REVUE DE STATISTIQUE APPLIQUÉE

SALIMA HASSANI
PASCAL SARDA
PHILIPPE VIEU

Approche non paramétrique en théorie de la fiabilité : revue bibliographique

Revue de statistique appliquée, tome 34, n° 4 (1986), p. 27-41
<http://www.numdam.org/item?id=RSA_1986__34_4_27_0>

APPROCHE NON PARAMÉTRIQUE EN THÉORIE DE LA FIABILITÉ : REVUE BIBLIOGRAPHIQUE

Salima HASSANI, Pascal SARDA, Philippe VIEU
Université Paul Sabatier, Toulouse

A la mémoire de Gérard COLLOMB
qui fut à l'initiative de ce travail

RÉSUMÉ

Nous passons en revue les estimateurs non paramétriques de la fonction de hasard proposés à ce jour. Nous donnons quelques exemples pratiques d'utilisation de ces estimateurs, puis les principaux résultats les concernant selon que le problème est à données non censurées ou à données censurées. Dans chacun de ces deux cas nous étudions séparément les estimateurs utilisant l'échantillon ordonné des observations et ceux construits directement à partir d'estimateurs de la densité. Les résultats concernent le biais, la déviation maximale et la convergence de ces estimateurs pour des observations indépendantes. Un résultat de convergence pour des observations dépendantes établi parallèlement à cette étude est donné dans la dernière partie.

ABSTRACT

We survey non-parametric estimators of the hazard function proposed at this day. We give some practical examples of application of these estimators, then principal results for them, with uncensored data or censored data. In each case we study estimators built with a orderly sample of observations and those built from density estimators. The results are about the bias, maximum deviation and convergence of these estimators from independant observations. A result of convergence from dependant observations established parallely with this study is given in the last part.

Mots clés : **Fonction de hasard, Taux de défaillance, Estimation non-paramétrique, Fiabilité.**

1. INTRODUCTION

On s'intéresse par exemple à la durée de vie de lampes d'éclairage. En prenant l'origine des temps au moment de la mise en service de la lampe on cherche à évaluer la probabilité qu'une lampe tombe en panne entre les instants t et t + Δt sachant qu'elle fonctionne encore au temps t ($\Delta t > 0$ est pris au sens physique du terme). On note X la variable aléatoire réelle positive « durée de vie d'une lampe ». On désigne par f (resp. F) la densité (resp. la fonction de répartition) de X.

Pour t fixé tel que $F(t) < 1$ on a

$$P[t < X < t + \Delta t \mid X > t] = \frac{P[t < X < t + \Delta t]}{P[X > t]}$$

$$= \int_t^{t+\Delta t} f(u)\, du \,/\, (1 - F(t)),$$

et dès que la densité f est continue on a,

$$\lim_{\Delta t \to 0} P[t < X < t + \Delta t \mid X > t] / \Delta t = f(t) / (1 - F(t)).$$

Pour une variable aléatoire réelle X positive de répartition F et de densité f on appelle fonction de hasard, ou parfois taux de hasard, hasard ou bien taux de défaillance, la fonction h définie par

$$h(t) = f(t) / (1 - F(t)), \qquad \forall\, t \in \mathbf{R}^{+*},\ F(t) \neq 1. \tag{1.1}$$

L'interprétation que l'on peut en faire dans le cas d'une densité f continue (voir exemple ci-dessus) révèle l'importance de l'étude de la fonction de hasard dans de nombreux problèmes pratiques, notamment dans les études de fiabilité ou de survie.

On appelle fonction de hasard cumulée la fonction H définie par

$$H(t) = \int_0^t h(u)\, du \quad , \qquad \forall\, t \in \mathbf{R}^{+*},\ F(t) \neq 1. \tag{1.2}$$

Dans les revues de langue anglaise H porte le nom de « Log-survival function » puisque,

$$H(t) = -\operatorname{Log}(1 - F(t)),\ \forall\, t \in \mathbf{R}^{+*},\ F(t) \neq 1. \tag{1.3}$$

Le problème de l'estimation de la fonction h a donné lieu à de nombreuses études dans le cadre de la statistique paramétrique. Cependant, dans la pratique, il est parfois difficile de trouver le modèle paramétrique convenable. Les méthodes non paramétriques lèvent cette difficulté puisqu'aucune hypothèse restrictive n'est faite sur la loi de X. Cette loi n'est pas supposée appartenir à une classe de lois indexées par un nombre fini de paramètres réels, mais seulement vérifier des hypothèses très générales de régularité. Le cas le plus général est celui où cette loi admet une densité nulle sur $\mathbf{R}^-$ et continue sur $\mathbf{R}^+$; pour certains résultats nous aurons toutefois besoin d'hypothèses plus fortes. Il est à noter cependant que ces méthodes nécessitent un nombre important d'observations (disons 1 000 et plus). Les résultats que nous exposons dans la suite trouveront des applications immédiates dans des domaines où ce nombre important de données existe, le plus souvent dans le domaine médical (la fonction de hasard correspond au risque instantané de décès) ou géophysique (cf. exemple 1 ci-dessous). Le problème de défaillance des logiciels est un exemple d'étude où, au moins à ce jour, les modèles paramétriques seront préférables à cause du peu d'observations disponibles. D'autre part, l'estimation non paramétrique de la fonction de hasard peut permettre de choisir un modèle paramétrique convenable et donc de déboucher sur une étude classique.

