République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université M'hamed Bougara Boumerdes
Faculté des Sciences
Département des Mathématiques

Mémoire présenté

Pour l'obtention du diplôme de Master en Recherche Opérationnelle
Option : Recherche Opérationnelle et Mathématiques de la Gestion

Par : Hadj Amar Karima

et : Khalfi Narimane

Etude comparative des méthodes de sélection du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau

Soutenu devant le jury composé de :

Président	M^{me} M.BEN MANSSOUR	M.A.B	U.M.B.B.
Promoteur	M^{r} B.ISSAADI	M.A.A	U.M.B.B.
Examinateur	M^{r} F.CHEURFA	M.A.B	U..M.B.B.

Année Universitaire 2015 – 2016

Remerciements

Nous remercions tout d'abord *ALLAH* de nous avoir donné la santé, la volonté, la force et le courage pour pouvoir surmonter les moments difficiles , et atteindre nos objectifs.

Nous témoignons une reconnaissance particulière à notre promoteur monsieur Issaadi Bedredine . Pour ses commentaires, remarques et suggestions qui ont donné une autre dimension à notre travail.

Nous tennons aussi à le remercier pour toutes ses efforts pour bien mener à notre formation durant notre curçus.

Nous adressons également nos remerciments à tous les membres de jury , nous sommes très reconnaissant à leurs remarques et commentaires qui nous aidions beaucoup pour mieux présenter ce document.

Nous avons beaucoup de reconnaissances envers tous les enseignants du département de Mathématique de la Faculté des Sciences de Boumerdes .

Que tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail, trouvent ici nos sincères reconnaissances.

Dédicace

Je dédie ce modeste travail à mes très chers parents pour leur amour, leur sotien, pour le courage et la volenté qu'ils m'ont inculqués.

A mes chers frères et sœurs.

A mon fiancé et mes beaux parents.

A toute ma famille chacun par son nom.

A ma binôme Narimane.

A tout mes amies.

Karima

Dédicace

J'ai le plaisir de dédier ce modeste travail :

aux êtres qui me sont les plus chers au monde, à ceux que je serai toujours reconnaissante,que Dieu tout puissant vous garde et vous procure santé, bonheur et longue vie.

Mes très chers parents.

A mes chers frères Abdou et Aala.

A mes chers sœurs Amina et Ikrame.

A ma grand-mère que j'aime.

A mes oncles et mes tantes.

A ma binôme Karima.

A tout mes amies.

Narimane

Table des matières

Introduction générale

La théorie de l'estimation est une des préoccupations majeures des statisticiens. On trouve dans la littérature deux types d'approches d'estimations de la densité de probabilité : l'approche paramétrique et l'approche non-paramétrique.

L'approche paramétrique suppose que les données sont issues d'une loi de probabilité de forme connue dont seuls les paramètres sont inconnus. Son objectif est de connaitre la vraie valeur du paramètre ou plus généralement une fonction de cette valeur. Le principal inconvénient de cette approche est qu'elle nécessite une connaissance préalable du phénomène aléatoire considéré.

Pour pallier les insuffisances et les défauts des familles paramétriques, une seconde approche dite non paramétrique propose de laisser parler les données, sans spécifier au préalable de forme sur f. L'outil d'estimation non paramétrique nous est fourni par l'histogramme : une fois les données regroupées en classes de valeurs, les fréquences empiriques sont représentées par des aires rectangulaires dont les bases correspondent aux classes elles mêmes. L'histogramme convient bien pour des analyses relativement grossières.Néanmoins, ses discontinuités n'apparaissent pas très naturelles et, ce qui est plus grave, les points tombant près des bords d'une classe et ceux tombant près du milieu ne sont pas différenciés, ceci explique la variabilité des interprétations statistiques que l'on peut faire d'un histogramme suivant le choix de l'origine et des classes. Pour des densités raisonnablement lisses, l'histogramme apparaît donc comme un estimateur sévèrement limité.

Il existe d'autres méthodes non paramétriques plus robustes que la méthode par l'histogramme : la méthode d'estimation par les séries orthogonales et la méthode du noyau. Nous regarderons brièvement en quoi consiste la méthode d'estimation par l'histogramme et la méthode par les séries orthogonales et en détail l'estimateur par la méthode du noyau vu sa souplesse d'utilisation et ses propriétés de convergence. Le succès rencontré par l'estimateur

à noyau auprès de la communauté des utilisateurs peut essentiellement s'expliquer en trois points :

- D'abord l'expression théorique de l'estimateur est extrêmement simple puisque il s'écrit sous la somme de n variables aléatoires indépendantes et identiquement distribuées.
- Ensuite l'estimateur converge vers la densité f en de nombreux sens, en particulier au sens L_1. d'autres part, il est convergent dans tous les modes : en probabilité, en moyenne, presque sûrement et presque completement.
- Enfin, l'estimateur à noyau est flexible, dans la mesure où il laisse à l'utilisateur une grande latitude non seulement dans le choix du noyau K, mais encore dans le choix du paramètre de lissage h.

C'est *Rosenblatt* [19] en 1956, suivi de *Parzen* [18] en 1962, qui ont proposé une classe d'estimateurs à noyau d'une densité univariée. Les estimateurs à noyau sont des fonctions de deux paramètres K, appelé noyau, et h dit paramètre de lissage (largeur de fenêtre). *Rosenblatt* reprenait l'idée de Fix et $Hodges$ 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.

Avant de construire les estimateurs à noyaux de la densité, d'en mesurer les perfomances théoriques et, le cas échéant, d'identifier le meilleur, il est nécessaire de spécifier un critère d'erreur qui puisse être éventuellement optimisé. Citons par exemple l'erreur quadratique moyenne intégrée $MISE$ et l'intégrale de l'erreur quadratique ISE.

Les propriétés de convergence de l'estimateur à noyau ont été établies par *Parzen* [18], *Silverman* [28] et *Nadaraya* [16]. Les théorèmes relatifs à l'erreur quadratique moyenne et l'erreur quadratique intégrée moyenne ont été obtenus sous forme élémentaire par *Parsen* [18]. Enfin, c'est *Epanechnikov* en 1969 [8] qui s'est rendu compte de l'existence d'un noyau asymptotiquement optimal K_e. Mais l'erreur quadratique moyenne asymptotiquement intégrée varie peu en fonction du choix de K.

Si le choix du noyau n'est pas un problème dans l'estimation de la densité, il n'en est pas de même pour le choix de la largeur de fenêtre qui ne dépend que de la taille n de l'échantillon. Plusieurs travaux ont montré que les estimateurs peuvent changer dramatiquement pour de petites variations du paramètre de lissage. Actuellement, il n'existe pas de choix optimal pour ce paramètre de lissage. Le choix optimal qui minimise l'erreur relative globale (MISE) dépend de la dérivée seconde de la densité inconnue. Les auteurs se sont alors attachés à introduire des procédures de sélection automatiques et donc moins subjectives que le simple choix à l'oeil.
L'étude de ce problème a nourri une littérature abondante, notamment vers le milieu des années quatre-vingt.

Le premier objectif de ce mémoire est de faire le point sur les procédures de sélection du paramètre de lissage h. Plusieurs méthodes ont été proposées dans la littérature :

- La classe des méthodes dites plug-in (ré-injection) : la méthode plug-in itéré proposée par *Scott*, *Tapia* et *Thompson* [25], la méthode rule of thumb proposée par *Silverman* [28], la méthode surlissage (oversmoothing) proposée par *Scott* et *Terrell* [26], la méthode plug-in itéré moderne proposée par *Park* et *Marron* [17] ou encore la méthode de *Sheather* et *Jones* [27].
Le principe de ces méthodes repose sur l'estimation d'une quantité qui dépend de la dérivée seconde de la densité de probabilité inconnue f qu'on désir estimer.

-Les méthodes reposant sur la validation croisée. D'un point de vue pratique, un des principaux intérêts de ces méthodes est leur caractère direct. Contrairement aux méthodes plug-in, le paramètre de lissage estimé dépend des observations. *Habbema*, *Hermans* et *Vandenbroek* [12] ont proposé une méthode fondée sur un critère non asymptotique du maximum de vraisemblance. Une autre méthode dite, validation croisée non biaisée, a été proposée par *Rudemo* [21] et *Bowman* [2]. Mais cette méthode possède de nombreux défauts (convergence très lente vers l'optimal paramètre de lissage,...). Pour remidier aux problèmes de cette méthode, deux autres approches ont été développées, la validation croisée biaisée proposée par *Scott* et *Terrell* [26] et la validation croisée lissée introduite par *Hall*, *Marron* et *Park* [11].

Le second objectif de ce mémoire est d'appliquer et de comparer les différentes méthodes de sélection du paramètre de lissage sur plusieurs distributions connues. Toutes ces méthodes nous donnent un paramètre de lissage optimal pour la distribution à estimer. Elles diffèrent au niveau du choix du critère à optimiser.
Le mémoire est structuré comme suit :
Une introduction générale pour situer notre étude ;
Un premier chapitre où nous présentons les différentes méthodes non paramétriques d'estimation de la densité de probabilité tel que la méthode des histogrammes, la méthode des séries orthogonales et la méthode du noyau qui sera étudiée en détail.
Le deuxième chapitre traitera les principales méthodes de sélection du paramètre de lissage par la méthode du noyau.
Le troisième chapitre regroupera les résultats de simulation des différentes méthodes de sélection du paramètre de lissage.
Nous terminerons par une conclusion générale.

1

Estimation non paramétrique de la densité de probabilité

1.1 Introduction

Soit X une variable aléatoire de fonction de densité f. Le fait d'attribuer à X la fonction f, nous permet d'obtenir une description de la distribution de X et ainsi de calculer les probabilités associées à X par la relation

$$p(a < X < b) = \int_a^b f(x)dx$$

Supposons maintenant qu'il y'a n observations $x_1, x_2, \ldots, x_n$ issues d'une variable aléatoire réelle X de densité de probabilité réelle $f(x)$ inconnue .Il est donc intéressant de construire une estimation de cette densité, à partir de ces observations. Deux approches existent : l'approche paramétrique et l'approche non paramétrique.

Dans l'approche paramétrique, nous supposons que les données proviennent d'une famille de distributions paramétriques connue. L'estimation de la densité est obtenue en estimant les paramètres de la distribution à partir des données et en substituant ces estimateurs dans la formule de densité pour cette distribution. Par exemple, supposons que les données proviennent d'une distribution normale de moyenne μ et d'écart type σ . L'estimation de la densité f relative

à ces observations est obtenue en trouvant les estimateurs de μ et de σ à partir de ces données et en substituant ces derniers dans la formule de densité pour une distribution normale.

Par opposition l'estimation non paramétrique ne fait aucune hypothèse à priori sur l'appartenance de f à une famille de lois connues. Dans cette approche, ce sont les observations qui vont nous permettre de déterminer l'estimation de la densité f.

Dans ce chapitre nous allons présenter une étude des différentes méthodes d'estimation, la méthode d'estimation par histogramme (estimation naturelle de la densité de probabilité), l'estimation par séries orthogonales, en nous concentrant sur la méthode du noyau qui peut être vue comme une extension de la méthode d'estimation par histogramme. Nous présentons également, les propriétés statistiques de quelques méthodes d'estimation.

1.2 Définitions et critères d'erreur :

Définition 1 :

Soit une variable aléatoire X absolument continue de densité de probabilité $f(x)$ .L'espérance mathématique de X est définie par :

$$\mathbb{E}(X) = \int xf(x)dx$$

Définition 2 :

La variance d'une variable aléatoire X absolument continue est définie par :

$$\mathbb{V}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

Définition 3 :

On appelle moment d'ordre k d'une variable aléatoire X , le nombre m_k défini par :

$$m_k = \mathbb{E}(X^k) = \int x^k f(x)dx$$

Et on appelle moment centré d'ordre k d'une variable aléatoire X , le nombre μ_k défini par :

$$\mu_k = \mathbb{E}(X - \mathbb{E}(X))^k = \int (x - \mathbb{E}(X))^k f(x)dx$$

Définition 4 :

Un estimateur T_n est dit sans biais lorsque son espérance mathématique est égale à la vraie valeur du paramètre.

$$E(T_n) = \theta$$

Un estimateur T_n est dit asymptotiquement sans biais si le biais diminue avec l'augmentation de la taille de l'échantillon :

$$\lim_{n\to\infty} E(T_n) = \theta$$

Critères d'erreur ($MSE, MISE$ *et* ISE) :

L'évaluation de la similarité entre l'estimateur f_h et la vraie densité f à estimer, nécessite des critères d'erreur. La mesure la plus naturelle utilisée est la moyenne intégrée des erreurs quadratiques. Ainsi, on défini d'abord la moyenne des erreurs quadratiques (Mean Squared Error MSE)

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

Une mesure globale de l'efficacité de l'estimateur f_h est obtenue en intégrant le MSE . Il s'agit de l'Erreur Quadratique Moyenne Intégrée en anglais Mean Integrated Squared Error ($MISE$), elle s'écrit :

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

Notons qu'il est plus pratique d'utiliser l'intégrale des erreurs quadratiques (ISE) pour l'évaluation du $MISE$. Le ISE est défini par :

$$\begin{aligned} ISE(f_h) &= \int [f(x) - f_h(x)]^2 dx \\ &= \int f(x)^2 dx - 2\int f(x) f_h(x)dx + \int f_h^2(x)dx. \end{aligned}$$

1.3 Estimation par les séries orthogonales :

Les estimations par séries orthogonales s'approchent des problèmes d'estimation de la densité de probabilité de différents point de vue . Ils sont mieux expliqués par un exemple spécifique.

Supposons que nous voulons estimer une densité f sur l'intervalle [0, 1]. L'idée de la méthode de série orthogonale est alors d'estimer f par l'estimation des coefficients de son expansion Fourier [28].
Définissons $\phi_v(x)$ par :

$$\phi_0(x) = 1$$

$$\forall r = 1, 2, .. \; On \; a \begin{cases} \phi_{2r-1}(x) = \sqrt{2}\cos 2\Pi r x & ; \\ \phi_{2r}(x) = \sqrt{2}\sin 2\Pi r x & . \end{cases}$$

f peut être représentée par une série Fourier comme suite
$\sum_{v=0}^{\infty} f_v \phi_v$. Où, pour chaque $v \geq 0$:

$$f_v = \int_0^1 f(x)\phi_v(x)dx$$

Supposons que X est une variable aléatoire avec la densité f, alors f_v peut être écrite :

$$f_v = \mathbb{E}\phi_v(X)$$

Par conséquent, l'estimateur de f_v basé sur un échantillon $X_1, ..., X_n$ généré par f est

$$\hat{f}_v = \frac{1}{n}\sum_{i=1}^{n} \phi_v(X_i)$$

La somme $\sum_{v=0}^{\infty} \hat{f}_v \phi_v$ ne sera pas une bonne estimation de f, mais elle convergera à une somme de fonctions delta .

$$\hat{f}_v = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i)$$

où δ est la fonction de delta Dirac. Alors, pour chaque v,

$$\hat{f}_v = \int_0^1 w(x)\phi_v(x)dx.$$

$\hat{f}_v$ sont exactement les coefficients Fourier de la fonction w.
Pour obtenir une bonne estimation de la densité f il est nécessaire de lisser w. La façon la plus facile de le faire est de tronquer l'expansion $\sum \hat{f}_v \phi_v$, à un certain point, en choisissant un entier K et définir l'estimateur $\hat{f}$ par :

$$\hat{f}(x) = \sum_{v=0}^{K} \hat{f}_v \phi_v(x).$$

Le choix de la limite K détermine la quantité de lissage.
Une approche plus générale introduit des poids λ_v, qui satisfait $\lambda_v \to 0$ quand $v \to \infty$, pour obtenir l'estimateur :

$$\hat{f}(x) = \sum_{v=0}^{\infty} \lambda_v \hat{f}_v \phi_v(x). \tag{1.1}$$

1.4 Estimation par les histogrammes

L'estimateur le plus ancien pour estimer une densité est l'histogramme des fréquences. D'après [14], l'origine des histogrammes est attribuée à *John Graunt* au $XVII^{eme}$ siècle répondant à l'objectif d'une représentation de la distribution de données. À ce titre, il peut être considéré comme un estimateur de la densité de probabilité sous-jacente à un ensemble fini d'observations. Supposons que l'on ait n observations $x_1, ..., x_n$ issues d'une même loi de probabilité inconnue de densité f, où f est à support borné $]a_0, a_k]$. Estimer cette densité f par la méthode de l'histogramme revient à approcher f par une fonction en escaliers. Pour cela, on partitionne l'intervalle de référence $]a_0, a_k]$ en $k \in N$ classe de la forme $]a_{j-1}, a_j]$, $j \in \{1, ..., k\}$. La largeur de la classe j est alors $h_j = a_j - a_{j-1}$ (si les classes sont de même largeur $h = \frac{a_k - a_0}{k}$ et on dit que c'est un histogramme à pas fixe, et si les h_j sont différentes, on dit que c'est un histogramme à pas variable).
L'estimateur histogramme s'écrit alors : $\forall\ x \in]a_0, a_k]\ \exists\ j \in \{1, ..., k\}$, tel que $\forall\ x \in]a_{j-1}, a_j]$

$$f_h(x) = \frac{f_j}{a_j - a_{j-1}}$$

où f_j est la fréquence empirique du nombre d'observations appartenant à la classe correspondante, tel que

$$f_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]a_{j-1}, a_j]}(X_i)$$

L'estimateur peut aussi s'écrire de la forme suivante

$$f_h(x) = \sum_{j=1}^{k} \frac{f_j}{h_j} \mathbb{1}_{]a_{j-1}, a_j]}(x)$$

Ce qui est équivalent à

$$f_h(x) = \frac{1}{n}\sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \mathbb{1}_{]a_{j-1},a_j]}(X_i)}{h_j} \mathbb{1}_{]a_{j-1},a_j]}(x) \tag{1.2}$$

On constate que cet estimateur dépend de plusieurs paramètres :

- Choix de bornes inférieures et superieures ;
- Le nombre de classes ;
- La largeur de classes.

Il s'avère que de mauvais choix de ces paramètres donne des estimateurs de très mauvaises qualités. Pour simplifier les notations, on supposera maintenant que les classes sont de même largeur, c'est-à-dire que pour tout $j = 1, .., k$, $a_j - a_{j-1} = h$, il est aisé de remarquer que f_h est une densité de probabilité, si l'on pense à la convergence de cet estimateur, il est clair que f_h sera d'autant plus proche de la vraie densité f que les largeurs de classe seront plus étroites, d'où la nécessite d'imposer que $h \to 0$ quand $n \to \infty$ en revanche, il ne faut pas que h tende trop vite vers 0, sinon on pourrait avoir des classes sans aucune observation, est donc une fonction en escalier f_h avec des marches d'ordonné nulle, très éloignée de la réalité il faut donc que, h tende vers 0 avec $n \to \infty$, et que malgré cela il tombe de plus en plus d'observations dans chaque classe.

1.4.1 Propriétés statistiques de l'estimateur par histogramme

Nous présentons, quelques propriétés statistiques de l'estimateur par histogramme f_h .
En statistiques, il est nécessaire de mesurer la qualité d'un estimateur. Pour cela, on évalue, d'une part, l'écart entre la moyenne de l'estimateur et la densité à estimer, ce critère d'évaluation est appelé *biais*[2] , et d'autre part, la variance de l'estimateur (due au caractère aléatoire d'observations) qui caractérise la dispersion des valeurs de l'estimateur dans l'ensemble d'observations.

- Le biais de l'estimateur f_h est donné [29], pour tout $x \in]a_{j-1},a_j]$, $\forall \;\; j \in \{1, ..., k\}$ par :

$$\begin{aligned} Biais(f_h(x)) &= \mathbb{E}(f_h(x)) - f(x) \\ &= \frac{1}{2} f'(h - 2(x - a_j)) + o(h^2) \end{aligned}$$

- La variance de l'estimateur f_h est donné [29] par :

$$\begin{aligned} \mathbb{V}(f_h(x)) &= \mathbb{E}((f_h(x))^2) - \mathbb{E}((f_h(x)))^2 \\ &= \frac{f(x)}{nh} + o(n^{-1}) \end{aligned}$$

Afin d'apprécier la qualité de l'estimateur, il est usuel d'évaluer la distance entre l'estimateur et la densité à estimer. La distance la plus couramment utilisée est l'erreur quadratique moyenne MSE.

• Le MSE de l'estimateur f_h est donné [29], pour tout $x \in]a_{j-1},a_j]$ par :

$$\begin{aligned} MSE &= \mathbb{V}(f_h(x)) + Biais^2(f_h(x)) \\ &= \frac{f(x)}{nh} + \frac{f'(x)^2}{4}[h - 2(x - a_j)]^2 + o(n^{-1}) + o(h^3) \end{aligned}$$

• Le $MISE$ de l'estimateur f_h est donné [29], pour tout $x \in]a_{j-1},a_j]$ par :

$$MISE = \frac{1}{nh} + \frac{h^2 \int f'(x)^2 dx}{12} + o(n^{-1}) + o(h^3)$$

1.4.2 L'histogramme mobile

L'estimateur histogramme précédent f_h n'est pas un bon estimateur. Considérons la classe $C_j =]a_{j-1}, a_j]$, et imaginons que le point $x \in C_j$ où l'on veut estimer $f(x)$ par $f_h(x)$ se situe près de l'extrémité. Alors, toutes les observations de la classe C_j interviennent dans le calcul de f_h, mais on se rend compte qu'une observation située près de a_j sera prise en compte, alors qu'elle est assez éloignée de x, et qu'une observation située tout près de x dans la classe C_{j-1} n'entre pas dans le calcul de f_h. Pour remédier cet estimateur, on peut alors utiliser l'histogramme mobile, qui est un translaté de l'histogramme de manière à ce que l'observation x où l'on estime, se retrouve au centre d'une classe, plus précisément au centre de la classe $]x - h, x + h]$ où h désigne la demi-largeur d'une classe. L'estimateur histogramme mobile s'écrit alors :

$$f_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{]x-h,x+h]}(X_i)$$

Remarquons que

$$\begin{aligned} x - h < X_i \leq x + h \quad &\Leftrightarrow \quad -1 < \frac{X_i - x}{h} \leq 1 \\ &\Leftrightarrow \quad -1 < \frac{x - X_i}{h} \leq 1 \end{aligned}$$

On peut écrire alors :

$$f_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{]-1,1]}\left(\frac{x - X_i}{h}\right)$$

En posant $W(x) = \frac{1}{2}\mathbb{1}_{]-1,1]}(x)$, l'estimateur s'écrit alors :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right) \tag{1.3}$$

où W est appelée fonction poids.

1.5 Estimation par la méthode du noyau :

Le premier document publié qui a traité explicitement l'estimation de densité de probabilité est dû à *Rosenblatt* en 1956 [19] qui a proposé l'estimateur naïf, suivi de $Parzen$ en 1962 [18]qui définit une classe de fonctions K, appelées noyau pour désigner la fonction de densité que l'on utilise dans les méthodes non paramétriques, remplaçons du même coup le terme 'fonction-poids' (weight function) qui était généralement utilisé par *Rosenblatt*.

L'estimateur précédemment construit peut encore être amélioré. En effet, maintenant que la classe est centrée en x, on peut tout de même remarquer que pour l'estimateur histogramme mobile, toutes les observations de cette classe ont le même rôle dans le calcul de f_h. Il serait plus judicieux de penser que plus une observation est proche de x, plus elle doit intervenir dans le calcul de $f_h(x)$. L'idée la plus naturelle alors est de pondérer les observations en mettant d'autant plus de poids qu'on se trouve proche de x, et d'autant moins qu'on s'en trouve éloigné.

On a déjà vu un exemple de fonction de poids, notée W au paragraphe précédent. C'était une densité de probabilité correspondant à la loi uniforme sur l'intervalle $[-1, 1]$. Cette fonction de poids est trop brutale et elle ne répond pas à nos préoccupations. On choisira donc des fonctions de poids (noyaux) dans une classe plus large de densités, comprenant notamment des densités à support non borné, et ayant un seul mode à l'origine comme par exemple la loi normale centrée réduite.

1.5.1 Estimateur de Parzen-Rosenblatt

Soit (X_n) une suite de variables aléatoires indépendantes et de même loi, de densité de probabilité f. On veut estimer f à partir d'un échantillon (X_1,X_2.....X_n). Soit h le paramètre de lissage tel que

$$h(n) \to 0 \; et \; nh(n) \to \infty \; quand \; n \to \infty$$

On peut estimer f par l'estimateur de $Parzen - Rosenblatt$ [19] [18].

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

Généralement la fonction K appelée noyau est une fonction positive et bornée tel que :

- ▷ K est symétrique ,i.e, $K(u) = K(-u)$;
- ▷ $\int_{\mathbb{R}} K(u)du = 1$;
- ▷ $\int_{\mathbb{R}} uK(u)du = 0$;
- ▷ $\int_{\mathbb{R}} u^2K(u)du = \mu_2 < \infty$.

Cet estimateur a été largement étudié par de nombreux auteurs, citons par exemple $Wolverton$ et $Wagner$ (1969), $Roussas$ (2000, 2001)[20], $Bosq$ et $al.$(1999) et Lu (2001). Une première justification concernant la forme de l'estimateur de $Parzen - Rosenblatt$, a été donnée précédemment. En voici une seconde [7], basée sur la fonction de répartition empirique associée à (X_n).

Fonction de répartition empirique :

Fonction de répartition empirique :

$F_n : \mathbb{R} \to [0, 1]$ définie pour tout $x \in \mathbb{R}$ par $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i<x\}}$. On peut également écrire de manière équivalente

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

La fonction de répartition empirique F_n est un estimateur simple de F, cette fonction est un très bon estimateur de F.

