*Séminaire Mathématique de Béjaïa (LaMOS)*
*Volume 3, 2005, pp. 77 - 84*

# 12

# Estimation de la fonction régression par la méthode du noyau. Propriétés statistiques

K. LAGHA

Laboratory of Modelisation and Optimisation of Systems
*University of Bejaia, 06000 Bejaia, Algéria*
email : karima_lagha@yahoo.com

**Résumé** Dans ce travail, nous considérons le problème d'estimation non paramétrique par la méthode du noyau de la fonction de régression. Nous étudions la consistance de l'estimateur de Nadaraya-Watson à travers les propriétés statistiques et asymptotiques de l'estimateur, à savoir le biais, la variance, l'erreur quadratique moyenne, asymptotique et intégrée, la normalité asymptotique ainsi que les lois uniformes du logarithme pour la fonction de régression.

**Mots clés :** Fonction de régression, estimation non paramétrique, Lois uniformes du logarithme, paramètre de lissage.

## 12.1 Introduction

On considère le modèle de régression suivant :
Soit un $n$-échantillon

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

On suppose qu'il existe une fonction $m(.)$ telle que pour $i, \quad 1 \leq i \leq n$

$$Y_i = m(X_i) + \epsilon_i, \qquad \epsilon_i \curvearrowright \mathcal{N}(\mu, \sigma^2)$$

et les $\epsilon_i$ sont II des $X_i$
Si $(X, Y)$ a une fonction densité conjointe $f_{X,Y}$ et $E|Y| < \infty$ alors $m(.)$ est définie par :

$$m(x) = E\left(Y \setminus X = x\right) = \frac{\int y f(x,y) dy}{\int f(x,y) dy} = \frac{r(x)}{f(x)}$$

### Estimation non paramétrique de $m(.)$

Soit $m(.)$ une fonction dans une classe de fonction $\mathcal{F}$ (exp : classe de fonctions continues sur $[0, 1]$). On considère l'estimateur ayant la forme suivante (estimateur linéaire par rapport à $Y_i$) :

$$\hat{m}_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

Où $W_{ni}(.)$ est II des obs $Y_i$.

### 12.1.1 Estimateur de Nadaraya-Watson NW

En utilisant l'estimateur à noyau de la fonction densité (proposé par Parzen-Rozenblatt), défini par :

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

tel que :

1. $h = h_n \longrightarrow 0 \quad \text{qd} \quad n \to \infty$
2. $K : R \longrightarrow R$ f$^{ct}$ mesurable vérifiant :
   (K.1) $K$ est bornée : $\sup_{u\in R} |K(u)| < \infty$
   (K.2) $\lim_{|u|\to\infty} |u|K(u) = 0$
   (K.3) $K(.) \in L_1(R) : \int_R |K(u)| < \infty$
   (K.4) $\int_R K(u) = 1$

L'estimateur de NW est se présente alors :

$$\hat{m}_n(x) = \begin{cases} \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} & \text{si} \quad \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \neq 0 \\ \\ \frac{1}{n}\sum_{i=1}^{n} Y_i & \text{sinon} \end{cases}$$

Le noyau $K$ détermine la forme du voisinage autour de x et la fenêtre $h$ contrôle la taille de ce voisinage.
Ou encore, $\hat{m}_n(x) = \frac{\hat{r}_n(x)}{\hat{f}_n(x)}$ avec

$$\hat{r}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) = \int y\hat{f}(x, y)dy$$

- Phénomène de sous lissage : Variance est trop grande
- Phénomène de sur lissage : Biais (erreur déterministe) est trop grande

Les propriétés statistiques de NW dépendent de h (à choisir de sorte à équilibrer le biais et la variance).

