Université de Abdelhamid Ibn Badis de Mostaganem

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et d'Informatique

Mémoire de master

Thème

Estimation non paramétrique de la regression

Larbi Hassiba

Mombre de jury :

Mr.medeghri.A Président

Mr.Mechdène.M Examinateur

Mr.mohammedi.M Encadreur

Soutenu le 26/05/2015

Année Universitaire 2014 -2015

Table des matières

Remerciements

Tout d'abord je remercie mon grand Dieu le tout puissant qui m'a donné son aide à completer ce travail et à chèrcher le savoir.

Je passe mes remèrciement à mes enseignant de l'univercité.

Je remèrcie chaleureusement mes parent qui m'ont encouragé en tout long de mes études et toute ma famille chaqu'un avec son non.

Sans oublier mes amis et tout ce qui m'aidait de loin ou de proche à terminer ce mémoire.

INTRODUCTION

Soit $M = (\Omega, A, P_\theta)$ une modèle statistique avec $\theta \in \Phi$ Lorsque Φ est vaste de dimension infini, M est une modèle non paramétrique. A ce modèle on associe les méthodes d'estimation non paramétriques, étudiées par [Bosq] et d'autres.

Pour estimer n'importe quel paramètre fonctionnel il suffit d'estimer la fonction de répartition F par la fonction de répartition empirique F_n, et par conséquent l'estimateur θ_n de θ est $T(F_n)$ où T est la fonctionnelle statistique.

La fonction de répartition empirique donc joue un rôle fondamental dans l'estimation fonctionnelle plus précisément dans l'estimation de la densité f, pour qu'on puisse tirer plus d'information sur la loi parente. La connaissance de l'estimateur de F et f mènent à résoudre un autre problème fondamental de la statistique non paramétrique, c'est le problème de la régression. Ce problème peut être généralisé en cherchant une application mesurable quelconque « r », telle que $r(x)$ soit la plus proche de Y au sens des moindres carrés. Si Y est de carré intégrable, l'espérance de Y conditionnelle à X notée $E^X(Y)$ est la solution de ce problème. C'est la raison pour laquelle on introduit la recherche de l'estimateur de la fonction de régression « r» et la convergence presque complète.Pour cela, la méthode du noyau introduite par Rosenblatt (1956) pour estimer la fonction de densité a été reprise simultanément par Waston (1964) et Nadaraja (1964) pour estimer la fonction de régression.

Nous avons choisi d'organiser notre mémoire selon le plan décrit ci-dessous :

Le premier chapitre introduit le modèl non paramétrique et présente une inégalité de grande déviation, qui s'appelle

« l'inégalité de Bernshtein .Frechet » , fondamentale dans l'étude de la vitesse de convergence ponctuelle des estimateurs fonctionnels. L'estimateur de F et f est donnés dans le chapitre 2 avec la vitesse de convergence. En se basant sur l'estimateur de la densité, on présente une étude asymptotique de la fonction de régression.

Finalement, dans le dernier chapitre on généralise la notion de l'estimateur de « r » au cadre vectoriel.

Chapitre 1

Introduction au cadre fonctionnel :

1.1 Préliminaires :

Définition 1.1.1 *Soit la structure statistique suivant* $(\Omega, A, P_\theta), \theta \in \Phi$ *et* $\Phi \subset \mathbb{R}^R$

Ω *:espace fondamental*

A *:tribu*

P *:ensemble de probabilité*

Φ *:ensemble des paramètres*

si $R < \infty$ *on dit que la statistique est paramétrique.*

si $R = \infty$ *on dit que la statistique est non paramétrique.*

Exemple 1.1.1 *On prend* $\Omega = \{(0,1)\}$

$A = \rho(\Omega)$: *c'est l'ensemble des parties de* Ω.

on a $P_p(0) = 1 - p = q$ *et* $P_p(1) = p$

c'est le modèl de Bernoulli

Donc ;

$(\Omega, A, P_\theta) : \{A = \rho(\Omega)\,;\ \Omega = \{(0,1)\}\ ;\ \theta = p\}$

Définition 1.1.2 *On appelle fonctionelle statistique toute application* T :

$$T : \mathbb{F} \to \Phi$$

$$F \to T(F) = \theta$$

$$F \to d(F) = F' = f$$

où ;

Ŧ : *l'espase des fonctions de répartition*

d : est la dérivé.

F : la fonction de répartition.

f :la densité.

On dit que $\theta = f$ un paramètre fonctionnel.

Exemple 1.1.2 *L'espérence de la variable X :*

$$E(X) = \int X \ dF = T(F)$$

Définition 1.1.3 ***La convergence prèsque complète :***

Soit la suite de variables aléatoires (x_n), $n \in \mathbb{N}$

x_n convèrge vers x complètement si :

Pour tout $\varepsilon > 0$

$$\sum_{n=0}^{\infty} P(|x_n - x|) > \varepsilon) < \infty$$

$$\Rightarrow |x_n - x| \to 0$$

$$\Rightarrow x_n \to x$$

Définition 1.1.4 ***La vitèsse de la convèrgence :***

Si on écrit

$$x_n = 0(y_n)$$

ça veut dire

Il existe $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P(|x_n| > \varepsilon \, |y_n|) < \infty$$

y_n est la vitèsse de la convèrgence.

