République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A/MIRA de Béjaia
Faculté des Siences Exactes
Département des Mathématiques/MI

Mémoire de fin de Cycle

En vue de l'obtention d'un Master en Mathématiques

Option

Statistique et Analyse Décisionnelle
SAD

Thème

Equivalence du choix de la fenêtre pour l'estimation des fonctions de densité et d'intensité par la méthode du noyau.

Présenté par :
M^{elle} **DERRASSE CHOUK Besma**

Devant le jury composé de :

Présidente	M^{me} **K. TIMRIDJINE**	**M.C.B**	**Université A/Mira, Béjaïa**
Rapporteur	M^{me} **A. BARECHE**	**M.C.A**	**Université A/Mira, Béjaïa**
Examinatrice	M^{me} **H. TABTI**	**M.A.A**	**Université A/Mira, Béjaïa**

Université de Béjaïa, Juin 2015.

Remerciements

Tout d'abord je tiens à remercier Dieu de m'avoir donné le courage, la morale et la santé pour mener à bien ce travail.

Je remercie particulièrement ma promotrice M^me A. BARECHE pour sa disponibilité, son soutien et ses remarques précieuses qui m'ont aidé à bien présenter ce travail.

Mes remerciements s'adressent aussi à M^me K. TIMRIDJINE d'avoir accepté de présider mon jury de soutenance.

Je suis également très reconnaissante à M^me H. TABTI pour son soutien et sa gentillesse et aussi d'avoir accepté d'examiner ce travail.

Je tiens à remercier toute ma famille, mes amies et mes condisciples de la promotion SAD 2015.

Enfin, je remercie chaleureusement toutes les personnes qui m'ont aidé, et qui ont contribué de proche ou de loin à la réalisation de ce travail.

Dédicaces

A mes très chers parents qui m'ont toujours encouragé

Que Dieu les protège.

A mes frères : Makhlouf, M^{ed} Teyeb et Hamid.

A mes sœurs : Hassiba et Nabila.

A mes grands parents, mes oncles et mes tantes.

A tous mes cousins et à toutes mes cousines et à toute ma famille.

A tous ceux qui me sont chers.

Table des matières

Table des figures

Liste des tableaux

Résumé

Dans ce mémoire, nous nous intéressons à l'estimation non paramétrique par la méthode du noyau. Nous adoptons cette méthode pour l'estimation de la fonction de densité de probabilité et celle de l'intensité d'un processus de Poisson non homogène.

L'estimation de la fonction de densité, à partir d'un échantillon $X_1, X_2, ..., X_n$ issu d'une variable aléatoire X, nécessite le choix du noyau K et du paramètre de lissage h. Le choix de ces deux paramètres est crucial pour la qualité de l'estimation. Nous donnons les différents noyaux utilisés dans la littérature, à savoir les noyaux symétriques pour les densités définies sur $\mathbb{R}$ et les noyaux asymétriques pour les densités définies sur $[0, +\infty[$ (problème des effets de bord). Nous présentons ensuite une brève synthèse sur les méthodes du choix du paramètre de lissage. Pour l'estimation de l'intensité d'un processus de Poisson non homogène, nous présentons deux modèles mathématiques, le modèle multiplicatif simple d'Aalen (1978) et le modèle de Cox (1980). Une méthode de sélection du paramètre de lissage est donnée par la démarche suivie par Diggle (1985) en se plaçant dans le cas du modèle de Cox.

Malgré que les deux méthodes de sélection de h dans l'estimation de la densité et de l'intensité sont motivées de manières différentes, il y a une équivalence du choix de ce paramètre dans les deux méthodes. Nous présentons cette équivalence et les profits q'elle permet d'avoir pour chaque méthode dans le contexte de l'autre. Une étude de simulation est donnée à la fin de ce mémoire pour illustrer les différents aspects théoriques présentés.
Mots clés : Estimation non paramétrique, méthode du noyau, paramètre de lissage, densité, intensité, processus de Poisson non homogène.

Abstract

In this work, we are interested in nonparametric estimation with the kernel method. We use this method for the estimation of the probability density function and the intensity function of a nonhomogeneous Poisson process. The estimation of the density function from a sample $X_1, X_2, ..., X_n$ from a random variable X, requires the choice of the kernel K and the smoothing parameter h. The choice of these two parameters is crucial for the quality of the estimation. We give the different kernels used in the literature which are the symmetric kernels for the densities defined on $\mathbb{R}$ and the asymmetric kernels for densities defined on $[0, +\infty[$ (problem of boundary effects).
For the estimation of the intensity function, we present two mathematical models, the Aalen's simple multiplicative model (1978) and the Cox model (1980). A method for the selection of the smoothing parameter is given by the technique introduced by Diggle (1985) using the Cox model.

Although the selection methods of the smoothing parameter in the estimation of the density and the intensity are motivated in different ways, there is an equivalence in the selection of this parameter. We describe this equivalence and its benefits for the estimation of both density and intensity. A simulation study is given at the end of this work to illustrate the different theoretic aspects presented.
Key words : Nonparametric estimation, kernel method, smoothing parameter, density, intensity, nonhomogeneous Poisson process.

Introduction générale

La théorie de l'estimation est l'une des branches les plus basiques de la statistique mathématique. Cette théorie est habituellement divisée en deux composantes principales, à savoir, l'estimation paramétrique et l'estimation non paramétrique. Depuis l'apparition de cette théorie, l'estimation de la densité de probabilité est l'un de ses principaux problèmes. L'estimation de la densité f par l'approche paramétrique a comme inconvénient principal de nécessiter une connaissance préalable sur la loi de probabilité du phénomène aléatoire que l'on étudie. Cependant, l'approche non-paramétrique estime la densité de probabilité directement à partir de l'information disponible sur l'ensemble des observations. On dit souvent que dans cette approche les données parlent d'elles mêmes. On s'intéresse dans le cadre de ce travail à l'approche non paramétrique.

Il existe plusieurs méthodes non paramétriques pour l'estimation de la densité de probabilité, on peut citer la méthode de l'histogramme, la méthode des series orthogonales, la méthode splines et la méthode du noyau. Cette dernière est la plus utilisée vu sa simplicité et la qualité de l'estimation qu'elle assure ; tout au long de ce travail, nous nous intéressons à l'estimation par la méthode du noyau.

Rosenblatt (1956) [31], et Parzen (1962) [28] sont les premiers à proposer une classe d'estimateurs à noyau d'une densité univariée. Cet estimateur est une fonction de deux paramètres, le noyau K et le paramètre de lissage h (ou la largeur de la fenêtre). Rosenblatt [31] reprenait l'idée de Fix et Hodges [10] en 1951, qui consistait à estimer la densité en un point, en comptant le nombre d'observations situées dans l'intervalle de longueur $2h$ et centré en ce point. L'estimateur à noyau en un point x, à partir d'un échantillon $X_1, X_2, ..., X_n$ issu d'une variable aléatoire X de densité f, est de la forme suivante :

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}).$$

Pendant plusieurs années, cet estimateur a été utilisé principalement pour l'estimation de la fonction de densité. Au début des années quatre-vingt, Leadbetter et Wold [26], Ramlau-Hansen [29] et Devroy et Gyorfi [30], proposent un estimateur à noyau pour une fonction d'intensité d'un processus stochastique. En 1985, Diggle [16] utilise cet estimateur pour l'estimation de l'intensité d'un processus de Poisson non homogène sous quelques conditions.

La qualité de l'estimateur à noyau dépend de deux paramètres : le noyau K et la fenêtre h. Les noyaux symétriques fournissent de bons résultats lorsque la densité à estimer est définie sur $\mathbb{R}$ [31, 28, 36]. Cependant, ils s'avèrent mal appropriés au cas de l'estimation des densités définies sur $[0, +\infty[$; cela est dû au poids qu'ils assignent en dehors du support positif de la densité à estimer [8, 9, 33, 2]. Par ailleurs, le choix de la fenêtre est d'une ultime importance pour la qualité du lissage. Dans le cas de l'estimation d'une densité, il s'est développé une importante bibliographie sur la sélection de ce paramètre. Plusieurs travaux ont montré que les estimateurs peuvent changer dramatiquement suite à une petite variation du paramètre de lissage. Dans la littérature, des méthodes diverses ont été proposées pour le choix de ce paramètre ; parmi elles, on peut citer les méthodes reposant sur la validation croisée, biaisée [34] et non biaisée [32, 4], la règle du pouce proposée par Silvermann (1986) [36], la méthode de ré-injection proposée par Scott, Tapia et Thomson [35]. Pour le cas de l'estimation de l'intensité, Diggle (1985)[16] a étudié les propriétés de l'estimateur à noyau dans le cas d'un processus de Cox, où il considère l'erreur quadratique moyenne comme une mesure de la qualité de l'estimateur. En 1990, Brooks et Marron [5] ont adapté les techniques de la validation croisée pour le choix du paramètre de lissage dans l'esimateur à noyau d'une intensité en se fixant comme objectif la minimisation de l'erreur quadratique intégrée. Malgré que les deux méthodes de sélection de h dans l'estimation de la densité et de l'intensité sont motivées de manières différentes, Diggle et Marron (1988)[15] ont prouvé qu'il y a une équivalence du choix de ce paramètre dans les deux méthodes.

Le premier objectif de ce travail est d'appliquer la méthode du noyau pour l'estimation des deux fonctions : densité et intensité. Deuxièmement, on présente l'équivalence qui existe entre les deux approches et les profits qu'elle permet d'avoir pour chaque approche dans le contexte de l'autre.

Ce mémoire est organisé en trois chapitres. Le premier est consacré à la présentation de l'estimateur à noyau d'une densité et d'une intensité en donnant ces propriétés statistiques et les méthodes du choix de la fenêtre pour les deux cas. L'équivalence du choix du paramètre de lissage fait l'objet du chapitre 2. Dans le dernier chapitre, nous présentons une étude de simulation pour l'estimation de la densité d'une loi normale et d'une loi exponentielle. On présente aussi les résultats de simulation d'une intensité d'un processus de Poisson non homogène qui sera utilisée pour établir l'équivalence présentée dans le deuxième chapitre. On termine ce mémoire par une conclusion générale et une bibliographie.

1

Estimation non paramétrique par la méthode du noyau

1.1 Introduction

La méthode du noyau est l'une des méthodes d'estimation non paramétrique la plus utilisée. Rosenblatt (1956)[31], suivi de Parzen (1962)[28], ont proposé une classe d'estimateurs à noyau d'une densité de probabilité. Cet estimateur est une fonction de deux paramètres : le noyau K et le paramètre de lissage h. Le succès rencontré par cet estimateur s'explique par sa simplicité, sa flexibilité et aussi ses propriétés de convergence. Il laisse à l'utilisateur une grande latitude non seulement dans le choix du noyau K, mais aussi dans le choix du paramètre de lissage h.
L'utilisation de cet estimateur ne se limite pas à l'estimation de la fonction de densité. En 1985, Diggle [16] utilise cet estimateur pour l'estimation de l'intensité d'un processus de Poisson non homogène sous quelques conditions. L'objectif de ce chapitre est de présenter les estimateurs à noyau des fonctions de densité et d'intensité, on donnera aussi les propriétés statistiques de ces estimateurs ainsi que les différentes méthodes du choix du paramètre de lissage.

1.2 Estimation de la densité

1.2.1 Introduction

Soit X une variable aléatoire de densité de probabilité inconnue f. Supposons que nous avons n observations $x_1, x_2, ..., x_n$ provenant de X. Le problème consiste à trouver un estimateur pour la fonction f à partir de cet échantillon issu de X. Pour cela, l'approche non paramétrique est la plus adéquate lorsqu'on ne possède aucune information précise sur la forme et la classe de la vraie densité. Dans cette approche, ce sont les observations qui vont nous permettre de déterminer un estimateur pour la densité f.

Dans cette section, on s'intéresse à la méthode du noyau pour l'estimation de la den-

sité de probabilité. L'estimateur à noyau sera présenté ainsi que ses différentes propriétés statistiques.

1.2.2 Définition du noyau de Parzen-Rosenblatt

En 1956, Rosenblatt [31] a proposé le premier estimateur à noyau pour la densité de probabilité $f(x)$. Six ans après, cet estimateur a été généralisé par Parzen(1962)[28] ; à partir de cette date, cet estimateur a pris le nom de l'estimateur de Parzen-Rosenblatt.

L'idée de l'estimateur par la méthode du noyau consiste à évaluer la densité $f(x)$ au point x en comptant le nombre d'observations tombées dans un certain voisinage de x sur $\mathbb{R}$.

Définition 1.1. Soientt $x_1, x_2, ..., x_n$ n observations d'une variable aléatoire X de densité de probabilité $f(x)$ et de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt$.
On appelle fonction de répartirion empirique associée à $x_1, ..., x_n$, la fonction aléatoire F_n : $\mathbb{R} \longrightarrow [0, 1]$ définie par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(x_i<x)}.$$

Ou encore :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i)_{]-\infty,x[}, \tag{1.1}$$

où $1(.)_{]-\infty,x[}$ est la fonction indicatrice sur $]-\infty, x[$.
La fonction de répartition empirique F_n est un estimateur simple de F [40].

$$nF_n(x) = \sum_{i=1}^{n} 1_{(x_i<x)} \ \underrightarrow{loi} \ \mathbf{B}(n, F(x)),$$

où **B** est la loi binomiale.
À partir de la définition d'une densité de probabilité (basée sur la dérivée de la fonction de répartition) et en utilisant l'équation (1.1), la densité f peut s'écrire en ses points de continuité :

$$f_h(x) = \lim_{h\longrightarrow 0} \frac{F_n(x+h) - F_n(x-h)}{2h}. \tag{1.2}$$

$$\begin{aligned} f_h(x) &= \frac{1}{2nh}\sum_{i=1}^{n} 1(x_i)_{]x-h,x+h[} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} 1_{\{x-h<x_i<x+h\}}. \end{aligned}$$

En posant :

$$\omega(u) = \begin{cases} 1/2, & \text{-1<u}\leq 1 \\ 0, & \text{sinon,} \end{cases}$$

on peut réecrire (1.2) sous la forme suivante :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \omega(\frac{x - x_i}{h}). \tag{1.3}$$

L'expression (1.3) définit l'estimateur à noyau (noyau uniforme ω) donné par Rosenblatt (1956)[31].
En 1962, Parzen [28] a généralisé cet estimateur en remplaçant la fonction ω (noyau uniforme) par une fonction noyau K satisfaisant la condition suivante :

$$\int_{-\infty}^{+\infty} K(u)du = 1.$$

Généralement, K est une densité de probabilité. Donc l'estimateur à noyau de Parzen est donné par :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{1.4}$$

où h est le paramètre de lissage et K est la fonction noyau définie sur $\mathbb{R}$ et vérifiant les conditions suivantes :

$$\int_{\mathbb{R}} K(y)dy = 1, \quad \int_{\mathbb{R}} yK(y)dy = 0, \quad \int_{\mathbb{R}} y^2K(y)dy = \sigma_K^2 < \infty.$$

L'estimateur donné par (1.4) est bien une densité de probabilité, en effet :

$$\begin{aligned}
\int_{\mathbb{R}} f_h(x)dx &= \int_{\mathbb{R}} \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})dx \\
&= \frac{1}{nh}\sum_{i=1}^{n}\int_{\mathbb{R}} K(\frac{x - x_i}{h})dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} K(u)du \\
&= 1.
\end{aligned}$$

1.2.3 Critères d'erreur

Lorsqu'on définit un estimateur, il y a un certain nombre de critères qui permettent d'évaluer la similarité de cet estimateur noté f_h par rapport à la vraie densité f à estimer. Parmi les nombreux critères proposés dans la littérature, on trouve ISE (Integrated Squared Error), MSE (Mean squared Error) et MISE (Mean Integrated Squared Error).
Dans ce qui suit, on donne la définition de chaque critère.

