République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

**UNIVERSITÉ MOHAMED KHIDER, BISKRA**

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

**DÉPARTEMENT DE MATHÉMATIQUES**

Mémoire présenté en vue de l'obtention du Diplôme **:**

**MASTER en Mathématiques**

Option : **statistique**

Par

**RABIE Imane**

Titre :

# Analyse de sensitivité des techniques de validation croisée pour le choix du paramètre de lissage.

Membres du Comité d'Examen :

| | | | |
|---|---|---|---|
| Pr. | **BRAHIMI Brahim** | UMKB | Président |
| Dr. | **CHERFAOUI Mouloud** | UMKB | Encadreur |
| Dr. | **BERKANE Hassiba** | UMKB | Examinatrice |

Juin 2019

# DÉDICACE

*Je dédie ce modeste travail*

*A mon cher père Tahar*

*A ma chère mère Hadda*

*A toute ma famille*

*A tous mes amis*

*A tous qui m'aiment*

*Imane.*

# REMERCIEMENTS

*Je remercie ALLAH le Tout-puissant de m'avoir donné le courage, la volonté et la patience de mener à terme ce présent travail.*

*La première personne que je tiens à remercier est mon encadreur Mr CHERFAOUI Mouloud. Pour ses orientations, sa confiance et sa patience qui ont constitué un apport considérable sans lesquels ce travail n'aurait pu être mené au bon terme. Qu'il trouve ici un hommage vivant à sa haute personnalité.*

*Mes respects et remerciements vont également aux membres du jury qui m'ont fait l'honneur d'évaluer mon travail.*

*Je tiens à exprimer mes sincères remerciements à tous les professeurs qui nous ont enseigné et qui par leurs compétences nous ont soutenu dans la poursuite de nos études.*

*Enfin, je remercie tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail*

Imane.

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction générale

En général, dans les statistiques de toute évidence la densité $f$ génère l'échantillon, mais la question qui se pose lorsque étant données un échantillon est-ce-que nous pouvons approximativement recrées leur fonction de densité ?

Afin d'estimer la densité inconnue $f$, une première approche dite paramétrique consiste à supposer que $f$ appartient à une famille de densités continues ou discrètes qui peuvent être décrites par un certain nombre de paramètres réels. Le statisticien qui opte pour une telle approche possède une bonne connaissance a priori du phénomène aléatoire. Ces modèles paramétriques peuvent être modifiés, lorsque les données présentent des phénomènes spécifiques. Cependant, lorsqu'aucune information n'est disponible sur le phénomène étudié ou le paramètre est de dimension infini, l'application d'un modèle paramétrique n'est pas satisfaisant. Pour pallier les insuffisances et les défauts des familles paramétriques, une seconde approche dite non-paramétrique consiste, à estimer, à partir des observations, une fonction inconnue sans spécifier de forme sur cette fonction à estimer.

Un petit survole de la littérature nous permet de rendre compte que de nombreuses approches non paramétrique ont été proposé pour l'estimation d'une densité de probabilité mais ça reste la technique qui a rencontré le plus de succès est bien la méthode du noyau vue la simplicité de sa forme, ses modes de convergence multiples et sa flexibilité. Cependant, la mise en œuvre de cette technique nécessite le choix d'un noyau $K$ et d'un paramètre de lissage $h$.

Le choix du noyau $K$ dans le cas de densités réelles à supports non bornés est très peu

influent et les critères du choix sont alors la simplicité et la vitesse de calcul. Les noyaux employés ici sont symétriques (dit aussi classiques). Cependant, lorsqu'on veut estimer des densités à support borné au moins d'un seul côté, l'estimateur à noyau classique devient non consistant, à cause des effets du bord. Ce problème est dû à l'utilisation d'un noyau symétrique qui assigne un poids en dehors du support lorsque le lissage est pris en compte près du bord. Plusieurs solutions ont été proposées dans la littérature pour remédier à cette difficulté. La solution la plus simple est de remplacer le noyau symétrique par un noyau asymétrique, qui n'assigne pas un poids en dehors du support de la densité que l'on veut estimer. Autrement dit, le choix doit être adapté selon le support de la fonction inconnue à estimer.

En revanche, le paramètre de lissage est un facteur important et crucial dans l'estimation de la fonction de densité par la méthode des noyaux associés (symétriques et asymétriques). De petites ou de grandes valeurs de $h$ peuvent conduire à une estimation sous ou sur-lissée. Deux catégories de méthodes classiques ont été proposées dans la littérature pour choisir ce paramètre. La première catégorie repose sur la minimisation de l'erreur quadratique moyenne intégrée ($MISE$). Cette classe de méthodes est intéressante en théorie, mais sa difficulté majeure réside dans les applications, en effet, le paramètre de lissage optimal dépend d'une ou plusieurs quantités inconnues. La deuxième catégorie est de type validation croisée, elle est intéressante en pratique car elle se laisse guider seulement par les observations. Cependant, dans le cas de densités à support réel non borné, des études ont montré que les techniques de validation croisée peuvent produire plusieurs optimums locaux.

L'objectif du présent travail est de vérifier, à base des échantillons simulés, si le phénomène des optimums locaux dans les méthodes de validation croisée ($UCV$ et $LCV$) persiste lorsque nous considérons l'estimation à noyau des densités à supports semi-bornés ($x \in \mathbb{R}_+$) ou des densités à supports discret ($x \in \mathbb{N}$). Pour répondre à notre objectif nous avons organiser ce mémoire comme suit :

Dans le premier chapitre nous allons présenter les principales notions de l'estimation de la

densité de probabilité par la méthode du noyau. Ensuite, ses propriétés (biais, variances, ....) et les inconvénients du choix des deux paramètres qui constituent un estimateur à noyau, à savoir : le paramètre de lissage $h$ et le noyau $K$.

Dans le deuxième chapitre, nous allons présenter brièvement l'idée et la notion de l'estimateur à noyau asymétrique dans le cas de variables définies sur $\mathbb{R}_+$ et le cas de variables définies sur $\mathbb{N}$. Par la suite, la question du choix de noyau et du paramètre de lissage ainsi que les propriétés des estimateurs conçu dans ce cadre sera présenté.

Enfin, avant de conclure, dans le troisième chapitre nous allons présenté une application numérique qui illustre le phénomène des optimums locaux dans les méthodes de validation croisée ($UCV$ et $LCV$) pour le choix du paramètre de lissage lors de l'estimation d'une densité de probabilité par la méthode du noyau, et cela dans le cas de : densités à support continu et non-borné ($x \in \mathbb{R}$), densités à support continu positif ($x \in \mathbb{R}_+$) et densités à support discret non-borné ($x \in \mathbb{N}$).

# Chapitre 1

# Estimation à noyau symétrique de la densité de probabilité

## Introduction

Dans ce chapitre, après avoir décrit l'origine de l'estimateur à noyau d'une densité de probabilité, nous avons énoncé ses propriétés. Par la suite, le problème du choix des paramètres de cet estimateur à savoir le choix du noyau et la sélection du paramètre de lissage ont été abordé.

## 1.1 Critères d'erreur

Il est intéressant de commencer par citer quelques normes de mesure d'erreur qui sont un critère de performance de cet estimateur.

Soit $\hat{f}$ un estimateur de la densité de probabilité $f$.

- **Les distances** $L_p$

La distance $L_p$ entre $f$ et $\hat{f}$ est définie par :

$$L_p(f,\hat{f}) = \begin{cases} \left(\int \left|f(x)-\hat{f}(x)\right|^p dx\right)^{1/p}, & \text{si } 0<p<\infty \\ \sup\limits_x \left|f(x)-\hat{f}(x)\right|, & \text{si } p=\infty \end{cases}$$

- **L'erreur quadratique moyenne (MSE)**

$$\begin{aligned} MSE(f(x),\hat{f}(x)) &= \mathbb{E}\left(\hat{f}(x)-f(x)\right)^2 = \left[f(x)-\mathbb{E}\left(\hat{f}\,(x)\right)\right]^2 + \mathbb{E}\left(\hat{f}^2(x)\right) - \left[\mathbb{E}\left(\hat{f}(x)\right)\right]^2 \\ &= \mathbb{V}ar(\hat{f}(x)) + Biais^2\hat{f}(x). \end{aligned} \tag{1.1}$$

- **L'erreur quadratique intégrée (ISE)**

$$ISE(f,\hat{f}) = \int [f(x)-\hat{f}(x)]^2 dx = \int \left[f(x)^2 - 2f(x)\hat{f}(x) + \hat{f}^2(x)\right] dx.$$

- **L'erreur quadratique moyenne intégrée (MISE)**

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

## 1.2 Estimateur à noyaux classiques

L'idée de construction de l'estimateur à noyau d'une densité consiste à évaluer la densité $f$ au point $x$, en comptant le nombre d'observations tombées dans un certain voisinage de $x \in \mathbb{R}$. Plus précisément, le principe de sa construction se base principalement sur le lien entre une densité de probabilité et la fonction de répartition lui associée, toute en exploitant l'estimateur empirique de cette dernière fonction.

**Définition 1** *Soit $x_1, ...., x_n$ un n-échantillon de loi $f(x)$ sur $\mathbb{R}$, de fonction de répartition*

*$F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associé à $x_1, ..., x_n$, la fonction aléatoire $F_n : \mathbb{R} \to [0, 1]$ définie, pour tout $x \in \mathbb{R}$, par :*

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

*Également, elle s'écrit comme suite :*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(x_i)_{]-\infty, x[}}. \tag{1.2}$$

À partir de la définition d'une densité de probabilité et en utilisant l'équation (1.2), on aura :

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} \quad avec \quad h \to 0; \tag{1.3}$$

cette dernière peut être réécrite, en ses points de continuité, sous la forme suivante :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \omega \left( \frac{x - x_i}{h} \right), \tag{1.4}$$

où,

$$\omega(u) = \begin{cases} 1/2, & \text{si } -1 \leq u \leq 1 \\ 0, & \text{sinon} \end{cases} \tag{1.5}$$

Ce dernier est l'estimateur à noyau uniforme dit de **Rosenblatt** [16].

Parzen [15] a étudié une classe plus générale d'estimateurs à noyau uniforme en remplaçant la fonction $\omega$ donnée dans la formule (1.5) par une fonction $K$ satisfaisant les conditions suivantes :

$$\int_{\mathbb{R}} K(u)du = 1, \qquad \int_{\mathbb{R}} uK(u)du = 0 \quad et \quad \sigma_K^2 = \int_{\mathbb{R}} u^2 K(u) < \infty. \tag{1.6}$$

Par analogie avec la définition de l'estimateur de Rosenblatt [16], l'estimateur à noyau (de Parzen [15]) s'écrit :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{1.7}$$

où $h = h(n)$ est appelé *paramètre de lissage* et la fonction $K$ est appelée *noyau*.
Les conditions données dans la formule (1.6) permettent de prouver plusieurs types de convergence (locale et globale) de l'estimateur définie dans la formule (1.7) (voir Silverman [23]).

## 1.3 Propriétés de l'estimateur à noyau

Juste après introduction de l'estimateur à noyau de la densité par Rosenblatt (1956) [16], Parzen (1962) [15] a étudié ses propriétés fondamentales. Depuis, cet estimateur est devenu un objet classique étudié par les statisticiens. L'estimateur de la densité de probabilité par la méthode du noyau est le plus répandu aujourd'hui, car il répond au problème du choix des différents paramètres dans l'estimation à histogramme et possède de bonnes propriétés. Dans cette section nous allons résumer quelques résultats théoriques obtenues sur les propriétés de l'estimateur en question.