Nous nous proposons dans cet article de passer en revue les différents travaux effectués à ce jour en estimation non paramétrique de la fonction de hasard. Ce travail complète la revue bibliographique de SINGPURWALLA et WONG (1983a).

Nous donnons dans le paragraphe 2 quelques applications de l'estimation de h selon que les données sont censurées ou non (définitions au § 2). Dans le paragraphe 3 nous donnons des généralités sur les noyaux puisque

ceux-ci interviennent dans la plupart des estimateurs non paramétriques introduits.

Ces estimateurs sont donnés dans le paragraphe 4 (resp. 5) ainsi que les résultats les concernant à partir de données non censurées (resp. censurées) et indépendantes. Dans chacun de ces paragraphes les estimateurs sont classés selon qu'ils sont construits à partir de l'échantillon ordonné ou bien directement à partir d'estimateurs de f et F. Cette classification relativement arbitraire trouve son origine dans l'état actuel du développement de l'estimation non paramétrique du taux de défaillance. Nous donnons dans le paragraphe 6 quelques nouveaux résultats qui ont été établis parallèlement à cette étude et qui concernent le cas d'observations dépendantes. Nous donnons enfin dans le paragraphe 7 quelques références sur des études par simulation.

2. EXEMPLES D'UTILISATION

Nous donnons brièvement deux exemples d'utilisation de l'estimation de h. Nous renvoyons aux auteurs pour plus de détails. Les estimateurs utilisés dans ces exemples sont définis aux paragraphes 4 et 5.

2.1. Problème à données non censurées

On dit que le problème est à données non censurées lorsque l'on peut observer toutes les variables.

RICE et ROSENBLATT (1976) traitent un problème relevant de la géophysique (secousses sismiques). La fonction de hasard h (t) représente le risque d'avoir une nouvelle secousse au temps t, l'instant 0 étant pris au moment de la dernière secousse enregistrée. Ils disposent des données des 4 764 dernières secousses enregistrées en Californie. Le résultat de leur étude montre que ce risque est très élevé dans les dix premières minutes puis décroît pour se stabiliser au bout de 25 h environ.

2.2. Problème à données censurées

Dans beaucoup de problèmes pratiques on ne dispose pas de toutes les données $X_1, \ldots, X_n$ mais uniquement des couples (Z_i, J_i), $i = 1, \ldots, n$, où

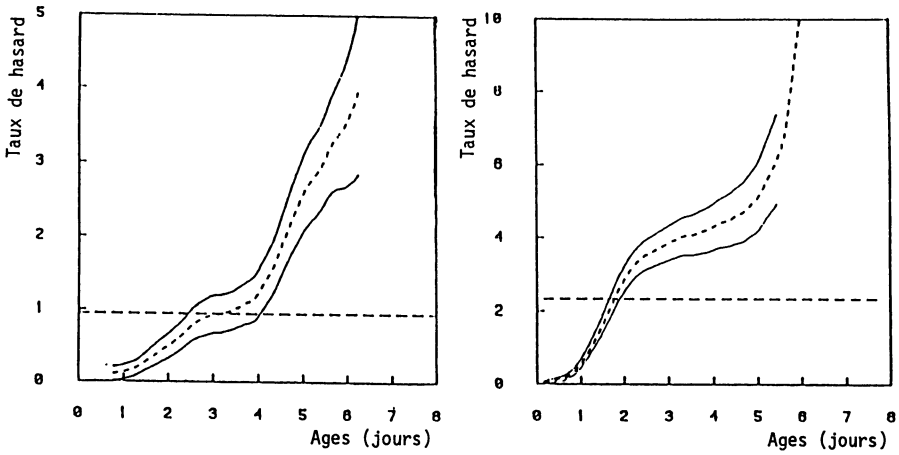
$$Z_i = \min (X_i, C_i),$$

et

$$J_i = \mathbf{1}_{[X_i \leq C_i]} = \mathbf{1}_{[Z_i = X_i]},$$

les C_i étant des variables aléatoires réelles positives. On dit qu'on a alors un problème à données censurées à droite par les variables $C_1, \ldots, C_n$.

YANDELL (1983) traite des données provenant d'une expérience de survie, pour rechercher l'effet d'une dose de 300 rad. d'une irradiation par les rayons gamma sur 2 534 souris. Les animaux sont répartis en deux groupes, « traitement » et « contrôle ». Ils meurent naturellement ou sont sacrifiés. X_i est ici la durée de vie (potentielle lorsque l'animal est sacrifié) et C_i la date à laquelle le $i^{ème}$ animal est sacrifié s'il est encore en vie. Les données sont constituées par le *moment* (c'est-à-dire $Z_i = \min(X_i, C_i)$) et la *nature* (qui peut s'exprimer par $J_i = \mathbf{1}_{[X_i \leq C_i]}$) du décès. Parmi les 1 080 souris du groupe de contrôle, 361 ont été sacrifiées ainsi que 343 parmi les 1 454 du groupe « traitement ». L'étude a pour but d'analyser le taux de mortalité dans le groupe de traitement tout au long de l'expérience et de le comparer à celui du groupe de contrôle. Les taux de mortalité obtenus montrent que l'irradiation augmente le taux de mortalité dans les deux premiers jours et au-delà du cinquième jour de traitement, alors que, entre les 2^e et 5^e jours, les taux semblent identiques. Ce résultat est illustré par les deux figures ci-dessous qui sont extraites de l'article de Yandell (1983) et qui représentent pour chacun des deux groupes, en pointillés, la fonction de hasard estimée en chaque point par $h_{n,9}$ (cf. définition (5.1)) ainsi que, en trait plein, des intervalles de confiance simultanés de niveau 80 %, intervalles obtenus à l'aide de la propriété de normalité asymptotique de $h_{n,9}$.



