

# Estimation non-paramétrique: Méthode d'estimation à Noyau.

STEPHANE SADIO

17/04/2021

## Contents

Méthode non paramétrique d'estimation de la densité_____	1
Risque quadratique ponctuel des estimateurs à noyau sur les classe des espaces de Hölder . . . . .	1

## Méthode non paramétrique d'estimation de la densité\_\_\_\_\_

### Méthode d'estimation à noyau

Principe:

### Risque quadratique ponctuel des estimateurs à noyau sur les classe des espaces de Hölder

Nous nous intéressons au risque quadratique ponctuel de  $\hat{f}_n$ , i.e étant donné  $x_0 \in \mathbb{R}$

$$R(\hat{f}_n, f) = \mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2]$$

Rappelons la décomposition “biais au carré + variance” du risque quadratique:

$$\mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2] = (\mathbb{E}[\hat{f}_n(x_0)] - f(x_0))^2 + \mathbb{V}(\hat{f}_n(x_0))$$

#### Majoration du biais et de la variance.

Dans cette section, nous allons nous intéresser au compromis biais-variance afin de minimiser le risque quadratique. Les deux propositions suivantes montrent que sous certaines hypothèses, on peut majorer le biais ainsi que la variance.

*Définition:* Soit  $l \in \mathbb{N}^*$ . On dit que le noyau  $K$  est d'ordre  $l$  si  $u^j K(u)$  est intégrable et  $\int u^j K(u) du = 0$ ,  $j = 1, \dots, l$ .

**Proposition:** Si  $f \in \sum(\beta, L)$  avec  $\beta > 0$  et  $L > 0$  et si  $K$  est un noyau d'ordre  $l = \lfloor \beta \rfloor$  tel que  $\int |u^\beta| |K(u)| du < \infty$  alors pour tout  $x_0 \in \mathbb{R}$ , et pour tout  $h > 0$  le biais peut être borné comme suit :

$$|\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)| \leq \frac{h^\beta L}{l!} \int |u|^\beta |K(u)| du$$

*Preuve:* (voir Esti-non para.pdf page 97, prop 4.10).

Le biais au carré tend vers zéro à la vitesse  $h^{2\beta}$ . Plus la fonction  $f$  est régulière, plus le biais tend vite vers zéro quand  $h$  tend vers zéro (à condition bien sûr que l'ordre du noyau soit suffisamment grand).

**Proposition:** Si  $f$  est bornée et si  $K$  est de carré intégrable alors

$$\mathbb{V}(\hat{f}_n(x_0)) \leq \frac{\|f\|_\infty \|K\|_2^2}{nh}$$

En particulier, si  $f \in \sum(\beta, L)$  alors

$$\mathbb{V}(\hat{f}_n(x_0)) \leq \frac{M(\beta, L)}{nh}$$

**Démonstration:**

$$\mathbb{V}(\hat{f}_n(x_0)) = \mathbb{V}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)\right) = \sum_{i=1}^n \mathbb{V}\left(\frac{1}{nh} K\left(\frac{X_i - x_0}{h}\right)\right) = \sum_{i=1}^n \mathbb{V}\left(\frac{1}{nh} K\left(\frac{X_i - x_0}{h}\right)\right) = \sum_{i=1}^n \frac{1}{n^2 h^2} \mathbb{V}(K\left(\frac{X_i - x_0}{h}\right)) = \frac{1}{nh^2} \mathbb{V}(K\left(\frac{X_i - x_0}{h}\right))$$

Et enfin, on utilise la proposition ?: il existe une constante positive  $M(\beta, L)$  tel que  $\|f\|_\infty \leq M(\beta, L)$ . Ceci implique que

$$\mathbb{V}(\hat{f}_n(x_0)) \leq \frac{1}{nh} M(\beta, L) \int K^2(v) dv$$

Pour que la variance tende vers zéro, il faut que  $nh$  tende vers l'infini. En particulier, à  $n$  fixé, la variance est une fonction décroissante de  $h$ . Il y a donc une valeur optimale de  $h$  qui doit réaliser l'équilibre entre le biais au carré et la variance. On peut à présent donner un contrôle du risque quadratique par le théorème suivant.