### 1.3.1 Espérance, Biais et Variance de l'estimateur

Les expressions de l'espérance, biais et de la variance de l'estimateur à noyau sont données respectivement par (pour plus de détails voir Silverman [23]) :

$$\begin{aligned}
\mathbb{E}\left(\hat{f}(x)\right) &= f(x) + \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2), \\
Biais\left(\hat{f}(x)\right) &= \mathbb{E}\left(\hat{f}(x)\right) - f(x) = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2), \\
\mathbb{V}ar\left(\hat{f}(x)\right) &= \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(y)dy - \frac{f'(x)}{n}\int_{-\infty}^{\infty} yK^2(y)dy - \frac{1}{n}\left(f(x) + biais\hat{f}(x)\right)^2,
\end{aligned}$$

où $\mu_2(K) = \int_{-\infty}^{\infty} y^2 K(y)dy$.

### 1.3.2 Comportement asymptotique de l'estimateur à noyau

Parzen [15] a élaboré les conditions de plusieurs types de convergence de l'estimateur à noyau ainsi que la convergence de ses propriétés. Les principaux résultats obtenus, par l'auteur, sont résumés dans le Théorème suivant :

**Théorème 1.3.1** *Sous les conditions suivantes :*

1. $\lim_{n\to+\infty} h(n) = 0$ *et* $\lim_{y\to+\infty} |yK(y)| = 0$,
2. $\sup_y |K(y)| < \infty$ *et* $\int_{-\infty}^{\infty} |K(y)|dy < \infty$,
3. $\int_{-\infty}^{\infty} K(y)dy = 1$.

*on a :*

$$\lim_{n\longrightarrow\infty} \mathbb{E}\left(\hat{f}(x)\right) = f(x) \text{ et } \lim_{n\longrightarrow\infty} nh\mathbb{V}ar\left(\hat{f}(x)\right) = f(x)\int_{-\infty}^{\infty} K^2(y)dy.$$

*Si de plus,* $\lim_{n\to\infty} nh(n) = \infty$, *alors*

- $\lim_{n\longrightarrow\infty} MSE(\hat{f}(x), f(x)) = 0$, *pour tout point* $x$ *pour lequel la densité* $f$ *est continue,*
- $\lim_{n\longrightarrow\infty} MISE(\hat{f}, f) = 0, \quad \forall f \in \mathbb{L}^p$.
- $\hat{f}(x) \xrightarrow{loi} \mathcal{N}\left(\mathbb{E}\left(\hat{f}(x)\right), \mathbb{V}ar\left(\hat{f}(x)\right)\right)$.
- $\forall \epsilon > 0, \quad P\left(\sup_{x\in\mathbb{R}} |\hat{f}(x) - f(x)| < \epsilon\right) = 1$, *si la transformée de Fourier* $\tilde{K}(z) = \int_{-\infty}^{\infty} \exp(-izy)K(y)dy$ *est absolument intégrable.*

*On note par* $\mathbb{L}^p$ *: l'ensemble des fonctions* $f$ *définies sur* $\mathbb{R}$*, telle que* $\int |f(x)|^p dx < \infty$*.*

Le Théorème suivant donne le résultat élaboré par Nadaraya [13] concernant la convergence uniforme presque complète de l'estimateur à noyau classique.

**Théorème 1.3.2** *Si* $h$*,* $K$ *et* $f$ *satisfaisant les conditions suivantes :*

1. $K$ *est un noyau positif à variation bornée,*
2. $f$ *est uniformément continue,*
3. $\lim\limits_{n\longrightarrow\infty} h(n) = 0,$
4. $\sum\limits_{n=1}^{\infty} \exp(-\gamma n h(n)^2) < \infty, \quad \forall \gamma > 0,$

*alors :*

$$\sup_x |\hat{f}(x) - f(x)| \longrightarrow 0, \quad \textit{avec une probabilité 1.}$$

Pour la convergence en $L_1$ presque complète, Devroye [6] a dégagé des conditions de convergence, qui sont résumées dans le Théorème suivant :

**Théorème 1.3.3** *Si,*

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

*alors,*

$$\forall f \in \mathcal{F}, \ \lim_{n\longrightarrow\infty} \int |\hat{f}(x) - f(x)| dx = 0, \ \textit{Presque Complètement,}$$

*où* $\mathcal{F}$ *est l'ensemble des densités de probabilité.*

### 1.3.3 Vitesse de convergence

Wahba [25] a montré qu'on ne peut pas améliorer indéfiniment la convergence d'un estimateur $\hat{f}$ vers $f$, même pour la fonction la plus régulière possible (indéfiniment dérivable, bornée). C'est-à-dire, $MSE(\hat{f}(x), f(x))$ ne peut tendre vers 0 que d'un ordre $\frac{c}{n}$, où $c$ est une constante.

## 1.4 Choix du noyau

Un noyau approprié aide à surmonter les problèmes des bosses (multi-modes) et de la discontinuité de la densité estimée. Par exemple, si $K$ est une distribution gaussienne, alors la fonction de densité estimée $\hat{f}$ sera lisse et admet des dérivées de toutes ordres. Dans la littérature, il existe plusieurs fonctions qui jouent le rôle d'un noyau, la Table 1.1 résume les noyaux les plus usuels dont leurs formes sont illustrées dans la Figure 1111.

TABLE 1.1: Noyaux usuels et leurs supports.

| Nom | Expression | Domaine |
|---|---|---|
| Noyau Uniforme (Rosenblatt) | $K(u) = \frac{1}{2}$ | $\|u\| \leq 1$ |
| Noyau Box (boite) | $K(u) = \frac{1}{2\sqrt{3}}$ | $\|u\| \leq \sqrt{3}$ |
| Noyau Triangulaire | $K(u) = (1 - \|u\|)$ | $\|u\| \leq 1$ |
| Noyau Cosine | $K(u) = \frac{\pi}{4}\cos(\frac{\pi u}{2})$ | $\|u\| \leq 1$ |
| Noyau Gaussien | $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$ | $u \in \mathbb{R}$ |
| Noyau Biweight (Tukey) | $K(u) = \frac{15}{16}(1-u^2)^2$ | $\|u\| \leq 1$ |
| Noyau Triweight | $K(u) = \frac{35}{32}(1-u^2)^3$ | $\|u\| \leq 1$ |
| Noyau Epanechnikov | $K_E(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)$ | $\|u\| \leq \sqrt{5}$ |

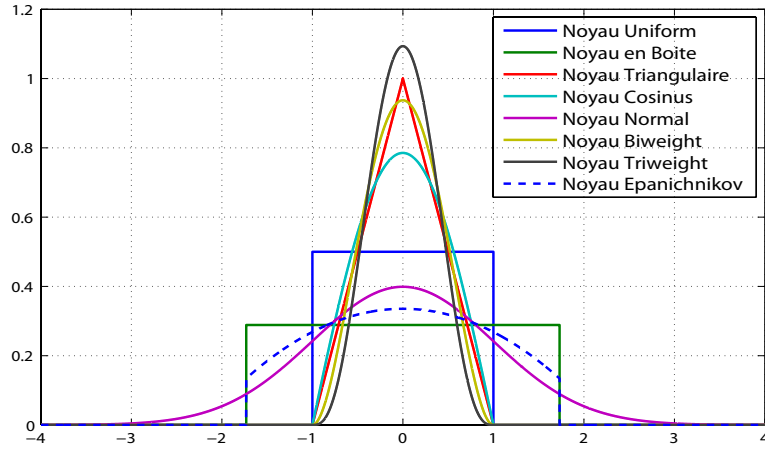


FIGURE 1.1 – La forme des noyaux usuels

## 1.5 Choix du paramètre de lissage

D'après la formule (1.7), on constate que l'estimateur $\hat{f}(x)$ de $f(x)$ ne dépend pas seulement du noyau $K$, mais aussi du paramètre de lissage $h$. Dans la pratique, l'étape critique dans l'estimation d'une densité par la méthode du noyau est le choix du paramètre $h$, qui contrôle la qualité graphique de l'estimateur $\hat{f}$ [15, 23]. En effet, le choix de $h$ est une étape importante lors de l'estimation par la méthode du noyau, dans le sens où une petite perturbation de $h$ est suffisante pour que les caractéristiques de $\hat{f}$ changent complètement (performances numériques et/ou graphiques). Par ailleurs, si $h$ est trop petit, le biais de l'estimateur devient petit devant sa variance et l'estimateur sera trop fluctuant. Pour cela, on obtient un phénomène de *sous-lissage*. Dans le cas contraire, lorsque $h$ est trop grand, le biais prend l'ascendant sur la variance et l'estimateur varie peu, alors on obtient un phénomène de *sur-lissage*.

L'exemple présenté dans la Figure 1.2, réalisé dans le cadre d'estimation d'une densité d'une loi normale, centrée réduite, à partir d'un échantillon de taille $n = 200$, est une illustration de l'influence du choix du paramètre de lissage sur les caractéristiques graphiques de l'estimateur en question. Les graphes des trois estimateurs présentés, mis en évidence le phénomène de sur-lissage dans le cas $h = 0.8$ (trop grand), le phénomène de sous-lissage dans le cas $h = 0.1$ (trop petit) et l'estimation idéale dans le cas $h^* = 0.3334$ ($h^*$ est l'optimal au sens du $ISE$).

Il existe Dans la littérature plusieurs techniques sont proposées pour la sélection de ce paramètre que l'on peut regrouper en deux familles :

- Méthodes de plug-in (re-injection)
- Méthodes de Cross-Validation (Validation-croisée).

La multitude des méthodes de sélection de paramètre de lissage et leurs diversités du point de vue de leurs principes, est due au fait que ces méthodes restent incomplètes. Autrement dit, ces méthodes ont toujours des inconvénients [26], soit au sens de la qualité numériques de l'estimateur $\hat{f}$, par rapport à une norme d'erreur bien déterminée, ou au sens de la

qualité graphique de l'estimateur (l'allure graphique de la courbe de $\hat{f}$ est sur-lissée ou sous-lissée).

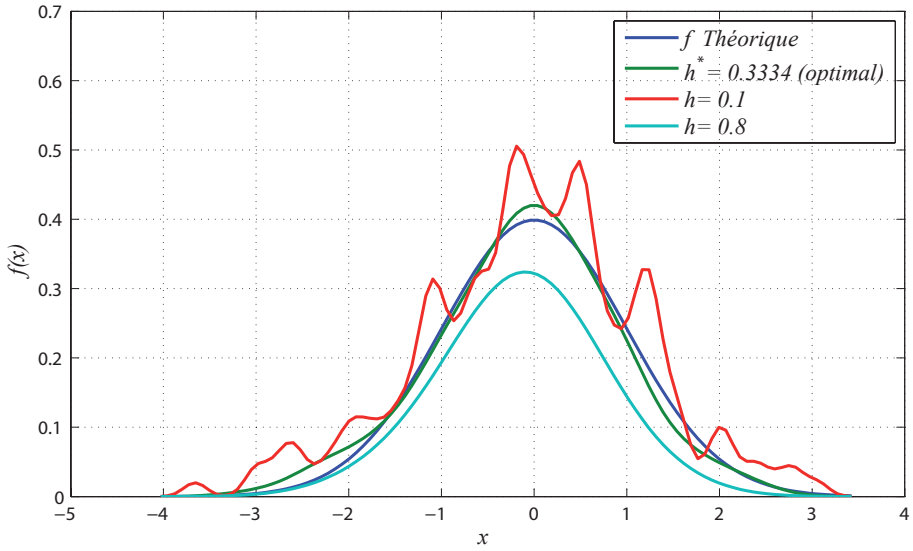


FIGURE 1.2 – Illustration des phénomènes sur-lissage ($h = 0.8$), sous-lissage ($h = 0.1$) et estimation idéal ($h^* = 0 : 3334$)

Dans ce qui suit nous allons présenter quelques méthodes de sélection les plus usuelles.

