République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Départementde Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option : **Statistique**

Par TELLI MOUFIDA

Titre :

Sur l'estimation non paramétrique

Devant le Jury :

Mr.	YAHIA DJABRANE	Pr	U. Biskra	Président
Mr.	NECIR ABDELHAKIM	Pr	U. Biskra	Encadreur
Mme.	ZOUAOUI NOUR EL HOUDA	Dr	U. Biskra	Examinatrice

Soutenu Publiquement le 27/06/2022

Dédicace

Je dédie ce modeste travail

A ma chère maman ***Kelthoum****, ma vie et mon bonheur pour son amour, ses encouragements, et pour tous ses sacrifices afin que je cibler ce travail.*

A l'homme de ma vie qui a souffert sans me laisser souffrir, qui était toujours à mes côtés, mon cher papa **Abdelkarim.**

A mes chers frères ***Abdelkamel****,* ***Nadhir****,* ***Ismail****,* ***Aissa****,* ***Moussa*** *pour leurs encouragements, soutiens et pour l'amour qu'ils m'ont donné.*

A ma chère sœur ***Hassina****, les mots ne peuvent pas exprimer mon amour et ma gratitude de t'avoir comme sœur, merci d'être la personne qui m'encourage et me donne la confiance et l'effort pour terminer mes études et devient la fille de mes rêves.*

A mes amies : **Wissal**, **Loubna**, **Lamia** et **Imene** qui m'ont toujours encouragé et qui m'ont donné le soutien moral pour finir ce travail.

Remerciements

J'ai pu accomplir cet humble travail grâce à Dieu qui m'a la donné force, la volonté et le courage.

Tout d'abord, je tiens à remercie le Professeur ***Necir Abdelhakim*** *pour tous ses efforts, son aide indéfectible, son écoute et ses conseils qui m'ont permis de cibler mes candidatures.*

J'adresse aussi mes remerciements les plus chaleureux au Professeur ***Djarbane Yahia*** *pour leurs conseils précieux, sa disponibilité, et son sens d'écoute et d'échange.*

Je remercie également tous les membres du comité de lecture pour avoir accepté l'évaluation et la discussion de ce mémoire.

De plus, je tiens aussi à remercier vivement tous mes enseignants qui m'ont formé et accompagné tout au long de mon parcours avec beaucoup de patience et de pédagogie. Grâce à leurs compétences et leurs savoirs, j'ai pu accomplir totalement dans mes années d'études.

Finalement, j'adresse mes sincères remerciements à tous mes proches qui m'ont aidé à finir ce travail.

Notations et symboles

Les abréviations et les symboles dierent utilisés tout au long de ce mémoire sont expliqués ci-dessous

$\mathcal{N}(0, \widehat{\sigma}^2)$	:	La loi normale centré à variance $\widehat{\sigma}^2$
$\xrightarrow{P}$	:	Converge en probabilité
$\xrightarrow{\mathcal{D}}$	:	Converge en distribution
ND	:	Nadaraya Watson
MCO	:	Moindre carré ordinaire
LL	:	Linéaire locale
AD	:	Ajustement des lois
$F_n(\cdot)$	:	Fonction distribution empirique
CV	:	Convergence
IC	:	Intervalle de canfiance
$E[X]$	:	Moyenne de X
$Var[X]$	:	Variance de X

TCL	:	Théorème centrale limite
iid	:	identiquement indépendant distribuées
$\mathbb{R}$	:	Ensemble de nombres réels
$\mathbb{N}$	:	Ensemble d'entiers non négatifs
$\exists$	:	Existe
$(X_1, ..., X_n)$	:	Échantillons de taille n de X

Table des matières

Table des figures

Liste des tableaux

Introduction

La méthode d'estimation non paramétrique la plus simple est celle de l'histogramme qui était introduit par John Graunt en 1962. Comme l'histogramme donne une fonction qui n'est pas continue, Rosenblatt en 1956[10], suivi de Parzen en 1962[8], ont proposé une classe d'estimateurs à noyau d'une densité univariée. Les estimateurs à noyau sont fonction de deux paramètres K, appelé noyau, et h dit paramètre de lissage (largeur de fenêtre). Rosenblatt reprenait l'idée de Fix et Hodges en 1951[5], qui consistait à estimer la densité en un point, en comptant le nombre d'observations situées dans l'intervalle de longueur $2h$ et centrées en ce point. Les propriétés de convergence de l'estimateur à noyau ont été établies par Parzen, Silverman et Nadaraya. Devroye en 1985[2] a fait une étude complète sur la convergence L_1. Les théorèmes relatifs à l'erreur quadratique asymptotique et l'erreur quadratique intégrée asymptotique ont été obtenus sous forme élémentaire par Parzen. Epanechnikov en 1969[4] a montré l'existence d'un noyau asymptotiquement optimal. L'erreur quadratique moyenne asymptotiquement intégrée varie peu en fonction du choix du noyau K.

L'objet du mémoire est de faire une synthèse sur l'estimation à noyau de la densité de probabilité et de la régression. Notre mémoire est composé de trois chapitres :

Le premier chapitre est réservé aux quelques notations et définitions dont nous présentons l'estimateur de Parzen-Rosenblatt. Le deuxième chapitre fait l'objet de l'estimation

à noyau de la régression (estimateur de Nadaraya-Watson). Dans le troisième chapitre nous représentons une étude de simulations à l'aide du langage R.

Chapitre 1

Estimateur de Parzen-Rosenblatt

1.1 Estimation à noyau de la densité de probabilité

1.1.1 Estimateur discret

Soit X est variable aléatoire de distribution continue $F(x)$ est de densité $f(x) = \frac{d}{dx}F(x)$. Le but est d'estimer $f(x)$ à partir d'un échantillon aléatoire $X_1, ..., X_n$.Cette fonction de distribution F est naturellement estimée par la fonction de distribution empirique (FDE) $F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$.

L'estimateur naturel la densité $f(x)$ est la dérivée de $F_n(x)$, notée $\frac{d}{dx}F_n(x)$. Mais ce dernier n'est pas continue donc n'est pas dérivable, en d'autres termes cette dérivation n'a pas un sens, nous devons alors chercher un autre moyen pour estimer f. Au lieu de la dérivée on a utilisé les accroissements de F_n sur des petits intervalles de longueur $2h$. L'estimateur de f obtenu, dit l'histogramme, définit pour $h > 0$, par

$$\widehat{f}(x) := \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

On remarque que

$$\frac{1}{2nh}\left(\sum_{i=1}^{n}\mathbf{1}\left\{X_i\leq x+h\right\}+\sum_{i=1}^{n}\mathbf{1}\left\{X_i\leq x-h\right\}\right)$$
$$=\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\mathbf{1}_{\{|X_i-x|/h\leq 1\}}=\frac{1}{nh}\sum_{i=1}^{n}k_0\left(\frac{X_i-x}{h}\right),$$

où $k_0(x)=2^{-1}\mathbf{1}_{\{|x|\leq 1\}}$, ainsi on a réécrit $\widehat{f}$ comme étant un estimateur à noyau. Il bien connu que l'estimateur $\widehat{f}$, en tant qu'histogramme, présente des sauts de discontinuités. Afin d'obtenir un estimateur lisse de f, il suffit alors de choisir une fonction noyau continue au lieu de k_0. Ainsi on définit de façon général, l'estimateur à noyau de f par

$$\widehat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{X_i-x}{h}\right).$$

1.1.2 Fonctions noyau

Une fonction noyau $k:\mathbb{R}\rightarrow\mathbb{R}$ est une fonction qui satisfait $\int_{-\infty}^{\infty}k(u)du=1$.

Un noyau est positif ou nul ; c'est à dire $k(x)\geq 0$ pour tout x, en d'autres termes k est une fonction de densité de probabilité.

On définit les moments d'un noyau k par $\mu_j(k):=\int_{-\infty}^{\infty}x^jk(x)dx$, $j=1,2,...$

Une fonction noyau est symétrique : $k(x)=k(-x)$ pour tout x. Dans ce cas, tous les moments impairs sont nuls :

$$\int x^m k(x)\,dx=0,\ \text{si } m \text{ impaire.}$$

Les noyaux usuels sont répertoriés dans le tableau suivant :

Noyaux	Équations
Uniform	$k_0(x) = \frac{1}{2}\mathbf{1}(\lvert x\rvert \leq 1)$
Epanechnikov	$k_1(x) = \frac{3}{4}(1-x^2)\mathbf{1}(\lvert x\rvert \leq 1)$
Biweight	$k_2(x) = \frac{15}{16}(1-x^2)^2\mathbf{1}(\lvert x\rvert \leq 1)$
Triweight	$k_3(x) = \frac{35}{32}(1-x^2)^3\mathbf{1}(\lvert x\rvert \leq 1)$
Gaussien	$k_\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$

TAB. 1.1 – Noyaux usuelles

1.1.3 Estimateur de densité

L'estimateur $\widehat{f}$ est bien une densité. En effet, par le changement de variables

$u = (X_i - x)/h$, on écrit

$$\frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} k\left(\frac{X_i - x}{h}\right)dx = \int_{-\infty}^{\infty} k(u)du = 1.$$

Le premier moment associé à $\widehat{f}$ est :

$$\begin{aligned}\int_{-\infty}^{\infty} x\widehat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x\frac{1}{h}k\left(\frac{X_i - x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(X_i + uh)k(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\int_{-\infty}^{\infty} k(u)du + \frac{1}{n}\sum_{i=1}^{n} h\int_{-\infty}^{\infty} uk(u)du = \frac{1}{n}\sum_{i=1}^{n} X_i.\end{aligned}$$

Le deuxième moment associé à $\widehat{f}$ est ;

$$\begin{aligned}\int_{-\infty}^{\infty} x^2\widehat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2\frac{1}{h}k\left(\frac{X_i - x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(X_i + uh)^2k(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{i=1}^{n} X_ih\int_{-\infty}^{\infty} uk(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2\int_{-\infty}^{\infty} u^2k(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\mu_2(k).\end{aligned}$$

La variance associée à la densité $\widehat{f}$ est :

$$\int_{-\infty}^{\infty} x^2 \widehat{f}(x)dx - \left(\int_{-\infty}^{\infty} \widehat{f}(x)dx\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\mu_2(k) - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$
$$= \widehat{\sigma}^2 + h^2\mu_2(k).$$

1.1.4 Biais d'estimation

Les prédictions des transformations du noyau peuvent être écrites d'intégrales forme d'une convolution de noyau et d'une fonction de densité

$$\mathbf{E}\left[\frac{1}{h}k\left(\frac{X_i - x}{h}\right)\right] = \int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{z-x}{h}\right) f(z)dz.$$

Changement de variables $u = (z - x)/h$,

$$\int_{-\infty}^{\infty} k(u)f(x + hu)du,$$

alors la linéarité de estimateur

$$\mathbf{E}\left[\widehat{f}(x)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[\frac{1}{h}k\left(\frac{X_i - x}{h}\right)\right] = \int_{-\infty}^{\infty} k(u)f(x + hu)du.$$

En effectuant un développement limité d'ordre 2 à f, il vient,

$$f(x + hu) = f(x) + f^{(1)}(x)hu + \frac{1}{2}f^{(2)}(x)h^2u^2 + \frac{1}{3!}f^{(3)}(x)h^3u^3$$
$$+ ... + \frac{1}{v!}f^{(v)}(x)h^vu^v + o(h^v).$$

Le reste est d'ordre plus petite que h^v aussi $h \to \infty$, qui s'écrit $o(h^v)$.

