

ESTIMATEUR PAR HISTOGRAMME : RISQUE

Sous de conditions de régularité de la densité f , et en supposant de définir ν en fonction de la taille de l'échantillon N de sorte que $\nu_N \mapsto 0$ quand $N \mapsto +\infty$, nous pouvons démontrer le résultat asymptotique suivant :

$$\text{MISE}_f(\nu) = \underbrace{\frac{\nu^2}{12} \int_{\text{support}} (f'(x))^2 dx + \frac{1}{N\nu}}_{\text{Terme principale du risque}} + \underbrace{\mathcal{O}(\nu^2) + \mathcal{O}\left(\frac{1}{N}\right)}_{\text{Terme résiduel}} \text{ lorsque } N \rightarrow \infty$$

Le terme principale du risque est minimisé pour :

$$\nu_N^{\text{opt}} = \left(\frac{N}{6} \int_{\text{support}} (f'(x))^2 dx \right)^{-1/3}$$

Même si cette fenêtre optimale ne peut pas être déterminée précisément (car f est inconnue), ce résultat nous permet de conclure que lorsque N est grand, la fenêtre optimale ν_N doit être de l'ordre de $N^{-1/3}$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Nous cherchons maintenant à estimer la fenêtre optimale de notre histogramme, en estimant le risque uniquement à partir des observations.

Nous cherchons à définir un estimateur \hat{J} de $MISE_f(\nu) - ||f||_2^2$ qu'il soit sans biais. Pour toute densité f et tout ν . Soit :

$$\begin{aligned} J(\nu, x_1, \dots, x_N) &= MISE_f(\nu) - ||f||_2^2 \\ &= \frac{1}{N\nu} - \frac{N+1}{N\nu} \sum_{j=1}^b p_j^2 \end{aligned}$$

Afin que \hat{J} estimateur de J soit sans biais, il suffit d'avoir un estimateur sans biais de p_j^2 pour tout j , en se rappelant que p_j correspond à la fréquence théorique des observations se situant dans le j -ème intervalle.

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Un approche naïf consiste à estimer p_j^2 par \hat{p}_j^2 , où \hat{p}_j est la fréquence empirique, c'est-à-dire $\hat{p}_j := C_j/N$, en suivant les notations vu plus tôt dans ce cours.

EXERCICE. Calculer le biais de \hat{p}_j^2

Rappel : $C_j \sim \text{Binom}(N, p_j)$

SOLUTION

Etant donné que $C_j \sim \text{Binom}(N, p_j)$, il en découle :

$$\mathbb{E}[\hat{p}_j] = p_j; \text{Var}[\hat{p}_j] = \frac{p_j(1 - p_j)}{N}$$

Et par conséquent :

$$\mathbb{E}[\hat{p}_j^2] = \text{Var}[\hat{p}_j] + (\mathbb{E}[\hat{p}_j])^2 = p_j^2 \left(1 - \frac{1}{N}\right) + \frac{p_j}{N}$$

QUESTION. L'estimateur \hat{p}_j^2 est-il un estimateur sans biais de p_j^2 ?

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Nous pouvons en déduire les observations suivantes :

- \hat{p}_j^2 est un estimateur biaisé de p_j^2
- $\hat{p}_j^2 - \hat{p}_j/N$ est un estimateur sans biais de $p_j^2(1 - 1/N)$
- D'où :

$$\tilde{p}_j^2 := \frac{\hat{p}_j^2 - \hat{p}_j/N}{1 - 1/N} = \frac{N}{N-1} \hat{p}_j^2 - \frac{1}{N-1} \hat{p}_j$$

est un estimateur sans biais de \hat{p}_j^2 .

Ceci nous permet enfin de proposer l'estimateur suivant (rappel : $\sum_{j=1}^b \hat{p}_j = 1$) :

$$\hat{J}(\nu, x_1, \dots, x_N) = \frac{2}{(N-1)\nu} - \frac{N+1}{(N-1)\nu} \sum_{j=1}^b \tilde{p}_j^2, \quad \hat{p}_j = C_j/N$$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

L'estimateur \hat{f} peut être utilisé pour déterminer automatiquement la fenêtre optimale ν par la méthode de validation croisée :

1. Fixer le support $[m, m + l]$ (e.g., $m = \min_i x_i; l = \max_i x_i - m, i = 1, \dots, N$)
2. Initialiser : $b_{CV} = 1, \hat{f}_{CV} = \hat{f}(l/1, x_1, \dots, x_N)$
3. Tant que $b < N$:
 - Calculer $J := \hat{f}(l/b, x_1, \dots, x_N)$
 - Si $J < \hat{f}_{CV}$:
 - $b_{CV} = b$
 - $\hat{f}_{CV} = J$
4. Et enfin : $\nu_{CV} = l/b_{CV}$

→ Reprendre le TP, **Partie II.**

Quels avantages/inconvénients de l'estimateur par histogramme ?

