

Projet MAT 4102

Estimations de densités

Alix Aouar, Océane Wauquier, Erwan Le Blévec

March 31, 2022

Abstract

Le but de ce projet est d'estimer la densité d'une variable aléatoire à partir d'échantillons. Les données seront ici réelles et correspondront à des mesures d'abondance (renormalisées) de dauphins. Elles sont issues des campagnes océanographiques menées par l'Ifremer depuis 2004 sur le navire Thalassa, permettant de recenser et dénombrer les dauphins communs. À partir de ces échantillons récoltés régulièrement, le but est un indicateur d'abondance est fourni et nous aller chercher à estimer sa densité.

1 Première partie - Estimateur à base d'histogrammes

1.1 Question T1

Soit $x \in [0, 1[$, notons j_0 l'entier tel que $x \in A_{j_0}$

$$\text{On a } \mathbb{E}[(\hat{p}_n^h(x) - p_\star(x))^2] = \{ \mathbb{E}[\hat{p}_n^h(x)] - p_\star(x) \}^2 + \text{Var}[\hat{p}_n^h(x)]$$

$$\text{Var}(\hat{p}_n^h(x)) = \left(\frac{1}{nh}\right)^2 \text{Var} \left(\sum_{i=1}^n \sum_{j=1}^m 1_{A_j}(X_i) 1_{A_j}(x) \right) \quad (1)$$

$$= \frac{1}{nh^2} \text{Var} \left(\sum_{j=1}^m 1_{A_j}(X) 1_{A_j}(x) \right) \quad (2)$$

$$= \frac{1}{nh^2} \text{Var} (1_{A_{j_0}}(X)) \quad (3)$$

$$= \frac{1}{nh^2} (\mathbb{E}[1_{A_{j_0}}(X)^2] - \mathbb{E}[1_{A_{j_0}}(X)]^2) \quad (4)$$

$$= \frac{1}{nh^2} \mathbb{P}(X \in A_{j_0}) (1 - \mathbb{P}(X \in A_{j_0})) \quad (5)$$

(1) - (2) Par indépendance entre les v.a on peut intervertir la variance et la somme, on utilise ensuite le fait que les v.a sont identiquement distribuées.

(2) - (3) Par définition de j_0

(4) - (5) Espérance d'une indicatrice.

Remarque : On pourrait remplacer les valeurs de la probabilités par un estimateur de cette dernière (celui empirique par exemple). Les probabilités étant bornées on remarque que la variance converge vers 0 lorsque n tend vers l'infini - ce qui est rassurant.

De plus :

$$\mathbb{E}[\hat{p}_n^h(x)] - p_\star(x) = \mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^m 1_{A_j}(X_i) 1_{A_j}(x)\right] - p_\star(x) \quad (6)$$

$$= \frac{1}{h} \mathbb{E}\left[\sum_{j=1}^m 1_{A_j}(X) 1_{A_j}(x)\right] - p_\star(x) \quad (7)$$

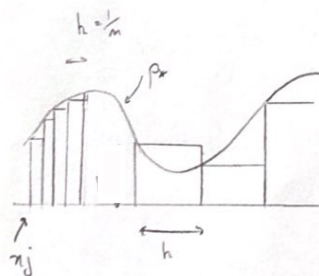
$$= \frac{1}{h} \mathbb{E}[1_{A_{j_0}}(X)] - p_\star(x) \quad (8)$$

$$= \frac{1}{h} \mathbb{P}(X \in A_{j_0}) - p_\star(x) \quad (9)$$

L'expression de l'erreur moyenne quadratique s'écrit alors :

$$\mathbb{E}[(p_n^h(x) - p_\star(x))^2] = \frac{1}{h^2 n} \mathbb{P}(X \in A_{j_0})(1 - \mathbb{P}(X \in A_{j_0})) + \left[\frac{1}{h} \mathbb{P}(X \in A_{j_0}) - p_\star(x)\right]^2$$

Remarques :



Augmenter n est bien le seul facteur qui permet de faire diminuer la variance sans toucher au biais.