Soient les deux suites numériques $(X_n)_{(n \in \mathbb{N})}; (Y_n)_{(n \in \mathbb{N})}$

On écrit :

$$X_n = o(Y_n); \ si \ \lim_{n \to \infty} \frac{X_n}{Y_n} = 0 \blacksquare$$

Et on écrit :

$$X_n = O(Y_n); \quad si \left| \frac{X_n}{Y_n} \right| \leq c$$

c est une constante.

1.2 L'inégalité de Bernshtein Frechet

Lemme 1.2.1 ***L'inégalité de Bernshtein Frechet :***

Soit $(X_1, X_2, ..., X_n)$ une suite de variables aléatoires indépendantes telque

$\alpha_i \leq X_i \leq \beta_i$ et $\alpha_i, \beta_i \in \mathbb{R}$

pour tout $t > 0$ on a :

$$P(\left| \sum_{i=1}^{n} (X_i - E(X_i)) \right| \geq t) \leq 2 \cdot \exp(\frac{-2t^2}{\sum_{i=1}^{n} (\beta_i - \alpha_i)^2}) \qquad (1)$$

Preuve. □

Posons $h > 0$

$$\mathbf{1}_{\mathbf{A}=(\sum_{i=1}^{n}(X_i - E(X_i)) - t \geq 0)} \leq \exp(h \sum_{i=1}^{n} (X_i - E(X_i)) - t) ...(1)$$

On sait que :

$$\mathbf{1}_A = \{0 \; si \; x \notin A \; ou \; 1si \; x \in A\}$$

$*Si \; \mathbf{1}_A = 0$

$$\begin{aligned} (\sum_{i=1}^{n} (X_i - E(X_i)) - t) &\leq 0 \\ et \; \exp(h \sum_{i=1}^{n} (Xi - E(Xi)) - t) &\geq 0 \end{aligned}$$

$*Si \; \mathbf{1}_A = 1$

$$\begin{aligned} \exp(h \sum_{i=1}^{n} (X_i - E(X_i)) - t) &\geq 1 \\ \exp(h \sum_{i=1}^{n} (Xi - E(Xi)) - t) &\geq 0 \end{aligned}$$

alors (1) est vrais dons les deux cas.

on sait que $E(\mathbf{1}_A(x)) = P(A)$

Donc :

$$\begin{aligned} P(A) &= P(\sum_{i=1}^{n}(X_i - E(X_i)) - t) \geq 0) \\ &\leq E\left[\exp(h\sum_{i=1}^{n}(X_i - E(X_i)) - t)\right] \\ &\leq E\left[\exp(h\sum_{i=1}^{n}(X_i - E(X_i)))\right].\exp(-ht) \\ &\leq \prod_{i=1}^{n} E\left[\exp(h(X_i - E(X_i)))\right].\exp(-ht) \\ &\leq \prod_{i=1}^{n} E(\exp(-h.E(X_i)).E(\exp(h.X_i))\ .\exp(-ht) \end{aligned}$$

On va poser $(2) = E(\exp(-h.E(X_i)).E(\exp(h.X_i))$

Pour $(h.x_i)$ on va utiliser le fait que cette fonction est convexe, on pose

$$\varphi(X_i) = \exp(h.X_i)$$

ou φ est convexe vérifie

$$\varphi(\alpha x + \beta y) \leq \alpha\varphi(x) + \beta\varphi(y)\ o\grave{u}\ \alpha, \beta \in \mathbb{R}$$

Posons

$$\alpha = \frac{\beta_i - X_i}{\beta_i - \alpha_i}\ et\ \beta = \frac{X_i - \alpha_i}{\beta_i - \alpha_i}$$

Il est claire que $\alpha + \beta$ =1

On pose

$$\begin{aligned} X_i &= \alpha x + \beta y \\ x &= \alpha i \\ y &= \beta i \end{aligned}$$

$$\varphi(\alpha x + \beta y) = \exp(h(\alpha x + \beta y)) \leq \alpha\varphi(h.x) + \beta\varphi(h.y)$$

$$\Rightarrow \exp(h.X_i) \leq \frac{\beta_i - X_i}{\beta_i - \alpha_i} \exp(h.\alpha_i) + \frac{X_i - \alpha_i}{\beta i - \alpha_i} \exp(h.\beta_i)$$

$$\Rightarrow E(\exp(h.X_i)) \leq \frac{\beta_i - E(X_i)}{\beta_i - \alpha_i} \exp(h.\alpha_i) + \frac{E(X_i) - \alpha_i}{\beta_i - \alpha_i} \exp(h.\beta_i)$$

$$\Rightarrow (2) \leq E(\exp(-h.E(X_i)).\left[\frac{\beta_i - E(X_i)}{\beta_i - \alpha_i} \exp(h.\alpha_i) + \frac{E(X_i) - \alpha_i}{\beta_i - \alpha_i} \exp(h.\beta_i)\right]$$