Groupe de Contrôle

Groupe Traitement

3. QUELQUES GÉNÉRALITÉS

La plupart des estimateurs non paramétriques de la fonction de hasard font intervenir la notion de noyau. Nous donnons tout d'abord la définition d'un noyau, ainsi que d'une δ-suite de fonctions qui peut être construite à partir d'un noyau.

On appelle δ-suite de fonctions (cf. WATSON et LEADBETTER (1964b)) une suite (δ_n) de fonctions réelles mesurables vérifiant les conditions suivantes

$$\begin{aligned}
&- \exists A < \infty, \int_{\mathbf{R}} |\delta_n(x)|\, dx < A, \forall n \in \mathbf{N}^*, \\
&- \int_{\mathbf{R}} \delta_n(x)\, dx = 1, \forall n \in \mathbf{N}^*, \\
&- \forall \lambda > 0, \delta_n(x) \underset{n \to \infty}{\longrightarrow} 0 \text{ uniformément sur } \{x, |x| \geqslant \lambda\}, \\
&- \forall \lambda > 0, \int_{|x| \geqslant \lambda} \delta_n(x)\, dx \underset{n \to \infty}{\longrightarrow} 0.
\end{aligned} \tag{3.1}$$

On appelle noyau de **R** une fonction K définie sur **R**, intégrable et bornée, qui vérifie

$$|x|\, K(x) \underset{|x| \to \infty}{\longrightarrow} 0. \tag{3.2}$$

On supposera dans la suite que le noyau est normé en ce sens que

$$\int_{\mathbf{R}} K(x)\, dx = 1.$$

A partir d'un noyau normé K ainsi défini on peut construire une δ-suite associée de fonctions (K_n) en posant

$$K_n(x) = (b_n)^{-1} K(x/b_n), \tag{3.3}$$

où (b_n) est une suite dans $\mathbf{R}^{*+}$ qui vérifie

$$\lim_{n \to \infty} b_n = 0.$$

Une δ-suite de fonctions (δ_n) (resp. un noyau K) est dite compatible avec une fonction de répartition ψ si pour tout $M > 0$, il existe $b > 0$ tel que pour tout x fixé (3.4)

$$b^{-1}\, \delta_n ((y-x)/b) / (1-\psi(y)) \text{ (resp. } b^{-1}\, K((y-x)/b) / (1-\psi(y))$$

est uniformément bornée sur $\{(x, y) : |y-x| > M\}$.

L'estimation de la fonction de hasard faisant intervenir des estimateurs de la densité f, signalons d'abord quelques revues bibliographiques concernant l'estimation non paramétrique de la densité : DEHEUVELS (1977), ERYER (1977), TAPIA et THOMPSON (1978), WERTZ (1978), BEAN et TSAKAS (1980). Notons enfin les travaux de ROSENBLATT (1956) et PARZEN (1962) qui ont introduit l'estimateur à noyau de la densité.

4. RÉSULTATS CONCERNANT LE CAS DE DONNÉES NON CENSURÉES

On suppose les variables aléatoires $X_1, \ldots, X_n$ indépendantes, définies sur le même espace probabilisé et ayant même loi que X. On note classiquement $X_{(1)}, \ldots, X_{(n)}$ les statistiques d'ordre.

4.1. Méthodes utilisant les statistiques d'ordre

Estimateur de WATSON et LEADBETTER

WATSON et LEADBETTER (1964 a, 1964 b) ont proposé l'estimateur suivant de h (x)

$$h_{n,0}(x) = \sum_{i=1}^{n} \delta_n(x - X_{(i)}) / (n - i + 1)$$

avec (δ_n) définie en (3.1). Ils établissent que le biais de $h_{n,0}$ tend vers 0 quand n tend vers l'infini dès que (δ_n) est compatible avec F.

Si de plus

$$u_n = 0(n) \qquad \text{avec} \qquad u_n = \int_R \delta_n^2(t)\, dt,$$

alors

$$\left(\frac{n}{u_n}\right) \text{Var}\, h_{n,0}(x) \underset{n\to\infty}{\to} h(x) / (1 - F(x)).$$

Cas particulier de l'estimateur $h_{n,0}$

Un cas particulier de $h_{n,0}$ est celui de l'estimateur noté $h_{n,1}$ défini par

$$h_{n,1}(x) = \sum_{i=1}^{n} K_n(x - X_{(i)}) / (n - i + 1)$$

avec K_n définie à partir d'un noyau K par (3.3). Le biais de cet estimateur proposé par RAMLAU-HANSEN (1983) converge donc vers 0.

