0 \geq X_{(i-1)})$ par $(\frac{n-i}{n-i+1})^{\delta_{(i)}} = \frac{n-i}{n-i+1}$, si l'observation $X_{(i)}$ n'est pas une censure et par $(\frac{n-i}{n-i+1})^{\delta_{(i)}} = 1$, si l'observation $X_{(i)}$ est une censure, d'où l'écriture de $\hat{S}_n$.

On définit aussi, la fonction Λ appelée taux de hasard cumulé, par $\Lambda(t) = -\log S(t)$. Cette fonction joue un rôle très important dans l'étude d'estimation du taux de hasard, puisque ce dernier n'est autre que la dérivée de Λ qu'on peut estimer facilement par $\hat{\Lambda}_n(t) = -\log \hat{S}_n(t)$. C'est à dire, en remplaçant S par l'estimateur de Kaplan-Meier $\hat{S}_n$.

1.2 Courbes d'influence

D'après l'expression de Peterson (1977), on peut exprimer la fonction de survie S, par rapport aux deux fonctions de subsurvie H^0 et H^1, définies par

$$H^i(t) = P(X_j > t, \delta_j = i) \qquad avec \quad i = 0, 1.$$

tel que :

$$S(t) = \exp \int_0^t \frac{dH^1(s)}{(H^1 + H^0)(s)} \times \exp \sum_{s \leq t} \log \frac{H^1(s^+) + H^0(s^+)}{H^1(s^-) + H^0(s^-)}$$

le premier terme correspond à une integration sur les intervalles de continuité de H^1, le second terme est une sommation sur les points de discontinuité de H^1.

En estimant les deux fonctions de subsurvie par :

$$H^i(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{X_j > t \,,\, \delta_j = i\}} \quad avec \quad i = 0, 1.$$

A. V. Peterson (1977), montre que l'estimateur de Kaplan-Meier, s'exprime aussi de la manière suivante:

$$\hat{S}_n(t) = \exp \int_0^t \frac{dH_n^1(s)}{(H_n^1 + H_n^0)(s)} \times \exp \sum_{s \leq t} \log \frac{H_n^1(s^+) + H_n^0(s^+)}{H_n^1(s^-) + H_n^0(s^-)}$$

En appliquant la fonction logarithme à S et $\hat{S}_n$, on peut écrire le taux de hasard cumulé Λ et son estimateur $\hat{\Lambda}_n$, tel que:

$$\Lambda(t) = -\int_0^t \frac{dH^1(s)}{(H^1 + H^0)(s)} - \sum_{s \leq t} \log \frac{H^1(s^+) + H^0(s^+)}{H^1(s^-) + H^0(s^-)}$$

et

$$\hat{\Lambda}_n(t) = -\int_0^t \frac{dH_n^1(s)}{(H_n^1 + H_n^0)(s)} - \sum_{s \leq t} \log \frac{H_n^1(s^+) + H_n^0(s^+)}{H_n^1(s^-) + H_n^0(s^-)}$$

N. Reid (1981), a utilisé ces deux expressions, pour faire une étude détaillée de l'estimateur de Kaplan-Meier, permettant de regrouper et d'améliorer les différents résultats de convergence déjà montrés. L'idée est la suivante, considérer la fonction de survie et son estimateur comme fonctionnelle de (H^1, H^0) et (H_n^1, H_n^0), ensuite utiliser le dévelpement de Von-Mises (1947). Ceci a été d'abord appliqué au taux de hasard cumulé afin de simplifier les calculs.

On peut donc écrire, $\Lambda(t) = T(H^1, H^0, t)$ et $\hat{\Lambda}_n(t) = T(H_n^1, H_n^0, t)$, où T est une fonctionnelle de $\mathcal{B} \times \mathcal{B}$ dans $\mathbf{R}$, avec $\mathcal{B}$, l'espace des mesures positives bornées. Le dévelopement de Von-Mises de cette fonctionnelle donne:

$$\begin{aligned}
\hat{\Lambda}_n(t) - \Lambda(t) &= T(H_n^1, H_n^0, t) - T(H^1, H^0, t) \\
&= \int IC_1(T, H^1, H^0; s)(t) d(H_n^1 - H^1)(s) \\
&\quad + \int IC_2(T, H^1, H^0; s)(t) d(H_n^0 - H^0)(s) \\
&\quad + R_n(t)
\end{aligned}$$

R_n est un reste, qu'on appelle terme d'ordre supérieur, IC_1 et IC_2 sont les courbes d'influence du taux de hasard cumulé définies par:

$$\frac{\partial}{\partial\varepsilon}T(H^1+\varepsilon(H_n^1-H^1),H^0+\lambda(H_n^0-H^0))\mid_{\varepsilon=0,\lambda=0} = \int IC_1(T,H^1,H^0;s)d(H_n^1-H^1)(s)$$

et

$$\frac{\partial}{\partial\lambda}T(H^1+\varepsilon(H_n^1-H^1),H^0+\lambda(H_n^0-H^0))\mid_{\varepsilon=0,\lambda=0} = \int IC_2(T,H^1,H^0;s)(t)d(H_n^0-H^0)(s)$$

N. Reid (1981), a montré que ces deux courbes d'influence de $\hat{\Lambda}_n$, s'expriment sous la forme:

$$IC_1(T,H^1,H^0;s)(t) = \int_0^{s\wedge t}\frac{h(u)}{1-L(u)}du+\frac{\mathbf{1}_{\{s\leq t\}}}{1-L(u)}$$

et

$$IC_2(T,H^1,H^0;s)(t) = \int_0^{s\wedge t}\frac{h(u)}{1-L(u)}du$$

($s\wedge t=\min(s,t)$), on notera par la suite ces deux fonctions par $k_1(t,s)$ et $k_2(t,s)$.

Les courbes d'influence de l'estimateur de Kaplan-Meier, sont définies de la même manière, mais cette fois ci on considère $\hat{S}_n(t)$ comme fonctionnelle $T^0(H_n^1,H_n^0,t)=\exp -T(H_n^1,H_n^0,t)$ (puisque $\hat{S}_n(t)=\exp-\hat{\Lambda}_n(t)$). On en déduit alors que:

$$IC_1(T^0,H^1,H^0;s)(t) = S(t)(\int_0^{s\wedge t}\frac{h(u)}{1-L(u)}du+\frac{\mathbf{1}_{\{s\leq t\}}}{1-L(u)})$$

et

$$IC_2(T^0,H^1,H^0;s)(t) = S(t)(\int_0^{s\wedge t}\frac{h(u)}{1-L(u)}du)$$

1.3 Définitions de $\hat{\Lambda}_n(t) - \Lambda(t)$ et $\hat{F}_n(t) - F(t)$

D'après le lemme 1 (cf. M. Delecroix et O. Yazourh (1992)) $\hat{\Lambda}_n(t) - \Lambda(t)$ s'exprime aussi de la manière suivante:

$$(1) \qquad \hat{\Lambda}_n(t) - \Lambda(t) = \frac{1}{n}\sum_{i=1}^{n} Z_i + R_n(t) = P_n(t) + R_n(t)$$

avec

$$Z_i = \int_0^{X_i \wedge t} \frac{h(u)}{1 - L(u)} du - \frac{1}{1 - L(X_i)} \mathbf{1}_{\{(\delta_i = 1) \cap (X_i \leq t)\}} \ , \ 1 \leq i \leq n.$$

Concernant le reste R_n, on peut se contenter ici, de citer le résultat de majoration suivant, qui joue un rôle important dans notre travail

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

où

$$\| F \|_\infty = \sup_{x \geq 0} \mid F(x) \mid$$

Cette majoration est valable, dès que l'hypothèse suivante, est vérifiée

$$(H_0) \qquad P(X_j > b, \delta_j = i) > 0 \ , \quad i = 0, 1.$$

(cf. lemme 2.1. J. Mielniczuk (1985), pour un développement et mise en forme des résultats de N. Reid).

En appliquant maintenant le lemme 2 (cf. chapitre III) on peut écrire:

$$(3) \qquad \hat{S}_n(t) - S(t) = \frac{1}{n}\sum_{i=1}^{n} Z_i^0 + R_n^0(t) = P_n^0(t) + R_n^0(t)$$

avec

$$Z_i^0 = \int_0^{X_i \wedge t} S(t) \frac{h(u)}{1 - L(u)} du + \int_0^b S'(u) \left(\int_0^{X_i \wedge u} \frac{h(v)}{1 - L(v)} dv \right) du - \delta_i \frac{S(t)}{1 - L(X_i)}$$

Et on montre que le reste R_n^0, se majore de la même manière que R_n.

Lemme 1 *Si b vérifie l'hypothèse (H_0), alors :*

$$\sup_{t \in [0,b]} \mid R_n^0(t) \mid = O\left(\sum_{i=0}^{1} \| H_n^i - H^i \|_\infty^2 + \frac{1}{n} \right) \tag{4}$$

Preuve :

$$\begin{aligned} \hat{S}_n(t) - S(t) &= \exp(-\hat{\Lambda}_n(t)) - \exp(-\Lambda(t)) \\ &= (\hat{\Lambda}_n(t) - \Lambda(t)) \exp(-\Lambda(t)) \\ &\quad + \frac{1}{2} (\hat{\Lambda}_n(t) - \Lambda(t))^2 \exp(-\Lambda(t) - \theta_n^t (\hat{\Lambda}_n(t) - \Lambda(t))) \end{aligned}$$

avec $\theta_n^t \in]0,1[$. On peut aussi écrire :

$$\begin{aligned} \hat{S}_n(t) - S(t) &= (\hat{\Lambda}_n(t) - \Lambda(t)) S(t) \\ &\quad + \frac{1}{2} (\hat{\Lambda}_n(t) - \Lambda(t))^2 S(t) \exp(-\theta_n^t (\hat{\Lambda}_n(t) - \Lambda(t))) \\ &= S(t) \int k_1(t,s) d(H_n^1 - H^1)(s) \\ &\quad + S(t) \int k_2(t,s) d(H_n^0 - H^0)(s) + S(t) R_n(t) \\ &\quad + \frac{1}{2} (\hat{\Lambda}_n(t) - \Lambda(t))^2 S(t) \exp(-\theta_n^t (\hat{\Lambda}_n(t) - \Lambda(t))) \end{aligned}$$

Donc, les courbes d'influence de $\hat{S}_n$, sont bien $S(t)k_1(t,s)$ et $S(t)k_2(t,s)$, et le reste R_n^0 s'écrit

$$R_n^0(t) = S(t) R_n(t) + \frac{1}{2} (\hat{\Lambda}_n(t) - \Lambda(t))^2 S(t) \exp(-\theta_n^t (\hat{\Lambda}_n(t) - \Lambda(t)))$$

Or, si b vérifie l'hypothèse (H_0), alors :

$$\sup_{t\in[0,b]} \mid \hat{\Lambda}_n(t) - \Lambda(t) \mid \xrightarrow[n\to\infty]{ps} 0$$

(même démonstration que le théorème 2, ci dessous).
Donc, $\exists \alpha > 0$ / $\sup_{t\in[0,b]} \exp(-\theta_n^t(\hat{\Lambda}_n(t) - \Lambda(t))) < \alpha$

On en déduit alors, que

$$\mid R_n^0(t) \mid \leq \mid R_n(t) \mid + \frac{1}{2}(\hat{\Lambda}_n(t) - \Lambda(t))^2 S(t)\alpha$$

Cherchons ensuite une majoration de $(\hat{\Lambda}_n(t) - \Lambda(t))^2 S(t)$.

$$\begin{aligned}
(\hat{\Lambda}_n(t) - \Lambda(t))^2 S(t) &= S(t)[\int k_1(t,s)d(H_n^1 - H^1)(s) + \\
&\quad \int k_2(t,s)d(H_n^0 - H^0)(s) + R_n^0(t)]^2 \\
&\leq 3S(t)[\int k_1(t,s)d(H_n^1 - H^1)(s)]^2 + \\
&\quad 3S(t)[\int k_2(t,s)d(H_n^0 - H^0)(s)]^2 \\
&\quad +3S(t)R_n^0(t)^2
\end{aligned}$$

Le premier terme, peut être majoré de la manière suivante :

$$\begin{aligned}
S(t)[\int k_1(t,s)d(H_n^1 - H^1)(s)]^2 &= S(t)[\int_0^\infty (\int_0^{s\wedge t} \frac{h(u)}{1-L(u)}du \\
&\quad + \frac{\mathbf{1}_{\{s\leq t\}}}{1-L(s)})d(H_n^1 - H^1)(s)]^2 \\
&= S(t)[-(H_n^1 - H^1)(0) \\
&\quad - \int_0^t (\frac{h(s)}{1-L(s)} + \frac{L'(s)}{(1-L(s))^2}) \\
&\quad (H_n^1 - H^1)(s)ds]^2
\end{aligned}$$

(Cette égalité est obtenue en utilisant une integration par partie). On peut donc majorer ce dernier terme par :

$$|| H_n^1 - H^1 ||_\infty^2 \frac{C_1}{(1-L(s))^2}$$

(C_1 est une constante positive), puisque:

$$\int_0^t \frac{h(s)}{(1-L(s))} ds \leq \frac{1}{1-G(t)} \int_0^t \frac{f(s)}{(1-F(s))^2} ds$$

et

$$\int_0^t \frac{f(s)}{(1-F(s))^2} ds = \frac{1}{1-F(t)} - 1$$

donc,

$$\int_0^t \frac{h(s)}{(1-L(s))} ds \leq \frac{1}{1-L(t)}$$

ainsi que,

$$\int_0^t \frac{L'(s)}{(1-L(s))^2} ds = \frac{1}{1-L(t)} - 1 \leq \frac{1}{1-L(t)}$$

De la même manière, on montre que:

$$S(t)[\int k_2(t,s) d(H_n^0 - H^0)(s)]^2 \leq || H_n^0 - H^0 ||_\infty^2 \frac{C_2}{(1-L(s))^2}$$

(C_2 est une constante positive)
En utilisant, maintenant la majoration (2) de R_n, on obtient:

$$\sup_{t\in[0,b]} | R_n^0(t) | \leq C'_b \{ \sum_{i=0}^{1} || H_n^i - H^i ||_\infty^2 + \frac{1}{n} \}$$

(Sous l'hypothèse (H_0), C'_b est une constante qui dépend de b)

1.4 Convergence de l'estimateur de Kaplan-Meier.

1.4.1 Convergence uniforme p.s.

Parmi les nombreux auteurs qui se sont interessés à l'étude de consistance de l'estimateur de Kaplan-Meier, on peut citer, N. Breslow et J. Crowley (1974), V. Susarla et J. Van Rysin (1979), A. Földes, L. Rejtö et B. B. Winter (1980), A. Földes et L. Rejtö (1981) et N. Reid (1981). Nous allons

alors, rappeler certains résultats de ces auteurs et donner quelques théorèmes déduits des travaux de N. Reid (1981) et la décomposition (3).

A. Földes, L. Rejtö et B. B. Winter (1980), ont montré l'inégalité suivante :

Théorème 1 *Il existe des constantes positives a, b, c indépendantes de F, G, ε, T, λ et n, tel que si $T > 0$ et $L(T^-) < 1$, alors pour tout $\varepsilon > 0$, pour tout n et λ, $0 < \lambda < 1 - L(T^-)$, on a :*

$$P(\sup_{0\leq t\leq T} \mid \hat{S}_n(t) - S(t) \mid > \varepsilon) \leq a\lambda\varepsilon^{-2}\exp(-n(\lambda\varepsilon)^4 b) + c(\lambda\varepsilon)^{-1}(1-L)^n$$

Un résultat du même genre qu'on montre à partir de la décomposition (3), et qu'on utilise dans notre travail, est le suivant :

Théorème 2 *Si b vérifie l'hypothèse (H_0), alors*
$\forall\, \varepsilon > 0,\ \varepsilon < 1,$

$$P(\sup_{0\leq t\leq b} \mid \hat{S}_n(t) - S(t) \mid > \varepsilon) \leq Cn\exp(\frac{-n\varepsilon}{C'})$$

où C et C', sont deux constantes positives.