On retrouve en revanche la balance entre le biais et la variance avec h . L'augmenter (ce qui revient à diminuer le nombre de classes dans l'histogramme) réduit l'erreur sur la variance mais augmente le terme du biais.

Si $h \ll 1$ $p_{j_0} = \int_{A_{j_0}} p_\star(x) dx \approx p_\star(x_{j_0}) h$ peut être correcte lorsque l'intervalle d'intégration est petit (donc $\frac{p_{j_0}}{h} \approx p_\star(x)$)

Dès que h augmente cette approximation n'est plus du tout précise \rightarrow le biais \uparrow .

T2. Evaluer $\mathcal{E} = \mathbb{E} \left[\int_0^1 [\hat{p}_{n,h}(x) - p_*(x)]^2 dx \right]$

$$\mathbb{E} \left[\int_0^1 [\hat{p}_{n,h}(x) - p_*(x)]^2 dx \right] = \underbrace{\int_0^1 p_*^2(x) dx}_{(1)} + \underbrace{\int_0^1 \mathbb{E} [\hat{p}_{n,h}^2(x)] dx}_{(2)} - 2 \int_0^1 p_*(x) \mathbb{E} [\hat{p}_{n,h}(x)] dx$$

On ne connaît pas p_* , le terme (1) ne peut pas être évalué.

Terme (3): $\mathbb{E} [\hat{p}_{n,h}(x)] = \frac{1}{nh} \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^m \mathbb{1}_{A_j}(x) \mathbb{1}_{A_j}(X_i) \right]$

(X_i) identiquement distribuées $= \frac{1}{h} \sum_{j=1}^m \mathbb{1}_{A_j}(x) \underbrace{\mathbb{E} [\mathbb{1}_{A_j}(X)]}_{= P_j}$

Il vient $2 \int_0^1 p_*(x) \mathbb{E} [\hat{p}_{n,h}(x)] dx = 2 \int_0^1 p_*(x) \frac{1}{h} \sum_{j=1}^m \mathbb{1}_{A_j}(x) P_j dx$

$$= \frac{2}{h} \sum_{j=1}^m P_j \int_{A_j} p_*(x) dx$$

def p_* $= \frac{1}{h} \sum_{j=1}^m P_j^2$

$$(3) = \frac{2}{h} \sum_{j=1}^m P_j^2$$

Terme (2): le carré d'une double somme n'étant pas pratique à manipuler on cherche à la transformer.

$$\mathbb{E} [\hat{p}_{n,h}^2(x)] = \text{Var} [\hat{p}_{n,h}(x)] + \mathbb{E} [\hat{p}_{n,h}(x)]^2$$

(q. T.1) $= \frac{1}{nh^2} P_{j_0}(x) (1 - P_{j_0}(x)) + \left[\frac{1}{h} P_{j_0}(x) \right]^2$

$j_0: \begin{cases} [0,1] \rightarrow [1,m] \\ n \mapsto j \text{ telle que } x \in A_j \end{cases}$

On passe par les $\mathbb{1}_{A_j}$ par ailleurs les $j_0(x)$ $= \frac{1}{nh^2} \sum_{j=1}^m P_j (1 - P_j) \mathbb{1}_{A_j}(x) + \frac{1}{h^2} \sum_{j=1}^m \mathbb{1}_{A_j}(x) P_j^2$

On intègre sur $[0,1]$ $\hookrightarrow \left[\frac{1}{nh} \sum_{j=1}^m P_j (1 - P_j) + \frac{1}{h} \sum_{j=1}^m P_j^2 = \int_0^1 \mathbb{E} [\hat{p}_{n,h}^2(x)] dx \right]$

Finalement $\left\| \mathcal{E} = \int_0^1 p_*^2(x) dx + \frac{1}{nh} \sum_{j=1}^m P_j (1 - P_j) - \frac{1}{h} \sum_{j=1}^m P_j^2 \right\| (1 \frac{2}{h})$

2 Deuxième Partie - Estimateur à base de noyaux

Dans le cours l'idée des estimateurs à noyaux a été introduite sous la forme suivante :

$$p_n^h : x \mapsto \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} 1_{-1 \leq \frac{x_i - x}{h} \leq 1}$$

Forme qui découle directement de la relation entre la fonction de répartition et la densité d'une variable aléatoire. Si cet estimateur est ainsi théoriquement appréciable l'un des problèmes qu'il rencontre dans la pratique est la faiblesse de sa continuité (constant par morceaux, d'où son nom d'"estimateur par histogrammes").