L'erreur quadratique intégrée (ISE) :

$$ISE(f(x), f_h(x)) = \int (f(x) - f_h(x))^2.$$

L'erreur quadratique moyenne (MSE) :

$$\begin{aligned} MSE(f(x), f_h(x)) &= \mathbb{E}[(f(x) - f_h(x))^2] \\ &= \mathbb{E}(f^2(x)) - 2\mathbb{E}(f(x).f_h(x)) + \mathbb{E}(f_h^2(x)) \\ &= \mathbb{E}(f^2(x)) - 2\mathbb{E}(f(x).f_h(x)) + \mathbb{E}(f_h^2(x)) + \mathbb{E}^2(f_h(x)) - \mathbb{E}^2(f_h(x)) \\ &= [\mathbb{E}(f_h(x)) - f(x)]^2 + \mathbb{E}(f_h^2(x)) - \mathbb{E}^2(f_h(x)) \\ &= [Biais(f_h(x))]^2 + Var(f_h(x)). \end{aligned}$$

L'erreur quadratique moyenne intégrée (MISE) :

$$\begin{aligned} MISE(f(x), f_h(x)) &= \int MSE(f(x), f_h(x))dx \\ &= \int \mathbb{E}[(f(x) - f_h(x))^2]dx \\ &= \int [Biais(f_h(x))]^2 + Var(f_h(x))dx. \end{aligned}$$

1.2.4 Propriétés de l'estimateur à noyau

Cette section est consacrée à la présentation de quelques résultats théoriques sur les propriétés statistiques de l'estimateur à noyau de Parzen-Rosenblatt donné par l'expression (1.4), à savoir :
-L'espérance, le biais et la variance.
-Le comportement asymptotique du biais et de la variance.
-Le MSE et le MISE de l'estimateur.
On donnera aussi la convergence d'un estimateur à noyau au sens du MSE et MISE.
L'espérance :

$$\begin{aligned} \mathbb{E}(f_h(x)) &= \mathbb{E}[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})] \\ &= \frac{1}{nh}\mathbb{E}[\sum_{i=1}^{n} K(\frac{x - x_i}{h})] \\ &= \frac{1}{h}\int_{\mathbb{R}} K(\frac{x-u}{h})f(u)du. \end{aligned}$$

On pose $y = \frac{x-u}{h} \quad \Rightarrow \quad dy = \frac{-du}{h}$.
Donc : $\mathbb{E}(f_h(x)) = \int_{\mathbb{R}} K(y)f(x - hy)dy$.
En faisant le développement limite de Taylor à l'ordre 2, au point y=0 de f(x-hy),
on obtient :
$f(x - hy) = f(x) - \frac{hy}{1!}f'(x) + \frac{h^2y^2}{2!}f''(x) + o(h^2)$.
Donc $\mathbb{E}(f_h(x))$ sera donné par :

$$\begin{aligned} \mathbb{E}(f_h(x)) &= \int_{\mathbb{R}} K(y)[f(x) - yhf'(x) + \frac{h^2y^2}{2!}f''(x) + o(h^2)]dy \\ &= f(x)\int_{\mathbb{R}} K(y)dy - hf'(x)\int_{\mathbb{R}} yK(y)dy + \frac{h^2y^2}{2!}f''(x)\int_{\mathbb{R}} y^2K(y)dy + o(h^2)]dy. \end{aligned}$$

Si le noyau K est symétrique par rapport à 0, ie :

$$\int_{\mathbb{R}} yK(y)dy = 0 \quad et \quad \int_{\mathbb{R}} y^2K(y)dy < \infty,$$

alors :

$$\mathbb{E}(f_h(x)) = f(x) + \frac{h^2y^2}{2}f''(x)\sigma_2(K) + o(h^2), \tag{1.5}$$

avec : $\sigma_2(K) = \int_{\mathbb{R}} y^2K(y)dy$.

Le Biais :

$$\begin{aligned} Biais(f_h(x)) &= \mathbb{E}(f_h(x)) - f(x) \\ &= \int_{\mathbb{R}} yK(y)f(x-hy)dy - f(x) \\ &= f(x) + \frac{h^2y^2}{2!}f''(x)\sigma_2(K) + o(h^2) - f(x) \\ &= \frac{h^2y^2}{2}f''(x)\sigma_2(K) + o(h^2). \end{aligned} \tag{1.6}$$

La variance :

$$\begin{aligned} Var((f_h(x)) &= Var[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})] \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n} Var(K(\frac{x-x_i}{h})) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}[\mathbb{E}(K(\frac{x-x_i}{h}))^2] - \frac{1}{n^2h^2}\sum_{i=1}^{n}[\mathbb{E}(K(\frac{x-x_i}{h}))]^2 \\ &= \frac{1}{nh^2}\int_{\mathbb{R}}[K(\frac{x-u}{h})]^2f(u)du - \frac{1}{n}[\frac{1}{h}\int_{\mathbb{R}} K(\frac{x-u}{h})f(u)du]^2. \end{aligned}$$

On pose $y = \frac{x-u}{h} \quad \Rightarrow \quad dy = \frac{-du}{h}$.
Donc :

$$Var((f_h(x)) = \frac{1}{nh}\int_{\mathbb{R}}(K(y))^2f(x-hy)dy - \frac{1}{n}[\int_{\mathbb{R}} K(y)f(x-hy)dy]^2.$$

En faisant le développement limite de Taylor à l'ordre 2, au point $y = 0$ de $f(x-hy)$, on obtient :
$f(x-hy)$=$f(x) - \frac{hy}{1!}f'(x) + \frac{h^2y^2}{2!}f''(x) + o(h^2)$.
D'où :

$$Var(f_h(x)) = \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy - \frac{f'(x)}{n}\int_{\mathbb{R}} yK^2(y)dy - \frac{1}{n}(f(x) + Biais^2(f_h(x))). \tag{1.7}$$

Remarque 1.1. Les expressions approchées du biais et de la variance données respectivement par les équations (1.6) et (1.7) montrent que le paramètre de lissage h est inversement corrélé

à la variance et corrélé au biais. Par conséquent, lorsque la fenêtre h est très petite, le biais de l'estimateur à noyau est très petit face à sa variance. Dans ce type de situation, l'estimateur est très volatile et on parle de sous-lissage (under-smoothing). En revanche, lorsque h grandit, la variance devient petite et c'est le biais qui devient dominant. L'estimateur est alors très peu variable. On parle alors d'un effet de sur-lissage (over-smoothing).
En pratique, il est primordial de trouver la bonne dose de lissage qui permet d'éviter le sous-lissage et le sur-lissage.

Le comportement asymptotique du biais :

Théorème 1.1. *(Parzen [28]) Si on a :*

1. $\lim_{n\longrightarrow+\infty} h(n) = 0 \quad et \quad \lim_{n\longrightarrow+\infty} |yK(y)| = 0;$
2. $Sup|K(y)| < \infty;$
3. $\int_{\mathbb{R}} K(y)dy = 1;$

alors : l'estimateur $f_h(x)$ est asymptotiquement sans biais, ie :

$$\lim_{n\longrightarrow+\infty} \mathbb{E}(f_h(x)) = f(x).$$

Le comportement asymptotique de la variance :
Si les conditions du théorème (1.1) sont satisfaites, alors :

$$\lim_{n\longrightarrow+\infty} nhVar(f_h(x)) = f(x)\int_{\mathbb{R}} K^2(y)dy;$$

telle que : f est une densité continue $\forall x \in \mathbb{R}$.

Le MSE et le MISE d'un estimateur à noyau :

$$\begin{aligned} MSE(f(x), f_h(x)) &= \mathbb{E}[(f(x) - f_h(x))^2] \\ &= [Biais(f_h(x))]^2 + Var(f_h(x)) \\ &= [\frac{h^2y^2}{2}f''(x)\sigma_2(K) + o(h^2)]^2 + \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(y)dy + o(\frac{1}{nh}) \\ &= \frac{h^4}{4}(f''(x))^2\sigma_2^2(K) + \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(y)dy + o(\frac{1}{nh}) + o(h^2). \end{aligned}$$

$$\begin{aligned} MISE(f(x), f_h(x)) &= \int_{\mathbb{R}} MSE(f(x), f_h(x))dx \\ &= \int_{\mathbb{R}} \{\frac{h^4}{4}(f''(x))^2\sigma_2^2(K) + \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(y)dy + o(\frac{1}{nh}) + o(h^2)\}dy \\ &= \frac{1}{nh}\int_{\mathbb{R}} K^2(y)dy + \frac{h^4}{4}\sigma_2^2(K)\int_{\mathbb{R}} (f''(x))^2dx. \end{aligned}$$

Sous les conditions $h \to 0$ et $nh \to \infty$ quand n tend vers l'infini, on a le développement asymptotique de MISE, que l'on note AMISE :

$$AMISE(f(x), f_h(x)) = \frac{1}{nh}R(K) + \frac{h^4}{4}R(f''(x))\sigma_2^2(K), \tag{1.8}$$

avec :

$$R(q(y)) = \int_{\mathbb{R}} q^2(y)dy.$$

L'estimateur de h qui minimise ce critère est donné par l'équation (1.9) suivante :

$$h_{AMISE} = [\frac{R(K)}{nR(f'')\sigma_2^2(K)}]^{1/5}. \tag{1.9}$$

La convergence de l'estimateur à noyau

• **Convergence en moyenne quadratique**
Le MSE d'un estimateur à noyau est donné par la formule suivante :

$$MSE(f(x), f_h(x)) = \frac{h^4}{4}(f''(x))^2\sigma_2^2(K) + \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(y)dy + o(\frac{1}{nh}) + o(h^2). \tag{1.10}$$

Théorème 1.2. *(Parzen [28])*
Si $\lim\limits_{n\to\infty} h(n) = 0$ *et* $\lim\limits_{n\to\infty} nh(n) = \infty$,
et K satisfait les conditions suivantes :
- $\sup\limits_y |K(y)| < \infty$ *et* $\lim\limits_{y\to\infty} |yK(y)| = 0$,
- $\int_{\mathbb{R}} |K(y)|dy < \infty$ *et* $\int_{\mathbb{R}} K(y)dy = 1$,
alors l'estimateur $f_h(x)$ *est consistant en moyenne quadratique, c'est-à-dire :*

$$\lim_{n\to\infty} MSE(f(x), f_h(x)) = 0,$$

pour tout x pour lequel la densité f est continue.

• **Convergence en moyenne quadratique intégrée**

Théorème 1.3. *(Parzen [28])*
Si K est un noyau de Parzen-Rosenblatt, et on a :

$$\lim_{n\to\infty} h(n) = 0 \quad et \quad \lim_{n\to\infty} nh(n) = \infty,$$

alors :

$$(\forall f \in \mathbb{L}^p), \lim_{n\to\infty} MISE(f(x), f_h(x)) = 0.$$

On note par $\mathbb{L}^p$ *: l'ensemble des fonctions réelles de puissance p-ième intégrable, c'est-à-dire l'ensemble des fonctions f définies sur* $\mathbb{R}$*, telles que* $\int |f(x)|^p dx < \infty$.

Après une présentation générale de la méthode d'estimation de la densité de probabilité, nous avons introduit et défini la méthode du noyau de Parzen-Rosenblatt donnée par la formule (1.4). Cette partie nous a permis d'identifier les paramètres de la méthode et de souligner la nécessité et l'intérêt du choix du couple (K, h). Le noyau K peut être discret ou continu, et selon la symétrie du domaine de définition, il existe deux catégories principales de noyaux, les noyaux symétriques et les noyaux asymétriques. Quant au choix du paramètre de lissage h, qui est crucial pour la qualité de l'estimateur, plusieurs méthodes ont été développées .
Dans ce qui suit, on présente quelque noyaux des deux catégories, ensuite on donnera deux classes de méthodes du choix de la fenêtre h.

1.2.5 Les noyaux usuels

On donne ici une brève présentation de quelques noyaux usuels de chaque catégorie dans le cas continu.

a- Noyaux symétriques

Noyau uniforme (noyau de Rosenblatt) :
Ce noyau a été proposé par Rosenblatt (1956)[31] (détaillé dans la section précédente). L'avantage de ce noyau est la simplicité de sa forme, qui est donnée comme suit :

$$K(u) = \begin{cases} 1/2, & \text{si } |u| \leq 1\,; \\ 0, & \text{sinon.} \end{cases}$$

Noyau triangulaire :
Ce noyau a un avantage par rapport au noyau uniforme, il est continu partout, ce qui conduit à une estimation continue de f, sa forme est :

$$K(u) = \begin{cases} (1 - |u|), & \text{si } |u| \leq 1\,; \\ 0, & \text{sinon.} \end{cases}$$

Noyau Gaussien :
Ce noyau s'écrit sous la forme :

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}, \quad u \in \mathbb{R}.$$

Noyau Epanechnikov :
Ce noyau a été proposé par Epanechnikov en 1969 [18], il est défini par :

$$K(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - |\frac{u^2}{5}|), & \text{si } |u| \leq \sqrt{5}\,; \\ 0, & \text{sinon.} \end{cases}$$

La représentation graphique des noyaux définis ci-dessus est donnée par la figure (1.1) :

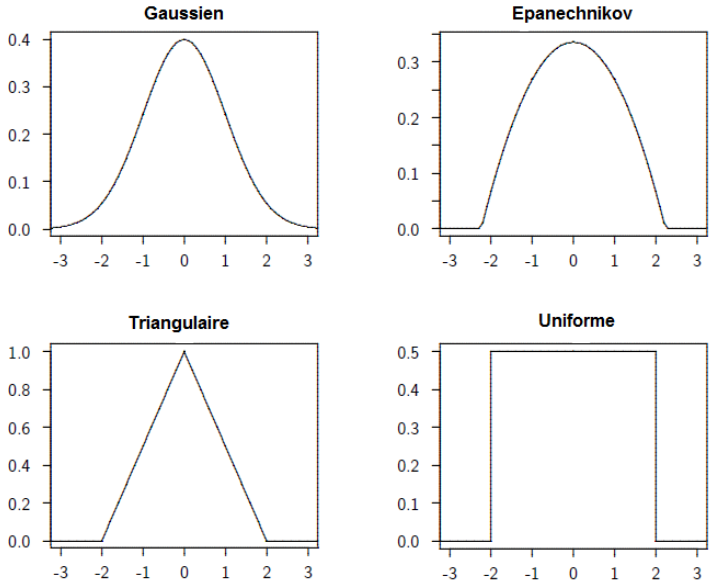

FIG. 1.1 – La représentation graphique des noyaux symétriques

b- Noyaux asymétriques

Effet du biais aux bornes

Généralement, le noyau gaussien est souvent utilisé en raison de sa simplicité et ses propriétés asymptotiques qui sont établies par plusieurs auteurs. Cependant l'inconvénient de ce noyau est qu'il attribue des poids positifs pour des valeurs de x qui sont à l'extérieur du support dans le cas de l'estimation d'une densité bornée ou semi bornée. Cela cause le problème du biais aux bornes ce qui réduit la qualité de l'estimateur.

