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

Université A. MIRA de Béjaia
Faculté des Sciences Exactes
Département de Mathématiques

Cours

Estimation non paramétrique

N. SAADI

Table des matières

Chapitre 1

Estimation non paramétrique de la densité de probabilité

1.1 Introduction

Un problème récurrent en statistique est celui de l'estimation d'une densité f à partir d'un échantillon de variables aléatoires réelles $X_1, ..., X_n$ indépendantes et de même loi inconnue. On peut se demander, pour quelles raisons on cherche à estimer la fonction de densité de probabilité. Les raisons sont multiples dont les plus importantes sont :

– La densité nous fournit des idées très claires sur le comportement de la distribution : maximaux locaux , points de symétrie, dispersion et localisation de la distribution ;

– Les ordinateurs utilisent pour la simulation des observations accordées pour une certaine distribution. La connaissance de la fonction densité de probabilité est alors indispensable ;

– Pour certains problèmes statistiques, les tests basés sur les estimateurs de densité sont supérieurs aux tests basés sur d'autres fonctions.

Il existe deux approches largement utilisées pour l'estimation de la densité de probabilité l'approche paramétrique et l'approche non paramétrique : L'approche paramétrique a comme inconvénient principal la connaissance au préalable de la loi du phénomène étudié. L'approche non paramétrique estime la densité à partir de l'information disponible. On dit souvent que dans cette approche les données parlent d'elles mêmes. L'avantage principal de l'estimation non paramétrique de la densité de probabilité est de ne pas nécessiter d'hypothéses à priori sur l'appartenance de cette densité à une famille de lois connues.

1.2 Critères d'erreurs

Pour mesurer les performances théoriques des estimateurs et identifier le meilleur, il est nécessaire de spécifier un critère d'erreur. Nous considérons la densité de probabilité f et

son estimateur f_n.

- **L'erreur quadratique intégrée ISE :**

$$ISE(f, f_n) = \int [f(x) - f_n(x)]^2 \, dx.$$

- **L'erreur quadratique moyenne MSE :**

$$MSE(f(x), f_n(x)) = \mathbb{E}\left[f(x) - f_n(x)\right]^2 = \mathbb{V}ar\left(f_n(x)\right) + \mathbb{B}iais^2\left(f_n(x)\right).$$

- **L'erreur quadratique moyenne intégrée MISE :**

$$MISE(f, f_n) = \int_{\mathbb{R}} \mathbb{E}\left[f(x) - f_n(x)\right]^2 dx = \int_{\mathbb{R}} \left[\mathbb{V}ar\left(f_n(x)\right) + Biais^2\left(f_n(x)\right)\right] dx.$$

1.3 Qelques définitions

Définition 1.1. On dit qu'un estimateur f_n de f est sans biais si : $\mathbb{E}(f_n) = f$.

Définition 1.2. On dit qu'un estimateur f_n de f est asymptotiquement sans biais si :

$$\lim_{n\to\infty} \mathbb{E}(f_n(x)) = f(x), \quad en \ \ tout \ \ point \ \ x \ \ pour \ \ lequel \ \ la \ \ densité \ \ f \ \ \ est \ \ continue.$$

Définition 1.3. On dit qu'un estimateur f_n de f est ponctuellement consistant en moyenne quadratique si :

$$\lim_{n\to\infty} MSE(f(x), f_n(x)) = 0, en \ \ tout \ \ point \ \ x \ \ pour \ \ lequel \ \ la \ \ densité \ \ f \ \ est \ \ continue.$$

Définition 1.4. On dit qu'un estimateur f_n de f est ponctuellement consistant en moyenne quadratique si :

$$\lim_{n\to\infty} MSE(f(x), f_n(x)) = 0, en \ \ tout \ \ point \ \ x \ \ pour \ \ lequel \ \ la \ \ densité \ \ f \ \ est \ \ continue.$$

Définition 1.5. On dit qu'un estimateur f_n de f est uniformément consistant en moyenne quadratique intégrée si :

$$\lim_{n\to\infty} MISE(f(x), f_n(x)) = 0, en \ \ tout \ \ point \ \ x \ \ pour \ \ lequel \ \ la \ \ densité \ \ \ f \ \ est \, écontinue.$$

1.4 Estimation non paramétrique par Histogramme

Etant données des observations $x_1, ..., x_n$ qui sont les réalisations des variables aléatoires réelles indépendantes et identiquement distribuées $X_1, X_2, ..., X_n$ de densité f inconnu sur l'intervalle $[a, b]$. Construire un histogramme consiste à partitionner l'intervalle $[a, b]$ en $p \in \mathbb{N}^*$ classes A_k, $k \in \{1, ..., p\}$, et à compter le nombre d'observations appartenant à

chaque classe A_k. Si toutes les classes de l'histogramme ont la même largeur, on dit que l'histogramme est régulier. On note $h \in \mathbb{R}$, la largeur des classes qui est alors appelée le paramètre de lissage. Le nombre d'observations appartenant à chaque classe A_k est appelé accumulateur de la classe A_k est noté $Acck = \sum_{i=1}^{n} 1_{A_k}(x_i)$, la probabilité de A_k (basée sur les observations x_i), notée $\mathbb{P}(A_k)$, est donnée par :

$$\mathbb{P}(A_k) = \frac{A_{cc_k}}{n}. \tag{1.1}$$

En émettant l'hypothèse, que les observations se répartissent uniformément dans la cellule A_k (de largeur h), on peut alors construire un estimateur de la densité f, pour tout $x \in [a, b]$, par :

$$f_h(x) = \frac{1}{h} \sum_{k=1}^{p} \mathbb{P}(A_k) 1_{A_k}(x) = \frac{1}{nh} \sum_{k=1}^{p} A_k 1_{A_k}(x). \tag{1.2}$$

Dans la suite, nous émettons l'hypothèse, que les classes $A_k \forall k \in \{1, \ldots, p\}$ forment une partition de $[a, b]$ et définissons pour chaque classe A_k, son centre a_k telles que

$$\forall k \in \{1, ..., p\}, A_k = [a_k - \frac{h}{2}, a_k + \frac{h}{2}] \quad et \quad \forall k \in 1, \ldots, p, a_{k+1} = a_k + h.$$

Remarque 1.4.1. Histogramme et densité de probabilité sont liés par des conditions aux limites : une densité de probabilité peut être vue comme la limite d'un histogramme lorsque le nombre d'observations tend vers l'infini et que la granularité de l'histogramme tend vers zéro. La figure [1.1] présente un histogramme de 100 observations tirées aléatoirement d'une loi normale centrée réduite $\mathcal{N}(0, 1)$. Ces observations sont réparties sur un intervalle de référence $A = [-5, 5]$. La largeur de l'histogramme est $h = 0.2$.

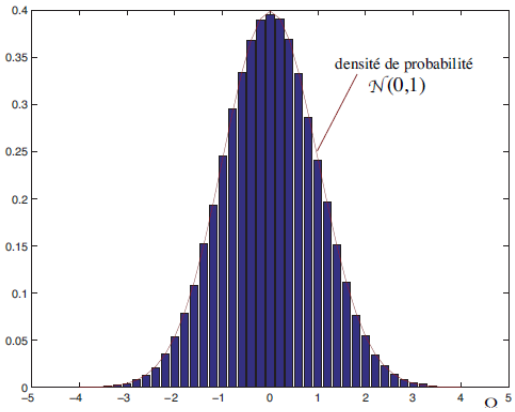


FIGURE 1.1 – Histogramme construit avec une largeur $h = 0.2$

.

1.4.1 Propriétés statistiques de l'histogramme

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, 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. On essaye généralement de réduire au mieux ces deux quantités.

- Le biais de l'estimateur est donné pour tout $x \in [a_k, a_{k+1}]$ par :

$$\mathbb{B}iais\left(f_h(x)\right) = \mathbb{E}\left(f_h(x)\right) - f(x) = \frac{1}{2}f'(x)(h - 2 - (x - a_k)) + O(h^2), \tag{1.3}$$

  ou O est un terme résiduel et f' est la dérivée de f. f doit être une fonction de $L^2([a, b])$ absolument continue et carrée integrable .

- La variance de l'estimateur est donnée pour tout $x \in [a_k, a_{k+1}]$ par :

$$\mathbb{V}ar\left(f_h(x)\right) = \mathbb{E}\left(f_h^2(x)\right) - \mathbb{E}^2\left(f_h(x)\right) = \frac{f(x)}{nh} + o(n^{-1}). \tag{1.4}$$

  Discussion du comportement du biais et de la variance :

  * Le biais décroit si h diminue mais la variance augmente.
  * Pour que la variance tende vers 0, il faut que $nh \to \infty$.
  * La variance diminue si h augmente mais le biais augmente.

  Il s'ensuit que :

$$MSE\left(f_h(x)\right) = \frac{f(x)}{nh} + \frac{f'(x)^2}{4}(h - 2 - (x - a_k))^2 + O(h^3) + O(n^{-1}), \tag{1.5}$$

$$MISE\,(f_h(x)) = \frac{1}{nh} + \frac{h^2 \int f'(t)^2 dt}{12} + O(h^3) + O(n^{-1}). \tag{1.6}$$

Remarque 1.4.2. Dans 1.6, on voit que : un petit h donne un histogramme peu biaisé, tandis qu'un grand h et un grand n déterminent un histogramme moins variable.

1.4.2 Choix du paramètre h

Définition 1.6. (**Règle de Scott.**) La valeur qui minimise l'erreur quadratique moyenne intégrée, $MISE$ est :

$$h_{opt} = \left[\frac{6}{\int f'(t)^2 dt}\right]^{\frac{1}{3}} n^{-\frac{1}{3}}.$$

Si f la densité de loi normale $\mathcal{N}(\mu, \sigma)$, alors

$$hopt = 3.491 \sigma n^{-\frac{1}{3}}.$$

En estimant σ par l'écart-type S de l'échantillon, on obtient ainsi la règle de Scott.

$$h_{opt} = 3.491 S n^{-\frac{1}{3}}.$$

Définition 1.7. (**Règle de Sturges.**) Prendre le nombre k de classes égal à $1 + \log_2 n$. En pratique, cela revient à prendre $h = \frac{x_{(k)} - x_{(1)}}{k}$, $x_{(i)}$ sont les valeurs d'observations d'un échantillon ordonné par ordre croissant. La règle de Sturges a tendance à produire des histogrammes trop lisses.

