Chapitre 2 : Estimateurs à noyau en dimension supérieure

A - Le cas multivarié (FV, chap IV)

On suppose à partir de maintenant que $X \in \mathbb{R}^d$ et $Y \in \mathbb{R}$ pour la regression (on pourrait avoir $Y \in \mathbb{R}^p$, ça ne changerait pas les vitesses de convergence).

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

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

$$\hat{r}(x) = \frac{\hat{g}(x)}{\hat{f}(x)} \quad \text{avec} \quad \hat{g}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)$$

On a $K : \mathbb{R}^d \longrightarrow \mathbb{R}$

h : paramètre de lissage

On peut définir (FV, chap IV, p. 78) l'équivalent d'un noyau d'ordre k en multivarié.

- On obtient sous hypothèses que f et r sont $\mathcal{C}^k$ et autres, en convergence presque complète : $\mathcal{O}(h^k) + \mathcal{O}\left(\sqrt{\frac{\ln(n)}{nh^d}}\right)$ pour convergence ponctuelle ou uniforme.

- Sous l'hypothèse f et r Hölder(β) et d'autres :

convergence p.co | uniforme / ponctuelle $\quad O(h^\beta) + O\left(\sqrt{\frac{\ln(n)}{nh^d}}\right)$

• Sous l'hypothèse f et r $\mathcal{C}^2$ et autres, convergence p.co :

$$MSE(\hat{r}(x)) = B^2(x)h^4 + V(x)\frac{1}{nh^d} + o\left(h^4 + \frac{1}{nh^d}\right)$$

$$MISE(\hat{r}) = B^2 h^4 + V\frac{1}{nh^d} + o\left(h^4 + \frac{1}{nh^d}\right)$$

Si on cherche le minimisant AMISE :

$$\frac{\partial AMISE}{\partial h}(h) = 4B^2h^3 - d\frac{V}{n}\frac{1}{h^{d+1}}$$

D'où $\frac{\partial AMISE}{\partial h}(h) = 0 \iff \frac{4B^2 n}{Vd} = \frac{1}{h^{d+4}}$

$$\iff h^{d+4} = \frac{Vd}{4B^2 n} \iff h = \left(\frac{Vd}{4B^2}\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}}$$

Plus d est grand, plus on prend h grand.

Pour ce h précis, on a :

$$MISE(\hat{r}) = B^2\frac{Vd}{4B^2 n h^d} + V\frac{1}{nh^d} + o\left(h^4 + \frac{1}{nh^d}\right)$$

$$= \frac{1}{nh^d}\left(\frac{Vd}{4} + V\right) + o\left(\frac{1}{nh^d}\right)$$

$$= \frac{V}{nh^d}\left(\frac{d}{4} + 1\right) + o\left(\frac{1}{nh^d}\right)$$

ou

$$MISE(\hat{r}) = \left(B^2 + V\frac{4B^2}{Vd}\right)h^4 + o(h^4)$$

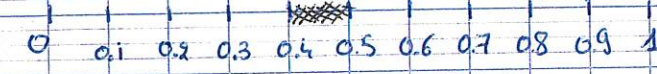
$$= B^2\left(1 + \frac{4}{d}\right)\left(\frac{Vd}{4B^2}\right)^{\frac{d}{d+4}} n^{-\frac{4}{d+4}} + o\left(n^{-\frac{4}{d+4}}\right)$$

On remarque que la dimension d dégrade la vitesse de convergence optimale pour la MISE, MSE, AMISE, ...

C'est dû au fléau de la dimension. Il y a une rarefaction des données lorsque d augmente.

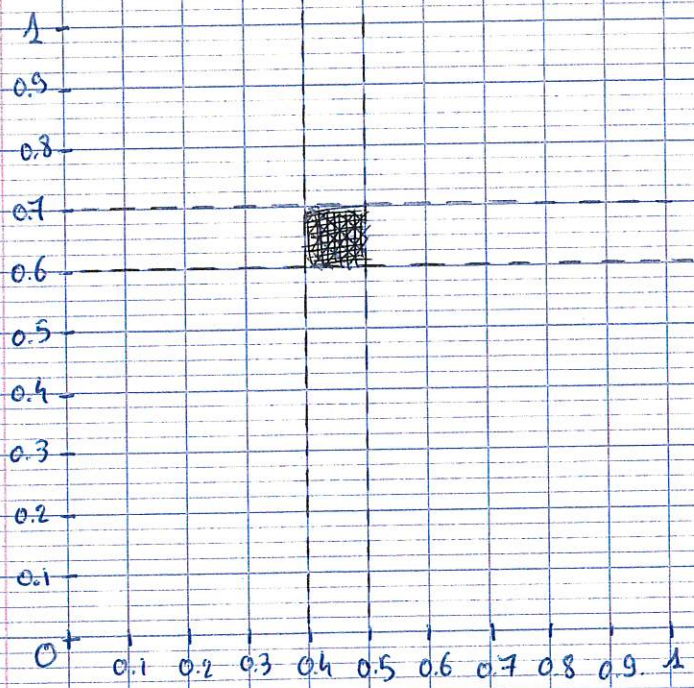
Supposons que l'on ait un échantillon de taille $n = 10^6$ iid de loi $\mathcal{U}_{[0,1]^d}$

Si $d = 1$:



N_1 le nombre d'individu sur un segment de longueur 0.1 est de loi $Bin(n, 0.1)$ dont l'espérance vaut $100\,000 =$

Si $d = 2$:



N le nombre d'individus sur un carré de longueur 0.1 suit une loi $Bin(n, 0.01)$ dont l'espérance est $n \times 0.01 = 10^4 = 10\,000$

Si $d > 2$, le nombre d'individu N sur une portion (pavé) de longueur de côté 0.1 a pour loi une $Bin(n, 0.$
d'espérance $n \times (0.1)^d = n\,10^{-d} = 10^{6-d}$.