L'auteur obtient la convergence en moyenne quadratique simple (resp. uniforme sur un compact) dès que l'on a

$$nb_n \underset{n\to\infty}{\to} \infty \qquad (\text{resp. } nb_n^2 \underset{n\to\infty}{\to} \infty).$$

SINGPURWALLA et WONG (1983b) donnent des résultats complémentaires sur le biais et la limite de l'erreur quadratique dès que le noyau K est symétrique (K (x) = K (−x), ∀x) et h est continuement différentiable à l'ordre m + 1 (m > 0) sur le compact [−c, +c]. Ils montrent alors

(i) $$\text{Biais}\, [h_{n,1}(x)] = \sum_{j=1}^{m} (b_n)^j h^{(j)}(x) \int_{-\infty}^{\infty} x^j K(x)\, dx/j! + O(b_n^m)$$

(ii) $$E(h_{n,1}(x) - h(x))^2 \underset{n\to\infty}{\to} (nb_n)^{-1} h(x)/(1 - F(x)) \int_{-c}^{c} K^2(u)\, du + b_n^{2m}(\lim_{n\to\infty} b_n^{-m}\, \text{Biais}\, h_{n,1}(x))^2$$

(iii) $$E(h_{n,1}(x) - h(x))^2_{\text{Optimal}} = O(n^{-2m/(2m+1)}).$$

Estimateur de RICE et ROSENBLATT (1976)

Cet estimateur est défini par

$$h_{n,2}(x) = \sum_{i=1}^{n} K_n(x - X_{(i)})\, \text{Log}\, [(n - i + 2)/(n - i + 1)].$$

RICE et ROSENBLATT (1976) montrent qu'il est asymptotiquement sans biais et qu'on a

$$h_{n,2}(x) = h_{n,1}(x) + O_p(n^{-1}).$$

SETHURAMAN et SINGPURWALLA (1981) donnent la propriété de convergence en loi dite de déviation maximale suivante

$$P(\beta_n(M_n - \alpha_n) \leqslant x) \underset{n\to\infty}{\longrightarrow} e^{-2e^{-x}} \tag{4.1}$$

sous l'hypothèse

$$n(b_n)^5 / \operatorname{Log} b_n \underset{n\to\infty}{\longrightarrow} 0$$

avec

$$M_n = \max_{b_n A \leqslant x \leqslant C} | \sqrt{(nb_n)((F(x)/1 - F(x))')^{-1}} (h_{n,2}(x) - h(x)) |.$$

Les suites réelles (α_n), (β_n) et les constantes A et C sont explicitées dans SETHURAMAN et SINGPURWALLA (1981).

Ce résultat (4.1) permet de définir directement des régions de sécurité asymptotique donnée pour la courbe de $h_{n,2}$.

Estimateurs de SETHURAMAN et SINGPURWALLA (1981)

L'estimateur ci-dessous, appelé estimateur trivial ou « naïf » a été proposé par SETHURAMAN et SINGPURWALLA (1981)

$$h_{n,3}(x) = \begin{cases} [(n - i + 1)(X_{(i)} - X_{(i-1)})]^{-1} & \text{si} \quad X_{(i-1)} \leqslant x < X_{(i)} \\ 0 & \text{si} \quad x \geqslant X_{(n)}. \end{cases}$$

Ces auteurs ont montré la non convergence de cet estimateur et en introduisant un noyau symétrique particulier tel que

$$\exists A > 0, \quad \forall u, \quad |u| \geqslant A, \quad K(u) = 0,$$

définissent un nouvel estimateur

$$h_{n,4}(x) = \frac{1}{b_n} \int K\left(\frac{x - z}{b_n}\right) h_{n,3}(z)\, dz \quad \text{pour} \quad x \geqslant b_n A.$$

Ils obtiennent ainsi sur $h_{n,4}$ le même résultat de déviation maximale que celui établi sur $h_{n,2}$ (4.1), en donnant un théorème général permettant d'établir facilement les résultats de type asymptotique pour $h_{n,4}$ à partir de ceux de $h_{n,2}$.

4.2. Méthodes utilisant les estimateurs de densité

Il est naturel d'estimer la fonction de hasard par

$$h_n(x) = f_n(x) / (1 - \hat{F}_n(x)) \tag{4.2}$$

où f_n et $\hat{F}_n$ sont respectivement des estimateurs de la densité f et de la fonction de répartition F. Quand $\hat{F}_n$ sera la fonction de répartition empirique on adoptera la notation F_n.

Estimateur de WATSON et LEADBETTER (1964a et b)

Cet estimateur, noté $h_{n,5}$, est défini par (4.2) avec

$$f_n(x) = n^{-1} \sum_{i=1}^{n} \delta_n(x - X_i)$$

et

$$\hat{F}_n(x) = \int_0^x f_n(t)\, dt\,.$$

WATSON et LEADBETTER (1964, a et b) montrent que le biais de $h_{n,5}$ tend vers 0 et obtiennent les deux résultats suivants

(i) $\mathrm{Var}\,(h_{n,5}(x)) \underset{n\to\infty}{\sim} (\alpha_n h(x))/(n(1 - F(x))$

(ii) $P((1 - F(x))(n/(c_n f(x)))^{1/2} (h_{n,5}(x) - h(x)) < t) \underset{n\to\infty}{\to} N(t)$

où N est la fonction de répartition de la loi normale réduite, (α_n) et (c_n) sont les suites supposées finies définies pour un $\varepsilon > 0$ par

$$\alpha_n = \int \delta_n^2(x)\, dx\,,$$

et

$$c_n = \int |\delta_n(x)|^{2+\varepsilon}\, dx\,,$$

avec l'hypothèse

$$c_n / n^{\varepsilon/2}\, c_n^{1+\varepsilon/2} \underset{n\to\infty}{\to} 0\,.$$