Remarque 1.4.3. La plus grande qualité de l'histogramme est sa simplicité. Parmi ses inconvénients importants, citons celui d'être trop peu sensible aux propriétés locales de f. En outre, alors que la plupart des fonctions de densité ne sont pas des fonctions en escalier, l'histogramme est un estimateur toujours de cette forme. L'application de certaines opérations sur l'estimé, comme par exemple une dérivée ou une intégration, devient alors impossible ou très difficile à effectuer. De plus, si on a un paramètre de lissage h trop petit, cela conduit à un histogramme plus découpé, tandis qu'à un paramètre de lissage h trop grand résulte un histogramme plus lissé.

Il existe d'autres méthodes non paramétriques plus robustes que la méthode par histogramme : la méthode d'estimation par des séries orthogonales et la méthode du noyau.

1.5 Estimation non paramétrique de la densité par des séries orthogonales

Définition 1.8. Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées de densité de probabilité f par rapport à la mesure de lebesgue sur $\mathbb{R}$, il s'agit d'estimer f à partir des observations $x_1, ..., x_n$. Pour cela on suppose que :

1. L'espace de Hilbert $\mathbb{L}^2$ est de dimension infinie ;
2. $\{e_k, k \in \mathbb{N}\}$ un système orthogornal dans $\mathbb{L}^2$;
3. $f \in \mathbb{L}^2$ tel que :

$$f(x) = \sum_{k=0}^{\infty} a_k e_k(x), k = 0, \ldots, x \in \mathbb{R}; \tag{1.7}$$

4. Avec $\{a_k, k \in \mathbb{N}\}$ sont les coefficients de Fourier associés à f donnés par :

$$a_k = \int_{\mathbb{R}} e_k(x) f(x) dx = \mathbb{E}[e_k(X)], k = 0, 1, ... \tag{1.8}$$

5. Considérant un sous espace vectorile G_{d_n} de $\mathbb{L}^2$ de dimension finie d_n. Le développement à l'ordre d_n de $f(x)$ dans G_{d_n} est donné par :

$$f_{d_n}(x) = \sum_{k=0}^{d_n} a_k e_k(x), k = 0, \ldots, x \in \mathbb{R}. \tag{1.9}$$

Pour estimer $f(x)$ dans $\mathbb{L}^2$ on se propose de constuire un estimateur sans biais de sa projection orthogonale $f_{d_n}(x)$ dans G_{d_n}. Par la méthode des moments, on peut estimer les coefficients $\{a_k, k \in \mathbb{N}\}$ par :

$$\hat{a}_k = \frac{1}{n} \sum_{i=1}^{n} e_k(X_i), k = 0, ... \tag{1.10}$$

Ainsi, $f(x)$ peut être estimée par :

$$\hat{f}_{d_n}(x) = \sum_{k=0}^{d_n} \hat{a}_k e_k(x). \tag{1.11}$$

"L'expression (1.11) explique pourquoi $\hat{f}_{d_n}(x)$ est souvent désigné comme l'estimateur de la densité de probabilité par la méthode des séries orthogonales".

1.5.1 Propriétés statistiques de l'estimateur

a. Les coefficients $(\hat{a}_k)_{k=0,\ldots,d_n}$ sont des estimateurs sans biais de $(a_k)_{k=0,\ldots,d_n}$.
En effet,

$$\begin{aligned} \mathbb{E}(\hat{a}_k) &= \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} e_k(X_i)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[e_k(X_i)] = \mathbb{E}[e_k(X)] && (1.12) \\ &= a_k. && (1.13) \end{aligned}$$

b. Le biais de $\hat{f}_{d_n}(x)$ est par définition :

$$\mathbb{B}iais(\hat{f}_{d_n}(x)) = \mathbb{E}(\hat{f}_{d_n}(x)) - h(x) = \sum_{k=0}^{d_n} a_k e_k(x) - f(x), \tag{1.14}$$

ce qui implique que $\hat{f}_{d_n}(x)$ est un estimateur biaisé de $f(x)$.

c. L'erreur qudratique moyenne intégrée de l'estimateur est donnée par le théorème suivant :

Théorème 1.5.1. *(Kronmal-Tarter)*
Si $\int f^2(x)dx < \infty$, alors :

$$MISE(\hat{f}_{d_n}(x)) = \int f^2(x)dx - \sum_{k=0}^{d_n} a_k^2 + \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{a}_k). \tag{1.15}$$

1.6 Choix pratique de la base

Le choix de la base dépend d'abord du support de la densité à estimer. Si le support de f est un intervalle compact, on pourra choisir les fonctions trigonométriques ou les fonctions de Legendre. Sur $\mathbb{R}_+$, on pourra utiliser les fonctions de Laguerre ou les fonctions d'Hermite. Quand on ne possède aucune information sur le support de f on peut utiliser les fonctions d'Hermite. Les fonctions d'Hermite donnent de bons résultats au voisinage de la loi normale réduite puisque le premier élément de la base $e_0(x) = \pi^{-\frac{1}{2}} \exp(-\frac{x^2}{2})$ est la densité d'une variable aléatoire de loi normale centrée réduite. Au voisinage d'une loi normale quelconque on peut considerer des fonctions d' Hermite modifiées données par

$$e_j^1(x) = e_j(\frac{x - \overline{X}}{S_n}), j \in \mathbb{N}, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n = [\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2]^{\frac{1}{2}}. \tag{1.16}$$

"On voit qu'il l n'y a pas de solution évidente qui se dégage. En effet, les systèmes orthonormaux sont très variés et il n'existe pas de théorèmes qui permettrait de conseiller un système particulier".

1.6.1 Exemples

Dans cette section, nous allons présenté quelques cas particuliers d'estimateurs basés sur des systèmes orthonormaux trigonométriques (Dirichlet et Fejer) et ceux associés aux fonctions d'Hermite, de Laguerre et de Legendre.

a. L'estimateur associé aux fonctions d'Hermite

Les fonctions d'Hermite sont données par les formules suivantes :

$$e_j(x) = (2^j j! \pi^{\frac{1}{2}})^{-\frac{1}{2}} Q_j(x) \exp(-\frac{x^2}{2}), x \in \mathbb{R}; j = 0, \ldots$$

où $Q_j(x)$ est le $j^{\text{ième}}$ polynôme d'Hermite défini par :

$$Q_j(x) = (-1)^j \exp(-x^2)\frac{d^j}{dx^j}\exp(x^2); x \in \mathbb{R}, \quad j = 0, \ldots$$

L'estimateur associé est donné par :

$$\hat{f}_{d_n}(x) = \frac{d_n+1}{2n}\sum_{i=1}^{n}[\frac{Q_{j+1}(X_i)Q_j(x) - Q_j(X_i)Q_{j+1}(x)}{X_i - x}]. \tag{1.17}$$

b. L'estimateur associé aux fonctions de Laguerre

Les fonctions de Laguerre sont données par les formules :

$$L_j(x) = [\frac{\Gamma(d_n+1)}{\Gamma(d_n+1+\alpha)}x^{-\alpha}\exp(x)]^{\frac{1}{2}}\frac{1}{i!}\frac{d^i}{dx^i}(x^{i+\alpha}\exp(-x)), i \geq 0, \alpha > 0.$$

L'estimateur de Laguerre associé est alors défini par :

$$\hat{f}_{d_n}(x) = \frac{\Gamma(d_n+1)}{n\Gamma(d_n+1+\alpha)}\sum_{i=1}^{n}[\frac{l_{j+1}(X_i)l_j(x) - l_j(X_i)l_{j+1}(x)}{X_i - x}].$$

c. L'estimateur associé aux fonctions de Legendre

Les fonctions de Legendre sont définies par :

$$p_i(x) = \sqrt{\frac{2i+1}{2}}\frac{1}{2^i i!}\frac{d^i}{dx^i}((x^2-1)^i), x \in [-1.1], i \geq 0.$$

L'estimateur associé est alors défini par :

$$\hat{f}_{d_n}(x) = \frac{d_n+1}{n\sqrt{2d_n+1}\sqrt{2d_n+3}}\sum_{i=1}^{n}\frac{p_{d_n}(X_i)p_{d_n+1}(x) - p_{d_n+1}(X_i)p_{d_n}(x)}{x - X_i}.$$

d. L'estimateur de Saadi-Adjabi

La base est donnée par :

$$e_k(x) = \frac{1}{\sqrt{2\pi}}(\cos(kx) + \sin(kx))1_{[-\pi,\pi]}(x), \ k = 0, \ldots. \tag{1.18}$$

L'estimateur de la densité associé à cette base trigonométrique est alors de la forme :

$$\hat{f}_{d_n}(x) = \frac{1}{4\pi n}\sum_{i=1}^{n}\left[\frac{\sin[\frac{(2d_n+1)(X_i-x)}{2}]}{\sin[\frac{X_i-x}{2}]} + \frac{\sin[\frac{(2d_n+1)(\frac{\pi}{2}-(X_i+x))}{2}]}{\sin[\frac{\frac{\pi}{2}-(X_i+x)}{2}]}\right]. \tag{1.19}$$

d. L'estimateur associé à la base de Dirichlet

Nous supposons que $I = [-\pi, \pi]$ un intervalle de $\mathbb{R}$ muni de la mesure de lebesgue et la base orthonormale est définie par :

$$e_0(x) = \frac{1}{\sqrt{2\pi}}, e_{2k}(x) = \frac{cos(kx)}{\sqrt{\pi}}, e_{2k+1}(x) = \frac{\sin(kx)}{\sqrt{\pi}}; k = 1, \ldots$$

pour d_n impair, l'estimateur de Dirichlet est donné par :

$$\hat{f}_{d_n,D}(X_i, x) \quad = \begin{cases} \frac{1}{2\pi n} \sum_{i=1}^{n} \frac{\sin[d_n \frac{(X_i - x)}{2}]}{\sin[\frac{X_i - x}{2}]} & \text{pour } X_i \neq x \\ \frac{d_n}{2\pi} \quad \text{sinon.} \end{cases}$$

e. L'estimateur de Fejer

Considérons les systèmes suivants dans $[a, b]$:

$$\{\cos k\pi[\frac{x-a}{b-a}], k = 0, 1, \ldots\}, \quad \{\sin k\pi[\frac{x-a}{b-a}], k = 0, 1, \ldots\}, \ \textit{et} \ \{\cos k\pi[\frac{x-a}{b-a}], \sin k\pi[\frac{x-a}{b-a}], k = 0, 1, \ldots\}. \tag{1.20}$$