Dans un contexte où il est préférable d'avoir un estimateur continu sur son intervalle de définition, cette forme intrinsèque de l'estimateur par histogrammes l'empêche d'être adapté.

Une considération classique pour approcher une densité avec un résultat davantage continu est d'utiliser noyau suivant :

$$K : u \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

L'expression de l'estimateur devient :

$$p_n^h : x \mapsto \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$$

Pour $i \in \{1, n\}$, $K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x - x_i)^2}{h^2}}$, chaque terme de la somme est donc proportionnel à une loi normale de moyenne de x_i et d'écart type h .

Pour des variations de x : Si x se situe dans un intervalle - dont la longueur est définie par la valeur de h - avec "beaucoup" de x_i les densités associées à ces x_i seront importantes en x donc l'estimateur le sera également.

Inversement si x est en quelque sorte isolé des x_i peu de normales vont "s'activer" et la valeur retournée par l'estimateur sera faible.

Pour des variations de h : Le paramètre h étant commun à tous les x_i , il agit de manière globale sur le graphe de l'estimateur. Son augmentation équivaut à celle de la variance de chaque gaussienne, une valeur élevée entrainera donc "l'activation" d'un plus grand nombre de gaussiennes en un même point, ce qui aura tendance à aplatir les variations de l'estimateur.

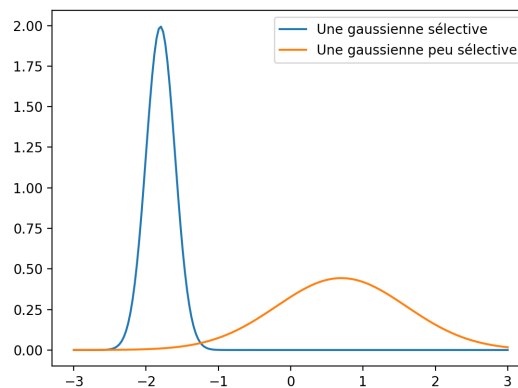


Figure 1: Sélectivité d'une gaussienne pour un x_i fixé

C'est pour cela que nous observons des courbes de plus en plus plates (*voir notebook, q.S3*) au fur et à mesure que la valeur de h augmente. Les x activent de plus en plus de gaussiennes et les valeurs de des densités s'uniformisent (on tend vers des indicatrices).

En revanche lorsque h diminue les gaussiennes deviennent si sélectives que le graphe de l'estimateur s'approche de celui d'un signal fortement bruité (voir figure ci dessous). Même si théoriquement diminuer h augmente la précision de l'approximation de la dérivée ($p_*(x) \approx \frac{F(x+h)-F(x-h)}{2h}$), en pratique cela fait tendre les gaussiennes vers des diracs, ce qui brise la continuité recherchée de l'estimateur (qui devient non nul ponctuellement, en les points de la base de données..)