On éssaye de mettre (3) sous la forme

$$\exp(\psi(h_i))$$

où

$$hi = h(\beta_i - \alpha_i)$$

i.e

$$(3)(hi) = \exp(\psi(h_i))$$

D'aprés le devleppement limitée de $\psi(hi)$ on a :

$$\psi(h_i) = \psi(0) + \psi'(0).h_i + \frac{1}{2}\psi"(\zeta).{h_i}^2 \text{ où } \zeta \in [0, h_i]$$

On trouve que :

$$\psi(0) = 0 \text{ et } \psi'(0) = 0 \text{ et } \psi"(h_i) \leq \frac{1}{4}$$

Donc :

$$|\psi(h_i)| \leq \frac{1}{4}.{h_i}^2 = \frac{h^2(\beta_i - \alpha_i)^2}{8}$$

On peut dire que :

$$(3)(h_i) = \exp(\frac{h^2(\beta_i - \alpha_i)^2}{8})$$

Celà veut dire que :

$$P(\sum_{i=1}^{n}(X_i - E(X_i)) - t \geq 0) \leq \exp(-ht). \prod_{i=1}^{n} E\left[\exp(h(X_i - E(X_i)))\right]$$

$$\leq \exp(-ht).\exp\left[\frac{h^2\sum(\beta_i-\alpha_i)^2}{8}\right]$$
$$\leq \exp\left[-ht+\frac{h^2\sum(\beta_i-\alpha_i)^2}{8}\right]$$

La relation est vraie pour $h \geq 0$.

Application :

On pose :

$h = \frac{4t}{\sum(\beta i-\alpha i)^2} \geq 0$

$$P(\sum_{i=1}^{n}(X_i - E(X_i)) - t \geq 0) \leq \exp\left[\frac{-4t^2}{\sum_{i=1}^{n}(\beta_i-\alpha_i)^2} + \frac{16t^2}{8\sum_{i=1}^{n}(\beta_i-\alpha_i)^2}\right]$$
$$\leq \exp\left[\frac{-2t^2}{\sum_{i=1}^{n}(\beta_i-\alpha_i)^2}\right]$$

On trouve le mème résultat pour

$$A = (\sum_{i=1}^{n}(X_i - E(X_i)) - t \leq 0).$$

Enfin on conclut :

$$P(\left|\sum_{i=1}^{n}(X_i - E(X_i))\right| \geq t) \leq 2\cdot\exp(\frac{-2t^2}{\sum_{i=1}^{n}(\beta_i-\alpha_i)^2})$$

C'est le résultat final.

Chapitre 2

L'estimation de la fonction de répartition et de la densité :

2.1 L'estimation de la fonction de répartition :

Définition 2.1.1 *Soit $(x_1, x_2, ..., x_n)$ un $n-$échantillon de x de fonction de répartition F, on note par Fn la fonction de répartition empérique s'écrit :*

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

Théorème 2.1.1 *$F_n(x)$ est un éstimateur sans biais de F.*

Preuve. Soit $(x_1, x_2, ..., x_n)$ un $n-$échantillon de x

$$\begin{aligned} F_n(x) &\in \{0,1\} \Rightarrow nF_n(x) \in \{0,n\}. \\ P(nF_n(x) &= 0) = P(x \leq x_i) \\ &= P(\bigcap x \leq x_i) \\ &= \prod_{i=1}^{n} P(x \leq x_i) \\ &= \prod_{i=1}^{n} (1 - P(xi \leq x)) \\ &= (1 - F(x))^n \end{aligned}$$

. □

Donc ;

nF_n est une valeur aléatoire binomiale des paramètres $(n, F(x))$.

Alors :

$$\begin{aligned} E(nF_n(x)) &= nF(x) \\ &\Rightarrow E(nF_n(x) = nE(F_n(x))) \\ &\Rightarrow E(F_n(x) = F(x) \\ &\Rightarrow E(F_n(x)) - F(x) = 0 \end{aligned}$$

Par suite, $F_n(x)$ est un estimateur sans biais de F.

Théorème 2.1.2 *Soit la fonction de répartition F,la fonction empérique F_n vérifie :*

$$F_n(x) - F(x) = O(\sqrt{\frac{\log n}{n}}) \ p.co$$

Preuve. □

D'aprés la définition de la vitèsse de convèrgence, on écrit :

Il existe $\varepsilon > 0$,

$$\sum_{i=1}^{n} P(|F_n(x) - F(x)| > \varepsilon \left|\sqrt{\frac{\log n}{n}}\right|) < \infty$$

On a :

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

$F(x) = E(F_n(x))$

Posons :

$$A = \sum_{n=1}^{\infty} P(|F_n(x) - F(x)| > \varepsilon\sqrt{\frac{\log n}{n}})$$

$$\Rightarrow A = \sum_{i=1}^{n} P(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(x_i) - \frac{1}{n}E(\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(x_i))\right| > \varepsilon\sqrt{\frac{\log n}{n}})$$

$$\Rightarrow A = \sum_{i=1}^{n} P(\left|\mathbf{1}_{]-\infty,x]}(x_i) - E(\mathbf{1}_{]-\infty,x]}(x_i))\right| > \varepsilon\sqrt{n\log n})$$

En va utiliser l'inégalité de Bernshtein Frechet :