Propriété de l'estimateur de la fonction de répartition :

On a :

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

- L'espérance de $F_n(x)$ est : $\mathbb{E}(F_n(x)) = F(x)$;
- la variance de $F_n(x)$ est : $\mathbb{V}(F_n(x)) = \frac{1}{n}[1 - F(x)]F(x)$;
- Le biais de $F_n(x)$ est : $Biais(F_n(x)) = \mathbb{E}(F_n(x)) - F(x) = F(x) - F(x) = 0$ donc $F_n(x)$ est un estimateur sans biais de $F(x)$;
- Le MSE de $F_n(x)$ est : $MSE(F_n(x)) = V(F_n(x)) = \frac{1}{n}[1 - F(x)]F(x)$.

On sait que :

$$\begin{aligned} f(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h} \\ &\simeq \frac{F(x+h)-F(x-h)}{2h} \textit{ pour } h \textit{ petit.} \end{aligned}$$

Rosenblatt(1956) a donné un estimateur de f, en remplaçant F par son estimateur F_n. D'où :

$$f_h(x) = \frac{F_n(x+h)-F_n(x-h)}{2h}$$

On obtient alors :

$$\begin{aligned} f_h(x) &= \sum_{i=1}^{n} \frac{\mathbb{1}_{\{x-h<X_i\leq x+h\}}}{2nh} \\ &= \frac{1}{2hn}\sum_{i=1}^{n} \mathbb{1}_{\{-1<\frac{x-X_i}{h}\leq 1\}} \\ &= \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \end{aligned}$$

Posons $u = \left(\frac{x-X_i}{h}\right)$
Alors $K(u) = \frac{1}{2}\mathbb{1}_{]-1,1]}(u)$ tel que K est le noyau uniforme .

1.5.2 Définition

Estimateur à noyau :

Un estimateur à noyau de la densité f est une fonction définie par :

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

où h est un paramètre appelé paramètre de lissage il dépend de n et il vérifie

$$h(n) \to 0 \textit{ lorsque } n \to \infty$$

et K est une densité de probabilité appelée noyau.

Noyau :

Un noyau K est une fonction mesurable, intégrable définie de $\mathbb{R} \to \mathbb{R}$ tel que $\int_{\mathbb{R}} K(x)dx = 1$

Remarques :

- Le noyau K détermine la forme de voisinage autour du point x et h controle la taille de ce voisinage.
- f_h possède les mêmes propriétés de continuité et de différentiabilité que le noyau K, par exemple si K est le noyau gaussien alors f admet des dérivées de tout ordre.

1.5.3 Noyaux usuels :

Noyau Uniforme (Rosenblatt) :[5]

$$K(x) = \begin{cases} \frac{1}{2}, & \text{Si } |x| \leq 1\,; \\ \\ 0, & \text{Sinon.} \end{cases}$$

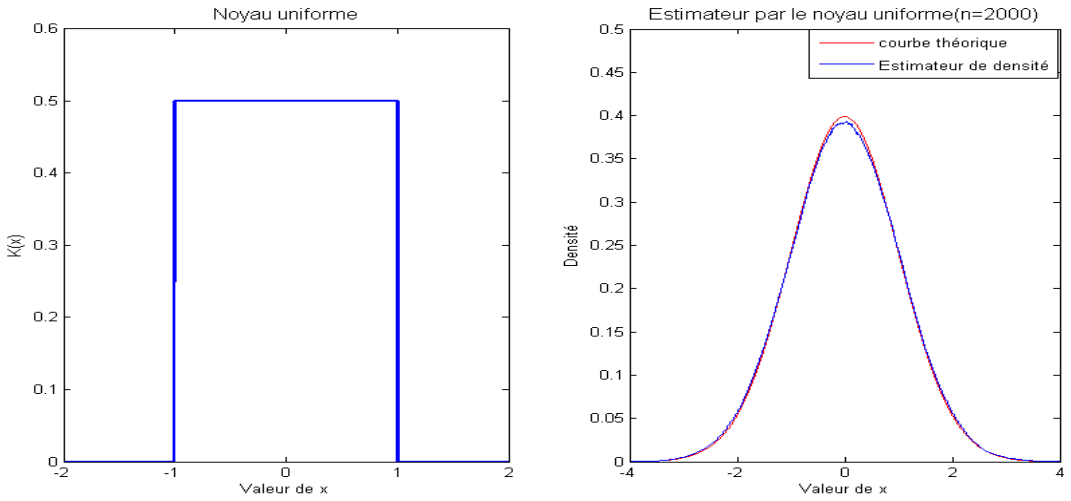

FIGURE 1.1 – Noyau Uniforme

Noyau triangulaire :[34]

L'avantage de ce noyau par rapport au précédent est sa continuité partout, ce qui conduit à un estimateur f_h continue. Ce noyau s'écrit sous la forme :

$$K(x) = \begin{cases} 1 - |x|, \ si \ |x| \ \leq 1 \\ 0, \ sinon \end{cases}$$

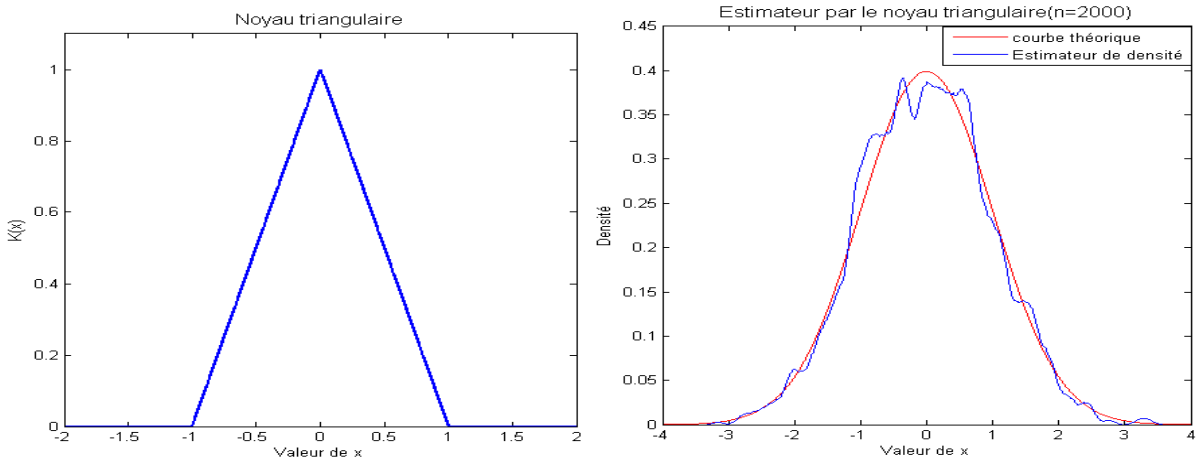

FIGURE 1.2 – Noyau Triangulaire

Noyau gaussien :

L'avantage du noyau gaussien est que plus la valeur de h est élevée plus on élargit la fenêtre, ce qui a un effet de lissage globale important ; mais le coût de calcul dans le cas de ce noyau est très élevé du fait de son support infini. Ce noyau s'écrit sous la forme :

$$K(x) = \frac{1}{\sqrt{2\Pi}} \exp\left(\frac{-1}{2}x^2\right), x \in \mathbb{R}$$

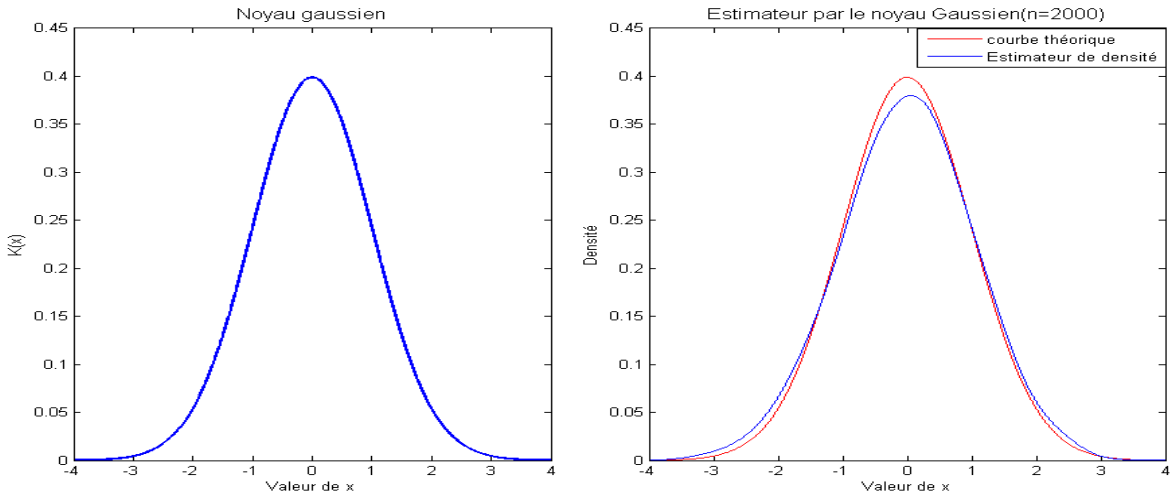

FIGURE 1.3 – Noyau Gaussien

Noyau quadratique (Biweight) :

Le noyau de biweight est très intéressant car il donne un estimateur dérivable partout. En fait, il s'agit du noyau le plus simple parmi les noyaux de forme polynômial dérivable partout. Ainsi, il assure le lissage locale de la fonction f_h. Ce noyau est d'une forme très proche du noyau gaussien, il est donc préférable de l'utiliser. il s'écrit sous la forme :

$$K(x) = \begin{cases} \frac{15}{16}(1-x^2)^2, \ si \ |x| \ \leq 1 \\ 0 \ , \ sinon \end{cases}$$

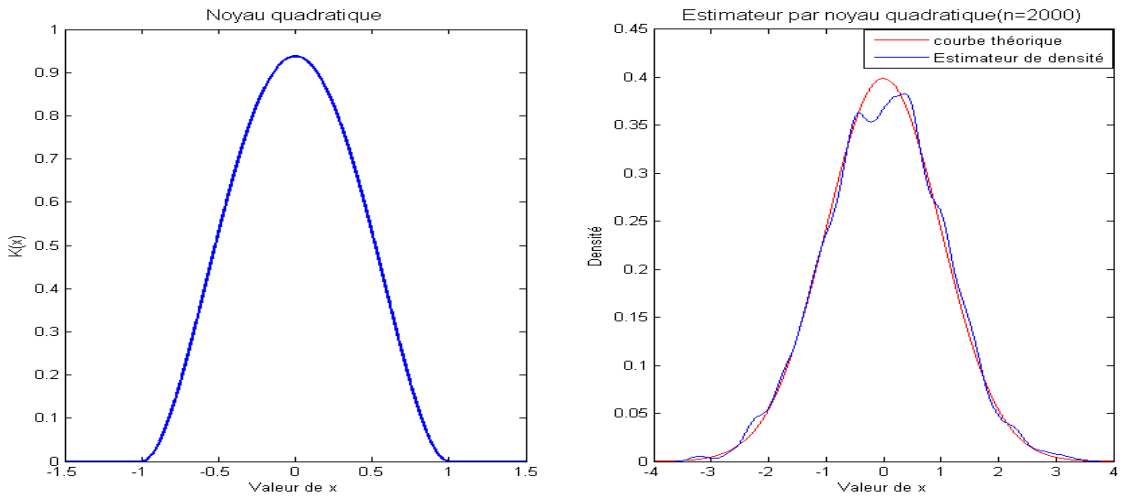

FIGURE 1.4 – Noyau quadratique (Biweight)

Noyau d'Epanechnikov ou parabolique :

En 1969, *Epanechnikov* [8], a donné la forme du noyau K_e défini par :

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

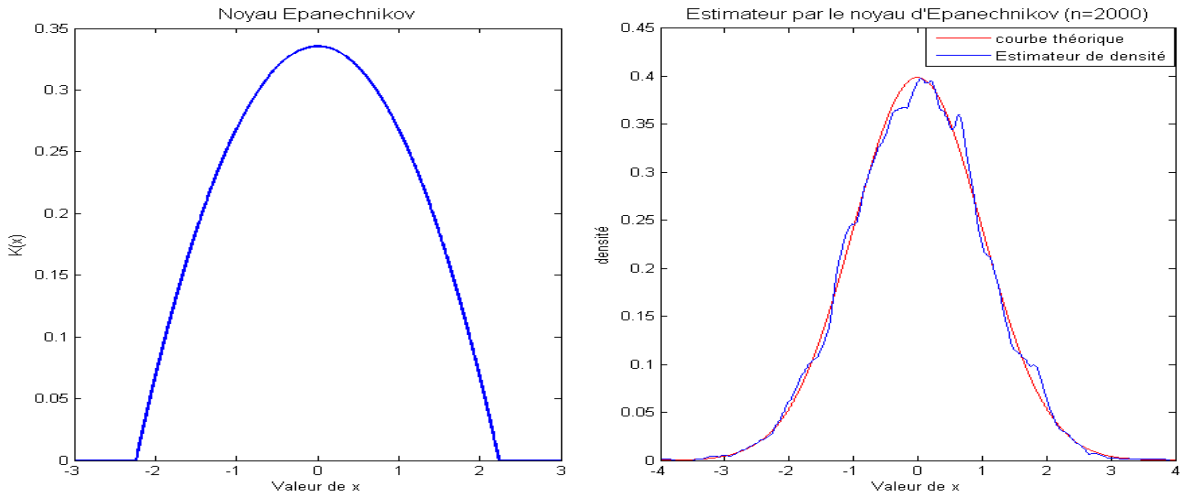

FIGURE 1.5 – Noyau d'Epanechnikov

Noyau sinus :

$$K(x) = \frac{1}{2\Pi} \left(\frac{\sin(\frac{x}{2})}{\frac{x}{2}} \right)^2, \ si \ x \neq 0$$

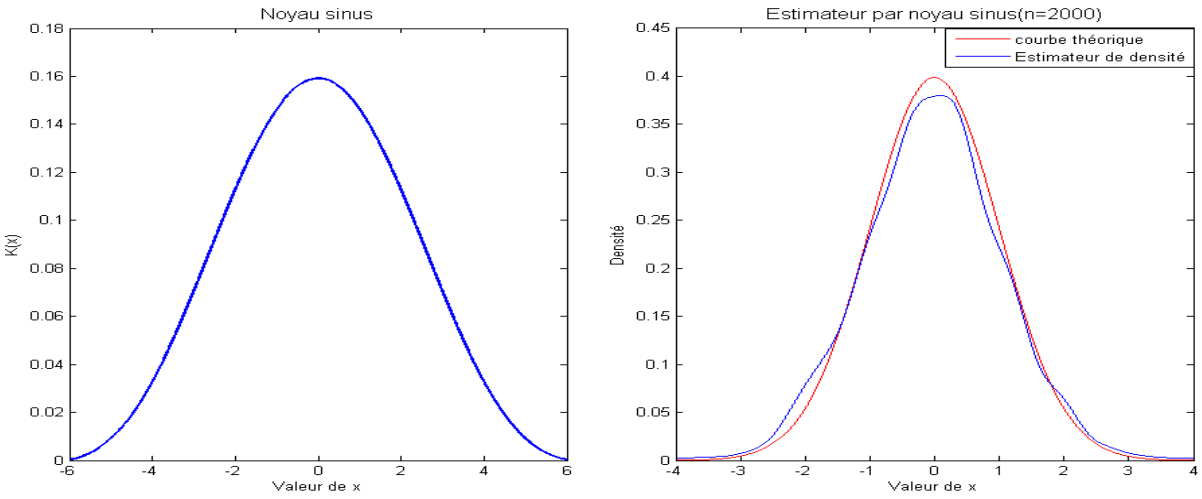

FIGURE 1.6 – Noyau Sinus

Noyau cosinus :[31]

$$K(x) = \begin{cases} \frac{\Pi}{4} \cos\left(\frac{\Pi x}{2}\right), \ si \ |x| \ \leq 1 \\ 0 \ , \ sinon \end{cases}$$

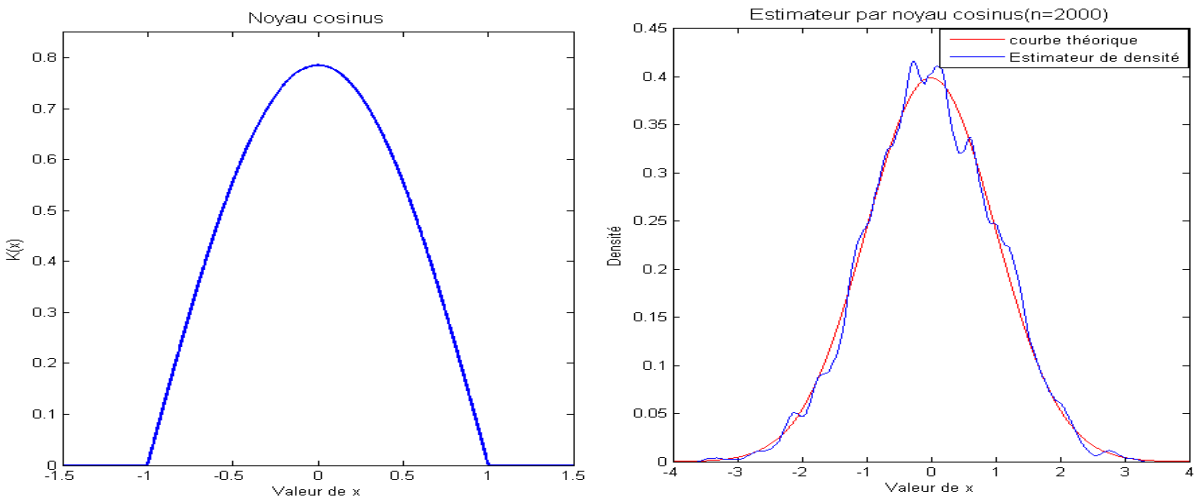

FIGURE 1.7 – Noyau Cosinus

Noyau de Silverman :[6]

$$K(x) = \frac{1}{2} \exp\left(\frac{-|x|}{\sqrt{2}}\right) \sin\left(\frac{|x|}{\sqrt{2}} + \frac{\Pi}{4}\right), x \in \mathbb{R}$$

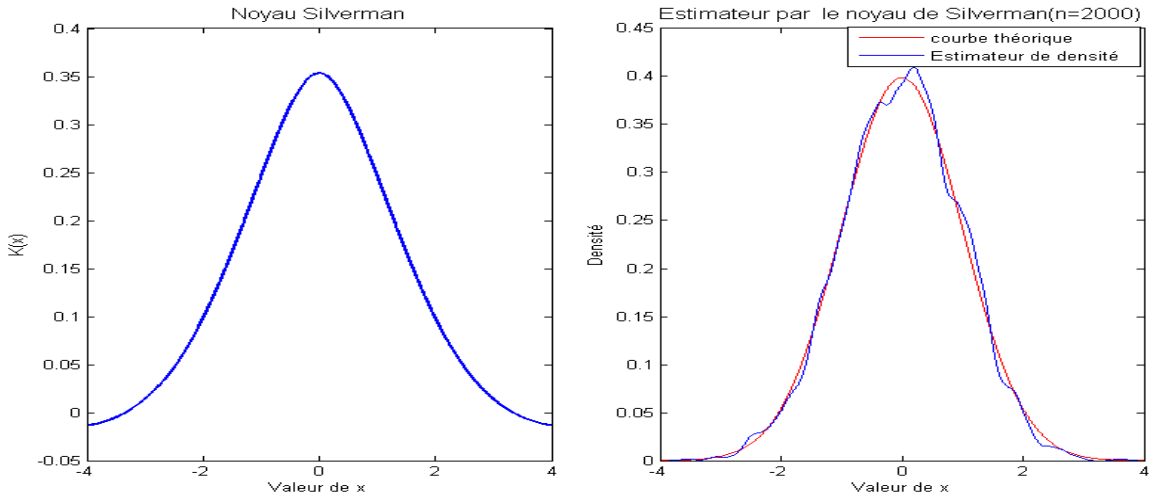

FIGURE 1.8 – Noyau Silverman

1.6 Propriétés d'un estimateur à noyau :

Nous présentons dans cette partie les propriétés statistiques de l'estimateur de densité f_h.

Lemme : Si le noyau K est positif et $\int K(u)du = 1$, alors l'estimateur f_h est une densité de probabilité. De plus, f_h est continue si K est continue.

Démonstration :

L'estimateur à noyau est positif et continu car la somme des fonctions positives et continues est elle-même une fonction positive et continue. Il faut donc vérifier que l'intégrale de f_h vaut un.
En effet,

$$\begin{aligned}
\int_{-\infty}^{+\infty} f_h(x) &= \int_{-\infty}^{+\infty} \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h}\Big) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K\Big(\frac{x-X_i}{h}\Big) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K(u)hdu \ (\textit{changement de variable } u = \frac{x-X_i}{h}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K(u)du \\
&= 1
\end{aligned}$$

1.6.1 Espérence , Biais et Variance de l'estimateur

Espérance de l'estimateur :

L'espérance mathématique de l'estimateur à noyau est définie par :[23]

$$\mathbb{E}(f_h(x)) = \ f(x) + \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2). \tag{1.5}$$

Démonstration :

$$\begin{aligned}
\mathbb{E}f_h(x) &= \mathbb{E}\Big[\tfrac{1}{nh}\sum_{i=1}^{n} K\Big(\tfrac{x-X_i}{h}\Big)\Big] \\
&= \tfrac{1}{h}\mathbb{E}\left(K\Big(\tfrac{x-X}{h}\Big)\right) \\
&= \tfrac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{x-t}{h}\right) f(t)dt.
\end{aligned}$$

en posant $u = \frac{x-t}{h}$:

$\mathbb{E}f_h(x) = \int_{-\infty}^{+\infty} K(u)f(x-hu)du$
D'après l'utilisation de la formule Taylor on obtient :

$$f(x-hu) = f(x) - huf'(x) + \frac{h^2u^2}{2!}f''(x) + o(h^2).$$

on obtient alors que :

$$\begin{aligned}
\mathbb{E}f_h(x) &= \int_{-\infty}^{+\infty} K(u)[f(x) - huf'(x) + \frac{h^2u^2}{2!}f''(x)]du + o(h^2) \\
&= f(x)\int_{-\infty}^{+\infty} K(u)du - hf'(x)\int_{-\infty}^{+\infty} uK(u)du + \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} u^2K(u)du + o(h^2)
\end{aligned}$$

On a :
$\int K(u)du = 1$, $\int uK(u)du = 0$ et $\int u^2K(u)du = \mu_2(K)$

Donc :

$\mathbb{E}f_h(x) = \ f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2).$

Biais de l'estimateur :

Le biais de l'estimateur $f_h(x)$ est :[23]

$$\begin{aligned} Biais f_h(x) &= \mathbb{E}f_h(x) - f(x) \\ &= f(x) + \tfrac{h^2}{2} f''(x)\mu_2(K) + o(h^2) - f(x). \end{aligned}$$

donc :

$$Biais f_h(x) = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2). \tag{1.6}$$

Remarque :

Le terme de droite de l'équation (1.6) est différent de zéro, ceci signifie que l'estimateur à noyau est un estimateur biaisé.

Le biais étant un $o(h^2)$, on voit que le biais optimal est un $o(n^{\frac{-2}{5}})$. Par conséquent, $f_h(x)$ est un estimateur asymptotiquement sans biais de $f(x)$ car $\lim\limits_{h\to 0} Biais f_h(x) = 0$, mais la convergence est lente car $n^{\frac{-2}{5}}$ tend lentement vers 0.

Variance de l'estimateur :

La variance de l'estimateur $f_h(x)$ [23] est définie par :

$$\mathbb{V}f_h(x) = \frac{1}{nh} f(x)R(K) + o\Big(\frac{1}{nh}\Big) \tag{1.7}$$

Démonstration :

$$\begin{aligned} \mathbb{V}f_h(x) &= \mathbb{V}\left(\tfrac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)\right) \\ &= \tfrac{1}{nh^2}\mathbb{V}\left(K\Big(\tfrac{x-X}{h}\Big)\right) \\ &= \tfrac{1}{nh^2}\left[\mathbb{E}\left(K\Big(\tfrac{x-X}{h}\Big)\right)^2\right] - \tfrac{1}{nh^2}\left[\mathbb{E}\left(K\Big(\tfrac{x-X}{h}\Big)\right)\right]^2 \\ &= \tfrac{1}{nh^2}\int_{-\infty}^{+\infty}\left[K\Big(\frac{x-t}{h}\Big)\right]^2 f(t)dt - \frac{1}{n}\left(\frac{1}{h}\int_{-\infty}^{+\infty} K\Big(\frac{x-t}{h}\Big)f(t)dt\right)^2. \end{aligned}$$

En utilisant la substitution $u = \frac{x-t}{h}$ on obtient :

$$\mathbb{V}f_h(x) = \frac{1}{nh}\int_{-\infty}^{+\infty} K^2(u)f(x-hu)du - \frac{1}{n}\left(\int_{-\infty}^{+\infty} K(u)f(x-hu)du\right)^2$$

Supposons que h est petit et que n est grand et développons $f(x-hu)$ en série de Taylor, on obtient :

$$\begin{aligned}\mathbb{V}f_h(x) &= \frac{f(x)}{nh}\int_{-\infty}^{+\infty} K^2(u)du + o(\frac{1}{nh}) \\ &= \frac{f(x)R(K)}{nh} + o(\frac{1}{nh})\end{aligned}$$

où $R(K) = \int_{-\infty}^{+\infty} K^2(u)du$.

Remarque :

La variance étant un $o((nh)^{-1})$, la variance optimale est un $o(n^{\frac{-4}{5}})$. Donc $f_h(x)$ est un estimateur convergent de $f(x)$, mais la convergence est plus lente car $n^{\frac{-4}{5}}$ tend plus lentement vers 0.