### 12.1.2 Consistance de l'estimateur

La décomposition biais-variance donne :

$$E\left[\{\hat{m}_n(x) - m(x)\}^2\right] = Var\left[\hat{m}_n(x)\right] + \{E[\hat{m}_n(x)] - m(x)\}^2$$

Lorsque cette espérance tend vers 0, on a

$$\hat{m}_n(x) \xrightarrow{L^2} m(x), \quad \text{d'où}, \;\; \hat{m}_n(x) \xrightarrow{P} m(x)$$

**Calcul de la variance**
On a

$$Var(\hat{f}_n(x)) = \frac{1}{nh} f(x) \int K^2(z)dz(1+o(h)), \quad h \to 0.$$
$$Var(\hat{r}_n(x)) = \frac{1}{nh} s(x) \int K^2(z)dz(1+o(h)), \quad h \to 0$$

où $s(x) = \int y^2 f(x,y)dy$ et

$$Cov\left(\hat{f}_n(x), \hat{r}_n(x)\right) = E\left[\hat{f}_n(x)\hat{r}_n(x)\right] - E\hat{f}_n(x)E\hat{r}_n(x) = E\left[\{\hat{f}_n(x) - E\hat{f}_n(x)\}\{\hat{r}_n(x) - E\hat{r}_n(x)\}\right]$$
$$= \frac{1}{nh} r(x) \int K^2(z)dz(1+o(h)), \;\; r(x) = \int y f(x,y)dy.$$

En utilisant la matrice variance-covariance on obtient

$$Var(\hat{m}_n(x)) = \frac{\sigma^2(x)}{nhf(x)} \int K^2(u)du(1+o(h))$$

**Calcul du biais**
On définit $\tilde{E}(\hat{m}_n(x)) = \frac{E(\hat{r}_n(x))}{E(\hat{f}_n(x))}$. D'où,

$$E(\hat{m}_n(x)) = \tilde{E}(\hat{m}_n(x)) + \frac{a_n(x) + b_n(x)}{\left(E\hat{f}_n(x)\right)^2}$$

avec,

$$\begin{cases} a_n(x) = -Cov(\hat{f}_n(x), \hat{r}_n(x)) = o\left(\frac{1}{nh}\right), \\ b_n(x) = E\left[\hat{m}_n(x)\left(\hat{f}_n(x) - E\hat{f}_n(x)\right)^2\right], \end{cases}$$

On a deux cas :

**1 cas** $Y$ est bornée : $\exists M$ tel que $|Y| \leq M$ alors $|\hat{m}_n(x)| \leq M$ d'où, $b_n(x) \leq o\left(\frac{1}{nh}\right)$.

$$E(\hat{m}_n(x)) = \tilde{E}(\hat{m}_n(x)) + o\left(\frac{1}{nh}\right)$$

**2 cas** $EY^2 < \infty$ et $nh^2 \to \infty$ alors $|\hat{m}_n(x)| \leq \sum_j |Y_j|$, d'où $b_n(x) \leq o\left(\frac{1}{n^{1/2}h}\right)$.

$$E(\hat{m}_n(x)) = \tilde{E}(\hat{m}_n(x)) + o\left(\frac{1}{n^{1/2}h}\right)$$

De plus,

$$\tilde{E}(\hat{m}_n(x)) - m(x) = \frac{h}{2}\left\{m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right\} \int u^2 K(u)du$$

### 12.1.3 Optimalité et Normalité asymptotique

Les C.N.S. sur $h_n$ pour obtenir la consistance de $\hat{m}_n(x)$ sont $h_n \to 0$ et $nh_n \longrightarrow \infty$ qd $n \to \infty$.

• L'erreur quadratique moyenne (MSE)

$MSE(\hat{m}_n(x)) = E\left[\{\hat{m}_n(x) - m(x)\}^2\right] = \frac{h^{2q}}{(q!)^2}\left\{m^{(q)}(x) + qm^{(q-1)}(x)\frac{f'(x)}{f(x)}\right\}^2 \times$
$\left(\int u^q K(u)du\right)^2 (1 + o(h)) + \frac{1}{nh}\left\{\frac{\sigma^2(x)}{f(x)}\right\}\int k^2(u)du(1 + o(h))$

• L'erreur quadratique moyenne Asymptotique(AMSE)

$$AMSE(\hat{m}_n(x)) = \frac{h^{2q}}{(q!)^2}\left\{m^{(q)}(x) + qm^{(q-1)}(x)\frac{f'(x)}{f(x)}\right\}^2 \times \left(\int u^q K(u)du\right)^2 + \frac{1}{nh}\left\{\frac{\sigma^2(x)}{f(x)}\right\}\int k^2(u)du$$

(Minimiser suivant $h$ la qtte ci-dessus).