### 1.5.1 Choix optimal

La décision du choix du paramètre de lissage $h$ suppose la spécification d'un critère d'erreur qui puisse être optimisé. De ce fait, il est claire que l'optimalité n'est pas un concept absolu, de fait que cette optimalité est étroitement liée au choix d'un critère, qui fait intervenir à la fois la densité inconnue $f$ et l'estimateur $\hat{f}$ (c'est-à-dire le paramètre $h$ et le noyau $K$). Supposons que nous nous intéressons au choix du paramètre de lissage $h$ qui minimise l'Erreur Quadratique Intégrée Moyenne ($MISE$), c'est-à-dire, la quantité suivante :

$$\arg\min_h MISE\left(f, \hat{f}\right) = \arg\min_h \int \mathbb{E}\left(\hat{f}(x) - f(x)\right)^2 dx.$$

Afin de déterminer le $h$ optimal, au sens du $MISE$, nous allons exploiter le résultat suivant :

**Théorème 1.5.1** *(Scott [19])*

*Si $f$ a une dérivée seconde absolument continue, $f^{(3)} \in \mathbb{L}^2$, le noyau $K \in \mathbb{L}^2$ et une densité de probabilité continue, symétrique de variance $\sigma_K^2 > 0$, alors, sous les conditions $h(n) \longrightarrow 0$ et $nh(n) \longrightarrow \infty$, on a le développement asymptotique :*

$$MISE = \frac{h^4}{4}\sigma_K^4 \int (f''(x)dx)^2 \; + \; \frac{\int K^2(x)dx}{nh} \; + \; o\left(h^5 + \frac{1}{n}\right), \tag{1.8}$$

*où, $\mathbb{L}^2$ est l'ensemble des fonctions $f$ définies sur $\mathbb{R}$, telles que $\int |f(x)|^2 dx < \infty$.*

A partir de l'expression (1.8) on définit la quantité suivante :

$$AMISE = \frac{h^4}{4}\sigma_K^4 \int (f''(x))^2 \, dx \; + \; \frac{\int K^2(x)dx}{nh}, \tag{1.9}$$

appelée l'Erreur Quadratique Intégrée Moyenne Asymptotique.

On remarque que le premier terme du membre à droite du développement (1.9) est un terme de biais, alors que le second est un terme de variance. De plus, on constate que, le terme du biais est une fonction croissante en $h$, alors que le terme de la variance est une fonction décroissante en $h$. C'est-à-dire, les deux termes varient dans le sens inverse par rapport à $h$. Une largeur de fenêtre $h$ trop importante entraînera une augmentation du biais et une diminution de la variance, alors qu'une largeur de fenêtre trop petite provoquera une augmentation de la variance et une diminution du biais (voir Figure 1.3). De ce fait, le paramètre de lissage $h^*$ optimal au sens du critère de l'$AMISE$, devra réaliser un compromis entre les valeurs de la variance et celle du biais.

Par ailleurs, pour obtenir le paramètre de lissage $h^*$ qui minimise l'Erreur Quadratique Intégrée Moyenne Asymptotique, il suffit de résoudre le système suivant :

$$\begin{cases} \frac{dAMISE}{dh} = 0, \\ \frac{d^2AMISE}{dh^2} > 0. \end{cases} \tag{1.10}$$

À partir de l'expression (1.9), on aura :

$$h^* = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{1/5} n^{-1/5}, \tag{1.11}$$

avec $R(g) = \int (g(x))^2 \, dx$.

Notons que $h^*$ est une quantité déterministe qui dépend du nombre d'observations $n$.

La valeur du $AMISE$ optimale ($AMISE^* = AMISE(h^*)$) est donnée par :

$$AMISE^* = \frac{5}{4}\left[\sigma_K R^4(K) R(f'')\right]^{1/5} n^{-4/5}. \tag{1.12}$$

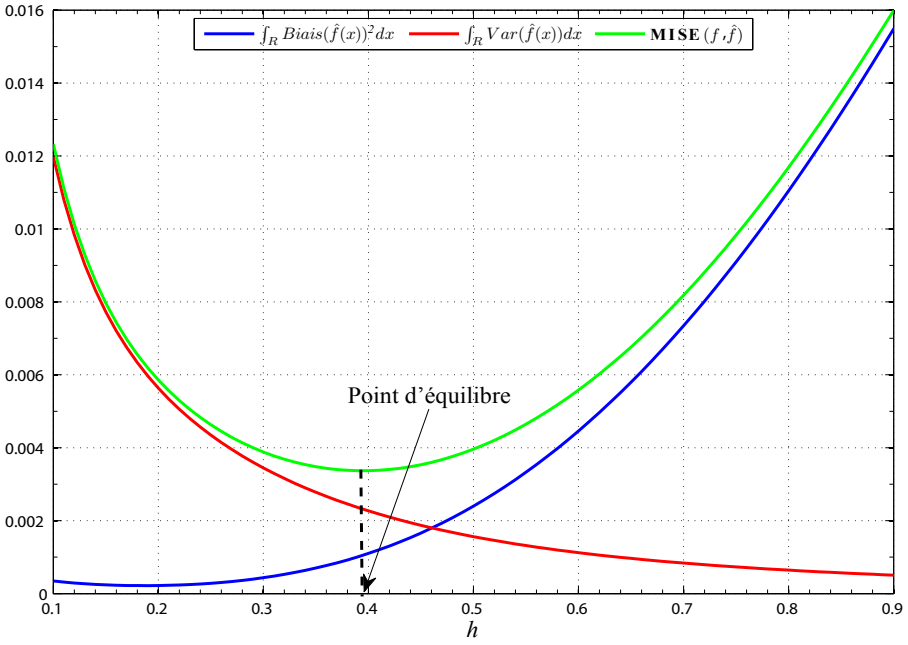


FIGURE 1.3 – Illustration du principe de l'équilibre biais-variance à l'aide d'une simulation.

En plus de sa nature asymptotique, la largeur de fenêtre optimale $h^*$ dépend de la densité inconnue $f$ au travers du paramètre $R(f'')$. Cette largeur de fenêtre "idéale" (relativement au critère d'erreur retenu) n'est donc pas directement calculable dans la pratique. Une façon classique de remédier à ce dernier problème consiste à remplacer la quantité $R(f'')$ par un estimateur approprié ou une quantité bien définie, d'où le principe des méthodes

plug-in.

### 1.5.2 Estimateur Rule Of Thumb (règle de référence)

L'estimateur "Rule Of Thumb" du paramètre de lissage, noté $h_{rot}$, suppose que nous utilisons le noyau gaussien pour estimer une densité $f$ d'une distribution normale centrée (la moyenne égale à 0) et de variance $\sigma^2$. De ce fait, la quantité $R(f'')$ est définie par :

$$R(f'') = \int (f''(x))^2 \, dx = \frac{3}{8}\sqrt{\pi}\sigma^{-5}. \tag{1.13}$$

En substituant $R(f'')$ et $K$ par leurs formules dans (1.11), on obtient :

$$h_{rot} = (4\pi)^{-1/10} \left[\frac{3}{8}\pi^{-1/2}\sigma\right] n^{-1/5} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} = 1.06\sigma n^{-1/5}.$$

Il suffit donc d'estimer $\sigma$ à partir des données et de substituer cet estimateur dans la formule ci-dessus. D'après Silverman (1986) [23], la formule (1.14) donnera de bons résultats si la population est réellement normalement distribuée. Par contre, elle peut donner une distribution trop lissée si la population est multimodale. Dans ce cas, de meilleurs résultats peuvent être obtenus, si on utilise l'interquartile $IQ$ définie par :

$$IQ = \frac{X_{(3n/4)} - X_{(n/4)}}{1.34},$$

où, $X_{(n/4)}$ et $X_{(3n/4)}$ sont respectivement le premier et le troisième quartile de l'échantillon observé.

Or, dans le cas où $X$ suit une loi normale, l'écart interquartile est $IQ = 1.394$ alors $h_{rot}$ de l'équation (1.14) devient

$$h_{rot} = 1.06 IQ n^{-1/5}.$$

Cette dernière formule peut aussi donner une distribution trop lissée si la vraie densité est multimodale. Parfois cette dernière donne des résultats moins bons que si l'on avait

utilisé l'écart type, d'où le meilleur des deux méthodes peut être obtenu en utilisant un estimateur adaptatif de l'étendue. C'est-à-dire, en utilisant $A$ au lieu de $\sigma$ dans la formule (1.14) où $A$ est défini par $A = \min(\sigma, IQ)$, donc la formule pour $h_{rot}$ devient alors :

$$h_{rot} = 1.06An^{-1/5}. \tag{1.14}$$

Cette correction est insuffisante dans de nombreux cas. Par exemple, si la vraie densité est multimodale.

### 1.5.3 Estimateur de Sheather et Jones

Sheather et Jones (1991) [22], recommandent l'utilisation de l'estimateur naturel $\hat{R}_a(f'')$, en faisant observer que le terme de biais $\frac{R(K'')}{na^5}$ est positif et peut donc servir à annuler le terme de biais (négatif) de l'erreur quadratique moyenne entre $\hat{R}_a(f'')$ et $R(f'')$. Afin de faire disparaître quelques effets indésirables du terme du biais. Les deux auteurs sont contraints de mettre en place une procédure de type plug-in en trois étapes.

Sheather et Jones [22], choisissent d'estimer $R(f'') = \int_{-\infty}^{+\infty}(f'')^2 dx$ par :

$$S(a) = \frac{1}{n(n-1)a^5} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(4)} \left( \frac{x_i - x_j}{a} \right), \tag{1.15}$$

où $L^{(4)}$ désigne la dérivée quatrième du noyau suffisamment lisse $L$ et $a$ est un nouveau paramètre de lissage appelé paramètre pilote.

### 1.5.4 Validation croisée non biaisée (UCV)

Cette méthode a été proposée par Rudemo (1982) [17] et Bowman (1984) [2]. Elle consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

$$\begin{aligned} UCV\,(h) &= \int_{\mathbb{R}} \left[\hat{f}(x) - f(x)\right]^2 dx - \int_{\mathbb{R}} f^2(x)dx \\ &= \int_{\mathbb{R}} \hat{f}^2(x)dx - 2\int_{\mathbb{R}} \hat{f}(x)f(x)dx. \end{aligned}$$

Puisque $\int_{\mathbb{R}} f^2(x)dx$ ne dépend pas du paramètre de lissage $h$. On peut choisir alors le paramètre de lissage de façon à ce qu'il minimise un estimateur de :

$$\int_{\mathbb{R}} \hat{f}^2(x)dx - 2\int_{\mathbb{R}} \hat{f}(x)f(x)dx.$$

Maintenant, on veut trouver un estimateur de $\int_{\mathbb{R}} \hat{f}(x)f(x)dx$. Pour cela, remarquons que

$$\int_{\mathbb{R}} \hat{f}(x)f(x)dx = \mathbb{E}\left(\hat{f}(x)\right).$$

L'estimateur empirique de $\int_{\mathbb{R}} \hat{f}(x)f(x)dx$, est alors

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(x_i),$$

et le critère à minimiser est :

$$UCV(h) = \int_{\mathbb{R}} \hat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_i(x_i), \tag{1.16}$$

avec,

$$\hat{f}_i(x_i) = \frac{1}{(n-1)h}\sum_{j=1, j\neq i}^{n} K\left(\frac{x_i - x_j}{h}\right), \tag{1.17}$$

est l'estimateur de la densité construit à partir de l'ensemble de points sauf au point $x_i$.