Kronmal et Tarter partir d'un développement en série de Fourier. Ils proposent trois estimateurs

$$\hat{f}_{d_n}(x) = \frac{c_0}{2} + \sum_{k=1}^{d_n} \overline{c_k} \cos k\pi[\frac{x-a}{b-a}], \ \ \hat{f^1}_{d_n}(x) = \frac{c_0}{2} + \sum_{k=1}^{d_n} \overline{s_k} \sin k\pi[\frac{x-a}{b-a}], \quad \textit{et} \ \ \hat{f^2}_{d_n}(x) = \frac{1}{2}(\hat{f}_{d_n}(x) + \hat{f^1}_{d_n}(x)). \tag{1.21}$$

Dans ces formules $\overline{c_k}$ et $\overline{s_k}$ représentent les moments trigonométriques de l'échantillon, c'est-à- dire :

$$\overline{c_k} = \frac{c_0}{n} \sum_{i=1}^{n} \cos k\pi[\frac{X_i - a}{b-a}], \quad \overline{s_k} = \frac{c_0}{n} \sum_{i=1}^{n} \sin k\pi[\frac{X_i - a}{b-a}] \ \ \textit{et} \ \ c_0 = \frac{2}{b-a}. \tag{1.22}$$

L'estimateur considéré est alors défini par :

$$g_{d_n}(x) = \frac{c_0}{2n(d_n + 1)} \sum_{i=1}^{n} [\frac{\sin((d_n + 1)\frac{\pi(X_i - a)}{b-a})}{\sin(\frac{\pi(X_i - a)}{b-a})}]^2, \tag{1.23}$$

dont l'erreur quadratique moyenne intégrée est donnée par

$$MISE(g_{d_n}(x)) = c_0 \int_a^b g^2(x)dx - \frac{c_0^2}{2} + \sum_{k=1}^{d_n} (1 - \frac{k}{d_n + 1})^2 [\mathbb{V}ar(\overline{c_k}) + \mathbb{V}ar(\overline{s_k}) - \frac{d_n + 1 + k}{d_n + 1 - k}(\mathbb{E}^2(\hat{c}_k) + \mathbb{E}^2(\hat{s}_k))] \tag{1.24}$$

1.7 Choix pratique du paramètre de lissage

La base étant supposée fixée, il reste à choisir le paramètre de lissage d_n. pour cela, on cherche à minimiser l'erreur quadratique moyenne intégée $MISE(\hat{f}_{d_n}(x))$. Il existe plusieurs méthodes pour le choix du parmètre de lissage : La méthode de Kronmal-Tarter et la méthode de Bosq

1.7.1 Méthode de Kronmal-Tarter

L'emploi de (1.11) pour estimer $f(x)$ n'est possible qu'après avoir déterminer le nombre optimum de terme d_n de la somme. Il est naturel de choisir d_n de sorte que l'erreur quadratique moyenne intégrée $MISE(\hat{f}_{d_n}(x))$ soit minimum. La règle adoptée pour déterminer la valeur optimum d_n repose sur l'algorithme suivant : A partir de $d_n = 1$ on augmente la valeur de d_n d'une unité jusqu'à ce que $MISE(\hat{f}_{d_n}(x))$ augmente on donne alors à d_n la valeur qui précède juste l'augmentation de $MISE(\hat{f}_{d_n}(x))$. On ajoutera donc à la somme (1.11) le $d_n^{\text{ième}}$ terme si et seulement si

$$\Delta_{d_n} = MISE(\hat{h}_{d_n}(x)) - MISE(\hat{h}_{d_n-1}(x)) \leq 0. \tag{1.25}$$

En tenant compte de (1.5.1), Δ_{d_n} se met sous la forme :

$$\begin{aligned}
\Delta_{d_n} &= MISE(\hat{f}_{d_n}(x)) - MISE(\hat{f}_{d_n-1}(x)) \\
&= \int f^2(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{a}_k) - a_k^2)] - \int h^2(x)dx + \sum_{k=0}^{d_n-1}[\mathbb{V}ar(\hat{a_k}) - a_k^2] \\
&= \mathbb{V}ar(\hat{a}_{d_n}) - a_{d_n}^2 \\
&= \mathbb{V}ar(\frac{1}{n}\sum_{i=1}^{n} e_{d_n}(X_i)) - a_{d_n}^2 \\
&= \frac{1}{n}\mathbb{V}ar(e_{d_n}(X)) - a_{d_n}^2 \\
&= \frac{1}{n}\int e_{d_n}(x)h(x)dx - \frac{1}{n}a_{d_n}^2 - a_{d_n}^2 \\
&= \frac{1}{n}[\int e_{d_n}(x)h(x)dx - (n+1)a_{d_n}^2] \\
&= \frac{n+1}{n}\mathbb{V}ar(e_{d_n}(X)) - \mathbb{E}(e_{d_n}(X))^2.
\end{aligned}$$

Posons alors

$$\theta_i = e_{d_n}(X_i), i = 1, ..., n, \quad \overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_i.$$

On peut alors définir un estimateur symétrique sans biais de Δ_{d_n} donné par :

$$\hat{\Delta}_{d_n} = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2 - \sum_{i=1}^{n}\theta_i^2]. \tag{1.26}$$

On se fixe maintenant un entier positif D , l'optimum d_n^* est alors de la forme :

$$d_n^* = \begin{cases} \inf\{d_n, 1 \leq d_n \leq D\} & \hat{\Delta}_d > 0 \\ D & sinon. \end{cases}$$

1.7.2 Méthode de Bosq

Bosq a proposé un nouveau estimateur de paramètre de lissage donné par :

$$\hat{d}_n = max\{j : 0 \leq j \leq d_n, |\hat{a}_j| \geq \gamma_n\}, \tag{1.27}$$

avec

$$\gamma_n = c\sqrt{\frac{logn}{n}}, c > 0. \tag{1.28}$$

Théorème 1.7.1.

1. *Si* $\frac{d_n}{n} \longrightarrow 0$ *on a*
$$MISE(\hat{h}_{\hat{d}_n}(x)) \longrightarrow 0. \tag{1.29}$$
2. *Si* $\sum_n \mid a_j \mid < \infty$ *et* $\sum_n d_n \exp[-\frac{n}{d_n^2}a] < \infty, a > 0,$
alors
$$\sup_{x \in E} \mid \hat{f}_{\hat{d}_n}(x) - f(x) \mid \underset{p.s}{\longrightarrow} 0. \tag{1.30}$$

1.8 Estimateur à noyau

Définition 1.9. Soit $(x_1, ..., x_n)$ un échantillon de loi $f(x)$ sur $\mathbb{R}$, de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associée à $(x_1, ..., x_n)$, la fonction aléatoire $F_n : \mathbb{R} \to [0, 1]$ définie pour tout $x \in \mathbb{R}$ par :

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

La densité est la dérivée de la fonction de répartition, ce qui permet d'écrire pour tout x :

$$f(x) = \lim_{h \to 0} \frac{F(x+h) + F(x-h)}{2h}.$$

Une des premières idées intuitives est de considérer pour $h > 0$:

$$f_h(x) = \lim_{h \to 0} \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{n}\sum_{i=1}^{n} 1(x_i)_{]x-h,x+h[} = \frac{1}{n}\sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right),$$

avec

$$w(u) = \begin{cases} \frac{1}{2} & si \quad -1 \leq u \leq 1, \\ 0, & sinon \end{cases}$$

Cet estimateur appelé estimateur de **Rosenblatt**, est le premier exemple d'estimateur à noyau construit à l'aide du noyau uniforme $K(u) = \frac{1}{2}1_{-1\leq u\leq 1}$. Parzen a étudié une classe générale d'estimateurs. la méthode de Parzen consiste à utiliser la formule ci-dessus pour tout $x \in \mathbb{R}$ et pas seulement pour la classe $[-1.1]$. Cette généralisation est certe utile, car elle conduit vers un estimateur qui est constant par morceaux comme les histogrammes, mais a l'avantage d'avoir des plateaux de longueurs variables. On remarque aisément que la discontinuité de l'estimateur défini ci-dessus est une conséquence de la discontinuité de la fonction indicatrice. Par conséquent, en remplaçant $1(z) \leq \frac{1}{2}$ par une fonction K quelconque, on obtient l'estimateur suivant :

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), \tag{1.31}$$

qui est continu et même l-fois continûment différentiable du moment où la fonction K l'est.
h est un paramètre qui est fonction de n, appelé paramètre de lissage.
K est une fonction définie sur $\mathbb{R}$ appelée noyau.

Lemme 1.8.1. Si le noyau K est une fonction positive et $\int_{-\infty}^{+\infty} K(\mu)d\mu = 1$, alors $f_h(x)$ est une densité de probabilité.

1.8.1 Noyaux usuels

Les noyaux les plus couramment utilisés en pratique sont :

– Le noyau uniforme(rectangulaire)

$$K(u) = \frac{1}{2}, \quad \mid u \mid \leq 1,$$

– Le noyau Triangulaire

$$K(u) = (1 - \mid u \mid), \quad \mid u \mid \leq 1,$$

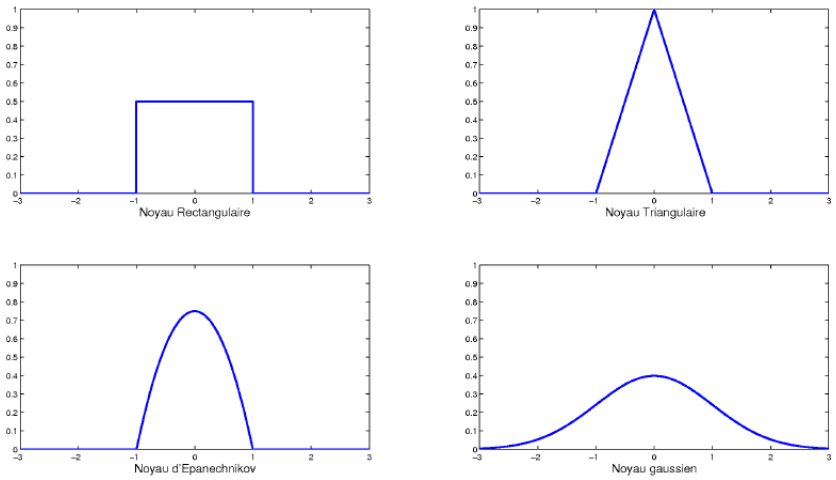
– Le noyau Gaussien

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

– Le noyau d'Epanechnikov

$$K(u) = \frac{3}{4}(1 - \mu^2), \quad \mid u \mid \leq 1$$

Les courbes de ces noyaux sont présentées ci-dessous :



1.8.2 Etude du biais et de la variance

Lorsqu'on définit un estimateur à noyau, on a non seulement le choix de la fenêtre $h > 0$ mais aussi celui du noyau K. Il y a un certain nombre de conditions qui sont considérées comme usuelles pour les noyaux et qui permettent d'analyser le risque de l'estimateur à noyau qui en résulte. On suppose que le noyau K vérifié les 4 conditions suivantes :

1. $\int_{-\infty}^{+\infty} K(\mu)d\mu = 1$, K est une densité de probabilité,
2. $\int_{-\infty}^{+\infty} \mu K(\mu)d\mu = 0$, K est symétrique,
3. $\int_{-\infty}^{+\infty} \mu^2 K(\mu)d\mu = \sigma_K^2 < +\infty$,
4. $\int_{-\infty}^{+\infty} K^2(\mu)d\mu < +\infty$.