On sait que :

$$\alpha_i = 0 \leq \mathbf{1}_{]-\infty,x]}(x_i) \leq \ \beta_i = 1$$

Posons :

$$t = \varepsilon\sqrt{n \log n}$$

on trouve

$$\sum_{i=1}^{n} P(|Fn(x) - F(x)| > \varepsilon \left|\sqrt{\frac{\log n}{n}}\right|) \leq \sum_{i=1}^{n} 2.\exp(\frac{-2\varepsilon^2 n \log n}{n})$$

$$\begin{aligned}\sum_{i=1}^{n} P(|Fn(x) - F(x)| > \varepsilon \left|\sqrt{\frac{\log n}{n}}\right|) &\leq \sum_{i=1}^{n} 2.\exp(\log n^{-2\varepsilon^2}) \\ &\leq \sum_{i=1}^{n} 2.n^{-2\varepsilon^2} \\ &\leq \sum_{i=1}^{n} 2.\frac{1}{n^{2\varepsilon^2}} < \infty\end{aligned}$$

(Série de Rieman convèrgente pour $\varepsilon > \frac{1}{\sqrt{2}}$)

D'où

$$\sum_{n=1}^{\infty} P(|F_n(x) - F(x)| > \varepsilon \left|\sqrt{\frac{\log n}{n}}\right|) < \infty$$

Par suite :

$$F_n(x) - F(x) = O(\sqrt{\frac{\log n}{n}}) \text{ en } (p.co)$$

2.2 Estimation à noyaux pour la fonction de répartition :

Soit $(X_1, X_2 \ldots X_n)$ un n-échantillon de x de fonction de répartition F.

On appelle estimateur à noyau :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - Xi}{h_n}\right)$$

Ou H est une fonction de répartition quelquanque et h_n est une suite des nombres réels positifs.

Théorème 2.2.1 *$F_n(x)$ est un estimateur sans biais de F.*

Preuve. □

Il suffit de montrer que :

$$E(F_n(x)) - F(x) \to 0 \; quand \; n \to \infty$$

On a :

$$\begin{aligned} E(F_n(x)) &= E\left(\frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x-x_i}{h_n}\right)\right) \\ &= \int_{-\infty}^{+\infty} H\left(\frac{x-z}{h_n}\right) f(z)dz \\ &= \left[H\left(\frac{x-z}{h_n}\right).F(z)\right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} H'\left(\frac{x-z}{h_n}\right) F(z)dz \end{aligned}$$

On pose :

$$\frac{x-z}{h_n} = y$$

$$\Rightarrow z = x - y.h_n$$

$$\Rightarrow dz = -h_n dy$$

$$\Rightarrow \left[H(\frac{x-z}{h_n}).F(z)\right]_{-\infty}^{+\infty} = 0$$

D'aprés le developpement limité de $F(x - y.h_n)$ on a :

$$F(x - y.h_n) = F(x) - h_n.y.F'(x) + o({h_n}^2)$$

Donc ;

$$E(F_n(x)) = \int_{-\infty}^{+\infty} H'(y)F(x - y.h_n)dy$$

$$= \int_{-\infty}^{+\infty} H'(y)(F(x) - h_n.y.F\prime(x) + o({h_n}^2))dy$$

$$= F(x).\int_{-\infty}^{+\infty} H'(y)dy - h_n.y.f(x)\int_{-\infty}^{+\infty} yH'(y)dy + o({h_n}^2)$$

$$= F(x) + o(h_n) + o({h_n}^2)$$

Donc ;

$$E(F_n(x)) - F(x) = 0(1) \to 0$$

2.3 Estimation non paramétrique de la densité :

Définition 2.3.1 *On a pour h assez petit :*

$$f(x) = \frac{dF(x)}{dx} \approx \frac{F(x+h) \ - \ F(x-h)}{2h}$$

La fonction de répartition est inconnue, mais on peut la remplacer par son estimateur.

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

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h}\mathbf{1}_{]x-h,x+h]}(x_i)$$

$$= \frac{1}{nh}\sum K^o\left(\frac{x-x_i}{h}\right)$$

où

$$K^o\left(\frac{x-x_i}{h}\right) = \frac{1}{2}\mathbf{1}_{]-1,1]}\left(\frac{x-x_i}{h}\right)$$

Généralisation :

On remplace K^o par un noyau plus général,

Soit $K : \mathbb{R} \to \mathbb{R}$ intégrable telle que

$$\int_{\mathbb{R}} K(t)dt = 1$$

On peut définir :

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right)$$

comme un estimateur de f.