Nous notons par $h_{ucv}$ l'estimateur de $h$ qui minimise $UCV(h)$.

La popularité de cette méthode est due à la motivation intuitive et au fait que cet estimateur est asymptotiquement optimal sous de faibles conditions. L'optimalité asymptotique

de la validation croisée non biaisée a été obtenue par Stone [24].

Cependant, cette méthode présente deux problèmes majeurs. D'une part, son manque de robustesse par rapport aux changements de la taille de l'échantillon c'est-à-dire, le résultat de simulation peut se révéler extrêmement variable d'un échantillon à un autre. D'autre part, la fonctionnelle à minimiser a souvent tendance à présenter plusieurs minimums locaux [9]. Pour d'autres études voir Hall [8], Burman [3], Scott et Terrell [20].

### 1.5.5 Validation croisée biaisée (BCV)

Le critère de validation croisée biaisée, a été introduit par Scott et Terrell (1987) [20] pour remédier aux problèmes de la méthode "validation croisée non biaisée". Il s'agit d'introduire un biais dans le $UCV$ afin de réduire sa variance.

Rappelons que l'Erreur Quadratique Intégrée Moyenne Asymptotique s'écrit sous la forme :

$$AMISE = \frac{h^4}{4}\sigma_K^4 R(f'') + \frac{R(K)}{nh}. \tag{1.18}$$

Le paramètre de lissage basé sur la méthode de validation croisée biaisée est la valeur de $h$ qui minimise un estimateur du $AMISE$. À partir de la formule (1.18), il est claire que afin d'estimer l'$AMISE$, il suffit d'estimer $R(f'')$. Un estimateur naturel de ce dernier terme est donné par $R(\hat{f}'')$ ($\hat{f}$ est l'estimateur de la densité $f$ obtenu par la méthode du noyau). Finalement, Scott et Terrell [20] ont proposé la forme de l'estimateur de $AMISE$ à minimiser (critère de $BCV(h)$), qui se résume comme suit :

**Proposition 1** *(Scott et Trell [20])*

*Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire $X$ de fonction de densité $f$. Pour un noyau $K$, on obtient :*

$$BCV(h) = \frac{R(K)}{nh} + h^4\frac{\mu_2^2(K)}{4n^2}\sum_i\sum_{j,j\neq i} K_h^{(2)}K_h^{(2)}(X_i - X_j). \tag{1.19}$$

Des résultats de simulations ont été obtenus pour la méthode de validation croisée biaisée dans le travail de Park et Marron [14]. Les auteurs ont constaté que la méthode validation croisée biaisée présente le même point faible que celui de la méthode validation croisée non biaisée. Cette méthode nous fournis plusieurs minimums locaux pour la fonctionnelle cible à minimiser.

### 1.5.6 Validation croisée par le maximum de vraisemblance ($LCV$)

Le paramètre de lissage basé sur la méthode de validation croisée par le maximum de vraisemblances est le paramètre qui maximise la fonction suivante :

$$LCV(h) = \left( \frac{1}{n} \sum_{i=1}^{n} \log \left( \hat{f}_i(X_i) \right) \right), \tag{1.20}$$

où $\hat{f}_i(X_i)$ est donné par (1.17).

## Conclusion

Dans ce chapitre, nous avons présenté les principales notions d'estimation de la densité de probabilité par la méthode du noyau. En effet, après présentation de la définition de l'estimateur à noyau d'une densité de probabilité et ses propriétés (biais, variances, ....), nous nous somme intéressés sur le problème de choix du noyau et de paramètre de lissage où nous nous somme focalisés principalement sur les avantages et les inconvénients des méthodes de selection du paramètre de lissage $h$ proposées dans la littérature.

# Chapitre 2

# Estimation à noyaux asymétrique d'une densité de probabilité

## Introduction

Dans ce chapitre, nous allons présenter brièvement l'idée et la notion de l'estimateur à noyau asymétrique (dans le cas de variables continues et le cas de variables discrètes) où notre intérêt sera orienté, par la suite, vers la question du choix du noyau et du paramètre de lissage ainsi que les propriétés des estimateurs conçu dans ce cadre.

## 2.1 Problème d'effet du biais aux bornes

Soient $X_1, X_2, ..., X_n$ un $n$-échantillon issu d'une densité de probabilité inconnue $f(x)$. L'estimateur classique de $f(x)$ obtenu par la méthode du noyau, proposé par Rosenblatt [16] suivi de Parzen [15], come cité dans le chapitre 1, s'écrit sous la forme :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

où $h$ représente le paramètre de lissage et $K$ est un noyau qui vérifie les conditions sui-

vantes :

$$\int_{\mathbb{R}} K(y)dy = 1, \quad \int_{\mathbb{R}} yK(y)dy = 0, \quad \text{et} \quad \int_{\mathbb{R}} y^2K(y)dy = \sigma_K^2 < \infty. \tag{2.2}$$

L'estimateur à noyau continu (2.1) a été développé primairement pour les densités à supports continus et non-bornés. La fonction noyau $K$ est classiquement symétrique (i.e. $K(-x) = K(x)$), elle est considérée comme moins importante que le paramètre de lissage $h$. Bien qu'un noyau symétrique soit approprié pour ajuster des densités à supports non-bornés, il ne l'est pas pour des densités à supports compacts ou bornés d'un côté et a fortiori à supports discrets.

En effet, lorsque on veut estimer des densités à support borné au moins d'un seul coté, l'estimateur à noyau classique devient non consistant, à cause des effets au bornes. Ce problème est dû à l'utilisation d'un noyau symétrique qui assigne un poids en dehors du support lorsque le lissage est pris en compte près de la borne. Plusieurs solutions ont été proposées dans la littérature pour remédier à cette difficulté. La solution la plus simple est de remplacer le noyau symétrique par un noyau asymétrique, qui n'assigne pas un poids en dehors du support de la densité que l'on veut estimer. Cette idée est due à l'origine aux travaux de Chen [4, 5]. Ainsi, la naissance de la notion des noyaux asymétrique où l'expression de l'estimateur, à noyau asymétrique d'une densité $f$, est donnée par :

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) \quad x \in \mathbb{R}, \tag{2.3}$$

avec $h$ est le paramètre de lissage et $K_{x,h}$ sera dit alors "noyau asymetrique" de cible $x$ et de fenêtre $h$.

## 2.2 Estimateur à noyaux asymétriques gamma

Soit $X_1,\ X_2, ...,\ X_n$ un $n$-échantillon issu d'une densité de probabilité inconnue $f$ qui est définie sur un support positif ($[0,\ \infty[$) et deux fois continûment derivable ($f \in \mathcal{C}^2([0,\infty[)$). Parmi les estimateurs proposés pour cette famille de distributions, on cite l'estimateur de Chen [5] qui suggère de remplacer l'estimateur classique par :

$$\hat{f}_G(x) = \frac{1}{n}\sum_{i=1}^{n} K_{(\rho_h(x),h)}(X_i), \tag{2.4}$$

où, $h = h(n)$ représente le paramètre du lissage satisfaisant les conditions $h \to 0$ et $nh \to \infty$ lorsque $n \to \infty$ et $K_{(\rho_h(x),h)}$ est la fonction de densité de la distribution gamma de paramètres $(\rho_h(x), h)$, donnée par la formule suivante :

$$K_{(\rho_h(x),h)}(t) = \frac{t^{\rho_h(x)-1}e^{-t/h}}{h^{\rho_h(x)}\Gamma(\rho_h(x))}, \tag{2.5}$$

avec

$$\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx, \quad p > 0.$$

La première version de l'estimateur à noyau gamma, notée $\hat{f}_{G_1}(x)$, est obtenue en remplaçant $\rho_h(x)$ par $x/h+1$, dans la formule (2.4) (voir [5]). Le biais et la variance asymptotiques de $\hat{f}_1(x)$ sont donnés respectivement par :

$$Biais(\hat{f}_{G_1}(x)) = h\left\{f'(x) + \frac{1}{2}xf''(x)\right\} + o(h), \tag{2.6}$$

$$Var\left(\hat{f}_{G_1}(x)\right) = n^{-1}\frac{h^{-1}\Gamma\left(2x/h+1\right)}{2^{2x/h+1}\Gamma^2(x/h+1)}f(x) + o(n^{-1}). \tag{2.7}$$

En raison de la contribution indésirable de $f'$ dans le biais de l'estimateur $\hat{f}_{G_1}(x)$ (voir l'expression du biais donnée par la formule (2.6)). Une autre version de $\hat{f}_{G_1}(x)$ appelée estimateur à noyau gamma modifié, notée $\hat{f}_{G_2}(x)$, est obtenue en remplaçant $\rho_h(x)$, dans

la formule (2.4), par la nouvelle quantité suivante :

$$\begin{aligned}\rho_h(x) &= (x/h)\mathbb{I}_{\{x\geq 2h\}} + \left(\frac{1}{4}(x/h)^2+1\right)\mathbb{I}_{\{x\in[0,2h[\}} \\ &= \begin{cases} x/h, & \text{si } x \geq 2h, \\ \frac{1}{4}(x/h)^2+1, & \text{si } x \in [0,2h[. \end{cases}\end{aligned} \tag{2.8}$$

Le biais asymptotique de $\hat{f}_{G_2}(x)$ est exprimé par la formule suivante :

$$Biais\left(\hat{f}_{G_2}(x)\right) = \begin{cases} \frac{h}{2}xf''(x), & \text{si } x \geq 2h, \\ h\xi_h(x)f'(x), & \text{si } x \in [0,2h[, \end{cases} \tag{2.9}$$

où, $\xi_h(x) = (1-x)\left\{\rho_h(x) - x/h\right\} / \left\{1 + h\rho_h(x) - x\right\}$, tandis que sa variance est similaire à celle de $\hat{f}_{G_1}(x)$ [5].

L'avantage de ces deux noyaux est que la forme et la qualité du lissage des estimateurs qu'ils nous fournis changent selon la position où la densité est estimée, ce qui fait la différence avec les noyaux symétriques. Ils sont exempts du problème de biais aux bornes, non négatif et réalisent un taux de convergence optimal pour l'erreur carrée intégrée moyenne (MISE). Également, ils atteignent 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, ils permettent une réduction de la variance lors du lissage en s'éloignant de la borne.

D'autres propriétés des estimateurs à noyaux gamma sont bien documentées dans la littérature. Bouezmarni & Scaillet [1] ont établi les conditions de convergence faible, de ce dernier, sur un compact $[0,+\infty[$, lorsque $f$ est continue sur ce support ainsi que la convergence faible au sens $MIAE$ (Mean Integer Absolute Error). Pour les densités non bornées à l'origine (au voisinage de zéro), les mêmes auteurs ont examiné les performances de cet estimateur par simulation et ont prouvé la convergence en probabilité. Fernandez & Monteira [7] ont établi le théorème central limite pour l'estimateur fonctionnel à noyaux gamma.

