Chapitre 1 : Estimateur à noyau pour Variables réelles

A - Historique

On se limite ici au cas de l'estimation de la densité (une évolution similaire s'est faite pour l'estimateur de la fonction de régression à partir du régressogramme)

• **Modèle paramétrique :**

On suppose $f_X = f_\theta$ connue à un paramètre $\theta \in \mathbb{R}^p$ près et on estime f_X en remplaçant θ par son estimateur :
$\forall x \in \mathbb{R}, \hat{f}_X(x) = f_{\hat{\theta}}(x)$

• **Viennent ensuite les modèles non-paramétriques.**

* Un premier estimateur non paramétrique est l'histogramme. Il consiste à estimer la densité au travers d'une fonction constante par morceaux (correspondant aux classes). Sur chaque classe, la densité f_X est estimée par la densité associée à cette classe (concentration des observations au sein de la classe).

Cet estimateur très rudimentaire présente un assez grand nombre de défauts.

↳ il est discontinu et constant par morceaux

↳ il dépend fortement du choix des classes (leurs nombres et leurs bornes).

$F_X(x) = \mathbb{P}(X \leq x)$

$X_1, \dots, X_n \overset{iid}{\sim} X$

Comment estimer F_X de manière non-paramétrique ?

Remarque : $\forall x \in \mathbb{R}$, $Z_x = \mathbb{1}_{X \leq x}$, on a $\forall x \in \mathbb{R}$, $F_X(x) = \mathbb{E}[Z_x]$

Donc, on pose naturellement, avec la méthode des moments,
$\forall x \in \mathbb{R}$, $\hat{F}_x(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$

Propriétés : 1 - $\hat{F}_x(x)$ est un estimateur sans biais de $F_x(x)$

2 - $\forall x \in \mathbb{R}$, $\mathrm{var}(\hat{F}_x(x)) = \frac{F(x)(1-F(x))}{n}$

3 - $\forall x \in \mathbb{R}$, $\hat{F}_x(x) \xrightarrow[n \to +\infty]{L^2} F_x(x)$

4 - $\forall x \in \mathbb{R}$, $\hat{F}_x(x) \xrightarrow[n \to +\infty]{p.s.} F_x(x)$ (LFGN).

5 - $\forall x \in \mathbb{R}$, $\sqrt{n}(\hat{F}_x(x) - F_x(x)) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, F_x(x)(1-F_x(x)))$

Démonstration :

1 - $\mathbb{E}[\hat{F}_x(x)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}_{X_i \leq x}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(X_1 \leq x) = \mathbb{P}(X_1 \leq x) = F_X(x)$

2 - $\mathrm{var}(\hat{F}_x(x)) = \frac{1}{n^2} \mathrm{var}\left(\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}\right) \overset{\perp\!\!\!\perp}{=} \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}(\mathbb{1}_{X_1 \leq x})$

$= \frac{1}{n} \mathrm{var}(X_1) = \frac{F(x)(1-F(x))}{n}$ car $X_1 \sim \mathcal{B}(F(x))$

3 - $\mathbb{E}[\hat{F}_x(x)^2] = \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}\right] + \frac{2}{n^2} \mathbb{E}\left[\sum_{i=2}^{n} \sum_{j=1}^{i-1} \mathbb{1}_{X_i \leq x} \mathbb{1}_{X_j \leq x}\right]$

$= \frac{1}{n} F(x) + \frac{2}{n^2} F(x)^2 \sum_{i=1}^{n-1} i = \frac{F(x)}{n} + \frac{n-1}{n} F(x)^2$

D'où :

$\mathbb{E}[|\hat{F}_x(x) - F(x)|^2] = \mathbb{E}[\hat{F}_x(x)^2] - 2\mathbb{E}[\hat{F}_x(x)] F(x) + F(x)^2$

$= \frac{F(x)}{n} + \frac{n-1}{n} F(x)^2 - F(x)^2$

$= \frac{1}{n}(F(x) - F(x)^2) \xrightarrow[n \to +\infty]{} 0.$

D'où $\hat{F}_x(x) \xrightarrow[n \to +\infty]{L^2} F(x)$ pour tout $x \in \mathbb{R}$.

4 - LGFN et 5 - TCL.

b- Estimateur de la densité par fenêtre mobile.

Tout part du fait que $f(x) = \lim_{h \to 0} \frac{F_X(x+h) - F_X(x-h)}{2h}$.

En effet, $\frac{F_X(x+h) - F_X(x-h)}{2h} = \frac{F_X(x+h) - F_X(x) + F_X(x) - F_X(x-h)}{2h}$.

$$= \frac{1}{2} \frac{F_X(x+h) - F_X(x)}{h} + \frac{1}{2} \frac{F_X(x) - F_X(x-h)}{h}.$$

$$\xrightarrow[h \to 0]{} \frac{1}{2}\left(f(x) + f(x)\right) = f(x).$$

Il suffit donc de prendre h assez petit et d'estimer F_X (via l'estimateur précédent) pour avoir un estimateur de f_X. On a alors :

$$\forall h > 0, \quad \tilde{f}_X(x) = \frac{\hat{F}_X(x+h) - \hat{F}_X(x-h)}{2h} = \frac{1}{2hn} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x+h} - \mathbb{1}_{X_i \leq x-h}.$$

$$= \frac{1}{2hn} \sum_{i=1}^{n} \mathbb{1}_{x-h < X_i \leq x+h}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{X_i - x}{h}\right) \quad \text{où} \quad \forall t \in \mathbb{R}, \; w(t) = \frac{1}{2} \mathbb{1}_{]-1;1]}(t)$$

Définition : h est appelé paramètre de lissage (cf. exemple)

Exemple : On prend $x = (2, 3, 3.5, 6, 7.5)$. On choisit $h = 1$.