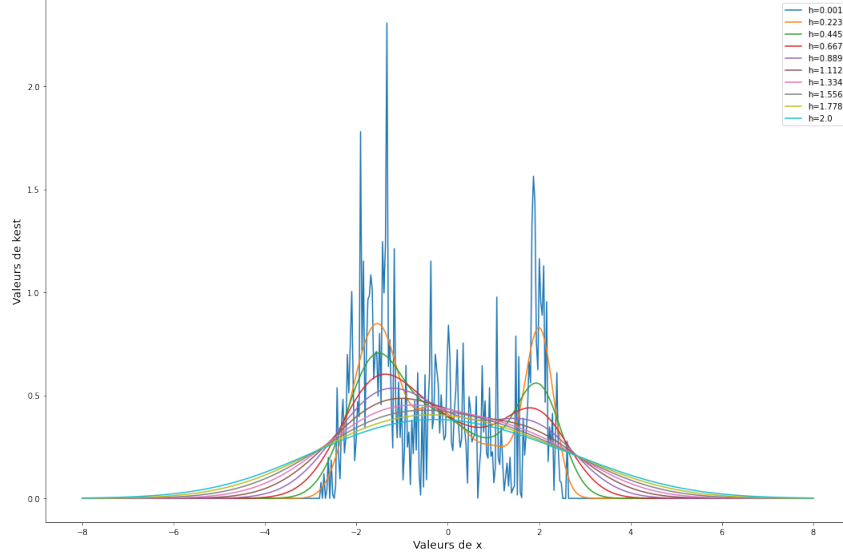


Figure 2: Variations de la courbe de kest en fonction de h

Remarque : Pour h très faible ($h = 0.001$) on remarque bien cette tendance "signal bruité" donnée par l'estimateur.

2.1 Question T3.1

Démonstration déjà réalisée en cours (un peu long de tout réécrire en Latex..)

$$\begin{aligned} \mathbb{V}[\hat{p}_n^h(x)] &= \frac{1}{nh_n^2} \mathbb{V} \left[K \left(\frac{X_1 - x}{h_n} \right) \right] \leq \frac{1}{nh_n^2} \mathbb{E} \left[K^2 \left(\frac{X_1 - x}{h_n} \right) \right] \\ &\leq \frac{1}{nh_n^2} \int_{\mathbb{R}} K^2 \left(\frac{u - x}{h_n} \right) p_*(u) du, \\ &\leq \frac{1}{nh_n} \int_{\mathbb{R}} K^2(u) p_*(x + uh_n) du, \\ &\leq \frac{1}{nh_n} \|p_*\|_{\infty} \int_{\mathbb{R}} K^2(u) du. \end{aligned}$$

Figure 3: Variance de l'estimateur à noyaux

Justifications :

1. Le caractère iid des $(X_i)_{1 \leq i \leq n}$ nous donne la première égalité
2. La première inégalité est obtenue avec $\mathbb{V}ar(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ donc $\mathbb{V}ar(X) \leq \mathbb{E}[X^2]$
3. Formule de transfert
4. Changement de variable, $u = x + z * h_n$
5. p_* bornée sur \mathbb{R}

2.2 Question T3.2

$$\begin{aligned}\mathbb{E}[\hat{p}_n^{h_n}(x)] &= \frac{1}{h_n} \mathbb{E} \left[K \left(\frac{X_1 - x}{h_n} \right) \right] \\ &= \frac{1}{h_n} \int_{\mathbb{R}} K \left(\frac{u - x}{h_n} \right) p_*(u) du, \\ &= \int_{\mathbb{R}} K(u) p_*(x + uh_n) du.\end{aligned}$$

Figure 4: Espérance de l'estimateur à noyaux

Première égalité : $(X_i)_i$ iid, puis changement de variable dans l'intégrale pour (2) - (3)
Le biais de l'estimateur peut alors s'écrire :

$$\mathbb{E}[\hat{p}_n^{h_n}(x)] - p_*(x) = \int_{\mathbb{R}} K(u) (p_*(x + uh_n) - p_*(x)) du.$$

Ce qui nous donne :

For all $u \in \mathbb{R}$, there exists $\xi_{u,x}$ between x and $x + uh_n$ such that

$$p_*(x + uh_n) - p_*(x) = uh_n p'_*(x) + \frac{u^2 h_n^2}{2} p''_*(\xi_{u,x}).$$

Using that $\int_{\mathbb{R}} u K(u) du = 0$ yields

$$\mathbb{E}[\hat{p}_n^{h_n}(x)] - p_*(x) = \int_{\mathbb{R}} K(u) \frac{u^2 h_n^2}{2} p''_*(\xi_{u,x}) du$$

and

$$\left| \mathbb{E}[\hat{p}_n^{h_n}(x)] - p_*(x) \right| = \frac{h_n^2}{2} \int_{\mathbb{R}} u^2 |K(u)| \frac{u^2 h_n^2}{2} p''_*(\xi_{u,x}) du \leq \frac{h_n^2}{2} \|p''_*\|_{\infty} \int_{\mathbb{R}} u^2 |K(u)| du.$$

Nous obtenons bien la majoration du biais par un terme d'ordre h_n^2 .