En remplaçant δ_n par K_n défini en (3.3) dans $h_{n,5}$, MURTHY (1965) établit la convergence presque sûre ponctuelle de $h_{n,5}$ sous l'hypothèse

$$nb_n \underset{n\to\infty}{\to} \infty\,,$$

et AHMAD (1976) établit la convergence presque sûre uniforme sur un compact (sur lequel $F(\cdot) < 1$) de cet estimateur sous l'hypothèse assez restrictive

$$\forall \gamma > 0, \quad \sum_{n=1}^{\infty} \exp\{-\gamma n b_n^2\} < \infty\,.$$

Notons que AHMAD (1976) a proposé un estimateur de la fonction de hasard généralisée définie par

$$h_{FG}(x) = f(x)(1 - G(x)) / (g(x)(1 - F(x)))\,,$$

où g (resp. G) est la densité (resp. la fonction de répartition) d'une loi connue.

Estimateurs à noyau

RICE et ROSENBLATT (1976) ont défini un estimateur $h_{n,6}$ à partir de (4.2) où

$$f_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K((x - X_i)/b_n)\,, \tag{4.3}$$

et $\hat{F}_n$ est la fonction de répartition empirique classique multipliée par le terme $n/(n + 1)$.

RICE et ROSENBLATT (1976) obtiennent sur $h_{n,6}$ le résultat de *déviation maximale* suivant,

$$P[(2 \log c_n)^{1/2} (M_n/((\alpha_n)^{1/2} - d_n)) < x] \underset{n\to\infty}{\longrightarrow} e^{-2e^{-x}}$$

avec

$$M_n = \max_{|x| \leqslant V_n} |(nb_n f^{-1}(x))^{1/2} (1 - F(x)) (h_{n,6}(x) - h(x))| .$$

Pour le détail des suites (c_n), (α_n), (d_n) et (v_n) on se reportera à RICE et ROSEMBLATT (1976, p. 69-71).

MURTHY (1965) avait introduit un estimateur très voisin.

Cet estimateur, noté $h_{n,7}$, est défini par (4.2) où f_n est donné par (4.3) et $\hat{F}_n$ est la fonction de répartition empirique classique.

La convergence ponctuelle presque sûre est établie par MURTHY (1965) sous l'hypothèse

$$nb_n \underset{n\to\infty}{\longrightarrow} \infty ,$$

et ce même auteur prouve la normalité asymptotique de $h_{n,7}$. La convergence uniforme presque complète sur un compact (sur lequel $F(\cdot) < 1$) est établie par COLLOMB *et al.* (1986) sous l'hypothèse

$$nb_n/\text{Log } n \underset{n\to\infty}{\longrightarrow} \infty .$$

Estimation par la méthode des k-points les plus proches

Cet estimateur, noté $h_{n,8}$, utilise F_n comme estimateur de la fonction de répartition et l'estimateur de f par la méthode des k-points les plus proches (LOFTSGAARDEN et QUESENBERRY (1965)) défini par

$$\tilde{f}_n(x) = (nR(k_n, x))^{-1} \sum_{i=1}^{n} K((x - X_i)/R(k_n, x)), \qquad \forall x \in \mathbf{R}^{+*}$$

où (k_n) est une suite entière positive vérifiant

$$\lim_{n\to\infty} k_n = \infty \qquad \text{et} \qquad \lim_{n\to\infty} n^{-1} k_n = 0$$

et $R(k_n, x)$ représente la distance euclidienne entre x et la $k_n^{\text{ième}}$ observation la plus proche de x. Dans COLLOMB *et al.* (1986) on obtient encore la convergence uniforme presque complète sur un compact lorsque

$$k_n/\text{Log } n \underset{n\to\infty}{\longrightarrow} \infty .$$

5. RÉSULTAT CONCERNANT LE CAS DE DONNÉES CENSURÉES

Pour tout ce qui concerne les variables $X, X_1, ..., X_n$ on adopte les mêmes notations qu'au paragraphe 4, et on note $\tilde{F}$ (resp. $\tilde{f}$) la fonction de répartition (resp. la densité) commune des variables $Z_1, ..., Z_n$, ainsi que G (resp. g) la fonction de répartition (resp. la densité) commune des variables $C_1, ..., C_n$.

On se place dans le cadre d'un modèle où les variables $X_1, ..., X_n$ sont indépendantes puisque cette hypothèse a été faite par tous les auteurs dont nous allons donner les résultats.

5.1. Méthode utilisant les statistiques d'ordre

TANNER et WONG (1983) ont introduit un estimateur basé sur les statistiques d'ordre $Z_{(1)}, ..., Z_{(n)}$ et qui généralise $h_{n,1}$ tel qu'il était défini pour des problèmes à données non censurées.