• L'erreur quadratique intégrée moyenne (MISE)

$$MISE(\hat{m}_n(x)) = E\left[\int \{\hat{m}_n(x) - m(x)\}^2 dx\right] = \int MSE(\hat{m}_n(x))dx$$

$$= \frac{h^{2q}}{(q!)^2}\left\{\int m^{(q)}(x)dx + q\int m^{(q-1)}(x)\frac{f'(x)}{f(x)}dx\right\}^2 \times$$

$$\left(\int u^q K(u)du\right)^2 + \frac{1}{nh}\int \frac{\sigma^2(x)}{f(x)}dx \int k^2(u)du(1 + o(h)).$$

### Normalité asymptotique (Schuster 1972)

Sous les conditions : $K$ noyau borné, à support compact et d'ordre 2 et $h_n$ égale à $cn^{-1/5}$, on a le théorème suivant :

**Théorème Hardle 1990**

Y bornée (ou de moment d'ordre $> 2$), $f(.)$ et $m(.)$ sont de classe $C^2(R)$, $\forall x$ tel que $\sigma^2(x)$ est continue, $f(x) > 0$,

$$(nh)^{1/2}\{\hat{m}_n(x) - m(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}\left(B(x), v^2(x)\right)$$

$$\begin{cases} B(x) = \left\{m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\int u^2 K(u)du\right\}, \\ \\ v^2(x) = \frac{\sigma^2(x)}{f(x)}\int K^2(u)du. \end{cases}$$

## 12.2 Lois Uniformes du logarithme pour la fonction régression

On utilise le principe d'invariance : approximation du processus empririque uniforme par une suite de ponts Browniens et obtention de lois uniforme du logarithme. Soit :

$$m_\psi(x) = E\left(\psi(Y) \setminus X = x\right) = \frac{\int \psi(y) f(x,y) dy}{\int f(x,y) dy} = \frac{r_\psi(x)}{f(x)}$$

$\psi(.)$ est une fct réelle mesurable et bornée sur des compacts.
• Hypothèses sur $f$ :
(F.1) $f_{X,Y}(.,.)$ est continue sur $I \times R$, $I = [a,b]$
(F.2) $f_X(.)$ est continue $> 0$ sur $I$
(F.3) $Y \mathbb{1}_{\{X \in I\}}$ est bornée
• Hypothèses sur $K$ :
(K.1) $K(.)$ est à variation bornée et continue sur $R$ :

$$\int |dK(u)| = |K|_v < \infty$$

(K.2) $K(u) = 0$ si $u \neq [-\epsilon/2, \epsilon/2]$, $0 < \epsilon < \infty$.
(K.3) $\int K(u) du = 1$.
• Hypothèses sur $h$ :
(H.1) $h_n \longrightarrow 0$ qd $n \longrightarrow \infty$
(H.2) $nh_n / \log n \longrightarrow \infty$ qd $n \longrightarrow \infty$
(C.N.S pour la cvgence uniforme en proba de la déviation associée aux estimateurs).
(H.3) $h_n \searrow 0$ et $nh_n \nearrow \infty$ qd $n \longrightarrow \infty$.
(H.4) $|\log h_n| / \log\log n \longrightarrow \infty$ qd $n \longrightarrow \infty$.
(C. reliées à la cvgence p.s.)
On considère le processus :

$$W_n(x,\psi) = \sum_{i=1}^{n} \psi(Y_i) K\left(\frac{x - X_i}{h}\right) - nE\left\{\psi(Y) K\left(\frac{x - X_i}{h}\right)\right\}$$