En intégrant terme par terme, en utilisant $\int_{-\infty}^{\infty} k(u)du = 1$

et la définition $\int_{-\infty}^{\infty} k(u)u^j du = \mu_j(k)$, on obtient

$$\begin{aligned}
&\int_{-\infty}^{\infty} k(u)f(x+hu)du \\
&= f(x) + f^{(1)}(x)h\mu_1(k) + \frac{1}{2}f^{(2)}(x)h^2\mu_2(k) + \frac{1}{3!}f^{(3)}(x)h^3\mu_3(k) \\
&\qquad\qquad + ... + \frac{1}{v!}f^{(v)}(x)h^v\mu_v(k) + o(h^v) \\
&= f(x) + \frac{1}{2}f^{(2)}(x)h^2\mu_2(k) + ... + \frac{1}{v!}f^{(v)}(x)h^v\mu_v(k) + o(h^v).
\end{aligned}$$

La moyenne est

$$\begin{aligned}
\mathbf{E}\left[\widehat{f}(x)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\frac{1}{h}k\left(\frac{X_i - x}{h}\right)\right] \\
&= f(x) + \frac{1}{v!}f^{(v)}(x)h^v\mu_v(k) + o(h^v).
\end{aligned}$$

Le biais de $\widehat{f}(x)$ est

$$\begin{aligned}
\mathbf{Bias}\left[\widehat{f}(x)\right] &= \mathbf{E}\left[\widehat{f}(x)\right] - f(x) \\
&= \frac{1}{v!}f^{(v)}(x)h^v\mu_v(k) + o(h^v).
\end{aligned}$$

Pour le noyau d'ordre 2, simplifié on a

$$\mathbf{Bias}\left[\widehat{f}(x)\right] = \frac{1}{2}f^{(2)}(x)h^2\mu_2(k) + O(h^4).$$

1.1.5 Variance d'estimation

L'estimateur à noyaux est un estimateur linéaire, et les $k\left(X_i - x/h\right)$ sont iid, alors

$$\begin{aligned}\mathbf{var}\left[\widehat{f}(x)\right] &= \frac{1}{nh^2}\mathbf{var}\left[k\left(\frac{X_i - x}{h}\right)\right] \\ &= \frac{1}{nh^2}\mathbf{E}\left[k\left(\frac{X_i - x}{h}\right)^2\right] - \frac{1}{n}\left(\frac{1}{h}\mathbf{E}\left[k\left(\frac{X_i - x}{h}\right)\right]\right)^2.\end{aligned}$$

De notre analyse de biais, nous avons $\frac{1}{h}\mathbf{E}\left[k\left(\frac{X_i - x}{h}\right)\right] = f(x) + o(1)$. Un développement de Taylor donne

$$\begin{aligned}\frac{1}{h}\mathbf{E}\left[k\left(\frac{X_i - x}{h}\right)^2\right] &= \frac{1}{h}\int_{-\infty}^{\infty} k\left(\frac{z - x}{h}\right)^2 f(z)dz \\ &= \int_{-\infty}^{\infty} k(u)^2 f(x + hu)du \\ &= \int_{-\infty}^{\infty} k(u)^2 (f(x) + O(h))du \\ &= f(x)R(k) + O(h).\end{aligned}$$

Lorsque $R(k) = \int_{-\infty}^{\infty} k(u)^2 du$ est la rugosité de noyaux,

$$\mathbf{var}\left[\widehat{f}(x)\right] = \frac{f(x)R(k)}{nh} + O(\frac{1}{n}).$$

1.1.6 Erreur quadratique moyenne

Une des mesures courantes et pratiques de la précision de l'estimation est l'erreur quadratique moyenne donnée par

$$\begin{aligned}\mathbf{MSE}\left[\widehat{f}(x)\right] &= \mathbf{E}\left[\widehat{f}(x) - f(x)\right]^2 \\ &= \mathbf{Bias}\left[\widehat{f}(x)\right]^2 + \mathbf{var}\left[\widehat{f}(x)\right] \\ &\backsimeq \left(\frac{1}{v!}f^{(v)}(x)h^v\mu_v(k)\right)^2 + \frac{f(x)R(k)}{nh} \\ &= \mathbf{AMSE}\left[\widehat{f}(x)\right],\end{aligned}$$

sous les conditions $h \longrightarrow 0$ et $nh \longrightarrow \infty$ quand $n \longrightarrow \infty$.

Une mesure globale de la précision est l'erreur carrée intégrée d'asymptotique moyenne :

$$\begin{aligned}\mathbf{AMISE}\left[\widehat{f}(x)\right] &= \int_{-\infty}^{\infty} \mathbf{AMSE}\left[\widehat{f}(x)\right] dx \\ &= \frac{\mu_v^2(k)}{(v!)^2}R(f^{(v)})h^{2v} + \frac{R(k)}{nh},\end{aligned}$$

telle que $R(f^{(v)}) = \int_{-\infty}^{\infty}(f^{(v)}(x))^2 dx$.

1.1.7 La fenêtre asymptotiquement optimale

On cherche la solution en prenant la dérivée de l'AMISE par rapport à h et en la fixant égale à zéro :

$$\begin{aligned}\frac{\partial \mathbf{AMISE}}{\partial h} &= \frac{\partial}{\partial h}\left(\frac{\mu_v^2(k)}{(v!)^2}R(f^{(v)})h^{2v} + \frac{R(k)}{nh}\right) \\ &= 2vh^{2v-1}\frac{\mu_v^2(k)}{(v!)^2}R\left(f^{(v)}\right) - \frac{R(k)}{nh^2} = 0.\end{aligned}$$

La solution de cette équation est $h_0 := C_v(k,f)n^{-1/(2v+1)}$, où

$$C_v(k,f) := R\left(f^{(v)}\right)^{-1/(2v+1)} A_v(k),$$

avec

$$A_v(k) := \left(\frac{(v!)^2 R(k)}{2v\mu_v^2(k)}\right)^{1/(2v+1)}.$$

La valeur de l'$\mathbf{AMISE}_0[k]$ en h_0, après simplification, est

$$\mathbf{AMISE}_0[k] = (1+2v)\left(\frac{R\left(f^{(v)}\right)\mu_v^2(k)R(k)^{2v}}{(v!)^2(2v)^{2v}}\right)^{1/(2v+1)} n^{-2v/(2v+1)}.$$

En utilisant les noyaux du second-ordre, on obtient

$$\mathbf{AMISE}_0[k] = \frac{5}{4}\left(\mu_2^2(k)\right)R^4(k)R(f^{(2)}))^{1/5}n^{-4/5}.$$

1.1.8 Noyaux asymptotiquement optimales

Le noyau d'Epanechnikov est souvent appelé le "noyau optimal". Pour comparer les noyaux, on utilise le principe d'efficacité relative :

$$\begin{aligned}\mathbf{eff}[k] &= \left(\frac{\mathbf{AMISE}_0[k]}{\mathbf{AMISE}_0[k_{v,1}]}\right)^{(1+2v)/2v} \\ &= \frac{\left(\mu_v^2(k)\right)^{1/2v} R(k)}{\left(\mu_v^2(k_{v,1})\right)^{1/2v} R(k_{v,1})}.\end{aligned}$$

1.1.9 Paramètre de lissage empirique

Le paramètre optimal dépend de l'inconnue $R\left(f^{(v)}\right)$. Silverman a proposé d'essayer la largeur de la fenêtre calculée en remplaçant $R\left(f^{(v)}\right)$ dans la formule optimale par $R\left(g_{\widehat{\sigma}}^{(v)}\right)$ où g_σ est une densité de référence. Un candidat plausible pour f est $\widehat{\sigma}^2$ (l'écart

type de l'échantillon). Le choix standard est de poser $g_\sigma = \phi_{\widehat{\sigma}}$, la densité $\mathcal{N}(0, \widehat{\sigma}^2)$.

Etant donné une densité g, on définit $g_\sigma(x) = \sigma^{-1} g(x/\sigma)$ et $g_\sigma^{(v)}(x) = \sigma^{-1-v} g^{(v)}(x/\sigma)$.

Ainsi

$$\begin{aligned} R\left(g_{\widehat{\sigma}}^{(v)}\right)^{-1/(2v+1)} &= \left(\int g_\sigma^{(v)}(x)^2\, dx\right)^{-1/(2v+1)} \\ &= \left(\sigma^{-2-2v}\int g^{(v)}(x/\sigma)^2\, dx\right)^{-1/(2v+1)} \\ &= \left(\sigma^{-1-2v}\int g^{(v)}(x)^2\, dx\right)^{-1/(2v+1)} = \sigma R\left(g^{(v)}\right)^{-1/(2v+1)}. \end{aligned}$$

En outre, on pose

$$\left(R\left(\phi^{(v)}\right)\right)^{-1/(2v+1)} = 2\left(\frac{\pi^{1/2} v!}{(2v)!}\right)^{1/(2v+1)},$$

et

$$R\left(\phi_{\widehat{\sigma}}^{(v)}\right)^{-1/(2v+1)} = 2\widehat{\sigma}\left(\frac{\pi^{1/2} v!}{(2v)!}\right)^{1/(2v+1)}.$$

Paramètre de lissage empirique est alors $h = \widehat{\sigma} C_v(k)\, n^{-1/(2v+1)}$ où

$$\begin{aligned} C_v(k) &= R\left(\phi^{(v)}\right)^{-1/(2v+1)} A_v(k) \\ &= 2\left(\frac{\pi^{1/2}(v!)^3 R(k)}{2v\,(2v)!\mu_v^2(k)}\right)^{1/(2v+1)}. \end{aligned}$$

Paramètre de lissage de Silverman Soit $h = \widehat{\sigma} C_v(k)\, n^{-1/(2v+1)}$ où $\widehat{\sigma}$ est l'écart type de l'échantillon, v est l'ordre du noyau, et $C_v(k)$ est la constante.

1.1.10 Dérivées de la densité

Considérons le problème de l'estimation de la r-ième dérivée de la densité

$$f^{(r)}(x) = \frac{d^r}{dx^r} f(x).$$

Un estimateur naturel prenant la forme

$$\widehat{f}^{(r)}(x) = \frac{d^r}{dx^r}\widehat{f}(x) = \frac{1}{nh^{1+r}}\sum_{i=1}^{n} k^{(r)}\left(\frac{X_i - x}{h}\right),$$

où $k^{(r)}(x) = \frac{d^r}{dx^r}k(x)$. On observe que

$$\mathbf{E}\left[\frac{1}{h^{1+r}}k^{(r)}\left(\frac{X_i - x}{h}\right)\right] = \int_{-\infty}^{\infty}\frac{1}{h^{1+r}}k^{(r)}\left(\frac{z - x}{h}\right)f(z)dz.$$

En faisant un changement de variables et r intégrations par parties on obtient

$$\int_{-\infty}^{\infty} k(u)\, f^{(r)}(x + hu)du.$$

En appliquant un développement de Taylor d'ordre v à $f^{(r)}(x + hu)$ on obtient

$$f^{(r)}(x + hu) = f^{(r)}(x) + \frac{1}{v!}f^{(r+v)}(x)\, h^v \mu_v(k) + o(h^v).$$

Ainsi le biais asymptotique est égal

$$\begin{aligned}\mathbf{Bias}\left[\widehat{f}^{(r)}(x)\right] &= \mathbf{E}\left[\widehat{f}^{(r)}(x)\right] - f^{(r)}(x) \\ &= \frac{1}{v!}f^{(r+v)}(x)\, h^v \mu_v(k) + o(h^v).\end{aligned}$$