Cet estimateur est donné à partir d'une δ-suite de la forme (3.3) par

$$h_{n,9}(x) = (b_n)^{-1} \sum_{j=1}^{n} (n - j + 1)^{-1} J_{(j)} K((x - X_{(j)}/b_n) \qquad (5.1)$$

où l'on convient que

$$J_{(j)} = \mathbf{1}_{[Z_{(j)} = X_{(j)}]} .$$

Les auteurs établissent leurs résultats pour une suite (b_n) vérifiant

$$\lim_{n \to \infty} nb_n = \infty . \qquad (5.2)$$

Ils montrent que $h_{n,9}(x)$ est un estimateur de h dont le biais tend vers 0 dès que le noyau K est compatible avec F. Si le noyau K est aussi compatible avec G, ils montrent que $h_{n,9}(x)$ a une distribution asymptotique normale et que $h_{n,9}$ est convergent en moyenne quadratique. Ils donnent une évaluation asymptotique de la variance de $h_{n,9}$,

$$\text{Var } h_{n,9}(x) = (nb_n)^{-1} h(x) (1 - \tilde{F}(x))^{-1} \int K^2(t)\, dt + 0((nb_n)^{-1}) . \qquad (5.3)$$

Toujours pour une suite (b_n) vérifiant (5.2), YANDELL (1983) donne une évaluation exacte de la variance de $h_{n,9}$ ainsi que de la covariance entre $h_{n,9}(x)$ et $h_{n,9}(y)$. Ainsi sous certaines autres hypothèses supplémentaires il montre que

$$nb_n \text{ Cov } (h_{n,9}(x), \quad h_{n,9}(y)) \underset{n \to \infty}{\to} 0, \qquad \forall x \neq y . \qquad (5.4)$$

Utilisant l'estimateur défini par (5.1), YANDELL (1983) propose un test non paramétrique de l'hypothèse $H_0 \equiv$ « $h_1 = h_2$ » où h_1 et h_2 sont les fonctions de hasard relatives à deux échantillons indépendants.

5.2. Méthode utilisant un estimateur de densité

BLUM et SUSARLA (1980) ont proposé un estimateur de $h_{n,6}$ tel qu'il était défini pour des problèmes à données non censurées. En notant $\tilde{F}_n$ la répartition empirique des Z_i, cet estimateur est défini par

$$h_{n,10}(x) = (nb_n)^{-1} \left[\sum_{j=1}^{n} J_j K((x - Z_j)/b_n) \right] / (1 - \tilde{F}_n(x)) . \qquad (5.5)$$

Sous certaines hypothèses de continuité et de dérivabilité de f, F, g et G, pour un noyau K défini par (3.2) absolument continu tel que $\int K(x)\, dx < \infty$ et $\int (K'(x))^2 dx < \infty$, et lorsque la suite (b_n) vérifie

$$\exists\, \tau \in]0,1[\qquad \text{tel que} \qquad b_n = n^{-\tau}.$$

il est montré que

$$P\left[(c_n/B)\left(\frac{R_n}{A} - c_n\right) \leqslant u\right] \underset{n\to\infty}{\to} \exp(-2e^{-u}), \tag{5.6}$$

avec

$$R_n = \sup_{x\in]0,1[} [nb_n(1 - F(x))(1 - G(x))]^{1/2} \left[h_{n,10}(x) - \frac{E[J_1 K((x - Z_1)/b_n)]}{b_n(1 - F(x))(1 - G(x))}\right],$$

et $c_n = (2\tau \operatorname{Log} n)^{1/2} B$.

Nous renvoyons aux auteurs pour les détails sur les constantes A et B, qui ne dépendent que du noyau K.

Méthode utilisant l'estimateur de KAPLAN-MEIER

Soient $\{X_j^1, X_j^2, ..., X_j^k\}_{j=1}^{\infty}$, k suites indépendantes de variables aléatoires indépendantes et identiquement distribuées (k est un nombre fixé). On suppose que X_j^i a une densité f^i continue et une fonction de répartition F^i. BURKE et HORVATH (1984) estiment

$$h^i(t) = f^i(t)/(1 - F^i(t)), \qquad \forall i = 1, ..., k.$$

On pose

$$X_j = \min(X_j^1, X_j^2, ..., X_j^k) \qquad \text{et} \qquad \delta_j^i = \mathbf{1}_{[X_j = X_j^i]}.$$

L'estimateur $h_{n,11}^i$ de h^i qui utilise la fonction empirique de hasard cumulée

$$\Lambda_n^i(t) = n^{-1} \sum_{\{1<j\leqslant n, X_n<t\}} \delta_j^i/(1 - F_n(X_j)), \qquad \forall i = 1, ..., k$$

est défini par

$$h_{n,11}^i(t) = \int_A^B \psi_n^i(t, s)\, d\Lambda_n^i(s), \tag{5.7}$$

où A et B sont des nombres réels convenablement choisis, pouvant dépendre de i, et $\{\psi_n^i\}_{n=1}^{\infty}$ des suites de fonctions mesurables sur $(A, B)^2$. BURKE et HORVATH (1984) définissent trois estimateurs différents en prenant des suites (ψ_n^i) différentes. Nous prendrons pour toute la suite

$$\psi_n^i(t, s) = (b_n^i)^{-1} K^i((t - s)/b_n^i)$$

où (b_n^i) est une suite réelle telle que

$$b_n^i = n^{-\beta_i}, \qquad 1/3 < \beta_i < 2/3, \qquad \forall i = 1, ..., k,$$

et K^i, un noyau dont la dérivée est de carré intégrable sur le support de K^i pour $i = 1, ..., k$.