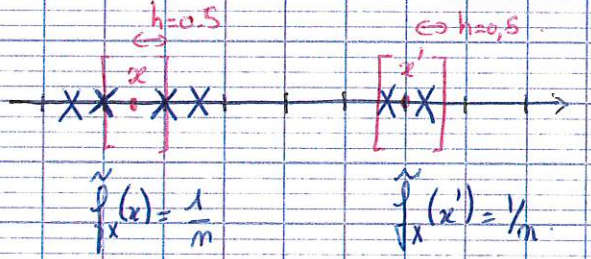
Remarque : Comme X est continue, $\mathbb{P}(X_i = x_0) = 0$ pour tout $x_0 \in \mathbb{R}$ et donc on considère souvent de manière équivalente l'estimateur par fenêtre mobile :

$$\tilde{f}'_X(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} \mathbb{1}_{[-1;1]}\left(\frac{X_i - x}{h}\right) \overset{p.s.}{=} \tilde{f}_X(x)$$

Exemple : On reprend les mêmes données

Exemple : Retour aux données calculés galaxies
Comparer les estimateurs par fenêtres mobiles pour des valeurs de $h = 0{,}1,\ 10,\ 100,\ 1000,\ 10\,000,\ 100\,000,$
Choisir le paramètre de lissage h. Choisir le plus adapté. Quels défauts reste-t-il à gommer ?

D'après nos expérimentations, on remarque :
* Avantage : • ne dépend plus du choix des classes (leurs nombres, leurs bornes) mais simplement des valeurs autour de x (et donc du choix de h
* Défaut : • discontinu, continu par morceaux
• même poids donné à tous les points se trouvant dans l'intervalle $[x-h ; x+h]$



$\hat{f}_X(x) = \frac{1}{n}$ $\quad$ $\hat{f}_X(x') = \frac{1}{n}$

→ on voudrait pondérer en fonction de la distance entre X_i et x.

Rappel : Propriété théorique d'un estimateur
A - Estimateur Ponctuel
1 - Biais d'un estimateur $\hat{f}$ de f au point x :
$$\text{Biais}_f(\hat{f}(x)) = \mathbb{E}_f[\hat{f}(x)] - f(x)$$
Un estimateur est dit (asymptotiquement) sans biais si son biais est 0 (tend vers 0 quand $n \to +\infty$).

2 - Variance d'un estimateur $\hat{f}$ de f au point x
$$\text{var}_f(\hat{f}(x)) = \mathbb{E}_f\left[\left(\hat{f}(x) - \mathbb{E}_f[\hat{f}(x)]\right)^2\right]$$

3 - Erreur quadratique moyenne (Mean Squared Error)
$$\text{MSE}(\hat{f}(x)) = \mathbb{E}_f\left[(\hat{f}(x) - f(x))^2\right]$$
$$= \text{var}(\hat{f}(x)) + (\text{Biais}_f(\hat{f}(x)))^2$$

Un estimateur est dit consistant en moyenne quadratique si son MSE tend vers 0 lorsque $n \to +\infty$.

B - Globales

1 - Integrated Mean Squared Error (ou Mean Integrated Squared Error)

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

on peut inverser $\int$ et $\mathbb{E}$ par Fubini-Tonelli d'où les deux noms équivalents.

2 - Integrated Squared Error

$$ISE(\hat{f}) = \int (\hat{f}(x) - f(x))^2 dx$$

3 - Mean Sum of Squared Error

$$MSSE(\hat{f}) = \mathbb{E}\left[\sum_{j=1}^{n} (\hat{f}(x_j) - f(x_j))^2\right]$$

(*) on remplace parfois les x_j par des points déterministes t_i formant une grille de discrétisation $t = (t_1, \dots, t_p)$.

4 - Sum of Squared Error

$$SSE(\hat{f}) = \sum_{j=1}^{n} (\hat{f}(x_j) - f(x_j))^2$$

Proposition : Propriété de l'estimateur de la densité par fenêtre mobile.

1 - Si $h = h_n \xrightarrow[n \to +\infty]{} 0$ alors, $\forall x \in \mathcal{D}_{f_X}$, $\tilde{f}_X(x)$ est un estimateur sans biais de $f_X(x)$.

2 - Si $h = h_n \xrightarrow[n \to +\infty]{} 0$ et $n h_n \xrightarrow[n \to +\infty]{} +\infty$. Soit $x \in \mathcal{D}_{f_X}$, si $f_X(x) > 0$ alors on a $\tilde{f}_X(x) \xrightarrow[n \to +\infty]{L^2} f_X(x)$

Démonstration :

1 - $$\mathbb{E}\left[\tilde{f}_X(x)\right] = \mathbb{E}\left[\frac{\hat{F}_X(x+h) - \hat{F}_X(x-h)}{2h}\right] = \frac{F_X(x+h) - F_X(x-h)}{2h} \xrightarrow[n \to +\infty]{} f_X(x)$$

2 - $$var(\tilde{f}_X(x)) = var\left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} \mathbb{1}_{[-1;1]}\left(\frac{X_i - x}{h}\right)\right)$$

$$\mathrm{var}\left(\tilde{f}_X(x)\right) = \frac{1}{4n^2h^2}\mathrm{var}\left(\sum_{i=1}^{n} \mathbb{1}_{[-1;1]}\left(\frac{X_i - x}{h}\right)\right)$$

$$\overset{\text{iid}}{=} \frac{1}{4n^2h^2}\sum_{i=1}^{n}\mathrm{var}\left(\mathbb{1}_{[-1;1]}\left(\frac{X-x}{h}\right)\right)$$