Preuve :

$$\sup_{t\in[0,b]} \mid \hat{S}_n(t) - S(t) \mid \leq \sup_{t\in[0,b]} \mid P_n^0(t) \mid + \sup_{t\in[0,b]} \mid R_n^0(t) \mid$$

et on montre que,

$$\sup_{t\in[0,b]} \mid P_n^0(t) \mid \leq C_1(\sum_{i=0}^{1} \mid\mid H_n^i - H^i \mid\mid)$$

avec C_1 une constante positive. (Majoration qui s'obtient de la même manière que celles dans la démonstration du Lemme 1). En utilisant le Lemme 1, on peut donc écrire :

$$\sup_{t\in[0,b]} \mid \hat{S}_n(t) - S(t) \mid \leq C_1(\sum_{i=0}^{1} \mid\mid H_n^i - H^i \mid\mid) + C_2(\sum_{i=0}^{1} \mid\mid H_n^i - H^i \mid\mid^2 + \frac{1}{n})$$

C_2 est une constante positive. On a donc :

$$\begin{aligned} P(\sup_{t\in[0,b]} \mid \hat{S}_n(t) - S(t) \mid > \varepsilon) &\leq P(C_1(\sum_{i=0}^{1} \| H_n^i - H^i \|) > \frac{\varepsilon}{2}) \\ &\quad + P(C_2(\sum_{i=0}^{1} \| H_n^i - H^i \|^2 + \frac{1}{n}) > \frac{\varepsilon}{2}) \\ &\leq Cn\exp(-\frac{n\varepsilon}{C'}) \end{aligned}$$

On utilise l'inégalité suivante :

$$(5) \qquad \forall \varepsilon > 0, \quad P(\| H_n^i - H^i \| > \varepsilon) \leq 8(n+1)\exp(-\frac{n\varepsilon^2}{32})$$

(cf. P. Polard(1984))

Un autre résultat du type loi du logarithme itéré, établi par A. Földes et J. Rejtö (1981), est donné dans le théorème ci dessous. On définit d'abord, pour toute fonction de répartition F :

$$T_F = \sup\{t, F(t) < 1\}$$

Théorème 3 *Si F et G, sont deux f. d. r. continues et telles que $T_F < T_G \leq +\infty$, alors*

$$\sup_{0\leq t<\infty} \mid \hat{S}_n(t) - S(t) \mid = O((\frac{\log\log n}{n})^{\frac{1}{2}}) \quad p.s.$$

Corollaire 1 *Si F et G, sont deux f. d. r. continues et si $G(T) < 1$, alors*

$$\sup_{0\leq t<T^*} \mid \hat{S}_n(t) - S(t) \mid = O((\frac{\log\log n}{n})^{\frac{1}{2}}) \quad p.s.$$

où $T^* = T \wedge T_F$

Remarques

1) En utilisant la majoration (4) et l'égalité:

$$\sqrt{n} \mid\mid H_n^i - H^i \mid\mid = O((\frac{\log\log n}{n})^{\frac{1}{2}}) \ , \ i = 0, 1$$

(cf. N. Reid (1981)), on a pour tout b vérifiant (H_0) (ce qui est vérifié si $F(b) < 1$ et $G(b) < 1$)

$$\sup_{t\in[0,b]} \mid \hat{S}_n(t) - S(t) \mid = O((\frac{\log\log n}{n})^{\frac{1}{2}}) \quad p.s.$$

2) Pour que l'estimateur $\hat{S}_n$ converge presque sûrement sur $\mathbf{R}^+$, on a vu théorème 3, que l'hypothèse: $T_F < \infty$ doit être vérifiée, ce qui s'explique par l'absence d'observations après l'instant T_F. Malheureusement ceci ne caractérise pas les lois usuelles dans les modèles de durées. La difficulté de montrer ce résultat pour $T_F = +\infty$, vient du fait qu'on ne peut pas contrôler l'écart entre $\hat{S}_n(t)$ et $S(t)$, pour de grandes valeurs de t. D'où l'idée de choisir une suite b_n qui tend vers l'infini et de montrer que le sup sur $[0, b_n]$ de $\mid \hat{S}_n(t) - S(t) \mid$ tend vers zéro, en utilisant la décomposition (3) et en cherchant une majoration du reste R_n^0 sur $[0, b_n]$ du type (4).

1.4.2 Variance de l'estimateur de Kaplan-Meier

A partir de la décomposition (3), on peut aussi déterminer l'équivalent asymptotique de la variance de $\hat{S}_n(t)$. Dans le Lemme 2 (cf. chapitre III), on calcul la variance de tout estimateur qui s'écrit sous la forme: $\int_0^b m(u)d\hat{F}_n(u)$. Comme cas particulier, on peut prendre $b = t$ et la fonction m égale à la constante 1. On en déduit alors:

Théorème 4 *Si t vérifie l'hypothèse* (H_0), *alors:*

$$V(\hat{S}_n(t)) = \frac{1}{n} S(t)^2 \int_0^t \frac{h(s)}{1 - L(s)} ds + o(\frac{1}{n})$$

En utilisant ensuite le Lemme 1 (cf. M. Delecroix et O. Yazourh (1992)) on montre, de la même manière, le résultat suivant:

Théorème 5 *Si t vérifie l'hypothèse* (H_0), *alors:*

$$V(\hat{\Lambda}_n(t)) = \frac{1}{n}\int_0^t \frac{h(s)}{1-L(s)}ds + o(\frac{1}{n})$$

1.4.3 Normalité asymptotique

N. Breslow et J. Crowley (1974), ont étudié le comportement asymptotique du processus $\sqrt{n}(\hat{S}_n - S)$. Ils ont, montré le résultat suivant :

Théorème 6 *Si la survie X^0 et la censure C sont indépendantes et si F et G n'ont aucune discontinuité commune, alors :*

$$\sqrt{n}(\hat{S}_n - S) \xrightarrow[n\to\infty]{\mathcal{L}} Z$$

où Z est un processus gaussien centré, de fonction de covariance :

$$Cov(Z(t_1), Z(t_2)) = S(t_1)S(t_2)\int_0^{t_1\wedge t_2} \frac{h(s)}{1-L(s)}ds$$

Un autre résultat, concernant la normalité asymptotique, qu'on peut déduire de la décomposition (3), est la loi limite de la quantité J_n, définie par :

$$J_n = \frac{\sqrt{n}(\hat{S}_n(x) - E(\hat{S}_n(x)))}{B_0}$$

où

$$B_0^2 = S(x)^2\int_0^x \frac{h(s)}{1-L(s)}ds$$

On montre alors:

Théorème 7 *si x vérifie l'hypothèse (H_0), alors :*

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

La démonstration de ce résultat se déduit facilement de celle du théorème 3 (cf. chapitre III).

Si on définit maintenant, A_n telle que:

$$A_n = \frac{\sqrt{n}(\hat{\Lambda}_n(x) - E(\hat{\Lambda}_n(x)))}{B_1}$$

où

$$B_1^2 = \int_0^x \frac{h(s)}{1 - L(s)} ds$$

alors on montre

Théorème 8 *Si x vérifie l'hypothèse H_0, alors:*

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

Là aussi, la démonstration se déduit facilement de celle du théorème 2 (cf. M. Delecroix et O. Yazourh (1992))

2 Synthèse des résultats obtenus en nonparamétrique

2.1 Estimateur de la densité f

La technique classique de l'estimation fonctionnelle, consiste si on veut estimer la densité d d'une mesure $\mathcal{T}$, à se donner une suite de fonctions $K_n(.,.)$ de $\mathbf{R^2}$ dans $\mathbf{R}$, telle que:

$$d(x) = \lim_{n \longrightarrow \infty} \int K_n(x,t) d(t) dt$$

Dans le cas usuel de données non censurées, on approxime $\mathcal{T}$ par la mesure empirique $\mathcal{T}_n$, et comme $\int K_n(x,t)d(t)dt$ est égale à $\int K_n(x,t)d\mathcal{T}_n(t)$, on estime $d(x)$ par:

$$\hat{d}_n(x) = \int K_n(x,t) d\mathcal{T}_n(t) = \frac{1}{n} \sum_{i=1}^{n} K_n(x, X_i)$$

(cf. D. Bosq et J. P. Lecoutre (1987), "Théorie de l'estimation fonctionnelle")

Si on note maintenant, μ la mesure dont la fonction de répartition est F, alors l'estimateur donné par Kaplan-Meier est la mesure discrète

$$\hat{\mu}_n = \sum_{i=1}^{n} p_i \Delta_{X_i}$$

avec

$$p_i = \begin{cases} 0 & \text{si } \delta_i = 0 \\ \frac{1}{n-R_i+1} \prod_{j/X_j \leq X_{i-1}} (\frac{n-R_j}{n-R_j+1})^{\delta_j} & \text{si } \delta_i = 1 \end{cases}$$

On peut donc définir des estimateurs de la densité par :

$$\hat{f}_n(x) = \int K_n(x,t) d\hat{\mu}_n(t) = \sum_{i=1}^{n} p_i K_n(x, X_i)$$

Dans la quasi totalité des travaux déjà effectués, dans ce cadre, on utilisait la méthode du noyau, où les fonctions K_n sont définies par :

$$K_n(x,y) = \frac{1}{h_n} K(\frac{x-y}{h_n})$$

h_n est un nombre réel positif dépendant de n, et K est par exemple la densité gaussienne.

De la méthode du noyau, découle aussi la méthode de l'histogramme classique, qui dans un premier temps a été utilisée par A. Földes, L. Rejtö et B. B. Winter (1981).

Définition 1 *On considère l'intervalle* $[0,T]$, $T < \infty$, *pour tout entier* N *positif, soit*

$$0 < t_0^{(N)} < t_0^{(N)} < \ldots < t_{\nu_N}^{(N)} = T$$

Une partition de $[0,T]$ *en* ν_N *subintervalles* $\Delta_i^{(N)}$, *définis par :*

$$\Delta_i^{(N)} = \begin{cases} [t_{i-1}^{(N)}, t_i^{(N)}] & si \;\; 1 \leq i \leq \nu_N \\ \\ [t_{\nu_N - 1}^{(N)}, T] & si \;\; i = \nu_N \end{cases}$$

L'estimateur de la densité f du type histogramme, correspondant à cette partition est défini par:

$$\hat{f}_N(x) = \frac{\hat{F}_N(t_i^{(N)}) - \hat{F}_N(t_{i-1}^{(N)})}{t_i^{(N)} - t_{i-1}^{(N)}} \qquad si \;\; x \in \Delta_i^{(N)}$$

Estimateur de la probabilité de tomber dans l'intervalle $]t_i^{(N)}, t_{i-1}^{(N)}]$ sur la largeur de cet intervalle.

Les principaux résultats de convergence de $\hat{f}_N$ vers f, sont donnés au théorème suivant, que montrent A. Földes, L. Rejtö et B. B. Winter (1981). Ils supposent d'abord que $T \in \mathbf{R}^+$, que $F'(x)$ existe pour chaque $x \in [0,T]$ et que $f = F'$ sur $[0,T]$.

Théorème 9 *On suppose que f est continue sur $[0,T]$ et $L(T^-) < 1$*

(I) si $\max\{|\, \Delta_i^{(N)} \,| : \; 1 \leq i \leq \nu_N\} \longrightarrow 0$
et si $(\frac{N}{\log N})^{\frac{1}{4}} \min\{|\, \Delta_i^{(N)} \,| : \; 1 \leq i \leq \nu_N\} \longrightarrow \infty$, *quand* $N \longrightarrow \infty$ *alors*

$$\sup_{0 \leq x \leq T} |\, \hat{f}_N(x) - f(x) \,| \longrightarrow 0 \quad p.s.$$

(II) si f a une dérivée bornée sur $[0,T]$, et
$\frac{N^{\frac{1}{8}}}{(\log N)^{-\frac{1}{4}}} \max\{|\, \Delta_i^{(N)} \,| : \; 1 \leq i \leq \nu_N\} \longrightarrow 0$

et $\inf\{(\frac{N}{\log N})^{\frac{1}{8}} \min\{|\, \Delta_i^{(N)} \,| : \; 1 \leq i \leq \nu_N\}\} \longrightarrow 0$ *alors*

$$\frac{N^{\frac{1}{8}}}{(\log N)^{-\frac{1}{4}}} \sup_{0 \leq x \leq T} |\, \hat{f}_N(x) - f(x) \,| \longrightarrow 0 \quad ps$$

Les auteurs ci-dessus, étudient ensuite l'estimateur $\tilde{f}_n$ de la densité, construit par la méthode du noyau, tel que:

$$\tilde{f}_n(x) = \int \frac{1}{h_n} K(\frac{x-y}{h_n}) d\hat{F}_n(y)$$

Ils montrent alors, que si $F'(x)$ existe pour chaque $x \in \mathbf{R}$ et $f = F'$ sur $\mathbf{R}$, alors on a le théorème suivant:

Théorème 10 *On suppose que f est bornée, $G(T_F^-) < 1$, K est continue à droite et à variation bornée sur $\mathbf{R}$ et*

$$h_n \longrightarrow 0 \quad avec \quad h_n(\tfrac{n}{\log n})^{\frac{1}{8}} \longrightarrow \infty \quad quand \quad n \longrightarrow \infty$$

(I) Si f est continue au point x alors $\tilde{f}_n(x) \longrightarrow f(x)$ p.s.

(II) Si $-\infty \leq a < b \leq \infty$ et f est uniformement continue sur (a,b), alors pour chaque $c > 0$

$$\sup_{x \in (a+c, a-c)} | \tilde{f}_n(x) - f(x) | \longrightarrow 0 \quad p.s.$$

(III) Si $-\infty \leq a < b \leq \infty$ et f a une dérivée bornée sur (a,b), alors:

$$\sup_{x \in (a,b)} | \tilde{f}_n(x) - f(x) | \longrightarrow 0 \quad p.s.$$

Plus récemment S. H. Lo, Y. P. Mack et J. L. Wong (1989), ont étudié un estimateur de la densité, construit à partir d'une version modifiée Γ_n de l'estimateur $\hat{F}_n$ de Kaplan-Meier, telle que:

$$(6) \quad \Gamma_n(x) = \begin{cases} 1 - \prod_{i/X_{(i)} \leq x} (\frac{n-i+1}{n-i+2})^{\delta_{(i)}} & \text{si } x \leq X_{(n)} \\ \Gamma_n(X_{(n)}) & \text{si } x > X_{(n)} \end{cases}$$

(On suppose que la plus grande observation est non censurée)

Dans leur analyse, les auteurs ci-dessus, se basent sur une décomposition de $F_n - F$ (cf. S. H. Lo et K. Singh (1986)), du même genre que (3). Ils supposent d'abord que la densité f est continue au point x, $f(x) > 0$, G est continue en x et $L(x) < 1$. L'estimateur de la densité proposé, est construit par la méthode du noyau, tel que:

$$f_n(x) = \frac{1}{h_n} \int K(\frac{x-u}{h_n}) d\Gamma_n(u)$$

Les suppositions suivantes, concernant le noyau K, seront utilisées dans l'étude de f_n.

(k_1) K est une fonction de densité symétrique.

(k_2) K est à support compact $[-c, c]$.

(k_3) K est continue.

(k_4) K est à variation bornée.

S. H. Lo, Y. P. Mack et J. L. Wong (1989), montrent que l'estimateur f_n vérifie les résultats suivants :

Théorème 11 *On suppose que K satisfait $(k_1), \ldots, (k_4)$, h_n est telle que, $h_n \longrightarrow 0$ et $\log n^2/nh_n \longrightarrow 0$ quand $n \longrightarrow \infty$ et f est deux fois differentiable en x, alors :*

$$E(f_n(x)) = f(x) + h_n^2 \frac{f"(x)}{2} \int v^2 K(v)dv + o(\frac{1}{\sqrt{nh_n}} + h_n^2)$$

$$var(f_n(x)) = \frac{1}{\sqrt{nh_n}}(\frac{f(x)}{1 - G(x)}) \int K^2(v)dv + o(\frac{1}{nh_n})$$

Corollaire 2 *(i) On suppose que K satisfait $(k_1), \ldots, (k_4)$, h_n est telle que, $h_n \longrightarrow 0$ et $\log n^2/nh_n \longrightarrow 0$ quand $n \longrightarrow \infty$, alors*

$$(nh_n)^{\frac{1}{2}}[f_n(x) - E(f_n(x))] \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

avec $\sigma^2 = (f(x)/1 - G(x)) \int K^2(u)du$.

(ii) Si en plus, on suppose que f est deux fois continument différentiable en x, et h_n satisfait $h_n = o(n^{-1/5})$, alors

$$(nh_n)^{\frac{1}{2}}[f_n(x) - f(x)] \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

2.2 Estimation du taux de hasard

Dans une approche non paramétrique, on cherche à estimer la fonction $h(t) = f(t)/(1 - F(t))$, sans hypothèses "apriori" sur l'ensemble de lois à la quelle appartient la loi des X_i^0, à partir de l'échantillon observé.

En pratique plusieurs estimateurs de h, ont été proposés. Cette multiplicité se justifie, d'une part par la diversité des techniques d'estimation fonctionnelle, d'autre part par l'utilisation des trois définitions différentes de h, que nous rappelons ci-dessous.

- La première définition que nous avons déjà vu, est $h(t) = f(t)/(1 - F(t))$. On est donc amené à chercher des estimateurs de f et de F.

- On peut aussi écrire

$$h(t) = \frac{f(t)(1 - G(t))}{(1 - F(t))(1 - G(t)} = \frac{l^1(t)}{1 - L(t)}$$

l^1 est la densité de la mesure ν définie par:

$$\nu(A) = P[(X_i^0 \in A) \cap (\delta_i = 1)]$$

puisque

$$\begin{aligned} P[(X_i^0 \leq t) \cap (\delta_i = 1)] &= P[(X_i^0 \leq t) \cap (X_i^0 \leq C_i)] \\ &= \int_0^t f(x)\{\int_x^\infty g(y)dy\}dx \\ &= \int_0^t l^1(x)dx \end{aligned}$$

- Enfin on a:

$$\int_0^t h(x)dx = \int_0^t \frac{f(x)}{1 - F(x)}dx = -\log(1 - F(t)) = \Gamma(t)$$

h est donc la dérivée du taux de hasard cumulé, on peut alors, l'estimer à partir de $\hat{\Lambda}_n$.