Proposition 2.3.1 *Soit la densité f,son estimateur f_n vérifie :*

$$E(f_n(x)) - f(x) = O(h_n^k), k \in \mathbb{Z}$$

Preuve.

$$\begin{aligned} E(f_n(x)) &= \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{x - z}{h_n}\right) f(z)dz \\ &= \int_{-\infty}^{+\infty} K(y).f(x - h_n y)dy \end{aligned}$$

□

On sait que :

$$f(x - h_n y) = f(x) - h_n y f'(x) + \cdots + \frac{(-1)^k h_n^k f^{(k)}(x)}{k!} + o(h_n^k)$$

Donc ;

$$\begin{aligned} E(f_n(x)) &= \int_{-\infty}^{+\infty} K(y)\left(f(x) - h_n y f'(x) + \cdots + \frac{(-1)^k h_n^k f^{(k)}(x)}{k!} + o(h_n^k)\right) dy \\ &= f(x)\int_{-\infty}^{+\infty} K(y)dy + \int_{-\infty}^{+\infty} K(y)\left(h_n y f'(x) + .. + \frac{(-1)^k h_n^k f^{(k)}(x)}{k!} + o(h_n^k)\right) dy \end{aligned}$$

Puisque ;

$$\int_{-\infty}^{+\infty} K(y)dy = 1$$

alors :

$$E(f_n(x)) - f(x) = -\int_{-\infty}^{+\infty} K(y)[h_n y f'(x) + .. + (-1)^k h_n^k f^{(k)}(x) + o(h^k)]dy$$

On Remarque que :

$$\left|\frac{E(f_n(x)) - f(x)}{h_n^k}\right| \leq c$$

c est une constante.

D'ou ;

$$E(f_n(x)) - f(x) = O(h_n^k), k \in \mathbb{Z}$$

Chapitre 3

L'estimation non paramétrique de la régression(Cas reèl) :

Dans ce chapitre on essayera à estimer la fonction de la régression notée $r(x)$ où x est un réel.

3.1 Le model de la régression :

On apelle model fonctionel de régression tout model s'écrivant sous la forme :

$$Y_i = r(X_i) + \varepsilon_i$$

Où ;

Y_i : la variable aléatoire à expliquer.

X_i : la variable aléatoire explicative.

ε_i : variable aléatoire centrée indépendante de X.

Dans la suite nous nous limiterons à une variable aléatoire Y,

X peut prendre différents natures.

On a :

$$Y = r(X) + \varepsilon$$

Si $X = x$,

$$Y \backslash X = x = r(x)$$

$$\Rightarrow E(Y \backslash X = x) = r(x)$$

Calculons $E(Y\backslash X = x)$:

$$\begin{aligned} E(Y\backslash X = x) &= \int_{-\infty}^{+\infty} y.f_{Y\backslash X=x}(x,y)dy \\ &= \frac{\int_{-\infty}^{+\infty} y.f_{Y,X}\ dy}{f_X(x)} \\ &= \frac{\int_{-\infty}^{+\infty} y.f_{Y,X}\ dy}{\int_{-\infty}^{+\infty} f_{Y,X}dy} \\ &\simeq \frac{\sum y_i n_i}{\sum n_i} \end{aligned}$$

Définition 3.1.1 *On appelle estimateur à noyau pour la fonction $r(x)$ l'estimateur définit par :*

$$r_n(x) = \frac{\sum_{i=1}^{n} Y_i\_K\left(\frac{x-xi}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-xi}{h}\right)} \simeq \frac{\sum y_i n_i}{\sum n_i}$$

où

$$n_i = K\left(\frac{x-xi}{h}\right)$$

On peut écrire :

$$r_n(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{x-xi}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-xi}{h}\right)} = \frac{g_n(x)}{f_n(x)}$$

Hypothèses :

Le modèl r est renforcé par les conditions :

1-r et f sont k fois continument dérivable autour de x.

2-

$$f(x) > 0$$

f est la densité de x.

3-

$$\lim_{n\to\infty} h = 0$$

et

$$\lim_{n\to\infty} \frac{nh}{\log n} = \infty$$

Ou h est le paramètre de lissage.

4- K est borné et intégrable et à support compact.

5- K vérifie :

$$\int t^j K(t)dt = 0 \quad \forall j = 1, \ldots, k-1$$

et

$$0 < |\int t^k K(t)dt| < +\infty$$

6-

$$|Y| < M < \infty$$

Théorème 3.1.1 *Sous les conditions précédentes on a :*

$$r_n(x) - r(x) = O(h^k) + O\left(\sqrt{\frac{\log n}{nh}}\right) \quad \textit{en (p.co)}$$

Preuve. [Preuve du théorème]

Dans cette preuve c désigne une constante, et $\int_{-\infty}^{+\infty} K(y)dy = 1$ et $rf = g$.

$$r_n(x) = \frac{g_n(x)}{f_n(x)}$$

Et

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

Et

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

On a :

$$r_n(x) - r(x) = \left(\frac{g_n(x)}{f(x)} - \frac{g(x)}{f(x)}\right)$$

$$= \frac{g_n(x) - g(x)}{f_n(x)} + (f(x) - f_n(x))\frac{r(x)}{f_n(x)}$$

$r_n(x) - r(x) \to 0$

Ceci implique que :

$$(1)\dots\ g_n(x) - g(x) \to 0 \ \ en\ (p.co)$$

$$(2)\dots\ f(x) - f_n(x) \to 0 en(p.co)$$

$$(3)\dots f_n(x) \nrightarrow 0 \text{ en (p.co)}$$

Preuve. [Preuve de (1)]

$$g_n(x) - g(x) = g_n(x) - E(g_n(x)) + E(g_n(x)) - g(x)$$

On a deux parties :

$$E(g_n(x)) - g(x)...a$$

c' est la partie de biais.

$$E(g_n(x)) - g_n(x)...b$$

c'est la partie dispèrssion .