Or $\mathrm{var}\left(\mathbb{1}_{[-1;1]}\left(\frac{X-x}{h}\right)\right) = \mathbb{E}\left[\mathbb{1}_{[-1;1]}\left(\frac{X-x}{h}\right)\right] - \mathbb{E}\left[\mathbb{1}_{[-1;1]}\left(\frac{X-x}{h}\right)\right]^2$

$$= \mathbb{P}\left(-1 \le \frac{X-x}{h} \le 1\right) - \left[\mathbb{P}\left(-1 \le \frac{X-x}{h} \le 1\right)\right]^2$$

$$= \mathbb{P}(x-h \le X \le x+h) - \mathbb{P}(x-h \le X \le x+h)^2$$

$$= F_X(x+h) - F_X(x-h) - \left(F_X(x+h) - F_X(x-h)\right)^2$$

$$\mathrm{var}\left(\tilde{f}_X(x)\right) = \frac{1}{4nh^2}\left(F_X(x+h) - F_X(x-h)\right)\left(1 - \left(F_X(x+h) - F_X(x-h)\right)\right)$$

$$= \frac{1}{2nh}\left(\frac{F_X(x+h) - F_X(x-h)}{2h}\right)\left(1 - 2h\,\frac{F_X(x+h) - F_X(x-h)}{2h}\right)$$

$$\underset{n\to+\infty}{\sim} \frac{1}{2nh} f_X(x) \xrightarrow[n\to+\infty]{} 0 \quad \text{car } nh_n \xrightarrow[n\to+\infty]{} +\infty$$

Cela implique directement que

$$MSE\left(\tilde{f}(x)\right) = \mathrm{var}\left(\tilde{f}_X(x)\right) + \mathrm{Biais}\left(\tilde{f}(x)\right)^2 \xrightarrow[n\to+\infty]{} 0$$

Remarque : A n fixé, on a intérêt à prendre :
- h petit pour rendre le terme de biais petit
- h grand pour rendre le terme de variance grand.

Il faudra donc trouver un équilibre entre le biais et la variance. Prendre h trop petit revient à avoir un surajustement aux données (biais faible et grande variance). Estimateur proche de 0 sauf près des observations où l'on observe des pics. Prendre h trop grand revient à avoir un sous-ajustement aux données (biais fort et variance faible). Estimation de la densité

trop applati, les variations de la densité sont sous-représentées

B - Estimateur à noyaux lissés de la fonction de densité

I - Définition

Idée : On veut remplacer l'indicatrice ω par une fonction plus lisse k, appelée noyau possédant les propriétés suivantes :

- $\forall x \in \mathbb{R}, \; k(x) \geq 0$ (positivité)
- $\int_{\mathbb{R}} k(x)\,dx = 1$ (densité)
- $\int_{\mathbb{R}} x\,k(x)\,dx = 0$ (symétrie)
- k maximale en 0 et $k(x) \xrightarrow[|x| \to +\infty]{} 0$ en décroissant

Définition : Estimateur à noyau de la densité : Parzen (1964)

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

où • k est la fonction définissant le poids de chaque observations (appelée noyau, en anglais "kernel").

• h est le paramètre de lissage (en anglais bandwith ou smoothing parameter).

Exemple : 1er exemple de noyau : $k(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

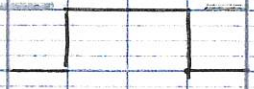
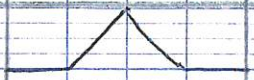
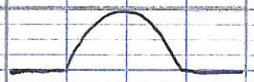
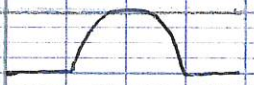
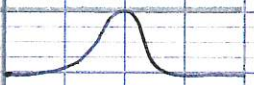
Dans ce cas précis,

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} e^{-\frac{(X_i - x)^2}{2h^2}} = \frac{1}{\sqrt{2\pi}h^2} \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{(x - X_i)^2}{2h^2}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi}h} e^{-\frac{(X_i - x)^2}{2h^2}}}_{f_{\mathcal{N}(X_i, h^2)}}$$

(voir exemple en code R).

Quelques exemples de noyau

noyau K	$K(t) =$	
Uniforme	$\frac{1}{2} \mathbb{1}_{\{\lvert t\rvert \leq 1\}}$	
Triangulaire	$(1-\lvert t\rvert)\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$	
Epanechnikov	$\frac{3}{4}(1-t^2)\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$	
Quartic	$\frac{15}{16}(1-t^2)^2\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$	
Triweight	$\frac{35}{32}(1-t^2)^3\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$	
Gaussien	$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$	
Cosinus	$\frac{\pi}{4}\cos\left(\frac{t\pi}{2}\right)\mathbb{1}_{\{\lvert t\rvert \leq 1\}}$	

Remarques : 1- On choisit un noyau qui soit maximal en 0 et dont la valeur diminue lorsque l'on s'écarte de 0 afin de moduler l'importance données aux observations en fonction de leur proximité avec le point x où l'on veut estimer f_x. Plus une observation sera proche, plus son poids sera élevée (car $\frac{x_i - x}{h}$ petit), et réciproquement.