**Théorème:** Soit  $\beta > 0$  et  $L > 0$  et  $K$  un noyau de carré intégrable et d'ordre  $\lfloor \beta \rfloor$  tel que  $\int |u^\beta| |K(u)| du < \infty$ . Alors, en choisissant une fenêtre de la forme  $h = cn^{-\frac{1}{2\beta+1}}$  avec une constante  $c > 0$ , on obtient pour tout  $x_0 \in \mathbb{R}$ ,

$$R(\hat{f}_n(x_0), \sum_d(\beta, L)) := \sup_{f \in \sum_d(\beta, L)} \mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2] \leq C n^{-\frac{2\beta}{2\beta+1}}$$

où  $C$  est une constante dépendant de  $L$ ,  $\beta$ ,  $c$  et  $K$ .

**Démonstration:** On a

$$R(\hat{f}_n(x_0), f(x_0)) = \text{Biais} + \text{Variance}$$

Si nous nous référons aux deux propositions précédentes, nous pouvons écrire

$$R(\hat{f}_n(x_0), f(x_0)) \leq \left(\frac{h^\beta L}{l!} \int |u|^\beta |K(u)| du\right)^2 + \frac{M(\beta, L) \|K\|_2^2}{nh}$$

On cherche ensuite la fenêtre  $h$  qui minimise cette quantité. Comme on ne se soucie pas vraiment des constantes exactes quand on cherche la vitesse de convergence d'un estimateur, on utilisera la notation  $c_1 = (\frac{L}{l!} \int |u|^\beta |K(u)| du)^2$  et  $c_2 = \frac{M(\beta, L) \|K\|_2^2}{nh}$ . On doit alors minimiser en  $h$  la quantité

$$c_1 h^{2\beta} + \frac{c_2}{nh}$$

On a une quantité croissante et une quantité décroissante en  $h$ . Encore une fois, comme on ne se soucie pas des constantes, donc on cherche la fenêtre  $h$  qui nous donne l'ordre minimal du risque. Quand  $h$  est trop grand, le biais est trop grand, et quand  $h$  est trop petit, c'est la variance qui est trop grande. On cherche donc la fenêtre  $h$  qui réalise un équilibre entre le biais au carré et la variance:

$$h^{2\beta} \approx \frac{1}{nh}$$

où le signe  $\approx$  signifie ici “de l'ordre de”. Cela donne

$$h \approx n^{-\frac{1}{2\beta+1}}$$

Autrement dit, pour une fenêtre  $h$  de l'ordre de  $n^{-\frac{1}{2\beta+1}}$ , le biais au carré et la variance sont de même ordre. Plus exactement, on choisit la fenêtre  $h_* = cn^{-\frac{1}{2\beta+1}}$ , avec  $c$  une constante positive, on a

$$\text{Biais au carré} \approx h_*^{2\beta} \approx \text{Variance} \approx \frac{1}{nh_*}$$

De plus on a alors

$$h_* \approx n^{-\frac{2\beta}{2\beta+1}}$$

Autrement dit, il existe une certaine constante  $C$  telle que, pour cette fenêtre  $h_*$ , on a

$$R(\hat{f}_n(x_0), \sum_d (\beta, L)) \leq C n^{\frac{-2\beta}{2\beta+1}}$$

Cette fenêtre est donc optimale à une constante près (si on change  $c$ , on change  $C$  ça ne change pas le taux qui est  $n^{\frac{-2\beta}{2\beta+1}}$ ).

**Remarque:** \_\_\_ l'estimateur dépend de  $\beta$  à travers la fenêtre  $h$ . Or, sans connaissance a priori sur les propriétés de la fonction  $f$ , on ne peut donc pas utiliser cet estimateur. On essaie alors de trouver un choix de fenêtre ne dépendant que des données et qui soit aussi performant (ou presque) que l'estimateur utilisant cette fenêtre optimale. A ce sujet, on introduira plus loin un choix de fenêtre ne dépendant que des données et qui est basé sur ce qu'on appelle la validation croisée (ou “cross validation” en Anglais).

\_\_\_ Nous avons vu plus haut que le biais au carré tend vers zéro quand  $h$  tend vers zéro (si  $\beta$  est suffisamment grand). Nous en déduisons la convergence de l'espérance de l'estimateur à noyau  $\hat{f}_n$  vers la fonction  $f$ . Et donc, l'estimateur à noyau est asymptotiquement sans biais,  $\hat{f}_n$  est consistante.