Pour la variance, on trouve

$$\begin{aligned}
\mathbf{var}\left[\widehat{f}^{(r)}(x)\right] &= \frac{1}{nh^{2+2r}}\mathbf{var}\left[k^{(r)}\left(\frac{X_i-x}{h}\right)\right] \\
&= \frac{1}{nh^{2+2r}}\mathbf{E}\left[k^{(r)}\left(\frac{X_i-x}{h}\right)^2\right] - \frac{1}{n}\left(\frac{1}{nh^{2+2r}}E\left[k^{(r)}\left(\frac{X_i-x}{h}\right)\right]\right)^2 \\
&= \frac{1}{nh^{2+2r}}\int_{-\infty}^{\infty} k^{(r)}\left(\frac{z-x}{h}\right)^2 f(z)dz - \frac{1}{n}f^{(r)}(x)^2 + O\left(\frac{1}{n}\right) \\
&= \frac{1}{nh^{1+2r}}\int_{-\infty}^{\infty} k^{(r)}(u)^2 f(x+hu)\,du + O\left(\frac{1}{n}\right) \\
&= \frac{f(x)}{nh^{1+2r}}\int_{-\infty}^{\infty} k^{(r)}(u)^2\,du + O\left(\frac{1}{n}\right) \\
&= \frac{f(x)\,R\left(k^{(r)}\right)}{nh^{1+2r}} + O\left(\frac{1}{n}\right).
\end{aligned}$$

AMSE et AMISE sont

$$\begin{aligned}
\mathbf{AMSE}\left[\widehat{f}^{(r)}(x)\right] &= \frac{f^{(r+v)}(x)^2 h^{2v}\mu_V^2(k)}{(v!)^2} + \frac{f(x)\,R\left(k^{(r)}\right)}{nh^{1+2r}} \\
&= \frac{R\left(f^{(r+v)}\right)h^{2v}\mu_v^2(k)}{(v!)^2} + \frac{R\left(k^{(r)}\right)}{nh^{1+2r}}.
\end{aligned}$$

Paramètre asymptotiquement optimale est

$$\begin{aligned}
h_r &= C_{r,v}(k,f)n^{-1/(1+2r+2v)}. \\
C_{r,v}(k,f) &= R\left(f^{(r+v)}\right)^{-1/(1+2r+2v)} A_{r,v}(k). \\
A_{r,v}(k) &= \left(\frac{(1+2r)\,(v!)^2 R(k^{(r)})}{2v\mu_v^2(k)}\right)^{1/(1+2r+2v)}.
\end{aligned}$$

L'AMISE avec la fenêtre optimale est

$$\mathbf{AMISE}\left[\widehat{f}^{(r)}(x)\right] = (1+2r+2v)\left(\frac{\mu_v^2(k)}{(v!)^2(1+2r)}\right)^{(2r+1)/(1+2r+2v)} \times \left(\frac{R\left(k^{(r)}\right)}{2v}\right)^{2v/(1+2r+2v)} n^{-2v/(1+2r+2v)}.$$

L'efficacité relative d'un noyau k est alors

$$\mathbf{eff}[k] = \left(\frac{\mathbf{AMISE}_0[k]}{\mathbf{AMISE}_0[k_{v,r+1}]}\right)^{(1+2v+2r)/2v} = \left(\frac{\mu_v^2(k)}{\mu_v^2(k_{v,r+1})}\right)^{(1+2r)/2v} \frac{R(k)}{R(k_{v,r+1})}.$$

Les efficacités relatives des différents noyaux sont présentées dans le tableau 2.

		Biweight	Triweight	Gaussien
	$v=2$	1.0000	1.0185	1.2191
$r=1$	$v=4$	1.0000	1.0159	1.2753
	$v=6$	1.0000	1.0136	1.3156
$r=2$	$v=2$		1.0000	1.4689
	$v=4$		1.0000	1.5592
	$v=6$		1.0000	1.6275

TAB. 1.2 – Efficacités relatives eff[K]

1.1.11 Estimation de la densité multivariée

Supposons maintenant que X_i, est un q-vecteur et que nous voulons estimer sa densité $f(x) = f(x_1, ..., x_q)$. Un estimateur à noyau multivarié prend la forme

$$\widehat{f}(x) = \frac{1}{n|H|}\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right),$$

où $K(u)$ est une fonction noyau multivariée dépendant d'un vecteur de largeur de fenêtre $H = (h_1, ..., h_q)^{'}$ et $|H| = h_1 h_2 ... h_q$. Un noyau multivarié satisfait

$$\int K(u)(du) = \int K(u)\, du_1 ... du_q = 1.$$

Typiquement, $K(u)$ prend la forme produit

$$K(u) = k(u_1)\, k(u_2) ... k(u_q).$$

Le biais de l'estimateur est

$$\textbf{Bias}\left[\widehat{f}(x)\right] = \frac{\mu_v(k)}{v!} \sum_{j=1}^{q} \frac{\partial^v}{\partial x_j^v} f(x)\, h_j^v + o\left(h_1^v + ... + h_q^v\right),$$

et la variance

$$\begin{aligned} \textbf{var}\left[\widehat{f}(x)\right] &= \frac{f(x)\, R(K)}{n|H|} + O\left(\frac{1}{n}\right) \\ &= \frac{f(x)\, R^q(k)}{n h_1 h_2 ... h_q} + O\left(\frac{1}{n}\right). \end{aligned}$$

D'où l'AMISE est

$$\textbf{AMISE}\left[\widehat{f}(x)\right] = \frac{\mu_v^2(k)}{(v!)^2} \int \left(\sum_{j=1}^{q} \frac{\partial^v}{\partial x_j^v} f(x)\, h_j^v \right)^2 (dx) + \frac{R^q(k)}{n h_1 h_2 ... h_q}.$$

Supposons que $h_1 = h_2 = ... = h_q = h$. Alors

$$\textbf{AMISE}\left[\widehat{f}(x)\right] = \frac{\mu_v^2(k)\, R(\nabla^v f)}{(v!)^2} h^{2v} + \frac{R(k)^q}{n h^q},$$

où

$$\nabla^v f(x) = \sum_{j=1}^{q} \frac{\partial^v}{\partial x_j^v} f(x).$$

Nous constatons que le paramètre de lissage optimal est

$$h_0 = \left(\frac{(v!)^2 qR(k)^q}{2v\kappa_v^2(k) R(\nabla^v f)} \right)^{1/(2v+q)} n^{-1/(2v+q)}.$$

Nous remplaçons f par la densité normale multivariée ϕ. Nous pouvons calculer que

$$R(\nabla^v \phi) = \frac{q}{\pi^{q/2} 2^{q+v}} \left((2v-1)! + (q-1)\left((v-1)!\right)^2 \right).$$

En faisant cette substitution, nous obtenons $h_0 = C_v(K,q)\, n^{-1/(2v+q)}$ où

$$C_v(k,q) = \left(\frac{\pi^{q/2} 2^{q+v-1} (v!)^2 R^q(K)}{v\mu_v^2(K) \left((2v-1)! + (q-1)\left((v-1)!\right)^2 \right)} \right)^{1/(2v+q)}.$$

Nous obtenons le paramètre de lissage empirique de base pour la variable j-ième

$$h_j = \widehat{\sigma}_j C_v(k,q)\, n^{-1/(2v+q)}.$$

1.1.12 Validation croisée des moindres carrés

Définir l'erreur quadratique intégrée moyenne (**MISE**)

$$\mathbf{MISE}[h] = \int \left(\widehat{f}(x) + f(x) \right)^2 (dx) = \int \widehat{f}^2(x)\,(dx) - 2\int \widehat{f}(x) f(x)\,(dx) + \int f^2(x)\,(dx).$$

Le premier terme peut être calculé directement. Pour le cas univarié

$$\begin{aligned}\int \widehat{f}^2(x)\,dx &= \int \left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right)^2 dx \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) dx.\end{aligned}$$

La convolution de k avec lui-même est $\overline{k}(x) = \int k(u)\,k(x-u)\,du = \int k(u)\,k(u-x)\,du$ avec changement de variables $u = \frac{X_i - x}{h}$,

$$\begin{aligned}\frac{1}{h}\int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) dx &= \int k(u)\,k\left(u - \frac{X_i - X_j}{h}\right) du \\ &= \overline{k}\left(\frac{X_i - X_j}{h}\right).\end{aligned}$$

Dans le cas multivarié,

$$\int \widehat{f}^2(x)\,dx = \frac{1}{n^2\,|H|}\sum_{i=1}^{n}\sum_{j=1}^{n}\overline{K}\left(H^{-1}(X_i - X_j)\right),$$

où $\overline{K}(u) = \overline{k}(u_1)...\overline{k}(u_q)$.

Le deuxième terme dans l'expression pour **MISE**$[h]$ qui dépend $f(x)$ est donc inconnu et doit être estimé. En général, une estimation raisonnable de l'intégrale $\int g(x)\,f(x)$ est $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$, suggérer l'estimation $\frac{1}{n}\sum_{i=1}^{n}\widehat{f}(X_i)$. Une façon de nettoyer cela est de remplacer $\widehat{f}(X_i)$ par "leave-one-out" estimation $\widehat{f}_{-i}(X_i)$, où

$$\widehat{f}_{-i}(x) = \frac{1}{(n-1)\,|H|}\sum_{j\neq i}^{n}\overline{K}\left(H^{-1}(X_j - x)\right),$$

est l'estimation de la densité calculée sans observation X_i, et donc

$$\widehat{f}_{-i}(X_i) = \frac{1}{(n-1)|H|} \sum_{j \neq i}^{n} \overline{K}\left(H^{-1}(X_i - X_j)\right).$$

Alors

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i) = \frac{1}{n(n-1)|H|} \sum_{i=1}^{n}\sum_{j \neq i}^{n} K\left(H^{-1}(X_i - X_j)\right),$$

dans le sens où

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i)\right] = \mathbf{E}\left[\int \widehat{f}(x) f(x)\, dx\right].$$

telle que

$$\begin{aligned}
\mathbf{E}\left[\widehat{f}_{-n}(X_n)\right] &= \mathbf{E}\left[\mathbf{E}\left[\widehat{f}_n(x) f(x)(dx)\right]\right] \\
&= \int \mathbf{E}\left[\widehat{f}_n(x)\right] f(x)(dx) \\
&= \mathbf{E}\left[\int \widehat{f}_n(x) f(x)(dx)\right].
\end{aligned}$$

Ensemble, le critère de validation croisée des moindres carrés est

$$CV(h_1, ..., h_q) = \frac{1}{n^2|H|}\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{K}\left(H^{-1}(X_i - X_j)\right) - \frac{2}{n(n-1)|H|}\sum_{i=1}^{n}\sum_{j \neq i}^{n} K\left(H^{-1}(X_i - X_j)\right).$$

Une autre façon d'écrire ceci est

$$\begin{aligned}
CV(h_1, ..., h_q) &= \frac{\overline{K}(0)}{n|H|} + \frac{1}{n^2|H|}\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{K}\left(H^{-1}(X_i - X_j)\right) \\
&\quad - \frac{2}{n(n-1)|H|}\sum_{i=1}^{n}\sum_{j \neq i}^{n} K\left(H^{-1}(X_i - X_j)\right). \\
&\simeq \frac{R^q(k)}{n|H|} + \frac{1}{n^2|H|}\sum_{i=1}^{n}\sum_{j=1}^{n} \overline{K}\left(H^{-1}(X_i - X_j)\right) - 2K\left(H^{-1}(X_i - X_j)\right).
\end{aligned}$$

Utilisons $\overline{K}(0) = \overline{k}(0)^q$ et $\overline{k}(0) = \int k^2(u)\,du$,et l'approximation est $(n-1)$ par n.

1.1.13 Noyaux de convolution

Si $k(x) = \phi(x)$ alors $\overline{k}(x) = \exp(-x^2/4)/\sqrt{4\pi}$. Lorsque $k(x)$ est un noyau Gaussien d'ordre supérieur, Wand et Schucany (Canadian Journal of Statistics,1990, p.201) donnent une expression pour $\overline{k}(x)$.