2- Le fait d'utiliser un noyau lisse permet d'obtenir un estimateur lisse de f_x qui n'est plus constant par morceaux (si K ne l'est pas)

3- h joue le même rôle que précédemment pour l'estimateur à fenêtre mobile. Prendre h trop grand fait que tous les $\frac{x_i - x}{h}$ sont proche de 0 et donc que

$$\mathbb{E}_f\left[\hat{f}_x(x)\right] = f_x(x) + \frac{h^2}{2} f_x''(x)\mu_2(k) + \int_{\mathbb{R}} k(t) R(x+th)\,dt$$

$$\Leftrightarrow \text{Biais}\left(\hat{f}_x(x)\right) = \frac{h^2}{2} f_x''(x)\mu_2(k) + \int_{\mathbb{R}} k(t) R(x+th)\,dt$$

Or $$\int_{\mathbb{R}} k(t) R(x+th)\,dt = h^2 \int_{\mathbb{R}} t^2 k(t) \underbrace{\frac{R(x+th)}{t^2h^2}}_{\to 0 \text{ car } th \to 0.} dt$$

On peut montrer que :
$$\int_{\mathbb{R}} k(t) R(x+th)\,dt = o(h^2)$$ (voir annexe 1 à la fin du cours)

D'où :
$$\text{Biais}\left(\hat{f}_x(x)\right) = \frac{h^2}{2} f_x''(x)\mu_2(k) + o(h^2).$$

2- $$\text{Var}\left(\hat{f}_x\right) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} k\left(\frac{X_i - x}{h}\right)\right)$$

$$\overset{\text{iid}}{=} \frac{1}{n^2}\sum_{i=1}^{n} \frac{1}{h^2} \text{var}\left(k\left(\frac{X-x}{h}\right)\right)$$

$$= \frac{1}{n} \text{var}\left(\frac{1}{h} k\left(\frac{X-x}{h}\right)\right)$$

On a :
$$\mathbb{E}_f\left[\frac{1}{h^2} k^2\left(\frac{X-x}{h}\right)\right] = \int_{-\infty}^{+\infty} \frac{1}{h^2} k^2\left(\frac{y-x}{h}\right) f_x(y)\,dy \qquad t = \frac{y-x}{h}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{h} k^2(t) f_x(x+th)\,dt$$

$$= \int_{-\infty}^{+\infty} \frac{1}{h} k^2(t) f_x(x)\,dt + \int_{-\infty}^{+\infty} \frac{1}{h} k^2(t)\left(f_x(x+th) - f_x(x)\right) dt$$

$$= \frac{1}{h} f_x(x) R(k) + \int_{-\infty}^{+\infty} \frac{1}{h} k^2(t)\left(f_x(x+th) - f_x(x)\right) dt$$

D'après convergence dominée : $\int_{-\infty}^{+\infty} \frac{1}{h} k^2(t)\left(f_x(x+th) - f_x(x)\right) dt = o($ (voir annexe 2 à la fin du cours).

D'où :
$$\mathbb{E}_f\left[\frac{1}{h^2} k^2\left(\frac{X-x}{h}\right)\right] = \frac{1}{h} f_x(x) R(k) + \frac{1}{h} o(1).$$

D'où $\mathrm{var}_f(\hat{f}_x(x)) = \frac{1}{n}\left(\frac{1}{h} f_x(x) R(k) + \frac{1}{h} o(1) - \left(f_x(x) + \frac{h^2}{2} f_x''(x) \mu_2(k) + o(h^2)\right)^2\right)$,

$$\mathrm{var}_f(\hat{f}_x(x)) = \frac{1}{n}\left[\frac{1}{h} f_x(x) R(k) + o\left(\frac{1}{h}\right) - \left(f_x(x) + O(h^2)\right)^2\right]$$

$$= \frac{1}{nh} f_x(x) R(k) + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n}\right) \quad \text{mais } O\left(\frac{1}{n}\right) = o\left(\frac{1}{nh}\right)$$

$$= \frac{1}{nh} f_x(x) R(k) + o\left(\frac{1}{nh}\right)$$

Donc, on déduis aisement :

$$MSE(\hat{f}_x(x)) = \frac{h^4}{4} (f_x''(x))^2 \mu_2^2(k) + o(h^4) + \frac{1}{nh} f_x(x) R(k) + o\left(\frac{1}{nh}\right)$$

Remarque : Si $h = h_n \xrightarrow[n \to +\infty]{} 0$ et $nh_n \xrightarrow[n \to +\infty]{} +\infty$ alors on a $MSE(\hat{f}_x(x)) \xrightarrow[n \to +\infty]{} 0$

Propriété : Convergence presque complète

Hypothèse : * $S \subset \mathbb{R}$, S compact

* f_x C^2 au voisinage de x, (resp. au voisinage de S)

* k a support compact, symetrique, de carré intégrable, d'intégrale 1 et bornée (et éventuellement $|k(x) - k(y)| < C|x-y|^2$, $\forall x, y \in S$)

* $f_x(x) > 0$ (resp. $\inf_{x \in S} f_x(x) > 0$)

* $h = h_n \xrightarrow[n \to +\infty]{} 0$ et $\frac{n}{\ln(n)} \xrightarrow[n \to +\infty]{} +\infty$

Si les hypothèses sont valides, alors

1 - $\hat{f}_x(x) - f_x(x) = \Theta_{p.com}(h^2) + \Theta_{p.com}\left(\sqrt{\frac{\ln(n)}{nh}}\right)$

2 - $\sup_{x \in S} |\hat{f}_x(x) - f_x(x)| = \Theta_{p.com}(h^2) + \Theta_{p.com}\left(\sqrt{\frac{\ln(n)}{nh}}\right)$

Démonstration : admise

tous les points auraient des poids similaires (→ estimateur aplatie de la densité). On parle de "sulissage". Si au contraire h est très petit, $\frac{x_i - x}{h}$ sera très grand sauf quand x_i est très proche de x (on aurait donc encore un estimateur proche de 0 sauf autour des observations où l'on observerait des pics). On parle de "sous lissage"

Exemple : Retour aux données galaxies

Essayer les différents noyaux. On choisira comme paramètre de lissage celui qui est le plus approprié pour la fenêtre mobile. Ensuite faire varier à noyau fixé le paramètre de lissage.

Avantages :

↳ estimateur lisse et non constant par morceaux (si k l'est).