Pour a on a :

$$E(g_n(x)) = \frac{1}{h}E(YK\left(\frac{x-X}{h}\right))$$

En conditionant par rapport à X on arrive à :

$$E(g_n(x)) \ = \frac{1}{h}\int_{-\infty}^{+\infty} r(u)K\left(\frac{x-u}{h}\right)f(u)du, z = \frac{x-u}{h}$$

$$= \int_{-\infty}^{+\infty} g(x-zh)K\left(z\right)dz$$

$$g(x-zh) \ = g(x) - hzg'(x) + \dots + \frac{(-1)^k z^k h^k g^{(k)}(\theta_z)}{k!} + o(h^k)\ \theta_z \in [x, x+zh]$$ □

On utilise l'hypothèse (5), on trouve :

$$E(g_n(x)) = g(x) + (-1)^k h^k \frac{\int_{-\infty}^{+\infty} z^k K\left(z\right) g^{(k)}(\theta_z)\ dz}{k!}$$

On arrive finalement à :

$$E(g_n(x)) = g(x) + (-1)^k h^k \int_{-\infty}^{+\infty} z^k K\left(z\right) dz \frac{g^{(k)}(x)}{k!} + o(h^k)$$

D'où,

$$E(g_n(x)) - g(x) = O(h^k) \ \ en \ \ (p.co)$$

Pour (b) on a besoin du lemme suivant :

Lemme 3.1.1 *Soit $\Delta_1, \Delta_2, ...\Delta_n$ des variable aléatoires centrées, indépendantes et de mème loitelles qu'il existe deux réels positifs d et δ^2 vérifiant :*

$$|\Delta_1| \leq d$$

et

$$\Delta_1^2 \leq \delta^2$$

Pour tout $\varepsilon \in \left]0, \frac{\delta^2}{d}\right[$ on a :

$$P\left[\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_i\right| > \varepsilon\right] \leq 2\exp\left(-\frac{n\varepsilon^2}{4\delta^2}\right)$$

pour appliquer ce lemme aux variables :

$$\Delta_i = \frac{1}{h}\left(Y_i K\left(\frac{x-Xi}{h}\right) - EY_i K\left(\frac{x-Xi}{h}\right)\right)$$

$$|\Delta_i| \leq \frac{c}{h}$$

car K et Y sont bornés

$$Var(\Delta_i) = E(\Delta_i^2) - (E(\Delta_i))^2$$

et

$$E(\Delta_i) = 0$$

on a

$$E(\Delta_i^2) \leq E(\mu_i^2)$$

avec :

$$\mu_i = \frac{1}{h}K\left(\frac{x-Xi}{h}\right)Y_i$$

En conditionant par rapport à X on arrive à :

$$E(\mu_i^2) = \frac{1}{h}E\left(\frac{1}{h}K^2\left(\frac{x-Xi}{h}\right)Y^2\right)$$

$$= \tfrac{1}{h^2}\int_{-\infty}^{+\infty}\phi(u)f(u)K^2\left(\tfrac{x-u}{h}\right)du$$

où

$$\phi(u) = E(Y^2 \backslash X = u)$$

Posons le changement de variable : $z = \frac{x-u}{h}$

$$E(\mu_i^2) = \frac{1}{h}\int_{-\infty}^{+\infty} \phi(x - zh) f(x - zh) K^2(z)\, du$$

Comme ϕ est bornée, et f l'est aussi car continue sur K il existe une constante c telle que

$$E(\mu_i^2) \leq \frac{c}{h}$$

et de manière évidente on a l'existance d'une constante c telle que :

$$E(\Delta_i^2) \leq \frac{c}{h}$$

on applique maintenant le lemme :

$$P\left[|E(g_n(x)) - g_n(x)| > \varepsilon\right] \leq 2\exp\left(-\frac{n\varepsilon^2 h}{4c}\right)$$

si on pose :

$$\varepsilon = \varepsilon_0\sqrt{\frac{\log n}{nh}}$$

On arrive finallement à :

$$P\left[|E(g_n(x)) - g_n(x)| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right] \leq 2\exp\left(-c\varepsilon_0^2 \log n\right)$$

$$\leq 2n^{-c\varepsilon_0^2}$$

Il suffit de prendre $\varepsilon_0 = \sqrt{c}$ pour obtenir une série de Rieman convèrgente.

D'ou

$$E(g_n(x)) - g_n(x) = O\left(\sqrt{\frac{\log n}{nh}}\right) \quad en\ (p.co)$$

Preuve. [Preuve de (2)]

$$f_n(x) - f(x) = f_n(x) - E(f_n(x)) + E(f_n(x)) - f(x)$$

On a deux parties :

$$E(f_n(x)) - f(x)...c$$

c' est la partie de biais.

$$E(f_n(x)) - f_n(x)...d$$

c'est la partie dispèrtion .