Ce problème a donné un essor pour de nouvelles méthodes et de nouveaux noyaux pour l'estimation de la densité pour le cas de données définies sur un support borné. Parmi ces méthodes, on peut citer :
La méthode de réflexion (noyau miroir) de Schuster (1985)[33], la méthode de renormalisation locale de Diggle (1985)[16] et Hardle (1990)[22].
L'inconvénient des estimateurs donnés par ces méthodes est qu'ils attribuent des poids négatifs aux valeurs du voisinage des bornes.

La solution la plus récente est d'utiliser des noyaux asymétriques qui n'assignent aucun poids à l'extérieur du support. Chen (1999)[8] et Chen (2000)[9] propose, respectivement, le noyau Beta pour les densités à support compact (exemple : [a,b]) et le noyau Gamma pour les densités à support positif (c'est-à-dire sur $[0, +\infty]$).
Ces deux noyaux asymétriques seront présentés dans le paragraphe suivant.

Noyau Gamma :

Soit $x_1, x_2, ..., x_n$, n observations d'une variable aléatoire X de densité de probabilité *f*, inconnue et définie sur un support $[0, +\infty[$.
L'objectif est d'estimer cette densité par un noyau Gamma.

La première forme du noyau Gamma est définie par (voir Bouezmarni et Scaillet [2]) :

$$K_{(\frac{x}{h}+1,h)}(t) = \frac{t^{(x/h)}e^{-(t/h)}}{h^{(x/h)+1}\Gamma((x/h)+1)}, \tag{1.11}$$

avec

$$\Gamma(a) = \int_0^{+\infty} e^{-x}x^{a-1}dx, \quad \forall a > 0 \quad et \quad t > 0.$$

Donc, l'estimateur à noyau Gamma est défini comme suit :

$$f_{G_1}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{(\frac{x}{h}+1,h)}(x_i). \tag{1.12}$$

Biais : le biais de $f_{G_1}(x)$ est [2] :

$$Biais(f_{G_1}(x)) = h(f'(x) + \frac{1}{2}xf''(x)) + o(h). \tag{1.13}$$

L'analyse du biais de l'éstimateur à noyau gamma est détaillée dans [17].
Variance : la variance de $f_{G_1}(x)$ est [2] :

$$Var(f_{G_1}(x)) = \frac{1}{n}Var(K_{(\frac{x}{h}+1,h)}(x_i)) = \frac{1}{n}\mathbb{E}(K_{(\frac{x}{h}+1,h)}(x_i))^2 + o(\frac{1}{n}). \tag{1.14}$$

Soit η_x une variable aléatoire de distribution gamma de paramètres $(2x/h+1, h)$, on aura :

$$\mathbb{E}\{K_{(\frac{x}{h}+1,h)}(X_i)\}^2 = B_h(x)\mathbb{E}\{f(\eta_x)\};$$

avec $B_h(x)$ est donnée par :

$$B_h(x) = \begin{cases} \frac{1}{2\sqrt{\pi}}n^{-1}h^{-1/2}x^{-1/2}f(x), & x/h \to \infty\,; \\ \frac{\Gamma(2k+1)}{2^{1+2x}\Gamma^2(k+1)}n^{-1}h^{-1}f(x), & x/h \to k\,; \end{cases}$$

et k>0.
MSE [2] :

$$MSE(f_{G_1}(x)) = h^2\{f'(x) + \frac{1}{2}xf''(x)\}^2 + n^{-1}B_hf(x). \tag{1.15}$$

MISE [2] :

$$MISE(f_{G_1}(x)) = h^2\int_0^{\infty}\{f'(x)+\frac{1}{2}xf''(x)\}^2dx+\frac{1}{2\sqrt{\pi}}n^{-1}h^{-\frac{1}{2}}\int_0^{\infty}x^{-\frac{1}{2}}f(x)dx+o(h^2+n^{-1}h^{-\frac{1}{2}}). \tag{1.16}$$

Le paramètre de lissage optimal [2] :

$$h^*_{G_1} = \frac{[\frac{1}{2\sqrt{\pi}}\int_0^{\infty}x^{-1/2}f(x)dx]^{\frac{2}{5}}}{4^{2/5}[\int_0^{\infty}(xf'(x)+\frac{1}{2}xf''(x))^2dx]^{\frac{2}{5}}}n^{-\frac{2}{5}}. \tag{1.17}$$

MISE optimal [2] :

$$MISE^*(f_{G_1}(x)) = \frac{5}{4^{4/5}}[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{-1/2}f(x)dx]^{\frac{4}{5}}[\int_0^\infty \{f'(x)+\frac{1}{2}xf''(x)\}^2dx]^{\frac{1}{5}}n^{-4/5}. \quad (1.18)$$

En raison de la contribution indésirable de f' dans le biais, donné par la formule (1.13), de l'estimateur $f_{G_1}(x)$, une autre version de cet estimateur notée $f_{G_2}(x)$ a été donnée par Chen.

$$f_{G_2}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{(\rho_h(x),h)}(x_i), \quad (1.19)$$

avec :

$$K_{(\rho_h(x),h)}(t) = \frac{t^{\rho_h(x)-1}e^{-t/h}}{h^{\rho_h(x)}\Gamma(\rho_h(x))},$$

et

$$\rho_h(x) = \begin{cases} x/h, & \text{si } x \geq 2h\,; \\ \frac{1}{4}(x/h)^2+1, & \text{si } x \in [0,2h[. \end{cases}$$

Biais : le biais de l'estimateur $f_{G_2}(x)$ est donné par [2] :

$$Biais(f_{G_2}(x)) = \begin{cases} \frac{1}{2}xf''(x)h+o(h), & \text{si } x \geq 2h\,; \\ \xi_h(x)hf'(x)h+o(h), & \text{si } x \in [0,2h[. \end{cases} \quad (1.20)$$

Où : $\xi_h(x) = (1-x)(\rho_h(x)-\frac{x}{h})/(1+h\rho_h(x)-x)$.
On a : $\int_0^\infty (xf''(x))^2dx < \infty$ et $xf''(x)$ converge vers 0 quand x tend vers l'infini.
Alors le biais sera minimal en augmentant x.
La variance de cet estimateur $(f_{G_2}(x))$ est la même que celle de l'estimateur $f_{G_1}(x)$.
MSE [2] :

$$MSE(f_{G_2}(x)) = \begin{cases} \frac{1}{4}h^2\{xf''(x)\}^2+n^{-1}B_h(x)f(x), & x \geq 2h\,; \\ h^2\{\xi_h(x)f'(x)\}^2+n^{-1}B_h(x)f(x), & x \in [0,2h). \end{cases} \quad (1.21)$$

MISE [2] :

$$MISE(f_{G_2}(x)) = \frac{1}{4}h^2\int_0^\infty \{xf''(x)\}^2dx + \frac{1}{2\sqrt{\pi}}n^{-1}h^{-\frac{1}{2}}\int_0^\infty x^{-\frac{1}{2}}f(x)dx + o(h^2+n^{-1}h^{-\frac{1}{2}}). \quad (1.22)$$

Le paramètre de lissage optimal [2] :

$$h^*_{G_2} = \frac{[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{-1/2}f(x)dx]^{\frac{2}{5}}}{[n\int_0^\infty (xf''(x))^2dx]^{\frac{2}{5}}}. \quad (1.23)$$

MISE optimal [2] :

$$MISE^*(f_{G_2}(x)) = \frac{5}{4^{4/5}}\{\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{-1/2}f(x)dx\}^{\frac{4}{5}}\{\int_0^\infty \{xf''(x)\}^2\}^{1/5}n^{-4/5}. \quad (1.24)$$

Convergence des noyaux Gamma :

Le noyau gamma est facile à établir, il élimine le biais aux bornes et est souvent non

négatif. Il atteint le taux de convergence optimal pour les variables $i.i.d$ au sens de MISE dans la classe des estimateurs à noyaux non négatifs. De plus, il permet une réduction de la variance lors du lissage en s'éloignant des bornes. Particulièrement lorsque on utilise la normalité asymptotique de l'estimateur à noyau gamma.

Bouezmarni et Scaillet (2003)[2] ont donné les conditions de convergence faible de l'estimateur à noyau gamma sur un impact $[0, +\infty[$ lorsque f est continue sur ce support et la convergence faible au sens MIAE (Mean Integrated Absolute Error). Pour les densités non bornées à l'origine (c'est-à-dire au voisinage de zero) ils ont examiné les performances de cet estimateur par simulation et ils ont prouvé sa convergence en probabilité vers l'infini à l'origine.

Fernandez et Monteiro (2005)[20] ont établi le théorème central limite pour l'estimateur fonctionnel par le noyau gamma. Bouezmarni et Ronbouts (2006)[3] ont démontré la convergence presque sûre au sens du MISE et la normalité asymptotique de cet estimateur.

Noyau Beta :

Le noyau Beta a été proposé par Brown et Chen (1999)[6], et Chen (1999,2000)[8, 9] pour l'estimation non paramétrique de la courbe de régression et des densités unidimensionnelles définies sur un support compact.

L'idée de Harrell et Davis (1982)[23] et Chen (1999)[8] est d'utiliser le noyau Beta pour estimer la densité à support compact [0,1] et de régler ainsi le problème du biais aux bornes. L'estimateur sera alors de la forme :

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K(x_i, \frac{x}{h} + 1, \frac{1-x}{h} + 1), \tag{1.25}$$

où $K(., \alpha, \beta)$ représente la densité de la distribution Beta de paramètres α et β,

$$K(x, \alpha, \beta) = \frac{x^{\alpha}(1-x)^{\beta}}{\mathbf{B}(\alpha, \beta)}, \qquad x \in [0, 1],$$

avec :

$$\mathbf{B}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad et \quad \Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx, \quad \forall a > 0.$$

Le noyau beta a deux avantages, premièrement il peut parfaitement estimer les densités à support compact et deuxièmement il possède une forme flexible qui change le lissage dans le sens naturel quand on s'éloigne des bornes. Par conséquent, le noyau beta élimine le biais aux bornes et fournit une réduction de la variance.

Charpentier, Fermanian et Scaillet [7] ont montré par simulation que l'estimateur à noyau beta est plus performant quand on le compare à d'autres estimateus avec des noyaux standards.

1.2.6 Le choix du paramètre de lissage

Le paramètre de lissage est le second élément de la méthode d'estimation à noyau. Ce paramètre est indispensable pour la convergence de l'estimateur à noyau et donc l'efficacité du lissage et la qualité de l'estimation (voir figure (1.2)). Plusieurs méthodes pour choisir ce

paramètre ont été développées dans la littérature et quelques études comparatives ont été effectuées sur ces méthodes. Ces techniques sont regroupées en deux classes. On présente ici les principales méthodes de chaque classe.

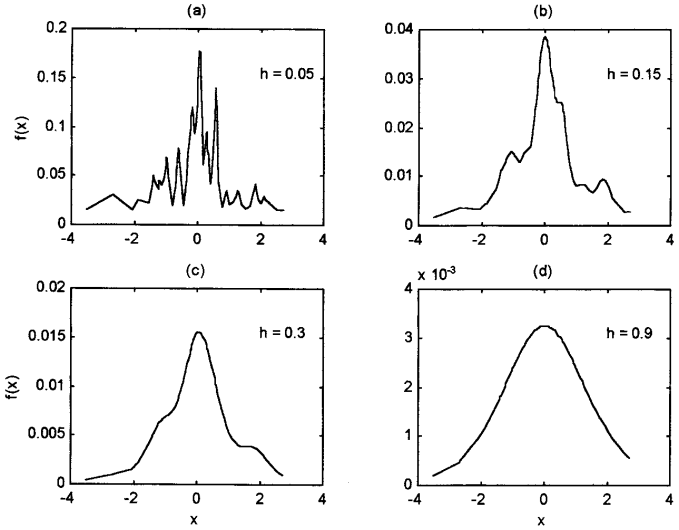

FIG. 1.2 – L'influence du paramètre de lissage h sur la qualité de l'estimation.

I- Première classe

La majorité des méthodes de cette classe ont été proposées avant 1990. Dans ce qui suit, on donne une présentation de quelques méthodes les plus connues.

a- La méthode de validation croisée par moindres carrés

Rudemo (1982)[32] et Bowman (1984)[4] ont donné l'idée de cette méthode. En utilisant la formule développée de l'erreur quadratique intégrée $ISE(h)$, on choisit le paramètre de lissage h qui minimise cette erreur.

$$ISE(h) = \int_{\mathbb{R}} (f(x) - f_h(x))^2 dx = \int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx + \int_{\mathbb{R}} f^2(x)dx. \qquad (1.26)$$

On remarque que le troisième terme de la formule (1.26) ne dépend pas de h, donc on peut choisir le h de façon à ce qu'il minimise le critère de la validation croisée défini par :

$$UCV(h) = ISE(h) - \int_{\mathbb{R}} f^2(x)dx = \int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx. \qquad (1.27)$$

On doit donc trouver un estimateur pour $\int_{\mathbb{R}} f_h(x)f(x)dx$. Remarquons que :

$$\int_{\mathbb{R}} f_h(x)f(x)dx = \mathbb{E}(f_h(x)).$$

Son estimateur empirique est alors : $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x)$,

avec : $f_{h,i}(x) = \frac{1}{(n-1)h}\sum_{\substack{j=1\\ i\neq j}}^{n} K(\frac{x-x_j}{h})$ est l'estimateur de la densité construit à partir de l'ensemble des points sauf le point x_i (la méthode de leave one out).

En remplaçant cet estimateur dans (1.27), pour un noyau K, le critère de la validation croisée par les moindres carrés est donné par :

$$LSCV(h) = \frac{R(K)}{nh} + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}[\int_{\mathbb{R}} \frac{1}{(nh)^2}K(\frac{x-x_i}{h})K(\frac{x-x_j}{h})dx - \frac{2}{n(n-1)h}K(\frac{x_i-x_j}{h})]. \tag{1.28}$$

On note h_{LSCV} l'estimateur de h qui minimise le critère LSCV(h). De plus, si K est un noyau gaussien le critère LSCV(h) est donné par la proposition suivante :

Proposition 1.1. *(Jean [24])*
Soit $x_1, x_2, ..., x_n$ n observations d'une variable aléatoire X de densité de probabilité f. En utilisant un noyau gaussien, on obtient :

$$LSCV(h) = \frac{1}{2n^2h\sqrt{\pi}}(n+2\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\exp(-(\frac{x_i-x_j}{2h})^2)) - \frac{2}{\sqrt{2}\pi n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\exp(-\frac{(x_i-x_j)^2}{2h^2}).$$

b- La méthode de validation croisée biaisée

Cette méthode a été proposée par Scott et Terrell (1987)[34] pour remédier aux problèmes de validation croisée non biaisée. Il s'agit d'introduire un biais dans la formule de LSCV afin de réduire sa variance.
Le paramètre de lissage choisi par cette méthode est la valeur de h qui minimise un estimateur du critère AMISE (donné par (1.8)). On peut estimer le AMISE si l'on estime $R(f'')$. Un estimateur naturel de ce terme est donné par $R(f''_h)$ où f_h est l'estimateur de la densité par la méthode du noyau.

