REVUE DE STATISTIQUE APPLIQUÉE

V. COUALLIER
P. SARDA
P. VIEU

Estimation non paramétrique de discontinuités d'une fonction d'intensité

Revue de statistique appliquée, tome 45, n° 3 (1997), p. 89-106
<http://www.numdam.org/item?id=RSA_1997__45_3_89_0>

Rev. Statistique Appliquée, 1997, **XLV** (3), 89-106

ESTIMATION NON PARAMÉTRIQUE DE DISCONTINUITÉS D'UNE FONCTION D'INTENSITÉ

V. Couallier, P. Sarda, P. Vieu

Laboratoire de Statistique et Probabilités
U.M.R. C55830
Université Paul Sabatier
118, Route de Narbonne 31062 TOULOUSE Cedex

RÉSUMÉ

Nous proposons un estimateur non paramétrique de points de discontinuité et des valeurs du saut en ces points, pour la fonction d'intensité d'un processus ponctuel. Nous obtenons une expression asymptotique de l'erreur quadratique de cet estimateur dans un modèle de Poisson à intensité multiplicative, sous des hypothèses de régularité de la fonction entre deux points de discontinuité. La méthode est illustrée à l'aide de données simulées. Nous proposons enfin une méthode d'estimation de la fonction d'intensité tenant compte des discontinuités. Nous utilisons cette méthode sur un jeu de données météorologiques.

Mots-clés : *Estimation non paramétrique, processus de Poisson, intensité, estimation de discontinuités.*

ABSTRACT

We propose a nonparametric estimator of jump points and corresponding sizes of jump values for the intensity of a point process. We obtain an asymptotic expression for a quadratic error of this estimator in a multiplicative intensity Poisson process under regularity conditions for the intensity between two jump points. The method is illustrated through simulated data. We propose finally a method for estimating the intensity which takes into account jump points. We use this method on meteorological data.

Keywords : *Nonparametric estimation, Poisson process, intensity, estimation of jump points.*

1. Introduction

On considère N points sur un intervalle $[0, T]$. Ces points peuvent être des dates d'occurences de certains événements captés dans le temps comme des émissions radioactives, des arrivées dans une file d'attente, ou encore des localisations d'événements ponctuels. Une manière d'appréhender ces données numériques, ordonnées sur $[0, T]$, est d'admettre qu'elles décrivent une réalisation d'un processus ponctuel sur

R, censuré à gauche en 0, et à droite en T. Pour la compréhension de tels processus, on connaît l'importance de la fonction d'intensité stochastique définie dans le cas général par

$$\lambda(t, \mathcal{H}_t) = \lim_{\delta \to 0} \frac{1}{\delta} P(N(t, t+\delta) > 1 | \mathcal{H}_t), \tag{1}$$

où $N(t, t+\delta)$ est le nombre de points tombant dans l'intervalle $[t, t+\delta]$ et $\mathcal{H}_t$ est l'histoire du processus jusqu'en t.

Afin d'exploiter cette définition, on fait généralement l'hypothèse minimale que le processus n'admet pas d'occurence multiple, ce qui s'écrit

$$P(N(t, t+\delta) > 1) = \circ(\delta).$$

Nous renvoyons à Cox et Isham (1980) pour une description des processus ponctuels.

On s'intéresse à l'estimation de cette intensité et pour cela on se place dans le cadre de l'estimation non paramétrique. En effet, les méthodes développées dans cette discipline ont l'avantage de ne pas supposer l'appartenance de l'objet estimé à une classe de fonctions indexées par un nombre fini de paramètres réels. L'estimateur le plus largement utilisé est l'estimateur à noyau introduit par Ramlau-Hansen (1983) et Diggle (1985). Ce dernier étudie les propriétés asymptotiques de cet estimateur dans le cas d'un processus de Cox. L'analyse de l'erreur quadratique moyenne montre que la qualité de l'estimateur dépend du choix capital du paramètre de lissage, comme c'est le cas dans les autres problèmes d'estimation à noyau. Il s'est développé une importante bibliographie sur la sélection de ce paramètre dans les problèmes voisins d'estimation de densité ou de fonction de régression (voir Marron, 1988, Vieu, 1993, et Broniatowski, 1993 pour des revues bibliographiques) et plus récemment dans le cadre qui nous intéresse. Ainsi, Brooks et Marron (1991) ont obtenu des résultats d'optimalité asymptotique pour le choix de ce paramètre et pour une mesure d'erreur quadratique, dans un modèle poissonien non homogène et moyennant quelques conditions de régularité de la fonction d'intensité à estimer.

Dans la pratique, il peut s'avérer que ces conditions ne soient pas remplies dans certaines régions de l'intervalle. Le travail présenté dans cet article est motivé par l'allègement des hypothèses de régularité de la fonction d'intensité ce qui permet d'envisager d'éventuels points de discontinuité pour cette fonction.

Le problème de l'estimation de discontinuités a été essentiellement envisagé dans le cas de la régression. Wu et Chu (1993a) ont proposé des estimateurs à noyau de points de discontinuités et de la valeur du saut en ces points. Nous nous inspirons ici de leur travail pour estimer les discontinuités éventuelles de la fonction d'intensité.

Dans la section 2, nous explicitons le cadre général d'estimation non paramétrique du processus d'intensité et nous précisons les estimateurs à noyau entrant en jeu. Nous présentons également l'idée générale de la méthode d'estimation d'une discontinuité. Dans la section 3, en supposant que le processus sous-jacent est un processus de Poisson non homogène à intensité multiplicative, nous donnons une écriture asymptotique de l'erreur quadratique ponctuelle de l'estimateur à noyau d'une discontinuité. Ce résultat est obtenu sous des hypothèses de régularité de l'intensité entre deux points

de discontinuité. Dans la section 4, nous mettons en œuvre la méthode proposée à la section 2 à l'aide de données simulées. Nous analysons les résultats obtenus en discutant en particulier les choix du noyau et de la largeur de fenêtre. Nous proposons une méthode alternative d'estimation pour une fonction d'intensité supposée discontinue s'appuyant sur un estimateur à noyau adapté à la correction des effets de bord (Diggle, 1985).