Ces troix définitions, ont été utilisées par de nombreux auteurs, afin de construir des estimateurs de h. A partir de la première écriture, A. Földes, L. Rejtö et B. B. Winter (1981) ont construit l'estimateur $\hat{f}_n/(1 - \hat{F}_n + \frac{1}{n})$, où $\hat{f}_n$ est un estimateur de la densité, $\hat{F}_n$ est l'estimateur de Kaplan-Meier et le terme $\frac{1}{n}$ au dénominateur permet d'éviter une division par zéro. Deux estimateurs de la densité ont été utilisés, celui du type histogramme et celui du

type noyau. Dans le premier cas, les auteurs ci-dessus montrent le théorème suivant :

Théorème 12 *Soit* $\hat{h}_n^1 = \hat{f}_n/(1 - \hat{F}_n + \frac{1}{n})$. *Si on suppose que* f *est continue et que* $L(T) < 1$, *alors les assertions* (I) *et* (III) *du théorème 9, restent vérifiées si on remplace* $\mid \hat{f}_N(x) - f(x) \mid$ *par* $\mid \hat{h}_N(x) - h(x) \mid$

Dans le cas, où la densité f est estimée par la méthode du noyau, c'est à dire que $\tilde{f}_n(x) = \int \frac{1}{h_n} K(\frac{(x-t)}{h_n}) d\hat{F}_n(t)$, A. Földes, L. Rejtö et B. B. Winter (1981), établissent les résultats suivants :

Théorème 13 *Soit* $\hat{h}_n^2 = \tilde{f}_n/(1 - \hat{F}_n + \frac{1}{n})$. *Si on suppose que* f *est bornée,* $G(T_F^-) < 1$, K *est continue à droite et à variation bornée sur* $\mathbf{R}$, *et* $h_n \longrightarrow 0$ *avec,* $h_n(n/\log n)^{\frac{1}{8}} \longrightarrow \infty$

(I) *si* $x < T_F$ *et* f *est continue au point* x *alors* $\tilde{h}_n(x) \longrightarrow h(x)$ *p.s.*

(II) *si* f *est continue sur* $[0, T_F]$ *alors, pour toutes constantes positives* c *et* d :

$$\sup_{x \in [c, T_F - d]} \mid \tilde{h}_n(x) - h(x) \mid \longrightarrow 0 \quad p.s.$$

(III) *Si* f *est à variation bornée sur* $(0, T_F)$ *alors, pour toute constante positive* c :

$$\sup_{x \in (0, T_F - c]} \mid \tilde{h}_n(x) - h(x) \mid \longrightarrow 0 \quad p.s.$$

L'utilisation de la deuxième définition de h a été initiée par J. Blum et V. Susarla (1980). L'estimateur construit à partir de la formulation: $h = l^1/(1 - L)$ est

$$\hat{h}_n^3 = \frac{l_n^1(t)}{1 - L_n(t)}$$

où L_n est la fonction de répartition empirique de l'échantillon $(X_1, \ldots, X_n)$ et l_n^1 est un estimateur de la densité des $(X_i, \delta_i = 1)$. J. Blum et V. Susarla (1980), ont étudié le comportement asymptotique de l'estimateur $\hat{h}_n^3$, quand l_n^1 est l'estimateur de la densité l^1 construit par la méthode du noyau:

$$\hat{l}_n^1(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \delta_i K(\frac{x - X_i}{h_n})$$

L'estimateur du taux de hasard ainsi construit, converge presque sûrement et vérifie la normalité asymptotique.

En 1983, H. Ramlau-Hansen a étudié le même estimateur de h, défini ci-dessus, il a montré que si h est continue alors :

$E(\hat{h}_n^3(t)) \longrightarrow h(t)$ pour tout $t \in]0,1[$ et $G(t^-) < 1$ et

$$Var(\hat{h}_n^3(t)) = \frac{1}{nh_n} \frac{h(t)}{1 - G(t)} \exp\{\int_0^t h(u)du\} \int K^2(u)du + o(\frac{1}{nh_n})$$

Dans la troisième définition, h représente la dérivée du taux de hasard cumulé Λ. Donc, on peut utiliser les méthodes d'estimation fonctionnelle pour construire des estimateurs de h à partir de ceux de Λ. En effet, A. Tanner et W. H. Wong(1983), ont introduit l'estimateur de h défini par :

$$\begin{aligned} \hat{h}_n^4(x) &= \frac{1}{h_n} \int K(\frac{x - t}{h_n}) d\Lambda_n(t) \\ &= \frac{1}{h_n} \sum_{i=1}^{n} \frac{\delta_{(i)}}{n - i + 1} K(\frac{x - X_{(i)}}{h_n}) \end{aligned}$$

et obtenu, en régularisant $d\Lambda_n$ par convolution avec un noyau K, où Λ_n est l'estimateur empirique du taux de hasard integré, tel que :

$$\Lambda_n(x) = \sum_{X_{(i)} \leq x} \frac{\delta_{(i)}}{n - i + 1}$$

Cet estimateur est au fait une approximation de $\hat{\Lambda}_n(x) = -\log \hat{S}_n(x) = \sum_{X_{(i)} \leq x} \delta_{(i)} \log(1 - \frac{1}{n-i+1})$, puisqu'il suffit de remplacer $log(1 - \frac{1}{n-i+1})$ par $-1/(n - i + 1)$, pour obtenir $\Lambda_n(x)$. L'estimateur Λ_n a été étudié, notamment par W. Nelson (1969,1972) et O. Aalen (1983). A. Tanner et W. H. Wong (1983), ont étudié le comportement asymptotique de $\hat{h}_n^4$ et ont montré les

résultats résumés dans le théorème ci-dessous. On donne d'abord la définition suivante :

Définition 2 *Le noyau K est dit compatible avec F, si pour tout $M > 0$, il existe h_n suffisament petit, tel que $\frac{1}{h_n}K(\frac{y-x}{h_n})/(1-F(y))$ soit uniformément borné pour $\mid y - x \mid > M$.*

Théorème 14 *Si $n \longrightarrow \infty$, $h_n \longrightarrow 0$, et $nh_n \longrightarrow \infty$, alors :*

a) Si K est compatible avec F, $E(\hat{h}_n^4(x)) \longrightarrow h(x)$.

b) Si K est compatible avec F et G, alors :

$$Var(\hat{h}_n^4(x)) = \frac{1}{nh_n}\frac{h(x)}{1-F(x)}\int K^2(t)dt + o(\frac{1}{nh_n})$$

$$et\ (nh_n)^{\frac{1}{2}}(\frac{1-F(x)}{h(x)\int K^2(t)dt})^{\frac{1}{2}}(\hat{h}_n^4(x) - h(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

A partir de l'estimateur de la densité f_n, défini par S. H. Lo, Y. P. Mack et J. L. Wong (1989), ces auteurs ont construit l'estimateur $\hat{h}_n^4$ du taux de hasard, du type h_n^1, tel que :

$$\hat{h}_n^4 = \frac{f_n}{1-\Gamma_n}$$

avec Γ_n, l'estimateur de F défini en (6). Ils montrent alors, que $\hat{h}_n^4$ vérifie les propriétés asymptotiques suivantes :

Théorème 15 *Sous les hypothèses du théorème 11, on a :*

$$\begin{aligned} E[(\hat{h}_n^4(x) - E(\hat{h}_n^4(x)))^2] &= h_n^4\{\frac{f''(x)}{2(1-F(x))}\int v^2K(v)dv\}^2 \\ &\quad + \frac{1}{nh_n}\{\frac{h(x)}{1-L(x)}\int K^2(v)dv\} + o(h_n^4 + \frac{1}{nh_n}) \end{aligned}$$

Théorème 16 *On pose $\tau^2 = \frac{h(x)}{1-L(x)}\int K^2(v)dv$*

(i) Sous les hypothèses du corollaire 2, (i), quand $n \longrightarrow \infty$

$$(nh_n)^{\frac{1}{2}}[\hat{h}_n^4(x) - E(\hat{h}_n^4(x))] \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\tau^2)$$

(ii) Sous les hypothèses du corollaire 2, (ii), quand $n \longrightarrow \infty$

$$(nh_n)^{\frac{1}{2}}[\hat{h}_n^4(x) - h(x))] \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\tau^2)$$

3 Méthode proposée et commentaire

3.1 La méthode des fonctions orthogonales

Dans les travaux cités précédemment, nous avons vu que la méthode d'estimation fonctionnelle utilisée, est celle du noyau. Seul D. K. Kimura (1972) et N. L. Hjort (1985), à notre connaissace ont proposé la méthode alternative des "fonctions orthogonales". Il s'agit au fait d'une approche semi paramétrique, puisqu'on se ramène dans cette méthode à l'estimation d'un nombre fini de paramètres de manière globale, c'est à dire on utilise toutes les observations de l'echantillon pour estimer chaque paramètre. Cette méthode semble donc plus robuste de ce point de vue par rapport à celle du noyau, on évite les problèmes de localisation, en particulier dans les zones où peu d'observations non censurées apparaissent. Notre travail, sera donc consacré à l'etude de l'estimation de la densité et du taux de hasard, par la méthode des fonctions orthogonales. Dans la première partie, on construit un estimateur du taux de hasard à partir de Λ_n l'estimateur du taux cumulé Λ, et on étudie les convergences L^2 et ponctuelles. Dans la deuxième partie, on construit de la même manière un estimateur de la densité à partir de $\hat{F}_n$, l'estimateur de Kaplan-Meier. On en déduit ensuite un estimateur du taux de hasard: $\hat{f}_n/(1-\hat{F}_n+\frac{1}{n})$. Les résultats de convergences établis pour l'estimateur de la densité $\hat{f}_n$ et l'estimateur $\hat{F}_n$, permettent enfin, d'étudier le comportement asymptotique de cet estimateur du taux de hasard.

En conclusion, on peut dire que la méthode des fonctions orthogonales, donne des estimateurs comparables à ceux construits par la méthode du noyau et dont l'efficacité peut être trés satisfaisante si on arrive à choisir la base la mieux adaptée au paramètre fonctionnel à estimer, problème qui devrait faire l'objet de travaux ultérieurs.

3.2 Commentaire général

Pour étudier les estimateurs construits par la méthode des "fonctions orthogonales", on utilise essentiellement la décomposition (9) (cf. M. Delecroix et O. Yazourh (1992)), et la décomposition (15), (cf. chapitre III) des quantités:

$$\int m(t)d\Lambda_n(t) - \int m(t)d\Lambda(t)$$

et

$$\int m(t)d\hat{F}_n(t) - \int m(t)dF(t)$$

où $\int m(t)d\Lambda_n(t)$ et $\int m(t)d\hat{F}_n(t)$, sont des estimateurs de h et de f. En remplaçant la fonction m par un noyau K et en utilisant les mêmes techniques, on peut étudier les estimateurs construits par la méthode du noyau et retrouver les résultats de convergence déjà montrés. Cette étude peut donc être étendue au cas général, c'est à dire aux estimateurs construits par la méthode du noyau généralisé (cf. D. Bosq et J .P. Lecoutre (1987)).

Un autre estimateur du type $\hat{h}_n^3$ proposé par J. Blum et V. Susarla (1980), peut être construit, en estimant la densité des v.a. $(X_i, \delta_i = 1)$ par la méthode des fonctions orthogonales, sans utiliser l'estimateur de Kaplan-Meier. Il serait donc intéressant de comparer ces deux estimateurs du taux de hasard à ceux dont la construction est basée sur l'estimateur de Kaplan-Meier.

Enfin pour faciliter le choix de l'estimateur en fonction du paramètre fonctionnel à estimer, il est nécessaire de compléter par la suite ce travail par une étude pratique basée sur des simulations.

Références

[1] O. Aalen (1978) : *Nonparametric estimator of partial transition probabilities in multiple decrement models.* Ann. Stat. 6, p 534-545.

[2] J. Begun, W. Hall, W. Huang, J. Wellner (1983) : *Information and asymptotic efficiency in parametric-nonparametric models.* Ann. Stat. 11, p 432-452.

[3] J. Blum, V. Susarla (1980) : *Maximal deviation theory of density and failure rate function estimates based on censored data, In Multivariate Analysis.* V, ed. par P. Krishnaiah, 213-222, North-Holland, Amsterdam.

[4] D. Bosq, J. P. Lecoutre (1987) : *Théorie de l'estimation fonctionnelle.* Paris, Economica.

[5] N. Breslow, J. Crowley (1974) : *A large sample study of the life table and product limit estimates under random censorship.* ann. of stat. 2, (1974), p 437-453.

[6] D. R. Cox, D. Oakes (1984) : *Analysis of survival data.* Chapman and Hall.

[7] M. Delecroix, O. Yazourh (1992) : *Estimation non paramétrique du taux de hasard en présence de censures droites: méthode des fonctions orthogonales.* Statistique et analyse des données, à paraître.

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

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

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

[11] J. D. Kalbfleisch, R. L. Prentice (1980) : *The statistical analysis of failure time data.* New-York, Wiley.

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

[13] D. K. Kimura (1972) : *Fourier series methods for censored data.* Thèse soutenue à l'Université de Washington.

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

[15] H. Los, K. Singh (1986) : *The product-limit estimator and the bootsrap. Some asymptotic representations.* Prob. Th. and Rel. Fields 71, p 455-465.

[16] J. Mielniczuk (1985) : *Properties of some kernelestimators and of a density function for censored ata.* Periodica Math. Hungarica, vol. 16, (2), p 69-80.

[17] W. B. Nelson (1972) : *Theory and applications of hazard plotting for censored failure data.* Thechnometrics 14, p 945-966.

[18] A. V. Peterson (1977) : *Expressing the Kaplan-Meier estimate as afunction of empirical subsurvival functions.* J.A.S.A. 72, p 854-858.

[19] D. Pollard (1984) : *Convergence of stochstic processes.* Springer Verlag.

[20] H. Ramlau-Hansen (1983) : *Smoothing counting processes intensities by means of kernel functions.* Ann. Stat. 11, p, 453-466.

[21] N. Reid (1981) : *Influence functions for censored data.* The annals of stat. vol. 9, n1, p 78-92.

[22] V. Susarla, J. Van Ryzin (1979) : *Large sample theory for survival curve estimators under variable censoring.* Optimization methods in statistics , p 475-508, New York, Academic Press.

[23] A. Tanner, W. H. Wong (1983) : *The estimation of the hazard function from randomly censored data by the kernel method.* The annals of statistics, vol. 11 n3, p 989-993.

CHAPITRE II

ESTIMATION NON PARAMETRIQUE DU TAUX DE HASARD EN PRESENCE DE CENSURES DROITES : LA METHODE DES FONCTIONS ORTHOGONALES

M. DELECROIX * et O. YAZOURH **

RESUME

On définit un estimateur non paramétrique du taux de hasard de variables aléatoires positives soumises à des censures droites. Cet estimateur, original, est construit par la méthode des fonctions orthogonales. On démontre sa convergence asymptotique en divers sens stochastiques.

Classification AMS 62 G05
STMA 04 180

Mots clés

Données censurées, estimateur non paramétrique du taux de hasard, méthode des fonctions orthogonales, estimateur de Kaplan-Meier, modèle à censures aléatoires.

NON PARAMETRIC INFERENCE FOR FAILURE RATES WITH RANDOM RIGHT CENSORSHIP : ORTHOGONAL FUNCTIONS METHOD

ABSTRACT

Let h be the common hazard-rate of right censored lifetimes. We define a new nonparametric estimator of h, based on the orthogonal functions method, and prove its pointwise and L^2 asymptotic consistencies.

Keywords

Censored data, nonparametric hazard rate estimation, Kaplan-Meier estimator, orthogonal functions methods, random censorship model

---

* GREMAQ - (URA 947) Université des Sciences Sociales Toulouse 1,
Place Anatole France
31042 TOULOUSE CEDEX
L.S.P., (URA DO745) Université Paul Sabatier, 118 route de Narbonne,
31062 TOULOUSE CEDEX

** Laboratoire de Statistique et Probabilité - U.F.R. de Mathématiques pures et
appliquées - U.S.T.L.F.A.
59655 - VILLENEUVE D'ASCQ CEDEX

I - INTRODUCTION

L'estimation fonctionnelle dans les modèles de durée a désormais une longue histoire : on peut dater de Kaplan-Meier (1958) la première étude sur le sujet. La spécificité des problèmes statistiques à résoudre dans ce cadre provient évidemment de la présence des "censures" apportées aux observations de la variable d'intérêt par les conditions de l'expérience. Divers choix de modèles, dépendant de la nature des censures, sont possibles à ce niveau : nous adopterons dans ce travail celui d'Efron (1967).

Appelons X_i^0, $i \geq 1$, les variables aléatoires représentatives des durées de vie que l'on veut étudier (temps de chômage, de survie, de non-défaillance, etc...). Elles sont à valeurs dans $\mathbb{R}^+$, i.i.d, on appelle 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 (droites) apportées aux durées de vie d'intérêt (dates d'échantillonnage etc...). On suppose que les C_i forment une suite de variables réelles i.i.d, à valeurs dans $\mathbb{R}^+$, 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 X_i 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).

Dans le cadre de ce modèle ont été étudiés des estimateurs de F, du "taux de hasard cumulé" Λ ($\Lambda = - \text{Log}\,(1 - F)$), de la densité f des X_i^0, enfin de leur taux de hasard h ($h = f/(1 - F) = \Lambda'$). On peut citer, sans souci d'exhaustivité, des travaux récents comme ceux de K. E. Gneyou (1990) ou G. Gregoire (1991), et renvoyer à Droesbeke et al (1990) chapitre 4, pour une synthèse de résultats obtenus sur le sujet avant 1988.

Dans les travaux concernant f et h, la méthode d'estimation fonctionnelle usuelle est celle du noyau. Quelques auteurs ont néanmoins déjà utilisé la méthode alternative des "fonctions orthogonales". Kimura (1972) semble être le pionner en ce domaine, mais sans aborder le problème du devenir asymptotique de l'estimateur. Hjört (1985, 1991) propose un estimateur de h déduit du classique estimateur de Nelson Aalen (Nelson (1972), Aalen (1978)), sans en étudier systématiquement les propriétés de convergence. Enfin Mc Keagne (1986), sous le nom d'"estimateur sieves", propose aussi de projeter un estimateur empirique sur une famille de fonctions orthogonales.

Dans la lignée de ces travaux nous étudions, dans ce papier, les propriétés de convergence, L^2 et ponctuelle, d'un estimateur h_n de h défini par la méthode des fonctions orthogonales à partir d'un estimateur classique Λ_n de Λ. Ces convergences établies, nous conclurons l'article en comparant les performances de h_n avec celles des estimateurs non paramétriques qu'on avait jusque là proposés, afin de justifier l'emploi de h_n dans des situations pratiques.