Le $MISE$, le $h$ optimal (au sens du MISE) ainsi que le $MISE$ associe a ce dernier correspondants aux deux noyaux sont comme suit :

**MISE :**

$$\begin{aligned} MISE\left(\hat{f}_{G_1}\right) &= h^2\int_0^\infty \{f'(x)+\frac{1}{2}xf''(x)\}^2dx + \frac{n^{-1}h^{\frac{-1}{2}}}{2\sqrt{\pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)dx + \circ(n^{-1}h^{\frac{-1}{2}}+h^2) \\ MISE\left(\hat{f}_{G_2}\right) &= \frac{1}{4}h^2\int_0^\infty \{xf''(x)\}^2dx + \frac{1}{2\sqrt{\pi}}n^{-1}h^{\frac{-1}{2}}\int_0^\infty x^{\frac{-1}{2}}f(x)dx + \circ(n^{-1}h^{\frac{-1}{2}}+h^2) \end{aligned}$$

**Paramètre de lissage optimal :**

$$\begin{aligned} h^*_{G_1} &= 4^{\frac{-2}{5}}\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)dx\right]^{\frac{2}{5}}\left[\int_0^\infty \{f'(x)+\frac{1}{2}xf''(x)\}^2dx\right]^{\frac{-2}{5}} n^{\frac{-2}{5}}. \\ h^*_{G_2} &= \left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)dx\right]^{\frac{2}{5}}\left[\int_0^\infty \{xf''(x)\}^2dx\right]^{\frac{-2}{5}} n^{\frac{-2}{5}}. \end{aligned}$$

**MISE Optimal :**

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

$$MISE^*\left(\hat{f}_{G_2}\right) = \frac{5}{4^{4/5}}\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{-1/2}f(x)dx\right]^{4/5}\left[\int_0^\infty \{xf''(x)\}^2dx\right]^{1/5} n^{-4/5}.$$

La méthode proposée ci-dessus qui consiste à choisir le paramètre de lissage $h$ de sorte à minimiser $MISE$, a un intérêt purement théorique. La difficulté majeure de cette méthode réside en fait dans les applications, car l'expression du $h$ optimal dépend généralement de trois quantités inconnues $f,\ f',\ f''$. Ceci rend plus difficile le choix du paramètre de lissage. A cet effet, pour la sélection du paramètre de lissage l'idée la plus naturelle est d'adopter les mêmes techniques exposées dans le cas des noyaux symétriques, à savoir : les méthodes plug-in et les méthode de validation croisée. Cependant, comme on la cité auparavant, l'$AMISE$ dépend généralement des quantités inconnues $f,\ f',\ f''$. Ceci rend plus difficile le choix du paramètre de lissage par les méthodes plug-in c'est l'une des raisons pour

laquelle les méthodes les plus répandues dans le cas d'estimation à noyaux asymétriques est les méthodes de validation croisée.

Considérons un $n$-échantillon $X_1, X_2, ..., X_n$ $.i.i.d$ issue de la variable aléatoire $X$, et un noyau asymétrique $K_{x,h}$ la méthode la plus utilisée est celle qui minimise un estimateur convenable du $ISE\,(h)$. Ainsi, avec le même raisonnement et la même démarche que dans le cas des noyaux symétriques on peut montrer que le critère à minimiser dans ce cas noyaux asymétriques d'une manière générale et en particulier le cas de noyaux gamma est bien que la fonctionnelle suivante :

$$UCV(h) = \int \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{x,h}\,(x_i) \right\}^2 dx - \frac{2}{n\,(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K_{x_i,h}\,(x_j)\,. \tag{2.10}$$

On peut également sélectionner le paramètre de lissage en utilisant la validation croisée par le maximum de vraisemblance. Le $h$ optimal dans ce cas est donné par :

$$h_{lcv} = \arg\max_{h>0} \left( \frac{1}{n} \sum_{i=1}^{n} \log\left( \hat{f}_i(X_i) \right) \right), \tag{2.11}$$

avec

$$\hat{f}_i(X_i) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} K_{X_i,h}(X_j). \tag{2.12}$$

## 2.3 Estimateur à noyaux asymétriques discret

Dans cette section notre intérêt porte sur la notion d'estimateur à noyau d'une densité discrète. Ci-dessous une définition liée à la notion de l'estimation à noyau discret.

**Définition 2** *Soit $X_1, X_2, ....., X_n$ un $n$-échantillon iid issu d'une variable aléatoire $X$ de la fonction de masse de probabilité inconnue $f$ sur $\mathbb{N}$. L'estimateur à noyau asymétrique discret de $f$ est défini par :*

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i), \tag{2.13}$$

*où $h$ est le paramètre de lissage (fenêtre) et $K_{x,h}$ est le noyau asymétrique discret dépendant de $x$ et $h$ de support $N_{x,h} = N_x$ (ne dépend pas de $h$).*

### 2.3.1 Propriétés de l'estimateur à noyau discret

Dans cette section nous allons introduire la définition quelques propriétés de l'estimateur à noyau discret, qui ont été établis principalement par Senga Kiessé [10] et Kokonendji et Senga Kiessé [11], ainsi que les conditions de convergence en moyenne, en moyenne quadratique et en moyenne quadratique intégrée.

**Proposition 2** *Soit $X_1, X_2, ....., X_n$ un n-échantillon iid issu d'une variable aléatoire $X$ de la fonction de mass de probabilité inconnue $f$ sur $\mathbb{N}$, si $\hat{f}$ est l'estimateur à noyau asymétrique discret de $f$, alors, pour $x \in \mathbb{N}$ et $h > 0$ on a :*

$$E(\hat{f}(x)) = E(\mathcal{K}_{x,h})$$

*où $\mathcal{K}_{x,h}$ est la variable aléatoire de loi $K_{x,h}$ définie sur $N_x$. De plus, on a $\hat{f}(x) \in [0\ ;\ 1]$ pour $x \in \mathbb{N}$ et*

$$E(\hat{f}(x)) = \sum_{y \epsilon N \cap N_x} f(y) K_{x,h}(y) \longrightarrow f(x) \text{ quand } h \to 0 \text{ lorsque } n \longrightarrow +\infty$$

– **L'erreur quadratique moyenne** $(MSE)$ :

$$MSE(\hat{f}(x)) = E[\hat{f}(x) - f(x)]^2 = Var(\hat{f}(x)) + Biais^2((\hat{f}_h(x)),$$

où $E[\hat{f}_h(x) - f(x)]^2 \longrightarrow 0$ quand $nh \longrightarrow +\infty$ et $h \to 0$.

– **L'erreur quadratique moyenne intégrée** $(MISE)$ :

$$
\begin{aligned}
MISE(\hat{f}_h(x)) &= \sum_{x\epsilon\mathbb{N}} MSE(f(x), \hat{f}_h(x)) = \sum_{x\epsilon\mathbb{N}} Var(\hat{f}(x)) + \sum_{x\epsilon\mathbb{N}} Biais^2((\hat{f}(x)), \\
&= \frac{1}{n}\sum_{x\epsilon\mathbb{N}_x} f(x)\left[(\Pr(K_{x,h}=x))^2 - f(x)\right] \\
&+ \sum_{x\epsilon\mathbb{N}_x}\left[f\left(E(K_{x,h})\right) - f(x) + \frac{1}{2}Var(K_{x,h})f^{(2)}(x)\right]^2 + o\left(\frac{1}{n}+h^2\right) \\
&= MISE(n,h,K,f), \qquad (2.14)
\end{aligned}
$$

avec $f^{(2)}$ représentent la différence finie d'ordre 2, donnée par :

$$
f^{(2)}(x) = \begin{cases} \left\{f\left(x+2\right) - 2f\left(x\right) + f\left(x-2\right)\right\}/4, & \text{si } \; x \in \mathbb{N}\backslash\left\{0,1\right\}; \\ \left\{f\left(3\right) - 3f\left(1\right) + 2f\left(0\right)\right\}/4, & \text{si } \; x = 1; \\ \left\{f\left(2\right) - 2f\left(1\right) + f\left(0\right)\right\}/2, & \text{si } x = 0. \end{cases}
$$

Rappelons que la différence finie d'ordre 1 est donnée par :

$$
f^{(1)}\left(x\right) = \begin{cases} \left\{f\left(x+1\right) - f\left(x-1\right)\right\}/2, & \text{si } x \in \mathbb{N}/\left\{0\right\}; \\ f\left(1\right) - f\left(0\right), & \text{si } x = 0. \end{cases}
$$

### 2.3.2 Choix du noyau

Dans la littérature plusieurs fonctions sont proposées pour jouer le rôle du noyau discret. Le table suivante résume les noyaux les plus usités dans le cadre d'estimation d'une densité discrète, ainsi que leurs caractéristiques.

| Noyau | La forme | $E\left(\mathcal{K}_{x,h}\right)$ | $Var\left(\mathcal{K}_{x,h}\right)$ |
|---|---|---|---|
| Poisson | $K_{p_0(x+h)}(y) = e^{-(x+h)}\frac{(x+h)^y}{y!}$ | $x+h$ | $x+h$ |
| Binomial | $K_{B(x+1,(x+h)/(x+1))}(y) = \frac{(x+1)!}{y!(x+1-y)!}\left(\frac{x+h}{x+1}\right)^y\left(\frac{1-h}{1+x}\right)^{x+1-y}\mathbf{1}_{\{y\leq x+1\}}$ | $x+h$ | $\frac{(x+h)(1-h)}{x+1}$ |
| Binomial négatif | $K_{BN\left(x+1,\frac{(x+1)}{(2x+1+h)}\right)}(y) = \frac{(x+y)!}{y!x!}\left(\frac{x+h}{2x+1+h}\right)^y\left(\frac{x+1}{2x+1+h}\right)^{x+1}$ | $x+h$ | $(x+h)+\frac{(x+h)^2}{x+1}$ |
| Triangulaire | $K_{T_{(a,h,x)}}(y) = \frac{(a+1)^h-\lvert y-x\rvert^h}{(2a+1)(a+1)^h-2\sum_{i=0}^{a} i^h}\mathbf{1}_{\{\lvert y-x\rvert<a\}}$ | x | $Var\left(K_{T_{(a,h,x)}}\right)$ |

TABLE 2.1: Quelques noyaux discrets usuels.

avec $\mathbf{1}(.)$ est la fonction indicatrice,

$$Var\left(K_{T_{(a,h,x)}}\right) = \frac{1}{P(a,h)}\left\{\frac{a(2a+1)(a+1)^{h+1}}{3} - 2\sum_{i=0}^{a} i^{h+2}.\right\}$$

et $P(a,b)$ est la constante de normalisation donnée par :

$$P(a,b) = (2a+1)(a+1)^b - 2\sum_{i=0}^{a} i^b,$$

### 2.3.3 Choix de paramètre de lissage

Dans cette section nous présentons quelques méthodes classiques pour le choix du paramètre de lissage dans l'estimation des fonctions discrètes.

1. **Minimisation du MISE**

   Soit $X = (X_1, \ldots\ldots, X_n)$ un $n$ échantillons fixé $iid$ de distribution inconnue $f$ alors l'erreur quadratique intégrée (ISE) est donné par :

$$ISE = \sum_{x\epsilon\mathbb{N}}(\hat{f}(x) - f(x))^2 = ISE(n,h,K,f), \tag{2.15}$$

ainsi (2.15) conduit à choisir une fenêtre adéquate :

$$h^* = \arg\min_h ISE(X, h, K, f), \tag{2.16}$$

pour laquelle la mesure est sur un seul échantillon et la fenêtre optimale $h^*_{opt}$ peut être obtenue, dans le cas de plusieurs échantillons, à travers $h^*_{opt} = \arg\min_{h>0} E(ISE(X, h, K, f))$
Ces techniques ont été développées et détaillées par Kokonendji et Senga Kiessé [12] ; Kokonendji et Senga Kiessé [11].