Ce problème de discontinuités dans la structure de processus ponctuels est motivé par l'intérêt pratique que lui porte le Centre de Météorologie Nationale qui cherche à localiser des ruptures dans la structure micrométrique des nuages. Nous conclurons donc par une validation numérique de notre méthode sur les données fournies par cet organisme.

2. Estimation à noyau de l'intensité stochastique

On considère N points X_i sur $[0, T]$ représentant une réalisation d'un processus de fonction d'intensité stochastique λ.

Afin d'estimer λ, Ramlau-Hansen (1983) et Diggle (1985) ont adapté un estimateur non paramétrique de la densité d'une variable aléatoire réelle, à savoir l'estimateur à noyau usuel, en posant

$$\hat{\lambda}(x) = \sum_{i=1}^{N} K_h(X_i - x), \tag{2}$$

avec $K_h(.) = h^{-1}K(./h)$, où le noyau K est une densité de probabilité supposée symétrique, et h est un réel strictement positif appelé paramètre de lissage ou largeur de fenêtre.

Dans le cas d'un processus de comptage, sous l'hypothèse d'une intensité multiplicative (Aalen, 1978) qui suppose la factorisation du processus d'intensité en un processus prévisible et une fonction déterministe, Ramlau-Hansen (1983) obtient la convergence uniforme et la normalité asymptotique d'un estimateur à noyau de cette fonction déterministe sous une hypothèse de continuité.

Dans cette classe de processus, Brooks et Marron (1991) s'intéressent plus particulièrement aux processus de Poisson (voir section 3). Sous l'hypothèse d'existence de deux dérivées de λ, ils obtiennent l'optimalité asymptotique de la fenêtre choisie par validation croisée globale. Moncaup *et al* (1995) comparent ce choix avec celui d'une fenêtre obtenue par validation croisée locale.

Diggle (1985) obtient l'expression de l'erreur quadratique moyenne définie par $EQM_x(h) = \mathbb{E}\left[\hat{\lambda}(x) - \lambda(x)\right]^2$ dans le cas d'un processus de Cox encore appelé processus doublement stochastique (Cox et Isham, 1980). La fonction λ est alors la réalisation d'un processus $\{\Lambda(x), x \in \mathbb{R}\}$ stationnaire et positif. Conditionnellement à la réalisation λ de Λ, le processus ponctuel est un processus de Poisson de fonction d'intensité λ. L'auteur étudie $EQM_x(h)$ dans le cas d'un noyau uniforme et à partir

de l'expression obtenue, propose un choix de la largeur de fenêtre du type «plug-in» qui consiste à prendre $\hat{h}$ minimisant un estimateur $\hat{M}(h)$ de $EQM_x(h)$. Dans le même contexte, Diggle et Marron (1988) poursuivent ce travail en comparant la fenêtre «plug-in» à celle obtenue en densité par validation croisée (Marron, 1988). Ils établissent tout d'abord une équivalence entre les deux approches puis, lorsque la fonction $\delta(|x-y|) = \mathbb{E}(\Lambda(x)\Lambda(y))$ se met sous la forme $\delta(x) = \mu^2\delta_0(x)$ (δ_0 fixée) et $\delta_0(|x|)$ admet une dérivée d'ordre 4 continue en 0, ils obtiennent une expression asymptotique de $EQM_x(h)$ sous la forme d'une somme dont les deux termes principaux dépendent respectivement de la variance et du carré du biais de l'estimateur.

Dans l'ensemble des travaux cités ci-dessus, les auteurs se placent sous des hypothèses de régularité. Entre autres, dans le cadre d'un processus de Poisson, une hypothèse minimale de continuité de la fonction d'intensité est requise et conduit de manière naturelle à considérer des noyaux symétriques, au moins en un point loin du bord de l'intervalle (voir Diggle, 1985, pour le problème des effets de bord).

Pour estimer la fonction d'intensité en un point de discontinuité, il est naturel de considérer des noyaux dissymétriques. En effet, nous privilégions ainsi le comptage des points d'un côté ou d'un autre du point x ce qui nous amène à estimer intuitivement les limites à droite et à gauche de λ en x.

Cette idée nous conduit à considérer deux estimateurs à noyau $\hat{\lambda}_1$ et $\hat{\lambda}_2$ définis par

$$\hat{\lambda}_1(x) = \sum_{i=1}^{N} K_{1,h}(X_i - x),$$

et

$$\hat{\lambda}_2(x) = \sum_{i=1}^{N} K_{2,h}(X_i - x),$$

où K_1 et K_2 sont deux noyaux dissymétriques à support $[-1, w]$ et $[-w, 1]$ respectivement, et w un réel dans $[0, 1[$. On définit également

$$\hat{S}(x) = \frac{\hat{\lambda}_2(x) - \hat{\lambda}_1(x)}{k_w}, \tag{3}$$

où k_w est une constante dépendant des noyaux K_1 et K_2 (voir ci-dessous).

Intuitivement, la valeur absolue de $\hat{S}$ est plus grande aux points de discontinuité de λ alors qu'en un point de continuité, $\hat{\lambda}_1$ est proche de $\hat{\lambda}_2$. La pondération k_w est introduite afin d'obtenir un estimateur du saut asymptotiquement non biaisé dans le cadre d'un processus de Poisson (voir Lemme 2).

Finalement, en supposant que la fonction λ admet un point de discontinuité θ, un estimateur de ce point est fourni par

$$\hat{\theta} = \underset{x \in]0,T[}{\arg\max} |\hat{S}(x)|. \tag{4}$$

Notons qu'un estimateur similaire a été introduit dans le cadre de la régression par Wu et Chu (1993a) qui en étudient les propriétés asymptotiques.

3. Estimation de discontinuités de la fonction d'intensité d'un processus de Poisson

3.1. Le modèle

Explicitons le processus d'intensité stochastique dans le cas d'un modèle poissonien non homogène. Nous supposons qu'il existe une mesure μ sur R telle que

- $\mathcal{N}(t) = N(0,t)$ est un processus de dénombrement à accroissements indépendants;
- $N(s,t)$ suit une loi de Poisson d'espérance $\mu([s,t[)$.

Si, en outre, il existe une fonction λ telle que $\mu([s,t[) = \int_s^t \lambda(x)dx$ alors on sait (Cox et Isham, 1980) que cette fonction λ est le processus d'intensité défini dans le cas général par (1) qui est donc dans ce cas une fonction déterministe.