Le second paragraphe sera consacré à la définition de h_n et à l'exposé des principaux résultats de convergence. Nous discuterons les propriétés obtenues dans le troisième paragraphe, le quatrième contiendra les démonstrations, le cinquième la bibliographie.

II - RESULTATS DE CONVERGENCE

2.1) La fonction h n'étant pas de carré intégrable sur $\mathbb{R}^+$ pour les lois de survie usuelle (la loi exponentielle par exemple), nous construirons un estimateur de la restriction de h à un intervalle [o,b], la continuité de h (donc celle de f) suffisant à obtenir son appartenance à $L^2([o,b])$. Ce point n'implique pas de restriction pratique : on peut toujours, pour réaliser l'estimation, choisir b supérieur au plus grand des X_i non censurés qu'on observe et même beaucoup plus grand, ce qui permet d'estimer h sur un intervalle "utile" suffisant au praticien.

b étant désormais fixé, nous supposerons d'autre part vérifiée l'hypothèse usuelle :
H0) : $P[X_j > b, \delta_j = 0] > 0$ et $P[X_j > b, \delta_j = 1] > 0$.
H0) assure simplement que l'on puisse obtenir des observations, censurées ou non, à droite de la valeur b, c'est-à-dire en pratique que ni la loi des X_i, ni celle des censures C_i ne sont concentrées sur un intervalle fini, ce qui est le cas usuel du modèle d'Efron.

2.2) Sans distinguer plus avant h et sa restriction à [0,b], supposons donc que $h \in L^2([0,b])$ et que (e_i) , $i \geq 1$ constitue une base orthonormale de cet ensemble. Alors h s'écrit $\sum_1^\infty a_i e_i$; avec :

$$a_i = \int_0^b e_i(x)\, h(x)\, dx = \int_0^b e_i \, d\,\bar{\Lambda}$$

où $\bar{\Lambda}$ est la mesure de densité h (i. e. de fonction de répartition Λ) sur [0,b].

L'estimateur h_n est alors défini , q(n) étant une suite croissante d'entiers, par :

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

$$\hat{a}_i = \int_0^b e_i \, . \, d(\bar{\Lambda}_n)$$

où $\bar{\Lambda}_n$ est une mesure empirique approximant $\bar{\Lambda}$; c'est la mesure de fonction de répartition Λ_n égale à - Log (1 - F_n), où F_n est l'estimateur de Kaplan-Meier de F (voir Droesbeke et al (1988) pour une synthèse). Pratiquement $\bar{\Lambda}_n$ met en chaque X_i non censuré la masse $- \text{Log} \left(1 - \frac{1}{n - R_i + 1}\right)$ où R_i est le rang de X_i dans l'échantillon ordonné. Dès lors

$$\hat{a}_j = - \sum_{i/x_i<b} \delta_i \, . \, e_j (X_i) \, . \, \text{Log} \left[1 - \frac{1}{n - R_i + 1} \right]$$

En posant $K_n (x, t) = \sum_{i=1}^{q(n)} e_i(x) \, e_i(t)$, on a donc :

$$h_n (x) = \int_0^b K_n (x, t) \, d \, \bar{\Lambda}_n (t) = - \sum_{i/x_i<b} \delta_i \, . \, K_n (X_i , x) \, . \, \text{Log} \left[1 - \frac{1}{n - R_i + 1} \right]$$

On notera que Tanner et Wong (1983) ont étudié le même type d'estimateur, s'écrivant $\int K_n^* (x, t) \, d \, \Lambda_n^* (t)$, K_n^* correspondant cette fois à la méthode du noyau et Λ_n^* dérivant de l'estimateur de Nelson-Aalen (Nelson (1972), Aalen (1978)) de Λ, asymptotiquement équivalent à $\bar{\Lambda}_n$.

2.3) Comme l'appartenance de h à L^2 [0,b] est la condition sine qua non de construction de h_n, nous étudierons d'abord la convergence de h_n vers h au sens de la norme de cet espace, notée désormais : $\| \ \|_{L^2[0,b]}$. Cette convergence sera obtenue sous l'hypothèse générale suivante, vérifiée par les bases classiques de $l^2[0,b]$:

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].
En posant alors, pour toute fonction réelle m dérivable :

$$(1)\quad N(m,b) = \left[|m(b)| + \int_0^b |m'(t)|\, dt \right]$$

nous pouvons définir la quantité D_n, caractéristique de la base (e_i), par :

$$(2)\quad D_n = \sum_1^{q(n)} N^2(e_i,b) = \sum_1^{q(n)} \left[|e_i(b)| + \int_0^b |e_i'(t)|\, dt \right]^2$$

et énoncer :

Théorème 1 : Soit b un réel positif tel que H_0 soit vérifiée ,(e_i) une base de L^2 [0,b] telle que H_1 le soit.

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

nous aurons : $\lim_{n\to\infty} E\left(\left| |h_n - h| \right|^2_{L^2[0,b]} \right) = 0$.

Si maintenant $q(n) \uparrow \infty$, de telle sorte que pour tous γ_1, et γ_2 strictement positifs les séries de termes généraux $q(n) \exp\{-\gamma_1 n / q(n)\}$ et $n\left\{ \sum_{i=1}^{q(n)} \exp\left[-\gamma_2 n/\sqrt{q(n)}.N(e_i,b) \right] \right\}$ convergent,

on aura $\lim_{n\to\infty}$ p.s. $\left[\left| |h_n - h| \right|^2_{L^2[0,b]} \right] = 0$.

2.4) Une base classique de $L^2[0,b]$ est constituée par les fonctions trigonométriques, définies par $e_1(t) = 1/\sqrt{b}$ et ($k \geq 1$, $t \in [0,b]$) :

$$(3)\quad e_{2k}(t) = \sqrt{\frac{2}{b}} \cdot \sin\left[\frac{2k\pi}{b} \left(t - \frac{b}{2} \right) \right], \quad e_{2k+1}(t) = \sqrt{\frac{2}{b}} \cos\left[\frac{2k\pi}{b} \left(t - \frac{b}{2} \right) \right]$$

Pour ces fonctions les hypothèses du théorème 1 sont vérifiées, le majorant M étant égal à $\sqrt{2/b}$, et de surcroît on a :

$$N(e_i,b) = O\left\{ \int_0^b \left| e_i'(t) \right| dt \right\} = O\ (i)$$

On peut alors énoncer le corollaire suivant, dont la démonstration découle directement du théorème, 1 et de l'égalité précédente :

Corollaire : Si la base (e_i) est celle des fonctions trigonométriques, la convergence vers 0 de $\left|\left| h_n - h \right|\right|^2_{L^2[0,b]}$ est assurée

* en moyenne si $q(n) \uparrow \infty$, $q(n) / n \to 0$, $(q(n))^3 (\text{Log}n)^4/n^2 \to 0$

* p.s. si $\forall\ \gamma\ ,\ \gamma > 0$, $\sum_1^\infty n\ q(n) \exp\left\{ - \gamma\ n / (q(n))^{3/2} \right\} < \infty$

(et en particulier, dans les deux cas, si q(n) est la partie entière de $n^{2/3} / (\text{Log}n)^{4/3+\alpha}$, $\alpha > 0$)

2.5) Pour un x quelconque de]0,b[posons alors :

$$(4) \qquad A_n = \sqrt{n}\ (h_n(x) - E\ (h_n(x))) / B_n$$

où $B_n^2 = \int_0^b K_n^2(x,t)\ \frac{h(t)}{1 - L(t)}\ dt$

Pour obtenir la normalité asymptotique de A_n, nous devons imposer à la suite q(n) et à la base (e_i) la condition suivante :

H2) : $q(n) / \sqrt{n}.B_n = o(1)$ et $N(\bar{K}_n,b)\ (\text{Log}n)^2 / B_n.\ \sqrt{n} = o(1)$, lorsque $n \to \infty$.

Ici, x étant fixé, $\bar{K}_n$ représente la fonction : $t \to K_n(x,t)$

On obtient alors :

Théorème 2

Sous les conditions H_0), H_1) et H_2) on a : $A_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, lorsque $n \to \infty$.

∎

2.6) Remarques

Si la quantité q(n) reste bornée, H_2) est automatiquement vérifiée ($N(\bar{K}_n,b)$) et B_n sont constantes , donc $A_n \to \mathcal{N}(0,1)$. Il n'y a d'ailleurs dans ce cas qu'à utiliser le théorème central limite standard en lieu de la démonstration du §IV.L'intérêt du résultat reste limité puisqu'alors la quantité $\overline{h_n}(x) - h_n(x) = \sum_{q(n)+1}^{\infty} a_i e_i(x)$ ne tend pas vers 0 ...

Dans le cas général, $q(n) \uparrow \infty$, mais il est difficile d'obtenir la loi limite de $A_n^* = (h_n(x) - h(x)) / \sigma(h_n(x))$. Il est facile (voir 4, § IV) de montrer que sous H_2) $V(h_n(x))$ est équivalent à B_n^2 / n mais $\sqrt{n}\left(\overline{h_n}(x) - h(x)\right) / B_n$ qui est $O\left[\sqrt{n}\left(\sum_{q(n)+1}^{\infty} a_i\right) / B_n\right]$ dépend entièrement de la série des coefficients a_i : sans hypothèse précise sur celle-ci, on ne peut obtenir la convergence à 0 de ce terme, donc la loi limite cherchée.

Si la base choisie est celle des fonctions trigonométriques on peut montrer (voir 5, § IV) qu'existe une constante A, A > 0, telle que $B_n^2 \geq A.q(n)$, pour n grand. Comme $N(\bar{K}_n,b) = O\left[(q(n))^2\right]$, H2) est vérifée en ce cas dès que $q(n)^{3/2}[\mathrm{Log}(n)]^2/\sqrt{n}$ tend vers zéro.

2.7) Dans les résultats qui précèdent, la convergence de h_n vers h résulte évidemment de la convergence uniforme de Λ_n vers Λ (Breslow - Crowley 1974), l'essentiel des démonstrations concernant le comportement asymptotique de quantités du type : $\int_0^b e_i(t)\,d(\bar{\Lambda}_n(t)-\bar{\Lambda}(t))$

Elles se basent pratiquement sur le lemme 1 ci-dessous, qui montre que ces quantités s'écrivent sous forme d'une somme de variables indépendantes, à un reste asymptotiquement négligeable près, à partir d'une évaluation de $\Lambda_n(t)-\Lambda(t)$ introduite par

N.Reid (1981).En utilisant la notion de "courbes d'influence", cet auteur a montré exactement que :

$$(5) \quad \forall\ t,\ t \geq 0,\ \Lambda_n(t) - \Lambda(t) = P_n(t) + R_n(t)$$

où

$$(6) \quad n.P_n(t) = \sum_{i=1}^{n} \left\{ I_{\{\delta_i = 1\}}\, k_1(t, X_i) + I_{\{\delta_i = 0\}}\, k_2(t, X_i) \right\}$$

avec

$$\begin{cases} k_1(t, s) = \displaystyle\int_0^{s \wedge t} \frac{h(u)}{1 - L(u)}\, du - \frac{I_{\{s \leq t\}}}{1 - L(s)} \\ k_2(t, s) = \displaystyle\int_0^{s \wedge t} \frac{h(u)}{1 - L(u)}\, du \end{cases}$$

($s \wedge t = \min(t, s)$, k_1 et k_2 sont les "courbes d'influence" de Λ). Le reste R_n se majore uniformément sur tout intervalle $[0, T]$. Il existe ainsi un réel C_T (variable avec T) tel que

$$(7) \quad \sup_{[0,T]} |R_n(t)| \leq C_T \left\{ \left[\sum_{i=0}^{1} \left|\left| H_n^i - H^i \right|\right| \right]^2 + \frac{1}{n} \right\}$$

où $H^i(t) = P(X_j \leq t,\ \delta_j = i)$ et $H_n^i(t) = \frac{1}{n} \sum_{j=1}^{n} I_{\{X_j \leq t,\ \delta_j = i\}}$, pour $i = 0,1$ et

$\left|\left| F \right|\right| = \sup_{0 \leq x \leq T} F(x)$. La majoration est valable dès que l'on a :

$$P(X_i > T,\ \delta_i = 0) > 0 \quad \text{et} \quad P(X_i > T,\ \delta_i = 1) > 0$$

(cf. Mielniczuk (1983) pour une synthèse et mise en forme des résultats de Reid), d'où l'hypothèse H0) introduite.

Soit alors m une application de $\mathbb{R}^+$ dans $\mathbb{R}^+$, supposée continuement différentiable, b un réel positif tel que $L(b)<1$.

Si nous posons :

$$C(m,b) = \int_0^b m(t)\ d\bar{\Lambda}_n(t) - \int_0^b m(t)\ d\bar{\Lambda}(t)$$

et

$$(8) \qquad Z_i = \int_0^{X_i \wedge b} m(t) \frac{h(t)}{1 - L(t)} dt - I_{\{(\delta_i = 1) \cap (X_i \leq b)\}} \frac{m(X_i)}{1 - L(X_i)}$$

nous obtenons :

Lemme 1

Les Z_i sont des variables centrées, de variance égale à $\int_0^b m^2(t) \frac{h(t)}{1 - L(t)} dt$, et on peut écrire

$$(9) \qquad C(m,b) = \frac{1}{n} \sum_1^n Z_i + R_n^*$$

où

$$(10) \qquad \left|R_n^*\right| \leq N(m,b) \cdot \left\{ \sup_{t \in [0,b]} \left|R_n(t)\right| \right\}$$

Dès lors on aura :

$$\begin{cases} E(C(m,b)) = o\left(N(m,b) \frac{(Logn)^2}{n}\right) \\ V(C(m,b)) = \frac{1}{n} V(Z_i) + o\left(\frac{1}{n}\right) \end{cases}$$

La démonstration de ce lemme, comme celle des théorèmes qui précèdent, est fournie au paragraphe IV. On notera que Lo et Singh (1986) avaient déjà introduit une décomposition du même type, pour $F_n(t) - F(t)$, et l'avaient utilisée pour étudier la convergence d'un estimateur de h basé sur la méthode du noyau (Lo, Mack et Wang (1989)).

III -QUELQUES COMMENTAIRES

1) Outre les convergences ponctuelles et L^2, on a déjà obtenu des théorèmes de convergence uniforme p.s. pour un estimateur s'écrivant $f_n / \left(1 - F_n + \frac{1}{n}\right)$ (f_n est un estimateur de f construit par la méthode du noyau, F_n l'estimateur de Kaplan Meier, le résultat est dans Földes et al. (1981)) et pour l'estimateur de Tanner-Wong, évoqué au §2 (le résultat est dans Gneyou (1991)). Les convergences obtenues ont lieu sur un intervalle borné [0,T].

Ce type de résultats est immédiat pour h_n : sous réserve que la série $\sum_{1}^{q(n)} a_i \, e_i$ converge uniformément vers h sur [0,b], il est clair que pour obtenir :

$$\sup_{0\le x\le b} \left| h_n(x) - h(x) \right| \xrightarrow[p.s.]{} 0$$

il suffira de prouver que $\sup_{0\le x\le b} \left| \sum_{1}^{q(n)} (\hat{a}_i - a_i)\, e_i(x) \right| \xrightarrow[p.s.]{} 0$

soit : $\forall\ \varepsilon > 0,\ \sum_{n} \left\{ \sum_{1}^{q(n)} P\left(\left| \hat{a}_i - a_i \right| > \varepsilon / q(n) \right) \right\} < \infty$

et cette condition est vérifiée (cf. (14)) si pour tous $\gamma_1 > 0,\ \gamma_2 > 0$, les séries

$$\sum_{1}^{\infty} n \sum_{1}^{q(n)} \exp\left\{ -\gamma_1\, n / q(n)\, N(e_i,b) \right\} \text{ et } \sum_{1}^{\infty} q(n) \exp\left\{ -\gamma_2\, n / q^2(n) \right\}$$

convergent, c'est-à-dire dans le cas des fonctions trigonométriques, en choisissant $q(n) = \sqrt{n} / \log n$ par exemple.

2) Les résultats évoqués ci-dessus ne permettent pas de comparer les efficacités des divers estimateurs existant de h, avec h_n, du point de vue des convergences fonctionnelles. On peut par contre établir une comparaison à partir des M.S.E et des M.I.S.E des estimateurs existant.

* Pour les convergences ponctuelles Grégoire (1991) et Mielniczuk (1983) ont montré que les estimateurs à noyau de la densité f et du taux de hasard h des X_i avaient un M.S.E asymptotique optimal en $n^{-4/5}$, la décomposition biais-variance obtenue s'écrivant, aux constantes près, comme dans le cas de données non censurées.Tanner et Wong (1983) ont, eux montré que, sous certaines conditions, on pouvait écrire la variance asymptotique de leur estimateur h_n^* sous la forme

$$V\left(h_n^*(x) \right) = \frac{A_1}{n\delta(n)} \frac{h(x)}{1 - L(x)} + o\left(\frac{1}{n\delta(n)} \right)$$

avec $\delta(n) \to 0$, $n\delta(n) \to \infty$, [$\delta(n)$ est la "fenêtre" au choix du statisticien, A_1 dépend du choix du "noyau"]. Nous pouvons ici écrire, quand la base (e_i) est celle des fonctions trigonométriques et q(n) choisi tel que H1) et H_2) soient vérifiées :

$$(11) \quad V(h_n(x)) \sim \frac{B_n^2}{n} \sim \frac{q(n)}{n} \cdot \frac{h(x)}{1 - L(x)} \cdot A_2$$

(voir 4) et 5), au § IV). La similitude est frappante.

h_n sera donc aussi performant que les estimateurs à noyau pour estimer les fonctions h telles que le biais $\sum_{q(n)+1}^{\infty} a_i e_i(x)$ converge assez rapidement vers 0 (par ex. en $(q(n))^{-2}$, ce qui correspond au biais usuel que donne un estimateur du noyau avec $\delta(n) = (q(n))^{-1}$) .