Pour la classe polynômiale, parce que le noyau $k(u)$ a le support sur $[-1,1]$, il s'ensuit que $\overline{k}(x)$ a le support sur $[-2,2]$ et pour $x \geq 0$ égale $\overline{k}(x) = \int_{x-1}^{1} k(u)\,k(x-u)\,du$. Cette intégrale peut être facilement résolu en utilisant un logiciel algébrique (Maple, Mathematica), mais l'expression peut être plutôt encombrante.

1.1.14 Normalité asymptotique

L'estimateur du noyau est la moyenne de l'échantillon

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|H|} K\left(H^{-1}(X_i - x)\right).$$

On peut donc appliquer le théorème limite central. Mais le taux de convergence n'est pas $\sqrt{n}$. Nous savons que

$$\mathbf{var}\left[\widehat{f}(x)\right] = \frac{f(x)\,R^q(k)}{nh_1h_2...h_q} + O\left(\frac{1}{n}\right).$$

Le taux de convergence est donc $\sqrt{nh_1h_2...h_q}$. Ainsi

$$\sqrt{nh_1h_2...h_q}\left(\widehat{f}(x)-\mathbf{E}\left[\widehat{f}(x)\right]\right)=\frac{\sqrt{nh_1h_2...h_q}}{n}\sum_{i=1}^{n}\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_{ni},$$

où

$$Z_{ni}=\sqrt{nh_1h_2...h_q}\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)-\mathbf{E}\left[\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right]\right).$$

On voit que

$$\mathbf{var}\left[Z_{ni}\right]\simeq f(x)R^q(k).$$

Donc par l'Théorème central limite

$$\sqrt{nh_1h_2...h_q}\left(\widehat{f}(x)-\mathbf{E}\left[\widehat{f}(x)\right]\right)\rightarrow_d\mathcal{N}\left(0,f(x)R^q(k)\right).$$

Utilisant l'hypothése que h est d'ordre plus petit que le taux optimal, $h=o\left(n^{-1/(2v+1)}\right)$.

Pour obtenir alors le résultat

$$\sqrt{nh}\left(\widehat{f}(x)-f(x)\right)\rightarrow_d N\left(0,f(x)R(k)\right).$$

1.1.15 Intervalles de confiance judicieux

Dans le cas univarié, les intervalles de confiance conventionnels prennent

$$\widehat{f}(x)\pm 2\left(\widehat{f}(x)R(k)/(nh)\right)^{1/2}.$$

Pour tester $H_0 : f(x) = f_0$, est

$$t(f_0) = \frac{\widehat{f}(x) - f_0}{\sqrt{nhf_0 R(k)}}.$$

Nous rejetons H_0 si $|t(f_0)| > 2$. Par la règle sans rejet, un intervalle de confiance asymptotique de 95% pour f est l'ensemble de f_0 qui ne rejette, c-à-d l'ensemble de f telle que $|t(f_0)| \leq 2$. Ceci

$$C(x) = \left\{ f : \left| \frac{\widehat{f}(x) - f}{\sqrt{nhfR(k)}} \right| \leq 2 \right\}.$$

Cet ensemble doit être trouvé numériquement.

Chapitre 2

Estimateur de Nadaraya-Watson

2.1 Estimation à noyau de la régression

2.1.1 Régression non paramétrique

Introduction

L'objectif d'une analyse de régression est de produire une analyse raisonnable de la fonction de réponse inconnue f, où pour les points N data (X_i, Y_i), la relation peut être modélisée comme suit : $y_i = m(x_i) + \varepsilon_i, \quad i = 1, ..., N$. Notée : $m(\cdot) = \mathbf{E}[y/x]$ if $\mathbf{E}[\varepsilon/x] = 0 - i.e., \varepsilon \perp x$

Nous avons différentes façons de modéliser la fonction d'attende conditionnelle (FAC), $m(\cdot)$:

-Approche paramétrique $y_i = x_i'\beta + \varepsilon_i, \quad i = 1, ..., N.$

-Approche non paramétrique $y_i = m(x_i) + \varepsilon_i, \quad i = 1, ..., N.$

-Approche semi-paramétrique $y_i = x_i'\beta + m_z(z_i) + \varepsilon_i, \quad i = 1, ..., N.$

Lissage

Nous voulons établir un lien entre y et x, sans prendre de forme fonctionnelle. Tout d'abord, nous considérons le cas d'un régresseur :

$$y_i = m(x_i) + \varepsilon_{i,} \quad i = 1, ..., N.$$

Dans l'modeles de lien cumulatif , une forme fonctionnelle linéaire est supposé : $m(x_i) = x_i^{'}\beta$.

Dans de nombreux cas, il n'est pas clair que la relation est linéaire.

Les modèles non paramétriques tentent de découvrir la relation (approximative) entre y_i et x_i. Approche très flexible, mais nous devons faire quelques hypothèses.

Une approximation raisonnable de la courbe de régression $m(x_i)$ sera la moyenne des variables de réponse près d'un point x_i.

$$\widehat{m}(x) = N^{-1} \sum_{i=1}^{N} W_{N,h,i}(x)\, y_i.$$

Interprétation de lissage Supposons que la somme des poids s'élève à 1 pour tous les x_i. Le $\widehat{m}(x)$ est une estimation des moindres carrés à x puisque nous pouvons $\widehat{m}(x)$ écrire comme solution à

$$\min_{\theta} N^{-1} \sum_{i=1}^{N} W_{N,h,i}(x)\,(y_i - \theta)^2.$$

C'est-à-dire qu'un estimateur de régression par noyau est une régression locale constante, puisqu'il définit $m(x)$ égal à une constante, θ, dans le très petit voisinage de x_0 :

$$\min_{\theta} N^{-1} \sum_{i=1}^{N} W_{N,h,i}(x)\,(y_i - \theta)^2 = N^{-1} \sum_{i=1}^{N} W_{N,h,i}(x)\,(y_i - \widehat{m}(x))^2.$$

Problèmes de lissage **Q :** Quel est l'effet du lissage sur les données ?

(1) Puisque la moyenne est faite sur les observations voisines, une estimation de $m(\cdot)$ aux pics ou aux fonds les aplatir. Ce biais d'échantillon fini dépend de la courbure locale de $m(\cdot)$. Solution : Quartier rétrécir !

(2) Aux points limites, la moitié des poids ne sont pas définis. Cela crée également un biais.

(3) Lorsqu'il y a des régions de données éparses, les poids peuvent être indéfinis aucune observation à la moyenne.

Solution : Définir les poids avec la portée variable.

L'efficacité informatique est importante.

2.1.2 Régression à noyau

Les régressions à du noyau sont des estimateurs moyens pondérés qui utilisent les fonctions du noyau comme poids

Le poids est défini par

$$W_{hi}(x) := K_h(x - X_i) / \widehat{f}_h(x),$$

où

$$\widehat{f}_h(x) = N^{-1} \sum_{i=1}^{N} K_h(x - X_i), \text{ et } K_h(u) = h^{-1} K(u/h).$$

Les formules statistiques standard nous permettent de calculer $\mathbf{E}[y/x]$:

$$\mathbf{E}[y/x] = m(x) = \int y \mathbf{f}_C(y/x)\, dy,$$

où $\mathbf{f}_C$ est la distribution de y conditionnelle à x. En particulier :

$$\mathbf{E}[y/x] = m(x) = \frac{\int_{-\infty}^{\infty} y\mathbf{f}_J(y,x)\,dy}{\mathbf{f}_M} = \frac{\int_{-\infty}^{\infty} y\mathbf{f}_J(y,x)\,dy}{\int_{-\infty}^{\infty} \mathbf{f}_J(y,x)\,dy}.$$

Lorsque les souscriptions M et J se réfèrent respectivement aux distributions marginales et aux distributions conjointes.

Estimateur Nadaraya-Watson

Tout d'abord, considérer d'abord $\mathbf{f}_M(x)$: $\widehat{\mathbf{f}}_M(x) = (Nh)^{-1}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)$.

Deuxièmement, considérer $\int \mathbf{f}_J(y,x_0)\,dy = (Nh)^{-1}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)$.

Ce qui suggère $\int y\mathbf{f}_J(y,x_0)\,dy = (Nh)^{-1}\sum_{i=1}^{N} y_i K\left(\frac{x_i - x_0}{h}\right)$.

En connectant ces deux estimations de noyau des termes dans le numérateur et le dénominateur de l'expression pour $m(x)$ donne l'estimateur de noyau Nadaraya-Watson(NW) :

$$\widehat{m}(x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)}.$$

Estimateur NW-Différent K(.)

La normalisation des poids $\widehat{\mathbf{f}}_h(x) = N^{-1}\sum_{i=1}^{N} K_h(x - X_i)$ est appelée l'estimateur de densité du noyau Rosenblatt-Parzon. Il s'assure que les poids s'additionnent pour atteindre 1.

Deux constantes importantes associées à une fonction noyau $K(\cdot)$ sont sa variance $\sigma_K^2 = d_K$ et la rugosité c_K, (également R_K), qui sont définies comme : $d_K = \int z^2 K(z)\,dz$ et $c_K = \int K^2(z)\,dz$.

Beaucoup de $K(\cdot)$ sont possible. Des considérations pratiques et théoriques limitent les choix. Choix habituels : Epanechnikov, Gaussien, Quatrique (Biweight) et Tricube

(Triweight).

Rappelons que le noyau Epanechnikov bénéfice de propriétés optimales.

Estimateur linéaire locale

Nous avons motivé l'estimateur NW à x comme moyenne de $y_i^{'}$ pour des observations dans un voisinage de x : une approximation de constante locale.

Au lieu de cela, nous pouvons faire moindre carré ordinaire (MCO) dans le même voisinage. Si nous utilisons une fonction de pondération, ceci est appelé l'estimateur linéaire locale (LL).

L'idée est de s'adapter au modèle local $y_i = \alpha + (x_i - x)^{'} \beta + \varepsilon_i$.

Nous utilisons $(X_i - x)$ plutôt que X_i pour avoir $m(x) = \mathbf{E}[y_i / X_i = x] = \alpha$.

Nous faisons MCO avec des observations telles que $|X_i - x| \leq h$ c'est-à-dire

$$\min_{\alpha,\beta} N^{-1} \sum_{i=1}^{N} \left(y_i - \alpha - (x_i - x)^{'} \beta \right)^2 I\left[|x_i - x| \leq h \right].$$

Nous avons un problème moindre carré (MC) pondéré, qui peut être généralisé à :

$$\min_{\alpha,\beta} N^{-1} \sum_{i=1}^{N} \left(y_i - \alpha - (x_i - x)\beta \right)^2 K\left(\frac{x_i - x}{h} \right).$$

Puis, en réglant $Z_i = \left[\mathbf{1}(X_i - x) \right]^{'}$ donne :

$$\begin{aligned} \begin{pmatrix} \widehat{\alpha}(x) \\ \widehat{\beta}(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} 1\left(|X_i - x| \leq h \right) Z_i Z_i^{'} \right)^{-1} \left(\sum_{i=1}^{n} 1\left(|X_i - x| \leq h \right) Z_i y_i \right) \\ &= \left(\sum_{i=1}^{n} K\left(H^{-1}(X_i - x) \right) Z_i Z_i^{'} \right)^{-1} \left(\sum_{i=1}^{n} K\left(H^{-1}(X_i - x) \right) Z_i y_i \right). \end{aligned}$$

La deuxième ligne est valide pour toute fonction du noyau (multivariée). Il s'agit d'une régression (localement) pondérée de y_i sur X_i.

L'estimateur LL préserve les données et se comporte mieux aux limites.