On cherche donc à estimer les discontinuités d'une fonction λ déterministe sachant que

$$N(s,t) \sim \mathcal{P}(\int_s^t \lambda(u)du) \tag{5}$$

où $\mathcal{P}(m)$ désigne la loi de Poisson de paramètre m.

Nous utilisons en outre une forme simplifiée du modèle à intensité multiplicative introduit par Aalen (1978) et fréquemment utilisée pour modéliser des processus de dénombrement. Nous supposons que la fonction d'intensité s'écrit

$$\lambda_c(x) = c.\alpha(x), x \in [0,T], \tag{6}$$

où c est une constante positive et α est une fonction déterministe inconnue telle que $\int_0^T \alpha = 1$. D'après (5) et (6) c est l'espérance du nombre d'observations N.

Cette décomposition est évidemment possible pour tout processus de Poisson. Dans les expressions asymptotiques, nous ferons tendre c vers l'infini, ce qui revient à augmenter le nombre d'observations sur l'intervalle $[0,T]$ sans changer la forme générale de la fonction d'intensité λ_c (voir à ce sujet la discussion dans Diggle et Marron, 1988, à propos de la nature des problèmes asymptotiques pour l'estimation d'une fonction d'intensité). Nous considérons pour cela une suite $\{c_s\}_{s=1}^{+\infty}$ de nombres réels pour laquelle est définie la suite de fonctions d'intensité $\lambda_{c_s}(x) = c_s\alpha(x)$ indexées par s. L'estimateur $\hat{S}$ est alors défini pour chaque fonction λ_{c_s} à partir de la largeur de fenêtre h_s.

3.2. Développement asymptotique de l'erreur quadratique moyenne

Nous nous intéressons à l'erreur quadratique moyenne $EQM_x(h)$ définie par

$$EQM_x(h) = \mathbb{E}\Big[\hat{S}(x) - \Big(\lambda(x^+) - \lambda(x^-)\Big)\Big]^2,$$

où $\lambda(x^+)$ et $\lambda(x^-)$ (respectivement $\alpha(x^+)$ et $\alpha(x^-)$) sont les limites à droite et à gauche de la fonction λ (respectivement α) au point x. Afin d'obtenir le développement asymptotique de $EQM_x(h)$, nous ferons les hypothèses suivantes :

H.1 – les noyaux K_1 et K_2 sont des fonctions de densité de probabilité à supports respectifs $[-1, w]$ et $[-w, 1]$ où $w \in [0, 1[$ et vérifient

$$k_w = \int_0^1 \Big(K_2(u) - K_1(u)\Big)du \neq 0,$$

$$K_1(-z) = K_2(z), \quad z \in I\!R;$$

H.2 – la fonction d'intensité λ admet un nombre fini de points de discontinuité et entre ces points, elle admet une dérivée lipschitzienne;

H.3 – les suites c_s et h_s vérifient

$$\lim_{s\to+\infty} c_s = +\infty, \quad \lim_{s\to+\infty} h_s = 0, \quad \lim_{s\to+\infty} c_s h_s = +\infty.$$

Nous aurons également besoin du résultat suivant :

Lemme 1 : Conditionnellement à N, les X_i sont distribuées comme les statistiques d'ordre de N variables aléatoires réelles Y_i, $(i = 1, .., N)$ indépendantes et identiquement distribuées de densité α.

Pour la démonstration de ce lemme, nous renvoyons à Kingman (1993) (p. 22).

Lemme 2 : expression et développement asymptotique du biais

Sous les hypothèses $H.1$, $H.2$ et $H.3$, en notant $K_1^2 = K_2 - K_1$, $\alpha'(x^-)$ et $\alpha'(x^+)$ les limites respectives à gauche et à droite en x de α' la dérivée de α, on a, pour $x \in]0, T[$ et pour s assez grand

$$\mathbb{E}\hat{S}(x) = (\lambda(x^+) - \lambda(x^-)) + c_s h_s A_x + c_s h_s O(h_s), \tag{7}$$

avec $A_x = (\alpha'(x^+) + \alpha'(x^-))\dfrac{\displaystyle\int_0^1 K_1^2(u)udu}{k_w}$.

Preuve : en utilisant la définition de K_1^2 on a $\forall x \in]0,T[, \forall s > 0$

$$\begin{aligned}\mathbb{E}\hat{S}(x) &= \mathbb{E}\left[\frac{1}{k_w}\sum_{i=1}^{N} K_{1,h_s}^2(X_i - x)\right] \\ &= \frac{1}{k_w}\mathbb{E}\left[\mathbb{E}(\sum_{i=1}^{N} K_{1,h_s}^2(X_i - x)|N)\right].\end{aligned}$$

D'après le lemme 1, on a

$$\mathbb{E}\hat{S}(x) = \frac{1}{k_w}\mathbb{E}(N)\int_0^T K_{1,h_s}^2(y-x)\alpha(y)dy.$$

En effectuant le changement $z = (y-x)/h_s$ pour $x \in]0,T[$ et pour s à partir d'un certain rang, on obtient, puisque $\mathbb{E}(N) = c_s$

$$\mathbb{E}\hat{S}(x) = \frac{c_s}{k_w}\int_{-1}^{1} K_1^2(z)\alpha(x+h_s z)dz.$$

A partir de cette écriture intégrale, nous obtenons un développement asymptotique du biais. Les hypothèses $H.2$ et $H.3$ nous donnent, pour s assez grand, la continuité de λ et de sa dérivée sur les intervalles $I_1 =]x-h_s, x[$ et $I_2 =]x, x+h_s[$.