Proposition 1.8.2. *Si les tois premières conditions sont remplies, alors*

$$\mathbb{B}iais\left(f_h(x)\right) = \mathbb{E}\left(f_h(x)\right) - f(x) = \frac{h^2}{2!} f''(x)\mu_2(K) + o(h^2), \quad \mu_2(K) = \int_{-\infty}^{\infty} y^2 K(y)dy. \tag{1.32}$$

Si, de plus la condition 4 est satisfaite, alors

$$\mathbb{V}\left(f_h(x)\right) = \frac{f(x)}{nh} \int K^2(u)du + O\left(\frac{1}{n}\right).$$

preuve. L'espérance mathématique de $f_h(x)$ est :

$$\begin{aligned}\mathbb{E}\left(f_h(x)\right) &= \frac{1}{nh}\mathbb{E}\left(\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\right) \\ &= \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{u-x}{h}\right) f(u)du\end{aligned}$$

En posant $y = \frac{u-x}{h} \Rightarrow dy = \frac{du}{h}$

$$\mathbb{E}\left(f_h(x)\right) = \int_{-\infty}^{+\infty} K(y) f(x+hy)dy.$$

En effectuant un développement de Taylor à l'ordre 2 au point $h = 0$ de la fonction $f(x+hy)$, il vient

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

Il en résulte que

$$\mathbb{B}iais\left(f_h(x)\right) = \mathbb{E}\left(f_h(x)\right) - f(x) = \frac{h^2}{2!}f''(x)\mu_2(k) + o(h^2), \;\; \mu_2(k) = \int_{-\infty}^{\infty} y^2 K(y)dy, \quad (1.33)$$

Pour prouver la seconde assertion, on utilise le faite que les variables aléatoires sont i.i.d. et que la variance de la somme de variables indépendantes coïncide avec la somme des variances :

$$\begin{aligned}\mathbb{V}\left(f_h(x)\right) &= \mathbb{V}\left(\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x_i - x}{h}\right)\right) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{V}\left(K\left(\frac{x_i - x}{h}\right)\right) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(K^2\left(\frac{x_i - x}{h}\right)\right)\right] - \frac{1}{n^2h^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(K\left(\frac{x - x_i}{h}\right)\right)\right]^2 \\ &= \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(y)dy - \frac{f'(x)}{n}\int_{-\infty}^{\infty} yK^2(y)dy - \frac{1}{n}\left(f(x) + \mathbb{B}iais\left(f_n(x)\right)\right)^2,\end{aligned}$$

ce qui nous donne :

$$\mathbb{V}ar\left(f_h(x)\right) = \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(y)dy + O\left(\frac{1}{nh}\right). \quad (1.34)$$

Proposition 1.8.3. *Si les tois premières conditions sont remplies et f est une densité bornée dont la dérivée seconde est bornée, alors*

$$| \mathbb{B}iais\,(f_h(x)) | \leq C_1 h^2, \tag{1.35}$$

où $C_1 = \frac{1}{2}\sup_{x\in\mathbb{R}} | f''(x) | \int \mu^2 | K(\mu) | d\mu$.
Si, de plus la condition 4 est satisfaite, alors

$$\mathbb{V}\,(f_h(x)) \leq \frac{C_2}{nh},$$

avec, $C_2 = \sup_{x\in\mathbb{R}} | f(x) | \int K^2(\mu) d\mu$.

preuve. Supposons f de de classe C^2 et telle que f'' soit bornée.

$$\begin{aligned}\mathbb{B}iais\,(f_h(x)) &= \mathbb{E}\,(f_h(x)) - f(x) = \frac{1}{nh}\mathbb{E}\left(\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\right) - f(x)\\ &= \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{u-x}{h}\right) f(u)du - f(x)\end{aligned}$$

En posant $y = \frac{u-x}{h} \Rightarrow dy = \frac{du}{h}$

$$\mathbb{B}iais\,(f_h(x)) = \int_{-\infty}^{+\infty} K(y)\,[f(x+hy) - f(x)]\,dy.$$

Puisque f est supposée de classe C^2, on peut appliquer la formule de Tylor à l'ordre 2, ce qui nous donne :

$$\mathbb{B}iais\,(f_h(x)) = \frac{h^2}{2!} f''(x)\mu_2(k) + o(h^2). \tag{1.36}$$

Ainsi, en ayant supposé de plus que le noyau k est symétrique et f'' est bornée,

$$| \mathbb{B}iais\,(f_h(x)) | \leq \frac{h^2}{2}\sup_{x\in\mathbb{R}} | f''(x) | \int_{-\infty}^{\infty} y^2 | K(y) | dy. \tag{1.37}$$

Pour prouver la seconde assertion, on utilise le faite que les variables aléatoires sont i.i.d. et que la variance de la somme de variables indépendantes coïncide avec la somme des variances :

$$\begin{aligned}\mathbb{V}\,(f_h(x)) &= \mathbb{V}\left(\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x_i - x}{h}\right)\right)\\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{V}\left(K\left(\frac{x_i - x}{h}\right)\right)\\ &\leq \frac{1}{nh^2}\left[\mathbb{E}\left(K^2\left(\frac{X-x}{h}\right)\right)\right] = \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{u-x}{h}\right) f(u)du.\end{aligned}$$

On en déduit la majoration :

$$\mathbb{V}\left(f_h(x)\right) \leq \frac{\parallel f \parallel_\infty}{nh} \int_{-\infty}^{\infty} K^2(y)dy.$$

Discussion du comportement du biais et de la variance :

* Le biais décroit si h diminue mais la variance augmente.
* Pour que la variance tende vers 0, il faut que $nh \to \infty$.
* La variance diminue si h augmente mais le biais augmente.

1.8.3 Risque quadratique ponctuel et risque quadratique intégré

Risque quadratique ponctuel :

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

$$MSE\left(f(x), f_h(x)\right) = \left[\mathbb{B}iais\left(f_h(x)\right)\right]^2 + \mathbb{V}\left(f_h(x)\right) \tag{1.38}$$

En remplaçant les expressions finales des deux termes, le biais et la variance dans l'équation (1.38) on obtient :

$$MSE\left(f(x), f_h(x)\right) = \frac{f(x)}{nh} \int_{-\infty}^{\infty} K^2(y)dy + \frac{1}{4}h^4(f''(x))^2 \left(\int_{-\infty}^{\infty} y^2 K(y)dy\right)^2 + O\left(\frac{1}{nh} + h^5\right). \tag{1.39}$$

Risque quadratique intégré :

$$\begin{aligned} MISE\left(f, f_h\right) &= \int_{-\infty}^{\infty} MSE\left(f(x), f_h(x)\right) dx \\ &= \int \mathbb{E}\left(f(x) - f_h(x)\right)^2 dx \end{aligned}$$

$$MISE(f, f_h) = \int \left[\left(\mathbb{B}iais\left(f_h(x)\right)^2 + \mathbb{V}\left(f_h(x)\right)\right] dx. \tag{1.40}$$

En remplaçant les expressions finales des deux termes, le biais et la variance dans l'équation (1.40) on obtient :

$$MISE(f(x), f_h(x)) = \frac{h^4}{4}\sigma_k^4 \int (f''(x))^2 dx + \frac{1}{nh} \int K^2(u)du + O\left(h^5 + \frac{1}{n}\right). \tag{1.41}$$

L'erreur quadratique moyenne intégrée asymptotique AMISE :

$$AMISE = MISE\left(f(x), f_h(x)\right) - O\left(h^5 + \frac{1}{n}\right),$$

$$AMISE = \frac{h^4}{4}\sigma_k^4 R(f'') + \frac{R(K)}{nh},$$

avec

$$R(s) = \int s^2(x)dx.$$

Théorème 1.8.4. *Si les 4 conditions sont remplies, alors Le paramètre de lissage h^* qui minimise l'erreur quadratique moyenne intégrée asymptotique est de la forme :*

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

La valeur du $AMISE$ optimale $AMISE^* = AMISE(h^*)$ est alors de forme

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

preuve. le paramètre de lissage h qui minimise l'erreur quadratique moyenne intégrée asymptotique est :

$$\begin{aligned} \frac{dAMIS}{dh} &= h^3\sigma_K^2 R\left(f''\right) - \frac{R(K)}{nh^2} = 0 \\ nh^5\sigma_K^4 - R(K) = 0 \;&\Rightarrow\; h^5 = \frac{R(K)}{n\sigma_K^4 R(f'')}. \end{aligned}$$

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

$$\begin{aligned} \frac{d^2AMIS}{dh^2} &= 3h^2\sigma_K^4 R\left(f''\right) + \frac{R(K)}{nh^3} > 0 \Rightarrow h^* \;\; minimise \;\; AMISE \\ nh^5\sigma_K^4 - R(K) = 0 \;&\Rightarrow\; h^5 = \frac{R(K)}{n\sigma_K^4 R(f'')}. \end{aligned}$$

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

$$AMISE^* = \frac{5}{4}\left[\sigma_K^4 R^4(K) R(f'')\right]^{1/5} n^{-4/5}. \tag{1.42}$$

Théorème 1.8.5. *(**Majoration du risque quadratique ponctuel**). Si les 4 conditions sont remplies et f est une densité bornée dont la dérivée seconde est bornée, alors le risque quadratique admet la majoration suivante :*

$$MSE\left(f(x), f_h(x)\right) \leq \frac{C^2h^4}{4} + \frac{C_2}{nh}.$$

Discussion : Sur-lissage et sous-lissage

Lorsque la fenêtre h est très petit, le biais de l'estimateur à noyau est très petit face à sa variance et c'est cette dernière qui détermine la vitesse de convergence du risque quadratique. Dans ce type de situation, l'estimateur est très volatile et on parle de **sous-lissage** (under-smoothing, en anglais). En revanche, lorsque h grandit, la variance devient petite et c'est le biais qui devient dominant. L'estimateur est alors très peu variable et est de moins à moins influencé par les données. On parle alors d'un effet de sur-lissage (over-smoothing en anglais). En pratique, il est primordial de trouver la bonne dose de lissage qui permet d'éviter le sous-lissage et le **sur-lissage.**

1.8.4 Noyau d'ordre l

Définition 1.10. On dit que $K : \mathbb{R} \to \mathbb{R}$ est un noyau d'ordre l si les fonctions

$$\begin{cases} \mathbb{R} \to \mathbb{R} \\ u \to u^j K(u) \end{cases}$$

sont intégrables pour $j = 0, 1, ..., l$ et vérfient $\int_{-\infty}^{+\infty} K(u)du = 1$, ainsi que

$$\int_{-\infty}^{+\infty} u^j K(u)du = 0, j =, 1, ..., l.$$

Définition 1.11. Soient T un intervalle de $\mathbb{R}$, et deux réels $\beta, L > 0$. La classe de Hölder $\Sigma(\beta, L)$ sur T est définie comme l'ensemble des fonctions $g : T \to \mathbb{R}$ telles que g est $l = \lfloor \beta \rfloor$ fois dérivale et telle que $\mid g^l(x) - g^l(y) \mid \leq \mid x - y \mid^{\beta - l}, \forall x, y \in T$.