2. **Validation croisée**

Nous proposons ici deux techniques qui se basent sur la méthode de validation croisée. Plus précisément, comme cité auparavant, est d'estimer la densité $f$ au point $x_i$ par la technique de validation croisée dont la forme est donner par :

$$\hat{f}_i(X_i) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} K_{X_i,h}(X_j),$$

(a) Validation croisée par les moindres carrées : le principe de cette méthode consiste à estimer le $ISE$ par la technique de validation croisée et par la suite de sélectionner le paramètre de lissage qui minimise l'estimateur en question.
La fenêtre optimale, dans ce cas, s'obtient par :

$$\begin{aligned} h_{cv} &= \arg\min_{h>0} CV\,(h) = \arg\min_{h>0} \left[ \sum_{x\epsilon\mathbb{N}} \left\{ \hat{f}(x) \right\}^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_i(X_i) \right] \\ &= \arg\min_{h>0} \left[ \sum_{x\epsilon\mathbb{N}} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) \right\}^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,\ j\neq i}^{n} K_{X_i,h}(X_j) \right]. \end{aligned} \tag{2.17}$$

(b) Validation croisée par le maximum de vraisemblance.

Ce critère consiste à choisir $h$ qui maximise la fonctionnelle

$$LCV(h) = \prod_{i=n}^{n} \hat{f}_i(X_i),$$

ou encore

$$LCV(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\hat{f}_i(X_i)\right). \tag{2.18}$$

Ainsi, on détermine une fenêtre optimale $h_{LCV}$ par :

$$h_{LCV} = \arg\max_{h>0} LCV(h).$$

3. **Excès des zéros** Cette technique pour le choix de la fenêtre repose sur une particularité des données de comptage qui n'est autre que l'excès des zéros dans l'échantillon, c'est-à-dire de choisir une fenêtre adaptés $h_0 = h_0(X, K)$ tel que $h$ satisfait :

$$\sum_{i=1}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) = n_0, \tag{2.19}$$

où $n_0 = card\{X_i = 0\}$ désigne le nombre des zéros dans l'échantillon $X_1, ....., X_n$ à condition que $n_0 > 0$. Ci-dessous quelques exemples de $h_0$ :

- Si le noyau utilisé est Poissonnien alors $h_0 = \log\left(\frac{1}{n_0}\sum_{i=1}^{n} e^{-X_i}\right)$.
- Si le noyau utilisé est Binomiale alors le $h_0$ est la solution de $n_0 = \sum_{i=1}^{n}\left(\frac{1-h_0}{X_i+1}\right)^{X_i+1}$.
- Si le noyau utilisé est Binomiale négative alors le $h_0$ est la solution de $n_0 = \sum_{i=1}^{n}\left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1}$.

**Remarque 1** *Le paramètre de lissage sélectionné par la méthode Excès des zéros n'existe pas toujours. En effet, pour certains noyaux l'équation 2.19 n'admet pas de solution. A titre d'exemple on peut cité le cas du noyau triangulaire (pour plus de détails voir Kokonendji et al [12] et Senga Kiessé [10]).*

# Conclusion

Pour les méthodes qui consiste à choisir le paramètre de lissage $h$ de sorte à minimiser le $MISE$ ou le $ISE$, l'intérêt est purement théorique. La méthode la plus répandue en

pratique est la technique de **validation croisée**, puisque elle est guidée seulement par les observations et n'utilise pas des approximations de $f$. Cependant la minimisation du critère $CV$ ne garantit pas l'existence d'un seul minimum i.e. la difficulté majeure de cette méthode est bien que le problème des minimums locaux. Dans le chapitre prochain, nous allons mettre en évidence et illustrer à travers des échantillons artificielles (simulation) ce phénomène de plusieurs minimums (minimum local) et cela dans le cas de noyaux classique et dans le cas de noyaux asymétriques (continus et discrets).

# Chapitre 3

# Phénomène des optimums locaux dans les méthodes $UCV$ et $LCV$

## Introduction

Dans ce chapitre notre objectif est d'illustrer, à base des échantillons simulés, le phénomène des optimums locaux dans les méthodes de validation croisée ($UCV$ et $LCV$) pour le choix du paramètre de lissage lors de l'estimation d'une densité de probabilité par la méthode du noyau, où nous allons considerer trois situations, à savoir : cas de densités à support continu et non-borné ($x \in \mathbb{R}$), cas de densités à support continu positif ($x \in \mathbb{R}_+$) et cas de densités à support discret non-borné ($x \in \mathbb{N}$).

## 3.1 Présentation de l'application

Afin de répondre à notre objectif nous avons implémenter un programme Matlab dont les principales étapes sont :

1. Fixer les paramètres $f$, $K$.
2. Générer 100 échantillons de taille $n$ à partir de la densité $f$.

3. Calculer $UCV(h)$ et $LCV(h)$ pour chaque échantillon.

4. Présenter graphiquement les quantités $UCV(h)$ et $LCV(h)$.

Il est à noter qu'au niveau de l'étape 3, lorsque la densité est à support continu (borné ou non borné), nous utilisons la méthode des trapèzes pour le calcul des intégrales intervenant dans la quantification de l'$UCV(h)$.

**Cas de noyaux symétriques :** Pour l'application numérique, dans ce cas, nous avons considéré les quatre distributions cibles suivantes :

$$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}, \quad x \in \mathbb{R}. \tag{3.1}$$

$$f_2(x) = 0.5\ \frac{1}{\sqrt{2\pi}} e^{\frac{(x+1)^2}{2}} + 0.5\ \frac{1}{(3/2)\sqrt{2\pi}} e^{\frac{(x-3)^2}{9/2}}, \quad x \in \mathbb{R}. \tag{3.2}$$

$$f_3(x) = 0.5\ \frac{1}{(3/2)\sqrt{2\pi}} e^{\frac{(x+3)^2}{9/2}} + 0.5\ \frac{1}{\sqrt{2\pi}} e^{\frac{(x-1)^2}{2}}, \quad x \in \mathbb{R}. \tag{3.3}$$

$$f_4(x) = 0.5\ \frac{1}{\sqrt{2\pi}} e^{\frac{(x+1)^2}{2}} + 0.5\ \frac{1}{\sqrt{2\pi}} e^{\frac{(x-1)^2}{2}}, \quad x \in \mathbb{R}. \tag{3.4}$$

Les courbes de ces densités tests sont présentées dans la figure suivante :

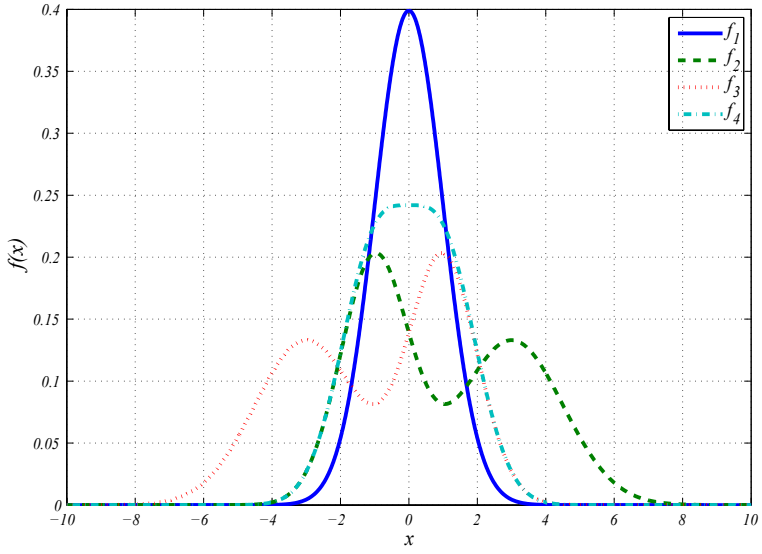


FIGURE 3.1 – Forme des densités cibles cas : continu à support non-borné

Pour le choix du noyau, dans cette situation nous avons opter pour le noyau gaussien

et le noyau d'Epanechnikov.

**Cas de noyaux asymétriques continus :** Pour l'application numérique, dans ce cas, nous avons considéré les quatre distributions suivantes :

- Une loi exponentielle de paramètre $\lambda = 1$ :

$$f_5(x) = \lambda e^{-\lambda x}, \quad x \geq 0. \tag{3.5}$$

- Une loi de Log-Normale de paramètres $(\mu, \sigma) = (0, 1)$ :

$$f_6(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln(x)-\mu)^2}{2\sigma^2}}, \quad x \geq 0. \tag{3.6}$$

- Une loi normale tronquée de paramètres $(\mu, \sigma) = (0, 1)$ :

$$f_7(x) = \frac{1}{\Phi(\mu/\sigma)\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad x \geq 0. \tag{3.7}$$

avec $\Phi(.)$ est la fonction de répartition de la loi normale centrée réduite.

Les courbes de ces densités tests sont données dans la figure suivante :

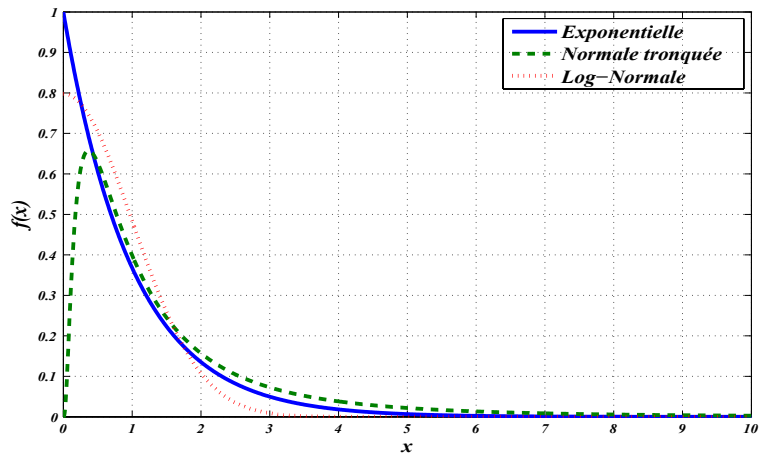


FIGURE 3.2 – Forme des densités cibles cas : continu à support positif

Les estimateurs dans ce passage sont conçus via les deux noyaux gamma et gamma modifié.

**Cas de noyaux asymétriques discrets :** Pour l'application numérique, dans ce cas, nous avons considéré les deux distributions suivantes :

- Une loi de Poisson de paramètre $\lambda = 3$ :

$$f_8(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{N}. \tag{3.8}$$

- Une loi Géométrique de paramètre $p = 0.3$ :

$$f_9(x) = (1-p)^x p, \quad x \in \mathbb{N}. \tag{3.9}$$

et pour le choix du noyau $K \in \{\text{Poisson, Biomiale Négative}\}$. Les diagrammes de ces densités tests sont présentés dans la figure 3.3.

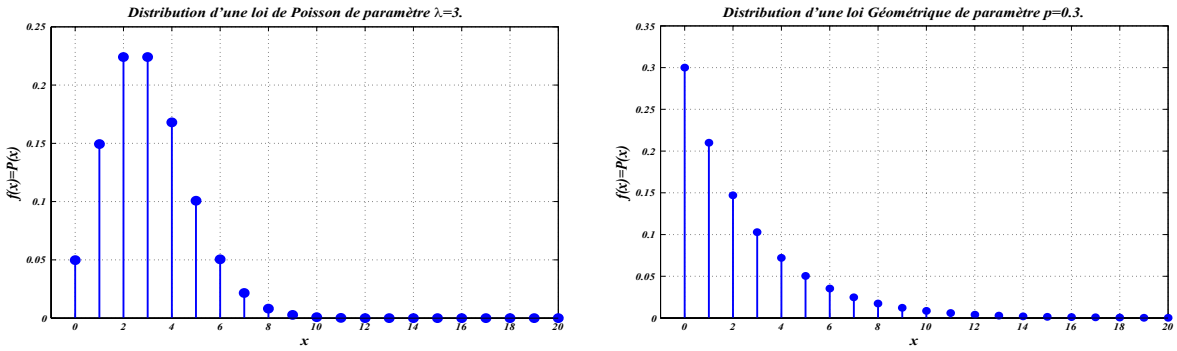


FIGURE 3.3 – Forme des densités cibles cas : discret

## 3.2 Résultats et discussion

Un échantillon de variation des expressions (1.16), (1.20), (2.10), (2.11), (2.17) et (2.18) en fonction du paramètre de lissage fournies par notre programme pour les paramètres précédents sur des échantillons de taille $n = 25$ est présenté dans les figures suivantes.