1.6.2 MSE et MISE de l'estimateur :

L'analyse de la performance de l'estimateur à noyau exige la spécification d'un critère d'erreur approprié, afin de mesurer l'erreur d'estimation à un simple point aussi bien que sur l'ensemble des points. Donc, nous étudierons dans un premier temps la proximité de notre estimateur f_h de la vraie densité f. L'estimateur f_h dépend du noyau K et du paramètre de lissage h, cette dépendance n'est généralement pas exprimée explicitement. Lorsque nous considérons l'estimation à un point, une mesure naturelle de la dispersion est l'erreur quadratique moyenne MSE définie par :

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

En reprenant l'expression du biais de l'estimateur f_h et la variance et en remplaçant dans l'expression MSE on obtient alors :

$$\begin{aligned}MSE(f_h) &= \frac{f(x)R(K)}{nh} + o(\frac{1}{nh}) + \left(\frac{h^2}{2}\mu_2(K)f''(x) + o(h^2)\right)^2 \\ &= \frac{R(K)}{nh}f(x) + \frac{h^4}{4}\mu_2^2(K)f''^2(x) + o(\frac{1}{nh} + h^4).\end{aligned}$$

Cependant, il peut être intéressant d'avoir une mesure globale de la précision de f_h comme estimateur de f au lieu d'avoir une mesure de précision à un point donné, notée $MISE$ est

considérée comme une mesure globale de la précision et est définie par [29] :

$$\begin{aligned} MISE(f_h) &= \int MSE(f_h) \\ &= \int \left(\frac{f(x)R(K)}{nh} + o(\frac{1}{nh}) \right) dx + \int \left(\frac{h^2}{2}\mu_2(K)f''(x) + o(h^2) \right)^2 dx \\ &= \frac{R(K)}{nh} + \frac{h^4}{4}\mu_2^2(K)R(f''(x)) + o(\frac{1}{nh} + h^4). \end{aligned}$$

où $R(f''(x)) = \int (f''(x))^2 dx$

1.6.3 Comportement asymptotique

Une approximation asymptotique de l'espérance de l'estimateur f_h est donnée sous les conditions suivantes sur f, h et K [33] :

1) la fonction de densité f admet la dérivé seconde f'' qui doit être une fonction absolument continue, de carré intégrable et monotone sur $(-\infty; -M)$ et $(M; +\infty)$ pour$M > 0$;

2) le paramètre de lissage h est positif et on suppose que h satisfait $\lim_{n\to\infty} h = 0$ et $\lim_{n\to\infty} nh = \infty$;

3) Pour que $f_h(x)$ soit une densité, on suppose que $K(u) \geq 0$ et $\int K(u)du = 1$, la fonction noyau est supposée être symétrique autour de zéro,soit $\int uk(u)du = 0$, et possède un moment d'ordre 2 fini,soit $\int u^2k(u)du < \infty$.

Nous avons établi que :

$$Biais f_h(x) = \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2)$$

$$\mathbb{V}f_h(x) = \frac{R(K)}{nh}f(x) + o(\frac{1}{nh})$$

Où $\mu_2(K) = \int u^2K(u)du$ et $R(g) = \int g^2(u)du$ pour une fonction g de carré intégrable.

Les expressions asymptotiques du biais et de la variance nous permettent de trouver des expressions asymptotiques pour MSE et $MISE$. Ces expressions ont été obtenues sous la condition (3) sur K et en supposant que la densité de probabilité f a des dérivées continues. On peut obtenir les approximations asymptotiques suivantes pour MSE et le $MISE$ sous des conditions appropriées d'intégrabilité de f et ses dérivées.
On note l'approximation asymptotique du MSE par :

$$AMSE(f_h) = \frac{R(K)}{nh}f(x) + \frac{h^4}{4}\mu_2^2(K)f''^2(x). \tag{1.8}$$

et l'approximation asymptotique du MISE par :

$$AMISE(f_h) = \frac{R(K)}{nh} + \frac{h^4}{4}\mu_2^2(K)R(f''(x)). \tag{1.9}$$

1.6.4 Criteres de convergence :

Parmi toutes les qualités que peut avoir un estimateur, on s'intéresse souvent à sa consistance, c'est à dire, au fait qu'un estimateur f_h converge ou non vers f. La convergence d'un estimateur peut être faible (en probabilité) ou forte (presque sûrement ou en moyenne quadratique).
On donne quelques résultats de convergence des estimateurs à noyaux de la littérature.

Convergence en moyenne quadratique :

Théorème : (Parzen)[18].
Soit f une densité continue et f_h son estimateur. Si le noyau K vérifie :

- $\int K(u)du = 1$ et $\int |K(u)|du < \infty$
- $\sup|K(u)| < \infty$ et $\lim\limits_{u\to\infty}|uK(u)| = 0$

Si le paramètre h satisfait :

$\lim\limits_{n\to\infty} h = 0$ et $\lim\limits_{n\to\infty} nh = \infty$

Alors f_h est un estimateur convergent en moyenne quadratique c'est à dire :

$$MSE(f_h(x)) \underset{n\to\infty}{\longrightarrow} 0.$$

Convergence en moyenne quadratique intégrée :

Théorème :(Parzen)[18]
Soit f une densité de puissance p^{eme}-intégrable et f_h son estimateur. Si le noyau K vérifie :

- $\int K(u)du = 1$ et $\int |K(u)|du < \infty$
- $\sup|K(u)| < \infty$ et $\lim\limits_{u\to\infty}|uK(u)| = 0$

Si le paramètre h satisfait :

$\lim\limits_{n\to\infty} h = 0$ et $\lim\limits_{n\to\infty} nh = \infty$

Alors f_h est un estimateur convergent en moyenne quadratique intégrée c'est à dire :

$$MISE(f_h(x)) \underset{n\to\infty}{\longrightarrow} 0.$$

Convergence uniforme en probabilité

Théorème(Parzen [18]).
Soit f la densité à estimer et f_h son estimateur, si les conditions suivantes sont vérifiées :

- $\int K(u)du = 1$ et $\int |K(u)|du < \infty$
- $\sup|K(u)| < \infty$ et $\lim\limits_{u\to\infty}|uK(u)| = 0$
- $\lim\limits_{n\to\infty} nh^2 = \infty$
- la transformée de Fourier $\mathcal{TF}(z) = \int \exp(-izu)K(u)du.$ est absolument intégrable,

alors f_h est convergent uniformément en probabilité c'est à dire :

$$\forall\varepsilon > 0, P\Big(\sup_{x\in\mathbb{R}}|f_h(x) - f(x)| < \varepsilon\Big) = 1.$$

Convergence uniforme presque complète

Théorème :(Nadaraya)[16].
Soit f une densité uniformément continue et son estimateur à noyau K positif et à variations bornées.
Si $\lim\limits_{n\to\infty} h = 0$ et $\sum\limits_{i=1}^{\infty}\exp(-\varepsilon nh^2) < \infty, \forall\varepsilon > 0$
alors f_h est convergent uniformément avec probabilité 1 c'est à dire :

$$\sup_{x\in\mathbb{R}}|f_h(x) - f(x)| \longrightarrow 0.$$

Théorème :(Silverman)[28]
Soit f une densité uniformément continue et f_h son estimateur à noyau K positif et à variations bornées.

Si $\lim\limits_{n\to\infty} h = 0$ et $\lim\limits_{n\to\infty}\frac{logn}{nh} = 0$

alors

$$\sup_{x\in\mathbb{R}}|f_h(x) - f(x)| \xrightarrow{p.s} 0.$$

où $\xrightarrow{p.s}$ est la convergence presque-sûre.

Convergence en loi

Ce dernier résultat est tiré des travaux de Parzen. Il montre que l'estimateur à noyau est asymptotiquement normale.

Théorème :(Parzen)[18]

Soit f une densité continue et f_h son estimateur si le noyau K vérifie

- $\int K(u)du = 1$ et $\int |K(u)|du < \infty$
- $\sup|K(u)| < \infty$ et $\lim\limits_{u\to\infty}|uK(u)| = 0$

Si $\lim\limits_{n\to\infty} h = 0$ et $\lim\limits_{n\to\infty} nh = \infty$
alors :

$$\frac{f_h(x) - \mathbb{E}\Big\{f_h(x)\Big\}}{\Big[\mathbb{V}\Big\{f_h(x)\Big\}\Big]^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

où $\xrightarrow{\mathcal{L}}$ désigne la convergence en loi.

1.6.5 Vitesse de convergence

n ne peut améliorer indéfiniment la convergence d'un estimateur f_h vers f, même pour la fonction la plus régulière possible (indéfiniment dérivable, bornée)[32], mais bien sûr inconnue, c'est-à-dire : $MSE(f_h(x))$ ne peut tendre vers 0 que d'un ordre $\frac{e}{n}$, où e est une constante.

Précisément [32] :

$$\sup_{f\in W(r,m,M)}\left(\Big|f_h(x) - f(x)\Big|^2\right) \text{ ne peut tendre vers 0 que } \frac{1}{n^{G(m,r)}},$$

$$\text{où } \frac{1}{n^{G(m,r)}} = \frac{2m - 2/r}{2m + 1 - 2/r} \text{ est une fonction croissante par rapport à } m \text{ et } r$$

$$\text{et } \lim_{m\longrightarrow\infty} G(m,r) = 1.$$

$f \in W(r,m,M)$ si :

a) les $(m-1)$ dérivées $f^{(i)}$ sont absolument continues ;
b) $\int\Big|f^{(m)}(x)\Big|^r dx < \infty$,
c) $\sup\limits_x\Big|f^{(m)}(x)\Big| \leq M < \infty$,

1.7 Noyau optimal basé sur le critère de $MISE$:

Epanechnikov (1969) [8]cite des travaux effectués par d'autres chercheurs sur des fonctions noyau. Ces auteurs travaillaient généralement avec un noyau de forme arbitraire. Dans son travail, Epanechnikov, étudie les propriétés de la fonction de densité empirique. Il étudie d'abord les propriétés asymptotiques de la fonction de densité empirique et ensuite il étudie l'erreur induite par l'estimation f_h sur la véritable fonction de densité f. À partir de cette fonction erreur, il tente de déterminer une forme de noyau optimal que l'on pourrait utiliser plutôt qu'un noyau arbitraire.

Tout d'abord, il impose certaines contraintes sur le noyau. En fait, il faut que le noyau soit symétrique, qu'il soit non-négatif sur tout le domaine D, que l'intégration sur tout son domaine D soit unitaire, qu'il ait une moyenne nulle et une variance finie et qu'il ait des moments donnés par :

$$K(u) = K(-u) \geq 0,\ \int_D K(u)d(u) = 1,\ \int_D uK(u)d(u) = 0,$$

$$\int_D u^2K(u)d(u) \leq a,\ \int_D u^mK(u)d(u) \leq \infty\ avec\ 0 \leq m \leq \infty$$

Epanechnikov (1969) exhibe un noyau optimal sous ces contraintes en minimisant l'erreur quadratique moyenne intégrée :

$$MISE(f_h(x)) = \tfrac{1}{4}h^4\mu_2^2(K)\int(f''(x))dx + \tfrac{1}{nh}\int K^2(u)du\int f(x)dx + o\left(h^4 + \tfrac{1}{nh}\right)$$

$$AMISE(f_h(x)) = \tfrac{1}{4}h^4\mu_2^2(K)\int(f''(x))dx + \tfrac{1}{nh}\int K^2(u)du\int f(x)dx$$

On constate que le $MISE$ s'écrit en fonction du biais et de la variance, le biais est une fonction croissante en h alors que le terme en variance est une fonction décroissante en h, si h est grand la variance sera petite (faible) et le biais sera fort, donc la valeur optimale de h qui minimise l'erreur quadratique moyenne intégrée $MISE$ réalise un compromis entre le biais et la variance.
Alors pour calculer le h optimale on dérive le $AMISE$ (asymptotique $MISE$) par rapport à h et on cherche le minimum comme suit :

$$\frac{\partial AMISE}{\partial h} = 0$$

$$\Rightarrow \quad h^3\mu_2^2(K)\int \Big(f''(x)\Big)^2 dx - \frac{1}{nh^2}\int K^2(u)du = 0$$

$$\Rightarrow \quad h^5 = \frac{\int K^2(u)du}{n\mu_2^2(K)\int \Big(f''(x)\Big)^2 dx}$$

D'où

$$h^* = \left[\frac{\int K^2(u)du}{\mu_2^2(K)\int (f''(x))^2 dx}\right]^{1/5} n^{-\frac{1}{5}}$$

donc :

$$h^* = \left[\frac{R(K)}{\mu_2^2(K)R(f'')}\right]^{1/5} n^{-\frac{1}{5}}$$

$$\frac{\partial^2 AMISE}{\partial^2 h} = 3h^2\mu_2^2(K)R(f'') + \frac{2}{nh^3}R(K) > 0 \Rightarrow h^* \text{ minimise la valeur de } \ AMISE$$

On remarque que l'expression du paramètre de lissage optimal au sens de l'erreur moyenne intégrée dépend encore de la fonction de densité inconnue, à cause du terme $f''(x)$. Par conséquent le calcul direct du paramètre de lissage optimal est impossible avec cette expression. En substituant h^* dans la formule $AMISE$ on obtient :

$AMISE^* = \frac{5}{4}C(K)R(f''(x)^{\frac{1}{5}}n^{\frac{-4}{5}}$
Où :
$C(K) = \mu_2^{\frac{2}{5}}R(K)^{\frac{4}{5}}$

Puisque l'on a aucune information sur $f''(x)$ et que le paramètre de lissage a été déja optimiser, alors pour minimiser le $AMISE$, il faut choisir le noyau K qui minimise la valeur de $C(K)$. Notre objectif est donc de minimiser C(K), Ce qui est équivalent à minimiser $\int (K(u))^2 du$ sous les contraintes :
$\int_D K(u)d(u) = 1$, $\int_D u^2K(u)d(u) = \mu_2(K)$
Hodges et *Lehmann* (1956)[13] montraient que ce problème de minimisation est résolu en choisissant $K(u)$ égale à :

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

On le note par K_e parce que ce noyau est appelé le noyau d'*Epanechnikov*. Nous pouvons donc considérer l'efficacité d'un noyau K (notée $eff(K)$) quelconque en le comparant avec K_e puisque ce dernier minimise le $AMISE$ si h est choisi de façon optimale, donc $eff(K) = \left[\frac{C(K_e)}{C(K)}\right]^{\frac{5}{4}}$.

Des études ont montré que l'écart de performance entre le noyau d'Epanechnikov et les noyau usuels est assez faible, on a tendance alors a utiliser des noyaux plus simples comme le noyau gaussien.
Le tableau suivant donne quelques noyaux et leurs efficacités respectives.

$Noyau$	$\{C(K_e)/C(K)\}^{5/4}$
$Epanchnikov$	1.000
$Quadratique$	0.9939
$Gaussien$	0.9512
$Triangulaire$	0.9859
$Uniforme$	0.9295

1.8 Conclusion

Dans ce chapitre nous avons parlé de l'estimation de la densité de probabilité, présenté les méthodes d'estimation non paramétriques de la densité de probabilité et nous avons détaillé celle du noyau, nous avons conclu d'après les études déja faites dans le domaine que le choix du noyau n'est pas très important car il n'a pas une grande influence sur la qualité de la fonction densité estimé par rapport à la vraie densité, mais le choix du paramètre de lissage h a une grande influence car les estimateurs peuvent changer dramatiquement par de petites variations de h .

2

Choix du paramètre de lissage

2.1 Introduction

L'estimation de la fonction de probabilité par la méthode des noyaux est principalement conditionnée par le paramètre de lissage h, ce paramètre est un facteur important dans l'estimation par la méthode des noyaux. Il représente en quelque sorte une fenêtre qui permet de déterminer le degré de lissage de l'estimation d'une fonction de densité.
Un faible paramètre de lissage implique un faible degré de lissage et résulte en une fonction de densité irrégulière. À l'opposé, une large valeur de h conduit à une estimation lisse.

La figure suivante illustre le rôle du paramètre de lissage dans l'estimation de la fonction de densité. Pour cet exemple, nous avons simuler des données à partir d'une distribution normale. On y remarque clairement que le degré de lissage s'accroît avec la largeur du paramètre de lissage. Cette figure illustre l'importance du paramètre de lissage dans l'estimation de la fonction de densité. Si h est trop faible, chacune des observations influence considérablement la forme de la densité. À l'opposé, si h est trop élevé, un surlissage risque de camoufler les particularités de la véritable fonction de densité.

Il ne faut pas oublier que ce choix dépend du but pour lequel l'estimation de la densité est

utilisée. Plusieurs méthodes pour choisir ce paramètre ont été développées dans la littérature et quelques études comparatives ont été effectuées pour ces méthodes. Deux études comparatives intéressantes ont été publiées. La première est celle de *Berlinet* et *Devroye* (1994)[1] et la deuxième est celle de *Cao* et *al.* (1994) [4]. Elles comparent plusieurs méthodes pour choisir le paramètre de lissage pour plusieurs distributions différentes. Toutes ces méthodes nous donnent un paramètre de lissage qui est optimale pour la distribution à estimer. Celles-ci diffèrent au niveau du choix du critère à optimiser. Dans ce chapitre nous allons parler de quelque méthodes permettant le calcul de ce paramètre.

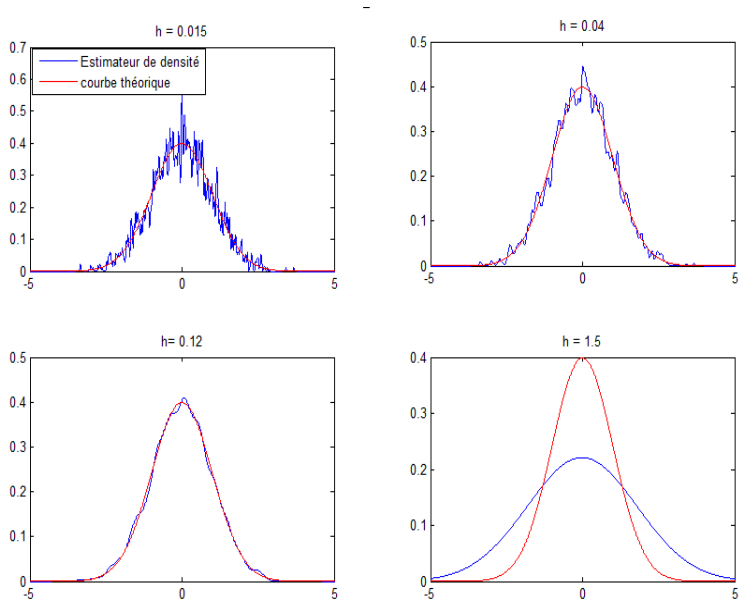

FIGURE 2.1 – influence du paraètre de lissage sur la fonction de densité estimée

2.2 Méthodes plug in (ré-injection) :

2.2.1 Estimateur optimal :

La décision d'un choix optimal pour le paramètre de lissage suppose la spécification d'un critère d'erreur qui puisse être optimisé. Bien sûr, l'optimalité n'est pas un concept absolu : elle est liée aux choix du critère qui peut faire intervenir à la fois la densité inconnue f et l'estimateur f_h(donc h et le noyau K).

Dans ce cas, on cherche à minimiser l'Erreur Quadratique Intégrée Moyenne(MISE)

$$MISE(f_h) = \int \mathbb{E}[f_h(x) - f(x)]^2 dx.$$

Le $MISE$ est défini par :

$$MISE = \frac{h^4}{4}\mu_2^2(K)R(f''(x)) + \frac{R(K)}{nh} + o(h^4 + \frac{1}{nh}). \tag{2.1}$$

et l'Erreur Quadratique Intégrée Moyenne Asymptotique $AMISE = MISE - o(h^4 + \frac{1}{nh})$ est :

$$AMISE = \frac{h^4}{4}\mu_2^2(K)R(f'') + \frac{R(K)}{nh} \tag{2.2}$$

avec $R(g) = \int g^2(x)dx$ pour toute fonction g

On peut remarquer que le premier terme du membre de droite du développement (2.1) est un terme de biais, alors que le second est un terme de variance. On constate que dans le $MISE$, le terme de 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 en sens inverse par rapport à h, une largeur de fenêtre h trop importante entraînera une augmentation du biais et une diminution de la variance (phénomène de surlissage), alors qu'une largeur de fenêtre trop petite provoquera une augmentation de la variance et une diminution du biais (phénomène de sous-lissage).

De l'expression (2.2),on peut déterminer le paramètre de lissage h^*qui minimise l'Erreur Quadratique Intégrée Moyenne Asymptotique :

$$h^* = \left[\frac{R(K)}{\mu_2^2(K)R(f'')}\right]^{\frac{1}{5}} n^{\frac{-1}{5}} \tag{2.3}$$

h^* peut s'écrire :

$$h^* = \psi(K)\varphi(f)n^{\frac{-1}{5}} \tag{2.4}$$

où $\psi(K) = \left[\frac{R(K)}{\mu_2^2(K)}\right]^{\frac{1}{5}}$ et $\varphi(f) = \left[\frac{1}{R(f'')}\right]^{\frac{1}{5}}$ avec $R(f'') \neq 0$.

Notons que h^* est une quantité déterministe qui dépend du nombre d'observations n.
La valeur du $AMISE$ optimale est donnée par :

$$AMISE^* = \frac{5}{4}\Big[\mu_2^2(K)R^4(K)R(f'')\Big]^{\frac{1}{5}} n^{\frac{-4}{5}} \tag{2.5}$$

le paramètre h^* optimal au sens du critère de l'erreur quadratique intégrée moyenne asymptotique devra réaliser un compromis entre les valeurs de la variance et celle du biais.

Outre sa nature asymptotique,la largeur de fenêtre optimale h^* dépend de la densité inconnue f à travers $R(f'')$. Cette largeur de fenêtre "idéale" (relativement au critère d'erreur retenu) n'est donc pas directement calculable. Une façon classique de remédier à ce dernier problème

consiste à remplacer la quantité $R(f'')$ par un estimateur approprié.

La première idée a été introduite par $Woodroofe$ [35].Il propose d'utiliser un paramètre de lissage h_1 pour calculer $f_{h_1}(x)$ et estimer $R(f'')$ pour calculer h^*.

2.2.2 Rule of Thumb

On a montré que si on choisit le paramètre de lissage de telle sorte que le $MISE$ soit minimum alors le paramètre de lissage optimal h^* est donnée par la formule (2.3). Dans cette formule il suffit d'assigner une valeur au terme $R(f'')$ pour obtenir une estimation de h^*.
Si on choisit f comme étant la distribution normale de moyenne 0 et de variance σ^2 on a alors :

$$R(f'') = \int (f''(x))^2 dx = \frac{3}{8}\pi^{\frac{-1}{2}}\sigma^{-5} \tag{2.6}$$

La formule (2.6) a été calculé de la façon suivante :
Supposons que f est une distribution normale de moyenne μ et de variance σ^2,on a alors :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

ainsi

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}}\left(-\left(\frac{x-\mu}{\sigma^2}\right) e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)$$

d'où :

$$f'(x) = -\left(\frac{x-\mu}{\sigma^2}\right) f(x)$$

ainsi

$$f''(x) = \frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\left(1 - \left(\frac{x-\mu}{\sigma}\right)^2\right)$$

d'où :

$$f''(x) = -\frac{1}{\sigma^2}\left(1 - \left(\frac{x-\mu}{\sigma}\right)^2\right) f(x)$$

La moyenne n'est qu'un paramètre de position, tandis que l'écart-type joue un rôle dans la forme de la distribution. On désire donc que h soit relié à l'écart-type. Ainsi, on pose la moyenne égale à zéro. Par conséquent, en considérant une distribution de variance σ^2 et de moyenne $\mu = 0$ on a alors :

$$\begin{aligned} f''(x) &= \frac{1}{\sigma^2}\left((\frac{x}{\sigma})^2 - 1\right) f(x) \\ &= \frac{1}{\sigma^3\sqrt{2\pi}}\left((\frac{x}{\sigma})^2 - 1\right) e^{-\frac{1}{2}(\frac{x}{\sigma})^2} \end{aligned}$$

En effectuant le changement de variable $z = x/\sigma$ on obtient :

$$\begin{aligned}
R(f'') &= \int (f''(x))^2 dx \\
&= \int \left[\frac{1}{\sigma^3 \sqrt{2\pi}} \left((\frac{x}{\sigma})^2 - 1 \right) e^{-\frac{1}{2}(\frac{x}{\sigma})^2} \right]^2 dx \\
&= \frac{1}{\sigma^5 2\pi} \int (z^2-1)^2 e^{-z^2} dz \\
&= \frac{1}{\sigma^5 2\pi} \int (z^4 - 2z^2 + 1) e^{-z^2} dz \\
&= \frac{1}{\sigma^5 2\pi} \left[\int z^4 e^{-z^2} dz - 2 \int z^2 e^{-z^2} dz + \int e^{-z^2} dz \right]
\end{aligned}$$

En effectuant le changement de variable $t = z\sqrt{2}$, on obtient :

$$\begin{aligned}
R(f'') &= \frac{1}{\sigma^5 2\pi} \left[\int \left(\frac{t}{\sqrt{2}} \right)^4 e^{-(\frac{t}{\sqrt{2}})^2} \frac{dt}{\sqrt{2}} - 2 \int \left(\frac{t}{\sqrt{2}} \right)^2 e^{-(\frac{t}{\sqrt{2}})^2} \frac{dt}{\sqrt{2}} + \int e^{-(\frac{t}{\sqrt{2}})^2} \frac{dt}{\sqrt{2}} \right] \\
&= \frac{1}{\sigma^5 2\pi} \left[\int \frac{t^4}{4\sqrt{2}} e^{-\frac{t^2}{2}} dt - 2 \frac{t^2}{2\sqrt{2}} e^{-\frac{t^2}{2}} dt + \int \frac{e^{-\frac{t^2}{2}}}{\sqrt{2}} dt \right] \\
&= \frac{1}{\sigma^5 2\pi} \left[\frac{\sqrt{\pi}}{4} \int \frac{t^4}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt - 2 \frac{\sqrt{\pi}}{2} \int \frac{t^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt + \sqrt{\pi} \int \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt \right] \\
&= \frac{1}{\sigma^5 2\pi} \left[\frac{\sqrt{\pi}}{4} \int \frac{t^4}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt - \sqrt{\pi} \int \frac{t^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt + \sqrt{\pi} \int \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt \right] \\
&= \frac{1}{\sigma^5 2\pi} \left[\frac{\sqrt{\pi}}{4} \mathbb{E}(T^4) - \sqrt{\pi} \mathbb{E}(T^2) + \sqrt{\pi} \int \phi(t) dt \right]
\end{aligned}$$

En considérant que $\mathbb{E}(T^2) = \mathbb{V}(T) = \int t^2 \phi(t) dt = 1$ pour une loi normale centrée réduite, et en considérant que l'intégrale de la densité ϕ sur tout son domaine est égale à 1, et $\mathbb{E}(T^4) = \int t^4 \phi(t) dt = 3$ pour la loi normale centrée réduite, on a finalement :

$$\begin{aligned}
R(f'') &= \frac{1}{\sigma^5 2\pi} \left[\frac{\sqrt{\pi}}{4} 3 - \sqrt{\pi} + \sqrt{\pi} \right] \\
&= \frac{3}{8\sigma^5 \sqrt{\pi}}
\end{aligned}$$

De plus si on utilise un noyau gaussien, on aura :

$$R(k) = \int K^2(u) du = \int \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \right)^2 du = \frac{1}{2\pi} \int e^{-u^2} du = \frac{1}{2\sqrt{\pi}}$$

Alors la valeur pour h^* notée dans ce cas h_{rot} est obtenue en substituant la valeur obtenue dans (2.6) dans la formule (2.3)

$$\begin{aligned}
h_{rot} &= \left(\frac{1}{2\sqrt{\pi}} \right)^{\frac{1}{5}} \left(\frac{3}{8\sigma^5 \sqrt{\pi}} \right)^{\frac{-1}{5}} n^{\frac{-1}{5}} \\
&= \left(\frac{4}{3} \right)^{\frac{1}{5}} \sigma n^{\frac{-1}{5}}
\end{aligned}$$

Donc :

$$h_{rot} = 1.06\sigma n^{\frac{-1}{5}} \tag{2.7}$$

Il suffit donc d'estimer σ à partir des données et de substituer cet estimateur dans la formule ci-dessus. D'après Silverman [28] (1986), cette formule donnera de bon résultat si la population est réellement normalement distribuée mais celle-ci peut donner une distribution trop lissée si la population est plutôt multimodale.