Deux développements de Taylor sur I_1 et I_2 définissent ξ_z^1 et ξ_z^2 tels que

$$\begin{aligned}-1 < z < 0 : \alpha(x+h_s z) &= \alpha(x^-) + h_s z\alpha'(\xi_z^1) \quad , x - h_s < x + h_s z < \xi_z^1 < x, \\ 1 > z > 0 : \alpha(x+h_s z) &= \alpha(x^+) + h_s z\alpha'(\xi_z^2) \quad , x < \xi_z^2 < x + h_s z < x + h_s.\end{aligned}$$

On a donc, en tenant compte de l'écriture de k_w

$$\mathbb{E}\hat{S}(x) = c_s(\alpha(x^+) - \alpha(x^-)) + \frac{c_s h_s}{k_w}\left[\int_{-1}^{0} K_1^2(z)z\alpha'(\xi_z^1)dz + \int_0^1 K_1^2(z)z\alpha'(\xi_z^2)dz\right].$$

Or, par $H.2$

$$\alpha'(\xi_z^1) = \alpha'(x^-) + O(h_s), \quad \alpha'(\xi_z^2) = \alpha'(x^+) + O(h_s).$$

Les $O(h_s)$ sont uniformes en z et en intégrant sur z nous obtenons

$$\begin{aligned}\mathbb{E}\hat{S}(x) &= c_s(\alpha(x^+) - \alpha(x^-)) + c_s h_s(\alpha'(x^+) \\ &\qquad + \alpha'(x^-))k_w^{-1}\int_0^1 K_1^2(u)udu + c_s h_s O(h_s) \\ &= (\lambda(x^+) - \lambda(x^-)) + c_s h_s A_x + c_s h_s O(h_s)\end{aligned}$$

Le théorème suivant nous fournit l'expression de $EQM_x(h)$.

Théorème : Développement asymptotique de EQM_x (h)

Sous les hypothèses $H.1$, $H.2$ et $H.3$, nous avons

$$\mathbb{E}\left[\hat{S}(x)-(\lambda(x^+)-\lambda(x^-))\right]^2 = B_x\frac{c_s}{h_s}+A_x^2h_s^2c_s^2+(\frac{c_s}{h_s}+h_s^2c_s^2)O(\sqrt{h_s})\,,\ x\in]0,T[, \tag{8}$$

où

$$B_x = \frac{1}{k_w^2}\left[\alpha(x^+)\int_0^1\Big(K_1^2(u)-(\alpha(x^+)-\alpha(x^-))k_w\Big)^2du \right.$$
$$\left. +\alpha(x^-)\int_{-1}^0\Big(K_1^2(u)-(\alpha(x^+)-\alpha(x^-))k_w\Big)^2du\right],$$

et

$$A_x = (\alpha'(x^+)+\alpha'(x^-))\frac{\displaystyle\int_0^1 K_1^2(u)udu}{k_w}.$$

Preuve du théorème : En décomposant de manière habituelle l'erreur quadratique en deux termes de biais au carré et variance on obtient :

$$\mathbb{E}\left[\hat{S}(x)-(\lambda(x^+)-\lambda(x^-))\right]^2 = Var(\hat{S}(x))+\left[E\hat{S}(x)-(\lambda(x^+)-\lambda(x^-))\right]^2.$$

On note

$$\mathbb{E}\hat{S}(x) = \frac{c_s}{k_w}I_s(x),$$

avec

$$I_s(x) = \int_{-1}^1 K_1^2(z)\alpha(x+h_sz)dz.$$

On a alors

$$\begin{aligned}\mathrm{Var}\,(\hat{S}(x)) &= \left[\frac{1}{k_w}\sum_{i=1}^N K_{1,h_s}^2(X_i-x)-\mathbb{E}\hat{S}(x)\right]^2\\ &= \mathbb{E}\left[\frac{1}{k_w}\sum_{i=1}^N\Big[K_{1,h_s}^2(X_i-x)-I_s(x)\Big]+\frac{N-c_s}{k_w}I_s(x)\right]^2\\ &= \mathbb{E}\left[\frac{1}{k_w}\sum_{i=1}^N\Big[K_{1,h_s}^2(X_i-x)-I_s(x)\Big]\right]^2+\frac{I_s^2(x)}{k_w^2}\mathbb{E}(N-c_s)^2\\ &\quad+\frac{2}{k_w^2}\mathbb{E}\left[(N-c_s)I_s(x)\sum_{i=1}^N\Big[K_{1,h_s}^2(X_i-x)-I_s(x)\Big]\right]. \qquad (9)\end{aligned}$$

En appliquant le changement de variable $z = (y - x)/h_s$ pour $x \in]0, T[$ et pour s assez grand, on obtient, en utilisant le Lemme 1

$$\mathbb{E}\left[\frac{1}{k_w}\sum_{i=1}^{N}\Big[K_{1,h_s}^2(X_i - x) - I_s(x)\Big]\right]^2 = \mathbb{E}\left[\mathrm{Var}\left(\Big[\frac{1}{k_w}\sum_{i=1}^{N} K_{1,h_s}^2(Y_i - x)\Big]|N\right)\right]$$
$$= \frac{1}{k_w^2}\frac{\mathbb{E}(N)}{h_s}\int_{-1}^{1}\Big[K_1^2(u) - I_s(x)\Big]^2\alpha(x + h_s u)du.$$

Le Lemme 2 nous donne une expression asymptotique de $I_s(x)$. En utilisant la même technique de développement de $\alpha(x + h_s u)$, on obtient

$$\mathbb{E}\left[\frac{1}{k_w}\sum_{i=1}^{N}\Big[K_{1,h_s}^2(X_i - x) - I_s(x)\Big]\right]^2$$
$$= \frac{1}{k_w^2}\frac{c_s}{h_s}\int_{-1}^{1}\Big[K_1^2(u) - (\alpha(x^+) - \alpha(x^-))k_w + O(h_s)\Big]^2\alpha(x + h_s u)du$$
$$= \frac{1}{k_w^2}\frac{c_s}{h_s}\left[\alpha(x^-)\int_{-1}^{0}\Big[K_1^2(u) - (\alpha(x^+) - \alpha(x^-))k_w\Big]^2 du\right.$$
$$\left. + \alpha(x^+)\int_{0}^{1}\Big[K_1^2(u) - (\alpha(x^+) - \alpha(x^-))k_w\Big]^2 du + O(h_s)\right]$$
$$= \frac{c_s}{h_s}B_x + \frac{c_s}{h_s}O(h_s). \tag{10}$$

Les deux autres termes du développement de la variance de $\hat{S}(x)$ s'expriment de la même façon. En effet, $I_s^2(x)$ est majoré par une constante indépendante de s et

$$E(N - c_s)^2 = \mathrm{Var}\,(\mathcal{P}(c_s)) = c_s = \frac{c_s}{h_s}O(h_s). \tag{11}$$