Lissage du nuage de points à pondération Un estimateur populaire de régression locale est le lissage de diagramme de dispersion à pondération locale (minuscules), introduit par Cleveland (1979)[1].

Il utilise une variable h, déterminée par la distance, et il utilise un noyau tricubique

$$K(z) = (70/81)\left(1 - |z|^3\right)^3 \mathbf{1}\left[|z| < 1\right].$$

Estimateur NW pondéré

Hall (1999)[6] proposé une estimation pondérée du NW définie par

$$\widehat{m}_h(x_0) = \frac{\sum_{i=1}^N p_i(x_0) K_h\left(\frac{x_0 - X_i}{h}\right) y_i}{\sum_{i=1}^N p_i(x_0) K_h\left(\frac{x_0 - X_i}{h}\right)},$$

où $p_i(x)$ sont des poids. Les poids satisfont :

$p_i(x) \geq 0$, $\sum_i p_i(x) = 1$, $\sum_i p_i(x) K\left(h^{-1}(x_i - x)\right)(x_i - x) = 1$.

Les poids sont déterminés par la probabilité empirique. Spécifiquement, pour chaque x; vous maximisez $\sum_i \ln p_i(x)$ .Les contraintes ci-dessus.

Les solutions prennent la forme

$$p_i(x) = \frac{1}{n\left(1 + \lambda'(X_i - x) K\left(H^{-1}(X_i - x)\right)\right)},$$

où λ est un ML, trouvé par optimisation numérique.

L'estimateur $\widehat{m}(x)$ à la même distribution asymptotique que l'estimateur LL. Lorsque

$y_i \geq 0$; les estimateurs NW standard et pondérés satisfont également $\widehat{m}(x) \geq 0$. C'est bon ($m(x)$ n'est pas négatif). D'autre côté, l'estimateur LL n'est pas nécessairement non négatif.

Résidus, concordance et convergence

Nous sommes habitués à utiliser les résidus ajustés pour construire des mesures l'ajustement des lois (AD). Les résidus sont définis comme d'habitude :

$$e_i = y_i - \widehat{m}(x_i), \ i = 1, ..., N.$$

Problème : En général, mais surtout quand il est petit, il est difficile de voir e_i comme une mesure du AD. Comme $h \to 0$, $\widehat{m}(\cdot) \to y_i$ (et $e_i \to 0$). Cela indique que l'erreur vraie n'est pas nulle.

Solution : Mesurer l'ajustement de la régression à $x = x_i$ en révaluant le modèle en excluant l'observation i-ième (notation :" $-i$", l'observation i-ième exclue). Pour la régression NW, on obtient :

$$\widehat{m}_{-i}(x) = \frac{\sum_{j \neq i}^{N} y_i K_h(x - X_j) y_j}{\sum_{j \neq i}^{N} K_h(x - X_j)} = \sum_{j \neq i}^{N} w_{N,h,-i}(x) y_j.$$

Maintenant, les résidus de congé unique sont définis comme suit :

$$e_{-i} = y_i - \widehat{m}_{-i}(x_i), \ i = 1, ..., N.$$

e_{-i} n'est pas une fonction de y_i; il n'y a pas de tendance à déborder pour les petits h.

Le résiduel moyen d'un seul carré est $CV := \frac{1}{N} \sum_{i=1}^{N} e_{-i}(h)^2$.

Cette fonction de son critère connu sous le nom de cross-validation. Ce critère peut être

utilisé pour sélectionner le lissage de paramètre.

Le lissage de paramètre convergence h_{CV} est la valeur qui minimise $CV(h)$. Habituellement, la restriction $h_{CV} \geq h_{LB}$ est imposée, où h_{LB} est une limite inférieure (lower bound) pour h_{CV}, pour s'assurer que le fenêtre n'est pas trop petite.

Il s'avère que $CV(h)$ est un estimateur de l'erreur de prévision moyenne au carré. C'est,

$$\mathbf{E}\left[CV\left(h\right)\right] = \mathbf{MSFE}_{N^{-1}}\left[h\right] = \mathbf{MISE}_{N^{-1}}\left[h\right] + \sigma^2.$$

Multivariée L'estimation NW définie par :

$$\widehat{m}_h\left(x\right) = \frac{\sum_{i=1}^{N} y_i K_h\left(x - X_i\right)}{\sum_{i=1}^{N} K_h\left(x - X_i\right)} = \sum_{i=1}^{N} y_i w_{N,h,i}\left(x\right).$$

La dernière expression montre simplement que cet estimateur peut être considéré comme une moyenne pondérée des observations de y. Dans la notation matricielle, nous pouvons écrire $\widehat{Y} = \mathbf{M}(h)\mathbf{Y}$, avec

$$M\left(h\right) = \begin{bmatrix} \frac{K\left(\frac{X_1 - x_1}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_1}{h}\right)} & \frac{K\left(\frac{X_2 - x_1}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_1}{h}\right)} & \cdots & \frac{K\left(\frac{Xn - x_1}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_1}{h}\right)} \\ \frac{K\left(\frac{X_1 - x_2}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_2}{h}\right)} & \frac{K\left(\frac{X_2 - x_2}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_2}{h}\right)} & \cdots & \frac{K\left(\frac{Xn - x_2}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_2}{h}\right)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{K\left(\frac{X_1 - x_n}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_n}{h}\right)} & \frac{K\left(\frac{X_2 - x_n}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_n}{h}\right)} & \cdots & \frac{K\left(\frac{Xn - x_n}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_n}{h}\right)} \end{bmatrix}$$

Prédictions de régression des grains $\widehat{Y} = \mathbf{M}(h)\mathbf{Y}$.

Prédictions de régression de la ligne : $\widehat{Y} = \mathbf{P}_x\mathbf{Y}$.

Un noyau multivarié est construit, ligne par ligne, en calculant le produit de densités

marginales pour chaque variable dans la matrice des régresseurs X. C'est-à-dire,

$$h^{-d}K\left(\frac{X-x_i}{h}\right)=\Pi_{j=1}^{d}h^{-1}K\left(\frac{x_j-x_{ji}}{h}\right).$$

2.1.3 Comparaison

Lissage moyen vs noyau

Moyenne (uniforme) plus lisse

$$\widehat{m}\left(x\right)=\frac{\sum_{i=1}^{N}w\left(\frac{x-x_i}{h}\right)x_i}{\sum_{i=1}^{N}w\left(\frac{x-x_i}{h}\right)},$$

où

$$w\left(u\right)=\begin{cases}1 & \text{si } \mid u\mid<1\\ 0 & \text{sinon.}\end{cases}$$

Noyau plus lisse

$$\widehat{m}\left(x\right)=\frac{\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right)x_i}{\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right)},$$

où $k\left(\cdot\right)$ est le noyau gaussien.

2.1.4 K-NN Estimations du voisinage le plus proche

Les méthode k-NN sont plus couramment utilisées pour la régression que pour l'estimation de la densité. Le lisseur k-NN classique est défini comme

$$\widehat{m}_k\left(x_0\right)=k^{-1}\sum_{i=1}^{N}I\left[\left(\left\|x_0-x_i\right\|\leq d_k\left(x_0\right)\right)\right]Y_i,$$

il s'agit de la valeur moyenne de y_i parmi les observations qui sont les k voisins les plus proches de x_0. (d_k est la distance entre x et x_0).

Un k-NN estimateur lisse est :

$$\widehat{m}_h(x_0) = \frac{\sum_{i=1}^{N} w_k(\|x_0 - x_i\| \leq d_k)\, y_i}{\sum_{i=1}^{N} w_k(\|x_0 - x_i\| \leq d_k)} = \sum_{i=1}^{N} W_{N,k,i}(x_0)\, y_i.$$

Une moyenne pondérée des k voisins les plus proches.

Le paramètre de lissage fixe le degré de lissage de la courbe estimée. Il joue un rôle similaire à h pour les lisseurs de grains.

L'influence de la variation de k sur les caractéristiques qualitatives de la courbe estimée est similaire à celle observée pour l'estimation du noyau avec un noyau uniforme.

Lorsque $k > N$, le lisseur k-NN est égal à la moyenne des variables de réponse. Lorsque $k = 1$, les observations sont reproduites à X_i, et pour un x entre deux variables prédictives adjacentes, une fonction pas à pas est obtenue avec un saut au milieu entre les deux observations.

Lorsque X est un vecteur, l'échelle importe. Ensuite, toujours l'échelle X.

Pour le cas d'un seul régresseur, nous avons des résultats asymptotiques similaires à ceux du cas de densité univariée.

Soit $N \to \infty, k \to 0$, et $Nk \to \infty$. Le biais et la variance de l'estimation k-NN avec des poids uniformes sont donnés par

$$\mathbf{E}\left[\widehat{m}_k(x) - m(x)\right] \approx \frac{1}{24\mathbf{f}(x)^3}\left[\left(m''\mathbf{f} + 2m'\mathbf{f}\right)(x)\right](k/n)^2.$$

$$var\{\widehat{m}_k(x)\} \approx \sigma^2(x)/k.$$

Calculs

Un grand avantage du lisseur k-NN est le calcul.

Les calculs peuvent être facilement mis à jour. L'algorithme nécessite des opérations

$O(N)$ pour calculer le lissage à tous les $x_i^{'}s$. comparez ceci aux calculs $O(N^2h)$ pour l'estimateur noyau.

La validation croisée est utilisée pour définir k, en utilisant laisser de côté les erreurs :

$$CV\left(k\right)=\frac{1}{N}\sum_{i=1}^{N}\left[y_i-\widehat{m}_{-i}\left(x_i\right)\right]^2.$$

2.1.5 Estimation de la variance non paramétrique

Supposons que nous ayons le modèle de régression suivant : $y_i=m_z\left(x_i\right)+x_i'\beta+\varepsilon_i$.

$\mathbf{E}\left[\varepsilon_i/X_i,Z_i\right]=0$.

$\varepsilon_i^2=\sigma^2\left(x_i\right)+\eta_i$, $\mathbf{E}\left[\eta_i/X_i\right]=0$.

$\sigma^2\left(x_i\right)$ est la fonction de régression de ε_i^2 sur x_i. Nous voulons l'estimer.

Problème : Si ε_i^2 a été observé =>régression NW ou LL.

Solution : Utiliser le résiduel non paramétrique e_i : $e_i=y_i-\widehat{m}_i\left(x_i\right)$.

Ensuite, nous pouvons utiliser l'estimateur NW :

$$\widehat{\sigma}\left(x_0\right)=\frac{\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h}\right)e_i^2}{\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h}\right)}.$$

Nous avons un estimateur en deux étapes. Situation semblable si nous utilisons l'estimateur LL. Le lissage de paramètre h ne sont pas les mêmes que pour l'estimation de $\widehat{m}\left(x\right)$.

2.1.6 Estimation des séries

L'estimation des séries est l'autre méthode de régression non paramétrique.

Les méthodes de séries se rapprochent d'une fonction inconnue, $m(x)$, avec une fonction

paramétrique flexible, avec le nombre de paramètres traités de la même manière que la fenêtre dans la régression du noyau.

Une approximation de série à $m(x)$ prend la forme générale : $m_K(x) = m_K(x, \beta)$, où $m_K(x, \beta)$ est une famille paramétrique connue et β est un vecteur d'inconnues k.

Une approximation de série linéaire prend la forme suivante :

$$\widehat{m}_k(x) = \sum_{j=1}^{K} z_{jK}(x)\beta_{jK} = z_K(x)' \beta_K.$$

Approximations uniformes

Une bonne approximation de série $m_K(x)$ aura la propriété qu'ellensenrapproche du vrai $m(x)$ à mesure que K augmente.