Lemme 1.1. *(Scott et Terrell [34])*
En supposant que le noyau K satisfait les conditions suivantes :
$\int K''(u)du = 0, \int uK''(u)du = 0, \int u^2K''(u)du = 2,$
on obtient le développement asymptotique :

$$\mathbb{E}(R(f''_h)) = R(f'') + \frac{R(K'')}{nh^5} + o(h^2).$$

Proposition 1.2. *(Scott et Terrell [34])*
Soit $x_1, x_2, ..., x_n$ n observations d'une variable aléatoire X de fonction de densité f. Pour un noyau K, on obtient :

$$BCV(h) = \frac{h^4}{4}\sigma_2^2(K)[R(f''_h) - \frac{R(K'')}{nh^5}] + \frac{R(K)}{nh}. \tag{1.29}$$

II- Deuxième classe

Beaucoup de méthodes ont été proposées dans cette classe (voir la monographie de Wand et Jones (1994)[38]). On présente ici quelques approches.

a- La règle du pouce (Rule of thumb)

L'idée de cette méthode revient à Deheuvels (1977)[12] avant d'être publiée par Silverman (1986)[36]. Le choix du paramètre de lissage par cette méthode consiste à remplacer la partie inconnue $R(f'')$ dans l'expression de l'estimateur optimal h_{AMISE}, donné par l'équation (1.9), par une distribution classique afin d'obtenir un estimateur pour h.
Si on choisit f comme étant la distribution normale de moyenne 0 et de variance σ^2, on a alors :

$$R(f'') = \int (f''(x))^2 dx = \frac{3}{8\sqrt{\pi}}\sigma^{-1/5}. \tag{1.30}$$

De plus, si on utilise un noyau gaussien, alors la valeur de h_{AMISE} que l'on note dans ce cas par h_{rot} et en substituant la valeur obtenue dans (1.30), on aura :

$$\begin{aligned} h_{rot} &= (4\pi)^{-1/10}[\frac{3}{8}\pi^{-1/2}\sigma]n^{-1/5} \\ &= (\frac{4}{3})^{1/5}\sigma n^{-1/5} \\ &= 1.06\sigma n^{-1/5}. \end{aligned} \tag{1.31}$$

Il suffit donc d'estimer σ à partir des données et de le remplacer dans la formule (1.31) de h_{rot}.

b- La méthode de ré-injection (plug-in)

En adoptant le critère de l'Erreur Quadratique Moyenne Intégrée (MISE), Scott, Tapia et Thomson [35] choisissent d'estimer la fonction $R(f'')$ dans l'expression de h_{AMISE} donnée par l'équation (1.9) à l'aide de l'estimateur naturel $\hat{R}_h(f'')$ défini comme suit :

$$\hat{R}_h(f'') = R(f_h''),$$

où f_h'' désigne la dérivée seconde de l'estimateur à noyau f_h. Avec un noyau K deux fois dérivable, on a :

$$f_h''(x) = \frac{1}{nh^3}\sum_{i=1}^{n} K''(\frac{x - x_i}{h}).$$

En choisissant par exemple le noyau gaussien :

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp(\frac{-u^2}{2}), u \in \mathbb{R},$$

l'estimateur $\hat{R}_h(f'')$ s'écrit comme suit :

$$\hat{R}_h(f'') = \frac{3}{8\sqrt{\pi}n^2h^9}\sum_{i=1}^{n}\sum_{j=1}^{n}[h^4 - (x_i - x_j)^2h^2 + \frac{1}{12}(x_i - x_j)^4]\exp[-\frac{(x_i - x_j)^2}{4h^2}].$$

Il est important de noter que la largeur de la fenêtre h contrôlant l'estimateur $\hat{R}_h(f'')$ de $R(f'')$ a été choisie identique à la largeur de la fenêtre intervenant dans l'estimateur f_h de f. En supposant que la quantité $R(f'')$ devrait être robuste par rapport à une erreur de spécification sur f, Scott, Tapia et Thomson [35] proposent finalement d'injecter l'estimateur $\hat{R}_h(f'')$ dans l'expression de h_{AMISE}, on obtient l'estimateur de h noté h_p :

$$h_p = [\frac{R(K)}{n\hat{R}_h(f'')\sigma_2^2(K)}]^{1/5}. \tag{1.32}$$

c- La méthode de bootstrap

La méthode bootstrap est une technique de ré-échantillonnage. Les premières versions de cette méthode ont été proposées par Taylor (1989)[37], Faraway et Jhun (1990)[19]. Elle consiste à considérer la fenêtre qui minimise l'approximation de MISE par bootstrap comme un estimateur de paramètre de lissage. Soit h_b le paramètre qui minimise l'erreur quadratique moyenne intégrée par l'une des méthodes précédentes.

Pour calculer la valeur de la fenêtre par la technique de bootstrap on doit ré-échantillonner par cette technique à partir de l'échantillon initial et construire ensuite l'estimateur de bootstrap qui s'écrit sous la forme suivante [19] :

$$f_h^j(x) = \frac{1}{nh}\sum K(\frac{x - x_i^j}{h}), \quad j = 1, 2, ...B;$$

où B est le nombre de réplications de bootstrap.

Nous construisons un estimateur initial de la densité f_{h_b} ensuite nous ré-échantillonnons par la technique de bootstrap à partir de l'échantillon initial pour construire les estimateurs $f_h^j(x)$,$j = 1, 2, ...B$.

À la fin, nous obtenons la fenêtre bootstrapée h_{boot} par la minimisation de $BMISE(h, h_b)$ par rapport à h. La quantité à minimiser, BMISE, est donnée par la formule suivante :

$$BMISE(h, h_b) = \frac{1}{B}\sum_{j=1}^{B}\int (f_h^j(x) - f_{h_b}(x))^2 dx. \tag{1.33}$$

1.3 Estimation de l'intensité

1.3.1 Introduction

On considère N points sur un intervalle [0,T]. Ces points peuvent être des dates d'occurrence 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 de modéliser ces données numériques, ordonnées sur [0,T], est d'admettre qu'elles décrivent une réalisation d'un processus ponctuel. Parmi les processus stochastiques les plus utilisés pour la modélisation de ces différentes situations, on trouve en premier lieu le processus de Poisson non homogène.

Ce processus est beaucoup utilisé dans plusieurs domaines grâce à son avantage d'avoir une structure probabiliste simple permettant de faire des analyses statistiques et de calculer

les grandeurs d'interêt.Un processus de Poisson non homogène est caractérisé par sa fonction d'intensité non constante. Dans cette section, on s'interesse à l'estimation de cette fonction par la méthode du noyau dans le cadre de l'estimation non paramétrique. Dans la littérature, on trouve deux modèles mathématiques pour l'estimation de la fonction d'intensité ; le premier est le modèle multiplicatif simple qui est une forme spécifique du modèle multiplicatif d'intensité donné par Aalen (1978)[1]. Le deuxième est celui introduit par Diggle (1985)[16] ; ce dernier a étudié 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 d'autres problèmes d'estimation à noyau. Ainsi, Brooks et Marron (1991) [5] ont obtenu des résultats d'optimalité asymptotique pour le choix de ce paramètre dans un modèle poissonien non homogène.

Dans cette section, après quelques définitions nécessaires pour la compréhension de la suite de travail, on présentera la démarche suivie par Diggle (1985) [16], Diggle et Marron (1988) [15] pour l'estimation de la fonction d'intensité d'un processus de Poisson non homogène. On donne aussi une brève présentation du modèle multiplicatif d'Aalen (1978)[1].

1.3.2 Processus de Poisson homogène

Les processus aléatoires (ou stochastiques) sont des processus qui décrivent l'évolution d'une variable aléatoire en fonction du temps. Soit $(Y_n)_{n\geq 1}$ le processus des temps d'occurrences qui représente les instants d'occurrences des événements. Par convention, $Y_0 = 0$. À partir de ce processus $(Y_n)_{n\geq 1}$, il est possible de définir le processus de comptage, noté $N(t)$, qui représente le nombre d'événements qui se sont produits entre 0 et t :

Définition 1.2. Un processus stochastique $\{N(t), t \geq 0\}$ est un processus de comptage si $N(t)$ représentant le nombre total d'événements qui se sont produits entre 0 et t est donné par :

$$N(t) = \sum_{i=1}^{n} 1_{(Y_i \leq t)}, \quad \forall t \geq 0.$$

N(t) doit donc satisfaire les quatre conditions suivantes :
i) N(t)≥0 ;
ii) N(t) a des valeurs entières uniquement ;
iii) Si s < t, alors N(s) ≤ N(t) ;
iv) Pour s < t, N(t) - N(s) est le nombre d'événements qui ont eu lieu entre s et t.

Un processus de comptage est un processus discret à temps continu.

Définition 1.3. On dit qu'un processus de comptage $\{N(t), t \geq 0\}$ est un processus de Poisson d'intensité $\lambda > 0$ si :
(H1) : N(0)=0 ;
(H2) : Le processus est à accroissements stationnaires et indépendants ;
(H3) : $\forall 0 \leq s < t$, la variable aléatoire $N(t) - N(s) \rightsquigarrow \mathcal{P}(\lambda(t-s))$:

$$\forall k \in \mathbb{N}, \mathbb{P}[N(t) - N(s) = k] = \frac{(\lambda(t-s))^k}{k!} e^{-\lambda(t-s)}.$$

Le paramètre d'intensité λ d'un processus de Poisson est tel que :
$\forall t \geq 0$

$$\lambda t = \mathbb{E}(N(t)),$$

car : $N(t) \rightsquigarrow \mathcal{P}(\lambda t)$.
Il vient que :

$$\mathbb{E}(N(1)) = \lambda.$$

Or : $N(1)$ représente le nombre d'événements dans l'intervalle $[0, 1]$.
La figure suivante donne la représentation d'une trajectoire d'un processus de Poisson homogène de paramètre $\lambda = 2$.

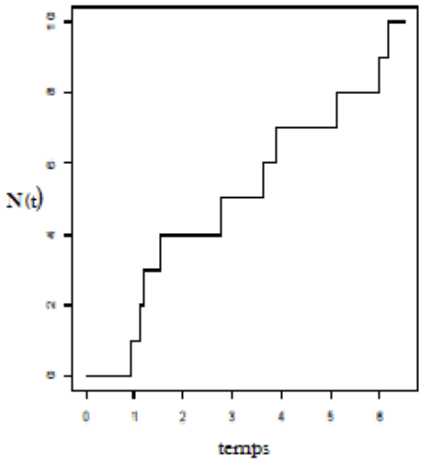

FIG. 1.3 – Trajectoire d'un processus de Poisson homogène.

1.3.3 Processus de Poisson non homogène

La différence entre un processus de Poisson homogène et un processus de Poisson non homogène réside dans le fait que les accroissements ne sont plus stationnaires. Cela est plus réaliste car, en général, le taux d'occurrence d'événements dépend du temps. Si on prend l'exemple d'un restaurant, ce taux sera élevé entre 11h et 13h puis faible dans l'après midi, puis à nouveau élevé entre 18h et 20h (c'est-à-dire, le taux d'occurrence d'événements n'est pas constant).

Il est courant de parler d'un processus de Poisson non homogène en terme de non stationnaire par opposition au processus de Poisson stationnaire.

Définition 1.4. Un processus de comptage $\{N(t), t \geq 0\}$ est un processus de Poisson non homogène s'il satisfait les conditions suivantes :
(NH1) : N(0)=0 ;
(NH2) : le processus est à accroissements indépendants ;

(NH3) : N(t) suit une loi de Poisson de paramètre $\Lambda(t)$:

$$\forall k \in \mathbb{N}, \mathbb{P}[N(t) = k] = \frac{(\Lambda(t))^k}{k!} e^{-\Lambda(t)},$$

avec

$$\Lambda(t) = \int \lambda(t) dt,$$

et

$$\lambda(t) = \lim_{h \to 0^+} \frac{1 - \mathbb{P}[N(t+h) - N(t) = 0]}{h}, \quad \forall t \geq 0. \tag{1.34}$$

La fonction $\lambda(t)$ représente la fonction d'intensité de ce processus. C'est l'estimation de cette fonction qui fait l'objet de cette section.

Remarque 1.2. Lorsque cette fonction est constante, il s'agit simplement d'un processus de Poisson stationnaire.

La trajectoire d'un processus de Poisson non homogène est représentée dans la figure suivante :

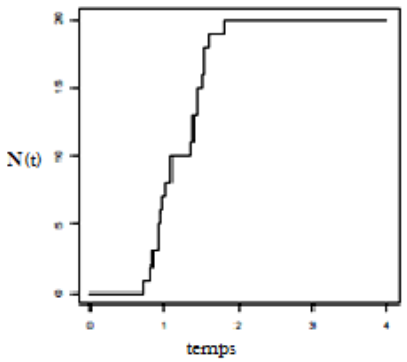

FIG. 1.4 – Trajectoire d'un processus de Poisson non homogène.

1.3.4 Le modèle de Cox et l'estimateur à noyau

On dispose de n points x_i sur un intervalle [0,T]. On veut déterminer l'intensité locale $\lambda(x)$ de ces points au voisinage de tout point x de [0,T]. Le modèle introduit consiste à supposer que ces points constituent une réalisation d'un processus de Cox stationnaire de fonction d'intensité $\{\Lambda(x), x \in \mathbb{R}\}$. Ce processus est également appelé "processus de Poisson doublement stochastique" (Cox et Isham 1980 [11]). Ainsi les deux conditions suivantes sont supposées être vérifiées :

– $\{\Lambda(x), x \in \mathbb{R}\}$ est un processus aléatoire réel stationnaire, à valeurs positives et de moyenne μ et on notera :

$$\mathbb{E}[\Lambda(x)\Lambda(y)] = \nu(|x - y|), \quad \forall (x, y) \in \mathbb{R}^2,$$

avec :

$$\nu(x) = \mu^2 \nu_0(x),$$

et $\nu_0(x)$ est une fonction fixée.

– conditionnellement à la réalisation $\lambda(x)$ de $\Lambda(x)$, les points x_i sont générés selon un processus de Poisson non homogène de fonction d'intensité $\lambda(x)$.

Diggle (1985) [16] a donné un estimateur non paramétrique pour la fonction d'intensité $\lambda(x)$ par la méthode du noyau :

$$\hat{\lambda}_h(x) = \frac{1}{h} \sum K(\frac{x - x_i}{h}), \quad x \in [0, T], \tag{1.35}$$

où h est le paramètre de lissage strictement positif.

1.3.5 Le modèle multiplicatif d'Aalen

Soient $x_1, x_2, ..., x_n$ n observations d'un processus de Poisson non homogène de fonction d'intensité $\lambda(x)$. Le modèle multiplicatif simple pour la fonction d'intensité est donné par l'équation suivante [5] :

$$\lambda_c(x) = c.\alpha(x), \quad x \in [0, T], \tag{1.36}$$

où c est une constante et $\alpha(x)$ est une fonction déterministe, non négative et inconnue ; avec :

$$\int_0^T \alpha(x)dx = 1.$$

Les n observations $x_1, x_2, ..., x_n$ ont la même fonction de densité que les statistiques d'ordre corréspondant à n variables aléatoires $i.i.d$ de densité $\alpha(x)$ sur $[0, T]$ (voir [5]).