Enfin, à l'aide de l'inégalité de Cauchy Schwartz et en utilisant (10) et (11) on obtient

$$\left|\frac{1}{k_w^2}\mathbb{E}\left[(N - c_s)I_s(x)\sum_{i=1}^{N}\Big[K_{1,h_s}^2(X_i - x) - I_s(x)\Big]\right]\right|$$
$$\leqslant \left|\frac{I_s(x)}{{k_w}^2}\right|\left(\mathbb{E}(N - c_s)^2\mathbb{E}\Big(\sum_{i=1}^{N} K_{1,h_s}^2(X_i - x) - I_s(x)\Big)^2\right)^{\frac{1}{2}}$$
$$\leqslant \left|\frac{I_s(x)}{{k_w}^2}\right| c_s^{\frac{1}{2}}\left[\frac{c_s}{h_s}(B_x + O(h_s))\right]^{\frac{1}{2}}$$
$$\leqslant \frac{c_s}{h_s}O(\sqrt{h_s}). \tag{12}$$

On a donc par (9), (10), (11) et (12) l'expression asymptotique de la variance, sachant que dans (11), $O(h_s)$ est un $O(\sqrt{h_s})$:

$$\text{Var}\,(\hat{S}(x)) = B_x \frac{c_s}{h_s} + \frac{c_s}{h_s} O(\sqrt{h_s}),$$

qui, combinée avec l'expression (7) du biais, amène directement le résultat (8) et achève ainsi la preuve du théorème. □

4. Études de simulations

Pour modéliser un processus de Poisson d'intensité connue, nous avons utilisé le Lemme 1 et généré N points Y_i de densité α connue. Cette fonction est construite par morceaux sur l'intervalle $[0, 20]$ à partir des densités de probabilités de variables aléatoires réelles gaussiennes et uniformes :

$$\alpha = a.f_{N(3,4)}.I_{[0,5]} + b.f_{\mathrm{U}_{[5,10]}}.I_{[5,10]} + c.f_{\mathrm{U}_{[10,13]}}.I_{[10,13]} + d.f_{N(17,5)}.I_{[13,20]},$$

où

- f_X désigne la densité de probabilité de la variable aléatoire réelle X ;
- $N_{(m,\sigma)}$ est la variable aléatoire normale de moyenne m et d'écart type σ ;
- $\mathrm{U}_{[\alpha,\beta]}$ est la loi uniforme sur l'intervalle $[\alpha, \beta]$;
- $I_{[a,b]}$ est la fonction indicatrice du segment $[a, b]$;
- a, b, c, d sont des constantes réelles positives qui font de α une densité de probabilité sur $[0, 20]$.

Les N points ordonnés constituent donc une réalisation d'un processus poissonien d'intensité $\lambda = N\alpha$ et la fonction d'intensité λ admet trois discontinuités en $x = 5$, $x = 10$ et $x = 13$. Dans cette simulation, $N = 2425$.

Afin d'estimer λ, nous avons utilisé l'estimateur $\hat{\lambda}$ défini en (2) sauf aux extrémités de l'intervalle où une correction des effets de bord est appliquée. Cet estimateur corrigé, introduit par Diggle (1985) s'écrit

$$\hat{\lambda}(x) = \frac{1}{p_h(x)} \sum_{i=1}^{N} K_h(X_i - x), \tag{13}$$

où $p_h(x) = \displaystyle\int_0^T K_h(x - u)du.$

Le problème du choix des noyaux est de moindre importance par rapport à celui du choix des fenêtres. Nous avons utilisé dans tous nos calculs des noyaux uniformes faciles à mettre en œuvre. Nous avons donc

$$K(x) = \frac{1}{2}\mathbb{I}_{[-1,1]}(x).$$

Les figures 1 à 3 sont les tracés de la fonction λ et de l'estimateur $\hat{\lambda}$ pour trois valeurs de la largeur de fenêtre. L'importance du choix du paramètre de lissage, mise en évidence dans les expressions d'erreurs quadratiques asymptotiques obtenues par Brooks et Marron (1991), peut être directement observée sur ces figures. Des valeurs de h trop petites font apparaître un sous-lissage ($\hat{\lambda}$ oscille beaucoup trop) caractérisant une variance importante de l'estimateur, tandis que des valeurs de h trop grandes gomment les variations brusques de $\hat{\lambda}$; il s'ensuit un surlissage de la courbe et c'est la contribution du biais à l'erreur quadratique qui devient trop importante (voir à ce sujet Moncaup *et al*, 1995). Les tracés de $\hat{\lambda}$ montrent que le choix de fenêtre mettant le mieux en évidence les discontinuités est $h = 0.4$, choix qui par ailleurs sous-lisse la fonction sur les sous-intervalles de continuité.