Il est possible d'obtenir de meilleurs résultats en utilisant une mesure de l'étendue de l'échantillon plus robuste que l'écart-type. L'écart inter-quartile R permet, entre autres, d'être moins sensible aux valeurs singulières. Si on modifie l'équation (2.7) en utilisant plutôt R que σ et en utilisant le fait que $R = 1.34\sigma$ dans le cas d'une distribution normale, on obtient :

$$h_{rot} = 0.79Rn^{\frac{-1}{5}} \tag{2.8}$$

où $R = X_{[3n/4]} - X_{[n/4]}$
et où $X_{[3n/4]}$ et $X_{[n/4]}$ représente le troisième quartile et le premier quartile respectivement.

Mais cette formule peut aussi donner une distribution trop lissée si la vraie densité est multimodale et parfois cette dernière donne des résultats moins bons que si l'on avait utilisé l'écart-type, d'où le meilleur des deux mondes peut être obtenu en utilisant un estimé adaptatif de l'étendue. C'est à dire, en utilisant A au lieu de σ dans la formule (2.7) où A est défini par $A = \min(\sigma, \frac{R}{1.34})$, la formule pour le h_{rot}, devient :

$$h_{rot} = 1.06An^{\frac{-1}{5}} \tag{2.9}$$

Cette expression performe relativement bien dans le cas où la distribution de la population est unimodale et symétrique, mais elle est moins efficace dans le cas des distributions multimodales et asymétriques.

Cette façon de calculer le paramètre de lissage peut être vraiment efficace dans certaines situations particulières, c'est-à-dire pour un certain nombre de distributions empiriques. De plus, elle est simple d'utilisation et rapide. Mais, comme dans la nature on est souvent confronté à des populations qui proviennent d'un mélange de plusieurs distributions, l'équation (2.9) n'est pas toujours appropriée.

2.2.3 Plug-in itéré

Description de la méthode "plug-in itéré"

En adoptant le critère d'Erreur Quadratique Intégrée Moyenne ($MISE$), *Scott*, *Tapia* et *Thompson* [25] choisissent d'estimer le paramètre $R(f'')$ de l'équation(2.3) 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 voit que :

$$f''_h(x) = \frac{1}{nh^3}\sum_{i=1}^{n} K''\left(\frac{x - x_i}{h}\right)$$

En choisissant par exemple le classique noyau gaussien

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

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

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

Il est important de noter que la largeur de fenêtre h contrôlant l'estimateur $\hat{R}_h(f'')$ de $R(f'')$ a été choisie identique à la largeur de 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 *Thompson* proposant finalement d'injecter l'estimateur $\hat{R}_h(f'')$ dans l'expression (2.3). Cette approche amène à considerer l'équation numérique suivante en h :

$$h = \psi(K)\varphi(f_h)n^{\frac{-1}{5}} \tag{2.10}$$

où $\varphi(f_h) = [\frac{1}{R(f''_h)}]^{\frac{1}{5}}$

Toute solution de l'équation (2.10)constitue un condidat potentiel à l'estimateur de la largeur de fenêtre asymptotique optimale h^*. Cette solution est notée par h_{stt}.

Lorsque l'équation admet plusieurs solutions, les auteurs proposent de choisir la plus grande (principe de surlissage), nous la noterons alors h_∞. Dans le cas contraire (absence de solution), nous dirons que l'algorithme de sélection est dégénéré.

La méthode de sélection suggérée par *Scott*, *Tapia* et *Thompson* revient à examiner les éventuels points fixes du système dynamique discret ϕ défini sur $\mathbb{R}^+$ de la façon suivante :

$$h_{i+1} = \phi(h_i) \tag{2.11}$$

où

$$\phi(h_i) = \psi(K)\varphi(f_{h_i})n^{\frac{-1}{5}}$$

Les étapes de ce processus itératif, appelé algorithme(S.T.T) sont :

Etape 1 : h_0 solution initiale, prenant par exemple l'étendue de l'échantillon ;

Etape 2 : $h_{i+1} = \psi(K)\varphi(f_{h_i})n^{\frac{-1}{5}}$;

Etape 3 : le critère d'arrêt est donné par la formule suivante : $\left|\frac{h_{i+1}-h_i}{h_{i+1}}\right| < \varepsilon$ où ε est une précision petite donnée.

2.2.4 Surlissage (Oversmoothing) :

L'estimateur h_{os}, dit paramètre de surlissage (oversmoothing parameter en anglais), est défini par :

$$h_{os} = 3\left[\frac{R(K)}{35\mu_2^2(K)}\right]^{\frac{1}{5}}\sigma n^{\frac{-1}{5}}.$$

Pour le noyau gaussien on a :

$$h_{os} = 1.144\sigma n^{\frac{-1}{5}}$$

où σ repésente l'écart-type empirique de l'échantillon. *Terrell* et *Scott* [26] ont montré sous les conditions suivantes :
f admet une dirivée second absolument continue, $f'' \in L^2$, le noyau $K \in L^2$ est une densité de probabilité continue, symétrique de variance $\mu_2(K)$ et sous conditions $h \to 0$ et $nh \to \infty$, que l'ensemble des largeurs de fenêtre optimales h^* admettent une borne supérieure. h_{os} est alors construit comme un estimateur de cette borne supérieure. La largeur de fenêtre h_{os} devrait donner de bons résultats lorsque la densité à estimer présente un faible niveau de complexité structurelle (unimodalité, par exemple). En revanche, sa qualité devrait se dégrader pour des densités cibles plus complexes, la fenêtre étant alors trop large pour permettre à l'estimateur à noyau de rendre compte des variations de la densité (phénomène de surlissage).

2.3 Méthodes de Validation Croisée (Cross Validation)

Parallèlement aux méthodes de sélection de type plug-in, d'autre méthodes sont également révélées performantes. L'idée de base des méthodes de validation croisée consiste à trouver une fonction de score $CV(h)$ ayant la même structure que le $MISE$ et dont le calcul soit plus simple. On sélectionne alors la fenêtre h minimisant ce critère dont on attend le même comportement asymptotique que h^*. La fenêtre h n'est alors plus déterministe, elle dépend des observations.

2.3.1 Validation croisée de la vraissemblance

Pour un estimateur à noyau f_h de f défini dans l'equation (1.4) et de largeur h, la sélection par validation croisée de la vraissemblance est une approche classique. C'est *Habbema*, *Hermans* et *Vandenbroek* [12] en 1974 qui ont proposé cette méthode fondée sur un critère non asymptotique du maximum de vraissemblance, mais l'interprétation entre les données est purement heuristique.

Il s'agit de maximiser par rapport à h la vraissemblance pour l'échantillon $(x_i)_{1\leq i\leq n}$ défini par :

$$LCV(h) = \prod_{i=1}^{n} f_{h,i}(x_i)$$

où $f_{h,i}(x_i) = \frac{1}{(n-1)h} \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} K\left(\frac{x_i - x_j}{h}\right)$, $i = \overline{1,n}$ est l'estimareur à noyau basé sur les $(n-1)$ observations différentes de x_i. La vraissemblance est alors :

$$LCV(h) = \prod_{i=1}^{n} \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left(\frac{x_i - x_j}{h}\right)$$

En utilisant un noyau gaussien on obtient :

$$LCV(h) = \prod_{i=1}^{n} \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right)$$

Un algorirhme peut être implémenter pour calculer le paramètre de lissage optimal noté h_{lcv}.

Algorithme 1 Likelihood validation croisée LCV

Début (Génération d'un échantillon $x_{1\leq i\leq n}$)

$LCV(h) = 1$;

Pour $i = 1$ à n faire

$Som = 0$;

Pour $j = 1$ à n faire

Si $i \neq j$ alors

$Som = Som + \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right)$;

Fin si

Fin pour $LCV(h) = \frac{1}{(n-1)h} LCV(h) Som$;

Fin pour

$h_{lcv} =$ **arg max(LCV(h))**

Cette méthode révèle un certain nombre de faiblesses pour les estimateurs non paramétriques tels que les estimateurs à noyau. Plusieurs études ont mis en avant la mauvaise robustesse de cette méthode ainsi que le risque qu'elle conduise à une estimée non consistante lorsque elle est appliquée a des observations dont la distribution présente des queues.

Une étude comparative a été faite par *Scott* et *Factor* [24] entre l'algorihtme de selection du plug-in itéré h_∞ et l'algorithme de pseudo maximum de vraissemblance h_{lcv}, cette étude confirme que les performances du plug-in itéré et du maximum de vraissemblance sont proches, avec néamoins un avantage accru pour l'estimateur h_{lcv} lorsque la taille de l'échantillion devient grande. L'estimateur h_{lcv} est pratiquement sensible à l'addition d'une donnée extrême dans l'échantillon, dont le comportement peut vite devenir irrégulier. De même cette méthode n'est pas valide que dans certaine cas (*Hall* [9], *Rudemo* [21], *Marron* [15])et peut conduire a des estimateurs non convergents lorsque la distribution des queues qui décroissent à une vitesse au plus exponentielle (cas d'une distribution exponentielle et student) (voir *Schuster* et *Gregory* [22]).

2.3.2 Validation croisée non biaisée

Description de la méthode "validation croisée non biaisée"

Une méthode appelée validation croisée non biaisée a été proposée par *Rudemo*[21] et *Bowman*[2] en 1984.
Dans la section (1.2), nous avons vu des mesures populaires de la divergence entre $f(x)$ et $f_h(x)$.Parmi ces mesures on trouve l'intégrale de l'erreur quadratique définie par :

$$\begin{aligned} ISE(f_h) &= \int (f_h(x) - f(x))^2 dx \\ &= \int f_h^2(x)dx - 2\int f_h(x)f(x)dx + \int f^2(x)dx \end{aligned}$$

Le paramètre de lissage choisi pour la méthode de la validation croisée est la valeur de ce paramètre h qui minimise un estimateur du ISE. Puisque $\int f^2(x)dx$ ne dépend pas du paramètre de lissage h. On peut choisir le paramètre de lissage de façon à ce qu'il minimise un estimateur de :

$$\begin{aligned} UCV(h) &= ISE(f_h) - \int f^2(x)dx \\ &= \int f_h^2(x)dx - 2\int f_h(x)f(x)dx \end{aligned}$$

On veut premièrement trouver un estimateur de $\int f_h(x)f(x)dx$. Remarquons que :

$$\int f_h(x)f(x)dx = \mathbb{E}(f_h(x))$$

L'estimateur empirique de $\int f_h(x)f(x)dx$, est alors $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(xi)$.

Le critère à optimiser est alors :

$$UCV(h) = \int f_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_{h,i}(x_i) \tag{2.12}$$

où $f_{h,i}(x) = \frac{1}{(n-1)h}\sum_{\substack{j=1\\ j\neq i}}^{n} K\left(\frac{x-x_j}{h}\right)$ est l'estimateur de la densité construit à partir de l'ensemble de points sauf le point x_i.
Montrons maintenant que $UCV(h)$ est un estimateur sans biais de $MISE(f_h) - R(f)$
On a :

$$\begin{aligned} MISE(f_h) - R(f) &= \mathbb{E}\int (f_h(x) - f(x))^2 dx - R(f) \\ &= \mathbb{E}\left[\int f_h^2(x)dx - 2\int f_h(x)f(x)dx\right] \end{aligned}$$

Il suffit de montrer que $\int f_h^2(x)dx$ et $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)$ sont des estimateurs sans biais de $\mathbb{E}\left[\int f_h^2(x)dx\right]$ et $\mathbb{E}\left[\int f_h(x)f(x)dx\right]$ respectivement. Or $\mathbb{E}\left[\int f_h^2(x)dx\right]$ admet l'estimateur sans biais trivial $\int f_h^2(x)dx$.

Il reste donc à montrer que $\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)$ est un estimateur sans biais de $\mathbb{E}\left[\int f_h(x)f(x)dx\right]$.
On a d'une part :

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)\right] &= \mathbb{E}\left[\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K\left(\frac{x_i - x_j}{h}\right)\right] \\ &= \frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathbb{E}\left(K\left(\frac{x_i - x_j}{h}\right)\right) \\ &= \frac{1}{h}\int f(z)\int K\left(\frac{x-z}{h}\right)f(x)dxdz. \end{aligned}$$

D'autre part :

$$\begin{aligned} \mathbb{E}\left[\int f_h(x)f(x)dx\right] &= \mathbb{E}\left[\int \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)f(x)dx\right] \\ &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left(\int K\left(\frac{x-x_i}{h}\right)f(x)dx\right) \\ &= \frac{1}{h}\mathbb{E}\left(\int K\left(\frac{x-z}{h}\right)f(x)dx\right) \\ &= \frac{1}{h}\int f(z)\int K\left(\frac{x-z}{h}\right)f(x)dxdz. \end{aligned}$$

ce qui implique que

$$\mathbb{E}\left[\int f_h(x)f(x)dx\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(x_i)\right]$$

on a pour un noyau K symétrique :

$$\begin{aligned}\int f_h^2(x)dx &= \int\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right)^2 dx \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx\end{aligned}$$

En effectuant le changement de variable $u = \frac{x-x_i}{h}$ on obtient :

$$\begin{aligned}\int f_h^2(x)dx &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(u)K\left(\frac{x_i-x_j}{h}+u\right)du \\ &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(u)K\left(\frac{x_j-x_i}{h}-u\right)du \text{ (\textit{puisque } K \textit{est un noyau symetrique})} \\ &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n} K*K\left(\frac{x_i-x_j}{h}\right)\end{aligned}$$

où $*$ représente le produit de convolution.

Ou bien, on peut aussi écrire :

$$\begin{aligned}\int f_h^2(x)dx &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx \\ &= \frac{1}{n^2h^2}\left(\sum_{i=1}^{n}\int K^2\left(\frac{x-x_i}{h}\right)dx + \sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\int K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx\right)\end{aligned}$$

Finalement, un estimateur sans biais de $MISE(f_h) - R(f)$ est donné donc par $UCV(h)$.

En utilisant l'équation (2.12),le critère $UCV(h)$ devient

$$\begin{aligned}UCV(h) &= \frac{1}{n^2h^2}\left(\sum_{i=1}^{n}\int K^2\left(\frac{x-x_i}{h}\right)dx + \sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\int K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx\right) \\ &\quad -\frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} K\left(\frac{x_i-x_j}{h}\right)\end{aligned}$$

Ou bien nous pouvons écrire aussi :

$$UCV(h) = \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} K\left(\frac{x_i-x_j}{h}\right) \quad (2.13)$$

Optimalité asymptotique

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 à été obtenue par *Stone* [30]

Théorème(*Stone*) :

Soit f_h l'estimateur de f et h_{ucv} le paramètre de lissage qui minimise le critère de validation croisée non biaisée, alors :

$$\frac{\int (f(x) - f_{h_{ucv}}(x))^2 dx}{\inf\limits_h \int (f(x) - f_h(x))^2 dx} \xrightarrow{p.s} 1$$

Propriétés de l'estimateur $UCV(h)$ **:**

Le théorème suivant donne la moyenne et la variance du $UCV(h)$ pour h fixé.

Théorème :

Si K est un noyau positif, alors

▷ $\mathbb{E}(UCV(h)) = AMISE - R(f) + o(h^4 + \frac{1}{nh})$;

▷ $\mathbb{V}(UCV(h)) = \frac{4}{n}[R(f^{\frac{3}{2}}) - R^2(f)] + o(\frac{1}{n^2 h} + \frac{h^4}{n})$.

Comporetement asymptotique de l'estimateur h_{ucv} **:**

Un résultat sur la normalité asymptotique et la vitesse de convergence sont disponibles pour la largeur de fenêtre optimale notée par h_{ucv}. Nous donnons le théorème sur la normalité asymptotique.

Théorème :

pour un noyau K positif satisfaisant les conditions suivantes :

1. f''' est absolument continue, $f^{(4)}$ intégrable, $R(f^{(4)}\sqrt{f}) < \infty$ et $R(\sqrt{f^{(4)}}f) < \infty$;
2. $K \geq 0$ un noyau symétrique avec $\mu_2 > 0$, K' est continue, alors h_{ucv} est asymptotiquement normal c'est à dire :

$$h_{ucv} \xrightarrow{cv.loi} N\left(h^*,\ \frac{2R(\rho)R(f)}{25n^2 h^{*7} \mu_2^4 R(f'')^2}\right) \tag{2.14}$$

où $\rho(c) = c\int K(w)K'(w+c)dw - 2cK'(c),\ -2 \leq c \leq 2$ et $h^* = Cn^{\frac{-1}{5}}$ le paramètre de lissage optimal défini dans (2.3).

L'écart type de $h_{ucv} - h^*$ est défini par :

$$\sqrt{\mathbb{V}(h_{ucv} - h^*)} = \frac{\sqrt{2}C^{-\frac{7}{2}}}{5\mu_2^2 R(f'')}[R(\rho)R(f)]^{\frac{1}{2}} n^{-\frac{3}{10}} \tag{2.15}$$

Remarque :

On retiendra aussi que, pour une densité f suffisament lisse, la vitesse de convergence de l'estimateur h_{ucv} est de l'ordre $n^{-\frac{1}{10}}$ c'est à dire

$$\frac{h_{ucv}}{h^*} = 1 + o(n^{-\frac{1}{10}})$$

(Pour la démonstration voire *Hall* et *Marron* [10])

Proposition :

Soit $X_1, X_2, ...X_n$ un n-échantillon indépendant et identiquement distribué issu d'une variable aléatoire X de fonction de densité f.

En utilisant le noyau gaussien on obtient :

$$\begin{aligned} UCV(h) &= \frac{1}{2n^2h\sqrt{\pi}}\left(n + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} exp\left(-\left(\frac{x_i - x_j}{2h}\right)^2\right)\right) \\ &- \frac{2}{\sqrt{2\pi}n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} exp\left(-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right) \end{aligned}$$

On a :

$$\begin{aligned} \int f_h^2(x)dx &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x - x_i}{h}\right)K\left(\frac{x - x_j}{h}\right)dx \\ &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(u)K\left(\frac{x_j - x_i}{h} - u\right)du \end{aligned}$$

Puisque le noyau K est un noyau gaussien de moyenne 0 et de variance 1 et en supposant $z = \frac{x_j - x_i}{h}$ alors :

$$\begin{aligned} \int K(u)K\left(\frac{x_j - x_i}{h} - u\right)du &= \int K(u)K(z-u)du \\ &= \int \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-u)^2}{2}}du \\ &= \frac{1}{2\sqrt{\pi}}e^{-\frac{z^2}{4}}\int \frac{1}{\sqrt{\pi}}e^{-(u-\frac{z}{2})^2}du \\ &= \frac{1}{\sqrt{2}\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{z}{\sqrt{2}})^2} \longrightarrow \mathcal{N}(0,2) \\ &= \frac{1}{2\sqrt{\pi}}exp\left(-\frac{1}{2}\left(\frac{x_j - x_i}{\sqrt{2}h}\right)^2\right) \end{aligned}$$

Donc :

$$\begin{aligned} \int f_h^2(x)dx &= \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2\sqrt{\pi}}exp\left(-\frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^2\right) \\ &= \frac{1}{2n^2h\sqrt{\pi}}\sum_{i=1}^{n}\sum_{j=1}^{n} exp\left(-\left(\frac{x_i - x_j}{2h}\right)^2\right) \end{aligned}$$

Finalement :

$$\begin{aligned} UCV(h) &= \frac{1}{2n^2h\sqrt{\pi}}\sum_{i=1}^{n}\sum_{j=1}^{n} exp\Big(-\big(\frac{x_i-x_j}{2h}\big)^2\Big) \\ &- \frac{2}{\sqrt{2\pi}n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} exp\Big(-\frac{1}{2}\big(\frac{x_i-x_j}{h}\big)^2\Big) \end{aligned}$$

ou bien :

$$\begin{aligned} UCV(h) &= \frac{1}{2n^2h\sqrt{\pi}}\Big(n+\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} exp\Big(-\big(\frac{x_i-x_j}{2h}\big)^2\Big)\Big) \\ &- \frac{2}{\sqrt{2\pi}n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} exp\Big(-\frac{1}{2}\big(\frac{x_i-x_j}{h}\big)^2\Big) \end{aligned}$$

Algorithme de la méthode validation croisée non biaisée :

Pour trouver le paramètre de lissage optimal, notée h_{ucv} par validation croisée non biaisée, on minimise numériquement $UCV(h)$. Par exemple on peut faire une recherche exhaustive du minimum $UCV(h)$ sur un grand nombre de valeurs possibles du paramètre de lissage ou il suffit de donner un algorithme afin de calculer le paramètre de lissage optimal. En utilisant le noyau gaussien, les étapes de l'algorithme sont :

Algorithme 2 validation croisée non biaisée UCV

Début (Génération d'un échantillon $x_{1\leq i\leq n}$)

$Somme1 = 0$, $Somme2 = 0$;

Pour $i = 1$ à n faire

Pour $j = 1$à n faire

Si $i \neq j$ alors

$Somme1 = Somme1 + \exp\Big(-\big(\frac{x_i-x_j}{2h}\big)^2\Big)$;

$Somme2 = Somme2 + \exp\Big(-\frac{1}{2}\big(\frac{x_i-x_j}{h}\big)^2\Big)$;

Fin Si

Fin pour

Fin pour

$UCV(h) = \frac{1}{2n^2h\sqrt{\pi}}(n + Somme1) - \frac{2}{\sqrt{2\pi}n(n-1)h}Somme2$;

$h_{ucv} = \arg\min_h UCV(h)$.

Avantages et inconvénients de la validation croisée non biaisée

Une étude de nature comparative a été donnée par *Park* et *Marron* [17] sur un certains nombres de densités loi normal (cas unimodale), mélange gaussien(cas bimodale), loi de Gumbel

et loi de Cauchy. Ils proposent de composer l'algorithme de la méthode de la validation croisée non biaisée avec la méthode de sélection plug-in itéré moderne. Les auteurs ont constaté qu'en moyenne, le paramètre de lissage h_{ucv} approche mieux que h_p la valeur asymptotique h^*, ils expliquent ce résultat par la présence du terme de biais perturbateur μ_∞ égal à 0 dans le cas de la validation croisée non biaisée.

Cette méthode présente deux problèmes majeurs(ou points faibles) :d'une part son manque de robustesse par rapport aux changements de taille de l'échantillon, c'est à dire le résultat de simulation peut se révéler extrêmement variable d'un échantillon à l'autre, d'autre part, la fonctionnelle à minimiser a souvent tendance à présenter plusieurs minimums locaux. Pour d'autres études, voire *Hall*[9], *Burman* [3], *Scott* et *Terrell* [26].