Théorème 1.8.6. *Soit f une densité et K un noyau tels que :f est bornée et dans une classe de Hölder $\Sigma(\beta, L)$ sur $\mathbb{R}$, K est un noyau d'ordre $l = \lfloor \beta \rfloor$ de carré intégrable et tel que $\int_{-\infty}^{+\infty} \mid \mu \mid^{\beta} \mid K(\mu) \mid d\mu < \infty$, alors*

$$MSE\left(\hat{f}_h(x)\right) \leq \frac{c_1}{nh} + c_2 h^{2\beta}.$$

Pour $h_{opt} = bn^{-\frac{1}{2\beta+1}}$, avec $b > 0$, il existe $C > 0$ tel que

$$MSE\left(\hat{f}_h(x)\right) \leq Cn^{-\frac{2\beta}{2\beta+1}}.$$

Définition 1.12. Soient deux réels $\beta, L > 0$ et soit $p = \lfloor \beta \rfloor$. On applle espace de **Niolski** l'ensemble, noté $\mathcal{H}(\beta, L)$, des fonctions $f : \mathbb{R} \to \mathbb{R}$ tel que f est p fois dérivable et tel que

$$\left(\int_{-\infty}^{\infty} (f^{(p)}(x+t) - f^{(p)}(x))^2 dx\right)^{\frac{1}{2}} \leq L \mid t \mid^{\beta - p}, \forall t \in \mathbb{R}.$$

Théorème 1.8.7. *Soit $f \in \mathcal{H}(\beta, L)$ et K est un noyau d'ordre $l = \lfloor \beta \rfloor$, tel que $\int_{-\infty}^{+\infty} \mid \mu \mid^{\beta} \mid K(\mu) \mid d\mu < \infty$, alors*

$$MISE\left(\hat{f}_h(x)\right) \leq \frac{\int K^2}{nh} + c_2 h^{2\beta},$$

avec

$$c_2 = \frac{1}{p!} \int_{-\infty}^{+\infty} \mid \mu \mid^{\beta} \mid K(\mu) \mid d\mu.$$

Pour $h_{opt} = bn^{-\frac{1}{2\beta+1}}$, avec $b > 0$, il existe $C > 0$ tel que

$$MISE\left(\hat{f}_h(x)\right) \leq Cn^{-\frac{2\beta}{2\beta+1}}.$$

Chapitre 2

Les méthodes de sélection du paramètre de lissage

2.1 Introduction

Le choix judicieux du paramètre de lissage permet une bonne utilisation de cette méthode dans la pratique. Dans ce chapitre, nous présenterons les méthodes suivantes :

- la méthode de "plug-in" (Rul of Thumb) ;
- la méthode de "plug-in" itéré ;
- la méthode de la validation croisée.

Définition 2.1. (Choix théorique optimal de h).
Le paramètre de lissage h est un réel positif dont le choix est dominant par rapport au choix du noyau K. Le paramètre de lissage h_{opt} qui minimise l'erreur quadratique moyenne intégrée asymptotique AMIS est de la forme :

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 R\left(f''\right)}\right]^{1/5} n^{-1/5} = \psi(K)\varphi(f'')n^{-1/5}, \tag{2.1}$$

avec

$$\psi(K) = \left[\frac{R(K)}{\sigma_K^4}\right]^{\frac{1}{5}} \quad et \quad \varphi(f'') = \left[\frac{1}{R\left(f''\right)}\right]^{1/5}, \varphi(f'') \neq 0.$$

Définition 2.2. La valeur du AMISE optimale $AMISE_{opt} = AMISE(h_{opt})$ est donnée par :

$$AMISE_{opt} = \frac{5}{4}\left[\sigma_K^4 R^4(K) R(f'')\right]^{1/5} n^{-4/5}. \tag{2.2}$$

Remarque 2.1.1. La convergence de l'estimateur à noyau est plus rapide que pour l'histogramme, étant d'ordre $n^{-4/5}$ au lieu de $n^{-2/3}$, mais ce résutat dépend de la quantité inconnue f'', c'est pourquoi on considère des méthodes pratiques pour le choix du paramètre de lisage.

2.2 Méthodes plug-in

2.2.1 Rule of Thumb

L'idée de base de la méthode Rul of Thumb pour le choix du paramètre h, est d'estimer dans l'expression de h_{opt} (2.1), la quantite inconnue $R(f'')$. Supposant que f appartient a une famille de distributions normales $\mathcal{N}(\mu;\sigma^2)$, de moyenne μ et variance σ^2 inconnues. Sous cette hypothese :

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

la densité de probabilité normale centré réduite et

$$f'(x)=\frac{1}{\sigma^2}\phi'\left(\frac{x-\mu}{\sigma}\right) \quad \Rightarrow \quad f''(x)=\frac{1}{\sigma^3}\phi''\left(\frac{x-\mu}{\sigma}\right)$$

la quantité inconnue $\int_{\mathbb{R}}[f''(x)]^2\,dx$, s'écrit alors

$$\int_{\mathbb{R}}[f''(x)]^2\,dx=\frac{1}{\sigma^6}\int_{\mathbb{R}}\left\{\phi''\left(\frac{x-\mu}{\sigma}\right)\right\}^2 dx$$

faisons le changement de variable $v=\frac{x-\mu}{\sigma}$, d'où $dv=\frac{1}{\sigma}\,dx$

$$\int_{\mathbb{R}}[f''(x)]^2\,dx=\frac{1}{\sigma^5}\int_{\mathbb{R}}\left\{\phi''(v)\right\}^2 dv$$

mais

$$\phi(v)=\frac{1}{\sqrt{2\pi}}e^{\frac{-v^2}{2}} \quad \Rightarrow \quad \phi'(v)=\frac{-v}{\sqrt{2\pi}}e^{\frac{-v^2}{2}} \quad \Rightarrow \quad \phi''(v)=\frac{1}{\sqrt{2\pi}}(v^2-1)e^{\frac{-v^2}{2}}$$

ce qui implique

$$\begin{aligned} h_n^{opt} &= 1 \left\{(4\pi)^{-1/2}\right\}^{1/5} \left\{\frac{3}{8}\ \pi^{-1/2}\widehat{\sigma}^{-5}\right\}^{-1/5} n^{-1/5} \\ &= 4^{-1/10}\ \pi{-}1/10 \left(\frac{3}{8}\right)^{-1/5} \pi^{1/10}\ \widehat{\sigma}\ n^{-1/5} \\ &= 2^{-2/5}\ 3^{-1/5}\ 2^{4/5}\ \widehat{\sigma}\ n^{-1/5} \\ &= 2^{2/5}\ 3^{-1/5}\ \widehat{\sigma}\ n^{-1/5} \\ &= 4^{1/5}\ 3^{-1/5}\ \widehat{\sigma}\ n^{-1/5} \\ &= \left(\frac{4}{3}\right)^{1/5}\ \widehat{\sigma}\ n^{-1/5} \\ &= 1.06\widehat{\sigma}\, n^{-1/5} \end{aligned}$$

2.2.2 Plug-in itéré

Définition 2.3. En adoptant le critère de l'erreur quadratique intégrée moyenne ($MISE$), Scott, Tapia et Thompson choisissent d'estimer le paramètre $R(f'')$ à l'aide de l'estimateur naturel $\hat{R}_h(f'')$ défini comme suit :

$$\hat{R}_h(f'') = R(f_h''), \tag{2.3}$$

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

$$f_h''(x) = \frac{1}{nh^3} \sum_{i=1}^{n} K'' \left(\frac{x - x_i}{h}\right). \tag{2.4}$$

En choisissant par exemple le classique noyau gaussien

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

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

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

Remarque 2.2.1. 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 $\hat{R}_h(f'')$ devrait être robuste par rapport à une erreur de spécification sur f, Scott, Tapia et Thompson proposent finalement d'injecter l'estimateur $\hat{R}_h(f'')$ dans l'expression (2.1). Cette approche amène à considérer l'équation numérique suivante en h :

$$h = \psi(K)\varphi(f_h)n^{-\frac{1}{5}}, \quad où \quad \varphi(f_h) = \left[\frac{1}{R(f_h'')}\right]^{\frac{1}{5}}. \tag{2.5}$$

Algorithme (S.T.T)

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

$$h_{i+1} = \phi(h_i), \tag{2.6}$$

où,

$$\phi(f_{h_i}) = \psi(K)\varphi(f_{h_i})n^{-\frac{1}{5}} = \psi(K)\left[\frac{1}{R(f_{h_i})}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

Les étapes de ce processus itératif, appelé algorithme (S.T.T) sont :
- h_0 : solution initiale, prenant par exemple l'étendue de l'échantillon.
- $h_{i+1} = \psi(K)\varphi(f_h)n^{-\frac{1}{5}}$.
- Le critère d'arrêt est donné par la formule suivante :

$$\left|\frac{h_{i+1} - h_i}{h_i}\right| \epsilon \tag{2.7}$$

où ϵ est une précision petite donnée.

2.2.3 Méthodes cross validation (validation croisée)

L'idée de base des méthodes validation croisée consiste à trouver une fonction de score $CV(h)$ ayant la même structure que le $MISE(h)$ et dont le calcul soit plus simple.