Remarque 1.3. Dans le modèle multiplicatif, l'estimateur à noyau de l'intensité $\lambda_c(x)$ est le même que celui donné dans le cas du modèle de Cox [5] (donné par la formule (1.35)). Dans la pratique, cette égalité est d'une très grande utilité ; ce que nous allons exploiter dans notre étude de simulation.

1.3.6 Le choix du paramètre de lissage

Le comportement de l'estimateur de la fonction d'intensité donné par l'équation (1.35) est lié au choix du paramètre de lissage. Diggle (1985) [16] a proposé une méthode bayesienne pour le choix du ce paramètre. Elle consiste a utiliser le modèle de Cox défini dans la section précédente. Si on considère l'erreur quadratique moyenne (MSE) comme une mesure de la qualité de l'estimateur, on aura [15] :

$$\begin{aligned} MSE(h) &= E[(\hat{\lambda}_h(x) - \Lambda(x))^2] \\ &= \nu(0) + \frac{\mu}{2h}[1 - 2\mu\mathcal{K}(h)] + \frac{\mu}{2h}^2 \int_0^{2h} \mathcal{K}(y)dy, \end{aligned} \tag{1.37}$$

avec

$$\mathcal{K}(h) = 2\mu^{-2} \int_0^h \nu(x)dx.$$

Alors la meilleure largeur de la fenêtre h est définie de la manière suivante :

$$h_0 = \underset{h}{Argmin}\, MSE(h). \tag{1.38}$$

Le problème du choix de la largeur de la fenêtre revient donc à estimer h_0.
Dans le cas d'un noyau uniforme $K(.) = \frac{1}{2} I_{[-1,1]}(.)$ utilisé pour obtenir l'estimateur $\hat{\lambda}_h$, la fonction $\mathcal{K}(h)$ est estimée par [15] :

$$\hat{\mathcal{K}}(h) = Tn^{-2} \sum \sum_{i \neq j} I_{[-t,t]}(x_i - x_j). \tag{1.39}$$

D'où la largeur de la fenêtre h_0 qui minimise le MSE, est bien estimée par le valeur $\hat{h}_M$ qui minimise :

$$\hat{M}(h) = \frac{1}{2\hat{\mu}h} - \frac{1}{h}\hat{\mathcal{K}}(h) + \frac{1}{(2h)^2} \int_0^{2h} \hat{\mathcal{K}}(y)dy, \tag{1.40}$$

avec $\hat{\mu}$ est l'estimateur de μ : $\hat{\mu} = n/T$.

Remarque 1.4. Si on utilise le critère de l'erreur quadratique intégrée (ISE) pour évaluer la performance de l'estimateur à noyau de l'intensité, la méthode de la validation croisée est la plus adéquate pour le choix optimal de cette fenêtre.

La méthode de la validation croisée

Cette technique est surtout développée en estimation de régression et de densité, et elle a été adaptée récemment par Brooks et Marron (1989)[5] au problème d'estimation de l'intensité d'un processus de Poisson non homogène, en se fixant comme objectif la minimisation de l'erreur quadratique intégrée.

$$\begin{aligned} ISE(h) &= \int_0^T (\hat{\lambda}_h(x) - \lambda(x))^2 dx \\ &= \int_0^T \hat{\lambda}_h^2(x)dx - 2\int_0^T \hat{\lambda}_h(x)\lambda(x)dx + \int_0^T \lambda(x)^2 dx. \end{aligned} \tag{1.41}$$

Ces auteurs proposent d'estimer la fenêtre optimale

$$h_1 = \underset{h}{Argmin}\, ISE(h), \tag{1.42}$$

en utilisant le critère de la validation croisée donné par :

$$CV(h) = \int_0^T \hat{\lambda}_h^2(x)dx - 2\sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h} K(\frac{x_i - x_j}{h}). \tag{1.43}$$

L'estimateur de h_1, sélectionné par le critère de la validation croisée, est le suivant :

$$h_{CV} = \underset{h}{Argmin}\, CV(h). \tag{1.44}$$

De manière intuitive, ce critère se justifie par le fait que chacun de ses termes est un estimateur des deux premiers termes que l'on obtient en développant l'expression de ISE(h)

(le troisième terme de ISE(h) ne dépend pas de h). Ainsi, la quantité de la validation croisée $CV(h)$ est un estimateur des termes dépendant de h dans l'équation de $ISE(h)$. Plus précisément, des résultats d'optimalité asymptotique du paramètre sont obtenus par Brooks et Marron (1989)[5] pour un processus de Poisson non homogène. Par conséquent, la valeur de h qui minimise $CV(h)$, minimise aussi $ISE(h)$.

1.3.7 Effets du biais aux bornes

Nous avons déjà parler du problème du biais aux bornes dans le cas de l'estimation de la densité. L'effet du biais aux bornes est aussi un problème pour l'estimation de la fonction d'intensité vu que cette dernière est définie sur un support positif $[0, T]$. Ce problème peut être corrigé par la méthode de réflexion (noyau miroir) de Schuster (1985) [33], ou bien par la méthode de renormalisation locale donnée par Diggle (1985) [16]. Cette renormalisation est utilisée aux extrémités de l'intervalle $[0, T]$ où une correction des effets de bords est appliquée. Cet estimateur corrigé, introduit par Diggle (1985) [16], est donné par l'expression suivante :

$$\hat{\lambda}(x) = \frac{1}{hp_h(x)} \sum_{i=1}^{N} K(\frac{x - x_i}{h}); \tag{1.45}$$

où $p_h(x) = \int_0^T \frac{1}{h} K(\frac{x-u}{h}) du$.

1.4 Conclusion

Dans ce chapitre l'estimateur non paramétrique par la méthode du noyau a été exposé. La première section a été consacrée à la présentation de l'estimateur à noyau de la fonction de densité de probabilité. Cet estimateur dépend de deux paramètres : la fonction K appelée noyau et le paramètre de lissage h (ou la largeur de la fenêtre). Deux catégories de noyaux ont été présentées : noyaux symétriques et noyaux asymétriques. Quant au deuxième paramètre, nous avons exposé deux classes de méthodes pour son choix. Le choix judicieux de ces deux paramètres permet une bonne utilisation de cette méthode dans la pratique.

Dans la deuxième section, un estimateur à noyau pour la fonction d'intensité a été présenté. En suivant la démarche donnée par Diggle (1985) [16], nous avons utilisé le modèle de Cox pour trouver une méthode du choix du paramètre de lissage optimal auquel la qualité de l'estimateur est subordonnée.

Dans le chapitre suivant, nous révélons une certaine équivalence du choix du paramètre de lissage entre les deux estimateurs présentés dans ce chapitre (l'estimateur de la densité et celui de l'intensité).

2

Equivalence du choix de la fenêtre dans l'estimation de la densité et l'intensité

2.1 Introduction

Comme on l'a déjà vu dans le chapitre précédent, la méthode du noyau est intéressante dans l'estimation non paramétrique ; soit pour la fonction de densité ou celle d'intensité d'un processus de Poisson non homogène. Dans chaque cas, la largeur de la fenêtre, dit le paramètre de lissage, est crucial pour les performances de l'estimateur. Plusieurs méthodes pour le choix de ce paramètre ont été présentées.

Le choix de la fenêtre par validation croisée a été largement étudié pour l'estimation de la densité de probabilité. Quant à l'estimation de l'intensité, le choix de la fenêtre minimise l'estimateur de l'erreur quadratique moyenne sous la supposition que les données sont générées par un processus de Cox. Malgré que les deux méthodes sont motivées de manières différentes, il y a une équivalence entre le choix de cette fenêtre.

Dans ce chapitre, des justifications pour cette équivalence seront données. On donne aussi les profits que cette équivalence permet d'avoir pour chaque méthode dans le contexte de l'autre.

2.2 L'équivalence

Les estimateurs

Pour l'estimation de la densité, Rudemo (1982) [32] et Bowman (1984) [4] ont proposé de choisir le paramètre de lissage h par la méthode des moindres carrés de validation croisée. Donc, l'estimateur de la densité de probabilité est donné par :

$$f_{h_{LSCV}}(x) = \frac{1}{nh_{LSCV}} \sum_{i=1}^{n} K(\frac{x - x_i}{h_{LSCV}}), \tag{2.1}$$

où h_{LSCV} est l'estimateur de h qui minimise le critère de la validation croisée donnée par la formule (1.28).

Pour l'estimation de la fonction d'intensité, nous avons présenté dans le chapitre précédent la démarche suivie par Diggle (1985) [16] pour le choix du paramètre de lissage contrôlant l'estimateur à noyau de cette fonction qui est donné par la formule suivante :

$$\hat{\lambda}_{\hat{h}_M}(x) = \frac{1}{\hat{h}_M} \sum_{i=1}^{n} K(\frac{x - x_i}{\hat{h}_M}), \quad x \in [0, T]. \tag{2.2}$$

où $\hat{h}_M$ est l'estimateur de h qui minimise le critère MSE(h) donné par (1.37).

L'équivalence

Théorème 2.1. *Dans le cas d'un noyau uniforme :*

$$h_{LSCV} = \hat{h}_M.$$

ie : h_{LSCV} est la valeur du paramètre de lissage qui minimise LSCV(h), elle minimise aussi $\hat{M}(h)$. Ce qui est équivalent à dire qu'elle minimise MSE(h) défini par l'équation (1.37), de même pour la valeur $\hat{h}_M$ qui minimise $\hat{M}(h)$, elle minimise aussi LSCV(h).

La preuve de ce théorème découle du lemme suivant :

Lemme 2.1. *(Diggle et Marron (1988)[15])*
Pour un noyau uniforme

$$\hat{M}(h) = T.LSCV(h)$$

Pour une démonstration détaillée du lemme (2.1), voir Diggle et Marron (1988)[15].

2.3 Les profits de l'équivalence

Cette dualité entre les deux problèmes montre comment appliquer la théorie asymptotique bien développée pour l'estimation de la densité de probabilité dans le contexte de l'estimation de l'intensité mais aussi elle motive l'application de la méthode du processus de Cox dans l'estimation de la densité.

2.3.1 Les profits pour l'estimation de l'intensité

La motivation pour $\hat{h}_M$ est appliquée uniquement dans le cas du noyau uniforme. Cependant la motivation pour h_{LSCV} fonctionne pour tout noyau. Cela suggère d'utiliser le lemme (2.1) pour trouver $\hat{M}(h)$ approprié dans le cas d'un noyau quelconque.

Lemme 2.2. *(Diggle et Marron (1988)[15])*
Pour un noyau K quelconque :

$$\hat{M}(h) = T.LSCV(h) = \frac{1}{\hat{\mu}h} \int K^2 + \frac{1}{h}\hat{\mathcal{K}}_{K*K}(h) - \frac{2}{h}\hat{\mathcal{K}}_K(h),$$

où :

$$\hat{\mathcal{K}}_K(h) = \frac{T}{n^2}\sum\sum_{i\neq j} K(\frac{x_i - x_j}{h}),$$

et

$$K * K(.) = \int K(. - x)K(x)dx.$$

Soit $\hat{h}_n$ la valeur qui minimise l'estimateur quelconque $\hat{M}(h)$, alors le minimum de $\hat{\mu}\hat{M}(h) + \nu(0)$ (qui est aussi minimum de $\hat{M}(h)$) est minimum de MSE(h) suivant :

Lemme 2.3. *(Diggle et Marron (1988)[15])*

$$MSE(\mathrm{h}) = \frac{\mu}{h}\int K^2 + \frac{\mu^2}{h}\mathcal{K}_{K*K}(h) - \frac{2\mu^2}{h}\mathcal{K}_K(h) + \nu(0),$$

avec

$$\mathcal{K}_K(h) = \frac{2h}{\mu}\int_0^\infty K(u)\nu(u)du.$$

Analogue de la théorie asymptotique de l'estimateur de densité dans l'estimateur d'intensité :

- Densité : les asymptotiques avec $n \to \infty$ ajoutent de l'information partout.
- Intensité : les asymptotiques avec $T \to \infty$ ajoutent de l'information seulement à l'extrémité droite de [0,T].

Pour ajouter de l'information partout, on prend $\mu \to \infty$ (μ est la moyenne du processus de Cox $\Lambda(x)$).
Pour empêcher le changement de la forme de la courbe $\Lambda(x)$ dans le processus limite, on prend :

$$\nu(x) = \mu^2\nu_0(x),$$

où $\nu_0(x)$ est une fonction fixée. Pour la densité $MSE = variance + biais^2$, et pour l'intensité on regarde $\mu^{-2}MSE$, où μ^{-2} est l'ajustement entre les estimateurs f_h et $\hat{\lambda}_h$.

Lemme 2.4. *(Diggle et Marron (1988)[15])*

$$\begin{aligned}\mu^{-2}MSE &= Variance(h) + biais^2(h)\\ &= [\frac{1}{\mu h}\int K^2] + [\frac{1}{h}\hat{\mathcal{K}}_{K*K}(h) - \frac{2}{h}\hat{\mathcal{K}}_K(h) + \nu(0)].\end{aligned}$$

Lemme 2.5. *(Diggle et Marron (1988)[15])*
Comportement de $biais^2(h)$:
Si $\nu_0(x)$ admet une $4^{ème}$ dérivée continue à l'origine, alors :
quand $t \to 0$

$$biais^2(h) = h^4\nu_0^4(0)[\frac{1}{2}\int u^2K(u)du]^2 + o(h^4).$$

La différence entre $biais^2(h)$ de la densité et $biais^2(h)$ de l'intensité est que $\nu_0^4(0)$ remplace $(f''(x))^2$.
Les démonstrations des lemmes (2.4) et (2.5) sont détaillées dans l'article de Diggle et Marron (1988) [15].

2.3.2 Les profits pour l'estimation de la densité

L'équivalence précédente motive la considération du problème de l'estimation de la densité du point de vue de la méthode du processus de Cox. Si $x_1, x_2, ..., x_n$ sont n observations de la variable aléatoire X de densité $\Lambda(x)$ sur [0,T], avec :

$$\mathbb{E}[\Lambda(x)] = \frac{1}{T},$$

et

$$\mathbb{E}[\Lambda(x)\Lambda(y)] = \nu(|x-y|),$$

une expression de l'erreur quadratique moyenne $MSE^*(h)$ de l'estimateur $f_h(x)$ est donné par le lemme suivant :

Lemme 2.6. *(Diggle et Marron (1988)[15])*

$$MSE^*(h) = \frac{1}{nhT}\int K^2 + \frac{(1-\frac{1}{n})}{hT^2}\mathcal{K}^*_{K*K}(h) - \frac{2}{hT^2}\mathcal{K}^*_K(h) + \nu(0),$$

où

$$\mathcal{K}^*_K(h) = 2T^2\int_0^\infty K(u)\nu(u)du.$$

et μ est identifié avec $\frac{1}{T}$.
La différence entre le lemme 2.3 et le lemme 2.6 est que n apparait dans le premier terme et $(1-1/n)$ dans le deuxième terme.
$\mathcal{K}^*_K(h)$ est estimé par [15] :

$$\hat{\mathcal{K}}^*_K(h) = \frac{1}{n(n-1)}hT\sum\sum_{i\neq j}K(\frac{x_i - x_j}{h}).$$

Le minimum $\hat{h}^*_M$ de :

$$\hat{M}^*(h) = \frac{1}{nh}T\int K^2 + \frac{(1-\frac{1}{n})}{h}\hat{\mathcal{K}}^*_{K*K}(h) - \frac{2}{h}\hat{\mathcal{K}}^*_K(h),$$

est aussi minimum de MSE^*.