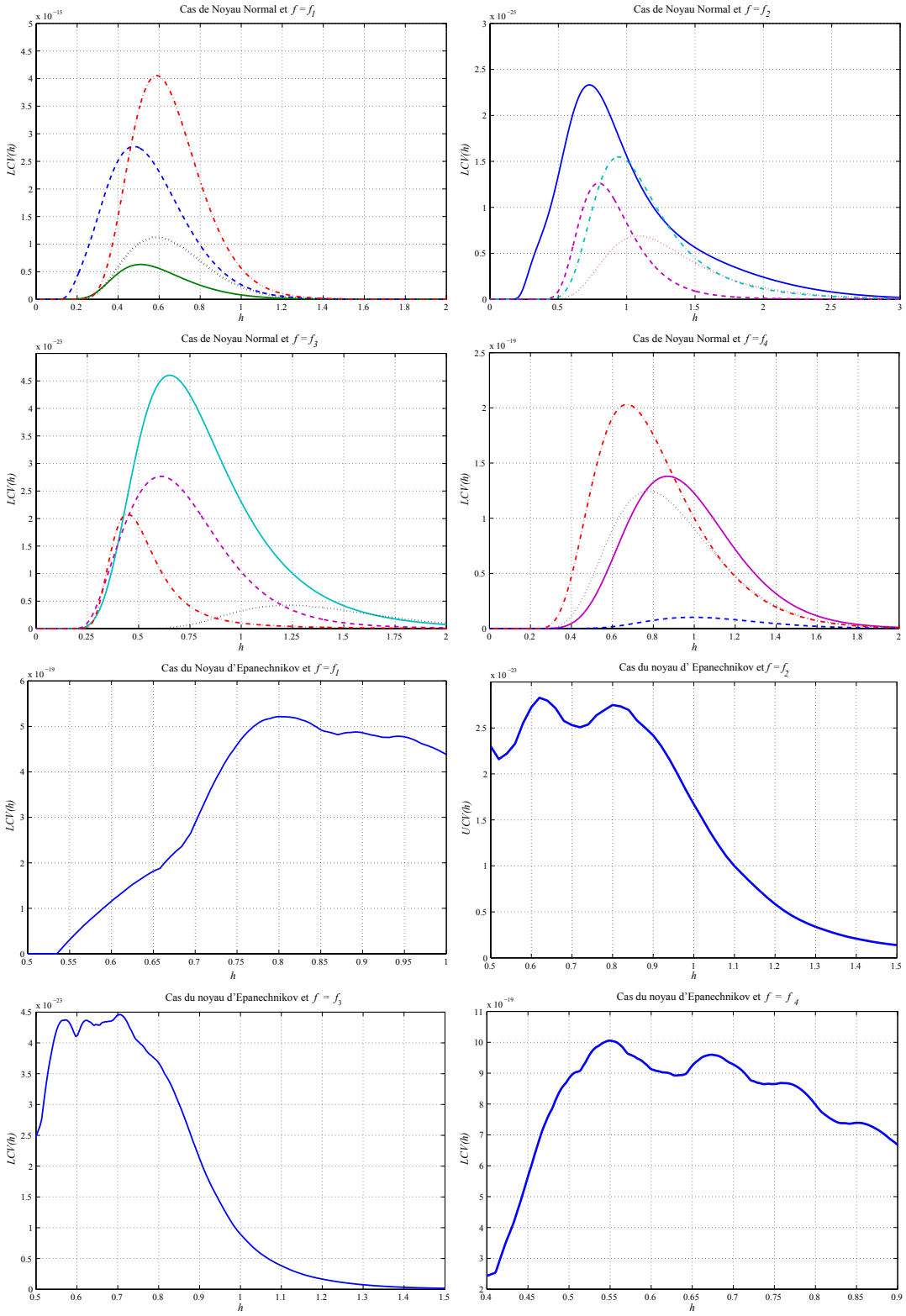


FIGURE 3.4 – Échantillon de variation de $LCV(h)$ en fonction de $h$ : cas de densités à support réel non borné.

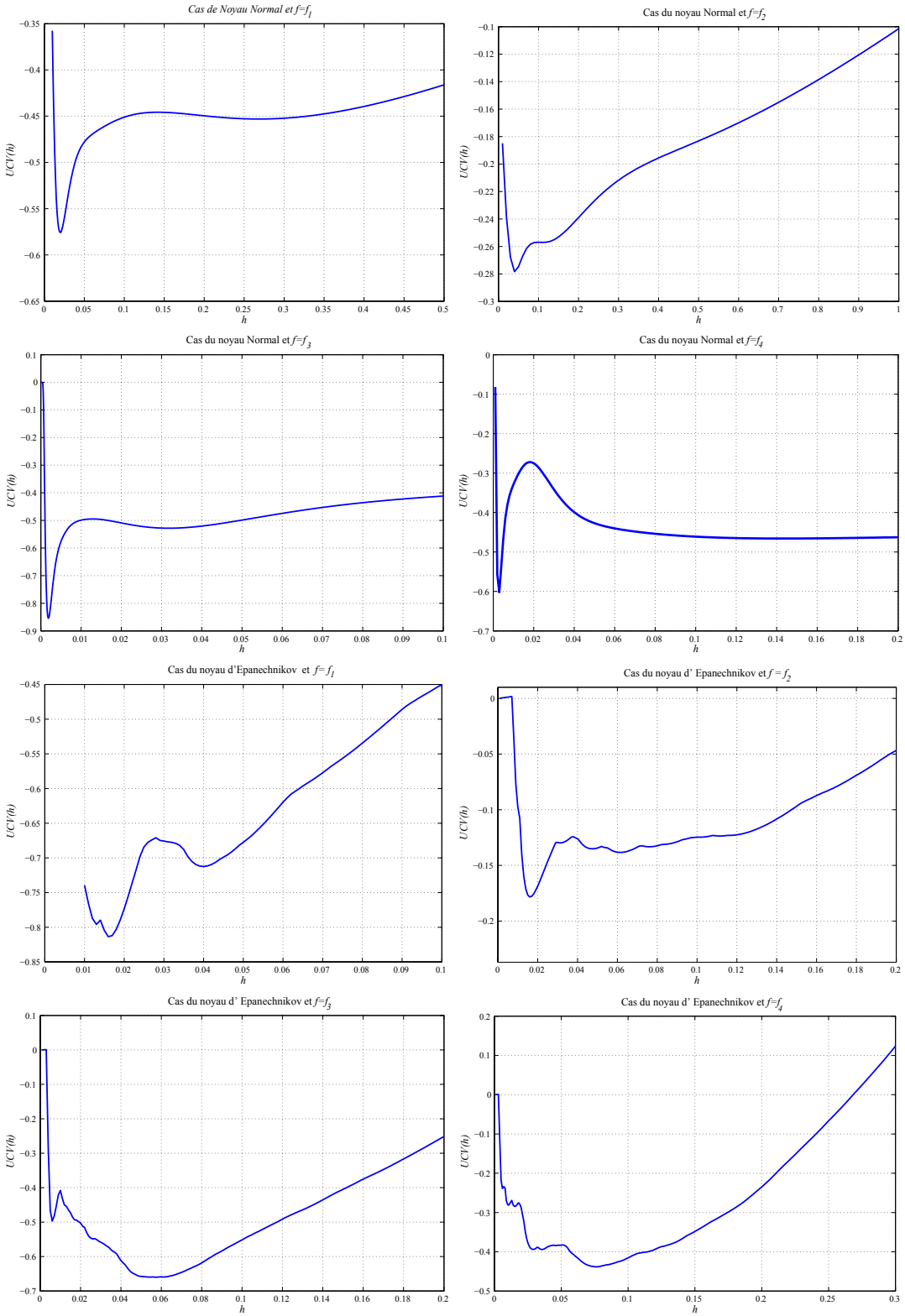


FIGURE 3.5 – Échantillon de variation de $UCV(h)$ en fonction de $h$ : cas de densités à support réel non borné.

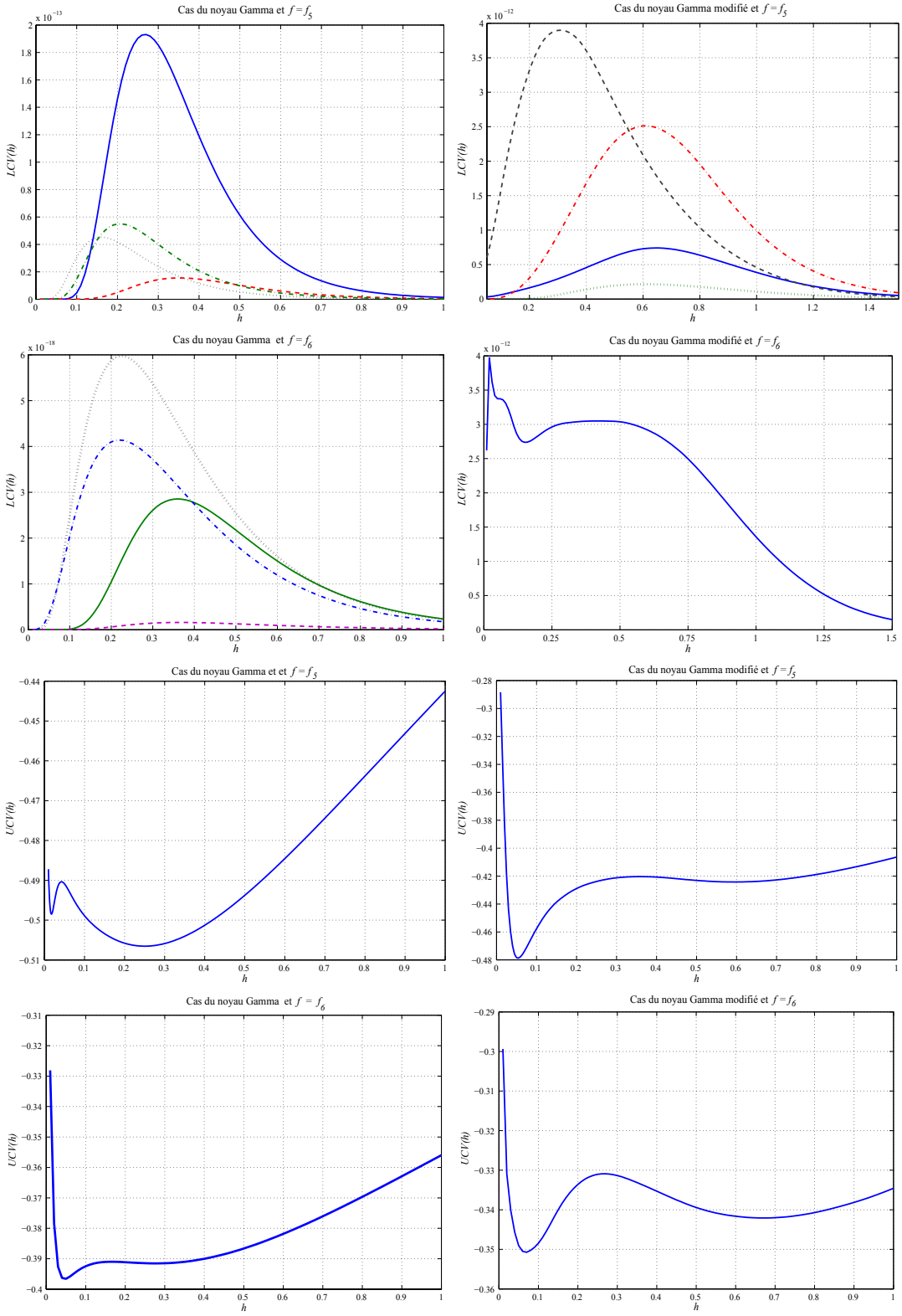


FIGURE 3.6 – Échantillon de variation de $UCV(h)$ et $LCV(h)$ en fonction de $h$ : cas de densités à support réel borné.