↳ poids des observations dépend de leur proximité par rapport à x.

Inconvenient :

↳ dépend du noyau k

↳ ⚠ dépend du paramètre de lissage h ! ⚠

II - Propriétés et vitesse de convergence

Propriété : 1 - $\forall$ $k \geq 0$ alors $\hat{f}_x \geq 0$

2 - $\forall$ k est une densité, alors $\hat{f}_x$ est une densité

3 - $\forall$ k est continue alors $\hat{f}_x$ est continue

4 - $\forall$ k est différentiable, alors $\hat{f}_x$ est différentiables

Démonstration : trivial !

Propriétés : Si f_X est $\mathcal{C}^2$ au voisinage de x et bornée (on peut s'en passer si k est à support compact). Si $h = h_n \xrightarrow[n \to +\infty]{} 0$ alors :

1 - $\text{Biais}\left(\hat{f}_X(x)\right) = \frac{h^2 f_X''(x)}{2} \mu_2(k) + o(h^2)$ où $\mu_2(k) = \int_{\mathbb{R}} s^2 k(s)\, ds < +\infty$

2 - Si $f_X(x) > 0$ et $R(k) = \int k^2(t)\, dt$ alors :

$$MSE\left(\hat{f}_X(x)\right) = \frac{1}{nh} f_X(x) R(k) + \int_{\mathbb{R}} \frac{h^4}{4} \left(f_X''(x)\right)^2 \mu_2^2(k) + o(h^4) + o\left(\frac{1}{nh}\right)$$

Démonstration :

1 - Calcul de l'espérance de $\hat{f}_X(x)$

$$\mathbb{E}_f\left[\hat{f}_X(x)\right] = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \mathbb{E}_f\left[k\left(\frac{X_i - x}{h}\right)\right]$$

$$\overset{iid}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_f\left[\frac{1}{h} k\left(\frac{X - x}{h}\right)\right] = \frac{1}{h} \mathbb{E}_f\left[k\left(\frac{X - x}{h}\right)\right]$$

$$= \int_{\mathbb{R}} \frac{1}{h} k\left(\frac{y - x}{h}\right) f_X(y)\, dy$$

$$= \int_{\mathbb{R}} k(t) f_X(x + th)\, dt.$$

Or, d'après le développement de Taylor-Young :

$$f_X(x + th) = f_X(x) + th f_X'(x) + \frac{(th)^2}{2} f_X''(x) + R(x + th)$$

où $\frac{|R(x+z)|}{z^2} \xrightarrow[z^2 \to 0]{} 0$ d'où :

$$\mathbb{E}_f\left[\hat{f}_X(x)\right] = \int_{\mathbb{R}} k(t) f_X(x)\, dt + \int_{\mathbb{R}} k(t)\, th f_X'(x)\, dt + \int_{\mathbb{R}} k(t) \frac{(th)^2}{2} f_X''(x)\, dt + \int_{\mathbb{R}} k(t) R(x + th)\, dt$$

$$= f_X(x) \underbrace{\int_{\mathbb{R}} k(t)\, dt}_{=1} + h f_X'(x) \underbrace{\int_{\mathbb{R}} t k(t)\, dt}_{=0} + \frac{h^2}{2} f_X''(x) \underbrace{\int_{\mathbb{R}} t^2 k(t)\, dt}_{\mu_2(k)} + \int_{\mathbb{R}} k(t) R(x + th)\, dt$$

Discussion du comportement biais / variance

- Si h diminue, le biais diminue mais la variance augmente
- Si h augmente, le biais augmente mais la variance diminue
- Pour que la variance tende vers 0, il faut $nh_n \xrightarrow[n\to+\infty]{} +\infty$
- Pour que le biais tend vers 0, il faut $h_n \xrightarrow[n\to+\infty]{} 0$.
- Pour une valeur de h et de k fixée:
  - ↳ plus la courbure de la densité ($=f_X''(x)$) est grande, plus le biais en x est grand
  - ↳ plus la densité est grande, plus la variance sera grande

Proposition : Si f est bornée et $\mathcal{C}^2$, $h = h_n \xrightarrow[n\to+\infty]{} 0$ et si $f_X \in L^2(\mathbb{R})$, $f_X' \in L^2(\mathbb{R})$, $f_X'' \in L^2(\mathbb{R})$.

Alors :

$$MISE\left(\hat{f}_X(x)\right) = \frac{1}{nh} R(k) + \frac{h^4}{4} \int_{\mathbb{R}} \left(f_X''(x)\right)^2 dx\, \mu_2^2(k) + o\left(h^4 + \frac{1}{nh}\right)$$

Démonstration : admise.

Maintenant, on cherche un choix optimal de h en fonction des expressions asymptotiques de MSE et MISE (h qui minimisent ces critères).

On note dans ce qui suit AMSE et AMISE les termes asymptotiquement dominants obtenus.

- $$\hat{h}_{AMSE} = \underset{h>0}{\operatorname{argmin}}\left(AMSE(h)\right) = \underset{h>0}{\operatorname{argmin}}\left(\frac{h^4}{4} f_X''^2(x)\mu_2^2(k) + \frac{1}{nh} f_X(x) R(k)\right)$$

Or, $$\frac{\partial AMSE}{\partial h}(h) = h^3 \left(f_X''(x)\right)^2 \mu_2^2(k) - \frac{1}{nh^2} f_X(x) R(k)$$

D'où $$\frac{\partial AMSE}{\partial h}(h) = 0 \Leftrightarrow h^5 = \frac{f_X(x) R(k)}{n\left(f_X''(x)\right)^2 \mu_2^2(k)} \quad \text{si } f_X''(x) \neq 0$$

D'où $$\hat{h}_{AMSE} = n^{-1/5} \sqrt[5]{\frac{f_X(x) R(k)}{\left(f_X''(x)\right)^2 \mu_2^2(k)}}$$

Au passage, on a $\frac{\partial^2 AMISE}{\partial h^2}(h) \geq 0$

On obtient de la même manière :

$$\hat{h}_{AMISE} = n^{-1/5} \sqrt[5]{\frac{R(k)}{\int_{\mathbb{R}} (f_X''(x))^2 dx \, \mu_2^2(k)}}$$

Remarque : Problème ! Ces expressions dépendent de f_X'' et de f_X qui sont inconnues. Une manière de s'en sortir serait de les remplacer par un estimateur (méthode plug-in sur laquelle on reviendra) ce qui est parfois lourd.