On pose

$$M_n^i = \sup_{C\leqslant t\leqslant D} [g_i(t)]^{-1/2} |h_{n,11}^i(t) - h^i(t)|, \qquad \forall i = 1, ..., k,$$

où

$$g_i(t) = f^i(t)(1 - F(t))^{-1}(1 - F^i(t))^{-1}$$

est supposée bornée sur [C, D] où $-\infty \leqslant A \leqslant C \leqslant D \leqslant B \leqslant \infty$.

Les auteurs précités donnent le résultat suivant

$$\lim_{n\to\infty} (2r_i \log(b_n^i)^{-1})^{-1/2} (b_n^i)^{1/2} M_n^i = 1 \qquad \text{p.s.}.$$

avec

$$r_i = \int_{-A_i}^{A_i} (K^i(t)')^2 \, dt, \qquad [-A_i, A_i] \text{ étant le support de } K^i.$$

Lorsque k = 1, on a un problème à données non censurées et on retrouve l'estimateur $h_{n,4}$.

Estimation par la méthode des k-points les plus proches

Avec les notations introduites pour l'estimateur $h_{n,8}$ on définit un estimateur des k-points les plus proches pour des données censurées par

$$h_{n,12}(x) = \frac{1}{R(k_n ; x)} \sum_{i=1}^{n} \frac{J_i}{n - i + 1} K\left[\frac{x - Z_{(i)}}{2R(k_n ; x)}\right].$$

La convergence ponctuelle presque sûre de cet estimateur a été établie par TANNER (1983) lorsque la suite (k_n) est de la forme $k_n = [n^\alpha]$ pour $1/2 < \alpha < 1$, h est continue et K est à support compact. Nous signalons aussi la note de SCHÄFER (1985) qui apporte une preuve simplifiée de ce même résultat.

Estimation par la méthode de l'histrogramme

Après avoir remarqué que h peut s'écrire sous la forme

$$h(x) = f^*(x)/(1 - \tilde{F}(x)),$$

où

$$f^*(x) = f(x)\,(1 - G(x)),$$

LIU and Van RYSIN (1985) proposent d'estimer h par

$$h_{n,13}(x) = f_n(x)/(1 - \tilde{F}_n(x)),$$

où $\tilde{F}_n$ est la fonction de répartition empirique des Z_i et f_n^* est un estimateur de f* qui constitue une généralisation de l'histogramme. Nous renvoyons à Van RYSIN (1973) pour la définition exacte de f_n^*. Dans LIU et Van RYSIN (1985) est établie la convergence presque sûre ponctuelle et la normalité asymptotique de $h_{n,13}$.

6. QUELQUES COMMENTAIRES ET DERNIERS RÉSULTATS

Outre les nombreux résultats sur l'évaluation asymptotique du biais et de la variance des estimateurs introduits, il a été prouvé que la plupart de ces estimateurs sont asymptotiquement non biaisés. Par contre il semble qu'il n'existe pas d'estimateur sans biais de h sous la seule hypothèse que f est continue. Ce résultat de non existence n'a pas encore été établi à notre connaissance mais il devrait être une conséquence du corollaire du théorème de BICKEL et LEHMAN (1969, p. 1525).

Par ailleurs pour tous les estimateurs introduits (exceptés $h_{n,0}$ et $h_{n,1}$) des résultats de déviation maximale ou de normalité asymptotique ont été établis.

Ces résultats sont intéressants puisqu'ils permettent de déterminer des intervalles de confiance, point par point et simultanés.

Pour ce qui concerne les problèmes de convergence d'estimateurs, les seuls résultats contenus dans les paragraphes 4 et 5 sont la convergence

— en moyenne quadratique, ponctuelle et uniforme pour $h_{n,1}$,
— en moyenne quadratique, ponctuelle pour $h_{n,10}$,
— presque sûre, ponctuelle pour $h_{n,5}$, $h_{n,7}$, $h_{n,12}$ et $h_{n,13}$,
— presque sûre, uniforme pour $h_{n,5}$.

Il faut donc constater que très peu de résultats de convergence forte ont été établis.

De même, mis à part RICE et ROSENBLATT (1976, p. 66) dans une remarque, aucun auteur ne s'est intéressé au problème de l'estimation du taux de hasard à partir d'observations dépendantes, si ce n'est COLLOMB *et al.* (1986).

Dans ce dernier article on démontre la convergence presque complète uniforme sur un compact C de $h_{n,7}$ et $h_{n,8}$ vers h lorsque (X_n) est un processus uniformément fortement mélangeant. (Nous renvoyons à BILLINGSLEY (1968) pour la définition).

Ces deux résultats de convergence uniforme sur un compact permettent d'estimer le mode θ de la fonction h (lorsque celui-ci est unique). Dans l'exemple introductif du paragraphe 1, θ représente l'instant auquel la lampe d'éclairage a le plus de chance de tomber en panne, sachant qu'elle ne l'a pas fait auparavant.

Remarquons que le modèle de dépendance introduit pour ces résultats englobe entre autres les processus markoviens qui vérifient la condition de Doeblin (voir DOOB (1953) p. 209).