Validation croisée non biaisée

Cette méthode a été proposée par Rudemo et Bowman en 1984. Le critère consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

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

Puisque $\int_{\mathbb{R}} f^2(x)dx$ 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 :

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

On veut premièrement trouver un estimateur de $\int_{\mathbb{R}} f_h(x)f(x)dx$. Remarquons que :

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

L'estimateur empirique de $\int_{\mathbb{R}} f_h(x)f(x)dx$, est alors $\frac{1}{n}\sum_{i=1}^n f_{h,i}(x_i)$. Le critère à optimiser est alors :

$$UCV(h) = \int_{\mathbb{R}} f_h^2(x)dx - \frac{2}{n}\sum_{i=1}^n f_{h,i}(x_i). \tag{2.8}$$

Où $f_{h,i}(x) = \frac{1}{(n-1)h}\sum_{1\leq i\leq n, i\neq j} k(\frac{x-x_j}{h})$ est l'estimateur de la densité construit à partir de l'ensemble de points sauf le point x_i.
En utilisant l'équation (2.8), le critére $UCV(h)$ devient :

$$UCV(h) = \frac{R(K)}{nh} + \sum_{i=1}^n \sum_{i\neq j, j=1} \left[\int \frac{1}{n^2h^2} K\left(\frac{x-x_i}{h}\right) K\left(\frac{x-x_j}{h}\right) dx - \frac{2}{n(n-1)h} K\left(\frac{x_i-x_j}{h}\right)\right]. \tag{2.9}$$

avec

$$R(K) = \int K^2(u)du.$$

Nous noterons h_{ucv} l'estimateur de h qui minimisr $UCV(h)$.

Algorithme de la méthode validation croisée non biaisée

Pour trouver le paramètre de lissage optimal, noté h_{ucv} par validation croisée non biaisée, on minimise numériquement $UCV(h)$. En utilisant le noyau gaussien. Les étapes de l'algorithme sont :

- **Début** (Génération d'un échantillon $(x_i)_{1\leq i\leq n}$)
- Somme1 $= 0$, Somme2 $= 0$;
- **Pour** $i = 1$ à n faire
- **Pour** $j = 1$ à n faire
- **Si** $i \neq j$

  $Som1 = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\frac{x_i - x_j}{h})^2\right)$;
  $Somme1 = Somme1 + Som1$;
  $Som2 = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{1}{4}(\frac{x_i - x_j}{h})^2\right)$;
  $Somme2 = Somme2 + Som2$;

- **Fin pour**

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

Validation croisée biaisée

La méthode de validation croisée biaisée, a été introduit par Scott et Terrell pour remédier aux problèmes de validation croisée non biaisée. Il s'agit d'introduire un biais dans le UCV afin de réduire sa variance.

Lemme 2.2.1. (Scott et Terrell)
Supposant que le noyau k satisfait aux conditions suivantes :

$$\int K''(u)du = 0, \quad u_1(K'') = \int uK''(u)du = 0, \quad u_2(K'') = \int u^2K''(u)du = 2.$$

On obtient le développement asymptotique :

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

avec :

$$R(f_h''(x)) = \int_R (f_h''(x))^2 dx, \qquad R(K'') = \int_R (K''(u))^2 du.$$

L'estimateur du $AMISE$:

$$BCV(h) = \frac{h^4}{4}\sigma_k^4 \left[R(f_h'') - \frac{R(K'')}{nh^5}\right] + \frac{R(K)}{nh}. \tag{2.10}$$

Algorithme de la méthode validation croisée biaisée

Il est facile aussi de donner un algorithme afin de calculer le paramètre de lissage optimal noté h_{bcv} qui minimise $BCV(h)$. En utilisant le noyau gaussin, les principales étapes de l'algorithme sont :

- **Début** (Génération d'un échantillon $(x_i)_{1\leq i\leq n}$)
- $BCV(h) = 0$;
- **Pour** $i = 1$ à n faire
- **Pour** $j = 1$ à n faire
- **Si** $i \neq j$ $x = \left(\frac{x_i - x_j}{h}\right)$;
  $BCV(h) = BCV(h) + \exp\left(-\frac{x^2}{4}\right)\left(3 - 3x^2 + \frac{1}{4}x^4\right)$;
- **Fin pour**
  $UCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{32n^2h}BCV(h)$;
  $h_{UCV} = \min_h BCV(h)$.

Validation croisée de la vraisemblance

Définition 2.4. Pour un estimateur à noyau f_h , la sélection par validation croisée de la vraisemblance consiste à maximiser par rapport à h la vraisemblance pour l'échantillon $(x_1)_{\leq i\leq n}$ définie par :

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

où ;

$$f_{h,i}(x_i) = \frac{1}{n(n-1)} \sum_{1\leq j\leq n, i\neq j} K\left(\frac{x - x_i}{h}\right),$$

est l'estimateur à noyau basé sur les $(n-1)$ observations différentes de x_i. La vraisemblance est alors :

$$LCV(h) = \prod_{i=1}^{n} \frac{1}{n(n-1)} \sum_{1\leq j\leq n, i\neq j} K\left(\frac{x - x_i}{h}\right).$$

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

Algorithme de validation croisée de la vraisemblance LCV

- **Début** (Génération d'un échantillon $(x_1)_{1\leq i\leq n}$)
- $LCV(h) = 1$;
- **Pour** $i = 1$ Som =0 ;
- **Pour** $j = 1$

. **Si** $i \neq j$
$Som = Som + K\left(\frac{x_i - x_j}{h}\right)$ (En général nous utilisons un noyau K gaussien) ;
$LCV(h) = \frac{1}{[(n-1)h]^n} LCV(h)Som$;
$h_{lcv} = \max_h LCV(h)$.

Avantages et Inconvénients de diverses méthodes de sélection du paramètre de lissage

- Pour les méthodes plug-in, plusieurs points importants peuvent etre imperativement soulignés. D'abord, cette technique est très satisfaisante théoriquement puisque l'expression de minimisant le MISE(h) qui est d'ordre O(n-1/5) est de la forme :

$$h_{opt} = \left[\frac{R(K)}{\sigma_K^4 R\left(f''\right)}\right]^{1/5} n^{-1/5} = \psi(K)\varphi(f'')n^{-1/5}.$$

- Les difficultés importantes se posent en pratique : pour calculer h_{opt} il faut, en effet, estimer $\varphi(f'')$ et donc les dérivées de la fonction de distribution f qui s'avère techniquement délicat.
- Le principal inconvénient des méthodes validation croisée est que 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.
- La méthode de Validation croisée de la vraisemblance 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 quand elle est appliquée a des observations dont la distribution présente des queues.

Chapitre 3

Régression non paramétrique

3.1 Introduction

Un des modèles le plus fréquemment rencontré en statistique paramétrique ou nonparamétrique est le modèle de régression. On dispose d'un échantillon, composé de n couples indépendants de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$, et on dénote par (X, Y) un élément générique de cet échantillon. Dans le modèle de régression non paramétrique, on suppose typiquement l'existence d'une fonction $R(.)$ qui exprime la valeur moyenne de la variable réponse Y en fonction de la variable d'entrée X :

$$Y_i = R(X_i) + \epsilon_i, \quad pour \quad 1 \leq i \leq \ n, \quad avec \quad \epsilon_i = \epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma^2) \tag{3.1}$$

Définition 3.1. Soit (X, Y) un couple de variables aléatoires réelles admettant une densité jointe sur $\mathbb{R}^2$ notée $f_{X,Y}$,et une densité marginale f_X. La variable Y est supposée intégrable,ie $\mathbb{E} \mid y \mid < \infty$. Nous pouvons alors définir proprement la fonction de régression ou espérance conditionnelle de Y sachant $X = x$, par

$$R(x) = \mathbb{E}[Y|X = x] = \frac{\int_{\mathbb{R}} y f_{X,Y}(x, y)dy}{\int_{\mathbb{R}} f_{X,Y}(x, y)dy} = \frac{r(x)}{f_X(x)}, \tag{3.2}$$

lorsque la densité $f_X(x)$ est différente de zéro. Le problème de l'estimation de $R(.)$ est du type non-paramétrique, i.e. la fonction de régression appartient à un ensemble nonparamétrique (infini-dimensionnel).

Définition 3.2. Un estimateur $\hat{R}(x)$ de $R(x)$ est dit estimateur linéaire de la régression non paramétrique si

$$\hat{R}(x) = \sum_{i=1}^{n} y_i W_{ni}(x), \tag{3.3}$$

où la fonction de poids $W_{ni}(.)$ ne dépend pas des observations y_i.

3.2 L'estimateur de Nadaraya-Watson

Supposons que l'on dispose d'un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$, de variables aléatoires à valeurs réelles, de même loi que le couple (X, Y). On se propose de construire un estimateur $\hat{R}(x)$ de la fonction de régression à partir des couples d'observations $(x_1, y_1), \ldots, (x_n, y_n)$. Le premier estimateur rencontré dans la littérature est l'estimateur à noyau de Nadaraya-Watson, noté estimateur [NW].

Définition 3.3. L'estimateur [NW] se présente sous la forme d'une moyenne locale pondérée des valeurs Y_i et est défini par,

$$\hat{R}_n^{NW}(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \times 1\left\{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \neq 0\right\}. \tag{3.4}$$

De manière similaire, nous pouvons définir l'estimateur [NW] par,

$$\hat{R}_n^{NW}(x) = \begin{cases} \frac{\sum_{i=1}^n y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}, & lorsque \quad \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \neq 0 \\ \\ \frac{1}{n}\sum_{i=1}^n y_i & sinon. \end{cases} \tag{3.5}$$

La fonction K $K : \mathbb{R} \to \mathbb{R}$ sera supposée mesurable et satisfaisant certaines hypothèses basiques parmi celles énoncées ci-dessous :

- K est bornée, ie. $\sup_{\mu\in\mathbb{R}} \mid K(\mu) \mid < \infty$;
- Le noyau K est symétrique ;
- $\int K(\mu)d(\mu) = 1$.

Remarque 3.2.1. Le noyau K détermine la forme du voisinage autour du point x et la fenêtre h contrôle la taille de ce voisinage, c'est à dire le nombre d'observations prises pour effectuer la moyenne locale.

Définition 3.4. L'estimateur à noyau de $r(x)$ est

$$\hat{r}_n(x) = \sum_{i=1}^n y_i K\left(\frac{x-x_i}{h}\right) \tag{3.6}$$

L'estimateur [NW] peut s'écrire

$$\frac{\hat{r}_n(x)}{f_{hX}(x)}, \tag{3.7}$$

avec $f_h(x)$ est l'estimateur à noyau de la densité $f(x)$ de la variable X.