Le théorème Stone-Weierstrass affirme que toute fonction continue peut être arbitrairement uniformément bien approximé par un polynôme d'ordre suffisamment élevé :

$$\sup_{x \in X} |m_K(x) - m(x)| \leq \varepsilon,$$

pour tout $\varepsilon > 0$.

C'est-à-dire, $m(x)$ peut être assez bien approximation en choisissant un polynôme approprié.

Le résultat ci-dessus peut être renforcé. Si le dérivé s-th de $m(x)$ est continu, alors l'erreur d'approximation uniforme, r_{Ki}, satisfait.

$$\sup_{x \in X} |r_{Ki} = m_K(x) - m(x)| = O\left(K^{-\alpha}\right),$$

comme $K \longrightarrow \infty$ où $\alpha = s/d$. $(\dim(X) = Nxd)$.

Résultat utile : Il donne un taux auquel l'approximation $m_K(x)$ approche $m(x)$ à mesure que K augmente.

Intuitivement, le nombre de dérivés s indexe le lissage de $m(x)$. Le meilleur taux auquel un polynôme (ou cannelure) se rapproche de $m(x)$.

Les deux résultats sont valables pour les approximations de cannelure.

Régression

Nous avons des observations sur (X, Y). Etapes :

1. Pour chaque i, construisez le vecteur de régresseur $z_{Ki} = z_K(x_i)$, en utilisant les transformations de séries.

2. Empiler les observations dans les matrices et z_K.

3. Faire $MCO \Rightarrow b = \left(Z_K' Z_K\right)^{-1} Z_K' y$.

4. Calculer la fonction de régression MC : $\widehat{m}_k(x) = z_k(x)' b_k$.

5. Calculer les erreurs estimés $e_{ki} = y_{ki} - \widehat{m}_k(x_i) = y_{ki} - z_k(x_i)' b_k$.

Régression-K

Soit β_K est une fonction de K. Cela reflète l'objectif d'être flexible pour intégrer une plus grande complexité lorsque les données sont suffisamment informatives. C'est-à-dire que K va généralement augmenter avec la taille de l'échantillon N.

K joue de rôle de h dans l'estimation du noyau.

Le nombre de termes de série, K, peut être déterminé par CV.

Régression-Asymptotiques

L'estimateur a la composante de biais asymptotique $r_K(x)$, en raison de la série d'ordre fini comme approximation de l'inconnu $m(x)$. La distribution asymptotique montre que le terme de biais est négligeable si K diverge assez vite pour que $NK^{-2\alpha} \to 0$. (En termes pratiques, cela signifie que K est plus grand qu'optimale).

Les erreurs-types asymptotiques pour $m(x)$ peuvent être estimées avec :

$$\widehat{s}(x) = \sqrt{\frac{1}{n} z_K(x)' \widehat{Q}_K^{-1} \widehat{\Omega}_K \widehat{Q}_K^{-1} z_K(x)},$$

où

$$\widehat{\Omega}_K = \frac{1}{n} \sum_{i=1}^{n} z_{Ki} z_{Ki}' \widehat{e}_{iK}^2.$$
$$\widehat{Q}_K = \frac{1}{n} \sum_{i=1}^{n} z_{Ki} z_{Ki}'.$$

Voir Newey (1997) pour plus de détails.

2.1.7 Lissage des splines

La détermination de K n'est pas facile. Un ajustement parfait peut être atteint en donnant beaucoup de flexibilité locale à $\widehat{m}(x)$. Le résultat de cette flexibilité sera un saccadé, difficile à interpréter $\widehat{m}(x)$.

Le lissage des splines quantifie la concurrence entre deux objectifs :

-produire un bon ajustement aux données-traditionnellement mesuré.

-produire une bonne courbe c'est-à-dire sans trop de variation locale rapide.

La courbe de régression $\widehat{m}_\lambda(x)$ est obtenue en minimisant la somme pénalisée des carrés

$$S_\lambda(m) = \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 + \lambda \int_a^b \left\{m''(x)\right\}^2 dx.$$

Où erreur de fonction deux fois différenciable sur $[a, b]$, et λ représente le taux de change entre l'erreur résiduelle et la rugosité de la courbe m.

Une variante du calcul des splines consiste à résoudre le problème équivalent

$\min_m \int \left|m''(x)\right|^2 dx$ soumis à $\sum_{i=1}^{n} (Y_i - m(X_i))^2 \leq \Delta$.

Les paramètres λ et Δ ont des significations similaires et sont liés par la relation

$\lambda := - \mid G'(\Delta) \mid^{-1}$ où $G(\Delta) := \int \left(\widehat{m}''_\Delta(x)\right)^2 dx$ et $\widehat{m}_\Delta(x)$ résout le problème ci-dessus.

2.1.8 Méthodes semi-paramétriques (MSP)

Un modèle est appelé semi-paramétriques s'il est décrit par θ et τ où θ est fini-dimensionnel (paramétrique) et θ est infini-dimensionnel (non paramétrique).

Tous les modèles de condition de moment sont semi-paramétriques en ce sens que la distribution des données (τ) est dimensionnelle non spécifiée et infini. Mais les paramètres plus typiquement appelés semi-paramétriques sont ceux où il y a une estimation explicite de τ.

Dans des nombreux contextes, la partie τ non paramétrique est une moyenne conditionnelle, une variance, une densité ou une fonction de distribution.

Souvent θ est le paramètre d'intérêt, et est un paramètre de nuisance, mais ce n'est pas nécessairement le cas.

Distribution asymptotique

D'après le document MINPIN d'Andrew (1994)[3]. Paramètre : $\widehat{\theta}$ minimise une fonction de critère, $Q_N(\theta, \widehat{\tau})$, qui dépend d'un estimateur de paramètre de Nuisance dimensionnel Infini Préliminaire.

$\Longrightarrow \widehat{\theta}$ est un estimateur en deux étapes.

La dérivation habituelle des distributions asymptotiques élargit l'équation $m(\theta, \tau) = 0$. Nous pouvons le faire pour θ, mais pas pour τ (il est dimensionnel infini).

Pour procéder, Andrews utilise une hypothèse. Maintenant, nous travaillons avec la version population de $m(\theta, \tau) = \mathbf{E}[m_i(\theta, \tau)]$ et étudions la convergence de

$$\begin{aligned} v_n(\tau) &= \sqrt{n}\left(\overline{m}_n(\theta_0, \tau) - m(\theta_0, \tau)\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_i(\theta_0, \tau) - \mathbf{E}[m_i(\theta_0, \tau)]\right). \end{aligned}$$

Dans un grand nombre d'hypothèses : $\widehat{\theta}$ et $\widehat{\tau}(x) \xrightarrow{P} \theta_0$ et τ_0 égal à 0 à (θ_0, τ_0) c'est-à-dire, condition d'identification-convergence de l'équation $m(\theta, \tau) = 0$ (fluidité des fonctions sous-jacentes ; et existence de moments),

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{\mathcal{D}} N(0, V),$$

où

$$\begin{aligned} V &= M^{-1}\Omega M^{-1'}. \\ M &= \mathbf{E}\frac{\partial}{\partial \theta'} m_i(\theta_0, \tau_0). \\ \Omega &= \mathbf{E} m_i(\theta_0, \tau_0)\, m_i(\theta_0, \tau)'. \end{aligned}$$

Modèle de régression partiellement linéaire

Il est facile de définir un modèle de régression ‹partiellement linéaire› :

$y_i = m_z(z_i) + x_i'\beta + \varepsilon_i \quad (\dim(Z) = Nxq)$.

$\mathbf{E}[\varepsilon_i / X_i Z_i] = 0$.

$\mathbf{E}[\varepsilon_i^2 / X_i = x, Z_i = z] = \sigma^2(x, z)$.

-Les régresseurs sont (X, Z).

-La moyenne conditionnelle est linéaire en X_i, mais peut-être non linéaire en Z_i.

-Les variables factices sont habituellement minces dans le vecteur X.

-Pour garder les choses simples, nous supposons une seule variable non linéaire : $q = 1$.

Objectif : Estimer β et $m_z(\cdot)$; et obtenir le IC

Enjeux : Identification, distribution des estimations.

Estimation

Robinson (Econometrica, 1988) montre que nous pouvons nous concentrer sur $m_z(z_i)$ en utilisant une généalogie de régression résiduelle. Commencez par : $y_i = m_z(z_i) + x_i'\beta + \varepsilon_i$ $(\dim(Z) = Nxq)$

Prise en compte des attentes conditionnelles n Z :

$$\begin{aligned}\mathbf{E}[y_i/z_i] &= \mathbf{E}[m_z(z_i)/z_i] + \mathbf{E}[x_i'\beta/z_i] \\ &= m_z(z_i) + \mathbf{E}[x_i'/z_i]\beta.\end{aligned}$$

-Deux moyens conditionnels : $m_y(z_i) = \mathbf{E}[y_i/z_i]$ et $m_x(z_i) = \mathbf{E}[x_i'/z_i]$ alors

$$m_y(z_i) = m_z(z_i) + m_x(z_i)'\beta.$$

Soustraire de l'équation originale ($m_z(z_i)$ disparaît) :

$$y_i - m_y(z_i) = \left[x_i^{'} - m_x(z_i)^{'}\right]\beta + \varepsilon_i.$$

Réécrire en termes de résidus : $y_i - m_y(z_i) = \left[x_i^{'} - m_x(z_i)^{'}\right]\beta + \varepsilon_i$

$$\begin{aligned}
\varepsilon_{yi} &= y_i - m_y(z_i). \\
\varepsilon_{xi} &= \left[x_i^{'} - m_x(z_i)^{'}\right]. \\
\varepsilon_{yi} &= \varepsilon_{xi}^{'}\beta + \varepsilon_i.
\end{aligned}$$

C'est-à-dire, β est le coefficient de la régression de ε_{yi} sur ε_{xi} Mais, nous n'observons pas les erreurs. C'est un estimateur MC irréalisable !

Robinson suggère les tapes suivants :

1. Estimer $m_y(z_i)$ et $m_x(z_i)$ par régression NW (différents $h^{'}s$, OK).

2. Obtenir les résidus, ε_{xi} et ε_{yi}.

3. À l'aide des résidus, faire une MCO pour estimer β.

Remarque : Nous pouvons utiliser *1* LL ou NW pondéré.

Coupe Les estimations de régression non paramétrique dépendent inversement de $\widehat{\mathbf{f}}_z(z)$.

Problème : Pour les valeurs de z où $\mathbf{f}_z(z)$. Est proche de $0, \widehat{\mathbf{f}}_z(z)$ n'est pas délimité par 0. Les estimations NW à ce stade peuvent être mauvaises.

Solution : coupe.

Soit $b > 0$ une constante de découpage. L'estimateur tronqué de β est :

$$\widehat{\beta} = \left(\sum_i \varepsilon_{xi} \varepsilon_{xi}^{'} I \left[\widehat{\mathbf{f}}_z (z) \geq 0 \right] \right)^{-1} \sum_i \varepsilon_{xi} \varepsilon_{yi} I \left[\widehat{\mathbf{f}}_z (z) \geq 0 \right]$$

$\Longrightarrow$Il s'agit d'une régression résiduelle MC réduite.

La théorie asymptotique exige que $b = b_N \rightarrow 0$, mais il n'est pas clair comment sélectionner la pratique b. Souvent, le découpage est ignoré dans les applications.

Suggestion : Modèle d'estimation avec et sans découpage.

Estimation de la partie non paramétrique

Le modèle $y_i = m_z (z_i) + x_i^{'} \beta + \varepsilon_i \quad (\dim (Z) = Nxq)$.