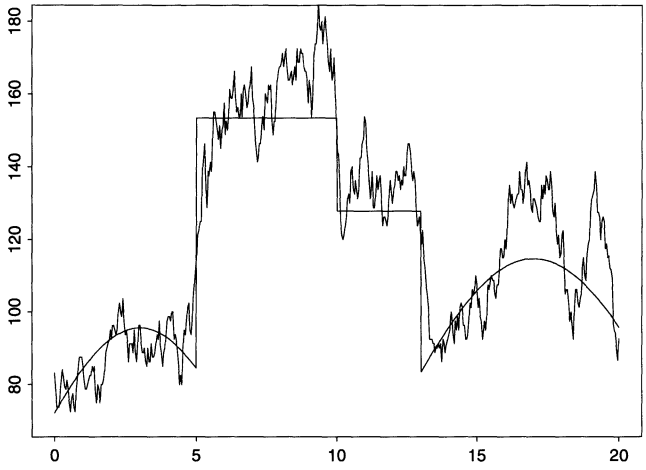

FIGURE 1
Estimateur à noyau $\hat{\lambda}$ pour $h = 0,4$

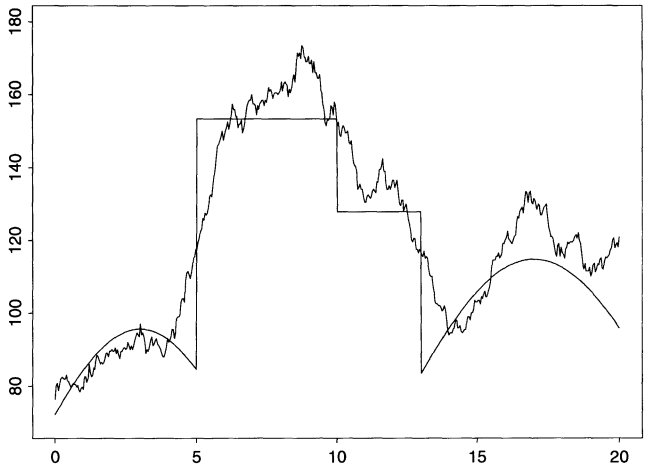

FIGURE 2
Estimateur à noyau $\hat{\lambda}$ pour $h = 1$

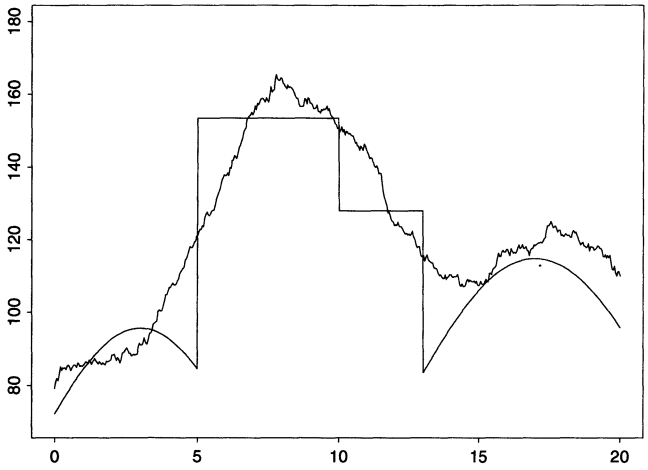

FIGURE 3
Estimateur à noyau $\hat{\lambda}$ pour $h = 2$

Nous avons tracé la fonction $\hat{S}$ définie en (3) en prenant $w = 0$. On obtient alors en utilisant des noyaux uniformes

$$k_w = 1, \quad K_1(x) = \mathbb{I}_{[-1,0]}(x), \quad K_2(x) = \mathbb{I}_{[0,1]}(x).$$

Les figures 4 à 6 montrent l'estimateur $\hat{S}$ pour trois largeurs de fenêtre. Les courbes de $\hat{S}$ mettent bien en évidence les effets de bord exposés plus haut et nous ne pouvons analyser ces courbes que sur l'intervalle $[0, T]$ diminué d'une bande de largeur h aux deux extrémités.

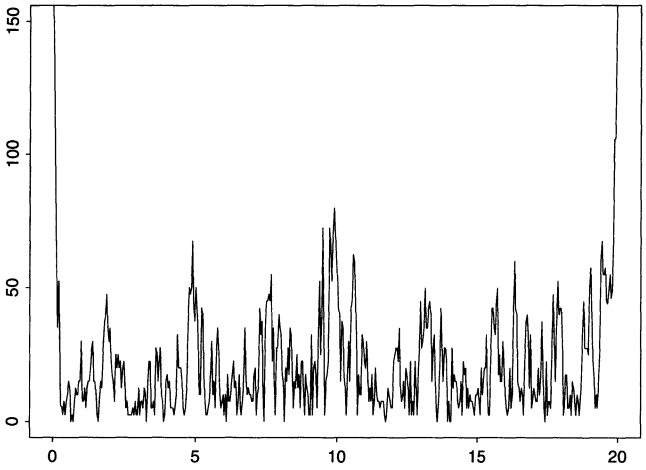

FIGURE 4
$\hat{S}$ pour $h = 0,4$

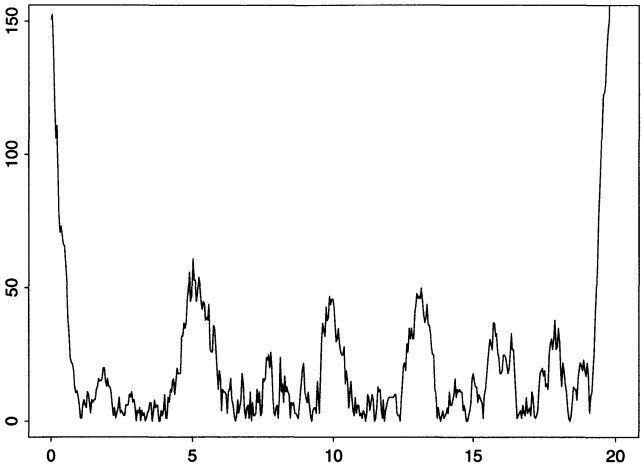

FIGURE 5
$\hat{S}$ *pour* $h = 1$

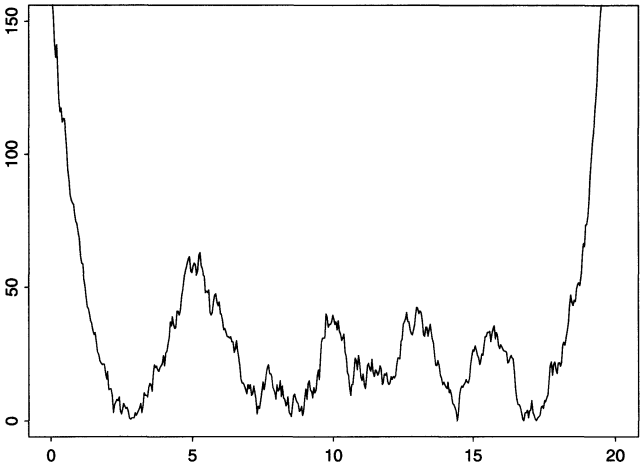

FIGURE 6
$\hat{S}$ *pour* $h = 2$

L'importance du choix de la fenêtre dans la localisation des points de discontinuité, qui apparaît sur ces figures a été mise en évidence dans l'expression asymptotique (8). Nous remarquons toutefois que dans la pratique, la dégradation de l'estimateur $\hat{\theta}$ défini en (4) est moins importante que celle observée en général sur l'estimation de la courbe λ elle-même. Sur toutes les simulations faites, nous avons observé un bon comportement de la localisation des discontinuités. Cependant, la qualité de l'estimation de la hauteur du saut, donnée ici par la valeur de $\hat{S}$ aux points de discontinuité (estimés par $\hat{\theta}$), est plus sensible au choix du paramètre de lissage.