* Pour les convergences dans $L^2[0,b]$, on obtient aussi pour le MISE une vitesse optimale en $n^{-4/5}$ avec les estimateurs à noyau de h (cf. Grégoire (1991), Singpurwalla et Wong (1983)). La vitesse de convergence de h_n paraît, a priori, moins rapide. Dans le cas des fonctions trigonométriques, à partir de (12), en choisissant q(n) impair de la forme $q(n) = 2\,\overline{q(n)} + 1$, on peut écrire :

$$\sum_{i=1}^{q(n)} E\left[(\hat{a}_i - a_i)^2\right] = \frac{1}{b} \cdot \frac{\overline{q(n)} + 1}{n} \int_0^b \frac{h(t)}{1 - L(t)}\, dt + O\left(\left(\overline{q(n)}\right)^3 \frac{(\log n)^4}{n^2} \right) + O\left(\left(\overline{q(n)}\right)^2 \frac{(\log n)^2}{n^{3/2}} \right)$$

quantité équivalente à $\frac{1}{b} \cdot \frac{\overline{q(n)} + 1}{n} \int_0^b \frac{h(t)}{1 - L(t)}\, dt$, pourvu que l'on choisisse $\overline{q(n)}$ tel que $\overline{q(n)}\ (\log n)^2 / \sqrt{n}$ tende vers 0 (ce qui n'est pas nécessaire à la convergence du MISE).

Comme de surcroît (cf. Szegö (1959)) on a ici : $\forall k,\ a_{2k}^2 + a_{2k+1}^2 = O\left(\frac{1}{k^2}\right)$, le MISE asymptotique s'écrit alors :

$$\text{AMISE} = \frac{1}{b} \cdot \frac{\overline{q(n)} + 1}{n} \int_0^b \frac{h(t)}{1 - L(t)}\, dt + O\left(1 / \overline{q(n)}\right)$$

et devient donc équivalent à $O\left(1 / \overline{q(n)}\right)$, sous la contrainte imposée à $\overline{q(n)}$.

Le résultat précédent est valable sans conditions particulières sur f, mais l'écart optimal obtenu apparaît plus grand que dans le cas de la méthode du noyau. Il dépend ici encore étroitement du reste de la série $\sum_{q(n)+1}^{\infty} a_i^2$, et de la base (e_i) choisie.

De façon générale, l'efficacité de h_n dépendra donc; comme dans le cas des échantillons non censurés, du choix d'une base (e_i) "la plus adaptée possible" à h, et d'un q(n) optimal. Hjört (1985) propose d'utiliser une méthode de validation croisée classique en ce domaine, mais n'en justifie pas l'usage, d'un point de vue théorique. Notre étude reste donc à compléter sur ce point, tout comme dans celui d'une mise en oeuvre pratique (par simulation), qui devrait faire l'objet de travaux ultérieurs.

IV - DEMONSTRATIONS

1) Théorème 1

1) Il s'agit d'abord de montrer que $\lim_{n\to\infty} M_n = 0$,où M_n est le "M.I.S.E." usuel, c'est-à-dire $M_n = E\left(\left\| h_n - h \right\|^2_{L^2[0,b]} \right)$.

Les (e_i) formant une base orthonormale de L^2 [0,b], on a :

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

Or $\hat{a}_i - a_i = \int_0^b e_i \, d\,\bar{\Lambda}_n - \int_0^b e_i \, d\,\bar{\Lambda}$, donc, d'après le lemme (1) (formules (26, (28)) :

$$E((\hat{a}_i - a_i)^2) = \frac{1}{n} \int_0^b e_i^2(t) \frac{h(t)}{1 - L(t)} dt + O\left\{ (N(e_i,b))^2 . \frac{(\log n)^4}{n^2} \right\}$$

$$+ O\left\{ N(e_i,b) . \left[\int_0^b e_i^2(t) \frac{h(t)}{1 - L(t)} dt \right]^{1/2} . \frac{(\log n)^2}{n^{3/2}} \right\}$$

soit, les e_i étant uniformément bornés :

$$(12) \quad \sum_{i=1}^{q(n)} E(\hat{a}_i - a_i)^2 = \frac{1}{n} \int_0^b \Big(\sum_{i=1}^{q(n)} e_i^2(t)\Big) \frac{h(t)}{1 - L(t)}\, dt + O\left\{D_n \cdot \frac{(\log n)^4}{n^2}\right\}$$

$$+ O\left\{\sum_{i=1}^{q(n)} (N(e_i,b))\right\} \cdot \frac{(\log n)^2}{n^{3/2}}$$

Le premier terme est $O(q(n) / n)$. S'il tend vers 0, comme le second, le carré du troisième terme qui est $O[q(n).D_n (\log n)^4 / n^3]$ tendra aussi vers 0. Enfin, comme $q(n) \uparrow \infty$, $\sum_{q(n)+1}^{\infty} a_i^2 \to 0$: c'est le reste d'une série convergente égale à $\int_0^b h^2(t)\, dt$.

La première partie du théorème est démontrée.

2) Pour la seconde partie, en utilisant les propriétés des e_i et le fait que $q(n) \uparrow \infty$, on est ramené à montrer que $\sum_1^{q(n)} (\hat{a}_i - a_i)^2 \underset{P.S}{\to} 0$, soit a fortiori :

$$\forall \varepsilon, \varepsilon > 0, \quad \sum_n P\left[\sum_1^{q(n)} (\hat{a}_i - a_i)^2 > \varepsilon\right] < \infty$$

ou encore

$$(13) \quad \forall \varepsilon, \varepsilon > 0, \quad \sum_n \left\{\sum_1^{q(n)} P\left(\left|\hat{a}_i - a_i\right| > \sqrt{\varepsilon} / \sqrt{q(n)}\right)\right\} < \infty$$

Mais d'après le lemme 1, on peut écrire, pour tout $\alpha > 0$:

$$P\left(\left|\hat{a}_i - a_i\right| > \alpha\right) \le P\left(\left|\sum_{j=1}^{n} Z_j^i\right| > \frac{\alpha n}{2}\right) + P\left(\left|R_{i,n}^*\right| > \frac{\alpha}{2}\right)$$

avec

$$Z_j^i = \int_0^{X_j \wedge b} e_i(t) \frac{h(t)}{1 - L(t)} dt - I_{\{(\delta_j = 1) \cap (X_j \le b)\}} \frac{e_i(X_j)}{1 - L(X_j)}$$

et (cf (7) et (10)), il existe un réel C_b tel que :

$$\left| R_{i,n}^* \right| \le C_b N(e_i,b) \left\{ \frac{1}{n} + \left[\sum_0^1 \left\| H_n^i - H^i \right\|^2 \right] \right\}$$

Pour n assez grand on peut minorer $[\alpha / 2\, C_b\, N(e_i,b) - n^{-1}]^{1/2}$ par $(\alpha / 9\, C_b\, N(e_i,b))^{1/2}$. On en conclut (cf. (24)) :

$$P\left(\left| R_{i,n}^* \right| > \frac{\alpha}{2} \right) \le \sum_0^1 P\left[\left\| H_n^i - H^i \right\| > \left(\alpha / 9\, C_b\, .\, N(e_i,b) \right)^{1/2} \right]$$

$$\le 16\,(n+1) \exp\left\{ - n\alpha / 288.C_b.N(e_i,b) \right\}$$

En appliquant l'inégalité de Bernstein (Pollard (1984), p 193), on a aussi :

$$P\left(\left| \sum_{j=1}^n Z_j^i \right| > \frac{\alpha n}{2} \right) \le 2 \exp\left[- \alpha^2 n^2 / 8K \right]$$

où $K \le n M^2 \int_0^b \frac{h(t)}{1 - L(t)} dt + \frac{1}{3} M \left(\int_0^b \frac{h(t)}{1 - L(t)} dt + \frac{1}{1 - L(b)} \right) \frac{\alpha n}{2}$

(M majore les (e_i), en valeur absolue, d'après H1)).

Finalement, en remplaçant α par $\sqrt{\varepsilon} / \sqrt{q(n)}$, on obtient :

$$(14) \quad P\left(\left| \hat{a}_j - a_i \right| > \sqrt{\varepsilon} / \sqrt{q(n)} \right) = O\left\{ n \exp\left[- \alpha_1\, n / \sqrt{q(n)} N(e_i,b) \right] \right\}$$

$$+ O\left\{ \exp\left[- \alpha_2 n / q(n) \right] \right\}$$

avec $\alpha_2 > 0$, $\alpha_1 > 0$. Les conditions du théorème assurent donc la convergence souhaitée.

Q.E.D. ∎

2) Démonstration du théorème 2

1) Posons d'abord $\overline{h_n}(x) = \sum_1^{q(n)} a_i\, e_i(x)$.

Comme $h_n(x) = \int_0^b K_n(x,t)\, d\,\bar{\Lambda}_n(t)$ et que $\overline{h_n}(x) = \int_0^b K_n(x,t)\, d\,\bar{\Lambda}(t)$ on aura (lemme 1) :

$$h_n(x) - \overline{h_n}(x) = \frac{1}{n} \cdot \sum_1^n Z_{i,n} + R_n^*(x)$$

où $Z_{i,n} = \int_0^{X_i \wedge b} K_n(x,t)\, \frac{h(t)}{1 - L(t)}\, dt - I_{\{(\delta_i=1) \cap (X_i \le b)\}}\, \frac{\bar{K}_n(X_i)}{1 - L(X_i)}$

$$\text{Donc } A_n = \sqrt{n} \left\{ \frac{1}{n} \sum_1^n Z_{i,n} + R_n^*(x) - E\left(h_n(x) - \overline{h_n}(x)\right) \right\} / B_n$$

D'autre part on a (cf. lemme 1)

$$* \sqrt{n} \cdot E\left(h_n(x) - \overline{h_n}(x)\right) / B_n = O\left\{\sqrt{n}\left[N(\bar{K}_n, b) \cdot \frac{(\log n)^2}{n}\right] / B_n\right\}$$
$$= o(1) \text{ sous } H_2)$$

$* \sqrt{n}\, R_n^*(x) / B_n \to 0$ en probabilité, puisque (voir la majoration de R_n^* dans la preuve du théorème 1) :

$$P\left[\sqrt{n}\, R_n^*(x) / B_n > \varepsilon\right] = O\left\{n \cdot \exp\left[-\gamma\left(\frac{\varepsilon B_n}{\sqrt{n}}\right) \cdot n \cdot / N(\bar{K}_n, b)\right]\right\}$$

$$= O\left\{n \cdot \exp\left[-\gamma_1 \sqrt{n}\, B_n / N(\bar{K}_n, b)\right]\right\}$$

et, sous H2, cette expression, $O\left\{n \cdot \exp\left\{-\gamma^* (\log n)^2\right\}\right\}$, tend vers 0.

La limite en loi de A_n est donc celle de $\sqrt{n}\left[(1/n)\sum_1^n Z_{i,n}\right] / B_n$, soit celle de $\sum_1^n Z'_{i,n}$ où $Z'_{i,n} = Z_{i,n} / \sqrt{n} \cdot B_n$.

2) On sait que $E(Z_{i,n}) = 0$ et $V(Z_{i,n}) = B_n^2$. Donc les $Z'_{i,n}$ sont centrées, de variances égales à $\frac{1}{n}$ et, $S_n^2 = \sum_1^n V(Z'_{i,n}) = 1$. Pour montrer que $\sum_1^n Z'_{i,n} \rightarrow$ $\mathcal{N}(0,1)$, il suffira de prouver (théorème. de Liapounov : Billingsley (1968) p 44) qu'il existe δ, $\delta > 0$ tel que :

$$(17) \quad \sum_1^n E\left(\left|Z'_{i,n}\right|^{2+\delta}\right) \rightarrow 0 \quad \text{si } n \rightarrow \infty$$

Or les v.a. $Z_{i,n}$ sont toutes majorées en valeur absolue par (e_i bornées) O (q(n)), donc $E\left(|Z_{i,n}|^{2+\delta}\right) = O\left[(q(n))^{\delta} . B_n^2\right]$ et

$$\sum_1^n E\left(\left|Z'_{i,n}\right|^{2+\delta}\right) = n . E\left(|Z_{i,n}|^{2+\delta}\right) / \left(\sqrt{n} . B_n\right)^{2+\delta}$$

$$= O\left\{(q(n)^{\delta}) / n^{\delta/2} . B_n^{\delta}\right\}$$

qui tend vers 0 sous la condition H_2) Q.E.D.

3) DEMONSTRATION DU LEMME 1

1) Le théorème d'intégration par parties permet d'écrire que C(m,b) vaut :

$$(18) \quad \left\{m(b)\,\Lambda_n(b) - \int_0^b \Lambda_n(t)\, m'(t)\, dt\right\} - \left\{m(b)\,\Lambda(b) - \int_0^b m'(t)\,\Lambda(t)\, dt\right\}$$

Λ_n et Λ étant les fonctions de répartition de $\bar{\Lambda}_n$ et $\bar{\Lambda}$. En décomposant $\Lambda_n - \Lambda$ selon (5), on obtient :

$$(19)\quad C(m,b) = \left\{ m(b)\, P_n(b) - \int_0^b P_n(t)\, m'(t)\, dt \right\} + R_n^*$$

avec

$$(20)\quad R_n^* = m(b)\, R_n(b) - \int_0^b R_n(t)\, m'(t)\, dt$$

La majoration de R_n^* donnée dans le lemme est immédiate. Reste à montrer que le premier terme constituant C(m,b) s'écrit $\frac{1}{n} \sum_1^n Z_i$.

2) D'après (6) , on a :

$$(21)\quad n\, P_n(t) = \sum_{i=1}^{n} \int_0^{X_i \wedge t} \frac{h(s)}{1 - L(s)}\, ds - \sum_{i=1}^{n} \frac{1}{1 - L(X_i)}\, I_{\{(X_i \le t) \cap (\delta_i = 1)\}}$$

soit

$$(22)\quad n\, P_n(t) = \int_0^t u(s)\, ds - \mu(]0,t])$$

en posant :

$$u(t) = \left\{ \sum_{i=1}^{n} (n - i + 1)\, I_{[X_{(i-1)}, X_{(i)}[}(t) \cdot \frac{h(t)}{1 - L(t)} \cdot \delta_{X_i} \right\}$$

et

$$\mu = \sum_{i=1}^{n} I_{\{\delta_i = 1\}} \frac{1}{1 - L(X_i)} \cdot \delta_{X_i}$$

($X_{(i)}$, i = 1,..., n, représente comme d'habitude l'échantillon ordonné, avec $X_{(0)} = 0$).

On obtient alors directement d'après (22) :

$$n \int_0^b P_n(t)\, m'(t)\, dt = \int_0^b (m(b) - m(s))\, u(s)\, ds - \int_0^b (m(b) - m(s))\, d\mu(s)$$

$$= n.m(b)\, P_n(b) - \sum_{i=1}^{n} \left\{ \int_0^{X_i \wedge b} m(s)\, \frac{h(s)}{1 - L(s)}\, ds \right\}$$

$$+ \sum_{i=1}^{n} m(X_i)\, \frac{I_{\{(\delta_i=1) \cap (X_i \le b)\}}}{1 - L(X_i)}$$

en remplaçant ν et μ par leurs expressions.

Donc $C(m,b) = \frac{1}{n} \sum_{i=1}^{n} Z_i + R_n^*$, les Z_i étant définis en (8).

3) Les variables X_i, i.i.d, ont la densité ℓ, égale à $f(1 - G) + g(1 - F)$, puisqu'elles ont la f.d.r. $1 - (1 - F)(1 - G)$. D'autre part, la mesure υ définie par $\upsilon(A) = P(X_i \in A, \delta_i=1)$ admet la densité $f(1 - G)$. Dès lors :

$$E(Z_i) = \int_0^{\infty} \left(\int_0^{s \wedge b} m(t)\, \frac{h(t)}{1 - L(t)}\, dt \right) \ell(s)\, ds - \int_0^{\infty} m(s)\, \frac{I_{\{s \le b\}}}{1 - L(s)}\, f(s)\, (1 - G(s))\, ds$$

Le premier des deux termes, valant : $\int_0^b \left(\int_t^{\infty} \ell(s)\, ds \right) m(t)\, \frac{h(t)}{1 - L(t)}\, dt$

(théorème de Fubini) annule le second, donc les Z_i sont centrées.