Nous avons estimé β. Maintenant, nous voulons estimer $m_z (z_i)$. Il semble qu'un algorithme itératif soit nécessaire, mais puisque β converge plus vite que le taux non paramétrique, nous pouvons prétendre qu'il est fixe. Ensuite,

$$\widehat{m}_z (z_0) = \frac{\sum_{i=1}^{N} K_h \left(\frac{z_0 - z_i}{h} \right) \left(y_i - X_i^{'} \widehat{\beta} \right)}{\sum_{i=1}^{N} K_h \left(\frac{z_0 - z_i}{h} \right)}.$$

Choix de fenêtre

Dans un contexte semi-paramétrique, il est important d'étudier l'effet d'une fenêtre sur la performance de l'estimateur d'intérêt avant de déterminer le paramètre de lissage.

Dans de nombreux cas, cela nécessite un débit le paramètre de lissage non conventionnel.

Toutefois, ce problème ne se pose pas dans les modèles partiellement linéaires. Les largeurs de fenêtre de première-étape h utilisées pour $\widehat{m}_y (z_i)$ et $\widehat{m}_x (z_i)$ sont de entrées pour le calcul de $\widehat{\beta}$.

h impacte la théorie pour $\widehat{\beta}$, grâce aux taux de convergence uniformes pour $\widehat{m}_y(z_i)$ et $\widehat{m}_x(z_i)$, suggérant que nous utilisions des règles de fenêtre conventionnelles, par exemple CV.

Chapitre 3

Etude de simulation

Dans les deux sections suivantes, nous étudions par le biais des simulations la performance de l'estimateur à noyau de la densité de probabilité et celui de la régression, et ce en utilisant le logiciel d'analyse statistique R.

3.1 Simulation d'estimation à noyau de la densité

Nous considérons ici différents noyaux K à savoir ceux de l'Epanechnikov, Gaussien, Uniform et Biweight en fixant le parametrer de lissage fixé $h = h_n$ qui dépend d'une taille n fixée de l'échantillon. En d'autres termes, nous avons à faire aux trois situations suivantes :

1. Paramètre de lissage fixé, noyau Epanechnikov et n fixé.

2. Paramètre de lissage fixé, noyau Gaussien et n fixé.

3. Paramètre de lissage fixé, noyau Uniform et n fixé.

4. Paramètre de lissage fixé, noyau Biweight et n fixé.

3.1.1 Le code R du programme

```
rm(list=ls())#effacerlamemoire

v=100

set.seed(123)

X=rnorm(v)

Y=runif(v)

Z=rexp(v)

T=rpois(v, lambda = 1)

K<-function(t){((3/4))*(1-t^2)*ifelse((abs(t))<1,1,0)} #Noyau de Epanchnikov

#K<-function(t){(1/sqrt(2*pi))*exp((-1/2)*t^2)} #Noyau de Gaussien

#K<-function(t){(1/2)*ifelse((abs(t))<1,1,0)} #Noyau de Uniform

#K<-function(t){(15/16)*((1-x^2)^2)*ifelse((abs(t))<1,1,0)} #Noyau de Biweight

par(mfrow=c(2,2))

fnX<-function(x,X){sum(K((v[1 :X]-x)/(2.34*sd(v)*X^(-1/5))))

/(X*(2.34*sd(v)*X^(-1/5)))}

hist(X,ylab="densité",freq=F,main="Estimation de la densité à noyau (cas

gaussien)",col="red")

lines(density(X),type="l",col="green") #Densité de noyau

curve(dnorm(x),add=TRUE,col="blue") #Densité théorique

fnY<-function(x,Y){sum(K((v[1 :Y]-x)/(2.34*sd(v)*Y^(-1/5))))

/(Y*(2.34*sd(v)*Y^(-1/5)))}

hist(Y,ylab="densité",freq=F,main="Estimation de

 la densité à noyau (cas uniforme)",col="green")
```

```
lines(density(Y),type="l",col="purple")

curve(dunif(x),add=TRUE,col="blue")

fnZ<-function(x,Z){sum(K((v[1 :Z]-x)

/(2.34*sd(v)*Z^(-1/5))))/(Z*(2.34*sd(v)*Z^(-1/5)))}

hist(Z,ylab="densité",freq=F,main="Estimation

de la densité à noyau (cas exepentielle)",col="pink")

lines(density(Z),type="l",col="black")

curve(dexp(x),add=TRUE,col="green")

fnT<-function(x,T){sum(K((v[1 :T]-x)/(2.34*sd(v)*T^(-1/5))))

/(T*(2.34*sd(v)*T^(-1/5)))}

hist(T,ylab="densité",freq=F,main="Estimation

de la densité à noyau (cas Poisson) ",col="blue")

lines(density(T),type="l",col="yellow")

curve(dpois(x,1),lamda=1,add=TRUE,col="brown")
```

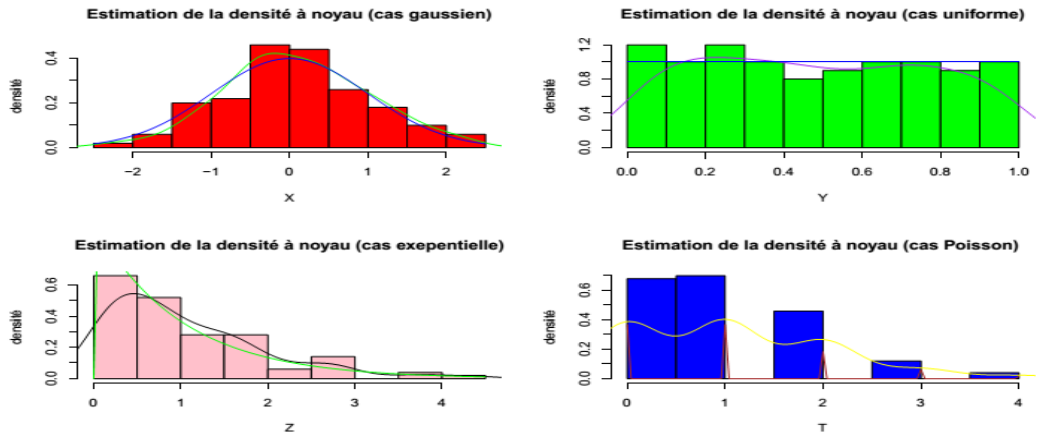

FIG. 3.1 – Histogramme, estimateur à noyau (package R, par défaut) estimateur à noyau (Epanechnikov), et la densité théorique

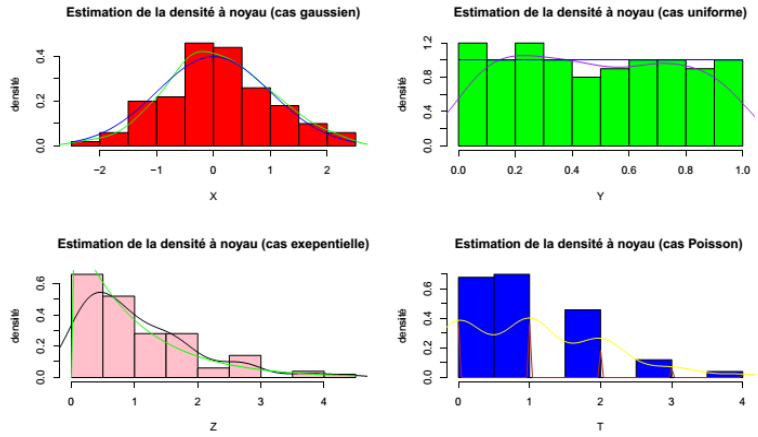

FIG. 3.2 – Histogramme, estimateur à noyau (package R, par défaut) estimateur à noyau (Gaussien), et la densité théorique

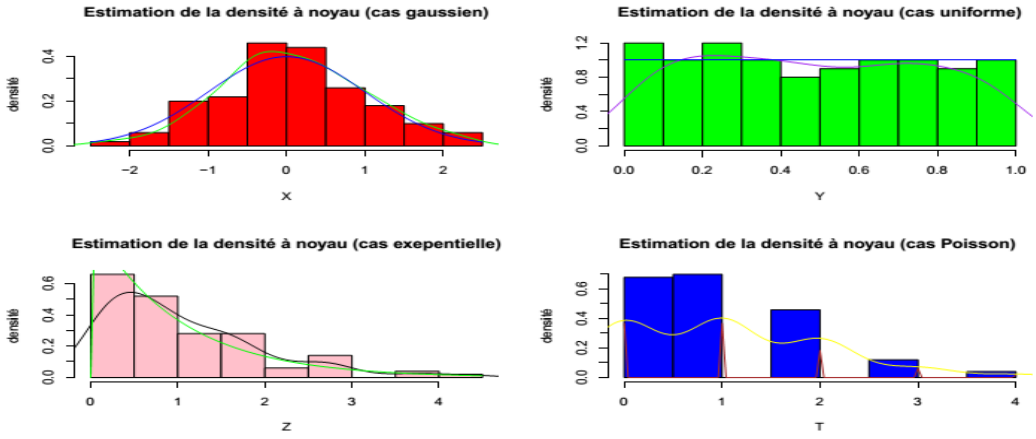

FIG. 3.3 – Histogramme, estimateur à noyau (package R, par défaut) estimateur à noyau (Uniform), et la densité théorique

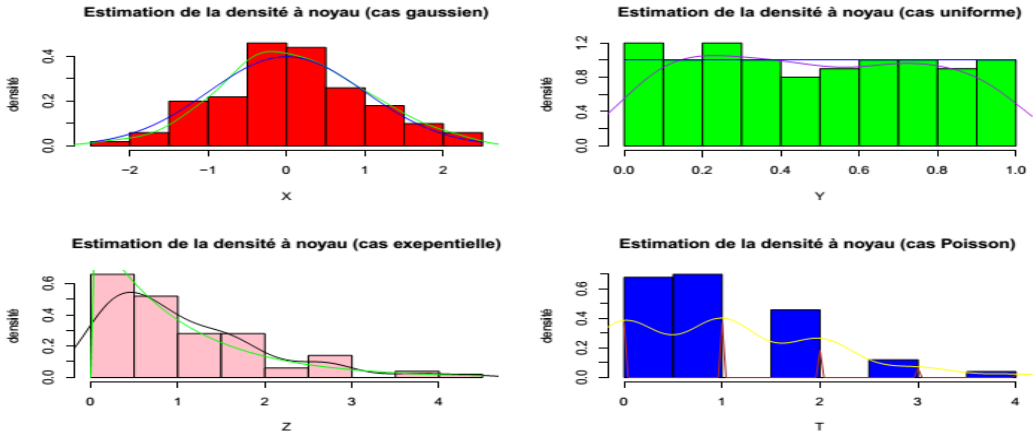

FIG. 3.4 – Histogramme, estimateur à noyau (package R, par défaut) estimateur à noyau (Bewieght), et la densité théorique

3.2 Application des donnes réelles

Aprés avoir validé les données de simulation, nous procéderons à l'applicationd'estimation de la densité f avec utilisant des données réelles (tableau ChickWeight)

3.2.1 Application d'estimation à noyau de la densité

```
rm(list=ls())#effacer la memoire

data("ChickWeight")

T=ChickWeight

X=T$weight

Y=T$Time

Z=Y

T=X

set.seed(123)

N=length(X)
```

```
K<-function(t){((3/4))*(1-t^2)*ifelse((abs(t))<1,1,0)}#Epanechnikov
#K<-function(t){(1/sqrt(2*pi))*exp((-1/2)*t^2)} #Noyau de Gaussien
#K<-function(t){(1/2)*ifelse((abs(t))<1,1,0)} #Noyau de Uniform
#K<-function(t){(15/16)*((1-x^2)^2)*ifelse((abs(t))<1,1,0)} #Noyau de Biweight
h=2.34*sd(X)*N^(-1/5)
fn<-function(x){sum(K((X[1 :N]-x)/h))/(N*h)}
par(mfrow=c(2,2))
hist(X,ylab="densité",freq=F,main="Estimation
de la densité à noyau Epanechnikov",col="red")
lines(density(X),type="l",col="green") #Densité de noyau
curve(dnorm(x),add=TRUE,col="blue") #Densité théorique
hist(Y,ylab="densité",freq=F,main="Estimation
de la densité à noyau cas Gaussien",col="green")
lines(density(Y),type="l",col="purple")
curve(dunif(x),add=TRUE,col="blue")
hist(Z,ylab="densité",freq=F,main="Estimation
de la densité à noyau cas Uniform",col="pink")
lines(density(Z),type="l",col="black")
curve(dexp(x),add=TRUE,col="green")
hist(T,ylab="densité",freq=F,main="Estimation
de la densité à noyau Biweight ",col="blue")
lines(density(T),type="l",col="yellow")
curve(dpois(x,1),add=TRUE,col="brown")
```