$d = 3$: $n(0.1)^d = 1000$ $\quad$ $d = 5$: $n(0.1)^d = 10$

$d = 4$: $n(0.1)^d = 100$ $\quad$ $d = 6$: $n(0.1)^d = 1$

Soit on augmente n (en général impossible), soit on prend h plus grand, ce qui détériore la vitesse de convergence.

Stone (1981, 1982) montre que la vitesse optimale d'un estimateur, quelqu'il soit, dans un modèle de regression avec $X \in \mathbb{R}^d$ et en supposant $f, r \in \mathcal{C}^k$, est donné par $n^{-\frac{k}{2k+d}}$ en norme $\mathbb{L}_q$, $q < \infty$. et $\left(\frac{n}{\ln(n)}\right)^{-\frac{k}{2k+d}}$ en norme $\mathbb{L}_\infty$

On remarque que notre estimateur à noyau atteint bien ces vitesses optimales :

MSE $\sim C n^{-\frac{4}{4+d}}$ $k=2$. vitesse $\sim C n^{-\frac{2}{4+d}}$.

AMISE $\sim C n^{-\frac{4}{4+d}}$ $k=2$ vitesse $\sim C n^{-\frac{2}{4+d}}$.

Le problème ne vient donc pas de l'estimateur mais du modèle qui est trop complexe à estimer. Peu considérée pour $d \geq 4$ ou 5.

On considère alors souvent des modèles semi-paramétrique

Modèle additif :

$$r(X) = \mu + \sum_{j=1}^{d} r_j(X^j)$$

Modèle additif étendu :

$$r(X) = L\left(\mu + \sum_{j=1}^{d} r_j(X^j)\right) \quad \text{où } L \text{ une fonction connue}$$

Modèle de transformations optimales :

$$\mathbb{E}[T(Y)|X] = \mu + \sum_{j=1}^{d} r_j(X^j) \quad \text{où } T \text{ fonction inconnue}$$

Modèle single index

$$Y = r(\langle X, \theta \rangle) + \varepsilon \quad \text{où } r, \theta \text{ inconnus}$$

Modèle multi-index :

$Y = \sum_{j=1}^{d} r_j(\langle X, \theta_j \rangle) + \varepsilon$ où $\begin{cases} r_j \\ \theta_j \end{cases}$ inconnu.

B - Le cas fonctionnel

$X \in (\mathcal{F}, d)$ où • $\mathcal{F}$ espace de fonction
• d pseudo-distance (même propriété de distance sauf $d(x,y) = 0 \not\Rightarrow x = y$).

Régression : Ferraty et Vieu (2000) $x \in \mathcal{F}$

$$\hat{r}(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i K\left(\frac{d(X_i, x)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{d(X_i, x)}{h}\right)} & \text{si } \sum_{i=1}^{n} K\left(\frac{d(X_i, x)}{h}\right) \neq 0 \\ \bar{Y} & \text{sinon} \end{cases}$$

Remarque : • En pratique, on dispose seulement d'observation discrétisée de cette courbes. On suppose que le passage discrétisé à courbe est effectué en amont.

• Le fléau de la dimension peut dans un premier temps faire penser que cela ne peut pas fonctionner. C'est vrai si on prend un processus gaussien et d une norme alors les vitesses de convergence sont en $(\ln(n))^{-\gamma}$. Mais les courbes que l'on a à traiter contiennent souvent une certaine regularité, une structure, une dynamique particulière qui lui confère un statut particulier ; qui font qu'une telle approche peut-être tout de même adaptée à condition de choisir d qui capte l'information utile de X. (une courbe ≠ ensemble de v.a. iid).

On peut par exemple choisir pour d

- $d_k(x_1, x_2) = \sqrt{\int \left(x_1^{(k)}(t) - x_2^{(k)}(t)\right)^2 dt}$
- $d_{PCA}^k(x_1, x_2) = \left\| \sum_{j=1}^{k} \langle x_1 - x_2, \varphi_j \rangle \varphi_j \right\|$ où φ_j $j^{ème}$ fct propres de l'ACP
- $d_\psi^k(x_1, x_2) = \left\| \sum_{j=1}^{k} \langle x_1 - x_2, \psi_j \rangle \psi_j \right\|$ où ψ_j base donné (Fourier, Ondelette

Le choix de d en pratique se fait par validation croisée :

$$CV(d) = \sum_{i=1}^{n} \omega(X_i) \left(Y_i - \hat{x}_d^{-i}(X_i)\right)^2$$