2.3.3 Validation croisée biaisée :

La notion de validation croisée biaisée, a été introduite par *Scott* et *Terrell* en 1987[26] , l'idée de cette méthode est de trouver la valeur de h qui minimise un estimateur du $AMISE$. Nous avons déja vu que :

$$AMISE(f_h) = \frac{h^4}{4}\mu_2^2(K)\int (f''(x))^2dx + \frac{1}{nh}\int K^2(u)du$$

Le paramètre de lissage basé sur la méthode de la validation croisée biaisée est la valeur de h qui minimise un estimateur du $AMISE$. On peut estimer le $AMISE$ si l'on connait $\int (f''(x))^2dx$, pour cette raison un estimateur de ce terme est donné par $\int (f_h''(x))^2dx$ *Scott* et *Terrell* (1987) [26] montraient que

$$\mathbb{E}\Big[\int (f_h''(x))^2dx\Big] = \int (f''(x))^2dx + \frac{1}{nh^5}\int (K''(u))^2du$$

Ils proposent alors d'estimer $\int (f''(x))^2dx$ par $\int (f_h''(x))^2dx - \frac{1}{nh^5}\int (K''(u))^2du$. On peut donc estimer le $AMISE$ par :

$$BCV(h) = \frac{h^4}{4}\mu_2^2(K)\Big[\int (f_h''(x))^2dx - \frac{1}{nh^5}\int (K''(u))^2du\Big] + \frac{1}{nh}\int K^2(u)du$$

Ou bien :

$$BCV(h) = \frac{h^4}{4}\mu_2^2(k)\Big[\hat{R}_h(f'') - \frac{1}{nh^5}R(K'')\Big] + \frac{R(K)}{nh}$$

Où $\hat{R}_h(f'') = \int (f_h''(x))^2dx$

Le paramètre de lissage h choisi par cette méthode est la valeur de h qui minimise $BCV(h)$, $h_{bcv} = \mathbf{arg\ \min_h(BCV(h))}$.

Proposition :

Soit $X_1, X_2, ...X_n$ un n-échantillon indépendant et identiquement distribué issu d'une variable aléatoire X de fonction de densité f.

En utilisant le noyau gaussien, on obtient :
$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$ et $\mu_2^2(K) = \int u^2 K(u) du = 1$
Donc la valeur de BCV est

$$BCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{16\sqrt{\pi}n^2h} \sum_{i=1}^{n} \sum_{j=i+1}^{n} e^{\frac{-1}{4}\left(\frac{x_i - x_j}{h}\right)^2} \left[3 - 3\left(\frac{x_i - x_j}{h}\right)^2 + \frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^4\right]$$

Démonstration :
Calculons la valeur des expressions $\int \left[K(x)\right]^2 dx$, $\int \left[K''(x)\right]^2 dx$ et $\int f_h''^2 dx$.

On a :

$$\int K^2(u) du = \frac{1}{2\sqrt{\pi}}$$

De plus, si on dérive $K(u)$ deux fois par rapport à u, on obtient
$K''(u) = \frac{-1}{\sqrt{2\pi}} \exp(-u^2/2) + \frac{1}{\sqrt{2\pi}} u^2 \exp(-u^2/2)$.
On obtient alors :

$$\begin{aligned}
\int \left[K''(u)\right]^2 du &= \int \left[\frac{-1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} + \frac{1}{\sqrt{2\pi}} u^2 e^{\frac{-u^2}{2}}\right]^2 du \\
&= \frac{1}{2\pi} \int e^{\frac{-(\sqrt{2}u)^2}{2}} du - \frac{1}{\pi} \int u^2 e^{\frac{-(\sqrt{2}u)^2}{2}} du + \frac{1}{2\pi} \int u^4 e^{\frac{-(\sqrt{2}u)^2}{2}} du \\
&= \frac{3}{8\sqrt{\pi}}.
\end{aligned}$$

Il faut maintenant évaluer $\int \left[f_h''(x)\right]^2 dx$. Puisque le noyau K est gaussien on a :

$$\begin{aligned}
f_h(x) &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \\
&= \frac{1}{nh} \sum_{i=1}^{n} e^{\frac{-1}{2}\left(\frac{x - x_i}{h}\right)^2}
\end{aligned}$$

Il est facile de montrer que si l'on dérive $f_h(x)$ deux fois par rapport à x, on obtient :

$$f_h''(x) = \frac{1}{nh^3\sqrt{2\pi}} \sum_{i=1}^{n} \left[\left(\frac{x - x_i}{h}\right)^2 e^{\frac{-1}{2}\left(\frac{x - x_i}{h}\right)^2} - e^{\frac{-1}{2}\left(\frac{x - x_i}{h}\right)^2}\right].$$

$$\begin{aligned}
\int \left[f_h''(x)\right]^2 dx &= \int \frac{1}{nh^3\sqrt{2\pi}} \sum_{i=1}^{n} \left[\left(\frac{x-x_i}{h} \right)^2 e^{\frac{-1}{2}\left(\frac{x-x_i}{h}\right)^2} - e^{\frac{-1}{2}\left(\frac{x-x_i}{h}\right)^2} \right] \\
&\quad \frac{1}{nh^3\sqrt{2\pi}} \sum_{j=1}^{n} \left[\left(\frac{x-x_j}{h} \right)^2 e^{\frac{-1}{2}\left(\frac{x-x_j}{h}\right)^2} - e^{\frac{-1}{2}\left(\frac{x-x_j}{h}\right)^2} \right] dx \\
&= \frac{1}{n^2h^6 2\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} \int \left[\left(\frac{x-x_i}{h} \right)^2 \left(\frac{x-x_j}{h} \right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} \right. \\
&- \left(\frac{x-x_i}{h} \right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} - \left(\frac{x-x_j}{h} \right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} \\
&+ \left. e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} \right] dx
\end{aligned}$$

Afin d'obtenir l'expression pour l'intégrale ci-dessus, il nous suffit de calculer les trois intégrales suivantes :

- $\int \left(\frac{x-x_i}{h}\right)^2 \left(\frac{x-x_j}{h}\right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} dx,$
- $\int \left(\frac{x-x_i}{h}\right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} dx,$
- $\int e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{x-x_j}{h}\right)^2\right]} dx.$

Les expressions obtenues pour chacune de ces intégrales sont respectivement :

- $e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2} \left[\frac{3\sqrt{\pi}h}{4} - \frac{\sqrt{\pi}}{4h}(x_i-x_j)^2 + \frac{\sqrt{\pi}}{16h^3}(x_i-x_j)^4 \right],$
- $e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2} \left[\frac{\sqrt{\pi}h}{2} + \frac{\sqrt{\pi}}{4h}(x_i-x_j)^2 \right],$
- $e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2} \left[\sqrt{\pi}h \right]$

Aprés remplacemment des intégrales ci-dessus on obtient :

$$
\begin{aligned}
\int \left[f_h''(x)\right]^2 dx &= \frac{1}{n^2h^6 2\pi}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \left(\frac{x-x_i}{h}\right)^2\left(\frac{x-x_j}{h}\right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2+\left(\frac{x-x_j}{h}\right)^2\right]} \\
&- \left(\frac{x-x_i}{h}\right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2+\left(\frac{x-x_j}{h}\right)^2\right]} - \left(\frac{x-x_j}{h}\right)^2 e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2+\left(\frac{x-x_j}{h}\right)^2\right]} \\
&+ e^{\frac{-1}{2}\left[\left(\frac{x-x_i}{h}\right)^2+\left(\frac{x-x_j}{h}\right)^2\right]} dx \\
&= \frac{1}{n^2h^6 2\pi}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2}\left[\frac{3\sqrt{\pi}h}{4} - \frac{3\sqrt{\pi}}{4h}(x_i-x_j)^2 + \frac{\sqrt{\pi}}{16h^3}(x_i-x_j)^4\right] \\
&= \frac{1}{n^2h^6 2\pi} 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2}\left[\frac{3\sqrt{\pi}h}{4} - \frac{3\sqrt{\pi}}{4h}(x_i-x_j)^2 + \frac{\sqrt{\pi}}{16h^3}(x_i-x_j)^4\right] \\
&+ \frac{n}{n^2h^6 2\pi}\frac{3\sqrt{\pi}h}{4} \\
&= \frac{1}{n^2h^6\sqrt{\pi}}\sum_{i=1}^{n}\sum_{j=i+1}^{n} e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2}\left[\frac{3h}{4} - \frac{3}{4h}(x_i-x_j)^2 + \frac{1}{16h^3}(x_i-x_j)^4\right] \\
&+ \frac{3}{8\sqrt{\pi}nh^5}
\end{aligned}
$$

On obtient l'expression de $BCV(h)$ donnée par :

$$
BCV(h) = \frac{1}{2\sqrt{\pi}nh} + \frac{1}{16\sqrt{\pi}n^2h}\sum_{i=1}^{n}\sum_{j=i+1}^{n} e^{\frac{-1}{4}\left(\frac{x_i-x_j}{h}\right)^2}\left[3 - 3\Big(\frac{x_i-x_j}{h}\Big)^2 + \frac{1}{4}\Big(\frac{x_i-x_j}{h}\Big)^4\right]
$$

Algorithme de la méthode validation croisée biaisée

Afin de calculer le paramètre de lissage optimal noté h_{bcv} qui minimise $BCV(h)$, en utilisant le noyau gaussien, les principales étapes de l'algorithme sont :

Algorithme 3 (validation croisée biaisée BCV)

Début (Génération d'un échantillon $x_{1\leq i\leq n}$)

$BCV(h) = 0$;

Pour $i = 1$ à n faire

Pour $j = i+1$ à n faire

$x = \left(\frac{x_i-x_j}{h}\right)$;

$BCV(h) = BCV(h) + \exp(-\frac{x^2}{4})\Big(3 - 3x^2 + \frac{1}{4}x^4\Big)$,

Fin pour

$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{16n^2h\sqrt{\pi}}BCV(h)$,

Fin pour

$h_{bcv} = \arg\min_h BCV(h)$.

2.3.4 Validation croisée lissée

On a vu à dans la section (1.2) que le $MISE(f_h)$ peut être exprimé sous la forme :

$$MISE(f_h) = \int \mathbb{V}[f_h(x)]dx + \int \Big[\mathbb{E}[f_h(x)] - f(x)\Big]^2 dx.$$

De plus, on a vu que l'on pouvait approximer $\int \mathbb{V}[f_h(x)]dx$ par $\frac{R(K)}{nh}$ qui ne dépend pas de f. Cependant, pour le deuxième terme de l'équation ci-dessus, on a vu que $\mathbb{E}[f_h(x)] = \int \frac{1}{h}K\big(\frac{x-t}{h}\big)f(t)dt$ d'où un estimateur naturel pour le deuxième terme du $MISE$ est donné par :

$$\hat{B}(h) = \int \left[\int \frac{1}{h}K\left(\frac{x-t}{h}\right)\hat{f}_g(t)dt - \hat{f}_g(x)\right]^2 dx$$

où $\hat{f}_g$ est un estimateur auxiliaire de f, c'est à dire $\hat{f}_g(x) = \frac{1}{ng}\sum_{i=1}^{n} L\left(\frac{x-x_i}{g}\right)$ avec $g \neq h$.
Nous avons alors un estimateur du $MISE(f_h)$ donné par :

$$SCV(h) = \frac{1}{nh}R(K) + \hat{B}(h)$$

On définit le paramètre de lissage choisi par cette méthode par la valeur de h qui minimise $SCV(h)$. *Hall*, *Marron* et *Park* (1992)[11] ont montré que lorsque $K = L = \phi$ où ϕ est la densité de la loi normale standard, alors $g = \hat{\sigma}(0.9266)n^{\frac{-2}{13}}$. On a vu dans la sous-section précédente que $R(K) = \frac{1}{2\sqrt{\pi}}$, évaluons maintenant l'expression du $\hat{B}(h)$.

$$\begin{aligned}
\hat{B}(h) &= \int \left[\int \frac{1}{h}K\left(\frac{x-t}{h}\right)\hat{f}_g(t)dt - \hat{f}_g(x)\right]^2 dx \\
&= \int \left[\sum_{i=1}^{n} \frac{1}{2\pi ngh}\int e^{-\frac{1}{2}\left(\frac{x-t}{h}\right)^2} e^{-\frac{1}{2}\left(\frac{t-x_i}{g}\right)^2} dt - \frac{1}{\sqrt{2\pi}ng}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x-x_i}{g}\right)^2}\right]^2 dx.
\end{aligned}$$

On peut évaluer $\int e^{-\frac{1}{2}\left(\frac{x-t}{h}\right)^2} e^{-\frac{1}{2}\left(\frac{t-x_i}{g}\right)^2} dt$ et l'on obtient que :

$$\begin{aligned}
\int e^{-\frac{1}{2}\left(\frac{x-t}{h}\right)^2} e^{-\frac{1}{2}\left(\frac{t-x_i}{g}\right)^2} dt &= \int e^{-\frac{1}{2}\left[\left(\frac{x-t}{h}\right)^2 + \left(\frac{t-x_i}{g}\right)^2\right]} dt \\
&= \frac{\sqrt{2\pi}gh}{(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x-x_i}{(g^2+h^2)^{\frac{1}{2}}}\right)^2}.
\end{aligned}$$

Donc :

$$\begin{aligned}
\hat{B}(h) &= \int \left[\sum_{i=1}^{n} \frac{1}{2\pi ngh} \int e^{-\frac{1}{2}\left(\frac{x-t}{h}\right)^2} e^{-\frac{1}{2}\left(\frac{t-x_i}{g}\right)^2} dt - \frac{1}{\sqrt{2\pi}ng} \sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x-x_i}{g}\right)^2} \right]^2 dx \\
&= \int \left[\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}n(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x-x_i}{(g^2+h^2)^{\frac{1}{2}}}\right)^2} - \frac{1}{\sqrt{2\pi}ng} \sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x-x_i}{g}\right)^2} \right]^2 dx \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \int \left[\frac{1}{\sqrt{2\pi}n(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x-x_i}{(g^2+h^2)^{\frac{1}{2}}}\right)^2} - \frac{1}{\sqrt{2\pi}ng} e^{-\frac{1}{2}\left(\frac{x-x_i}{g}\right)^2} \right] \\
&\quad \left[\frac{1}{\sqrt{2\pi}n(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x-x_j}{(g^2+h^2)^{\frac{1}{2}}}\right)^2} - \frac{1}{\sqrt{2\pi}ng} e^{-\frac{1}{2}\left(\frac{x-x_j}{g}\right)^2} \right] dx
\end{aligned}$$

Pour évaluer cette dernière équation, il nous suffit d'évaluer les intégrales suivantes :

- $\int \frac{1}{2\pi n^2(g^2+h^2)} e^{-\frac{1}{2}\left(\frac{x-x_i}{(g^2+h^2)^{\frac{1}{2}}}\right)^2 - \frac{1}{2}\left(\frac{x-x_j}{(g^2+h^2)^{\frac{1}{2}}}\right)^2} dx$
- $\int \frac{-1}{2\pi n^2 g(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x-x_i}{(g^2+h^2)^{\frac{1}{2}}}\right)^2 - \frac{1}{2}\left(\frac{x-x_j}{g}\right)^2} dx$
- $\int \frac{1}{2\pi n^2 g^2} e^{-\frac{1}{2}\left(\frac{x-x_i}{g}\right)^2 - \frac{1}{2}\left(\frac{x-x_j}{g}\right)^2} dx$

Les valeurs respectives pour chacune de ces intégrales sont :

- $\frac{1}{2\sqrt{\pi}n^2(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2(g^2+h^2))^{\frac{1}{2}}}\right)^2}$
- $\frac{-1}{\sqrt{2\pi}n^2(2g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2g^2+h^2)^{\frac{1}{2}}}\right)^2}$
- $\frac{1}{2\sqrt{\pi}n^2 g} e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2g^2)^{\frac{1}{2}}}\right)^2}$

D 'où

$$\begin{aligned}
\hat{B}(h) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \left[\frac{1}{2\sqrt{\pi}n^2(g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2(g^2+h^2))^{\frac{1}{2}}}\right)^2} - \frac{2}{\sqrt{2\pi}n^2(2g^2+h^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2g^2+h^2)^{\frac{1}{2}}}\right)^2} \right. \\
&\quad \left. + \frac{1}{2\sqrt{\pi}n^2 g} e^{-\frac{1}{2}\left(\frac{x_i-x_j}{\left(2g^2\right)^{\frac{1}{2}}}\right)^2} \right]
\end{aligned}$$

On obtient alors la formule pour 1a méthode de la validation croisée lissée qui est :

$$\begin{aligned}
SCV(h) &= \frac{1}{2\sqrt{\pi}nh} + \hat{B}(h) \\
&= \frac{1}{2\sqrt{\pi}nh} + \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{1}{2\sqrt{\pi}n^2(g^2+h^2)^{\frac{1}{2}}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2(g^2+h^2))^{\frac{1}{2}}}\right)^2} - \frac{2}{\sqrt{2\pi}n^2(2g^2+h^2)^{\frac{1}{2}}}\right. \\
&\quad \left. e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2g^2+h^2)^{\frac{1}{2}}}\right)^2} + \frac{1}{2\sqrt{\pi}n^2 g}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{(2g^2)^{\frac{1}{2}}}\right)^2}\right] \\
&= \frac{1}{2\sqrt{\pi}nh} + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}n^2}\left[\frac{1}{\sqrt{2(g^2+h^2)}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{\sqrt{2(g^2+h^2)}}\right)^2} - \frac{2}{\sqrt{2g^2+h^2}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{\sqrt{2g^2+h^2}}\right)^2}\right. \\
&\quad \left. + \frac{1}{\sqrt{2g^2}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{\sqrt{2g^2}}\right)^2}\right].
\end{aligned}$$

Tout comme précédemment, il est alors facile d'implanter un algorithme afin de calculer le paramètre de lissage qui minimise le $SCV(h)$.

Algorithme 4 validation croisée lissée SCV

Début (Génération d'un échantillon $x_{1\leq i\leq n}$)

$\bar{x} = 0$, $y = 0$;

Pour $j = 1$ à n faire

$\bar{x} = \bar{x} + x_j$

$y = y + x_j^2$

Fin pour

$s = (\frac{1}{n}y - (\frac{\bar{x}}{n})^2)^{\frac{1}{2}}$

$g = s(0.9266)n^{\frac{-2}{13}}$

$g_1 = 2(h^2 + g^2)$

$g_2 = h^2 + 2g^2$

$g_3 = 2g^2$

$SCV(h) = 0$

Pour $i = 1$ à n faire

Pour $j = 1$ à n faire

$x = \frac{(x_i - x_j)^2}{2}$

$SCV(h) = SCV(h) + \frac{1}{\sqrt{g_1}}e^{\frac{-x}{g_1}} - \frac{2}{\sqrt{g_2}}e^{\frac{-x}{g_2}} + \frac{1}{\sqrt{g_3}}e^{\frac{-x}{g_3}}$

Fin pour

Fin pour

$SCV(h) = (\frac{1}{n}SCV(h) + \frac{1}{h\sqrt{2}})\frac{1}{n\sqrt{2\pi}}$

$h_{scv} = arg\min_h SCV(h)$.

2.4 Conclusion :

Dans ce chapitre nous avons traité deux familles de procedure permettant le calcul du paramètre de lissage : famille des méthodes plug-in (ré-injection) et famille des méthodes validation croisée (cross validation), d'aprés les études déjà faittes dans le domaine nous avons conclus qu'aucune méthode n'est meilleure que les autres. Ces méthodes de sélection du paramètre de lissage ne fournissent pas une estimation graphiquement satisfaisante, laissant subsister des variations locales. L'estimateur a tendance à dévier systématiquement de la vraie valeur de la densité au voisinage de certains points critiques. Des travaux ont montré que le choix de la méthode de sélection dépend de la forme de la densité que l'on cherche à estimer.

3
Simulation

3.1 Introduction

Nous présentons dans ce chapitre le travail de simulation effectué pour étayer les différents aspects théoriques abordés dans notre étude.
L'expérimentation numérique nous servira à :

- Comparer les différents algorithmes de sélection du paramètre de lissage.
- Etudier la performance de ces algorithmes.
- Etudier l'influence de la taille de l'échantillon sur ces différents algorithmes.

3.2 Plan de simulation :

Nous nous contenterons de faire des simulations et d'observer le comportement asymptotique de l'estimateur à noyau calculé à partir d'échantillons simulés, censés représenter une loi connue, donc la densité de probabilité f. Ceci, nous permettra de savoir si l'estimateur f_h converge vers f.

Nous utilisons pour les simulations des échantillons de lois connues de taille de plus en plus grande (1000, 2500, 4000 et 6000).

Afin d'illustrer les performances des méthodes de sélection présentées dans le chapitre 2, nous utilisons trois densités tests. Nous avons choisi des densités présentant différents aspects :

D1 : La loi Normale $\mathcal{N}(0,1) : x \longmapsto \frac{1}{\sqrt{2\pi}} exp(\frac{-x^2}{2})$, $x \in]-\infty, +\infty[$,

D2 : La loi Exponentielle de paramètre $\lambda = 1$:$x \longmapsto exp(-x), x \geq 0$,

D3 : La loi de Khi-Deux χ^2 à 4 degrés de liberté : $x \longmapsto \frac{x exp(\frac{-x}{2})}{2^2 \Gamma(2)}$ $x \geq 0$
où $\forall \alpha > 0, \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} exp(-x) dx$.

Ces trois densités sont représentées sur la figure (3.1). Afin de rendre l'étude la plus pertinente possible, nous avons choisi des densités présentant différents aspects.
Ainsi nous remarquons que :

- La densité D1 est unimodale et suffisamment lisse ;
- La densité D2 est à queue lourde ;
- La densité D3 est unimodale avec queue.

Pour chaque densité, nous avons simulé un échantillon pour des tailles n=1000, 2500, 5000 et 6000.

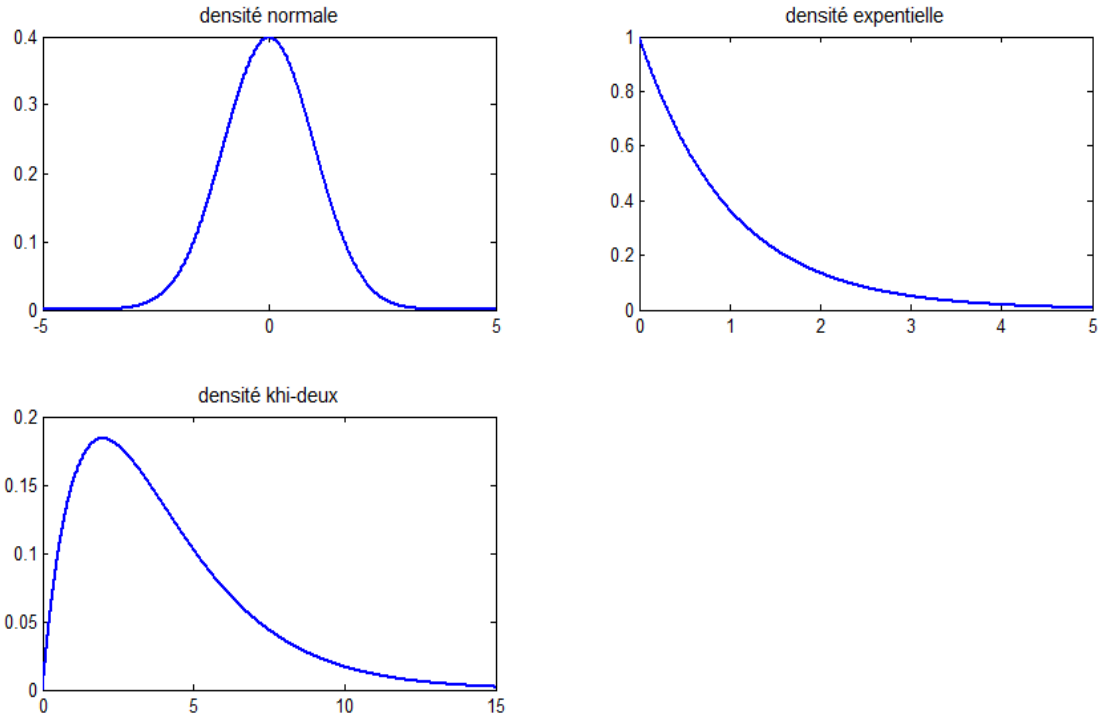

FIGURE 3.1 – Les densités tests

3.3 Algorithme de simulation

L'algorithme de simulation que nous avons utilisé comporte quatre phases :

Algorithme 5 Algorithme de simulation

-Simuler un échantillon de taille n pour une densité qui appartient à la famille paramétrique comme : D1, D2 ou D3.

-Calculer le paramètre de lissage optimal h_{opt} en utilisant les différentes méthodes de sélection (famille des méthodes plug-in, famille des méthodes validation croisée)

-Construire l'estimateur par la méthode du noyau à partir des observations.

-Tracer les deux courbes : densité test et la densité estimée.

Notons que dans la programmation des méthodes de sélection du paramètre de lissage h, nous avons choisi comme noyau, le noyau gaussien. L'estimateur f_h de f a les mêmes propriétés de continuité et de différentiabilité que K, et f_h admet aussi des dérivées de tout ordre comme K.

De plus il présente une trés petite différence au niveau de l'efficacité avec les autres noyaux en particulier le noyau d'Epanechnikov (K_e) qui minimise le $MISE$.

Les simulations et les graphiques ont été réalisés à l'aide du logiciel MATLAB. Nous avons utilisé la version 7.9.0 pour la programmation.

Matlab ("matrix laboratory") est un langage de programmation de quatrième génération émulé par un environnement de développement du même nom ; il est utilisé à des fins de calcul numérique. Développé par la société The $MathWorks$, Matlab permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en oeuvre des algorithmes, de créer des interfaces utilisateurs, et peut s'interfacer avec d'autres langages comme le C, $C++$, $Java$, et $Fortran$. Les utilisateurs de Matlab (environ un million en 2004) sont de milieux très différents comme l'ingénierie, les sciences et l'économie dans un contexte aussi bien industriel que pour la recherche. Matlab peut s'utiliser seul ou bien avec des toolbox ("boîte à outils").

3.4 Résultats de simulations

Les résultats de simulation sont donnés sous forme de tableaux et de graphiques. les tableaux (3.1) à (3.3) contiennent les résultats suivants :

1) h_{opt} : est le paramètre de lissage asymptotique optimal pour le modèle de densité testé, pour chaque méthode de sélection.