Propriété 3.2.1. *L'estimateur [NW] (3.4) est bien linéaire avec comme fonction de poids* $W_{n_i}^{NW}(.)$ *définie par,*

$$W_i^{NW}(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} \times 1\left\{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \neq 0\right\}. \tag{3.8}$$

Définition 3.5. On appelle estimateur à noyau de la densité conjointe du couple (X, Y)

$$\hat{f}_{X,Y}(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) K\left(\frac{y-y_i}{h}\right). \tag{3.9}$$

Proposition 3.2.2. *Si le noyau* K *est symétrique, nous obtenons les égalités suivantes*

$$\hat{R}_n(x) = \frac{\int_{\mathbb{R}} y\hat{f}_{X,Y}(x,y)dy}{\int_{\mathbb{R}} \hat{f}_{X,Y}(x,y)dy} = \frac{\int_{\mathbb{R}} y\hat{f}_{X,Y}(x,y)dy}{f_{X,h}(x)}. \tag{3.10}$$

3.3 Propriétés de l'estimateur à noyau

Nous obtenons la consistance des estimateurs à noyau, via la décomposition biais-variance suivante,

$$MSE\left(\hat{R}_n(x)\right) = \mathbb{E}\left(R(x) - \hat{R}_n(x)\right)^2 = \mathbb{V}ar\left(\hat{R}_n(x)\right) + \mathbb{B}iais^2\left(\hat{R}_n(x)\right). \tag{3.11}$$

3.3.1 Calcul de la variance

Nous posons, par convenance

$$\sigma^2(x) = \mathbb{V}ar\left[Y|X=x\right] = \frac{1}{f_X(x)}\int_{\mathbb{R}} y^2 f_{X,Y}(x,y)(x,y)dy - \left[R(x)\right]^2,$$

lorsque cette expression est bien définie.

Proposition 3.3.1. *On suppose* $\mathbb{E}\left[Y^2\right] < \infty$*. A chaque point de continuité des fonctions* $R(x)$*,* $f_X(x)$*, et* $\sigma^2(x)$*, tel que* $f_X(x) > 0$*,*

$$\mathbb{V}ar\left[\hat{R}_n(x)\right] = \frac{1}{nh} \times \left\{\frac{\sigma^2(x)}{f_X(x)}\int_{\mathbb{R}} K^2(\mu)d\mu\right\}(1 + o(1)), \tag{3.12}$$

où le terme $o(1)$ *tend vers* 0 *lorsque* $h \to 0$*.*

Proposition 3.3.2. *Si la fenêtre* $\hat{h}_n$ *satisfait les conditions* $h_n \to 0$ *et* $nh_n \to \infty$ *lorsque* $n \to \infty$*, la variance de l'estimateur à noyau de la fonction de régression tend vers zéro.*

3.3.2 Calcul du biais

L'estimateur à noyau de la fonction de régression se présente sous la forme d'un quotient aléatoire, c'est pourquoi on utilise généralement comme terme de centrage l'approximation suivante :

$$\tilde{\mathbb{E}}\left[\hat{R}_n(x)\right] = \frac{\mathbb{E}\left[\hat{r}_n(x)\right]}{\mathbb{E}\left[f_{h,X}(x)\right]} \tag{3.13}$$

Proposition 3.3.3. *Lorsque Y est bornée et $nh_n \to \infty$,*

$$\mathbb{E}\left[\hat{R}_n(x)\right] = \tilde{\mathbb{E}}\left[\hat{R}_n(x)\right] + O\left((nh)^{-1}\right). \tag{3.14}$$

Lorsque $\mathbb{E}(Y^2) < \infty$ et $nh_n^2 \to \infty$,

$$\mathbb{E}\left[\hat{R}_n(x)\right] = \tilde{\mathbb{E}}\left[\hat{R}_n(x)\right] + O\left((n^{\frac{1}{2}}h)^{-1}\right). \tag{3.15}$$

Proposition 3.3.4. *Supposons que $R(.)$ et $f_X(.)$ sont de classe $C^2(\mathbb{R})$ et que le noyau K est d'ordre 1, i.e. tel que*

$$\int_{\mathbb{R}} K(\mu)d\mu = 1, \int_{\mathbb{R}} \mu K(\mu)d\mu = 0 \quad et \quad \int_{\mathbb{R}} \mu^2 K(\mu)d\mu < \infty.$$

Nous avons alors, lorsque $h \to 0$ et $nh_n \to \infty$,

$$\mathbb{B}iais\left[\hat{R}_n(x)\right] = \frac{h^2}{2}\left\{\left\{R''(x) + 2R'(x)\frac{f'_X(x)}{f_X(x)}\right\}\int_{\mathbb{R}} \mu^2 K(\mu)d\mu\right\}(1 + o(1)). \tag{3.16}$$

Proposition 3.3.5. *Supposons que $R(.)$ et $f_X(.)$ sont de classe $C^q(\mathbb{R})$ et que le noyau K est d'ordre $q-1$.*
Nous avons alors, lorsque $h \to 0$ et $nh_n \to \infty$,

$$\mathbb{B}iais\left[\hat{R}_n(x)\right] = \frac{h^q}{q!}\left\{\left\{R^{(q)}(x) + qR^{(q-1)}(x)\frac{f'_X(x)}{f_X(x)}\right\}\int_{\mathbb{R}} \mu^q K(\mu)d\mu\right\}(1 + o(1)). \tag{3.17}$$

3.4 Optimalité asymptotique et choix des paramètres

On désigne par $\mathcal{K}[q]$ la classe des noyaux d'ordre q à support compact et bornés. Nous supposons, tout au long de cette section, que le noyau $K \in \mathcal{K}[q]$. L'hypothèse K borné et à support compact est très classique en régression non-paramétrique, elle implique notamment l'intégrabilité des divers moments de la fonction noyau $K(.)$.

Proposition 3.4.1. *Sous les hypothèses de la proposition (3.3.5), nous avons,*

$$\mathbb{B}iais\left[\hat{R}_n(x)\right] = \frac{h^q}{q!}\left\{\left\{R^{(q)}(x) + qR^{(q-1)}(x)\frac{f'_X(x)}{f_X(x)}\right\}\int_{\mathbb{R}} \mu^q K(\mu)d\mu\right\}(1 + o(1)) \tag{3.18}$$

$$= \frac{h^q}{q!}\left[b(x,q)\right](1 + o(1)). \tag{3.19}$$

Sous les hypothèses de la proposition (3.3.1), il s'ensuit

$$\begin{aligned}\mathbb{V}ar\left[\hat{R}_n(x)\right] &= \frac{1}{nh}\times\left\{\frac{\sigma^2(x)}{f_X(x)}\int_{\mathbb{R}}K^2(\mu)d\mu\right\}(1+o(1))\\ &= \frac{1}{nh}\times\left\{\frac{\sigma^2(x)}{f_X(x)}\left[\mu_0\left(K^2\right)\right]\right\}(1+o(1))\\ &= \frac{1}{hn}\times\left[\upsilon^2(x)\right](1+o(1)).\end{aligned}$$

Théorème 3.4.2. Sous les hypothèses des propositions (3.3.5) et (3.3.1), nous obtenons,

$$MSE\left(\hat{R}_n(x)\right) = \mathbb{V}ar\left[\hat{R}_n(x)\right]+\mathbb{B}^2iais\left[\hat{R}_n(x)\right] \tag{3.20}$$

$$= \left\{\frac{1}{hn}\times\left[\upsilon^2(x)\right]+\frac{h^{2q}}{(q!)^2}\left[b^2(x,q)\right]\right\}(1+o(1)) \tag{3.21}$$

Sous les hypothèses des propositions (3.3.5), (3.3.1) et (3.4.2) , nous obtenons, l'expression de l'erreur quadratique moyenne asymptotique ou AMSE

$$AMSE\left(\hat{R}_n(x)\right)=\frac{1}{hn}\times\left[\upsilon^2(x)\right]+\frac{h^{2q}}{(q!)^2}\left\{b(x,q)\right\}^2=AMSE(h), \tag{3.22}$$

La fenêtre optimale, au sens du critère local de minimisation de l'AMSE au point x, est alors obtenue en minimisant suivant h la quantité (3.22), c'est à dire

$$h_{n,opt}=\arg\min_h\left[AMSE(h)\right] \tag{3.23}$$

Lorsque $b(x,q)\neq 0$, nous obtenons :

$$h_{n,opt}=n^{-\frac{1}{(2q+1)}}\left\{\frac{q!(q+1)!\left\{\frac{\sigma^2(x)}{f_X(x)}\left[\mu_0\left(K^2\right)\right]\right\}}{2\left\{R^{(q)}(x)+q\times R^{(q-1)}(x)\frac{f'_X(x)}{f_X(x)}\right\}^2\left[\mu_q(K)\right]^2}\right\}^{\frac{1}{(2q+1)}} \tag{3.24}$$

La fenêtre $h_{n,opt}$ minimise donc asymptotiquement la MSE de l'estimateur [NW] au point x (critère local). Après calculs, il s'ensuit

$$\begin{aligned}AMSE^*(h) &= \left\{(q!)^{-\frac{2(q+1)}{2q+1}}\left\{\frac{(q-1)!}{2}\right\}^{\frac{2q}{2q+1}}+\left\{\frac{q!(q-1)!}{2}\right\}^{-\frac{1}{2q+1}}\right\}\times\\ &\quad\left\{\upsilon^2(x)\right\}^{\frac{2q}{2q+1}}\left[b(x,q)\right]^{\frac{2}{2q+1}}n^{\frac{-2q}{2q+1}}.\end{aligned}$$

A présent, on s'intéresse à l'estimation de la fonction de régression sur un intervalle $I\in\mathbb{R}$ et au risque global de l'estimateur de Nadaraya watson sur cet intervalle. On introduit pour cela l'erreur quadratique intégrée moyenne ou MISE ("mean integrated squared error"),

Théorème 3.4.3. Sous les hypothèses des propositions (3.3.5), (3.3.1) et (3.4.2

$$MISE\left(\hat{R}_n(x)\right) = \int_I MSE\left(\hat{R}_n(x)\right) dx = \left\{\frac{1}{hn} \times \left[\int_I v^2(x)dx\right] + \frac{h^{2q}}{(q!)^2}\int_I \left[b^2(x,q)\right] dx\right\}(1+o(1)) \tag{3.25}$$

La fenêtre optimale, au sens du critère global de minimisation de l'AMISE ("asymptotic mean integrated squared error") sur l'intervalle$I$, est donnée par,

$$h_{n,opt} = n^{-\frac{1}{(2q+1)}}\left\{\frac{q!(q+1)!\left\{\int_I v^2(x)dx\right\}}{2\int_I \left\{b(x,q)\right\}^2 dx}\right\}^{\frac{1}{(2q+1)}}. \tag{3.26}$$

3.5 Choix du paramètre de lissage

Dans cette section, nous supposons le noyau K fixé, et on ne s'intéresse qu'au choix de la fenêtre h.