7. QUELQUES RÉFÉRENCES POUR DES ÉTUDES PAR SIMULATION

RICE et ROSENBLATT (1976) donnent des courbes estimées de la fonction de hasard h par $h_{n,2}$ en simulant 200 nombres aléatoires suivant la loi exponentielle et la loi de RAYLEIGH (Loi de WEIBULL avec $\alpha = 1$, $\beta = 2$).

YANDELL (1983) à partir de simulations, traite les cas non censuré et censuré à 50 %, et obtient des courbes estimées de h, le nombre d'observations s'élevant à 200.

Nous signalons aussi les études de RACINE-POON et HOEL (1984). Ils utilisent dans leurs simulations une généralisation au cas censuré de l'estimateur de KAPLAN-MEYER. Ils mettent en évidence le risque que représente le choix d'un modèle non censuré en montrant que celui-ci peut aboutir à une augmentation importante du biais lorsque les observations sont effectivement censurées.

RÉFÉRENCES BIBLIOGRAPHIQUES

I.A. AHMAD (1976). — Uniform strong convergence of the generalized failure rate estimated. *Bull. Math. Stat, 17,* p. 77-84.

S.J. BEAN et C.P. TSAKAS (1980). — Developments in non-parametric density estimation. *Inter. Stat. Review, 48,* p. 267-287.

P.J. BICKEL et E.L. LEHMAN (1969). — Unbiased estimate in convexe families. *Annals of Mathematical Statistics, vol. 40,* p. 1523-1525.

P. BILLINGSLEY (1968). — Convergence of Probability Measures. *Wiley.*

S.R. BLUM et V. SUSARLA (1980). — Maximal deviation theory of density and failure rate function estimates based on censored data. *Multivariate Analysis, V,* p. 213-222.

M.D. BURKE et L. HORVATH (1984). — Density and failure rate estimation in a Competing risks model. *Sankya, vol. 46, Série A,* p. 135-154.

G. COLLOMB, S. HASSANI, P. SARDA et P. VIEU (1986). — Estimation non paramétrique de la fonction de hasard pour des observations dépendantes. *Statistique et Analyse des données, vol. 10, n° 3,* pp. 42-49.

P. DEHEUVELS (1977). — Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée, vol. XXV, n° 3.*

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

J. ERYER (1977). — Revue of some non-parametric methods of density estimation. *J. Inst. Math. Applic. 20,* p. 335-354.

R. LIU et J. VAN RYZIN (1985). — A histogram estimator of the hazard rate with censored data. *Ann. Stat. 13,* p. 592-605.

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

V.K. MURTHY (1965). — Estimation of jumps, reliability and hazard rate. *Ann. Stat. 36,* p. 1032-1040.

E. PARZEN (1962). — On estimation of probability density function and mode. *Annals of Mathematical Statistics, 33,* p. 1065-1076.

A.H. RACINE-POON et D.G. HOEL (1984). — Non-parametric estimation of the survival function when cause of death is uncertain. *Biometrics 40,* p. 1151-1158.

H. RAMLAU-HANSEN (1983). — Smoothing counting processes intensities by means of kernel functions. *Ann. Stat. 11,* p. 453-466.

J. RICE et M. ROSENBLATT (1976). — Estimation of the log survivor function and hazard function. *Sankhya, 38 A,* p. 60-78.

M. ROSENBLATT (1956). — Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics, Vol. 27,* p. 642-669.

H. SCHÄFER (1985). — A note on data-adaptive kernel estimation of the hazard and density function in the random censorship situation. *Annals of Statistics, Vol. 13, n° 2,* p. 818-820.

J. SETHURAMAN et N.D. SINGPURWALLA (1981). — Large sample estimates and uniform confidence bounds for the failure rate function based on a naive estimator. *Ann. Stat. 9,* p. 628-632.

N.D. SINGPURWALLA et M.Y. WONG (1983 a). — Estimation of the failure rate : a survey of non-parametric methods. Part I : Non bayesian Methods. *Commun. Statist. Theor. Math., 12 (5),* p. 559-588.

N.D. SINGPURWALLA et M.Y. WONG (1983 b). — Kernel estimators of the failure rate function and density estimation : an analogy. *J.A.S.A. 83, Vol. 78, n° 382.*

M.A. TANNER (1983). — A note on the variable kernel estimator of the hazard function from randomly censored data. *Ann. Stat., Vol. 11, n° 3,* p. 994-998.

M. TANNER et W.H. WONG (1983). — The estimation of the hasard function from randomly censored data by the kernel method. *Ann. Stat. 11,* p. 989-993.

R.A. TAPIA et J.R. THOMPSON (1978). — Non-parametric probability density estimation. *Baltimore, London : Johns Hopkins University Press.*

J. VAN RYSIN (1973). — A histogram method of density estimation. *Comm. Statist. 2,* p. 493-506.

G.S. WATSON et M.R. LEADBETTER (1964 a). — Hazard analysis I. *Biometrika, 51,* p. 175-184.

G.S. WATSON et M.R. LEADBETTER (1964 b). — Hazard analysis II. *Sankhya, 26 A,* p. 110-116.

W. WERTZ (1978). — Statistical density estimation a survey. *Vandenboeck et Ruprecht, Göttingen.*

B.S. YANDELL (1983). — Non-parametric inference for rates with censored survival data. *Ann. Stat. 11, 4,* p. 1119-1135.

Nous remercions l'I.M.S. pour l'autorisation qui nous a été accordée de reproduire les figures illustrant nos exemples.