Sous les conditions :
(F.1-3), (H.1-2) et (K.1-4) qd $n \longrightarrow \infty$ on a :

$$\left| \sup_{x \in I} \frac{\{\pm W_n(x,\psi)\}}{\{2nh \log 1/h\}^{1/2}} - \sigma_W(I) \right| \xrightarrow{p} 0$$

Sous les conditions :
(F.1-3), (H.2-4) et (K.1-4) qd $n \longrightarrow \infty$ on a :

$$\left|\sup_{x\in I}\frac{\{\pm W_n(x,\psi)\}}{\{2nh\log 1/h\}^{1/2}}-\sigma_W(I)\right|\xrightarrow{p.s.}0$$

où

$$\sigma_W^2(I)=\sup_{x\in I}E\left[\psi^2(Y)\setminus X=x\right]f(x)\left(\int K(t)dt\right)^2$$

Il s'agit de montrer que $\forall\epsilon>0$,

$$P\left\{\sup_{x\in I}\frac{|W_n(x,\psi)|}{\{2nh\log 1/h\}^{1/2}}>(1+\epsilon)\sigma_W(I)\right\}\overset{p(oup.s.)}{\longrightarrow}0$$

et

$$P\left\{\sup_{x\in I}\frac{|W_n(x,\psi)|}{\{2nh\log 1/h\}^{1/2}}<(1-\epsilon)\sigma_W(I)\right\}\overset{p(oup.s.)}{\longrightarrow}0$$

**Discrétisation**

Examiner le processus empirique fonctionnel $W_n(.,\psi)$ pour un nombre fini de points de I.

$$\begin{aligned}W_n(x,\psi)&=\sum_{i=1}^{n}\left[\psi(Y_i)K\left(\frac{x-X_i}{h}\right)-E\left(\psi(Y_i)K\left(\frac{x-X_i}{h}\right)\right)\right]\\&=n^{1/2}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(g(X_i,Y_i)-Eg(X_i,Y_i)\right)\right]=n^{1/2}\alpha_n(g_{n,x})\end{aligned}$$

où $g_{n,x}(u,v)=\psi(v)K\left(\frac{x-u}{h}\right)$. On considère pour $0<\delta<1$, $I=[a,b]$

$$z_{n,i}=a+i\delta h_n,\quad 0\le i\le l_n=\lceil(b-a)/\delta h_n\rceil.$$

D'où l'étude du processus empirique indéxé par la classe $G_n\left\{g_{n,i},\ 0\le i\le l_n\right\}$ et

$$P\left\{\max_{0\le i\le l_n}\frac{|\alpha_n(g_{n,i})|}{\{2nh\log 1/h\}^{1/2}}>(1+\epsilon)\sigma_W(I)\right\}\overset{p(oup.s.)}{\longrightarrow}0$$

**Proposition**

$Z_j$ sont des v.a. centrées de variance $\sigma^2<\infty,\ \forall j$ et $\exists M>0,\ |Z_j|<M\ \forall j$ alors $\forall t>0$,

$$P\left\{\sum_{j=1}^{n}Z_j>t\sqrt{n}\right\}\le exp\left(\frac{-t^2}{2\sigma^2+\frac{2}{3}Mn^{-1/2}t}\right)$$

Dans notre cas, on pose : $Z_j=g_{n,i}(X_j,Y_j)-Eg_{n,i}(X_j,Y_j)$.

$Z_j$ sont centrées et bornées par $2\|\psi\|.\|K\|=M$ ($K$ étant à variation bornée donc $\|K\|<\infty$).