2) $crit$: est la valeur du critère, les critères utilisés sont :

- $AMISE(h)$: est le critère utilisé pour les méthodes plug-in .
- $LCV(h)$: est le critère utilisé pour la méthode validation croisée de la vraisemblance.
- $UCV(h)$: est le critère utilisé pour la méthode validation croisée non biaisée .
- $BCV(h)$: est le critère utilisé pour la méthode validation croisée biaisée .
- $SCV(h)$: est le critère utilisé pour la méthode validation croisée lissée.

3) h^* : est la valeur optimale théorique calculée à partir de l'equation (2.4) .

4) $AMISE^* = AMISE(h^*)$: est la valeur exacte du $AMISE$. Cette valeur exacte est donnée par :

$$AMISE(h^*) = \frac{5}{4}\Big[\mu_2^2(K)R^4(K)R(f'')\Big]^{\frac{1}{5}} n^{\frac{-4}{5}}$$

5) Eff : est une estimation de l'efficacité de l'estimateur de h définie par :

$$\frac{AMISE(h^*)}{AMISE(h_{opt})}$$

Les tableaux (3.1) à(3.3) nous permettent de comparer les résultats obtenus par les méthodes plug-in et validation croisée, pour les différentes densités tests : D1, D2 et D3.

Densité normale

n	Méthode	h_{opt}	$crit$	$AMISE(h_{opt})$	h^*	$AMISE(h^*)$	$Eff(\%)$
1000	h_{rot}	0.2555	0.0013	0.001329	0.2660	0.001325	99.68
	h_{stt}	0.2011	0.0017	0.001488			89.01
	h_{os}	0.2757	0.0012	0.001328			99.73
	h_{lcv}	0.283	0.2504	0.001336			99.19
	h_{ucv}	0.269	-0.2910	0.001326			99.97
	h_{bcv}	0.28	0.0012	0.001332			99.45
	h_{scv}	0.274	0.0013	0.001327			99.82
2500	h_{rot}	0.2163	0.0006	$0.6374.(10^{-3})$	0.2215	$0.6367.(10^{-4})$	99.89
	h_{stt}	0.1968	0.0007	$0.6525.(10^{-3})$			97.57
	h_{os}	0.2335	0.0005	$0.6404.(10^{-3})$			99.41
	h_{lcv}	0.237	0.2473	$0.6429.(10^{-3})$			99.03
	h_{ucv}	0.249	-0.2867	$0.6564.(10^{-3})$			96.99
	h_{bcv}	0.224	0.0005	$0.6369.(10^{-3})$			99.97
	h_{scv}	0.229	0.0006	$0.6382.(10^{-3})$			99.77
4000	h_{rot}	0.2017	0.0004	$0.43719.(10^{-3})$	0.2016	$0.43718.(10^{-3})$	99.99
	h_{stt}	0.1026	0.0008	$0.6930.(10^{-3})$			63.08
	h_{os}	0.2177	0.0004	$0.4428.(10^{-3})$			98.73
	h_{lcv}	0.199	0.2415	$0.4373.(10^{-3})$			99.96
	h_{ucv}	0.141	-0.2801	$0.5210.(10^{-3})$			83.90
	h_{bcv}	0.233	0.0004	$0.4585.(10^{-3})$			95.33
	h_{scv}	0.21	0.0004	$0.4386.(10^{-3})$			99.65
6000	h_{rot}	0.1880	0.0003	$0.3161.(10^{-3})$	0.1859	$0.3160.(10^{-3}$	99.99
	h_{stt}	0.1553	0.0003	$0.3334.(10^{-3})$			94.80
	h_{os}	0.2029	0.0002	$0.3213.(10^{-3})$			98.35
	h_{lcv}	0.203	0.2390	$0.3214.(10^{-3})$			98.33
	h_{ucv}	0.2	-0.2785	$0.3197.(10^{-3})$			98.86
	h_{bcv}	0.19	0.0003	$0.3163.(10^{-3})$			99.90
	h_{scv}	0.193	0.0003	$0.3169.(10^{-3})$			99.71

TABLE 3.1 – **Résultats des simulations effectuées sur la loi normale pour déterminer les largeurs de fenêtre optimales**

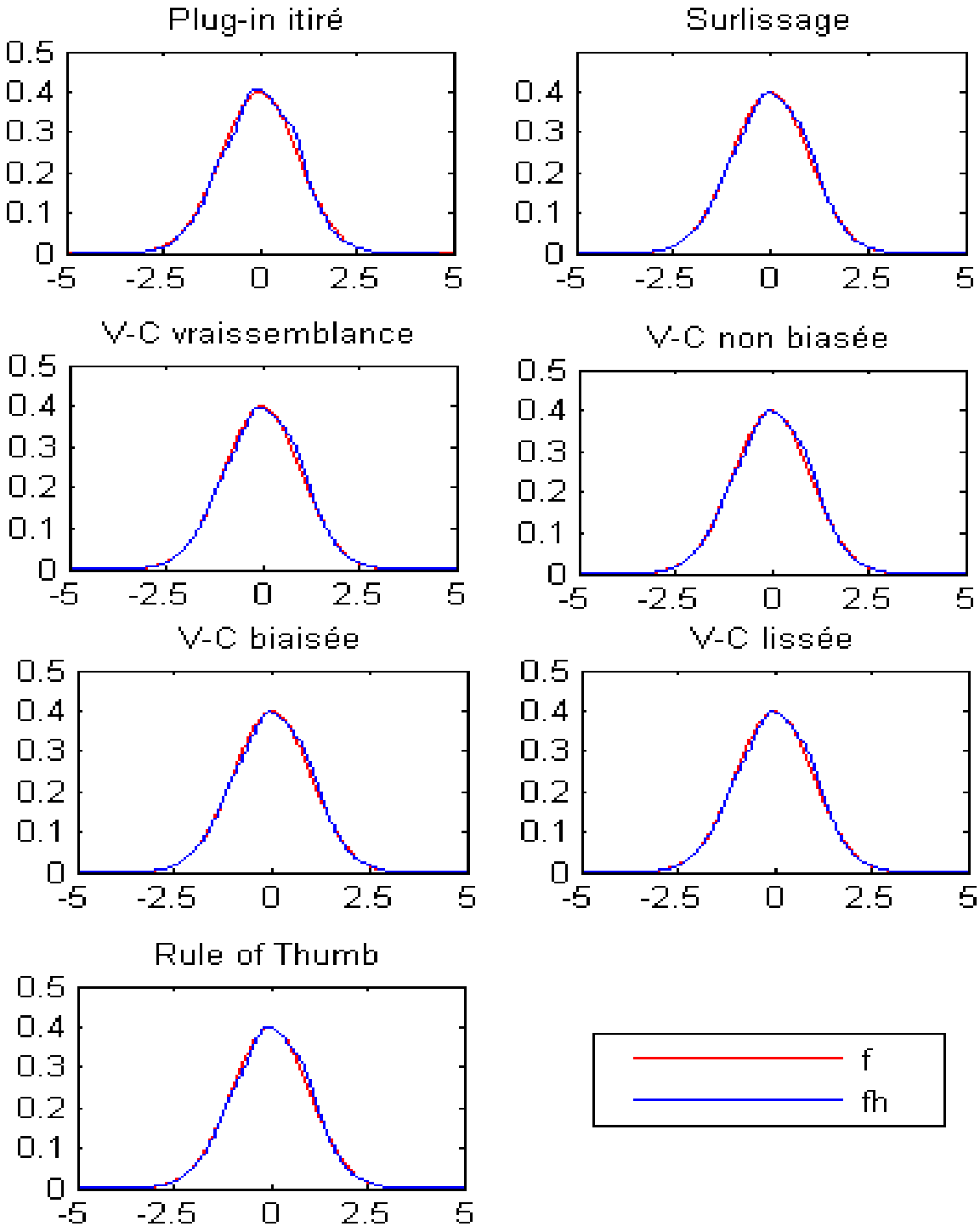

FIGURE 3.2 – Illustration des résultats de simulations effectuées sur la loi normale pour n=1000

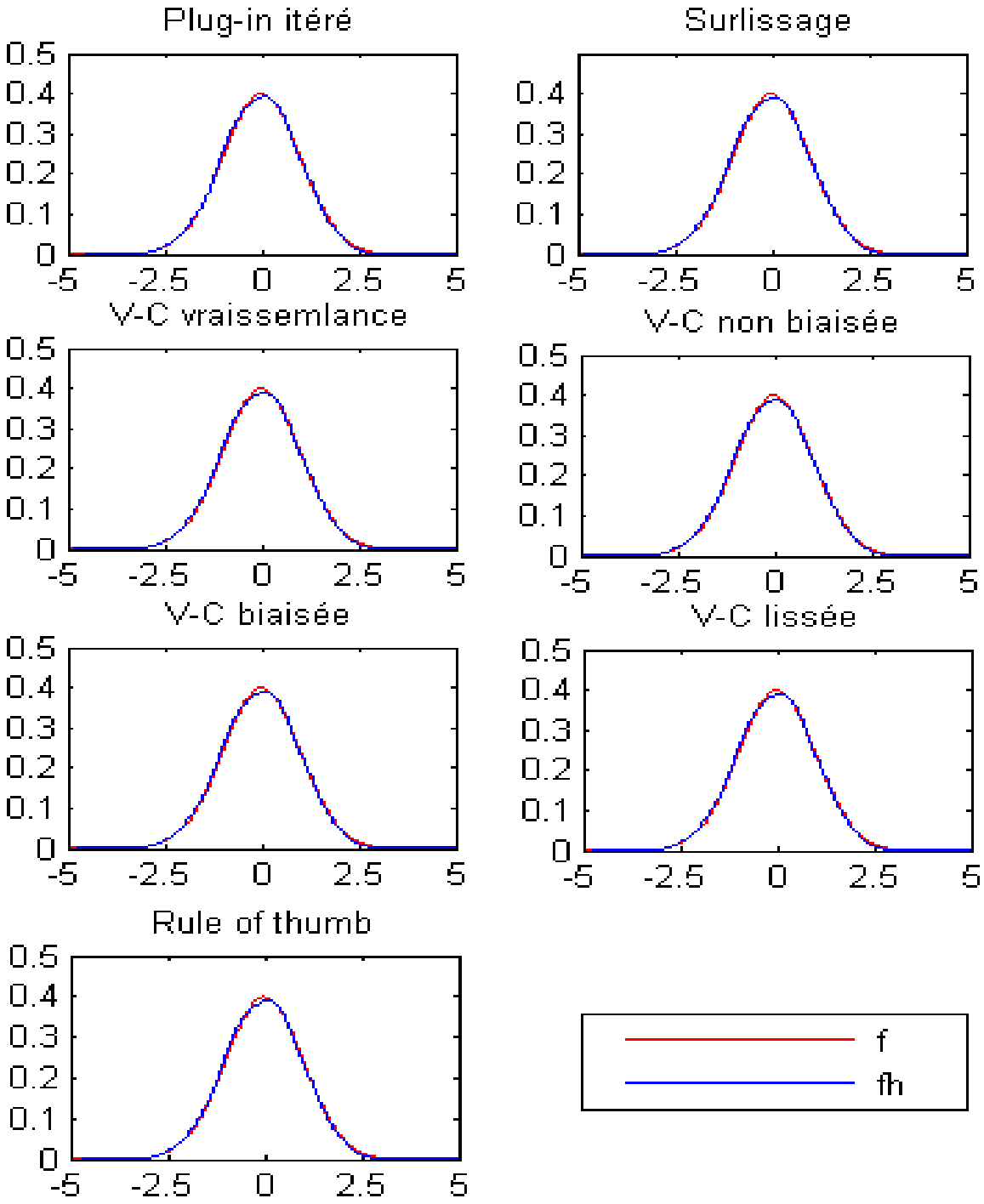

FIGURE 3.3 – Illustration des résultats de simulations effectuées sur la loi normale pour n=2500

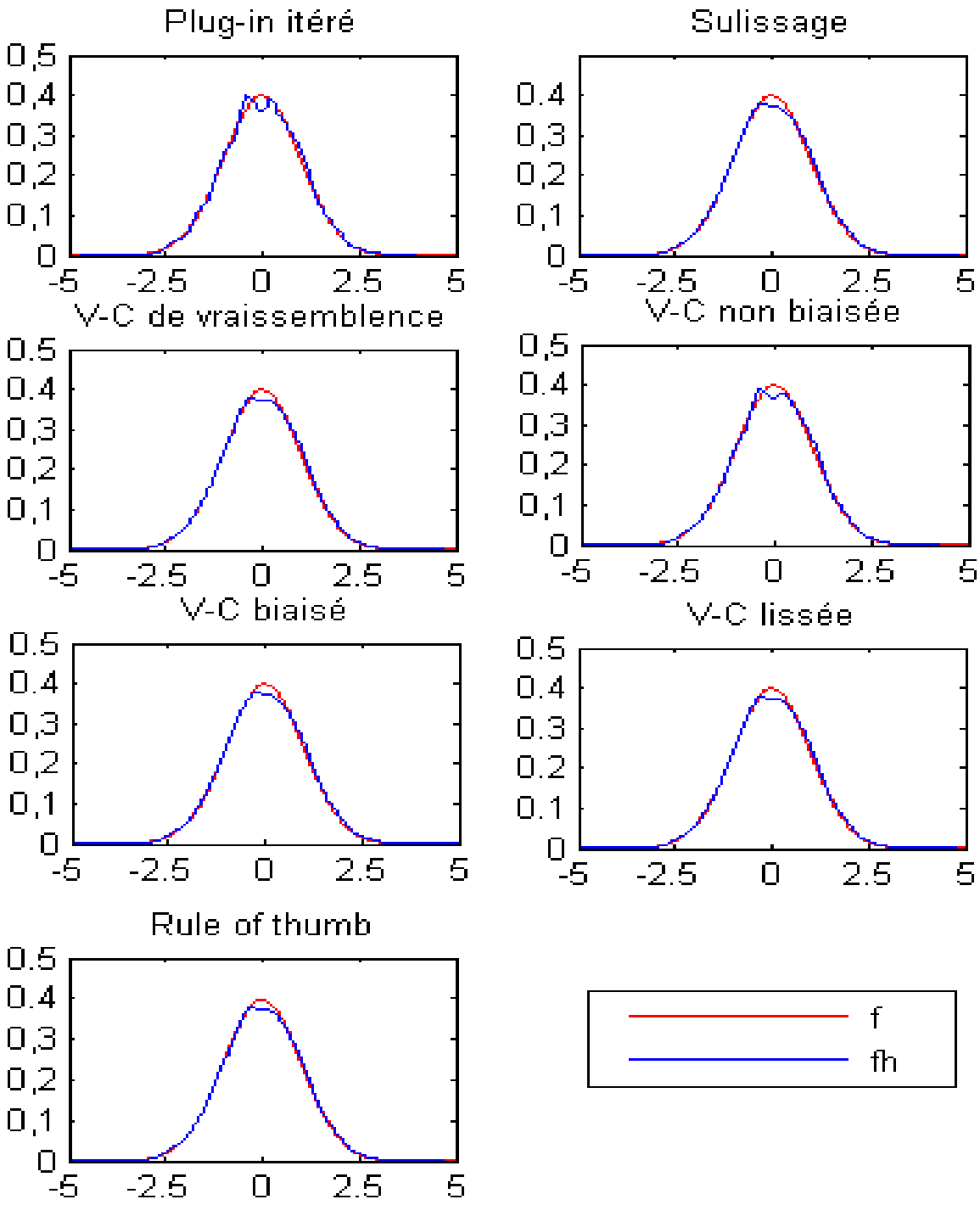

FIGURE 3.4 – Illustration des résultats de simulations effectuées sur la loi normale pour n=4000

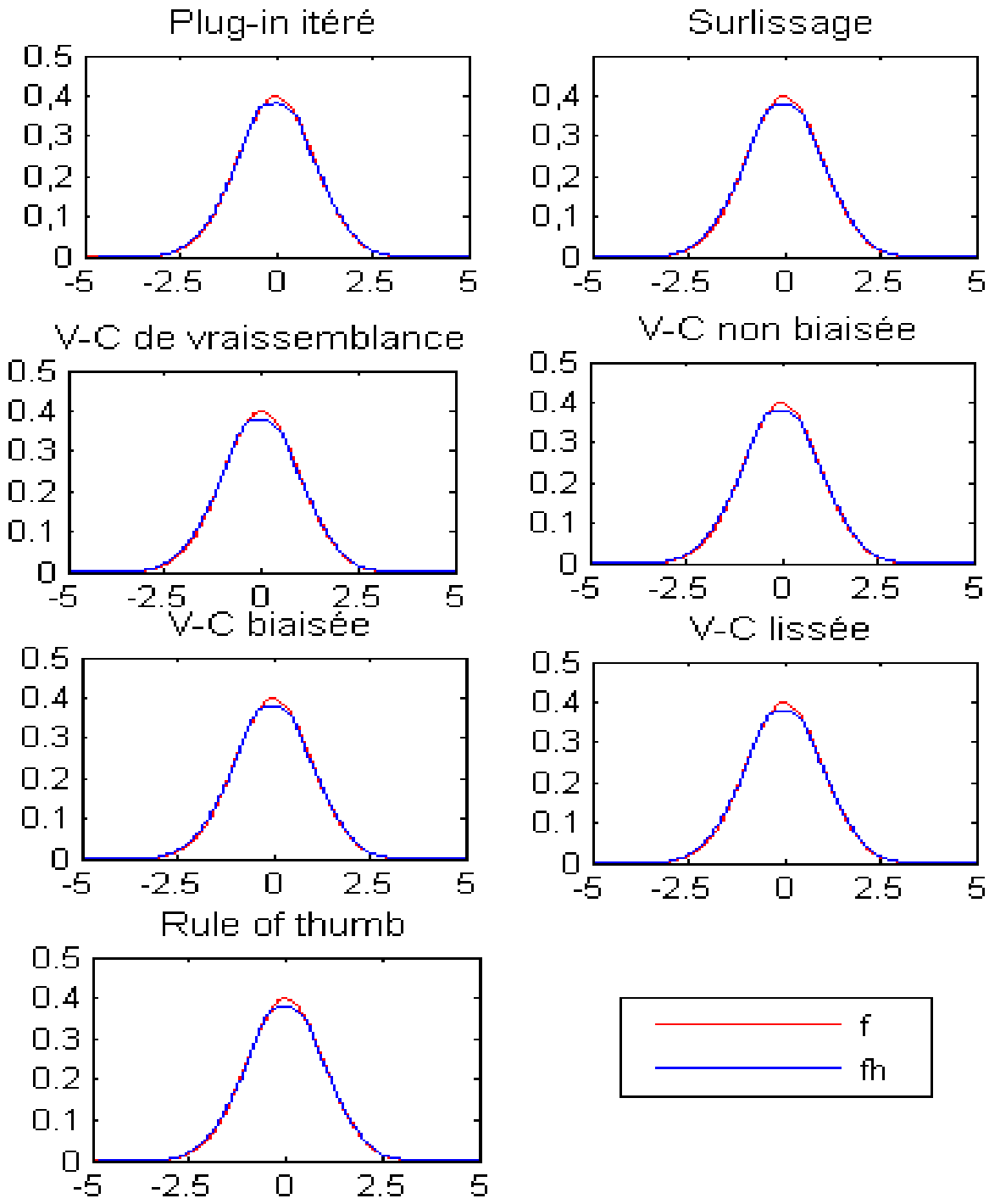

FIGURE 3.5 – Illustration des résultats de simulations effectuées sur la loi normale pour n=6000

Densité exponentielle

n	Méthode	h_{opt}	$crit$	$AMISE(h_{opt})$	h^*	$AMISE(h^*)$	$Eff(\%)$
1000	h_{rot}	0.2252	0.0015	0.0016	0.2240	0.0016	99.99
	h_{stt}	0.0324	0.0108	0.0087			18.08
	h_{os}	0.2685	0.0012	0.0017			92.54
	h_{lcv}	0.1630	0.3346	0.0018			86.54
	h_{ucv}	0.0470	-0.4608	0.0060			26.22
	h_{bcv}	0.1040	0.0045	0.0027			57.72
	h_{scv}	0.1510	0.0023	0.0019			81.42
2500	h_{rot}	0.1804	0.0007	$7.5783.(10^{-4})$	0.1865	$7.5625.(10^{-4})$	99.79
	h_{stt}	0.0048	0.0293	$2.3442.(10^{-2})$			3.22
	h_{os}	0.2442	0.0005	$9.0674.(10^{-4})$			83.40
	h_{lcv}	0.0630	0.3562	$1.7930.(10^{-3})$			42.17
	h_{ucv}	0.0200	-0.5075	$5.6419.(10^{-3})$			13.40
	h_{bcv}	0.0450	0.0041	$2.5080.(10^{-3})$			30.15
	h_{scv}	0.1180	0.0012	$9.8048.(10^{-4})$			77.13
4000	h_{rot}	0.1669	0.0005	$5.1952.(10^{-4})$	0.1697	$5.1924.(10^{-4})$	99.94
	h_{stt}	0.0170	0.0051	$4.1391.(10^{-3})$			12.54
	h_{os}	0.2230	0.0004	$6.2544.(10^{-4})$			83.02
	h_{lcv}	0.0920	0.3368	$7.7551.(10^{-4})$			66.95
	h_{ucv}	0.0240	-0.4769	$2.9385.(10^{-3})$			17.67
	h_{bcv}	0.0550	0.0031	$1.2833.(10^{-3})$			40.45
	h_{scv}	0.1050	0.0008	$6.8684.(10^{-4})$			75.59
6000	h_{rot}	0.1526	0.0003	$3.7586.(10^{-4})$	0.1565	$3.7540.(10^{-4})$	99.87
	h_{stt}	0.0100	0.0058	$4.6740.(10^{-3})$			8.03
	h_{os}	0.2025	0.0002	$4.4249.(10^{-4})$			84.83
	h_{lcv}	0.0620	0.3506	$7.6016.(10^{-4})$			49.38
	h_{ucv}	0.0150	-0.4896	$3.1343.(10^{-3})$			11.97
	h_{bcv}	0.0360	0.0023	$1.3062.(10^{-3})$			28.74
	h_{scv}	0.0910	0.0006	$5.2522.(10^{-4})$			71.47

TABLE 3.2 – **Résultats des simulations effectuées sur la loi exponentielle pour déterminer les largeurs de fenêtre optimales**

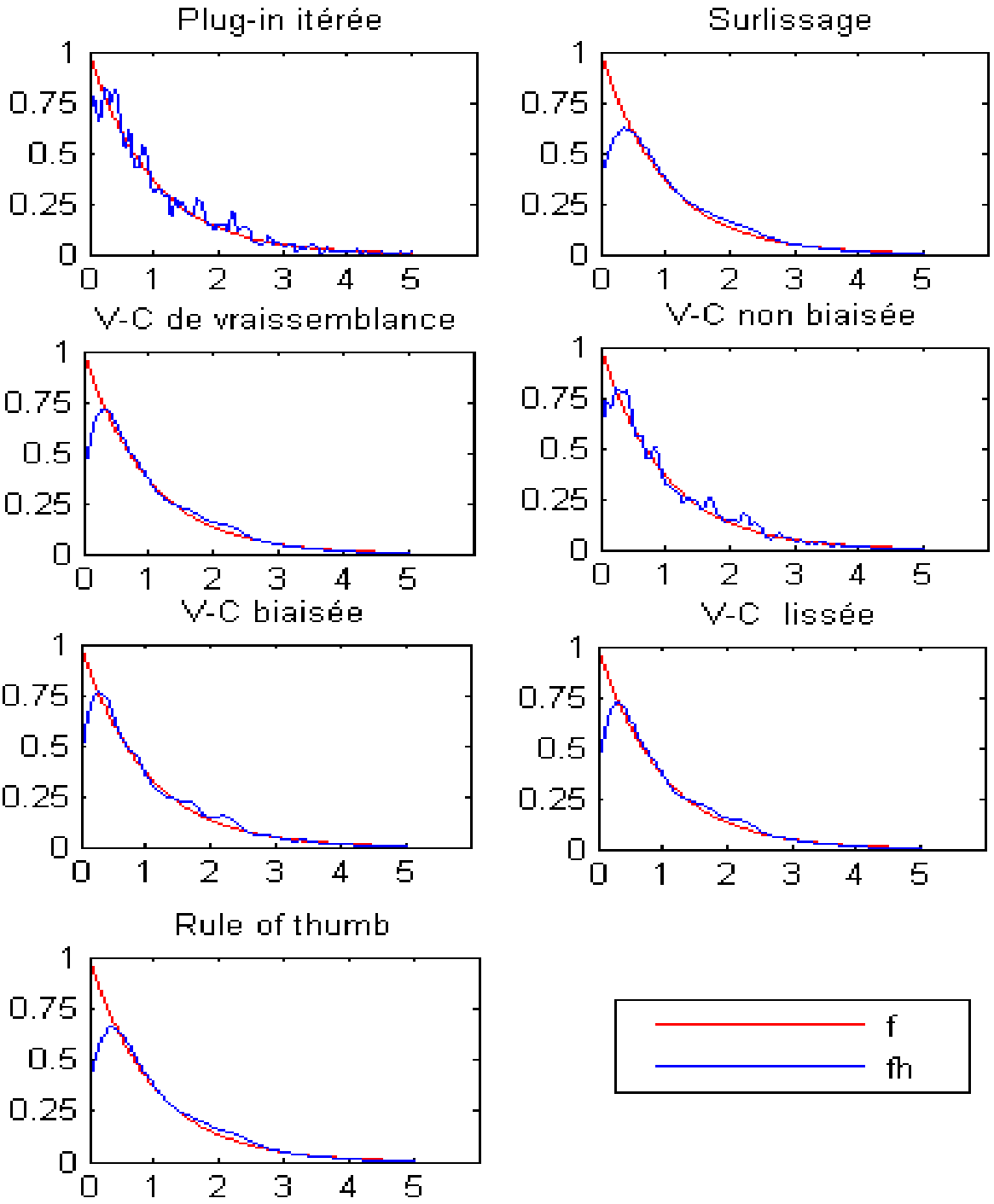

FIGURE 3.6 – Illustration des résultats de simulations effectuées sur la loi exponentielle pour n=1000