De même $E(Z_i^2) = \sum_{i=1}^{3} A_i$, avec

$$A_1 = \int_0^{\infty} \left(\int_0^{s \wedge b} m(t)\, \frac{h(t)}{1 - L(t)}\, dt \right)^2 \ell(s)\, ds$$

$$A_2 = -2 \int_0^{\infty} \left(\int_0^{s \wedge b} m(t)\, \frac{h(t)}{1 - L(t)}\, dt \right) \cdot \left(m(s)\, \frac{I_{\{s \le b\}}}{1 - L(s)} \right) f(s)\, (1 - G(s))\, ds$$

$$A_3 = \int_0^{\infty} \left(\frac{m(s)\, I_{\{s \le b\}}}{1 - L(s)} \right)^2 f(s)\, (1 - G(s))\, ds = \int_0^b m^2(s)\, \frac{h(s)}{1 - L(s)}\, ds$$

On veut montrer que $V(Z_i) = A_3$. Reste donc à prouver que $A_2 = - A_1$. Il suffit d'intégrer par parties A_1 : - (1 - L) étant primitive de $\hat{\ell}$; on aura bien :

$$A_1 = \left[- \left\{ \int_0^{s \wedge b} m(t) \cdot \frac{h(t)}{1 - L(t)} \, dt \right\}^2 (1 - L(s)) \right]_{s=0}^{s=\infty}$$

$$+ 2 \int_0^{\infty} (1 - L(s)) \left\{ \int_0^{s \wedge b} m(t) \, \frac{h(t)}{1 - L(t)} \, dt \right\} \left\{ \frac{m(s)\, h(s)}{1 - L(s)} \, I_{\{s \le b\}} \right\} ds$$

$$= 0 - A_2$$

3) Comme $E(Z_i) = 0$, $E(C(m,b)) = E(R_n^*)$, et donc, d'après (20) et (7), il existe un réel positif C_b tel que :

$$(23) \quad |E\, C(m,b)| \le N(m,b) \cdot C_b \left\{ \frac{1}{n} + \sum_{i=0}^{1} E \left[\left\| H_n^i - H^i \right\|^2 \right] \right\}$$

Une étude attentive de Pollard (1984), p 13-16, montre que, bien que les H_n^i et H^i ne soient pas des f.d.r. "standard" on peut écrire :

$$(24) \quad \forall \ \varepsilon > 0, \ P\left(\left\| H_n^i - H^i \right\| > \varepsilon \right) \le 8 (n + 1) \exp (- n \varepsilon^2 / 32)$$

En "coupant" alors $\left\| H_n^i - H^i \right\|^2$ par $I\left\{ \left\| H_n^i - H^i \right\| > \mathrm{Logn} / \sqrt{n} \right\}$, on obtient :

$$E\left(\left\| H_n^i - H^i \right\|^2 \right) \le \frac{(\mathrm{logn})^2}{n} + 32 (n + 1) \exp (- (\mathrm{logn})^2 / 32)$$

puisque $\| H_n^i - H^i \|$ se majore par 2.

On aurait de même :

$$E\left(\left\| H_n^i - H^i \right\|^4 \right) \le \frac{(\mathrm{logn})^4}{n^2} + 64 (n + 1) \exp (- n \varepsilon^2 / 32)$$

Alors (23) permet d'écrire :

$$(25) \quad \left|E(C(m,b))\right| = O\left\{ N(m,b) \cdot C_b \frac{(\log n)^2}{n} \right\}$$

En utilisant encore (9), on a aussi :

$$(26) \quad E\{[(C(m,b)]^2\} = \frac{1}{n} V(Z_i) + E\left[(R_n^*)^2 \right] + 2E\left[R_n^* \cdot \frac{1}{n} \left(\sum_1^n Z_i \right) \right] .$$

avec

$$(27) \quad E\left[(R_n^*)^2 \right] = O\left\{ [N(m,b) \cdot C_b]^2 \frac{(\log n)^4}{n^2} \right\}$$

et donc (inégalité de Schwarz)

$$(28) \ \frac{1}{n} E\left(R_n^* \left(\sum_1^n Z_i \right) \right) = O\left\{ \{N(m,b)\, C_b\} \times \frac{(\log n)^2}{n} \left\{ \frac{1}{n} \int_0^b m^2(t) \frac{h(t)}{1 - L(t)} dt \right\}^{1/2} \right\}$$

Le lemme s'ensuit ... ∎

4) <u>Calcul du biais et de la variance de</u> $h_n(x)$

1) On peut écrire :

$$V\left(h_n(x)\right) = E\left\{ \left(h_n(x) - \bar{h}_n(x) \right) - E\left(h_n(x) - \bar{h}_n(x) \right) \right\}^2$$

avec :

$$h_n(x) - \bar{h}_n(x) = \int_0^b K_n(x, t)\, d\,\overline{\Lambda}_n(t) - \int_0^b K_n(x, t)\, d\overline{\Lambda}(t)$$

D'après (25), on a donc

$$E\left(\bar{h}_n(x)-h_n(x)\right) = O\left\{N(\bar{K}_n,b).(\log n)^2 / n\right\}$$

Sous la condition H_2) (2ème partie) cette quantité est négligeable devant $B_n/\sqrt{n}$, alors que $E\left[\left(h_n(x) - \bar{h}_n(x)\right)^2\right]$ est elle-même équivalente à cette quantité (cf.26-27). Finalement $V(h_n(x)) = \frac{B_n^2}{n}(1 + o(1))$.

$$2)\ E\left(h_n(x) - h(x)\right) = E\left(h_n(x) - \bar{h}_n(x)\right) + \sum_{q(n)+1}^{\infty} a_i\ e_i\ (x)$$

Les e_i étant bornés, d'après ce qui précède, on obtient le biais de $h_n(x)$ égal à

$$O\left(\sum_{q(n)+1}^{\infty} |a_i|\right) + O\left(N(\bar{K}_n,b)\ .\ (\mathrm{Log}n)^2 / n\right)$$

5) Equivalent asymptotique de B_n^2 (base des fonctions trigonométriques).

Lorsque les (e_i) sont les fonctions définies en (3), et si q(n) est impair, un calcul classique donne :

$$K_n\ (x,t) = \sum_{1}^{q(n)} e_i(x)\ e_i(t) = \frac{1}{b}\ .\ \sin\left\{q(n)\ \frac{\pi}{b}\ (x - t)\right\} / \sin\left\{\frac{\pi}{b}\ (x - t)\right\}$$

quotient défini sans ambigüité pour tout x, $0 < x < b$: pour tout $t \in [0,b]$ le dénominateur ne s'annule que pour $x = b$, auquel cas on prolonge le quotient à q(n), par continuité.

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

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

où

$$g_n(u) = \frac{1}{\pi b}\ I_{\left\{q(n)\frac{x-b}{b}\pi,\ q(n)\frac{x}{b}\pi\right\}}(u)\ \left\{\frac{1}{q(n)}\ \frac{\sin u}{\sin(u/q(n))}\right\}^2\ \frac{h\left(x - \frac{bu}{\pi q(n)}\right)}{1 - L\left(x - \frac{bu}{\pi q(n)}\right)}$$

Soit $a = \sup\left(\left|\frac{x - b}{b}\right| , \frac{x}{b} \right)$. Pour $g_n(u) \neq 0$, nous aurons $- a\pi < u / q(n) \leq a\pi$

et comme $a < 1$, sin u/ sin (u/ q(n)), ici encore défini sans ambigüité, est tel que :

$$(29) \quad \left\{ \sin (u / q(n)) / (u / q(n)) \right\} \in \left[\frac{\sin a}{a} , 1 \right]$$

Alors :

* pour tout réel a, $\lim_{n \to \infty} g_n(u) = \frac{2}{\pi b} \left(\frac{h(x)}{1 - L(x)}\right) \left(\frac{\sin u}{u}\right)^2$

($q(n) \uparrow \infty$, h et L sont continues).

* pour n assez grand $g_n(u) \leq K \left(\frac{\sin u}{u}\right)^2 \left(\sup_{0 \leq t \leq b} h(t) \right) / (1 - L(b))$

(cf. 29)).

Donc, le théorème de Lebesgue s'appliquant, on peut écrire :

$$\int g_n(u) \, du \underset{n \to \infty}{\to} \frac{2}{\pi b} \frac{h(x)}{1 - L(x)} \int_{\mathbb{R}} \left(\frac{\sin u}{u}\right)^2 du$$

et B_n^2 est équivalent à A . q(n), $A > 0$, pour n grand Q.E.D.

∎

V - BIBLIOGRAPHIE

AALEN O. (1978)

"Nonparametric estimation of partial transition probabilities in multiple decrement models." - Ann. Stat. 6, p 534-545.

BILLINGSLEY P. (1968)

"Convergence of probability measures" - Wiley New-York.

BRESLOW N., CROWLEY J. (1974)

"A large sample study of the life table and product limit estimates under random censorship" - Ann. of statist. 2 (1974), p 437-453.

DROESBEKE J.J., FICHET B., TASSI P., éditeurs

"Analyse statistique des durées de vie." - Economica (1989).

EFRON B. (1967)

"The two sample problem with censored data." - Proc. 5th Berkeley symp, Vol. 4, p 831-853.

FOLDES A., REJTO L., WINTER B.B. (1981)

"Strong consistency properties of nonparametric estimators for randomly censored data II. Estimation of density and failure rate"- Period. Math. Hungar 12, p 15-29.

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.

HJORT N.L. (1985)

"Discussion contribution to Andersen and Borgan's review article." Scand. J. Statis. - 12 - p 141-150.

HJORT N.L. (1991)

"Semiparametric estimation of parametric hazard rates." Invited paper presented at the Advanced Study Workshop on Survival Analysis and Related Topics.

KAPLAN E., MEIER P. (1958)

"Nonparametric estimation from incomplete observations." - JASA 53, p 457-481.

KIMURA D.K. (1972)

"Fourier series methods for censored data." - Thèse soutenue à l'Université de Washington.

LO S.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 - p 461-473.

LO S.H., SINGH K. (1986)

"The product-limit estimator and the bootstrap. Some asymptotic representations. Prob. Th.and Rel. Fields 71 - p455-465.

McKEAGUE I.W. (1986)

"Estimation for a semimartingale regression model using the method of sieves". Ann. Statis. 14 - p 579-589.

MIELNICZUK J. (1983

"Properties of some kernel estimators and of the adopted Loftgarten-Quesenberry estimator of a density function for censored data." - Periodica Math. Hungarica, Vol. 16 (2), p 69-80

NELSON W.B. (1972)

"Theory and applications of hazard plotting for censored failure data." - Technometrics 14, p 945-966.

POLLARD D. (1984)

"Convergence of stochastic processes." - Springer Verlag.

PONS O. (1986)

"Vitesse de convergence des estimateurs à noyau pour l'intensité d'un processus ponctuel". Statistics.17 - p 577-584.

REID N. (1981)

"Influence Functions for censored data." - The annals of Stat., Vol. 9 n° 1, p 78-92.

SINGPURWALLA N.D., WONG W.H. (1983)

"The estimator of the hazard function from randomly censored data by kernel method." JASA, 78 p 478-481.

SZEGO G. (1959)

"Orthogonal polynomials." - Amer. Math. Soc. Coll. Publ. 32.

TANNER A., WING HUNG WONG (1983)

" The estimation of the hazard function from randomly censored data by the kernel method." - The annals of statistics, Vol. 11 n° 3, p 989-993.

CHAPITRE III

ESTIMATION FONCTIONNELLE DANS LES MODELES DE DUREE: METHODE DES FONCTIONS ORTHOGONALES.

M. DELECROIX * et O. YAZOURH **

RESUME

On définit un estimateur non paramétrique de la densité de variables aléatoires X_i° 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°.

Classification AMS 62 G05
STMA 04 180

Mots clés

Données censurées, estimateur non paramétrique du taux de hasard, méthode des fonctions orthogonales, estimateur de Kaplan-Meier, modèle à censures aléatoires.

DENSITY AND HASARD RATE ESTIMATEUR FOR CENSORED DATA: ORTHOGONAL FUNCTIONS METHOD.

ABSTRACT

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

Keywords

Censured data, non parametric density estimation, Kaplan-Meier estimator, orthogonal functions methods, random censorship model.

---

* GREMAQ - (URA 947) Université des Sciences Sociales Toulouse 1,
Place Anatole France
31042 TOULOUSE CEDEX
L.S.P., (URA DO745) Université Paul Sabatier, 118 route de Narbonne,
31062 TOULOUSE CEDEX

** Laboratoire de Statistique et Probabilité - U.F.R. de Mathématiques pures et appliquées - U.S.T.L.F.A.
59655 - VILLENEUVE D'ASCQ CEDEX

I - Introduction

Le but de ce travail est l'étude de la convergence d'estimation de la densité et du taux de hasard de variables aléatoires X_i° 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).

Appelons X_i°, $i>1$, les variables aléatoires représentent des durées de vie que l'on veut étudier (durée de chômage, durée de survie, durée de fonctionnement sans panne, etc...). Elles sont à valeurs dans $\mathbb{R}^+$, i.i.d, on appelle f leur densité, supposée continue, F leur fonction de répartition. Le statisticien ne peut alors observer que des variables X_i, $i>1$, où $X_i=\inf\{X_i^\circ, C_i\}$, les C_i symbolisant les censures apportées aux durées de vie d'intérêt (dates d'échantillonnages etc...). On suppose que les C_i forment une suite de variables réelles, i.i.d, à valeurs dans $\mathbb{R}^+$, de densité g, de fonction de répartition G, et indépendantes des X_i°. 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 X_i dont il dispose est censurée ou non, l'échantillon est finalement constitué de n couples (X_i,δ_i) où

$\delta_i=I_{\{X_i^\circ<C_i\}}$ (δ_i est l'indicateur de non censure).

L'estimation non paramétrique de f, de h, de F et du taux de hasard cumulé Λ, a été récemment abordée par de nombreux auteurs parmi lesquels ont peut citer, K.E.Gneyou (1991) et G.Gregoire (1991) pour des travaux récents, en renvoyant à Droesbeke et al (1990) chapitre 4, pour une synthèse de résultats obtenus avant 1988.

Dans la quasi totalité de ces travaux, on utilise la méthode du noyau pour estimer le paramètre fonctionnel d'intérêt. Kimura (1972) et Hjört (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 M.Delecroix et O.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 performances 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.

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. On pourra aussi, choisir une suite $(b_n)_{n \geq 0}$, telle que $b_n \longrightarrow \infty$, et construire l'estimateur f_n sur $[0,b_n]$ par la méthode des fonctions orthogonales, en utilisant une base orthonormale de $L^2_{([0,bn])}$.

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 coefficient 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 estimateurs utilisés et à l'énoncé des principaux théorèmes de convergence. Les démonstrations seront exposées dans le second paragraphe. Le troisième, consacré à quelques commentaires, sera suivi d'un appendice contenant les preuves de quelques lemmes techniques.

II - Notations, hypothèses et résultats

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, si nous appelons $(e_i), i \geq 1$, une base orthogonale de cet espace $L^2_{([0,b])}$, on peut 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. F_n est ici le classique estimateur de Kaplan-Meier (1958) de F. On sait que F_n est la fonction de répartition d'une mesure qui met la masse:

$$\frac{1}{n-R_i+1} \prod_{j/X_j \leq X_{i-1}} \left(\frac{n-R_j}{n-R_j+1} \right)^{\delta_j}$$

en chaque X_i non censuré. R_i représente le rang dans l'échantillon observé. Dès lors

$$\hat{a}_i = \sum_{k/X_k \leq b} \delta_k\, e_i(X_k) \left(\frac{1}{n-R_k+1}\right) \prod_{j/X_j \leq X_{k-1}} \left(\frac{n-R_j}{n-R_j+1}\right)^{\delta_j}$$

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

(1) $$h_n = \frac{f_n}{1-F_n+\frac{1}{n}}$$

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

(2) $$H^i(t)=P(X_i>t,\ \delta_j=i)$$

et $$H_n^i(t) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{X_j>t,\ \delta_j=i\}} \quad , \quad \text{pour } i=0,1.$$

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

(3) $$N(m,b) = \left[|m(b)| + \int_0^b |m'(t)|\,dt\right].$$

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

(4) $$D_n = \sum_{i=1}^{n} N(e_i,b)^2 = \sum_{i=1}^{q(n)} \left[|e_i(b)| + \int_0^b |e'_i(t)|\,dt\right]^2$$

(5) $$W_n(x,t) = \sum_{i=1}^{q(n)} e_i(x)\, e_i(t) \quad , \quad \text{pour tout x et t réels.}$$

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

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

H_0): $H^i(b) > 0$, $i=0,1$, pour tout $b > 0$.

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_0) traduit simplement le fait que les distributions des X_i et C_i n'ont pas un intervalle borné comme support. H_1) est liée à la base (e_i) 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) Résultats de convergence de l'estimateur de la densité:

Dans tout ce travail, ou supposera que b vérifie H_0).

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

Si l'on suppose que $q(n) \longrightarrow \infty$, $\lim_{n \longrightarrow \infty} \frac{q(n)}{n} = 0$ ainsi que,

$$\lim_{n \to \infty} \left[\frac{D_n (\log n)^4}{n^2}\right] = 0,$$

nous aurons: $\lim_{n \to \infty} E\left(||f_n - f||^2_{L^2[0,b]}\right) = 0.$

Si maintenant $q(n) \to \infty$, de telle sorte que pour tous γ_1 et γ_2 strictement positifs les séries de termes généraux $q(n) \exp\left\{\frac{-\gamma_1 n}{q(n)}\right\}$ et $n\left\{\sum_{i=1}^{q(n)} \exp\left[\frac{-\gamma_2 n}{\sqrt{q(n)}\, N(e_i,b)}\right]\right\}$ convergent,

on aura $\lim_{n \to \infty} p.s\left[\,||f_n - f||^2_{L^2[0,b]}\right] = 0.$

Remarque 1: Les convergences recherchées sont obtenues si "grosso modo" les fonctions e_i ne varient pas trop brutalement sur [0,b] et si q(n) croit 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 $L^2[0,b]$, $(e_i), i \geq 1$, est définie par $e_1(t) = \frac{1}{\sqrt{b}}$ et ($k \geq 1$, $t \in [0,b]$):

(7) $e_{2k}(t) = \sqrt{\frac{2}{b}} \sin \left[\frac{2k\pi}{b} (t - \frac{b}{2}) \right]$, $e_{2k+1}(t) = \sqrt{\frac{2}{b}} \cos \left[\frac{2k\pi}{b} (t - \frac{b}{2}) \right]$

On obtient:

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

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

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

On montrera aussi le résultat suivant, concernant la convergence uniforme p.s.