2.3 Question T3.3

Finalement :

Si on introduit $f_n: u \mapsto \frac{c_1}{nu} + c_2 u^4$

On a, $\forall u \in \mathbb{R}$ $f'_n(u) = \frac{-c_1}{nu^2} + 4c_2 u^3$ et $f'_n(u) = 0 \Leftrightarrow c_2 u^3 = \frac{c_1}{nu^2}$

On peut donc minimiser notre borne supérieure en choisissant $u = \left(\frac{c_1}{4c_2 n}\right)^{1/5}$

On a $h_n \in \left(\frac{c_1}{4c_2 n}\right)^{1/5}$ (voir des exp. précédentes pour les constantes)

La convergence faible

Alors $\sup_{x \in \mathbb{R}} \mathcal{E}(x) \leq c_1 n^{-1/5}$

Ce qui nous donne un choix optimal de h_n (condition de proportionnalité) et une borne supérieure de l'erreur quadratique moyenne.

2.4 Question T4

Voir page suivante

Th. On obtient que l'estimateur $\hat{h} \mapsto \int_0^1 \hat{p}_{n,h}^2 dx - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i}^n K\left(\frac{X_i - X_j}{h}\right)$

est un estimateur sans biais de $\mathbb{E}\left[\int_0^1 p_n^2(x) dx\right] - 2 \int_0^1 p_n(x) \mathbb{E}[p_{n,h}(x)] dx$

$$\mathbb{E}\left[\frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i}^n K\left(\frac{X_i - X_j}{h}\right)\right]$$

$$\stackrel{\text{lin}}{\text{esp}} = \frac{2}{n(n-1)h} \sum_{i=1}^n \mathbb{E}\left[\sum_{j \neq i}^n K\left(\frac{X_i - X_j}{h}\right)\right]$$

$$\stackrel{\text{i.i.d}}{(X_i)} = \frac{2}{(n-1)h} \sum_{j=2}^n \mathbb{E}\left[K\left(\frac{X_1 - X_j}{h}\right)\right]$$

$$\stackrel{\text{Transf}}{=} \frac{2}{(n-1)h} \sum_{j=2}^n \int_{[0,1]^2} K\left(\frac{x_1 - x_j}{h}\right) p(x_1, x_j) d(x_1, x_j)$$

$$\stackrel{\text{Fubini et}}{=} \frac{2}{(n-1)h} \sum_{j=2}^n \int_0^1 p_n(x_1) \left[\int_0^1 K\left(\frac{x_1 - x_j}{h}\right) p_n(x_j) dx_j \right] dx_1$$

$$\stackrel{(X_i)_{i \geq 2} \text{ i.i.d}}{=} \frac{2}{(n-1)} \sum_{j=2}^n \int_0^1 p_n(x_1) \mathbb{E}[p_{n,h}(x_1)] dx_1 = \mathbb{E}\left[K\left(\frac{x_1 - X_j}{h}\right)\right] = \mathbb{E}\left[K\left(\frac{X_1 - X_j}{h}\right)\right] = h \cdot \mathbb{E}[p_{n,h}(x)]$$

$$\stackrel{k \text{ paire}}{=} \frac{2}{n(n-1)h} \int_0^1 p_n(x) \mathbb{E}[p_{n,h}(x)] dx$$

$$\text{On obtient bien } \left[\mathbb{E}\left[\int_0^1 \hat{p}_{n,h}^2(x) dx - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i}^n K\left(\frac{X_i - X_j}{h}\right)\right] - \left[\int_0^1 \hat{p}_{n,h}^2(x) dx - 2 \int_0^1 p_n(x) \mathbb{E}[p_{n,h}(x)] dx\right] = 0 \right]$$

L'estimateur est sans biais.