3.5.1 La validation croisée

La fenêtre optimale qui minimise le risque quadratique intégré (MISE) est obtenue sous des hypothèses de régularité spécifiques et dépend alors de quantités inconnues, fonctionnelles de la distribution du couple (X, Y). Afin de construire un estimateur non oracle qui minimise l'erreur quadratique, il faut utiliser d'autres méthodes dont la plus commune est appelée la procédure de validation croisée. L'idée principale de la validation croisée consiste à minimiser, par rapport à h, l'estimé d'une mesure de la MISE.

La procédure de sélection du paramètre de lissage

Soient (X_1, Y_1), (X_2, Y_2),..., des variables aléatoires i.i.d. à valeurs dans $\mathbb{R}^p \times \mathbb{R}$. Nous considérons des estimateurs à noyaux, avec fenêtre h de la forme :

$$\hat{h} = \hat{h}_n = h_n\left\{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n); x\right\} \in H_n, x \in \mathbb{R}^p,$$

lorsque H_n désigne un sous-ensemble de $\mathbb{R}_n^+$ (i.e.,la zone de variation de $\hat{h}_n$).
Soit

$$d_I\left(\hat{R}_n(x), R(x)\right) = \int_{\mathbb{R}^p}\left\{R(x) - \hat{R}_n(x)\right\}^2 f_X(x)dx.$$

Maintenant, nous allons présenter la procédure de sélection de la fenêtre aléatoire $\hat{h}_n$ pour la distance d_I . On peut décomposer $d_I\left(\hat{R}_n(x), R(x)\right)$ de la manière suivante :

$$\begin{aligned} d_I\left(\hat{R}_n(x), R(x)\right)_I &= \int_{\mathbb{R}^p}\left\{R(x) - \hat{R}_n(x)\right\}^2 f_X(x)dx \\ &= \int_{\mathbb{R}^p} R_n^2(x) f_X(x)dx - 2\int_{\mathbb{R}^p} R(x)R_n(x)f_X(x)dx + \int_{\mathbb{R}^p} R^2(x)f_X(x)dx. \end{aligned}$$

Comme la dernière intégrale est indépendante de h, pour minimiser la perte associée à la distance d_I en fonction de h, il suffit de minimiser

$$\int_{\mathbb{R}^p} \hat{R}_n^2(x) f_X(x) dx - 2 \int_{\mathbb{R}^p} R(x) \hat{R}_n(x) f_X(x) dx. \tag{3.27}$$

Nous remarquons que le deuxième terme de l'intégrale

$$\int_{\mathbb{R}^p} R(x) R_n(x) f_X(x) dx = \mathbb{E}\left[\hat{R}_n(X) Y\right]$$

Il s'ensuit comme estimateur naturel,

$$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{R}_n(x_i) y_i\right],$$

où $\hat{R}_i(.)$ est l'estimateur défini par,

$$\hat{R}_i(x) = \frac{\sum_{j \neq i}^{n} y_j K\left(\frac{x - x_j}{h}\right)}{\sum_{j \neq i}^{n} K\left(\frac{x - x_j}{h}\right)}.$$

Il est possible d'approximer le premier terme intégrale de (3.35) par,

$$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{R}_n^2(x_i)\right],$$

En somme, il paraît raisonnable de choisir la fenêtre h qui minimise :

$$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{R}_n^2(x_i)\right] - \frac{2}{n} \sum_{i=1}^{n} \left[\hat{R}_n(x_i) y_i\right].$$

Cette dernière quantité est égale à

$$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{R}_n(x_i) - y_i\right]^2 - \frac{1}{n} \sum_{i=1}^{n} \left[y_i^2\right],$$

où le deuxième terme ne dépend pas de h et n'intervient donc pas dans la minimisation. Le critère de sélection de la fenêtre se réduit à choisir $\hat{h}$ qui minimise

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \hat{R}_n(x_i)\right]^2. \tag{3.28}$$

3.6 Estimation par la méthode des polynômes locaux

L'estimation de la fonction de régression par la méthode des polynômes locaux est fondée sur une simple généralisation de l'estimateur [NW]. L'idée maîtresse de l'approche localement polynomiale est de considérer le problème de la régression sous l'angle des moindres carrés. Intuitivement, cette démarche est pleine de bon sens, en dénotant que la fonction de régression $R(.)$ est elle même solution d'un problème de moindres carrés.

Définition 3.6. Nous rappelons la définition de l'estimateur [NW] : lorsque $K \geq 0$,

$$\hat{R}_n(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} = \frac{\hat{r}_n(x)}{f_{h,X}(x)}. \tag{3.29}$$

Nous avons, lorsque $K > 0$,

$$\left\{\hat{r}_n(x) - \hat{R}_n(x) f_{h,X}(x)\right\} = 0.$$

L'estimateur de la régression $\hat{R}_n(x)$ peut donc être regardé comme la solution du problème de moindres carrés pondérés suivant :

$$\arg\min_{\beta} \sum_{i=1}^n \{y_i - \beta\}^2 K\left(\frac{x - x_i}{h_n}\right)$$

Plus généralement, on introduit une fonction de poids $(w_i(x))$ construite à partir d'un noyau :

$$w_i(x) = K\left(\frac{x_i - x}{h}\right)$$

qui va attribuer un poids plus important aux observations pour lesquelles x_i est "proche" de x, et on minimise (en β) la somme des carrés pondérée :

$$\sum_{i=1}^n w_i \left(y_i - \beta\right)^2.$$

La solution est donnée par

$$\bar{R}_n(x) = \frac{\sum_{i=1}^n w_i(x) y_i}{\sum_{i=1}^n w_i(x)} \tag{3.30}$$

ce qui correspond à l'estimateur à noyau de la fonction de régression. On peut généraliser la formule ci-dessus en remplaçant la constante β par un polynôme de degré p : on se donne un point x en lequel on souhaite estimer la fonction de régression. Pour z dans un voisinage de x, on considère le polynôme

$$p(z, \beta) = \beta_0 + \beta_1(z - x) + \cdots + \frac{\beta_n}{p!}(z - x)^p.$$

On cherche à estimer la fonction de régression au voisinage de x par le polynôme $P_x(z;\beta)$ où le vecteur $\beta = (\beta_0, \ldots, \beta_p)$ est obtenu par minimisation de la somme des carrés pondérée

$$\sum_{i=1}^{n} w_i(x) \left(y_i - \beta_0 + \beta_1(z-x) + \cdots + \frac{\beta_n}{p!}(z-x)^p \right)^2. \tag{3.31}$$

La solution obtenue est le vecteur $\hat{\beta}(x) = (\hat{\beta}_0(x), \ldots, \beta_p(x))$, l'estimateur local de la fonction de régression R est

$$\hat{R}_n(z) = \hat{\beta}_0(x) + \hat{\beta}_1(z-x) + \cdots + \frac{\hat{\beta}_n}{p!}(z-x)^p, \tag{3.32}$$

Au point x, où l'on souhaite réaliser l'estimation, on obtient :

$$\hat{R}_n(z) = \hat{\beta}_0(x). \tag{3.33}$$

Remarque 3.6.1. Si $p = 1$, on parle de régression linéaire locale.

On peut expliciter la valeur de $\hat{\beta}_0(x)$ à partir d'un critère des moindres carrés pondérés : soit X_x la matrice

$$X_x = X = \begin{pmatrix} 1 & (x_1 - x) & \ldots & \frac{(x_1-x)^p}{p!} \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x) & & \frac{(x_n-x)^p}{p!} \end{pmatrix}$$

Nous posons

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad et \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

On désigne par W_x la matrice diagonale $n \times n$ de poids :

$$W_x = W = dia \left\{ K\left(\frac{x_i - x}{h_n} \right) \right\}. \tag{3.34}$$

On a alors :

$$\sum_{i=1}^{n} w_i(x) \left(y_i - \beta_0 + \beta_1(x_i - x) + \cdots + \frac{\beta_n}{p!}(x_i - x)^p \right)^2 = (Y - X_x\beta)^t W_x (Y - X_x\beta).$$

Minimiser l'expression ci-dessus conduit à l'estimateur des moindres carrés pondérés :

$$\hat{\beta} = \{X^t W X\}^{-1} X^t W Y, \tag{3.35}$$

et l'estimateur par polynômes locaux au point x correspond $\hat{R}_n(x) = \hat{\beta}_0(x)$, c'est-à-dire au produit scalaire du vecteur Y avec la première ligne de la matrice $\{X^t W X\}^{-1} X^t W$

On obtient le théorème suivant :

Théorème 3.6.1. *L'estimateur par polynômes locaux au point x est*

$$\hat{R}_n(x) = \sum_{i=1}^{n} l_i(x) y_i.$$

où, $l(x)^t = (l_1(x), \ldots, l_n(x))$,

$$l(x)^t = e_1^t \{X^t W X\}^{-1} X^t W,$$

avec $e_1^t = (1, 0, , ..., 0)$

$$\mathbb{E}\left(\hat{R}_n(x)\right) = \sum_{i=1}^{n} l_i(x) f(x_i).$$

$$\mathbb{V}\left(\hat{R}_n(x)\right) = \sigma^2 \sum_{i=1}^{n} l_i^2(x).$$

Bibliographie

[1]Arnak S.Statistique avancée : méthodes non-paramétriques. Ecole Centrale de Paris.

[2] D. Bosq and J.P. Lecoutre. Théorie de l'estimation fonctionnelle. Economica Edition, 1987.

[3] D. Blondin. Lois limites uniformes et estimation non-paramétrique de la régression. Mathématiques [math]. Université Pierre et Marie Curie - Paris VI, 2004.

[4] F. Comte. Estimation non paramétrique. Spatacus supérieur. Collection Recherche. 2005.

[5] N. Cencov. Estimation of an unknown distribution density from observations. Sovet.Math, 3, p. 1559-1562, 1962.

[6] E. Nadaraya, (1989). Nonparametric Estimation of Probability Densities and Regression Curves. Kluwer, Dordrecht.

[7] E. Parzen. On estimation of a probability density function and mode. Ann. Math. Statist, 33 :1065–1076, 1962.

[8] D.W. Scott and G.R. Terrell. Oversmoothed nonparametric density estimates. Journal of the American Statistical Association, 80 :209–214, 1985.

[9] M. Rosenblatt. Remarks in some nonparametric estimates of a density function. Ann. Math. Statist, 27 :832–837, 1956.

[10]N. Zougab. Etude comparative des méthodes de sélection du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau. université de Béjaia.2007.