D'après la proposition

$$P(\sum_j Z_j>t\sqrt{n})=P(\alpha_n(g_{n,i})>t)$$

On prend $t = \sigma_W(I)(1+\epsilon)\sqrt{2h_n \log 1/h_n}$, d'où

$$P\left(\frac{|\alpha_n(g_{n,i})|}{\sqrt{2h_n \log 1/h_n}} > \sigma_W(I)(1+\epsilon)\right) \leq 2(l_n+1)exp\left(-(1+\epsilon)^2 \log 1/h_n\right) = o(h_n^{(1+\epsilon)^2-1}).$$

Le résultat s'ensuit en utilisant le lemme suivant :

**Lemme de Cantelli**

Soit $(\sum, \mathcal{A}, P)$ un espace de probabilité quelconque, $\forall\{A_n, n \geq 1\} \subset \mathcal{A}$ (evts mesurables)

$\sum_{n\geq 1} P(A_n) < \infty \implies P(\limsup_n A_n) = 0$

(quand $\sum A_n$ est Convergente $A_n^C$ est vrai p.s).

## 12.3 Application Statistique

**Fenêtre adaptive et Intervalle de confiance**

• Pour construire un intervalle de confiance on utilise souvent la normalité asymptotique et les lois qui en découlent.

• Du point de vue statistique, la convergence en probabilité est une notion suffisante. La convergence p.s. nous oblige souvent à supposer des conditions plus fortes sur le paramètre de lissage.

• On considère les lois limites uniformes du logarithme pour le mode de convergence en proba afin de déterminer l'intervalle de confiance (car ça necessite des hypothèses moins restrictives sur la fenêtre).

**Corollaire**

Sous les hyp (F.1-3), (H.1-3) et (K.1-3), qd $n \longrightarrow \infty$ (Déviation maximale par rapport à l'espérance)

$$\left\{\frac{nh_n}{2\log 1/h_n}\right\}^{1/2} \sup_{x\in I} \pm \left\{\frac{\hat{f}_n(x)}{\hat{\sigma}_n^2(x)}\right\}^{1/2} \{\hat{m}_n(x) - \tilde{E}[\hat{m}_n(x)]\} \xrightarrow{P} \int K^2(u)du$$

(le biais est négligeable qd $h_n$ est d'ordre $n^{-\delta}, (2l+1)^{-1} \leq \delta \leq 1$.

$$\sup_{x\in I}\left\{\tilde{E}[\hat{m}_n(x)] - m(x)\right\} = O(h_n^l), l > 1$$

Si $h_n = n^{-1/(2l+1)}, n \to \infty$

$$\left\{\frac{nh_n}{2\log 1/h_n}\right\}^{1/2} \sup_{x\in I} \pm \left\{\frac{\hat{f}_n(x)}{\hat{\sigma}_n^2(x)}\right\}^{1/2} \{\hat{m}_n(x) - m_n(x)\} \xrightarrow{P} \int K^2(u)du$$

Où,

$$\hat{\sigma}_n^2(x) = \frac{\sum_{i=1}^n \left\{\psi(Y_i) - \hat{m}_n(x)\right\}^2 K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}$$

Si $\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \neq 0$. D'où, Pour tout $\epsilon \in ]0,1[$ qd $n \to \infty$

$$P\left\{m(x) \in \left[\hat{m}_n(x) - (1+\epsilon)L_n(x), \hat{m}_n(x) + (1+\epsilon)L_n(x)\right], \forall x \in I\right\} \longrightarrow 1 \quad \text{et}$$

$$P\left\{m(x) \in \left[\hat{m}_n(x) - (1-\epsilon)L_n(x), \hat{m}_n(x) + (1-\epsilon)L_n(x)\right], \forall x \in I\right\} \longrightarrow 0$$

avec

$$L_n(x) = \left\{\frac{2\log 1/h_n}{nh_n} \times \frac{\hat{\sigma}_n^2(x)}{\hat{f}_n(x)}\right\}^{1/2} \left[\int K^2(u)du\right]^{1/2}$$

D'où, L'intervalle de confiance asymptotiquement optimale pour $\hat{m}_n(x), x \in I$ est

$$\left[\hat{m}_n(x) - L_n(x), \hat{m}_n(x) + L_n(x)\right]$$