On déduit des expressions de $\hat{h}_{AMISE}$ et $\hat{h}_{AMISE}$ les expressions asymptotiques correspondantes (de AMSE et AMISE).

- $AMSE\left(\hat{f}_X, \hat{h}_{AMISE}\right) = \frac{1}{n \hat{h}_{AMISE}} f_X(x) R(k) + \frac{\hat{h}_{AMISE}^4}{4} \left(f_X''(x)\right)^2 \mu_2^2(k)$

$$= n^{-1} n^{1/5} \left(\frac{f_X(x) R(k)}{f_X''(x)^2 \mu_2^2(k)} \right)^{-1/5} f_X(x) R(k) + \frac{1}{4} n^{-4/5} \left(\frac{f_X(x) R(k)}{f_X''(x)^2 \mu_2^2(k)} \right)^{+1/5} \left(f_X''(x)\right)^2 \mu_2^2(k)$$

$$= n^{-4/5} \left(f_X(x) R(k)\right)^{4/5} \left(f_X''(x)^2 \mu_2^2(k)\right)^{1/5} + \frac{1}{4} n^{-4/5} \left(f_X(x) R(k)\right)^{4/5} \left(f_X''(x)^2 \mu_2^2(k)\right)^{1/5}$$

$$= \frac{5}{4} n^{-4/5} \left[\left(f_X(x) R(k)\right)^4 f_X''(x)^2 \mu_2^2(k) \right]^{1/5}$$

- $AMISE\left(\hat{f}_X, \hat{h}_{AMISE}\right) = \frac{5}{4} n^{-4/5} \left(R(k)^4 R(f_X'') \mu_2^2(k) \right)^{1/5}$ où $R(f_X'') = \int_{\mathbb{R}} f_X''(x)^2 dx$

Remarque : Comparaison entre les convergence (en vitesse) optimale :

* Dans le cas paramétrique, on a en général : $\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \sim \frac{C}{n}$

et si $\theta \mapsto f_\theta$ est $\mathcal{C}^2$ et si $\exists h \in \mathbb{L}^2(\mathbb{R})$ telle que $\frac{\partial^2 f_\theta}{\partial \theta^2}(x) \leq h(x)$
on a alors la même vitesse de convergence pour $f_{\hat{\theta}} \to f_\theta$ ($MSE[\hat{f}_\theta] \sim C_1 n^{-1}$ et $MISE[\hat{f}_\theta] \sim C_2 n^{-1}$).

* Dans le cas d'un histogramme à classes d'amplitudes égales à h, on a $MSE \sim C_1' n^{-2/3}$ et $MISE \sim C_2' n^{-2/3}$

III Choix optimal du noyau par rapport à AMISE et de l'AMISE

On voit que le noyau K influence l'expression asymptotique de la MISE et MSE optimales au travers du terme $C(K) = \left[\mu_2^2(K) R^4(K)\right]^{1/5}$

Le noyau optimal est donc celui qui minimise cette quantité tout en vérifiant les hypothèses faites sur le noyau

On montre que ce problème de minimisation sous contrainte a pour solution le noyau d'Epanechnikov (cf Epanechnikov, 1969)

$$K_{opt}(t) = \frac{3}{4}(1-t^2)\mathbb{1}_{\{|t|<1\}}$$

Afin de comprendre ce que on risque en n'utilisant pas le noyau optimal, on considère la notion d'efficacité d'un noyau K par rapport au noyau optimal K_{opt} qui se définie par :

$$eff(K) = \left(\frac{C(K_{opt})}{C(K)}\right)^{5/4}$$

Remarque : Par définition, $eff(K) \leq 1$

Ce rapport s'interprète comme la proportion d'observations

nécessaires pour atteindre avec K_{opt} la même AMISE qu'avec le noyau K.

En effet,

$$AMISE(K) := AMISE\left(\hat{f}_{\nabla x}, \hat{h}_{AMISE}, K, n\right)$$
$$AMISE(K_{opt}) := AMISE\left(\hat{f}_{\nabla x}, \hat{h}_{AMISE}, K_{opt}, n_0\right).$$

$$AMISE(K) = \frac{5}{4} n^{-4/5} R(f_x'')^{1/5} C(K).$$

$$AMISE(K_{opt}) = \frac{5}{4} n_0^{-4/5} R(f_x'')^{1/5} C(K_{opt})$$

$$AMISE(K) = AMISE(K_{opt}) \iff \left(\frac{n_0}{n}\right)^{4/5} = \frac{C(K_{opt})}{C(K)}$$

$$\iff \frac{n_0}{n} = \left(\frac{C(K_{opt})}{C(K)}\right)^{5/4} = eff(K) \leq 1.$$

Efficacité relative de quelques noyaux

Noyau	$K(t)$	$eff(K)$
Epanechnikov	$\frac{3}{4}(1-t^2)\mathbb{1}_{\{\lvert t\rvert\leq 1\}}$	1
Quartic	$\frac{15}{16}(1-t^2)^2\mathbb{1}_{\lvert t\rvert\leq 1}$	0.994
Triweight	$\frac{35}{32}(1-t^2)^3\mathbb{1}_{\lvert t\rvert\leq 1}$	0.987
Gaussien	$\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$	0.951
Triangulaire	$(1-\lvert t\rvert)\mathbb{1}_{\lvert t\rvert\leq 1}$	0.986
Uniforme	$\frac{1}{2}\mathbb{1}_{\lvert t\rvert\leq 1}$	0.930

On voit finalement que le choix du noyau n'a pas un si gros impact que cela sur les performances de notre estimateur (sauf peut-être pour uniforme et gaussien). Comme nous le verrons, un mauvais choix de

h est beaucoup plus néfaste qu'un mauvais choix de k.