2.4 Conclusion

Dans ce chapitre, on a présenté l'équivalence du choix de la fenêtre qui existe entre les estimateurs à noyaux des fonctions de densité et d'intensité. Nous avons donné des justifications ainsi que les profits tirés de cette équivalence.

3

Simulation

3.1 Introduction

Nous présentons dans ce chapitre une étude de simulation effectuée à l'aide du logiciel R, pour essayer d'illustrer les différents aspects théoriques abordés dans les chapitres précédents. Cette illustration numérique nous servira à :

- Voir les résultats de l'estimation de la densité pour différentes méthodes de sélection du paramètre de lissage.
- Estimer la fonction d'intensité par un noyau gamma avec un paramètre de lissage choisi par la méthode proposée dans la section (1.3.6).
- Etablir l'équivalence du choix de la fenêtre qui existe entre les deux estimateurs.

L'outil statistique R est un logiciel d'analyse statistique et graphique créé par Robert Gentleman et Ross Ihaka [41]. Ce logiciel est équipé de nombreux packages qui offrent une grande variété de procédures mathématiques et statistiques, incluant les méthodes économétriques complexes, le traitement des séries chronologiques et l'analyse des données.

3.2 Plan de simulation

Premièrement, pour l'estimation de la densité nous avons choisi de simuler les densités de probabilité de deux types de lois, la loi normale qui a une fonction de densité (**D1**) définie sur $\mathbb{R}$ et la deuxième est la loi exponentielle (de paramètre 1) dont la densité (**D2**) est définie sur un support positif $[0, +\infty[$. Les deux fonctions de densité sont définies comme suit :

D1. La loi normale N(0,1) :

$$x \mapsto \frac{1}{\sqrt{2\pi}} exp(\frac{-x^2}{2}), x \in]-\infty, +\infty[.$$

D2. La loi exponentielle de paramètre $\lambda > 0$:

$$x \mapsto \lambda exp(-\lambda x), x \in [0, +\infty[.$$

Pour $\lambda = 1 : x \mapsto exp(-x), x \in [0, +\infty[$.

Nous utilisons pour la simulation des échantillons de taille 100, 500 et 1000 pour chaque densité. Notons aussi que dans la programmation, nous avons utilisé le noyau gaussien et le noyau gamma pour la correction des effets de bords dans l'estimation de la loi exponentielle. Pour le choix du paramètre de lissage, nous avons utilisé trois méthodes parmi celles présentées dans le premier chapitre : la validation croisée par moindres carrés, la validation croisée biaisé et la méthode de ré-injection (plug-in).

Deuxièmement, pour modéliser un processus de Poisson non homogène de fonction d'intensité connue, nous avons choisi le modèle d'Aalen pour la fonction d'intensité. Nous avons généré N points y_i de densité α connue, on utilise deux fonction de densité : $\alpha_1 = exp(1)$ et $\alpha_2 = gamma(2, 5)$.
Les N points ordonnés constituent donc une réalisation d'un processus de poisson de fonction d'intensité λ telle que :

$$\lambda = N\alpha.$$

Pour l'estimation de la fonction d'intensité λ, on utilise l'estimateur à noyau donné par la formule (1.35) et celle donnée par Diggle [16] en prenant compte des effets de bord (1.45). Pour le choix du paramètre de lissage, nous avons utilisé la méthode de la validation croisée donnée par Brooks et Marron [5].

3.3 Résultats de la simulation

Les résultats de la simulation sont donnés sous forme de tableaux et de graphiques.

3.3.1 L'estimation d'une densité par la méthode du noyau

On commence par les résultats concernant l'estimation de la densité de probabilité ; les tableaux (3.1) et (3.2) résument les valeurs du paramètre de lissage par trois méthodes pour chacune des lois simulées, où :

- h_{BCV} : représente la valeur de h qui minimise le critère de la validation croisée biaisée.
- h_{LSCV} : représente la valeur de h qui minimise le critère de la validation croisée par les moindres carrés.
- $h_{plug-in}$: représente la valeur de h choisie par la méthode de ré-injection (plug-in).
- h_{AMISE} : représente la valeur de h théorique calculée par l'équation (1.9).

Les figures (3.1) à (3.6) représentent les graphiques des estimateurs à noyau gaussien obtenus pour chacune des lois simulées en utilisant les valeurs de h données dans les tableaux (3.1) et (3.2).

a- Cas de la loi normale $\mathcal{N}(0,1)$:

n \ h	h_{BCV}	h_{LSCV}	$h_{plug-in}$	h_{AMISE}
100	1.52817	1.52903	1.29829	1.51272
500	1.22802	1.25748	1.21273	1.23229
1000	1.09346	1.14429	1.10139	1.11397

TAB. 3.1 – Les valeurs de la fenêtre pour les échantillons simulés à partir de la loi $\mathcal{N}(0,1)$.

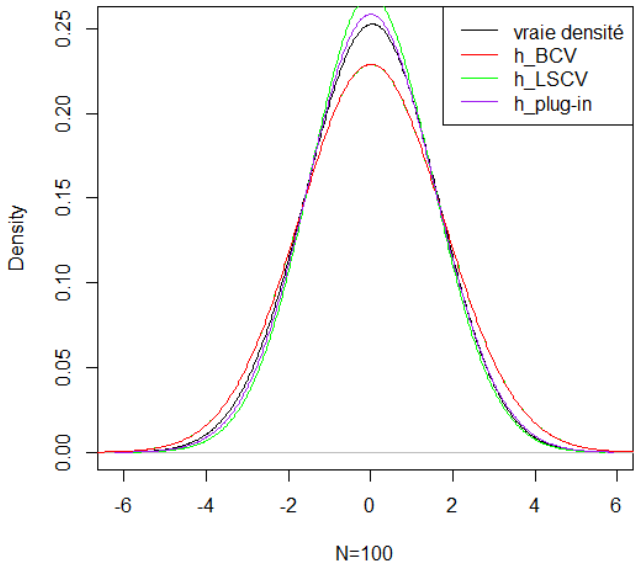

FIG. 3.1 – Comparaison entre la densité théorique $\mathcal{N}(0,1)$ et celle estimée par un noyau gaussien avec les paramètres de lissage h_{BCV}, h_{LSCV} et $h_{plug-in}$ pour N=100.

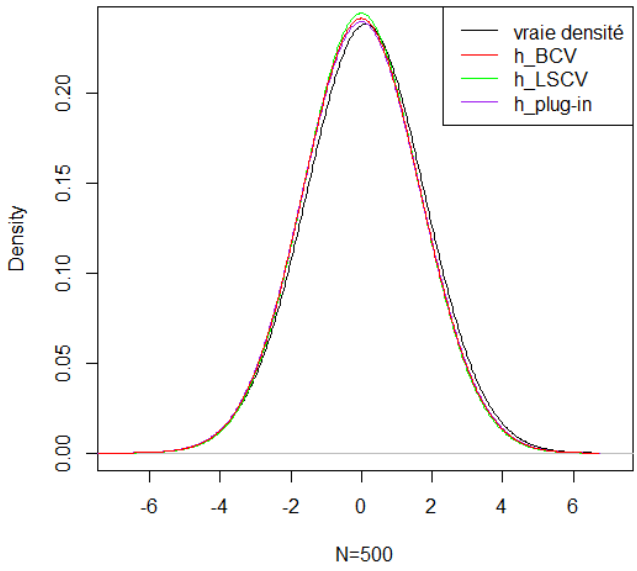

FIG. 3.2 – Comparaison entre la densité théorique $\mathcal{N}(0,1)$ et celle estimée par un noyau gaussien avec les paramètres de lissage h_{BCV}, h_{LSCV} et $h_{plug-in}$ pour N=500.

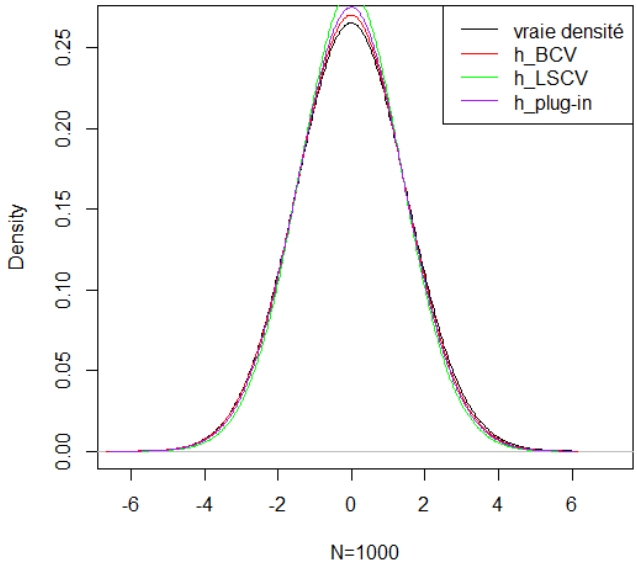

FIG. 3.3 – Comparaison entre la densité théorique $\mathcal{N}(0,1)$ et celle estimée par un noyau gaussien avec les paramètres de lissage h_{BCV}, h_{LSCV} et $h_{plug-in}$ pour N=1000.

a- Cas de la loi exponentielle $exp(1)$ **:**

n \ h	h_{BCV}	h_{LSCV}	$h_{plug-in}$	h_{AMISE}
100	0.56962	0.62865	0.61249	0.61833
500	0.45722	0.280502	0.363276	0.3592
1000	0.42378	0.33165	0.3476	0.3197

TAB. 3.2 – Les valeurs de la fenêtre pour les échantillons simulés à partir de la loi $exp(1)$ pour un noyau gaussien.

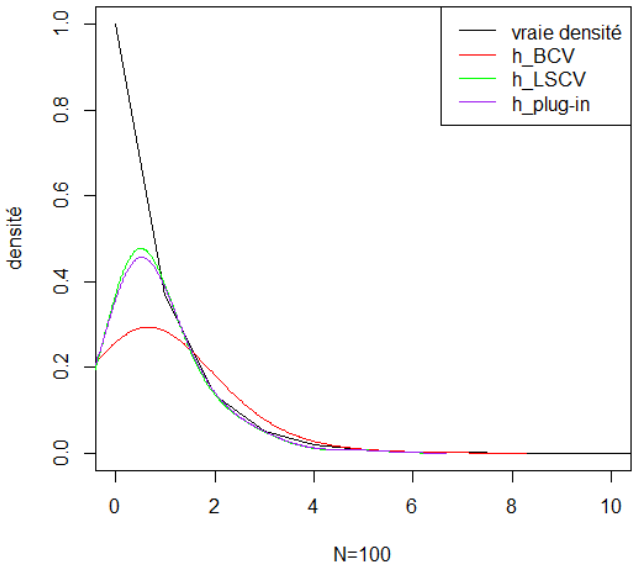

FIG. 3.4 – Comparaison entre la densité théorique $exp(1)$ et celle estimée par un noyau gaussien avec les paramètres de lissage h_{BCV}, h_{LSCV} et $h_{plug-in}$ pour N=100.

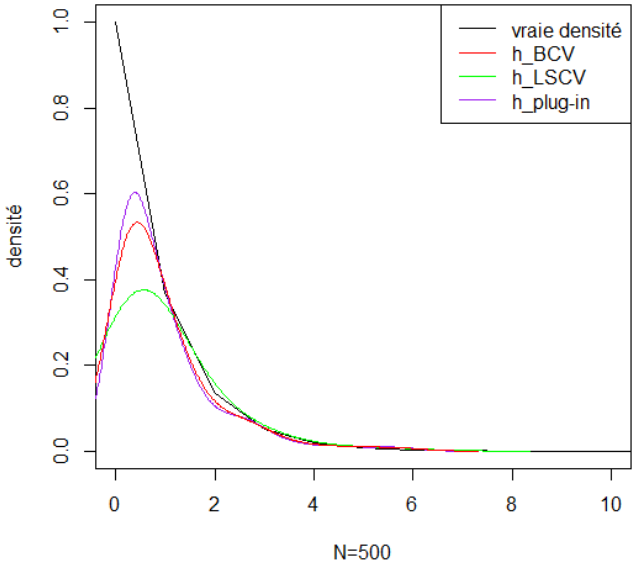

FIG. 3.5 – Comparaison entre la densité théorique $exp(1)$ et celle estimée par un noyau gaussien avec les paramètres de lissage h_{BCV}, h_{LSCV} et $h_{plug-in}$ pour N=500.

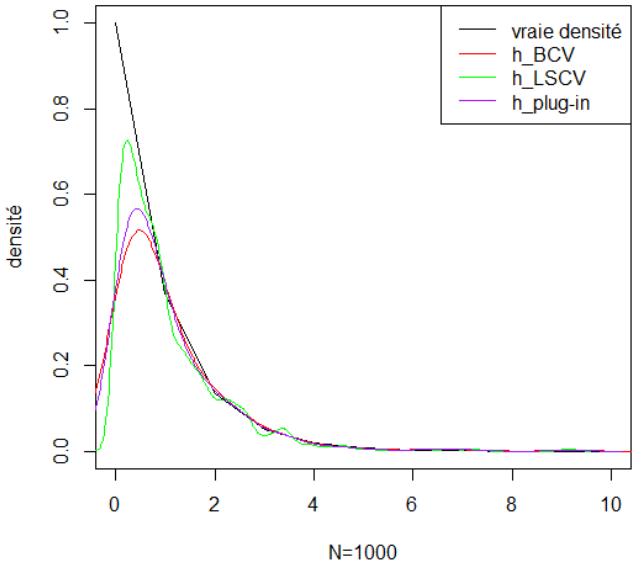

FIG. 3.6 – Comparaison entre la densité théorique $exp(1)$ et celle estimée par un noyau gaussien avec les paramètres de lissage h_{BCV}, h_{LSCV} et $h_{plug-in}$ pour N=1000.

Discussion

◇ Les résultats de l'estimation de la densité **D1** (d'une loi normale) montrent que le noyau gaussien est très performant. Les données du tableau (3.1) indiquent que les valeurs de h_{BCV}, h_{LSCV} et $h_{plug-in}$ sont très proches de la valeur de h_{AMISE} pour les différentes tailles de l'échantillon, c'est-à-dire que les trois méthodes de sélection du paramètre h estiment correctement sa valeur.
Ceci est confirmé par les graphes données par les figures (3.1), (3.2) et (3.3) où on peut voir que les courbes des estimateurs, pour les différents échantillons, sont très proches de celle de la vraie densité simulée **D1**.

◇ Contrairement au premier cas, les résultats de l'estimation de la densité **D2** ($exp(1)$) par un noyau gaussien sont de moindre qualité et les valeurs du paramètre de lissage données dans le tableau (3.2) indiquent que h_{BCV}, h_{LSCV} et $h_{plug-in}$ ne s'approchent pas de la valeur de h_{AMISE} ce qui conduit à une mauvaise estimation, particulièrement au voisinage de zero ; les graphes (3.4), (3.5) et (3.6) le confirment.