Théorème 2: Si $\sup_{x \in [0,b]} | \bar{f}_n(x) - f(x)| \longrightarrow 0$, lorsque $n \longrightarrow \infty$, $H_1)$ est vérifiée, et si pour tout γ_1 et γ_2 strictement positifs les séries de termes généraux:

$$n \sum_{n=1}^{q(n)} \exp \left\{ \frac{-\gamma_1 n}{q(n)\, N(e_i,b)} \right\} \quad \text{et} \quad q(n) \exp \left\{ \frac{-\gamma_2 n}{q(n)^2} \right\} \quad \text{convergent, alors:}$$

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

Corollaire 2: Dans le cas de la base trigonométrique (7), si $\sup_{x \in [0,b]} | \bar{f}_n(x) - f(x)| \longrightarrow 0$ quand $n \longrightarrow \infty$, alors si, $\forall \gamma > 0$, $\sum_{n=1}^{\infty} n\, q(n) \exp \left\{ \frac{-\gamma n}{q(n)^2} \right\} < \infty$, on a,

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

Enfin, on cherchera la limite de J_n définie par:

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

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

et $\quad H_n(x,s) = S(s)\, W_n(x,s) + \int_s^b S'(t)\, W_n(x,t)\, dt.$

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

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

$$J_n \xrightarrow{\mathcal{L}} N(0,1), \qquad \text{lorsque } n \longrightarrow \infty.$$

Ici encore, un exemple d'application nous est fourni par la base trigonométrique, on obtient alors:

Corollaire 3: Si les (e_i), $i \geq 1$, sont définies en (7) et si $\dfrac{q(n)^3\, (\log n)^4}{n} \longrightarrow 0$,

lorsque $n \longrightarrow \infty$, alors $J_n \xrightarrow{\mathcal{L}} N(0,1)$.

La démonstration de ce résultat, se base essentiellement sur le lemme suivant (cf. appendice).

Lemme 1: Si la base (e_i), $i \geq 1$, est celle des fonctions trigonométriques, alors:

$$B_n^2 > q(n)\, A \quad \text{avec, } A > 0.$$

3) 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 Lo, Mark et Wang (1989), sera $\dfrac{f_n}{1-F_n+\frac{1}{n}}$, où f_n est l'estimateur de la densité construit par la méthode des fonctions orthogonales, étudié ci dessus, et F_n est l'estimateur de Kaplan-Meier de F. On montrera alors que l'estimateur h_n, ainsi construit, possède les propriétés de convergence suivantes.

Théorème 4: Si $\sup_{x \in [0,b]} | \bar{f}_n(x) - f(x)| \longrightarrow 0$, lorsque $n \longrightarrow \infty$ et 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 $n \sum_{n=1}^{q(n)} \exp\left\{\frac{-\gamma_1 n}{q(n)\, N(e_i,b)}\right\}$ et $q(n) \exp\left\{\frac{-\gamma_2 n}{q(n)^2}\right\}$ convergent, alors , on a

$$\sup_{x \in [0,b]} |h_n(x) - h(x)| \longrightarrow 0 \quad \text{p.s,} \quad \text{lorsque } n \longrightarrow \infty$$

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

Corollaire 4: Si $\sup_{x \in [0,b]} | \bar{f}_n(x) - f(x)| \longrightarrow 0$, lorsque $n \longrightarrow \infty$, et si f vérifie H_2), alors si $\forall \gamma > 0$, $\sum_{n=1}^{\infty} n\, q(n) \exp\left\{\frac{-n\gamma}{q(n)^2}\right\} < \infty$, on a

$$\sup_{x \in [0,b]} |h_n(x) - h(x)| \longrightarrow 0 \quad \text{p.s,} \quad \text{lorsque } n \longrightarrow \infty$$

Un autre résultat, concernant la variance de l'estimateur h_n, est donné par:

Théorème 5: $\forall x \in [0,b]$, on a: $\mathrm{Var}(h_n(x)) = 0\,(\mathrm{Var}(f_n(x))) + 0\,(q(n)^2\, \mathrm{var}\,(S_n(x)))$.

En choisissant une autre fois la base des fonctions trigonométriques, on obtient:

Corollaire 5: Si la base $(e_i), i \geq 1$, est celle des fonctions trigonométriques définies sur [0,b], alors:

$$\forall x \in [0,b],\ \mathrm{Var}(h_n(x)) = 0\left(\mathrm{Var}(f_n(x)) + q(n)^2\, \mathrm{Var}\,(S_n(x))\right) = 0\left(\frac{q(n)^2}{n}\right).$$

III - Etude des convergences

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 Zi sont indépendantes, et R_n tend asymptotiquement vers

zéro. Cette technique a été récemment utilisée par Lo, Mark et Wang (1989), dans un autre cadre. Notre décomposition se basera, elle, essentiellement sur les résultats de N.Reid, permettant de développer $F_n(t)-F(t)$ en fonction des "courbes d'influence" de F_n. N.Reid a montré exactement que:

(9) $\forall t,\ t \geq 0,\ F_n(t)-F(t) = P_n(t)+R_n(t)$, où

(10) $$nP_n(t) = \sum_{i=1}^{n} 1_{\{\delta i=1\}} K_1(t,X_i) + 1_{\{\delta i=0\}} K_2(t,X_i)\ .$$

avec:

(11)
$$K_1(t,s) = S(t) \int_0^{s\wedge t} \frac{h(u)}{1-L(u)}\,du - S(t)\,\frac{1_{\{s\leq t\}}}{1-L(s)}$$
$$K_2(t,s) = S(t) \int_0^{s\wedge t} \frac{h(u)}{1-L(u)}\,du$$

(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:

(12) $$\sup_{t\in[0,b]} |R_n(t)| \leq C_b \left\{ \sum_{i=0}^{1} ||H_n^i - H^i||^2 + \frac{1}{n} \right\}.$$

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:

(13)
$$C(m,b) = \int_0^b m(t)\,dF_n(t) - \int_0^b m(t)\,dF(t)$$
$$= \int_0^b m(t)\,d(P_n(t)+R_n(t)).$$

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

(14) $$Z_i = \int_0^{X_i\wedge b} S(t)\,m(t)\,\frac{h(t)}{1-L(t)}\,dt + \int_0^b S'(t)\,m(t) \left(\int_0^{X_i\wedge b} \frac{h(u)}{1-L(u)}\,du \right) dt$$

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

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:

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

et on peut écrire:

$$C(m,b) = \frac{1}{n}\sum_{i=1}^{n} Z_i + R_n^*, \quad \text{où} \tag{15}$$

$$|R_n^*| < N(m,b)\, \{ \sup_{t \in [0,b]} |R_n(t)| \}. \tag{16}$$

Dès lors on aura:

$$E(C(m,b)) = 0\left\{ N(m,b)\, \frac{(\log n)^2}{n} \right\}.$$

$$V(C(m,b)) = \frac{1}{n} V(Z_i)\, (1 + o(1)). \tag{17}$$

2) Démonstration du théorème 1

1°) On commence d'abord par montrer que $\lim_{n \to \infty} M_n = 0$

où $M_n = E\,(\,|\,|f_n - f\,|\,|^2_{L^2[0,b]}\,)$.

Les (e_i) forment 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 (formules (23), (24) et (25)), on a:

$$E[(\hat{a}_i - a_i)^2] = \frac{1}{n} \int_0^b \left(S(s)\, e_i(s) + \int_s^b S'(t)\, e_i(t)\, dt \right)^2 \frac{h(s)}{1-L(s)}\, ds$$

$$+ 0\left\{ [(N(e_i,b))^2 \frac{(\log n)^4}{n^2}] \right\} + 0\left\{ N(e_i,b) \frac{(\log n)^2}{n} [\frac{1}{n} V(Z_i)]^{1/2} \right\}.$$

L'hypothèse H_1) est supposée vérifiée, donc les (e_i) sont uniformément bornés et par suite, on peut écrire:

$$(18) \qquad \sum_{i=1}^{q(n)} E(\hat{a}_i - a_i)^2 = \frac{1}{n} \int_0^b \sum_{i=1}^{q(n)} \left(S(s)\, e_i(s) + \int_s^b S'(t)\, e_i(t)\, dt \right)^2 \frac{h(s)}{1-L(s)}\, ds$$

$$+ 0\left(D_n(\frac{(\log n)^4}{n^2}) \right) + 0\left(\sum_{i=1}^{q(n)} N(e_i,b) \frac{(\log n)^2}{n^{3/2}} \right).$$

Le premier terme est un $0\left(\frac{q(n)}{n}\right)$, s'il tend vers zéro ainsi que le second , le carré du troisième terme qui est un $0\left(\frac{q(n)\, D_n\, (\log n)^4}{n^3}\right)$ tendra aussi vers zéro.

La démonstration de la deuxième partie du théorème, est la même que celle du théorème 1 (cf. M. Delecroix et O. Yazourh (1992)), il suffit de remplacer les Z_j^i par:

$$Z_j^i = \int_0^{X_i \wedge b} S(t)\, e_i(t) \frac{h(t)}{1-L(t)}\, dt + \int_0^b S'(t)\, e_i(t) \left(\int_0^{X_i \wedge b} \frac{h(u)}{1-L(u)}\, du \right) dt$$

$$- S(X_j)\, e_i(X_j) \frac{1_{\{(\delta_j=1) \cap (X_i<t)\}}}{1-L(X_i)} - \int_0^b S'(t)\, e_i(X_j) \frac{1_{\{(\delta_j=1) \cap (X_i<b)\}}}{1-L(X_i)}\, dt.$$

3) Preuve du corollaire 1

Dans le cas de la base trigonométrique, $\forall i, i \geq 1$, $a_i = 0\left(\frac{1}{i^2}\right)$ (cf. Sansone p. 106),

donc $\sum_{i=q(n)+1}^{\infty} a_i^2 = 0\left(\frac{1}{q(n)}\right)$ qui tend vers zéro quand n tend vers l'infini, et $N(e_i,b) = 0$ (i) ainsi que, $D_n = 0$ $(q(n)^3)$, en appliquant alors le théorème 1, on obtient les deux convergences souhaitées pour l'estimateur f_n, construit à partir de cette base.

4) Preuve du théorème 2

Il suffit de prouver que $\sup_{x \in [0,b]} |f_n(x) - f(x)| \longrightarrow 0$ p.s, quand $n \longrightarrow \infty$, et

donc $\sup_{x \in [0,b]} \left| \sum_{i=1}^{q(n)} (|\hat{a}_i - a_i|)\, e_i(x) \right| \longrightarrow 0$ p.s.

Soit: $\forall \varepsilon > 0, \quad \sum_n \left\{ \sum_{i=1}^{q(n)} P\left(\hat{a}_i - a_i > \frac{\varepsilon}{q(n)} \right) \right\} < \infty$,

cette condition est vérifiée (cf. (14), M. Delecroix et O. Yazourh (1992)), si pour tous $\gamma_1 > 0$ et $\gamma_2 > 0$ les séries de termes généraux: $n \sum_{n=1}^{q(n)} \exp\left\{ \frac{-\gamma_1 n}{q(n)\, N(e_i,b)} \right\}$ et

$q(n) \exp\left\{ \frac{-\gamma_2 n}{q(n)^2} \right\}$ sont convergentes.

5) preuve du corollaire 2

Dans le cas où les (e_i), sont les fonctions de la base trigonométrique,
$\forall i \geq 1, \quad N(e_i,b) = 0\,(i)$ et donc, $N(e_i,b) = 0\,(q(n))$.

6) Preuve du théorème 3

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

$$Z_{i,n} = \int_0^{X_i \wedge b} S(t)\ \overline{W_n}(t) \frac{h(t)}{1-L(t)}\, dt + \int_0^b S'(t)\ \overline{W_n}(t) \left(\int_0^{X_i \wedge b} \frac{h(u)}{1-L(u)}\, du \right) dt$$

$$- S(X_i)\ \overline{W_n}(X_i) \frac{1_{\{(\delta_j = 1) \cap (X_i \leq b)\}}}{1-L(X_i)} - \int_0^b S'(t)\ \overline{W_n}(t) \frac{1_{\{(\delta_j = 1) \cap (X_i \leq t)\}}}{1-L(X_i)}\, dt.$$

7) Preuve du corollaire 3

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

8) Preuve du théorème 4

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

Le premier terme se majore par:

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

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

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

D'autre part, $F(b) < 1$: $\exists \varepsilon > 0 \,/\, F(b) < 1-\varepsilon$, et $F_n \longrightarrow F$ p.s, quand $n \longrightarrow \infty$, donc $\forall \varepsilon' > 0$, $\exists \eta \,/\, \forall n \geq \eta$, $1 - F_n > 1 - F - \varepsilon'$ p.s.

En choisissant, $\varepsilon' = \frac{\varepsilon}{2}$, on a, $1 - F_n(b) + \frac{1}{n} \geq \frac{\varepsilon}{2}$ p.s.

c.à.d $\frac{1}{1-F_n(b)+\frac{1}{n}} \leq \frac{2}{\varepsilon}$ p.s. On en déduit finalement que:

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

Donc, il suffit que les hypothèses du théorème 2, soient vérifiées pour avoir:

$\sup_{t\in[0,b]}|f(t) - f_n(t)| \longrightarrow 0$ p.s. Et comme b vérifie H_0) : $\sup_{t\in[0,b]}|F(t) - F_n(t)| \longrightarrow 0$ p.s, on obtient donc la convergence souhaitée.

9) Preuve du corollaire 4

Voir preuve du corollaire 2.

10) Preuve du théorème 5

$$E\{[\, h_n(t) - E(h_n(t))\,]^2\}$$

$$= E\left\{\left[\frac{f_n(t)}{E(S_n(t)) + \frac{1}{n}} - \frac{1}{E(S_n(t)) + \frac{1}{n}}\,[h_n(t)\,(S_n(t) - E(S_n(t)))] - E(h_n(t))\right]^2\right\}$$

$$= E\left\{\left[\frac{f_n(t) - E(f_n(t))}{E(S_n(t)) + \frac{1}{n}} + \left(\frac{E(f_n(t))}{E(Sn(t)) + \frac{1}{n}} + E(h_n(t))\right)\right.\right.$$

$$\left.\left. - \frac{1}{E(S_n(t)) + \frac{1}{n}}\,(h_n(t)\,(S_n(t) - E(S_n(t))))\right]^2\right\}$$

$$\leq 3\,E\left[\left(\frac{f_n(t) - E(f_n(t))}{E(S_n(t)) + \frac{1}{n}}\right)^2\right] + 3\left[\frac{E(f_n(t))}{E(S_n(t)) + \frac{1}{n}} - E(h_n(t))\right]^2$$

$$+ 3\left[\frac{1}{E(S_n(t)) + \frac{1}{n}}\,\{h_n(t)\,(S_n(t) - E(S_n(t)))\}\right]^2.$$

Cherchons alors une majoration de chacun de ces trois termes, qu'on notera respectivement B_1, B_2 et B_3.

$$\frac{B_1}{3} = \frac{1}{(E(S_n(t)) + \frac{1}{n})^2}\,E[(f_n(t) - E(f_n(t)))^2].$$

$E(S_n(t)) + \frac{1}{n} \longrightarrow S(t)$, quand $n \longrightarrow \infty$, donc pour tout t fixé, tel que $S(t) > 0$:

$$B_1 = 0\ (\mathrm{Var}(f_n(t)).$$

$$\frac{B_2}{3} = \left[E\left(\frac{fn(t)}{E(S_n(t)) + \frac{1}{n}} - \frac{f_n(t)}{E(S_n(t)) + \frac{1}{n}}\right)\right]^2.$$

$$= \frac{1}{(E(S_n(t)) + \frac{1}{n})^2}\left[E\left(f_n(t)\,\frac{S_n(t) - E(S_n(t))}{S_n(t) + \frac{1}{n}}\right)\right]^2$$

$$\leq \frac{1}{(S(t) - \varepsilon)^4}\,E(f_n{}^2(t))\,E[(S_n(t) - E(S_n(t)))^2].$$

avec $\varepsilon > 0$ tel que $E(S_n(t)) + \frac{1}{n} > S(t) - \varepsilon$ et $S_n(t) + \frac{1}{n} > S(t) - \varepsilon$ p.s, pour $n \geq \eta > 0$.

Donc, $B_2 = 0\ (q(n)^2\ \mathrm{Var}\ (S_n(t)))$, puisque $f_n(t) = 0\ (q(n))$.

Concernant le dernier terme, on a:

$$\frac{B_3}{3} = \left[\frac{1}{E(S_n(t)) + \frac{1}{n}} \{h_n(t)\,(S_n(t) - E(S_n(t)))\}\right]^2.$$

$$\leq \frac{K\, q(n)^2}{(S(t) - \varepsilon)^4} \text{Var}\,(S_n(t)) \quad \text{avec } K>0,$$

puisque, $h_n(t) = 0\,(\frac{q(n)}{S(t) - \varepsilon})$, finalement on obtient:

$$\frac{B_3}{3} = 0\,(\,q(n)^2\ \text{Var}\,(S_n(t))), \quad \text{et par suite}$$

$$\text{Var}\,(h_n(t)) = 0\,(\text{Var}(f_n(t))) + 0\,(q(n)^2\ \text{Var}(S_n(t))).$$

11) Preuve du corollaire 5

En appliquant le lemme 2, on trouve:

$$\text{Var}\,(S_n(t)) = \frac{1}{n} S(t)^2 \int_0^t \frac{h(s)}{1-L(s)} + 0\,(\frac{1}{n}).$$

Il suffit d'écrire: $S_n(t) - E(S_n(t)) = S_n(t) - S(t) + S(t) - E(S_n(t))$

$$= \left(\int_0^t dF(u) - \int_0^t dF_n(u)\right) + E\left(\int_0^t dF_n(u) - \int_0^t dF(u)\right).$$

et utiliser (17), en remplaçant la fonction m par la fonction constante 1 et b par t.