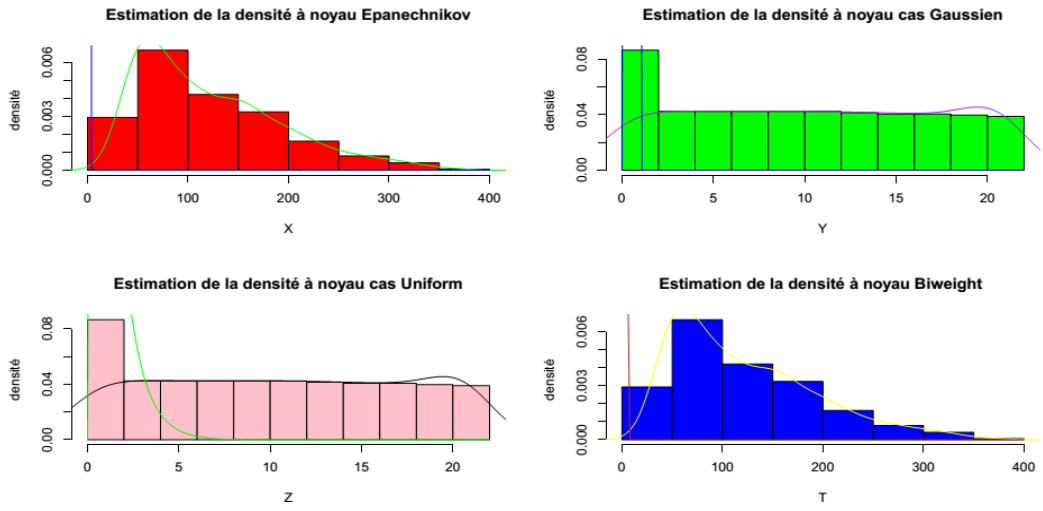

FIG. 3.5 – Differents estimateurs de la densité

3.3 Simulation de régression à noyau

Une estimation de la régression linéaire simple avec utilisant des données dans la banque mondiale, où nous avons pris deux échantillons sous la forme d'un tableau d'Excel CFE pour raccourci ≪Chômage, femmes (% de la population active féminine) (estimation modélisée OIT)≫, et CJH pour raccourci ≪Chômage, jeunes hommes (% de la population active masculine de 15 à 24 ans) (estimation modélisée OIT)≫,

3.3.1 Application

```
####La régression linéaire simple paramétrique####

library(readxl)

Classeur1 <-read_excel("C :/Users/elathir/Desktop/Classeur1.xlsx")

print(Classeur1)

attach(Classeur1)#Rend visible les vecteurs constituant Classeur1
```

```
#La reggression

Model<-lm(formula=CFE~CJH,data=Classeur1)

summary(Model)#Commande permettant de stocker la regreesion effectuée dans

Model

confint.default(Model)

attributes(Model)

Model$coefficients

Model$residuals

library(car)

library(carData)

scatterplot(CFE~CJH,data=Classeur1,xlab="CJH",ylab="CFE",

main="regression de CJH sur CFE",regLine=TRUE,ellipse=FALSE,smooth=FALSE,

grid=TRUE)
```

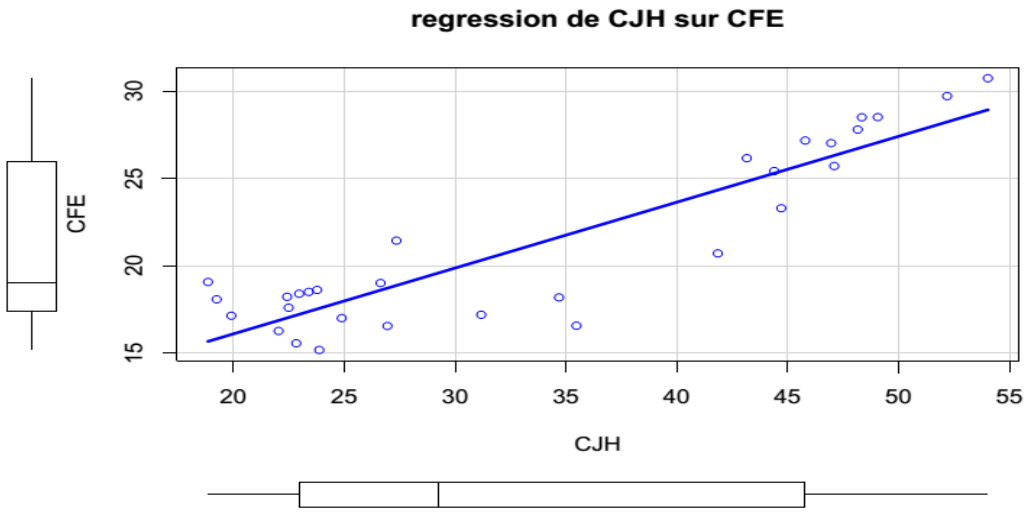

FIG. 3.6 – La régression linéaire simple, chômage femmes(% de la population active féminine) et juennes hommes (% de la population active masculine de 15 à 24 ans)

3.3.2 Réression à noyau non paramétrique

Nous avons présenté les résultats obtenus pour la régression à noyau non paramétrique (Dans la figure correspondante)

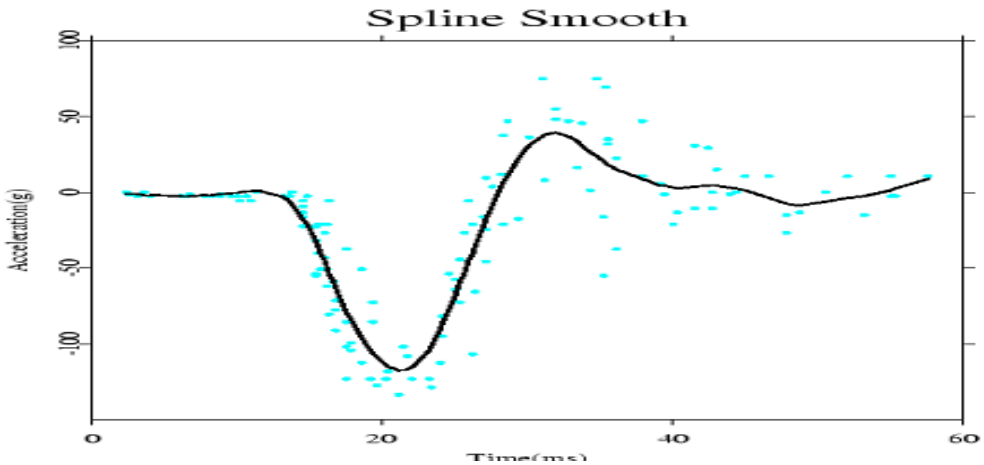

FIG. 3.7 – Spline lisse (jeu de donnéees moto). Tiré de Härdle(1990)

Conclusion

Nous avons discuté deux estimateurs non-paramétriques a savoir l'estimateur de la densité et de la régression à noyau. Les avantages des ces deux estimateurs ce résument en deux points :

- Aucune paramétrisation, à priori, sur la densité est exigée, contrairement à l'estimation paramétrique ;
- Les deux estimateurs sont des fonctions indéfiniment dérivables ; contrairement aux estimateurs basés sur l'histogramme.

Cette méthode d'estimation est basée sur la fonction noyau K et le paramètre de lissage h. Nous avons remarqué la qualité de l'estimation ne dépend pas de K mais plutôt de h. Plusieurs méthodes d'estimation de ce paramètre sont proposées dans littératures ; la plus utilise jusqu'a ce jour est celle de la validation croisée.

Bibliographie

[1] Cleveland, William S. Robust loccally weighted regression and smoothing scatterplots. J.Amer. Statist. Assoc. 74(1979), no. 368,829- -836.

[2] Devroye, L. and Györfi, L. (1985). Nonparametric Density Estimation : The L_1 Veiw. New York : Wiley.

[3] EconometricabVol.62,No.6(Nov., 1994).

[4] Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. Theor. Probab. Appl., 14, 153-158.

[5] Fix, E. and Hodges, J.L.(1951). Discriminatory analysis, nonparametric estimation : consistency properties. Report No. 4, Project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

[6] Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation, Ann. Statist., 11, 1156-1174.

[7] Nadaraya, E.A. (1965).On nonparametric estimates of density functions and regression curves. Theor. Probab. Appl., 10, 186-190.

[8] Newey,Whitney K. convergence rates and asymptotic normality for series estimators. J.Econometrics 79(1997), no.1, 147- -168.

[9] Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statis., 33, 1065-1076.

[10] Parzen, E. (1979). Nonparemtric statistical data modeling. J. Amer. Statist. Assoc., 74, 105-131.

[11] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., 27, 832-837.

[12] Silverman, B.W. (1986). Monographs on statistics and applied probability.Density estimation for statistics and data analysis, 26.

[13] Watson, G.S. and Wright, I.W. (1983). Hazard analysis I. Biometrika, 51, 175-184.

[14] Wolfgang Härdle Humboldt-Universität zu Berlin Wirtschaftswissenschaftliche Fakultät Instiotut für Statistik und Ökonometrie Spandauer Str. D-10178 Berlin 1994

[15] Härdle, Wolfgang Applied nonparametric regression. Econometric Society Monographs, 19. Cambridge Universityn Press, Combridge, 1990.

Résumé :

Ce mémoire porte sur l'estimation non paramétrique de la densité de probabilité et de la régression. Nous avons considéré l'estimateur à noyau de Parzen-Rosenblatt (pour la densité) et l'estimateur à noyau de Nadaraya-Watson (pour la régression). Une étude sur les propriétés asymptotiques, telles que la consistance et la normalité de ces derniers, sont données. Des simulations étudient la performance de ses deux estimateurs sont illustrées. Une application basée sur des données réelles est présentée.

Mots-clés : **Estimation non paramétrique, estimateur à noyau.**

ملخص:

تركز هذه المذكرة عن التقدير اللاوسيطي لكثافة الاحتمالات و الانحدار. اعتبرنا هذا المقدر ذو نواة ل Parzen-Rosenblatt (الكثافة) و مقدر ذو نواة ل Nadaraya-Watson (الانحدار) . درسنا الخصائص التقاربية كالمتانة و النظامية لهذان الاخيران. انجزنا محاكاة تدرس دقة هذان المقدران، مع تطبيقات مبنية على معطيات حقيقية.

***كلمات مفتاحية*: التقدير اللاوسيطي، قدر ذو نواة**

Abstract :

This memory deals with the nonparametric estimation of the probability density and regression. We considered the Parzen-Rosenblatt kernel estimator (for density) and the Nadaraya-Watson kernel estimator (for regression). A study of the asymptotic properties, such as the consistency and normality of these ones, are given. A simulation study is carried out to evaluate the finite sample behavior of the two estimators. An application based on real data is presented.

Key words : **nonparametric estimation, kernel estimators.**