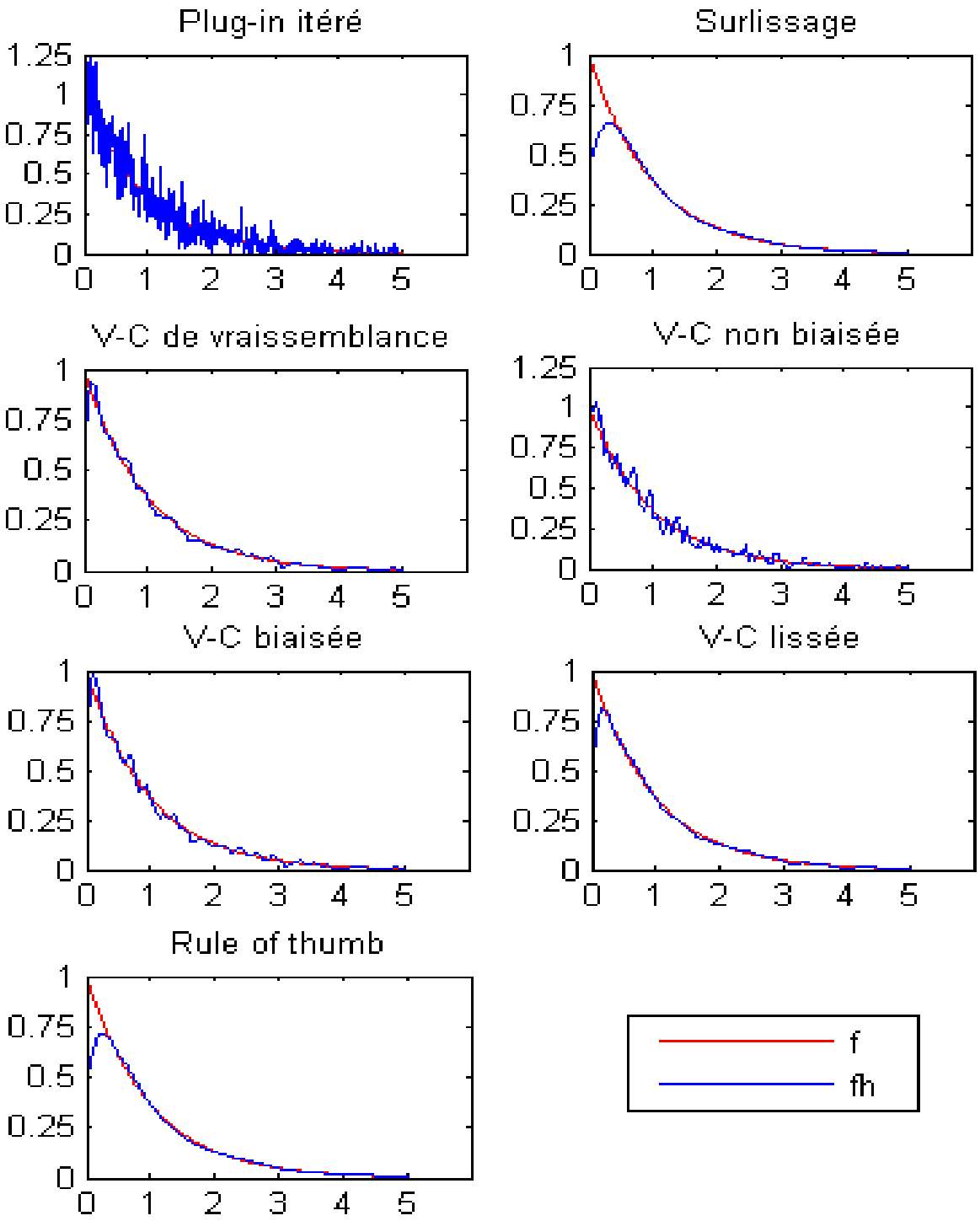

FIGURE 3.7 – Illustration des résultats de simulations effectuées sur la loi exponentielle pour n=2500

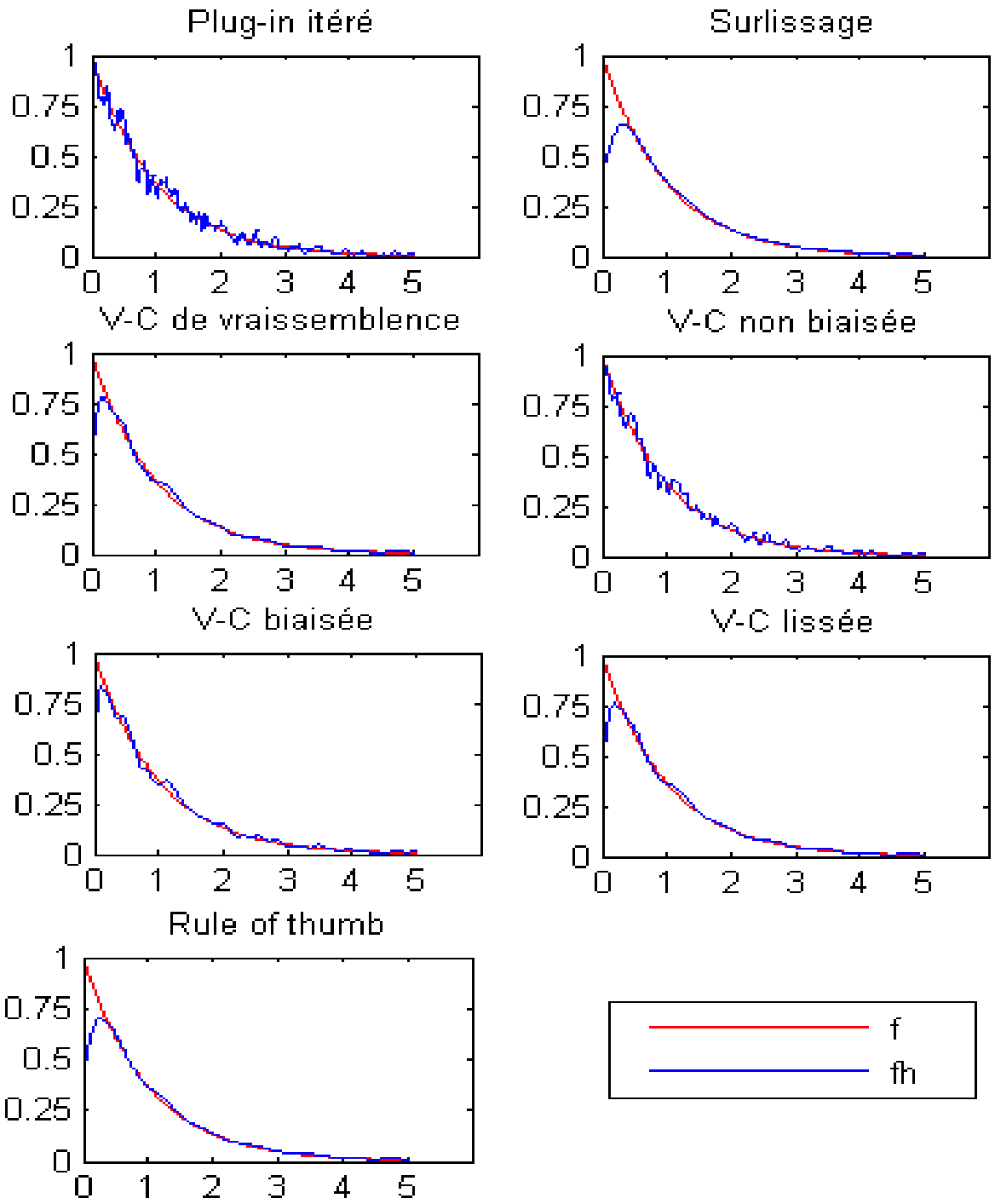

FIGURE 3.8 – Illustration des résultats de simulations effectuées sur la loi exponentielle pour n=4000

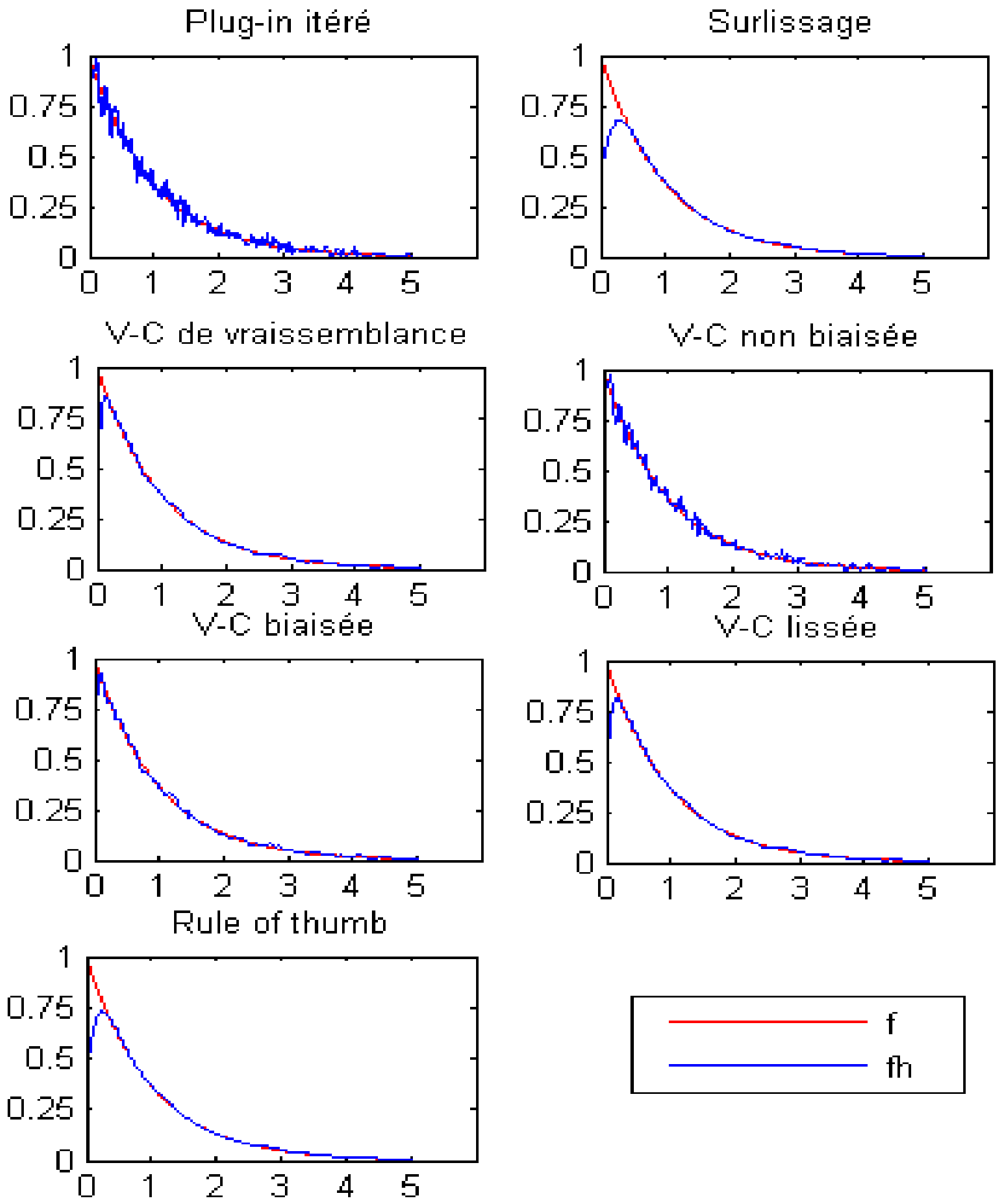

FIGURE 3.9 – Illustration des résultats de simulations effectuées sur la loi exponentielle pour n=6000

Densité khi-deux

n	Méthode	h_{opt}	$crit$	$AMISE(h_{opt})$	h^*	$AMISE(h^*)$	$Eff(\%)$
1000	h_{rot}	0.6466	0.0005	$2.1441.(10^{-3})$	0.3730	$9.4531.(10^{-4})$	44.08
	h_{stt}	0.3818	0.0009	$9.4637.(10^{-4})$			99.88
	h_{os}	0.8237	0.0169	$4.8385.(10^{-3})$			19.53
	h_{lcv}	0.4400	0.1009	$1.0071.(10^{-3})$			93.86
	h_{ucv}	0.4950	-0.1254	$1.1561.(10^{-3})$			81.76
	h_{bcv}	0.4930	0.0007	$1.1490.(10^{-3})$			82.26
	h_{scv}	0.5700	0.0006	$1.5257.(10^{-3})$			61.95
2500	h_{rot}	0.5533	0.0002	$1.1192.(10^{-3})$	0.3105	$4.5417.(10^{-4})$	40.58
	h_{stt}	0.2269	0.0006	$5.2311.(10^{-4})$			86.82
	h_{os}	0.6688	0.0073	$2.1234.(10^{-3})$			21.38
	h_{lcv}	0.3850	0.1038	$5.0764.(10^{-4})$			89.46
	h_{ucv}	0.3040	-0.1289	$4.5458.(10^{-4})$			99.91
	h_{bcv}	0.3260	0.0004	$4.5642.(10^{-4})$			99.50
	h_{scv}	0.4190	0.0003	$5.7029.(10^{-4})$			79.63
4000	h_{rot}	0.5316	0.0001	$9.1284.(10^{-4})$	0.2826	$3.1183.(10^{-4})$	34.16
	h_{stt}	0.2144	0.0004	$3.4950.(10^{-4})$			89.22
	h_{os}	0.6245	0.0055	$1.5985.(10^{-3})$			19.50
	h_{lcv}	0.3660	0.1000	$3.6792.(10^{-4})$			84.75
	h_{ucv}	0.2810	-0.1219	$3.1186.(10^{-4})$			99.99
	h_{bcv}	0.3200	0.0003	$3.2278.(10^{-4})$			96.60
	h_{scv}	0.3960	0.0002	$4.1823.(10^{-3})$			74.56
6000	h_{rot}	0.4725	0.0001	$5.8644.(10^{-3})$	0.2606	$2.2545.(10^{-4})$	38.44
	h_{stt}	0.1976	0.0002	$2.5282.(10^{-3})$			89.17
	h_{os}	0.5462	0.0032	$9.5534.(10^{-3})$			23.59
	h_{lcv}	0.312	0.1039	$2.4322.(10^{-3})$			92.69
	h_{ucv}	0.253	-0.1254	$2.25844.(10^{-3})$			99.82
	h_{bcv}	0.25	0.0002	$2.2621.(10^{-3})$			99.66
	h_{scv}	0.346	0.0001	$2.7584.(10^{-3})$			81.73

TABLE 3.3 – **Résultats des simulations effectuées sur la loi khi-deux pour déterminer les largeurs de fenêtre optimales**

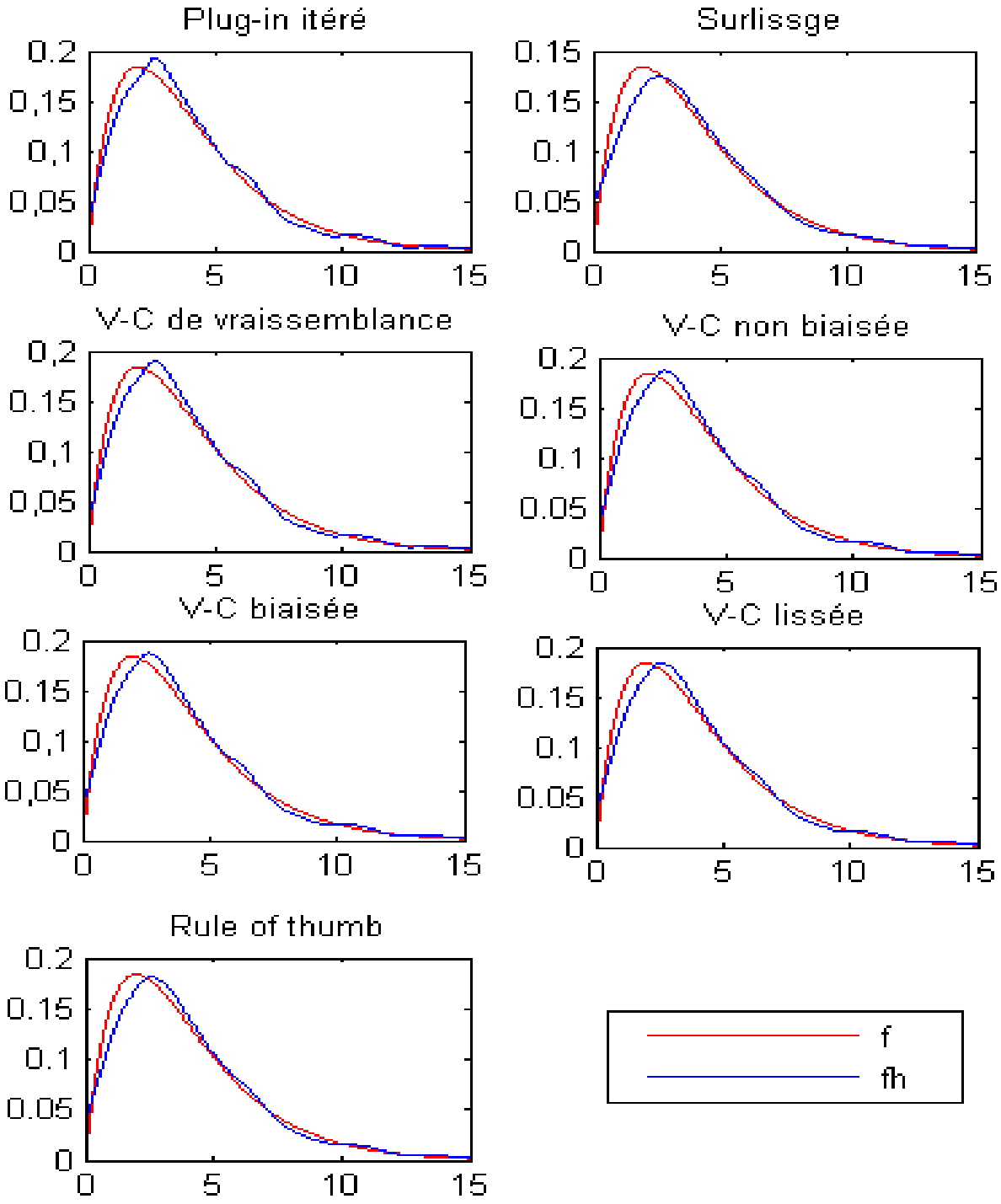

FIGURE 3.10 – Illustration des résultats de simulations effectuées sur la loi khi-deux pour n=1000

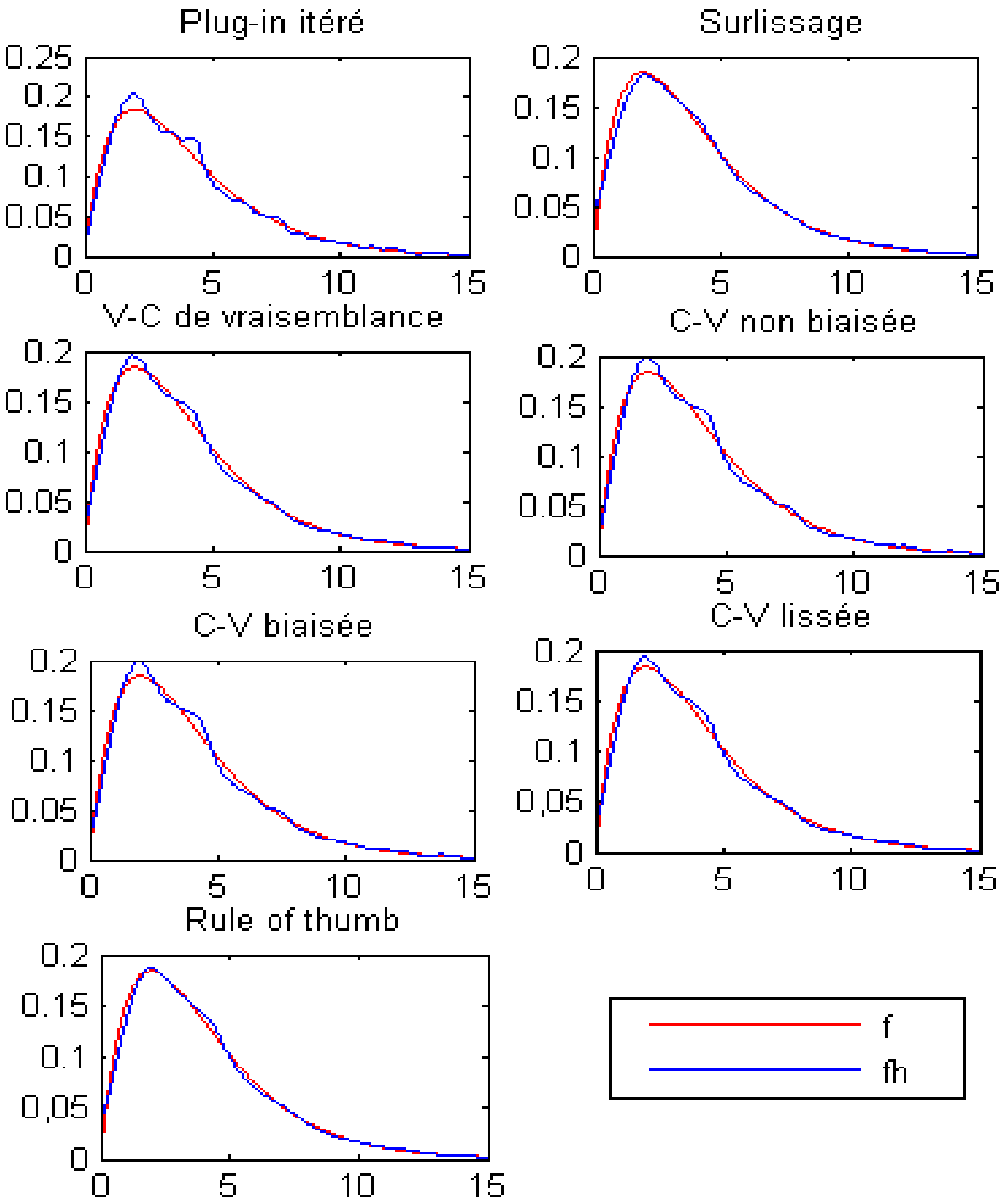

FIGURE 3.11 – Illustration des résultats de simulations effectuées sur la loi khi-deux pour n=2500

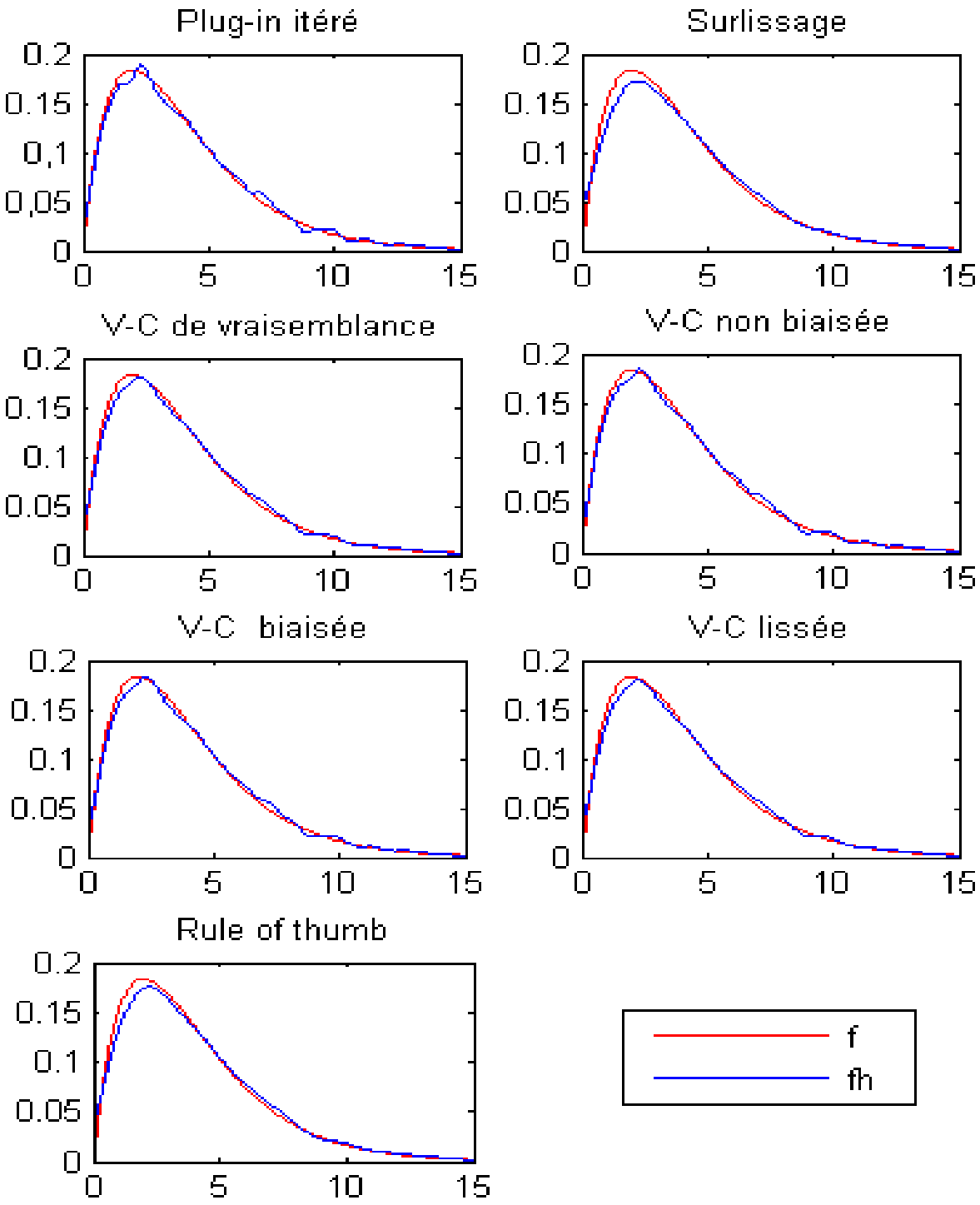

FIGURE 3.12 – Illustration des résultats de simulations effectuées sur la loi khi-deux pour n=4000