- Très naturel et intuitif → Simple à réaliser et à analyser
- Le résultat dépend fortement de l'échantillon ainsi que de certains paramètres qui peuvent être choisis a priori, comme en particulier la partition considéré du support (ou plus précisément, de la partie du support où l'on possède des données)

Une première idée pour résoudre un des limites de cette méthodes (le choix du support et de sa partition), et celle de supposer que chaque x_i dans notre échantillon est le centre d'une classe de l'histogramme de longueur ν .

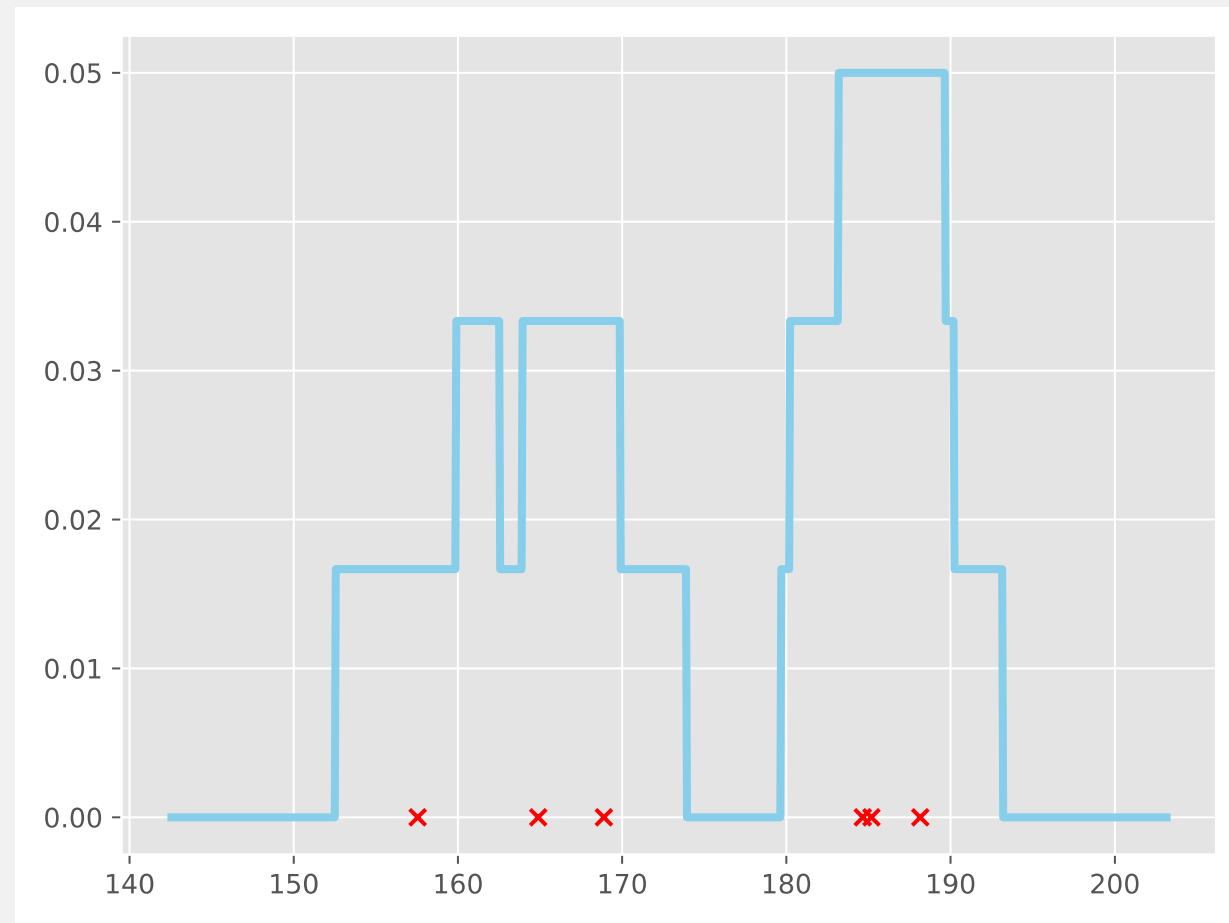
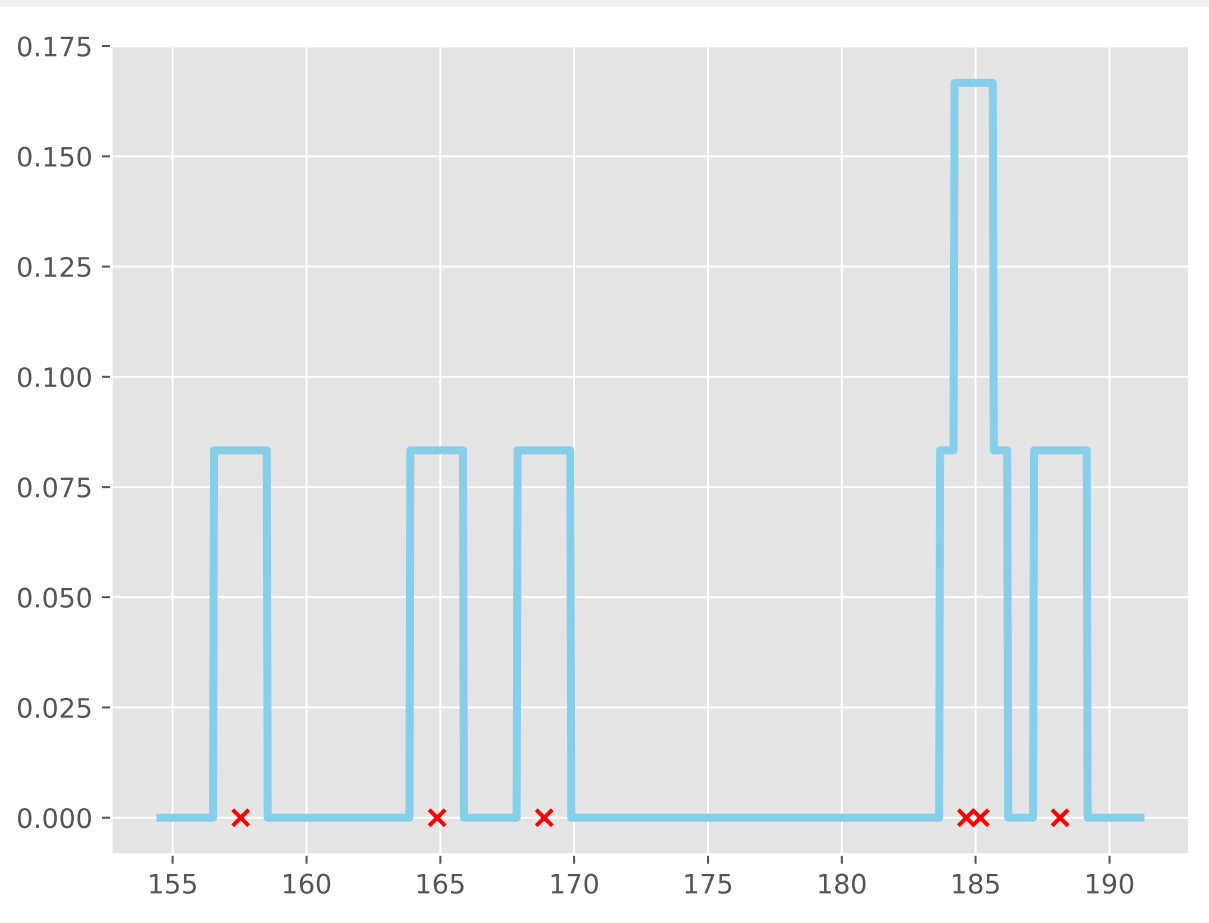
Cela nous amène à l'expression suivante :

$$\hat{f}_\nu^{\mathcal{U}}(x) := \frac{1}{N\nu} \sum_{i=1}^N \mathbb{I}(|x_i - x| \leq \nu/2) = \frac{1}{N\nu} \sum_{i=1}^N \mathbb{I}\left(\frac{|x_i - x|}{\nu} \leq \frac{1}{2}\right)$$

Cela revient, en pratique, à « construire » autour de chaque observation un « bloc » dont l'aire est égale à $1/N$, ce qui conduit à un estimateur toujours constant par morceaux, mais avec des plateaux de longueur variable.

→ Reprendre le TP

ESTIMATEUR PAR HISTOGRAMME : CONCLUSION



Pouvons-nous rendre cette estimation plus « lisse » ?



Estimateur par noyaux