Nous présentons ci-dessous une seconde méthode d'estimation des sauts en les discontinuités estimées qui s'appuie sur une estimation de la fonction d'intensité en la supposant discontinue.

Dans notre exemple, pour $h = 1$, les trois premiers maximums locaux et les estimations des sauts correspondants donnent :

$$\begin{array}{llll} \hat{\theta}_1 = 5.09 & (\theta_1 = 05) & \hat{S}(\hat{\theta}_1) = 61 & (S(\theta_1) = 71), \\ \hat{\theta}_2 = 13.22 & (\theta_2 = 13) & \hat{S}(\hat{\theta}_2) = 52 & (S(\theta_2) = 27), \\ \hat{\theta}_3 = 9.93 & (\theta_3 = 10) & \hat{S}(\hat{\theta}_3) = 47 & (S(\theta_3) = 45). \end{array}$$

En supposant le nombre p de discontinuités connu et après estimations à l'aide de $\hat{S}$ des p discontinuités, nous pouvons appliquer sur les $p+1$ intervalles ainsi formés une estimation de λ qui tient compte des effets de bord : ainsi l'estimateur défini en (13) est calculé sur chaque sous-intervalle. La fonction étant continue sur chacun des sous-intervalles, nous utilisons pour chaque estimateur $\hat{\lambda}_i$, $i = 1, ..., (p+1)$ la largeur de fenêtre obtenue par validation croisée exposée dans Brooks et Marron (1991). La figure 7 est le tracé des quatres estimations de λ sur les intervalles estimés $[0, 5.09]$, $[5.09, 9.93]$, $[9.93, 13.22]$ et $[13.22, 20]$.

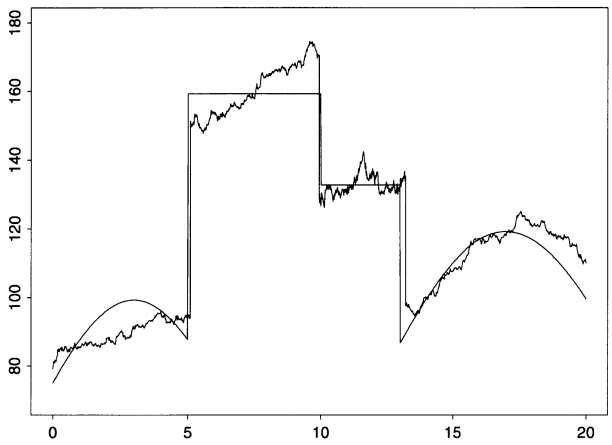

FIGURE 7
Estimation de λ par intervalle de continuité

Cette méthode nous permet de définir un nouvel estimateur de la hauteur du saut au point de discontinuité $\hat{\theta}_i$ estimé en posant

$$\hat{s}(\hat{\theta}_i) = \hat{\lambda}_{i+1}(\hat{\theta}_i) - \hat{\lambda}_i(\hat{\theta}_i).$$

5. Application aux données météorologiques

On dispose d'une série $X_i, i = 1, ..., N$, $N = 10229$, de gouttes d'eau le long de la trajectoire d'un avion qui traverse un nuage. Nous renvoyons à Brenguier (1993) et Moncaup *et al* (1995) pour une description détaillée du jeu de données.

On suppose que ces données sont une réalisation censurée d'un processus de Poisson non homogène décrit dans la section 3. L'estimateur de la fonction d'intensité λ est l'estimateur corrigé (13) avec $h = 1.335$, valeur qui minimise le critère de validation croisée (Moncaup *et al*, 1995). Les fortes pentes décelées en 47000 et 47015 (*cf.* fig.8) par cette méthode d'estimation, qualifiée de directe, laissent penser que la fonction d'intensité est discontinue.

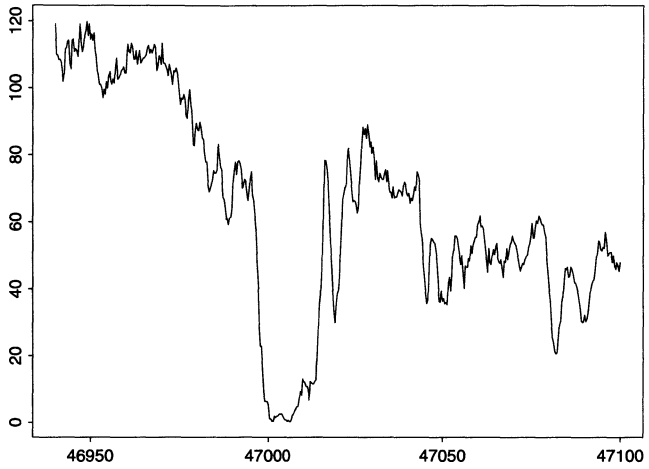

FIGURE 8
Estimation «directe» de λ pour $h = 1.335$

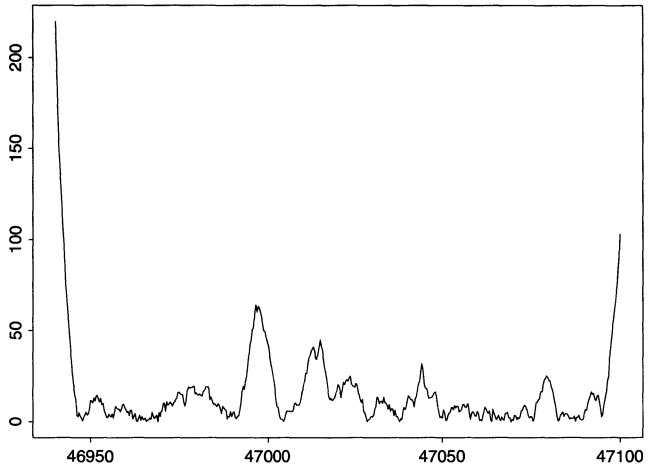

FIGURE 9
Tracé de $\hat{S}$ pour $h = 6$

Nous avons calculé $\hat{S}$ avec la valeur arbitraire $h = 6$. Sur la figure 9, nous voyons deux pics bien définis et deux autres plus discrets. Nous avons donc fait l'hypothèse de

l'existence de quatre discontinuités estimées par $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ et $\hat{\theta}_4$. Sous cette hypothèse, en utilisant la méthode précédente, nous obtenons une nouvelle estimation de λ obtenue par morceaux sur les cinq intervalles définis par les $\hat{\theta}_i$, présentée à la figure 10. Par comparaison avec la méthode d'estimation directe (figure 8) on obtient une courbe plus lisse entre les points de discontinuités. Cela tient essentiellement au fait que les fenêtres obtenues par validation croisée sur chaque intervalle sont plus larges que la fenêtre globale $h = 1.335$.