Cette mauvaise qualité de l'estimation par un noyau gaussien, pour une telle densité (à support positif), est dûe au biais apporté par les noyaux symétriques sur ce type de densité. Pour résoudre ce problème, nous faisons appel aux noyaux asymétriques. Le noyau gamma décrit dans la section (1.2.5) est le plus adéquat pour notre cas. En reprenant les mêmes étapes de l'estimation précédente, et en utilisant le noyau gamma, nous avons réestimé la densité **D2**.
Les résultats de cette estimation sont résumés dans le tableau (3.3) et représentés par les figures (3.7) à (3.9).

n \ h	h_{BCV}	h_{LSCV}	$h_{plug-in}$	h_{AMISE}
100	1.33204	0.80069	0.76623	0.71836
500	0.5817	0.23429	0.44608	0.29972
1000	0.34717	0.14876	0.3128	0.20197

TAB. 3.3 – Les valeurs de la fenêtre pour les échantillons simulés à partir de la loi $exp(1)$. Cas d'un noyau gamma.

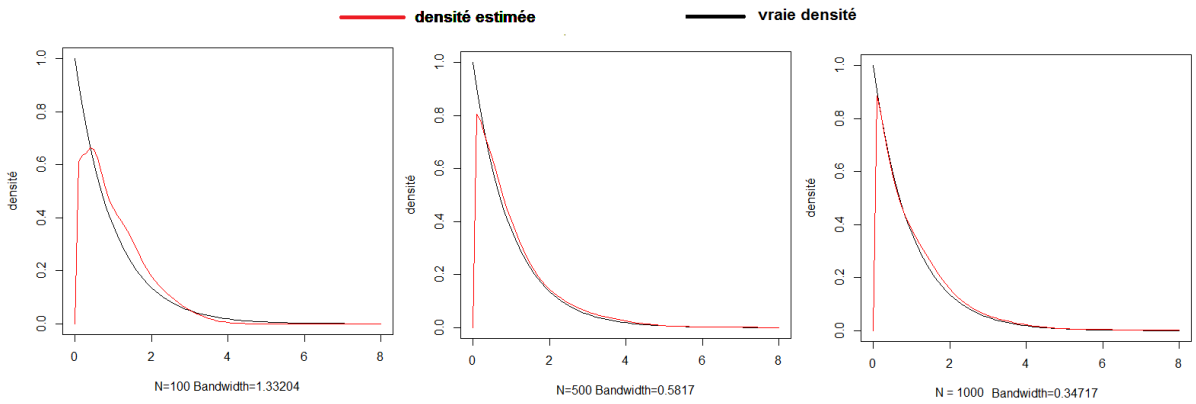

FIG. 3.7 – Comparaison entre la densité théorique $exp(1)$ et celle estimée par un noyau gamma avec le paramètre de lissage h_{BCV}.

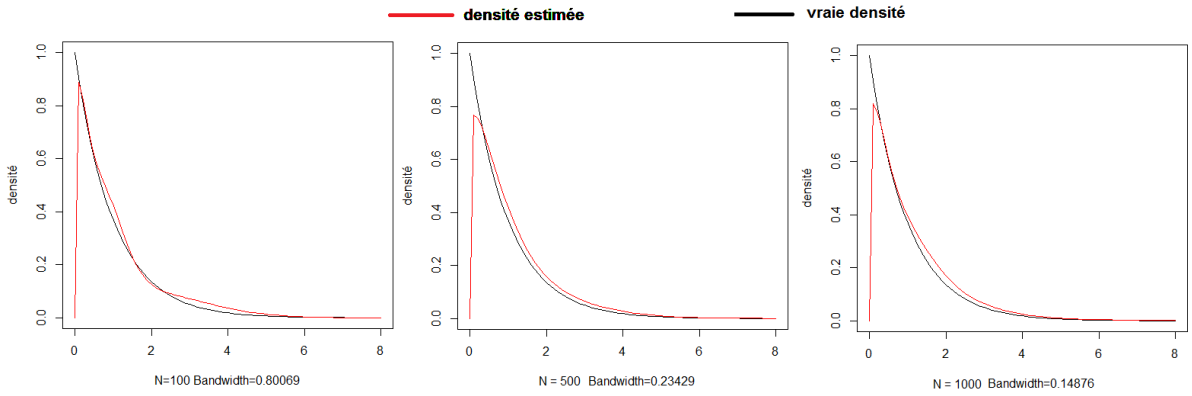

FIG. 3.8 – Comparaison entre la densité théorique $exp(1)$ et celle estimée par un noyau gamma avec le paramètre de lissage h_{LSCV}.

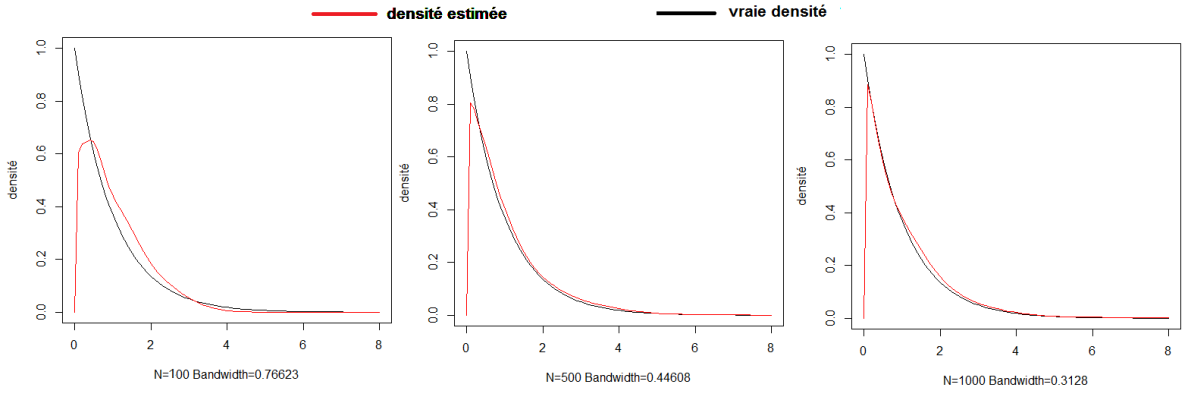

FIG. 3.9 – Comparaison entre la densité théorique $exp(1)$ et celle estimée par un noyau gamma avec le paramètre de lissage $h_{plug-in}$.

◇ Les graphes (3.7), (3.8) et (3.9) montrent que l'utilisation du noyau gamma a donné des résultats meilleurs pour l'estimation de la densité $exp(1)$. Ce résultat confirme l'efficacité des noyaux asymétriques pour la correction des effets du biais aux bornes.

3.3.2 L'estimation d'une intensité par la méthode du noyau

Les résultats de l'estimation de la fonction d'intensité sont comme suit :
Les figures (3.10) et (3.11) représentent les trajectoires des processus de Poisson de fonction d'intensité donnée par la formule :

$$\lambda_i = N\alpha_i,$$

avec $\alpha_1 = exp(1)$, $\alpha_2 = gamma(2, 5)$ et N=20.

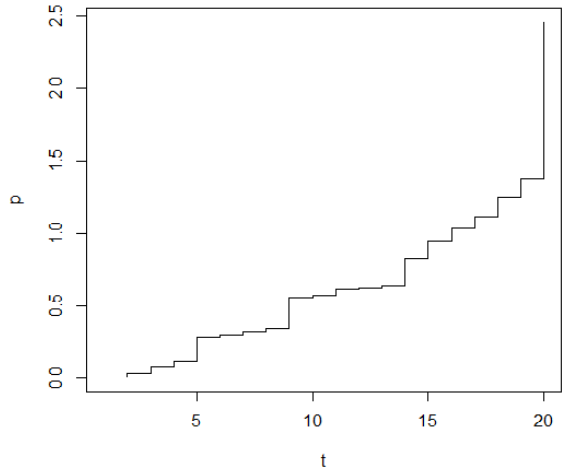

FIG. 3.10 – La trajectoire du processus de Poisson non homogène de fonction d'intensité λ_1.

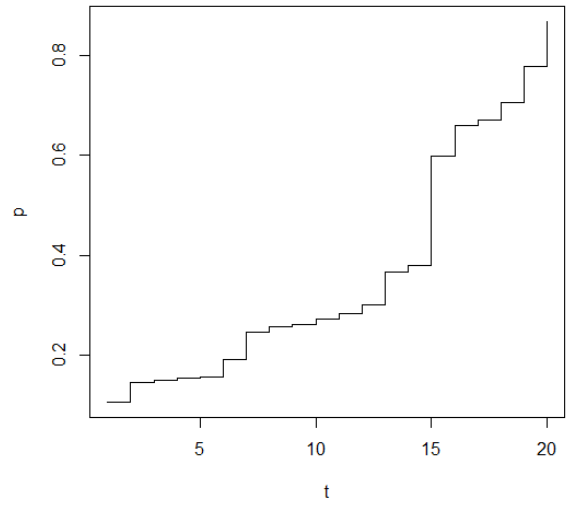

FIG. 3.11 – La trajectoire du processus de Poisson non homogène de fonction d'intensité λ_2.

Les représentations graphiques des estimateurs à noyau gaussien, gamma et par la formule de renormalisation de Diggle (1.45) pour les fonctions d'intensité λ_1 et λ_2 sont données par les figures (3.12) et (3.13) avec les largeurs de la fenêtre données dans le tableau suivant :

Intensité	h_{CV} (noyau gaussien)	h_{CV} (noyau gamma)	h_{CV} (Diggle)
λ_1	1.0631	0.9661	0.9334
λ_2	0.8916	0.5075	0.5213

TAB. 3.4 – Les valeurs de la fenêtre intervenant dans les estimateurs à noyau des fonctions d'intensité.

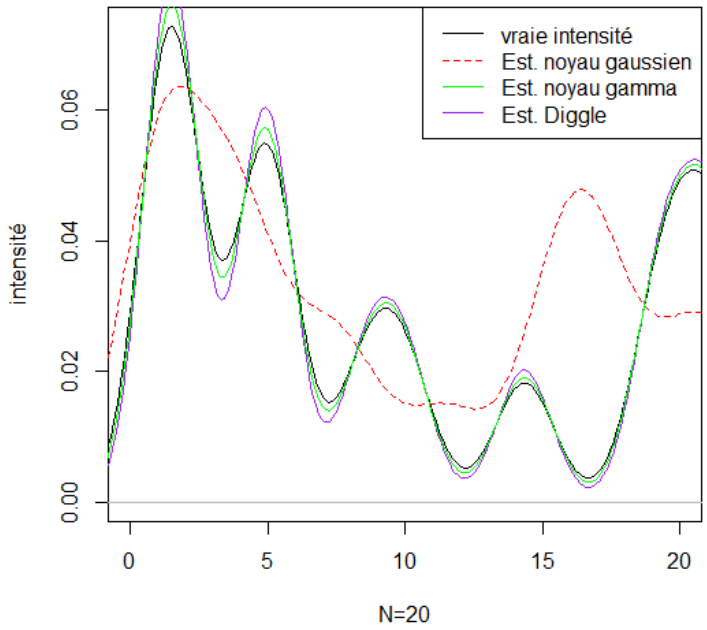

FIG. 3.12 – Comparaison des graphes des estimateurs à noyau avec celui de la vraie fonction d'intensité λ_1.

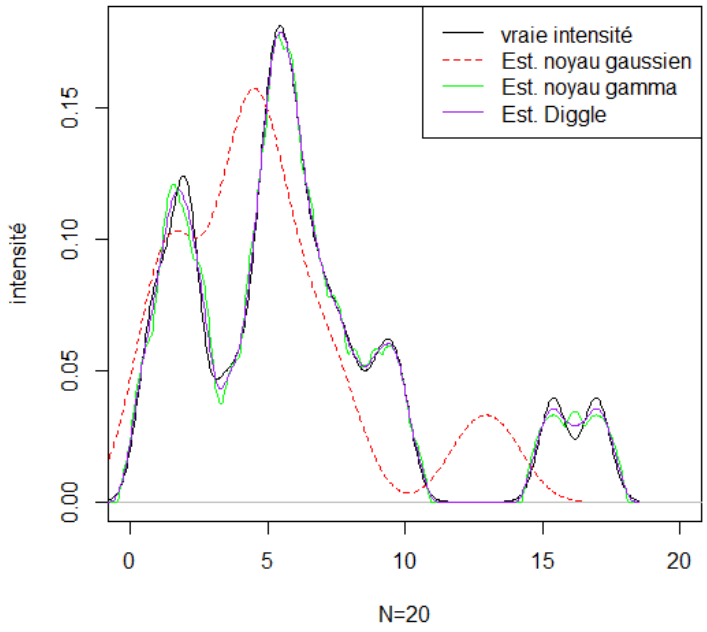

FIG. 3.13 – Comparaison des graphes des estimateurs à noyau avec celui de la vraie fonction d'intensité λ_2.

Discussion

D'après les graphes des figures (3.12) et (3.13), on remarque qu'avec un noyau gaussien (le trait discontinu) on a pas une bonne estimation des fonctions d'intensité ; ceci est à cause des effets du biais aux bornes. Pour cela, nous avons utilisé deux méthode de correction de ces effets indésirables ; la méthode de renormalisation de Diggle et un noyau asymétrique de type gamma.

Les représentations graphiques des estimateurs par un noyau gamma (le trait vert) sont très proches de ceux des vraies intensités λ_1 et λ_2. Le résultat est obtenu en utilisant la formule (1.45) donnée par Diggle [16]. Cette bonne qualité de lissage montre aussi que la méthode de sélection de h introduite par Brooks et Marron [5] donne les mêmes résultats asymptotiques que celle utilisée par Diggle er Marron [15] pour établir l'équivalence.

L'équivalence :

Le tableau (3.5) résume les valeurs de la fenêtre h intervenant dans l'estimation par un noyau gamma des intensités λ_1 et λ_2 et par la renormalisation de Diggle [16] en comparaison avec celles intervenant dans l'estimation des densités α_1 et α_2 par un noyau gamma.

Densité	h_{LSCV}	Intensité	h_{CV} (noyau gamma)	h_{CV} (Diggle)
α_1	0.9661	λ_1	0.9611	0.9334
α_2	0.5075	λ_2	0.5035	0.5213

TAB. 3.5 – Comparaison des valeurs de la fenêtre.

À partir du tableau (3.5), on remarque que h_{LSCV} et h_{CV} (pour les deux méthodes d'estimation de l'intensité) ont des valeurs très proches dans les deux cas (λ_1 et λ_2). Ceci montre bien évidemment qu'il existe une équivalence du choix de h dans l'estimation de l'intensité λ_1 et la densité α_1, de même pour λ_2 et α_2.

3.4 Conclusion

Dans cette étude de simulation, nous avons appliqué la méthode du noyau pour l'estimation des fonction de densité et d'intensité pour arriver à concrétiser l'équivalence du choix de la fenêtre entre le deux estimateurs. Dans une première étape, nous avons estimer deux types de fonctions de densité par un noyau gaussien ; une densité $\mathcal{N}(0,1)$ définie sur $\mathbb{R}$ et une $exp(1)$ définie sur un support positif ($[0,+\infty[$). Cette étape a permis de voir que le noyau gaussien est très performant dans le cas d'une densité définie sur $\mathbb{R}$, ce qui n'est pas le cas pour celle définie sur $[0,+\infty[$ vu les effets indésirables du biais aux bornes, ceci est à cause de l'utilisation d'un noyau symétrique. Ce résultat rend l'application d'une correction des effets du biais aux bornes d'une première priorité. De meilleurs résultats ont été obtenus après l'application du noyau asymétrique de type gamma dans l'estimation de **D2**.