Pour c on suit les mèmes démarches que a en posant $Y = 1$.Finallement on obtient :

$$E(f_n(x)) - f_n(x) = O(h^k), k \in \mathbb{Z} \text{ en (p.co)}$$

Pour (d) il suffit de reprendre les calculs précédents de (b) mais dans le cas particulier ou la variable $Y = 1$.on arrive à :

$$P\left[|E(f_n(x)) - f_n(x)| > \varepsilon_0\sqrt{\frac{\log n}{nh}}\right] \leq 2\exp\left(-c\varepsilon_0^2 \log n\right)$$

Enfin ;

$$E(f_n(x)) - f_n(x) = O\left(\sqrt{\frac{\log n}{nh}}\right)$$

□

Preuve. [Preuve de (3)] $(3) \Leftrightarrow \exists\varepsilon > 0, \sum_{n=1}^{\infty} P(f_n(x) \leq \varepsilon) < \infty$si (3) est vérifié alors :

$$\exists\varepsilon > 0,\ P(f_n(x) \leq \varepsilon) \to 0$$

$$\Leftrightarrow \exists\varepsilon > 0,\ P(f_n(x)) > \varepsilon) \to 1 \Leftrightarrow \exists\varepsilon > 0,\ f_n(x) > \varepsilon$$

$i.e : f_n(x) \nrightarrow 0$ en (p.co) □

si

$$|f_n(x)| \leq \frac{f(x)}{2}$$

alors

$$f_n(x) - f(x) > \frac{f(x)}{2}$$

$$\sum_{n=1}^{\infty} P\left(\left|f_n(x) \leq \frac{f(x)}{2}\right|\right) < \sum_{n=1}^{\infty} P\left(\left|f_n(x) - f(x) \leq \frac{f(x)}{2}\right|\right) < \infty$$

car $f_n \to f$ en (p.co)

Donc,

$$P\left(|f_n(x) >| \frac{f(x)}{2}\right) = 1$$

alors

$$\exists\varepsilon = \frac{f(x)}{2} > 0 \text{ pour que } f_n(x) \nrightarrow 0 \text{ en (p.co)}$$

□

Des résultat précédentes on arrive à :

$$r_n(x) - r(x) = O(h^k) + O\left(\sqrt{\frac{\log n}{nh}}\right) \text{ en (p.co)}$$

Chapitre 4

L'estimation non paramétrique de larégression(Cas vectoriel) :

Soit le model de régression fonctionel suivant :

$$Y_i = r(X_i) + \varepsilon_i$$

Où ;

Y_i : la variable aléatoire à expliquer (réel).

X_i : la variable aléatoire explicative à valeur sur $\mathbb{R}^p$

p un entier strictement positif.

ε_i : variable aléatoire centrée indépendante de X.

L'estimateur à noyau pour la fonction r est :

$$r_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-Xi}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-Xi}{h}\right)} = \frac{g_n(x)}{f_n(x)}, x \in \mathbb{R}^p$$

$$g_n(x) = \frac{1}{nh^p} \sum_{i=1}^{n} Y_i K\left(\frac{x-Xi}{h}\right)$$

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

Hypothèses :

Le modèl r est renforcé par les conditions :

1-r et f sont k fois continument dérivable autour de x.

x étant un point fixé de $\mathbb{R}^p$.

2-

$$f(x) > 0$$

f est la densité de x.

3-

$$\lim_{n\to\infty} h = 0$$

et

$$\lim_{n\to\infty} \frac{nh^p}{\log n} = \infty$$

Ou h est le paramètre de lissage.

4- K est borné et intégrable et à support compact, pour tout p-uplet d'entier positifs $(i_1, i_2, ..., i_p)$

5- K vérifie :

$$T(i_1, i_2, ..., i_p) = \int_{\mathbb{R}^p} u_1^{i_1} ... u_p^{i_p} K(u_1 ... u_p) du_1 ... du_p$$

$\forall j \in \{1, \ldots, p\}$ on a :

$$T_k(j) = \int_{\mathbb{R}^p} u_j^k K(u_1 ... u_p) du_1 ... du_p$$

et

$$0 < |\int t^k K(t) dt| < +\infty$$

Nous dirons q'une fonction K de $\mathbb{R}^p$ dans $\mathbb{R}$ est un noyau d'ordre $k, k \in \mathbb{N}^*$, lorsque :

$$\forall (i_1, i_2, ..., i_p) \in \mathbb{N}^{*p}, (\forall j, i_j < k) \Rightarrow (T_k(i_1, i_2, ..., i_p)) = 0$$

+

$$\forall (i_1, i_2, ..., i_p) \in \mathbb{N}^{*p}, \qquad \forall j, T_k(j) \in \mathbb{R}^*$$

6-

$$|Y| < M < \infty$$

Théorème 4.0.2 *Sous les hypothèses précédentes on arrive à :*
$r_n(x) - r(x) = O(h^k) + O\left(\sqrt{\frac{\log n}{nh^p}}\right)$ *en (p.co)*