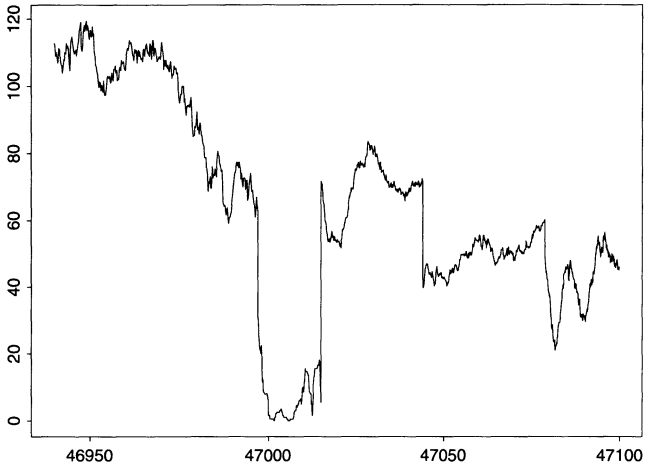

FIGURE 10
Estimation de λ par intervalle de continuité (quatre discontinuités)

6. Conclusion

Notre étude nous a permis de mettre en évidence la validité de la méthode d'estimation de discontinuités proposée à la fois pour l'estimation des cassures elles-mêmes mais aussi pour l'amélioration apportée à l'estimation d'une fonction d'intensité supposée discontinue. Le bon comportement sur les simulations de l'estimateur d'une intensité discontinue ne peut que renforcer la nécessité d'apporter des réponses théoriques à certains problèmes non encore résolus. En particulier, il serait intéressant d'établir des résultats de convergence de nos estimateurs $\hat{S}(x)$ et $\hat{\theta}$ du type de ceux obtenus par Wu et Chu (1993a, 1993b) en régression.

Le problème du choix du paramètre de lissage est également capital, tant pour la première méthode d'estimation de discontinuités (section 3) que pour la seconde méthode d'estimation du saut s'appuyant sur l'hypothèse que λ admet p discontinuités. Nous avons utilisé un choix automatique des largeurs de fenêtre sur chaque sous-intervalle qui s'inspire des travaux de Brooks et Marron (1991). Cependant, les bornes des intervalles étant aléatoires, nous ne pouvons pas déduire directement l'optimalité asymptotique de ces paramètres de lissage à partir des résultats de Brooks et Marron. Pourtant, tout porte à croire qu'en prouvant la

convergence presque sûre des estimateurs $\hat{\theta}$ sous l'hypothèse que p est bien le nombre de points de discontinuités, un résultat similaire d'optimalité asymptotique pourrait être obtenu. Dans notre simulation, p est évidemment connu. Dans les applications, ce nombre est en général inconnu, et la méthode proposée à la section 2 n'en fournit pas un estimateur. Il semble cependant que le choix de la valeur de p est crucial (voir à ce sujet les résultats de Wu et Chu, 1993a et 1993b qui obtiennent en régression des convergences différentes selon que le nombre de discontinuités est sous-évalué ou sur-évalué). Enfin il serait intéressant d'élargir notre modèle aux processus de Cox car souvent les applications pratiques nécessitent un caractère aléatoire de la fonction d'intensité et le modèle de Cox à intensité multiplicative fournit toute la structure mathématique désirée.

Remerciements

Nous tenons à remercier Jean-Louis Brenguier et Anna Pawlovska du Centre National de la Recherche Météorologique de Toulouse. Les nombreuses discussions que nous avons eues avec eux sont une des origines de ce travail. Nous tenons aussi à remercier le rapporteur dont les commentaires pertinents ont permis d'améliorer la présentation de ce travail.

Bibliographie

AALEN, O. (1978), Nonparametric inference for a family of counting processes, *Ann. Statis.*, **6**, 701-726.

BRENGUIER, J.L. (1993), Observations of cloud microstructure at the centimeter scale, *J. of Applied Meteorology* , **32**, 783-793.

BRONIATOWSKI, M. (1993), Cross validation in kernel nonparametric density estimation : a survey, *Publi. de l'I.S.U.P.*, **vol XXXVII**, 3-4, 3-28.

BROOKS, M.M. and MARRON, J.S. (1991), Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions, *Stochastic Processes and their Applications*, **38**, 157-165, North-Holland.

COX, D.R. and ISHAM, V. (1980), *Point processes*, Chapmann and Hall, London.

DIGGLE, P. (1985), A Kernel Method for Smoothing Point Process Data, *Appl. Statist.*, **34**, 138-147.

DIGGLE, P. and MARRON, J.S. (1988), Equivalence of Smoothing Parameters Selectors in Density and Intensity Estimation, *J. of the American Statistical Association*, **83**, 793-800.

KINGMAN, J.F.C. (1993), *Poisson Processes*, Oxford University Press Inc., New-York.

MARRON, J.S. (1988), Automatic smoothing parameter selection : a survey, *Empirical Economics*, **13**, 187-208.

MONCAUP, S., SARDA, P. et VIEU P. (1995), Une mise en œuvre d'estimateurs non paramétriques sur des données météorologiques, *Revue de Statistique Appliquée*, **Vol XLIII(4)**, 77-88.

RAMLAU-HANSEN, H. (1983), Smoothing counting process intensities by mean of kernel functions, *Ann. Statist.*, **11**, 453-466.

VIEU, P. (1993), Bandwidth selection for kernel regression : a survey. In "*Computer Intensitive Methods in Statistics*". Ed. W. HÄRDLE and L. SIMAR, Statistics and Computing, Physica Verlag, **1**, 134-149.

WU, J.S. and CHU, C.K. (1993a), Kernel-Type Estimators of Jump Points and Values of a Regression Function, *Ann. Statist.*, **21**, 1545-1566.

WU, J.S. and CHU, C.K. (1993b), Nonparametric Function Estimation and Bandwidth Selection for Discontinuous Regression Functions, *Statistica Sinica*, **3**, 557-576.