IV - Choix automatique du paramètre de lissage h.

Nous allons voir maintenant trois méthodes permettant de choisir une valeur "appropriée" du paramètre de lissage en pratique :

1 - La règle simple de référence à une distribution normale
2 - La méthode plug-in
3 - La méthode de validation croisée.

1) La règle simple de référence à une distribution normale

Rappelons, comme nous l'avons vu précédemment que

$$\hat{h}_{AMISE} = \left[\frac{R(k)}{R(f_X'') \mu_2^2(k)} \right]^{1/5} n^{-1/5}$$

Si on fait l'hypothèse que la densité que l'on cherche à estimer est de loi normale alors on peut calculer $R(f_X'')$. On a :

$$f_X''(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \right]$$

D'où :

$$R(f_X'') = \int_{\mathbb{R}} f_X''(x)^2 dx$$

$$= \int_{\mathbb{R}} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \right) \right)^2 dx \qquad t = x - \mu$$

$$= \int_{\mathbb{R}} \left(\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} \left(\frac{t^2}{\sigma^2} - 1 \right) \right)^2 dt$$

$$= \int_{\mathbb{R}} \frac{1}{2\pi\sigma^6} e^{-\frac{z^2}{2\sigma^2}} \left(\frac{z^2}{2\sigma^2} - 1 \right)^2 \frac{1}{\sqrt{2}} dz \qquad z = t\sqrt{2}$$

$$R(f_x'') = \frac{1}{\sqrt{2\pi}\sqrt{2}\sigma} \int_{\mathbb{R}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} \left(\frac{z^4}{4\sigma^4} - \frac{z^2}{\sigma^2} + 1 \right) dz.$$

$$= \frac{1}{2\sqrt{\pi}\sigma^5} \left[\frac{1}{4} \underbrace{\mathbb{E}[U^4]}_{=3} - \underbrace{\mathbb{E}[U^2]}_{=1} + 1 \right] \quad \text{où } U \sim \mathcal{N}(0,1).$$

$$= \frac{1}{2\sigma^5\sqrt{\pi}} \left[\frac{3}{4} - 1 + 1 \right]$$

$$= \frac{3}{8\sigma^5\sqrt{\pi}}.$$

On obtient alors :

$$\hat{h}_{AMISE}^{NR} = \left(\frac{8\sqrt{\pi}\sigma^5}{3} \frac{R(k)}{\mu_2^2(k)} \right)^{1/5} n^{-1/5}$$

$$= \sigma \left(\frac{8\sqrt{\pi} R(k)}{3\mu_2^2(k)} \right)^{1/5} n^{-1/5}.$$

On proposera donc de prendre comme paramètre de lissage

$$\hat{h}_{NR} = \left(\frac{8\sqrt{\pi} R(k)}{3\mu_2^2(k)} \right)^{1/5} \hat{\sigma}\, n^{-1/5}$$ où $\hat{\sigma}$ est un estimateur de l'écart-type σ.

Par exemple, $\hat{\sigma} = S_{n-1}$ où $\hat{\sigma} = \min\left(S_{n-1}, \frac{R}{1.349} \right)$ où R est l'écart interquartiles $Q_{0.75} - Q_{0.25}$ empirique et S_{n-1} est l'écart-type corrigé.

Remarque : D'où sort $\hat{\sigma} = \min\left(S_{n-1}, \frac{R}{1.349} \right)$?

Si $X \sim \mathcal{N}(0,1)$, $Q_X(0.75) - Q_X(0.25) = 1.349$.

Si $X \sim \mathcal{N}(\mu, 1)$, $Q_X(0.75) - Q_X(0.25) = 1.349$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, $Q_X(0.75) - Q_X(0.25) = \sigma \times 1.349$.

En effet, si $Z = aX$, $a > 0$

$Q_X(\alpha)$ est solution en x de $F_X(x) = \alpha$.

Donc $Q_Z(\alpha)$ solution de $F_Z(z) = \alpha = F_X(z/a)$

d'où $\frac{z}{a} = Q_X(\alpha) \iff z = a \times Q_X(\alpha)$.

d'où $Q_Z(\alpha) = a\, Q_X(\alpha)$.

est un estimateur à noyau (4 fois différentiable) de la densité utilisant un paramètre de lissage sous-optimal et construit à partir des $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$.

Les méthodes peuvent éventuellement être itérées en utilisant comme paramètre de lissage celui trouvé à l'étape précédente : $h_0 \longrightarrow h_1 \longrightarrow h_2 \longrightarrow \dots$

3) La méthode de validation croisée.

L'idée est assez différente de ce que nous avons fait pour les deux méthodes précédentes. On part pas de l'expression de h_{AMISE} mais on estime directement $MISE(\hat{f}_x) - \int f_x^2(x)dx$. Puis on prend comme valeur de h celui qui minimise ce critère.