Preuve. *[Preuve de théorème] La preuve est trés similaire du celle du théorème (3.1.1) en nous contenant d'insister sur les parties de la démonstration pour lesquelles l'aspect multi-dimentionnel. Dans cette preuve c désigne une constante, et* $\int_{\mathbb{R}^p} K(u)du = 1$ *et* $rf = g$*.Le résultat final sera prouvé dés que seront vérifiées les 5 prapriétés suivant :*

$$E(g_n(x)) - g(x) = O(h^k) \quad en\ (p.co) \quad (e)$$

$$E(g_n(x)) - g(x) = O(h^k) \quad en\ (p.co) \quad (f)$$

$$E(g_n(x)) - g_n(x) = O\left(\sqrt{\frac{\log n}{nh^p}}\right) \quad en\ (p.co) \quad (g)$$

$$E(f_n(x)) - f_n(x) = O\left(\sqrt{\frac{\log n}{nh^p}}\right) \quad en\ (p.co) \quad (h)$$

$$\exists \varepsilon > 0, \sum_{n=1}^{\infty} P(f_n(x) \leq \varepsilon) < \infty \quad en\ (p.co) \quad (i)$$

$$r_n(x) = \frac{g_n(x)}{f_n(x)}$$

Et

$$g_n(x) = \frac{1}{nh^p} \sum_{i=1}^{n} Y_i K\left(\frac{x - xi}{h}\right)$$

Et

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

On a :

$$E(g_n(x)) = \frac{1}{h^p} E(Y K\left(\frac{x - X}{h}\right))$$

En conditionant par rapport à X *on arrive à :*

$$E(g_n(x)) = \frac{1}{h^p} \int_{-\infty}^{+\infty} r(u) K\left(\frac{x-u}{h}\right) f(u)du, z = \frac{x-u}{h}$$

$$E(g_n(x)) = \int_{-\infty}^{+\infty} g(x - zh) K\left(z\right) dz$$

$$g(x - zh) - g(x) = \sum_{j=1}^{k} \frac{(-h)^j}{j!} \sum_{i_1,+...+i_p} \left(T_k(i_1, ..., i_p) \left[\frac{\partial^j g}{\partial x_1^{i_1} ... \partial x_p^{i_p}}\right](z) + o(h^j)\right)$$

On utilise la condition (5), on trouve :

$$E(g_n(x)) - g(x) = \frac{(-h)^k}{k!} \sum_{j=1}^{k} \left[\frac{\partial^k g}{\partial x_j^k}\right](x) T_k(j) + o(h^k)$$

ceci achève la preuve.

$$E(g_n(x)) - g(x) = O(h^k)$$

Pour montrer "f" on suit les mèmes étapes que "e" on arrive à :

$$E(f_n(x)) - f(x) = \frac{(-h)^k}{k!} \sum_{j=1}^{p} \left[\frac{\partial^k g}{\partial x_j^k} \right] (x) T_k(j) + o(h^k)$$

ceci achève la preuve.

$$E(f_n(x)) - f(x) = O(h^k)$$

Pour "f" on a besoin du lemme (3.1.1) :

si on applique ce lemme aux variables :

$$\Delta_i = \frac{1}{h^p} \left(Y_i K \left(\frac{x - Xi}{h} \right) - E Y_i K \left(\frac{x - Xi}{h} \right) \right)$$

$$|\Delta_i| \leq \frac{c}{h^p}$$

car K et Y sont bornés $Var(\Delta_i) = E(\Delta_i^2) - (E(\Delta_i))^2 et\ E(\Delta_i) = 0$

on a

$$E(\mu_i^2) \leq \frac{c}{h^p}$$

on applique maintenant le lemme :

$$P\left[|E(g_n(x)) - g_n(x)| > \varepsilon \right] \leq 2 \exp\left(-\frac{n\varepsilon^2 h^p}{4c} \right)$$

si on pose :

$$\varepsilon = \varepsilon_0 \sqrt{\frac{\log n}{n h^p}}$$

on arrive finallement à :

$$P\left[|E(g_n(x)) - g_n(x)| > \varepsilon_0 \sqrt{\frac{\log n}{n h^P}} \right] \leq 2 n^{-c\varepsilon_0^2}$$

on peut choisir $\varepsilon_0 > \frac{1}{\sqrt{c}}$

D'où le résultat.

Pour (h) il suffit de reprendre les calcul précédents de (g)

$$P\left[E\left| g_n(x) - g_n(x) \right| > \varepsilon_0 \sqrt{\frac{\log n}{n h^p}} \right] \leq 2 n^{-c\varepsilon_0^2}$$

Pour (i) il suffit de poser $\varepsilon = \frac{f(x)}{2}$ *pour que*

$$f_n \neq 0 \text{ } en \text{ } (p.co)$$

□

Des résultats précédents on arrive à

$$r_n(x) - r(x) = O(h^k) + O\left(\sqrt{\frac{\log n}{nh^p}}\right) \quad \text{en (p.co)}$$

CONCLUSION

D'aprés notre mémoire on a conclut que la fonction de répartition est la base de l'estimation fonctionnelle

On a montrer l'estimation de la régression en un point comme étant un rapport de densités f et g.

Bibliographie

[1] Frédéric Ferraty et Philippe Vieu, Cours de DEA, modèles statistique fonctionnel. Modèles de régression pour variables aléatoires uni, multi et $\infty-$dimentionnées.

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

[3] Les cours de statistique 3 (Master 2), pour l'année univèrcitaire 2014-2015.