### Choix de la fenêtre $h$ par validation croisée.

\_\_\_ Le choix de la fenêtre dans la section précédente est criticable: comme on l'a mentionné, il dépend de la régularité la fonction  $f$  qui est inconnue dans notre cas. On peut donc essayer d'estimer cette fenêtre idéale par un estimateur  $\hat{h}$ . De façon à souligner la dépendance à la fonction, on va noter  $\hat{f}_{n,h}$  l'estimateur associé à un choix de fenêtre  $h$ . L'estimateur final sera  $\hat{f}_{n,\hat{h}}$ , une fois le choix de  $\hat{h}$  fait.

\_\_\_ On cherche à minimiser en  $h$  le risque quadratique pour la distance  $L_2$ :

$$R(\hat{f}_{n,h}) = \mathbb{E}[\|\hat{f}_{n,h} - f\|_2^2] = \mathbb{E}[\|\hat{f}_{n,h}\|_2^2] - 2\mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx] + \|f\|_2^2$$

Or la fonction  $f$  étant inconnue, ce risque n'est pas calculable à partir des données. On cherche donc à estimer ce risque en utilisant uniquement les données. Remarquons tout de suite que minimiser en  $h$  la quantité  $R(\hat{f}_{n,h}, f)$  est équivalent à minimiser en  $h$  la quantité  $R(\hat{f}_{n,h}, f) - \|f\|_2^2$ . On va en fait remplacer la

minimisation de la quantité inconnue  $R(\hat{f}_{n,h}, f) - \|f\|_2^2$  par la minimisation d'un estimateur  $\hat{R}(h)$  de cette quantité. Plus précisément on va chercher un estimateur sans biais de cette expression:

$$\mathbb{E}[\|\hat{f}_{n,h}\|_2^2] - 2\mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx]$$

Le premier terme admet  $\|\hat{f}_{n,h}\|_2^2$  comme estimateur trivial (d'après la propriété des estimateurs sans biais:  $\mathbb{E}[\hat{\beta}] = \beta$ ).

Il reste à trouver un estimateur sans biais du second terme. Pour cela, nous admettons par construction l'estimateur sans biais  $\hat{G}$  défini en tout points sauf en  $X_i$  (c'est le principe du Leave-one-out):

$$\hat{G} = \frac{1}{n} \sum_{i=1}^n \hat{f}_{n,h}^{(-i)}(X_i)$$

avec

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

Montrons que  $\mathbb{E}(\hat{G}) = \mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx]$ .

Comme les  $X_i$  sont i.i.d., d'une part nous avons

$$\mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx] = \mathbb{E}[\int \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)f(x)dx] = \frac{1}{h}\mathbb{E}[\int K\left(\frac{x - x_1}{h}\right)f(x)dx] = \frac{1}{h} \int f(x) \int K\left(\frac{x - X_1}{h}\right)f(x_1)dx_1 dx$$

D'autre part, nous avons

$$\mathbb{E}[\hat{G}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \hat{f}_{n,h}^{(-i)}(X_i)\right] = \mathbb{E}[\hat{f}_{n,h}^{(-1)}(X_1)] = \mathbb{E}\left[\frac{1}{n(n-1)h} \sum_{j \neq 1} K\left(\frac{X_j - X_1}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h} K\left(\frac{X - X_1}{h}\right)\right] = \frac{1}{h} \int f(x) \int K\left(\frac{x - x_1}{h}\right)f(x_1)dx_1 dx$$

Donc,  $\hat{G}$  est un estimateur sans biais de  $\int \hat{f}_{n,h}(x)f(x)dx$ . Finalement, l'estimateur sans biais de  $R(\hat{f}_{n,h}, f) - \|f\|_2^2$  est donné par:

$$\hat{R}(h) = \|\hat{f}_{n,h}\|_2^2 - \frac{2}{n(n-1)} \sum_{i=1} \sum_{j=1, j \neq i} \frac{1}{h} K\left(\frac{X_i - X_j}{h}\right)$$

On définit alors

$$\hat{h} = \arg \min_{h \in H} \hat{R}(h)$$

Si ce minimum est atteint. On cherche une fenêtre parmi une grille finie de valeurs, grille qu'on a notée  $H$  dans la formule ci-dessus.

L'estimateur  $\hat{f}_{n,\hat{h}}$  a de bonnes propriétés pratiques et de consistence. La validation croisée est une méthode très générale mais nous l'utilisons ici pour le choix la fenêtre  $h$  optimale.