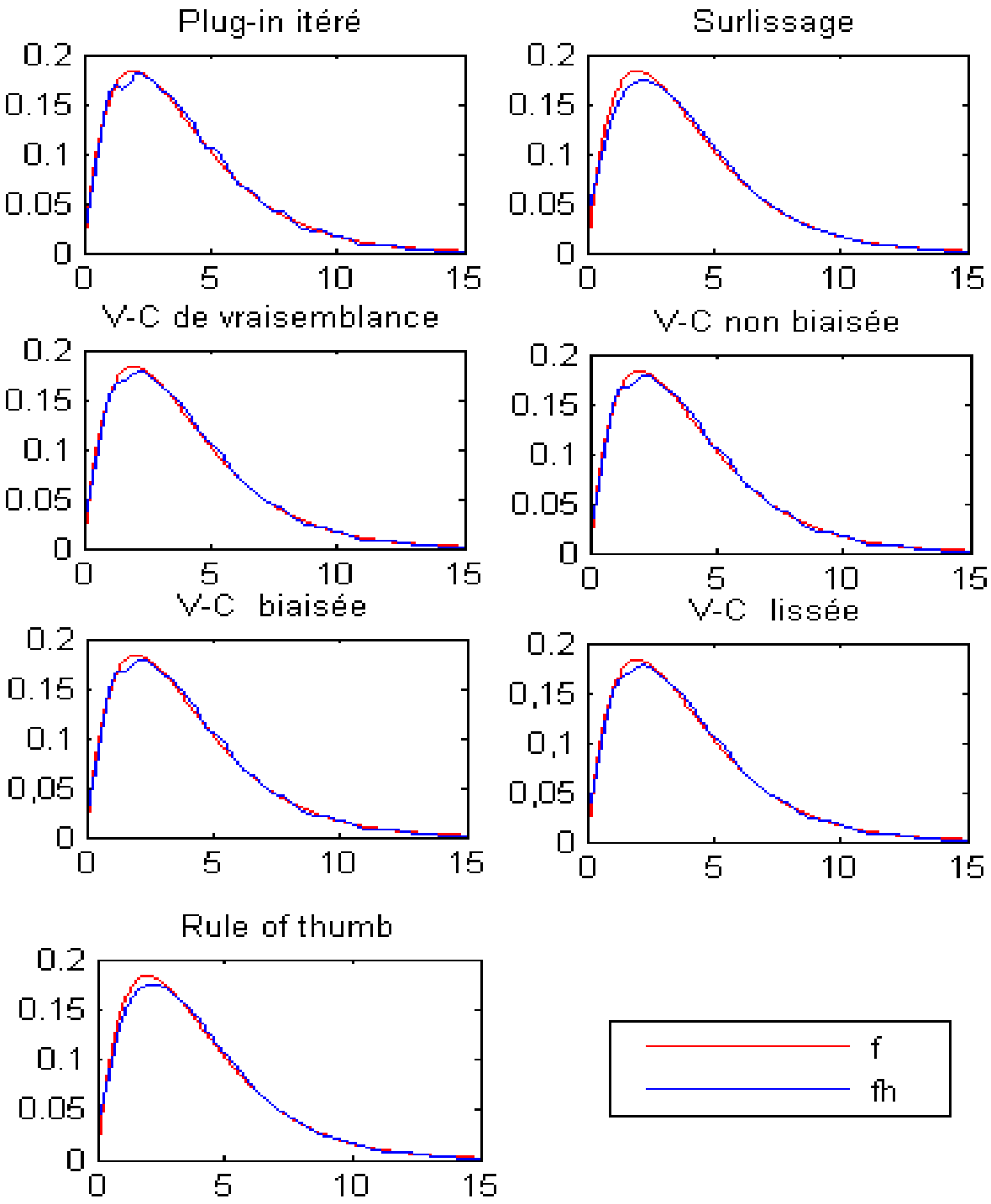

FIGURE 3.13 – Illustration des résultats de simulations effectuées sur la loi khi-deux pour n=6000

3.5 performances des estimateurs

Nous souhaitons tester les différentes méthodes de sélection considérées dans notre étude sur des densités de probabilité connues. Notre objectif est de :

- Vérifier les propriétés d'efficacité et de robustesse de ces méthodes de sélection.
- Comparer ces méthodes de sélection du paramètre de lissage.

Pour cela, nous avons pris comme critère de référence l'estimation de l'efficacité de l'estimateur. Ce critère a été proposé par Scott, Tapia et Thompson [25], Park et Marron [17].

3.5.1 Loi normale centrée réduite

Les figures [3.2] à [3.5] montrent les courbes de densité test et celles estimées par la méthode du noyau pour des données de nature gaussienne. La première lecture du tableau (3.1) indique que :

- Les méthodes de sélection plug-in et validation croisée fournissent des valeurs très voisines des valeurs h^* ;
- La taille de l'échantillon n influe sur les valeurs h_{opt}, h^*, Crit, AMISE*. On constate que l'augmentation de n entraine la décroissance de h_{opt}, h^*, $crit$ et $AMISE^*$;
- Les performances moyennes des estimateurs s'améliorent lorsque la taille de l'échantillon augmente. Cette observation est confirmée sur les graphes, où la courbe de l'estimateur f_h se rapproche de la densité cible f quand la taille de l'échantillon augmente : les estimateurs f_h semblent appartenir à une suite convergeant vers f.

Estimation de l'efficacité :

Les résultats indiquent que, toutes les méthodes de sélection estiment correctement la valeur du paramètre de lissage h. Cette constatation est confirmée par une analyse de la dernière colonne du tableau (3.1), qui donne une valeur élevée de l'efficacité de l'estimateur h. Cette efficacité croît quand on augmente la taille de l'échantillon.

On peut aussi remarquer que :

- les performances de l'estimateur h_{stt} se dégradent lorsque la taille d'échantillon augmente. Cette observation est confirmée par l'analyse de la dernière colonne du tableau (3.1), par exemple pour n=2500(4000),Eff= 0.97(0.63). ce résultat est expliqué par le défaut de convergence de cette méthode.
- les plus grandes valeurs de Eff sont obtenues par la méthode "rule of thumb", où par exemple pour la taille de l'échantillon n= 6000, la valeur de Eff est égale à 0.99. Ce

résultat est expliqué par le principe de la méthode rule of thumb : qui consiste à choisir f comme étant une loi gaussienne ;

- le paramètre de lissage h_{os} a donné de bons résultats dans ce cas, car la densité test à estimer présente un faible niveau de complexité structurelle (unimodalité, lisse et régulière) ;
- Parallèlement aux méthodes de sélection plug- in, les méthodes de validation croisée se sont également révélées performantes. les résultats obtenus par ces méthodes sont proches des résultats obtenus par les méthodes plug-in. La valeur du paramètre de lissage estimé par les différentes méthodes de sélection validation croisée est très proche de la valeur h^* et la qualité des estimations h_{ucu}, h_{bcu}, h_{lcu} et h_{scv} s'améliorent lorsque la taille d'echantillion devient grande ;

Les méthodes de validation croisée en particulier, les méthodes validation croisée non biaisée et validation croisée biaisée souffrent d'un défaut important dit de variabilité qu'on discutera plus bas.

3.5.2 Loi exponentielle

Nous supposons maintenant que l'ensemble des densités possibles est réduit à la densité $f(x) = exp(-x)$. On dit que ces distributions sont à queue lourde.
Les figures [3.6] à [3.9] montrent les courbes de densité test et celles estimées par la méthode de noyau pour des données de nature exponentielle.
La première visualisation des graphes et la lecture du tableau (3.2) nous permet de remarquer que :

- Les échantillons suivant une telle loi ne peut conduire à une estimation correcte de cette loi ;
- Dans ce cas, les valeurs du paramètre de lissage calculées par les différentes méthodes de sélection ont tendance à une surestimation (donnant un lissage un peu plus accentué que l'optimum h^*), ou un sous lissage (donnant des valeurs largement trop petites que h^*) ;
- Excepté pour les méthodes rule of thumb, validation croisée lissée et validation croisée de la vraissemblance qui rapprochent la partie droite de la courbe théorique, les résultats obtenus par les autre méthodes (plug-in et validation croisée) sont fort décevants. En effet, l'estimateur de la densité présente des pics significatifs ;
- Les performances de l'estimateur calculées par les différentes méthodes se dégradent lorsque la taille de l'échantillon augmente, les estimateurs f_h ne semble pas appartenir à une seule suite qui converge vers f.

Estimation de l'efficacité :

Les résultats obtenus dans ce cas sont portés dans le tableau (3.2). Ces résultats confirment le comportement de la courbe de la densité estimée. Les valeurs de l'estimateur de l'efficacité demeurent faibles dans ce cas. La plus grande valeur de l'estimateur de l'efficacité Eff est obtenue par la méthode rule of thumb, la méthode oversmoothing et la méthode validation croisée lissée. Pour expliquer ces mauvais résultats, il suffit de constater que pour la densité exponentielle, le paramètre de lissage asymptotique optimal h_{opt} est manifestement trop petit (les courbes d'estimation sont très irrégulières), même pour une grande taille d'échantillon.
La difficulté de l'estimateur à noyau apparaît lorsque la méthode est appliquée à des observations dont la distribution présente de grandes queues. Il est en effet très difficile dans ce cas de choisir le paramètre de lissage h de façon optimale, vu le phénomène de sous lissage, et avoir une très grande erreur sur les queues.

3.5.3 Loi de khi-deux :

On se place maintenant dans le cas où l'ensemble des densités est réduit à la loi du χ^2 à 4 degrés de liberté (ddl=4). On dit que ces distributions sont unimodales convergent lentement vers 0.
La loi khi-deux est définie comme suit :

$$f(x) = \frac{xexp(\frac{-x}{2})}{4} \ , x \geq 0$$

Les figures [3.10] à [3.13] sont composées de graphe de l'estimateur considéré et le graphe de la densité f de loi χ^2.
La première lecture du tableau et des graphes indique que :

- Les valeurs du paramètre de lissage estimées par les différentes méthodes de sélection sont nettement plus grandes que la largeur de fenêtre optimale h^* ;
- L'augmentation de la taille de l'échantillon entraîne la décroissance des valeurs théoriques h^*, $AMISE^*$ et h_{opt}, $crit$;
- Comme le montre les figures, ce genre d'estimation n'est pas vraiment satisfaisant. Deux défauts peuvent être notés : L'estimateur obtenue ne converge que très lentement et ne converge pas au voisinage de 0.

Estimation de l'efficacité :

Les performances des estimateurs plug-in et validation croisée sont nettement moins bonnes que celle obtenues dans le cas où la densité est gaussienne. Cette constatation est confirmée par les valeurs moins élevées de l'estimation de l'efficacité Eff données dans le tableau (3.3).

En effet, il est très peu probable que des observations de l'échantillon proviennent de la queue. L'estimation est tronquée malgré l'augmentation du nombre d'observations n.
Concrètement, dans le cas d'une augmentation du nombre d'observation n, l'estimation de la fonction prend une allure multimodale avec un vrai mode, puis un ou plusieurs modes générés par les données issues de la queue.
Nous pouvons aussi observer que les meilleurs résultats dans ce cas sont obtenus par les deux méthodes plug-in et par les deux méthodes validation croisée (validation croisée non biaisée, validation croisée biaisée) et on notera que les performance relatives des estimateurs s'améliorent avec la taille de l'échantillon. En revanche les résultats obtenus par les autres méthodes plug-in et validation croisée semblent de mauvaises qualités par rapport aux approches précédentes.

3.6 Conclusion :

Les résultats de simulation mettent en relief le problème bien connu du choix de la fenêtre. En effet, la performance de la procédure de sélection du paramètre de lissage varie en fonction de la densité à estimer. Cependant, plusieurs points importants peuvent être soulignés sur les méthodes plug-in et validation croisée :

- Les formules qui sont au cœurs des méthodes plug-in imposent des restrictions sur la densité inconnue f (par exemple f doit être suffisamment lisse et régulière) souvent difficiles à vérifier. Par exemple, L'expression classique (2.3) de la largeur de fenêtre asymptotique optimale n'est plus valide pour une densité à queue lourde (loi exponentielle) ;
- L'intérêt principal des méthodes de validation croisée est d'être complétement automatique. Un inconvénient de ces méthodes est que le paramètre de lissage calculé par ces méthodes pourrait ne pas convenir d'un échantillon à un autre ;
- Les méthodes plug-in se révèlent plus stables que les méthodes validation croisée en particulier, validation croisée biaisée et non biaisée ;
- L'expérience montre qu'il n'existe pas de méthode de sélection du paramètre de lissage qui soit intrinsèquement meilleure que toutes les autres ;
- Enfin, d'une certaine manière, ces résultats nous amènent à considérer que les performances des estimateurs à noyau varient suivant la densité à estimer. Ces estimateurs se comportent en effet très bien pour les densités gaussiennes (D_1). Il sont moins performants pour les cibles plus complexes (unimodale avec queue par exemple D_3) et ils sont en revanche non valide pour une densité cible exponentielle (D_2 avec queue lourde). L'utilisation d'un nombre d'observation de plus en plus important, ne garantit pas une bonne estimation.

Conclusion générale

Ce travail est une contribution au problème du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau. Nous avons étudié l'applicabilité, l'efficacité et la robustesse des différentes techniques de sélection du paramètre de lissage. Ces performances ont été mesurées numériquement à l'aide de jeux de données simulés.

Dans une première partie, nous avons exposé les différentes méthodes d'estimation de la densité de probabilité à savoir l'estimation par l'histogramme et l'estimation par la méthode du noyau. Nous nous sommes intéressés à la méthode du noyau parce qu'elle est "populaire" vu sa souplesse d'utilisation et elle présente de bonnes propriétés asymptotiques. L'estimateur à noyau est une fonction de deux paramètres : la fonction K appelée noyau et h appelé paramètre de lissage ou fenêtre. Si le choix du noyau K n'est pas un problème dans l'estimation de la densité de probabilité, il n'en est pas de même pour le choix du paramètre de lissage h qui ne dépend que de la taille de l'échantillon.

Dans la deuxième partie, nous avons exposé les différentes méthodes de sélection du paramètre de lissage, les méthodes reposant sur la validation croisée dont l'intérêt est le caractère direct, et l'autre classe de méthode dite plug-in (ré-injection) qui reposent sur l'estimation de la quantité $R(f'')$ inconnue.

Enfin, afin de tester l'applicabilité des différentes méthodes de sélection du paramètre de lissage, nous avons simulé des densités de probabilité tests présentant différents aspects (loi normale, loi exponentielle et loi Khi-Deux). L'algorithme de simulation que nous avons utilisé a permis :

- D'estimer le paramètre de lissage par les différentes méthodes de sélection ;
- De construire l'estimateur de la densité de probabilité par la méthode du noyau et de donner une représentation graphique de la densité test et celle estimée.

Les résultats numérique obtenus montrent que :

- La sélection du paramètre de lissage n'est valable qu'asymptotiquement. A taille d'échantillion fixée, l'analyse s'avère délicate. Ensuite, les formules qui sont au cœurs des méthodes plug-in imposent des restrictions sur la densité inconnue f souvent difficiles à vérifier dans la pratique ;

IL est essentiel de comprendre qu'il n'existe pas de méthode de sélection du paramètre de lissage qui soit meilleure que les autres. L'expérience montre que chaque algorithme de sélection possède, en quelque sort, ses densités de prédilection : un algorithme fonctionnera par exemple correctement pour des densités unimodales, alors qu'un second fournira de bien meilleurs résultats pour d'autres densités tests.

Les résultats de simulation confirment également l'inconvénient principal des méthodes validation croisée, la largeur de fenêtre estimée par cette technique présente une grande variabilité, c'est à dire que pour deux échantillons distincts issus de la même distribution, les fenêtres obtenues seront très différentes. Cette méthode présente cependant de nombreux avantages : outre le fait qu'elle ne demande pour être applicable, que des hypothèses faibles sur le degré de différentiabilité de f, c'est une méthode automatique entièrement guidée par les données.

Annexe

Loi normale

En théorie des probabilités et en statistique, la loi normale est l'une des lois de probabilité les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. Elle est également appelée loi gaussienne, loi de Gauss ou loi de Laplace-Gauss.
Plus formellement, c'est une loi de probabilité absolument continue qui dépend de deux paramètres : son espérance, un nombre réel noté μ, et son écart type, un nombre réel positif noté σ.
Une variable aléatoire réelle X suit une loi normale d'espérance μ et d'écart type σ, si cette variable aléatoire réelle X admet pour densité de probabilité la fonction $f(x)$ définie, pour tout nombre réel x, par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Une loi normale sera notée de la manière suivante $N(\mu, \sigma)$.

▷ Espérance et variance :

Soit X une variable aléatoire qui suit $N(\mu, \sigma)$ alors
$\mathbb{E}(x) = \mu$.
$\mathbb{V}(x) = \sigma^2$.

Lorsque la moyenne μ vaut 0, et l'écart-type vaut 1, la loi sera notée $N(0, 1)$ et sera appelée loi normale standard. Seule la loi $N(0, 1)$ est tabulée car les autres lois (c'est-à dire avec d'autres paramètres) se déduise de celle-ci à l'aide du théorème suivant :
Si Y suit $N(\mu, \sigma)$ alors $Z = \frac{Y-\mu}{\sigma}$ suit $N(0, 1)$.
On note Φ la fonction de répartition de la loi normale centrée réduite :
$\Phi(x) = P(Z < x)$ avec Z une variable aléatoire suivant $N(0, 1)$.

Loi exponentielle

Une loi exponentielle modélise la durée de vie d'un phénomène sans mémoire, ou sans vieillissement, ou sans usure : la probabilité que le phénomène dure au moins $s + t$ heures sachant qu'il a déjà duré t heures sera la même que la probabilité de durer s heures à partir de sa mise en fonction initiale. En d'autres termes, le fait que le phénomène ait duré pendant t heures ne change rien à son espérance de vie à partir du temps t.

La densité de probabilité de la distribution exponentielle de paramètre $\lambda > 0$ prend la forme :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \ x \geqslant 0, \\ 0 & , \ x < 0. \end{cases}$$

La distribution a pour support l'intervalle $[0,+\infty[$.

La fonction de répartition est donnée par :

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & , \ x \geqslant 0, \\ 0 & , \ x < 0. \end{cases}$$

.

▷ L'espérance et la variance :

$\mathbb{E}(X) = \frac{1}{\lambda}$.
$V(X) = \dfrac{1}{\lambda^2}$.

Loi du χ^2(khi-deux) :

Définition :

Soit Z_1, Z_2,.....Z_n une suite de variables aléatoires indépendantes de même loi $N(0,1)$.
Alors la variable aléatoire $\sum_{i=1}^{n} Z_i^2$ suit une loi appelée loi du $Khi - deux$ à n degrés de liberté, notée $\chi^2_{(n)}$

▷ Sa fonction caractéristique est $(1 - 2it)^{\frac{-n}{2}}$.

▷ La densité de la loi du χ^2 est

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{\frac{-x}{2}}, \ pour x > 0, \\ 0 \qquad\qquad , \ sinon. \end{cases}$$

où Γ est la fonction Gamma d'Euler définie par $\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x} dx$

▷ L'espérance et la variance de la loi du χ^2 :

$\mathbb{E}(x) = n.$

$\mathbb{V}(x) = 2n.$

▷ La somme de deux variables aléatoires indépendantes suivant respectivement $\chi^2_{(n_1)}$ et $\chi^2_{(n_2)}$ suit aussi une loi du χ^2 avec $n_1 + n_2$ degrés de liberté.

Bibliographie

[1] Berlinet, A. Devroye, L. A comparison of kernel density estimates. *Pub.Inst. Stat. Univ. Paris XXXVIII.* (1994).

[2] Bowman, A. W. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*, 2 (1984), 353–360.

[3] Burman, P. A date dependent approach to density estimation. *Zeitschrift Für Wahrscheinlichkeitstheorie and Verwandte Gebiete, 69* (1985), 609–628.

[4] Cao, R. Cuevas, A., and Gonzdez-Manteiga, W. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis 17(1)* (1994), 153–176.

[5] Coudret, R., Durrieu, G., and Saracco, J. Estimateurs a noyau bimodaux d'une densité bimodale et comparaison avec d'autres estimateurs non paramétriques. *Proc de la société Française de Statistique* (2012).

[6] Cybakov, A. B. *Introduction à l'estimation non paramétrique*, vol. 41. Springer Science & Business Media, 2003.

[7] Deheuvels, P., and Hominal, P. Estimation non paramétrique de la densité compte tenu d'informations sur le support. *Revue de Statistique Appliquée*, 27, 47,68 (1979).

[8] Epanechnikov, V. Nonparametric estimation of a multidimensional probability density. *Teoriya Veroyatnostei i ee Primeneniya 14*, 1 (1969), 156–161.

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

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

[11] Hall, P., Marron, J. S., and Park, B. U. Smoothed cross-validation. *Probability Theory and Related Fields 92*, 1 (1992), 1–20.

[12] Hebbema, J.D.F. Hermans, J., and Vandenbrok, K. A stepwise discriminant analysis program using density estimation. *in compstat, ed. G.Bruckmann, Wien Physica-Verlag* (1974), 101–110.

[13] Hodges, J., and Lehmann, E. The efficiency of some nonparametric competitors of the t-test. *Annals Mathematicals statistzcs 27(1)* (1956), 324–335.

[14] Lejeune, M. *Statistique. La théorie et ses applications.* Springer, France, 2004.

[15] Marron, J. S., and Ruppert, D. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* (1994), 653–671.

[16] Nadaraya, E. On non parametric estimation density function and regression. *Theory Probab P.P.L* (1965).

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

[18] Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statist. 33*, 3 (09 1962), 1065–1076.

[19] Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics 27*, 3 (Sept. 1956), 832–837.

[20] Roussas, G. G. Asymptotic normality of the kernel estimate of a probability density function under association. Statist. Probab. Lett. *Statistics &amp ; Probability Letters 50*, 1 (2000), 1–12.

[21] Rudemo, M. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* (1982), 65–78.

[22] Schuster, E.F. Gregory, G. On the nonconsistency of maximum likelihood nonparametric density estimators. *Computer Science and Statistics : Poceeding ofthe* 13^{th} *Symposium on the Interface, ed. W.F.Eddy, New York Spring Verlag* (1981), 295–298.

[23] Scott, D. W. *Multivariate Density Estimation :Theory, Practice, and Visualization.* Wiley Interscience, New York, 1992.

[24] Scott, D. W., and Factor, L. E. Monte carlo study of three data−based nonparametric probability density estimators. *Journal of the American Statistical Association 76*, 373 (1981), 9–15.

[25] Scott, D. W., Tapia, R. A., and Thompson, J. R. Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *The annals of statistics* (1980), 820–832.

[26] Scott, D. W., and Terrell, G. R. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association 82*, 400 (1987), 1131–1146.

[27] Sheather, S.J. Jones, M. A reliable data-based bandwith selection method for kernel density estimation. *Journal of the Royal Statistical Society,B53* (1991), 683–690.

[28] Silverman, B. W. *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[29] Simonoff, J. S. *Smoothing methods in statistics*. Springer Science & Business Media, 2012.

[30] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)* (1974), 111–147.

[31] Strauss, O. Estimation imprécise de densité de probabilité par transfert imprécis de comptage. GRETSI − CNRS, Saint Martin d'Hères (France)(2005), pp. 165–168.

[32] Wahba, G. Optimal properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Of Statist,3* (1975), 15–29.

[33] Wand, M. P., and Jones, M. C. *Kernel smoothing*. Crc Press, 1994.

[34] Wansouwé, W. E., Kokonendji, C. C., and Kolyang, D. T. Nonparametric estimation for probability mass function with Disake.

[35] Woodroofe, M. On choosing a delta sequence. *Annals Mathematicals statistzcs 49* (1970), 1665–1671.

Résumé : L'estimation non paramétrique de la densité de probabilité par la méthode du noyau est caractérisé par le choix du noyau K et du paramètre de lissage h appelé aussi fenêtre de lissage.
Le choix du paramètre de lissage est crucial pour la précision locale que pour la précision globale de l'estimateur.
Dans ce travail, nous avons étudié les méthodes de sélection de ce paramètre les plus fréquentes dans la littérature à savoir la famille des méthodes plug-in et la famille des méthodes cross validation.
L'étude comparative entre ces méthodes montre que la performance de chacune d'elle est principalement dépendante de la densité de probabilité à estimé par la méthode.
Chaque algorithme de sélection possède ses densités de prédilection.

Mots clés : estimation non paramétrique de densité, noyaux, paramètre de lissage, plug-in, validation croisée.

Abstract : In non-parametric estimation of the probability density by the kernel method , characterized by selecting the kernel K and the smoothing parameter h, also called smoothing window. The choice of the latter is crucial, both for local accuracy for the overall accuracy of the estimator. In this paper we study the selection methods most frequent in the literature : family of plug-in methods and family of cross validation methods. comparing these methods to simulated data shows that the performance of each method is mainly dependent on the density felt by this last , each selection algorithm possesses , somehow , his favorite densities.

Key-words : Nonparametric estimates of a density function, kernels, Smoothing parameter, plug-in, cross validation.