↳ Estimation de $MISE(\hat{f}_x) - \int f_x(x)^2 dx$

On note $MISE(h) = MISE(\hat{f}_x, h)$.

$$MISE(\hat{f}_x) = \mathbb{E}\left[\int (\hat{f}_x - f_x)(x)^2 dx\right]$$

$$= \mathbb{E}\left[\int \hat{f}_x(x)^2 dx\right] + \int f_x(x)^2 dx - 2\mathbb{E}\left[\int \hat{f}_x(x) f_x(x)\, dx\right]$$

D'où minimiser $MISE(h)$ en h revient à minimiser :

$$MISE(h) - \int f_x^2(x)dx = \mathbb{E}\left[\int \hat{f}_x(x)^2 dx\right] - 2\mathbb{E}\left[\int \hat{f}_x(x) f_x(x) dx\right]$$

Il nous suffit alors de trouver un estimateur sans biais de ce dernier critère :

$$LSCV(h) = \int \hat{f}_x(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_x^{-i}(X_i).$$

où $\hat{f}_x^{-i}(X_i) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \frac{1}{h} K\left(\frac{X_j - X_i}{h}\right)$

D'autre part, on peut calculer le coefficient $\left(\frac{8\sqrt{\pi}R(k)}{3\mu_2^2(k)}\right)^{1/5}$ en fonction du noyau utilisé :

Noyau	$\frac{8\sqrt{\pi}R(k)}{3\mu_2^2(k)}$
Epanechnikov	2.34
Gaussien	1.06
Triweight	2.78

↳ Atout : facile à calculer (et à implémenter)
↳ Défaut : on part de l'hypothèse que l'on cherche à estimer est de loi normale. Cela ne marchera pas bien si la vraie densité (que l'on estime) est de nature différente.

2) La méthode plug-in.

Elle consiste à estimer la constante $R(f'')$ qui apparaît dans l'expression de $\hat{h}_{AMISE}$ à l'aide d'un estimateur à noyau utilisant un paramètre de lissage sous-optimal.

• Méthode de Sheater et Jones (1991).
$\widehat{R(f''_x)} = R(\hat{f}''_x)$ où $\hat{f}_x$ est un estimateur à noyau de la densité utilisant un noyau deux fois différentiable et un paramètre de lissage h_0 sous-optimal

• Une autre méthode provient du fait que si $\lim_{x\to\pm\infty} f_x^{(3)}(x) f_x(x) = \lim_{x\to\pm\infty} f_x^{(2)}(x) f_x^{(1)}(x) = 0$ alors $\mathbb{E}[f_x^{(4)}(X)] = R(f_x^{(2)})$

On peut ensuite estimer $\mathbb{E}[f_x^{(4)}(X)]$ par $\frac{1}{n}\sum_{i=1}^{n} (\hat{f}_x^{-i})^{(4)}(X_i)$ où $\hat{f}_x^{-i}$

vérifie est un $\hat{f}^{SB}$ de $\mathbb{E}[f_x^{(4)}(X)]$

$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)$ est un estimateur sans biais de type "leave-one-out" de $\mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right]$

Proposition : $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)\right] = \mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right]$

Démonstration :

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\hat{f}_X^{-i}(X_i)\right] \overset{i.i.d}{=} \mathbb{E}\left[\hat{f}_X^{-1}(X_1)\right]$$

$$= \mathbb{E}\left[\frac{1}{n-1}\sum_{j=1,\, j\neq i}^{n} \frac{1}{h} k\left(\frac{X_j - X_1}{h}\right)\right] \overset{i.d}{=} \frac{1}{n-1}(n-1)\,\mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right)\right]$$

$$= \mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right)\right] = \int \frac{1}{h} k\left(\frac{x_2 - x_1}{h}\right) f_X(x_1) f_X(x_2)\,dx_1 dx_2$$

Or $\mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \int \frac{1}{h} k\left(\frac{X_i - x}{h}\right) f_X(x)\,dx\right]$

$\overset{i.d}{=} \mathbb{E}\left[\int \frac{1}{h} k\left(\frac{X_1 - x}{h}\right) f_X(x)\,dx\right]$ → on fait le changement de variable $u = \frac{X_1 - x}{h}$ (on suppose $h > 0$)

$= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right) \mid X_2\right]\right] = \mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right)\right]$

D'où $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)\right] = \mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right]$.

La méthode "least squares cross validation" consiste donc à prendre $\hat{h}_{LSCV} = \underset{h>0}{\operatorname{argmin}}\, LSCV(h)$

V. Estimation du mode

On va proposer une méthode d'estimation non-paramétrique du mode θ.

Rappel : $\theta \in \operatorname{argmax}\{t \in \mathbb{R} : f_X(t)\}$

On propose donc de manière assez naturelle l'estimateur du mode suivant

$$\hat{\theta} = \operatorname{argmax}\{t \in \mathbb{R} : \hat{f}_X(t)\}$$

Et argmax n'est pas forcement unique. On prend alors l'une des valeurs qui maximise $\hat{f}_X$.

En pratique, la méthode la plus basique consiste à estimer $\hat{f}_X$ sur une grille $t_1, \dots, t_M$ de points equi-espacés d'un pas ε petit et de prendre comme estimateur du mode le point de la grille où $\hat{f}_X$ est maximale

$$\hat{\theta} = \operatorname{argmax}\{j \in [\![1, M]\!],\ \hat{f}_X(t_j)\}.$$

On peut bien entendu utiliser des méthodes plus raffinées :

↳ Newtown - Raphson

↳ méthode du nombre d'or

↳ ...

C Estimateur à noyau de la fonction de régression

On suppose maintenant que l'on observe pour chaque individu une observation (X_i, Y_i) de variables réelles. On suppose $(X_i, Y_i) \overset{iid}{\sim} (X, Y)$.
$1 \le i \le n$

I - Définition

On souhaite étudier la manière dont la variable d'intérêt (ou réponse) Y dépend de la variable