De la même manière, si on remplace m(.) par $W_n(x,.)$, on obtient:

$$\text{Var}\,(f_n(t)) = \frac{1}{n} \int_0^b \left(S(s)\, W_n(x,s) + \int_s^b S'(s)\, W_n(x,t)\ dt\right)^2 \frac{h(s)}{1-L(s)} ds + 0\,(\frac{q(n)^2}{n})$$

$$= 0\,(\frac{q(n)^2}{n}).$$

IV - Commentaires

En utilisant les convergences uniformes p.s., ponctuelle et L^2, l'estimateur f_n peut être comparer à f_n^*, estimateur construit par la méthode du noyau et étudié par Földes et Al (1981), Mielniczuk (1983) et Lo, 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^* ,Földes et Al

(1981) ont obtenu cette convergence, en imposant plus de conditions sur les distributions des X_i° et C_i.

* Mielniczuk (1985) et Lo, 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:

$$V(f_n^*(x)) = \frac{A_1}{nh_n}\frac{f(x)}{1-G(x)} + 0\left(\frac{1}{n\,h_n}\right)$$

avec $h_n \longrightarrow 0$ et $n\,h_n \longrightarrow \infty$ (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) est la base trigonométrique, on montre que:

$V(f_n(x)) \sim \frac{q(n)^2}{n}\frac{f(x)}{1-G(x)} A_2$ (voir 11) § III). Il suffit donc, de choisir $h_n = q(n)^{-2}$, pour avoir la même vitesse de convergence des deux variances.

Si en plus la densité f 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 $L^2_{[o,b]}$ dans le cas de la base trigonométrique par exemple, on montre d'après (18), en choisissant q(n) impair de la forme $q(n) = 2\,\overline{q(n)}+1$, que

$$\sum_{i=1}^{q(n)} E[(\hat{a}_i - a_i)^2] = \frac{2}{nb}\sum_{i=1}^{\overline{q(n)}}\int_0^b \left[S^2(s)\,S(s)\,S(b) + \left(\int_s^b S'(t)\sin\left(\frac{2i\pi}{b}\left(t-\frac{b}{2}\right)\right)dt\right)^2\right.$$

$$\left. + \left(\int_s^b S'(t)\cos\left(\frac{2i\pi}{b}\left(t-\frac{b}{2}\right)\right)dt\right)^2\right]\frac{h(s)}{1-L(s)}\,ds + 0\left(\frac{\overline{q(n)}^3}{n^2}(\log n)^4\right) + 0\left(\overline{q(n)}^2\frac{(\log n)^4}{n^{3/2}}\right)$$

Le premier terme est un $0\left(\frac{\overline{q(n)}}{n}\right)$ et les deux derniers sont des $0\left(\frac{\overline{q(n)}}{n}\right)$, pourvu que l'on choisisse $\overline{q(n)}$ tel que $\frac{\overline{q(n)}\ (\log n)^2}{\sqrt{n}}$ tende vers zéro. Donc le MISE asymptotique s'écrit:

$$\text{AMISE} = \left(\frac{\overline{q(n)}}{n}\right) + 0\left(\frac{1}{\overline{q(n)}}\right)$$

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 p. 106 , on montre que:

$$\forall k,\quad a_{2k}^2 + a_{2k+1}^2 = 0\left(\frac{1}{k^2}\right),\quad \text{donc}\quad \sum_{i=q(n)+1}^{\infty} a_i^2 = 0\left(\frac{1}{\overline{q(n)}}\right).$$

Finalement, on a le MISE asymptotique qui équivaut à un $0\left(\frac{1}{\overline{q(n)}}\right)$, sous la contrainte imposée à $\overline{q(n)}$. 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) 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 Lo, 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 hn 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 basée sur des simulations.

V - Appendice

1) Preuve du lemme 1

On suppose que les (e_i), $i>1$, 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).

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

$$= \int_0^b \Big(\frac{1}{b}\, S(s)\frac{\sin\{q(n)\frac{\pi}{b}(x-s)\}}{\sin\{\frac{\pi}{b}(x-s)\}} + \frac{1}{b}\int_0^b S'(t)\,\frac{\sin\{q(n)\frac{\pi}{b}(x-t)\}}{\sin\{\frac{\pi}{b}(x-t)\}}\, dt\Big)^2 \frac{h(s)}{1-L(s)} ds.$$

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

$$B_n^2 = q(n)\int_R g_n(u)du.$$

où,

$$g_n(u) = \frac{1}{\pi b} 1_{\{q(n)\frac{x-b}{b}\pi,\, q(n)\frac{x}{b}\pi\}} \Big[S\Big(x - \frac{bu}{q(n)\pi}\Big)\frac{\sin u}{q(n)\sin(\frac{u}{q(n)})}$$

$$+ \frac{1}{q(n)}\int_{x-\frac{bu}{q(n)\pi}}^b S'(t)\,\frac{\sin q(n)\frac{\pi}{b}(x-t)}{\sin\frac{\pi}{b}(x-t)}dt\Big]^2 \frac{h(x-\frac{bu}{q(n)\pi})}{1-L(x-\frac{bu}{q(n)\pi})}.$$

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

$$g_n^1(u) = \frac{1}{\pi b} 1_{[-a\pi,0]}(u) \Big[S\Big(x - \frac{bu}{q(n)\pi}\Big)\frac{\sin u}{q(n)\sin(\frac{u}{q(n)})}$$

$$+ \frac{1}{q(n)}\int_{x-\frac{bu}{q(n)\pi}}^b S'(t)\,\frac{\sin q(n)\frac{\pi}{b}(x-t)}{\sin\frac{\pi}{b}(x-t)}dt\Big]^2 \frac{h(x-\frac{bu}{q(n)\pi})}{1-L(x-\frac{bu}{q(n)\pi})}.$$

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

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

(C est une constante positive), car si, $-a\pi \leq u \leq 0$, alors $-a\pi \leq \frac{u}{q(u)} < 0$,

$$\frac{\sin \frac{u}{q(n)}}{\frac{u}{q(n)}} \in \left[\frac{\sin a\pi}{a\pi}, 1\right] \text{ et } x \leq x - \frac{bu}{q(n)\pi} \leq x + a\, \frac{b}{q(n)} \leq x + a.$$

On utilise aussi pour cette majoration, le fait que:

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

La fonction g_n^1 est donc majorée par une fonction integrable. 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 C f 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:

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

et $\int g_n^1(u)\, du \xrightarrow[n \dashrightarrow \infty]{} \frac{1}{\pi} S(x)^2 \frac{h(x)}{1-L(x)} \int_{-a\pi}^{0} \left(\frac{\sin u}{u}\right)^2 du > 0.$

(On utilise le théorème de Lebesgue).

Donc , $B_n^2 = q(n) \int_{\mathbb{R}} g_n(u)\, du > q(n) \int_{\mathbb{R}} g_n^1(u)\, du,$

et par suite, $\exists\, A>0 \,/\, B_n^2 > q(n)\, A$.

2) Preuve du lemme 2

1) En faisant une intégration par parties, on a:

$$C(m,b) = m(b)\,(F_n(b)-F(b)) - \int_0^b m'(t)\,(F_n(t)-F(t))\,dt.$$

$$= m(b)\,P_n(b) - \int_0^b m'(t)\,P_n(t)\,dt + R_n^*.$$

(voir décomposition (9) de F_n-F), avec

$$(19) \qquad R_n^* = m(b)\,R_n(b) - \int_0^b m'(t)\,R_n(t)\,dt.$$

D'où la majoration de R_n^* donnée dans le lemme. Montrons ensuite que

$C(m,b) - R_n^*$, s'écrit sous la forme: $\frac{1}{n}\sum_{i=1}^{n} Z_i$.

2) D'après (10) et (11) on a:

$$(20) \qquad nP_n(t) = \sum_{i=1}^{n} S(t) \int_0^{X_i \wedge t} \frac{h(s)}{1-L(s)}\,ds - \sum_{i=1}^{n} \frac{S(t)}{1-L(X_i)}\,1_{\{(Xi\le t)\cap(\delta i=1)\}}\,, \text{ soit}$$

$$(21) \qquad nP_n(t) = S(t)\,\Big(\int_0^t u(s)ds - \mu(]0,t[)\Big).$$

En posant, $u(t) = \Big\{\sum_{i=1}^{n} (n-i+1)\,1_{[X(i-1),X(i)[}(t)\Big\}\,\frac{h(t)}{1-L(t)}$.

et $\mu = \sum_{i=1}^{n} 1_{\{\delta i=1\}}\,\frac{1}{1-L(X_i)}\,\delta_{Xi}$. On obtient, d'après (21):

$$n\int_0^b P_n(t)\,m'(t)\,dt = \int_0^b \big(S(b)\,m(b) - S(t)\,m(t)\big)\,u(t)\,dt - \int_0^b \big(S(b)\,m(b) - S(t)\,m(t)\big)\,d\mu(t)$$

$$- n\int_0^b \frac{S'(t)\,m(t)\,P_n(t)}{S(t)}\,dt$$

$$= n\,m(b)\,P_n(b) + \sum_{i=1}^{n} \int_0^b S'(t)\,m(t)\Big(\int_0^{X_i\wedge t} \frac{h(s)}{1-L(s)}\,ds - \frac{1_{\{(Xi\le t)\cap(\delta i=1)\}}}{1-L(X_i)}\Big)\,dt$$

$$- \sum_{i=1}^{n} \Big\{ \int_0^{X_i \wedge t} S(s)\, m(s)\, \frac{h(s)}{1-L(s)} ds \Big\} + \sum_{i=1}^{n} S(X_i)\, m(X_i)\, \frac{1_{\{(\delta i=1)\cap(Xi<b)\}}}{1-L(X_i)}$$

On en déduit alors le résultat cherché.

3) Les v. a. X_i, i.i.d ont la densité l égale à f(1-G) + g(1-F), puisque leur f.d.r est 1-(1-F)(1-G) = L. D'autre part la mesure ν définie par $\nu(A) = P(X_i \in A,\ \delta_i=1)$ admet la densité f(1-G). Donc,

$$E(Z_i) = \int_0^{\infty} \Big(\int_0^{s\wedge b} S(t)\, m(t) \frac{h(t)}{1-L(t)} dt \Big) \ell(s)\, ds - \int_0^{\infty} S(s)\, m(s) \frac{1_{\{s\le b\}}}{1-L(s)} f(s)\,(1-G(s))\, ds$$

$$+ \int_0^{\infty} \int_0^{b} S'(t)\, m(t) \Big(\int_0^{s\wedge b} \frac{h(u)}{1-L(u)} du \Big) dt \Big\} \ell(s)\, ds$$

$$- \int_0^{\infty} \Big(\int_0^{b} S'(t)\, m(t)\, \frac{1_{\{s\le t\}}}{1-L(s)} dt \Big) f(s)\,(1-G(s))\, ds.$$

On peut écrire le premier terme sous la forme:

$$\int_0^{b} \Big(\int_t^{\infty} \ell(s)\, ds \Big) S(t)\, m(t)\, \frac{h(t)}{1-L(t)} dt \qquad \text{(Théorème de Fuibini).}$$

il annule donc le second terme. En integrant par parties le troisième, on trouve:

$$\int_0^{\infty} \int_0^{b} S'(t)\, m(t)\, 1_{\{s\le t\}}\, h(s)\, dt\, ds,$$ ce qui annule le dernier et donc les Z_i sont bien centrés.

Pour calculer $E(Z_i^2)$, on pose: $Z_i = A_i + B_i$, avec

$$A_i = - \int_0^{X_i \wedge b} S(t)\, m(t) \frac{h(t)}{1-L(t)} dt - \int_0^{b} S'(t)\, m(t) \Big(\int_0^{X_i \wedge t} \frac{h(u)}{1-L(u)} du \Big) dt.$$

et $$B_i = S(X_i)\, m(X_i)\, \frac{1_{\{(\delta i=1)\cap(Xi\le b)\}}}{1-L(X_i)} + \int_0^{b} S'(t)\, m(t) \frac{1_{\{(\delta i=1)\cap(Xi\le t)\}}}{1-L(X_i)} dt.$$

donc, $E(Z_i^2) = E(A_i^2) + 2\, E(A_i\, B_i) + E(B_i^2)$.

$$* E(A_i^2) = \int_0^\infty \Big(\int_0^{s\wedge b} S(t)\, m(t) \frac{h(t)}{1-L(t)} dt \Big)^2 \ell(s)\, ds$$

$$+ 2 \int_0^\infty \Big\{ \int_0^{s\wedge b} S(t)\, m(t) \frac{h(t)}{1-L(t)} dt \Big\} \Big\{ \int_0^b S'(t)\, m(t) \int_0^{s\wedge t} \frac{h(u)}{1-L(u)} du\, dt \Big\} \ell(s)\, ds$$

$$+ \int_0^\infty \Big\{ \int_0^b S'(t)\, m(t) \Big(\int_0^{s\wedge t} \frac{h(u)}{1-L(u)} du \Big) dt \Big\}^2 \ell(s)\, ds.$$

$$* E(A_i B_i) = - \int_0^\infty \Big(\int_0^{s\wedge b} S(t)\, m(t) \frac{h(t)}{1-L(t)} dt \Big) \Big(\frac{S(s)\, m(s)\, 1_{\{s\le t\}}}{1-L(s)} \Big) f(s)\, (1-G(s))\, ds$$

$$- \int_0^\infty \Big(\int_0^{s\wedge b} S(t)\, m(t) \frac{h(t)}{1-L(t)} dt \Big) \Big(\int_0^b S'(t)\, m(t) \frac{1_{\{s\le b\}}}{1-L(s)} dt \Big) f(s)\, (1-G(s))\, ds$$

$$- \int_0^\infty \Big\{ \int_0^b S'(t)\, m(t) \Big(\int_0^{s\wedge t} \frac{h(u)}{1-L(u)} du \Big) dt \Big\} \Big\{ S(s)\, m(s) \frac{1_{\{s\le b\}}}{1-L(s)} \Big\} f(s)\, (1-G(s))\, ds$$

$$- \int_0^\infty \Big\{ \int_0^b S'(t)\, m(t) \Big(\int_0^{s\wedge t} \frac{h(u)}{1-L(u)} du \Big) dt \Big\} \Big\{ \int_0^b S'(t)\, m(t) \frac{1_{\{s\le t\}}}{1-L(s)} dt \Big\} f(s)\, (1-G(s))\, ds.$$

$$* E(B_i^2) = \int_0^\infty \Big(S(s)\, m(s) \frac{1_{\{s\le b\}}}{1-L(s)} \Big)^2 f(s)\, (1-G(s))\, ds$$

$$+ 2 \int_0^\infty \Big(S(s)\, m(s) \frac{1_{\{s\le b\}}}{1-L(s)} \Big) \Big(\int_0^b S'(t)\, m(t) \frac{1_{\{s\le t\}}}{1-L(s)} dt \Big) f(s)\, (1-G(s))\, ds$$

$$+ \int_0^\infty \Big(\int_0^b S'(s)\, m(s) \frac{1_{\{s\le t\}}}{1-L(s)} dt \Big)^2 f(s)\, (1-G(s))\, ds.$$

Une intégration par partie de tous les termes de $E(A_i^2)$, nous permet de conclure que, $E(A_i^2) = -2\, E(A_i B_i)$, d'où le résultat, puisque:

$$E(Z_i^2) = E(B_i^2) = \int_0^b \left(S(s)\, m(s) + \int_s^b S'(t)\, m(t)\, dt\right)^2 \frac{h(s)}{1-L(s)}\, ds.$$

Pour la suite de la démonstration, on se réfère à, (M.Delecroix et O.Yazourh (1992), lemme 1). On obtient finalement:

(22) $$\left|E(C(m,b))\right| = 0\left\{N(m,b)\,\frac{(\log n)^2}{n}\right\}.$$

(23) $$E\{[C(m,b)]^2\} = \frac{1}{n} V(Zi) + E[(R_n^*)^2] + 2\, E\left[\, R_n^* \frac{1}{n} \sum_1^n Z_i \,\right]$$

avec

(24) $$[E(R_n^*)^2] = 0\left\{\, N(m,b)^2\, \frac{(\log n)^4}{n^2}\right\}$$

(25) $$\frac{1}{n} E\left[R_n^* \sum_1^n Z_i\right] = 0\left\{N(m,b)\, \frac{(\log n)^2}{n} \left(\frac{1}{n} V(Z_i)\right)^{1/2}\right\}$$

D'où la conclusion.

BIBLIOGRAPHIE

DELECROIX M., YAZOURH O. (1992)

"Estimation non paramétrique du taux de hasard en présence de censures droites: la méthode des fonctions orthogonales", Statistique et analyse des données, à paraître.

DROESBEKE J.J., FICHET B., TASSI P., éditeurs

"Analyse statistique des durées de vie." - Economica (1989).

EFRON B. (1967)

"The two sample problem with censored data." - Proc. 5th Berkeley symp, Vol. 4, p 831-853.

FOLDES A., REJTO L., WINTER B.B. (1981)

"Strong consistency properties of nonparametric estimators for randomly censored data II. Estimation of density and failure rate"- Period. Math. Hungar 12, p 15-29.

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.

HJORT N.L. (1985)

"Discussion contribution to Andersen and Borgan's review article." Scand. J. Statis. - 12 - p 141-150.

KAPLAN E., MEIER P. (1958)

"Nonparametric estimation from incomplete observations." - JASA 53, p 457-481.

KIMURA D.K. (1972)

"Fourier series methods for censored data." - Thèse soutenue à l'Université de Washington.

LO S.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 - p 461-473.

REID N. (1981)

"Influence Functions for censored data." - The annals of Stat., Vol. 9 n° 1, p 78-92.

MIELNICZUK J. (1983)

"Properties of some kernel estimators and of the adopted Loftgarten-Quesenberry estimator of a density function for censored data." - Periodica Math. Hungarica, Vol. 16 (2), p 69-80

SANSONE G. (1959)

"Orthogonal functions" Pure and Applied Mathematics, Vol IX, Interscience Publishers, New York.