Dans la deuxième étape de notre simulation, nous avons généré deux processus de Poisson non homogènes. Le premier a comme fonction d'intensité λ_1 simulée à partir d'une loi $exp(1)$, et le deuxième est de fonction d'intensité λ_2 simulée à partir d'une loi $gamma(2,5)$. La simulation a été faite pour $N = 20$ sur un intervalle $[0,20]$ en utilisant de la notion des statistiques d'ordre introduite dans le cas du modèle multiplicatif de l'intensité. Une estimation de ces deux intensités par la méthode du noyau gamma a donné des résultats d'une bonne qualité, ce que nous pouvons voir sur les graphes (3.12) et (3.13).

Les résultats de la simulation montrent le problème bien connu du choix de la fenêtre. En effet, la performance de la procédure de sélection du paramètre de lissage varie selon la taille de l'échantillon et la fonction densité à estimer. Les trois méthodes appliquées dans cette simulation montrent qu'elles estiment toutes le paramètre de lissage correctement dans le cas de la loi normale. Cependant, dans le cas de la loi exponentielle une estimation optimale est obtenue pour un paramètre de lissage choisi par la méthode de la validation croisée des moindres carrés (LSCV) pour un noyau gamma. Ceci est le même cas dans l'estimation des densités α_1 et α_2.

Pour l'estimation des fonctions d'intensité λ_1 et λ_2 nous avons appliqué la méthode de la validation croisée introduite par Brooks et Marron [5].

Les résultats obtenus , dans l'estimation de ces fonctions, montrent que h_{LSCV} est très proche de h_{CV}. Ce résultat est obtenu pour deux fonction de densité (α_1 et α_2) en comparaison avec les fonction d'intensité λ_1 et λ_2. C'est-à-dire que l'équivalence présenté théoriquement dans le chapitre précédent est bien illustrée par cette similarité.

Conclusion générale

Ce travail est une contribution au problème du choix du paramètre de lissage dans l'estimation non paramétrique par la méthode du noyau. Ce problème n'est pas seulement posé dans le cas de l'estimation d'une densité de probabilité, mais aussi dans le cas de l'estimation de l'intensité d'un processus stochastique. L'objectif souligné est donc d'établir une équivalence du choix de ce paramètre dans les deux cas.

Dans un premier chapitre, l'estimateur non paramétrique par la méthode du noyau a été exposé. Une section a été consacrée pour la présentation de l'estimateur à noyau de Parzen-Rosenblatt d'une densité de probabilité et ses différentes propriétés statistiques. Deux classes de méthodes pour le choix du paramètre de lissage ont été exposées ; et nous avons aussi présenté les différents noyaux : symétriques et asymétriques. Dans une autre section, un estimateur à noyau pour la fonction d'intensité a été présenté. En suivant la démarche donnée par Diggle (1985) [16], nous avons utilisé le modèle de Cox pour trouver une méthode du choix du paramètre de lissage optimal auquel la qualité de l'estimateur est subordonnée.

Dans le deuxième chapitre, nous avons présenté l'équivalence du choix de la fenêtre qui existe entre les deux problème de l'estimation à noyau. Les profits tirés de cette équivalence pour l'estimation de la densité et de l'intensité ont été présentés.

Le troisième chapitre a été consacré pour une étude de simulation. Dans cette dernière, nous avons appliqué la méthode du noyau pour l'estimation des fonction de densité et d'intensité pour arriver à concrétiser l'équivalence du choix de la fenêtre. Dans une première étape, nous avons estimer deux types de fonctions de densité par un noyau gaussien ; une densité $\mathcal{N}(0,1)$ définie sur $\mathbb{R}$ et une $exp(1)$ définie sur un support positif ($[0,+\infty[$). L'examen des résultats de cette étape, nous a conduit à constater que l'effet du biais aux bornes, dû à l'usage d'un noyau symétrique, rend l'estimateur à noyau d'une mauvaise qualité, ceci veut dire que l'application d'une correction des effets du biais aux bornes est d'une première priorité. Dans cette étude, la correction a été faite par un noyau asymétrique de type gamma.

Dans la deuxième étape de notre simulation, nous avons généré deux processus de Poisson non homogènes. Le premier a comme fonction d'intensité λ_1 simulée à partir de la densité d'une loi $exp(1)$, et le deuxième est de fonction d'intensité λ_2 simulée à partir de la densité d'une loi $gamma(2, 5)$. La simulation a été faite pour $N = 20$ sur un intervalle $[0, 20]$ en utilisant de la notion des statistiques d'ordre introduite dans le cas du modèle multiplicatif de l'intensité. Une estimation de ces deux intensités par la méthode du noyau a donné des résultats de bonne qualité, ce que nous avons vu voir sur les graphes (3.12) et (3.13).

Les résultats de la simulation montrent le problème bien connu du choix de la fenêtre. En effet, la performance de la procédure de sélection du paramètre de lissage varie selon la taille de l'échantillon et la fonction densité à estimer. Les trois méthodes appliquées dans cette simulation montrent qu'elles estiment toutes le paramètre de lissage correctement dans le cas de la loi normale. Cependant, dans le cas de la loi exponentielle une estimation optimale est obtenue pour un paramètre de lissage choisi par la méthode de la validation croisée des moindres carrés (LSCV) pour un noyau gamma. Ceci est le même cas dans l'estimation des densités α_1 et α_2.

Pour l'estimation des fonctions d'intensité λ_1 et λ_2 nous avons appliqué la méthode de la validation croisée introduite par Brooks et Marron [5]. Les résultats obtenus, dans l'estimation de ces fonctions, montrent que h_{LSCV} est très proche de h_{CV}. Ce résultat est obtenu pour deux fonctions de densité (α_1 et α_2) en comparaison avec les fonction d'intensité λ_1 et λ_2. Cette grande similarité des deux valeurs (h_{LSCV} et h_{CV}) a permis d'illustrer la notion d'équivalence abordée théoriquement dans le deuxième chapitre.

Perspectives :

Parmi les perspectives de ce travail, nous pouvons dégager deux points intéressants :
-Il serait intéressant d'appliquer cette étude sur des données réelles.
-Il est intéressant aussi d'approfondir et d'utiliser concrètement les profits de l'équivalence de chacune des fonctions de densité et d'intensité dans le contexte de l'autre.
-Essayer de développer d'autres résultats théoriques et pratiques en utilisant d'autres méthodes de sélection du paramètre de lissage dans le cas de l'estimation d'une fonction d'intensité.

Bibliographie

[1] O. Aalen. *Nonparametric inference for a family of counting processes. The Annals of Statistics, (6) : 701-726, 1978.*

[2] T. Bouezmarni, O. Scaillet. *Consistency of asymetric kernel density estimators and smoothed histograms with application to income data. Econometric Theory, (21) : 390-412, 2003.*

[3] T. Bouezmarni, J.V.K. Ronbouts. *Nonparametric density estimation for positive time series. Econometric Theory, 2006.*

[4] A.W. Bowman. *An alternative method of cross-validation for the smoothing of density estimates. Biometrika, (71) : 353-360, 1984.*

[5] M.M. Brooks, J.S. Marron. *Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions, Departement of Statistics, University of North Carolina, 1989.*

[6] B.M. Brown, S. Chen. *Beta-bernstein smoothing for regression curves with compact supports. Scandinavian Journal of Statistics, (26) : 47-59, 1999.*

[7] A. Charpentier, J.D. Fermanian and O. Scaillet. *The estimation of copulas : theory and practice. Ensae-Crest and Katholieke Universiteit Leuven, BNP-Paribas and Crest, HEC Genève and Swiss Finance Institute, 2006.*

[8] S. Chen. *A beta kernel estimation for density functions. Computational Statistics and Data Analysis, (31) : 131-145, 1999.*

[9] S. Chen. *Beta kernel for regression curve. Statistica Sinica, (10) : 73-92, 2000.*

[10] V. Couallier, P. Sarda, P. Vieu. *Estimation non paramétrique de discontinuités d'une fonction d'intensité. Rev. Stat. App., (3) : 89-106, 1997.*

[11] D.R. Cox, V. Isham. *Point processes. Chapman and Hall, London, 1980.*

[12] P. Deheuvels. *Estimation non paramétrique de la densité par histogrammes généralisés. Revue de Statistique Appliqué, (25) : 5-42, 1977.*

[13] L. Devroye, A. Berlinet. *Estimation d'une densité : un point sur la méthode du noyau. Statistique et Analyse des Données, (14) : 1-32, 1989.*

[14] L. Devroy, L. Gyorfi. *Nonparametric density estimation : The L1 View. John Wiley and Sons, New York, 1984.*

[15] P. Diggle, J.S. Marron. *Equivalence of smoothing parameters selectors in density and intensity estimation. Journal of the American Statistical Association, 1988.*

[16] P. Diggle. *A kernel method for smoothing point process data. Applied Statistics, (34), 138-147, 1985.*

[17] E. Doucet. *Estimateurs à noyau et théorie des valeurs extrêmes : comparaison de leur pouvoir prédictif dans l'analyse du coût des réclamations en assurance automobile. Mémoire de maitrise en mathématiques, univ de Quebec, 2014.*

[18] V.A. Epanechnikov. *Nonparametric estimation of a multidimensional probability density. Theory Probab. Appl., (14) : 153-158, 1969.*

[19] J.J. Faraway, M. Jhung. *Bootstrap choice of bandwidth for density estimation. Journal of the American Statistical Association, (85) : 1119-1122, 1990.*

[20] M. Fernandez, P. Monteiro. *Central limit theorem for asymetric kernel functionals. Annals of the Institute of Statistical Mathematics, (57) : 425-442, 2005.*

[21] E. Fix, J.R. Hodges. *Discriminatory analysis, nonparametric discrimination : consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.*

[22] W. Härdle. *Applied nonparametric regression. Combridge University Press, UK, 1990.*

[23] F.E. Harrell, C.E. Davis. *A new distribution-free quantile estimator. Biometrika, (69) : 635-640, 1982.*

[24] S. Jean. *Reconnaissance d'objets astronomiques HI par déconvolution et par estimation de densité. Thèse de PhD, Math. et Stat., Univ Laval, 1997.*

[25] M.C Jones, J.S Marron and S.J Sheather. *A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association, 1996.*

[26] M.R Leadbetter, D. Wold. *On estimation of point process intensities. Contributions to Statistics : Essays in honor of Norman L. Johnson (P.K. Sen., ed.). North-Holland, Amsterdam, 299-312, 1983.*

[27] S. Moncaup, P.Sarda, P. Vieu. *Une mise en œuvre d'estimation non paramétrique d'une fonction d'intensité sur des données météorologiques. Rev. Stat. App., (4) : 77-88, 1995.*

[28] E. Parzen. *On estimation of a probability density function and mode. Ann. Math. Stat., (33) : 1065-1076, 1962.*

[29] Ramlau-Hansen, Henrik. *Smoothing counting process intensities by means of kernel functions. The Annals of Statistics, (11) : 453-466, 1983.*

[30] Ramlau-Hansen, Henrik. *The choice of a kernel function in the graduation of counting process intensities. Scandinavian Actuarial Journal, 165-182, 1983.*

[31] M. Rosenblatt. *Remarks on some nonparametric estimates of a density function. Ann. Math. Stat., (27) : 832-837, 1956.*

[32] M. Rudemo. *Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics, (9) : 65-78, 1982.*

[33] E.F. Schuster. *Incorporating support constraints into nonparametric estimators of densities. Communications in statistics-Theory and Methodes, (14), 1123-1136, 1985.*

[34] D.W. Scott, G.R. Terell. *Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association, (82) : 1131-1146, 1987.*

[35] D.W. Scott, R.A. Tapia and J.R. Thomson. *Kernel density estimation revisited. Nonlinear Analysis, Theory , Methods and Applications, (1) : 339-372, 1977.*

[36] B.W. Silverman. *Density estimation for statistics and data analysis. Chapman and Hall, London, 1986.*

[37] C. Taylor. *Bootstrap choice of smoothing parameter in kernel density estimation. Biometrika, (76) : 705-712, 1989.*

[38] M.P. Wand, M.C. Jones. *Kernel smoothing. Chapman and Hall, London, 1994.*

[39] E. Youndjé. *Estimation non paramétrique de la densité conditionnelle par la méthode du noyau. Thèse de PhD, Université de Rouen, 1993.*

[40] N. Zougab. *Etude comparative des méthodes de sélection du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau. Mémoire de Magistère, Math. App., Université de Béjaia, 2007.*

[41] *http ://www.cran.r-project.org/*

Résumé

Dans ce mémoire, nous nous intéressons à l'estimation non paramétrique par la méthode du noyau. Nous adoptons cette méthode pour l'estimation de la fonction de densité de probabilité et celle de l'intensité d'un processus de Poisson non homogène.

L'estimation de la fonction de densité, à partir d'un échantillon $X_1, X_2, ..., X_n$ issu d'une variable aléatoire X, nécessite le choix du noyau K et du paramètre de lissage h. Le choix de ces deux paramètres est crucial pour la qualité de l'estimation. Nous donnons les différents noyaux utilisés dans la littérature, à savoir les noyaux symétriques pour les densités définies sur $\mathbb{R}$ et les noyaux asymétriques pour les densités définies sur $[0, +\infty[$ (problème des effets de bord). Nous présentons ensuite une brève synthèse sur les méthodes du choix du paramètre de lissage. Pour l'estimation de l'intensité d'un processus de Poisson non homogène, nous présentons deux modèles mathématiques, le modèle multiplicatif simple d'Aalen (1978) et le modèle de Cox (1980). Une méthode de sélection du paramètre de lissage est donnée par la démarche suivie par Diggle (1985) en se plaçant dans le cas du modèle de Cox.

Malgré que les deux méthodes de sélection de h dans l'estimation de la densité et de l'intensité sont motivées de manières différentes, il y a une équivalence du choix de ce paramètre dans les deux méthodes. Nous présentons cette équivalence et les profits q'elle permet d'avoir pour chaque méthode dans le contexte de l'autre. Une étude de simulation est donnée à la fin de ce mémoire pour illustrer les différents aspects théoriques présentés.

Mots clés : Estimation non paramétrique, méthode du noyau, paramètre de lissage, densité, intensité, processus de Poisson non homogène.

Abstract

In this work, we are interested in nonparametric estimation with the kernel method. We use this method for the estimation of the probability density function and the intensity function of a nonhomogeneous Poisson process. The estimation of the density function from a sample $X_1, X_2, ..., X_n$ from a random variable X, requires the choice of the kernel K and the smoothing parameter h. The choice of these two parameters is crucial for the quality of the estimation. We give the different kernels used in the literature which are the symmetric kernels for the densities defined on $\mathbb{R}$ and the asymmetric kernels for densities defined on $[0, +\infty[$ (problem of boundary effects).
For the estimation of the intensity function, we present two mathematical models, the Aalen's simple multiplicative model (1978) and the Cox model (1980). A method for the selection of the smoothing parameter is given by the technique introduced by Diggle (1985) using the Cox model.

Although the selection methods of the smoothing parameter in the estimation of the density and the intensity are motivated in different ways, there is an equivalence in the selection of this parameter. We describe this equivalence and its benefits for the estimation of both density and intensity. A simulation study is given at the end of this work to illustrate the different theoretic aspects presented.

Key words : Nonparametric estimation, kernel method, smoothing parameter, density, intensity, nonhomogeneous Poisson process.