A partir des résultats obtenus on constate que :

- Dans le cadre d'estimation à noyaux classiques la méthode $UCV$ peut nous fournir des paramètres de lissage optimaux locaux qui peut influer la qualité de l'estimateur conçu dans ce cas.
- Le problème du phénomène des optimums locaux est rare dans le cas de la méthode $LCV$. D'autre part, le problème de la méthode $LCV$ réside essentiellement dans la valeur de l'optimum globale qui est très grande d'où le paramètre de lissage sélectionné par cette technique nous fournit des estimateurs sur-lissés (voir section 1.5).
- Le phénomène des optimums locaux, dans la méthode $UCV$, dépend de la nature de la densité cible à estimer où le phénomène est très fréquent dans le cas de densité à support réel ($x \in \mathbb{R}$) et cette dernière réduit dans le cas de densité définies sur $\mathbb{R}^+$. Dans le cas de densités discrètes ($x \in \mathbb{N}$) sur 100 échantillons de taille $n = 25$, on n'a rencontré aucun cas où il existe plusieurs optimums.
- Le choix du noyau, lorsque le paramètre de lissage est sélectionné par la méthode $UCV$, a un impact sur la fréquence du phénomène des optimums locaux. En effet, dans le cas de noyaux symétriques, l'utilisation du noyau d'Epanechnikov, pour $n = 25$, la totalité des échantillons générés montre l'existence de plusieurs optimums. Tandis que lorsqu'on utilise le noyau gaussien certains cas sont exempts de ce phénomène. Dans le cas d'utilisation des noyaux gamma, le phénomène persiste beaucoup plus dans le cas du noyau gamma que celui du noyau gamma modifié.
- Certes on a signalé auparavant que le phénomène des optimums locaux, dans la méthode $UCV$, dépend du type de support de la densité cible et du noyau utilisé pour la construction de l'estimateur mais le paramètre essentiel qui contrôle le phénomène en question est bien que la taille de l'échantillon. En effet, on peut passer d'une grande fréquence d'existence de ce phénomène, dans une certaine situation, à une fréquence nulle simplement en augmentant la taille de l'échantillon. A titre d'exemple, sur 100 échantillons de taille $n = 200$ on n'a rencontré aucun cas d'existence de plusieurs optimums

et cela dans la totalité des situations considérées.

# Conclusion

Dans ce chapitre, à base des exemples numérique nous avons illustré deux points essentiels :

1. Le phénomène des optimums locaux, dans le cadre de la méthode $UCV$ pour la sélection du paramètre de lissage, ainsi que les paramètres influant sur la fréquence de son existence.
2. Le problème majeur de la méthode $LCV$ est qu'elle a tendance à nous fournir des paramètres de lissage à grandes valeurs autrement dit le problème de cette méthode est le phénomène de sur-lissage des estimateurs.

# Conclusion générale

Le paramètre de lissage est un facteur important et crucial dans l'estimation de la fonction de densité par la méthode du noyau le fait que de petites ou de grandes valeurs de $h$ peuvent conduire à une estimation sous ou sur-lissée.

Dans ce mémoire nous avons considéré l'étude d'une catégorie de procédures de sélection du paramètre de lissage dans l'estimation de la fonction de densité par la méthode du noyau qui est d'une grande importance dans la pratique, à savoir : les méthodes de validation croisée. Plus précisément, notre intérêt est d'illustrer par simulation le phénomène des minimums locaux dans les deux méthodes de validation croisée $UCV$ et $LCV$ toute en prenant en considération le support de la densité cible (support réel non-borné, support réel semi-borné, et support discret).

Dans un premier lieu, la notion originale de l'estimateur à noyau d'une densité de probabilité, et qui se base sur des noyaux symétriques, à était mis en évidence et cela en introduisant sa forme, ses propriétés, le choix du noyau ainsi que certaines des procédures de sélection du paramètre de lissage proposées dans la littérature.

Dans un second lieu, après avoir souligner l'inconvénient de la version originale de l'estimateur à noyau dans le cas de densités à support borné au moins d'un côté sa nouvelle version qui se base principalement sur les noyaux asymétriques a été présenté. En particulier, nous avons mis l'accent sur les noyaux gamma qui sont adéquats pour le cas de densités à support réel positif ($x \in \mathbb{R}^+$) et les noyaux discrets destiner à l'estimation des densités définies sur l'ensemble des nombre naturelle.

Finalement, l'étude de simulation réalisée dans ce mémoire montre d'une part que le phénomène des optimums locaux dans les méthodes de validation croisée est très fréquent dans la méthode $UCV$ que dans la méthode $LCV$, de plus la fréquence de ce phénomène dépend de plusieurs paramètres, à savoir : la distribution réelle de l'échantillon et son type (à support borné ou non, continu ou discrète,...), la taille de l'échantillon et le noyau utilisé pour la construction de l'estimateur. D'autre part, le problème réel de la méthode $LCV$ n'est pas celui des optimums locaux mais plutôt c'est le phénomène des estimateurs sur-lissé car cette technique nous fournis des paramètres de lissage à grandes valeurs.

Il sera intéressant de compléter ce travail par :

- Réaliser une simulation extensive toute en considérant d'autres noyaux et d'autres lois.
- Revoir le présent travail lorsque nous considérons d'autres méthodes de validation croisée telle la $BCV$.
- Revoir le même travail afin évaluer la fréquence de l'existence du phénomène.

# Bibliographie


[1] T. Bouezmarni, and O. Scaillet, (2003). Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. Econometric Theory, 21, 390–412.

[2] Bowman, A. W. (1984) *An alternative method of cross-validation for the smoothing density estimates*. Biometrika **71** : 553 – 560.

[3] Burman, P. (1985) *A Date Dependent Approach to Density Estimation*. of Zeitschrift Für Wahrscheinlichke its theorie and Verwandte Gebiete **69** : 609 – 628.

[4] Chen, S. X. (1999). Beta kernel estimators for density functions. Computational Statistics & Data Analysis, 31(2), 131-145.

[5] Chen, S. X. (2000). Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics, 52(3), 471-480.

[6] Devroye, L. (1983) *The Equivalence of Weak, Strong and Complete Convergence $L^1$ for Kernel Density Estimates*. The Annals of Statistics **11** : 896 – 904.

[7] Fernandez, M. and Monteiro, P. (2005). Central Limit Theorem for Asymmetric Kernel Functionals, Annals of the Institute of Statistical Mathematics, 57, 425-442.

[8] Hall, P. (1982) *Cross-validation in density estimation*. Biometrika **69** : 383 – 390.

[9] Hall, P., Marron, S. J. (1991) *Local minima in cross-validation function*. Journal of the royal statistical society **90** : 149 – 173.

[10] Kiessé, T. S. (2008). Approche non-paramétrique par noyaux associés discrets des données de dénombrement (Doctoral dissertation, Université de Pau et des Pays de l'Adour).

[11] Kokonendji, C. C., & Kiesse, T. S. (2011). Discrete associated kernels method and extensions. Statistical Methodology, 8(6), 497–516

[12] Kokonendji, C. C., Senga Kiessé, T., & Zocchi, S. S. (2007). Discrete triangular distributions and non-parametric estimation for probability mass function. Journal of Nonparametric Statistics, 19(6-8), 241-254.

[13] Nadaraya, E. (1965) *On nonparametric estimation density function and regression.* Theory Probab. P.P.L **10** : 186 − 190.

[14] Park, B. U., Marron, S. J. (1990) *Comparison of data-driven bandwidth selectors.* Journal of the American Statistical Association **85** : 66 − 72.

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

[16] Rosenblatt, M. (1956) *Remarks in some nonparametric estimates of a density function.* Ann. Math. Statist. **27** : 832 − 837.

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

[18] O. Scaillet, (2004). Density estimation using inverse and reciprocal inverse gaussian kernels. Journal of Nonparametric Statistics, 16, 217–226.

[19] Scott, D. W. (1985) *Averaged shift histograms : effective nonparametric density estimators in several dimensions.* The Annals of Statistics **13** : 1024 − 1040.

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

[21] T. Senga Kiessé, Approche non-paramétrique par noyaux associés discrets des données de dénombrement. Thèse de Doctorat, Université de Pau, France (2008).

[22] Sheather,S. J., Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. **B 53** : 683 − 690.

[23] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[24] Stone, C. (1984) *An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates*. The Annals of Statistics **12** : 1285 − 1297.

[25] Wahba, G. (1975) *Optimal properties of variable knot, kernel and orthogonal series methods for density estimation*. Ann. Stat. **3** : 15 − 29.

[26] Zougab, N. (2007) *étude comparative des méthodes de sélection du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau*. Thèse de Magister en Mathématiques Appliquées, Université de Béjaïa.

**Résumé**

L'objectif du présent travail est d'illustrer, à base des échantillons simulés, le phénomène des optimums locaux dans les méthodes validation croisée ($UCV$ et $LCV$) pour le choix du paramètre de lissage dans l'estimation d'une densité de probabilité par la méthode du noyau. L'application numérique réalisée, sur trois types de densités à savoir : densités à support continu et non-borné ($x \in \mathbb{R}$), densités à support continu positif ($x \in \mathbb{R}_+$) et densités à support discret non-borné ($x \in \mathbb{N}$), montre d'une part que le phénomène des optimums locaux dans ces techniques dépend du type de la densité, du noyau utilisé et surtout de la taille de l'échantillon. D'autre part, les paramètres de lissage sélectionnés par la méthode $LCV$ nous fournissent des estimateurs sur-lissés.

**Mots clés :** Estimation à noyaux, Paramètre de lissage, $UCV$, $LCV$, optimum local.

**Abstract**

The objective of the present work is to illustrate, based on the simulated samples, the phenomenon of local optimums in the cross validation methods ($UCV$ and $LCV$) for the choice of the smoothing parameter in the estimation of a probability density by the kernel method. The numerical application realized, on three types of densities namely : densities with unbounded continuous support ($x \in \mathbb{R}$), densities with positive continuous support ($x \in \mathbb{R}_+$) and densities with unbounded discrete support ($x \in \mathbb{N}$), shows on the one hand that the phenomenon of the local optimums in these techniques depends on the type of density, the kernel used and especially the size of the sample. On the other hand, the smoothing parameters selected by the $LCV$ method provide us with over-smoothed estimators.

**Key words :** Kernel estimation, Smoothing parameter, $UCV$, $LCV$, local optimum.