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 $\nu_N \longmapsto 0$ quand $N \longmapsto +\infty$, nous pouvons démontrer le résultat asymptotique suivant :

$$\mathrm{MISE}_f(\nu) = \underbrace{\frac{\nu^2}{12}\int_{\text{support}} \left(f'(x)\right)^2 dx + \frac{1}{N\nu}}_{\text{Terme principale du risque}} + \underbrace{\mathcal{O}(\nu^3) + \mathcal{O}\left(\frac{1}{N}\right)}_{\text{Terme résiduel}} \text{ lorsque } N \to \infty$$

Le terme principale du risque est minimisé pour :

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

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 :

$$J(\nu, x_1, \ldots, x_N) = \mathrm{MISE}_f(\nu) - ||f||_2^2$$

$$= \frac{1}{N\nu} - \frac{N+1}{N\nu} \sum_{j=1}^{b} p_j^2$$

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 corresponds à a 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 : $C_j \sim Binom(N, p_j)$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

SOLUTION

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

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

Et par conséquent :

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

QUESTION. L'estimateur $\hat{p}_j^2$ est-il un estimateur sans biais de 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 : $\sum_{j=1}^{b} \hat{p}_j = 1$) :

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

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

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

1. Poser : $m = \min_i x_i; l = \max_i x_i - m, i = 1, \dots, N$
2. Initialiser : $b_{CV} = 1, \hat{J}_{CV} = \hat{J}(l/1, x_1, \dots, x_N)$
3. Tant que $b < N$:
   - Calculer J $:= \hat{J}(l/b, x_1, \dots, x_N)$
   - Si $j < \hat{J}$:
     - $b_{CV} = b$
     - $\hat{J}_{CV} = J$
4. Et enfin : $v_{CV} = l/b_{CV}$

→ Reprendre le TP, **Partie II.**

ESTIMATEUR PAR HISTOGRAMME : CONCLUSION

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)

ESTIMATEUR PAR HISTOGRAMME : CONCLUSION

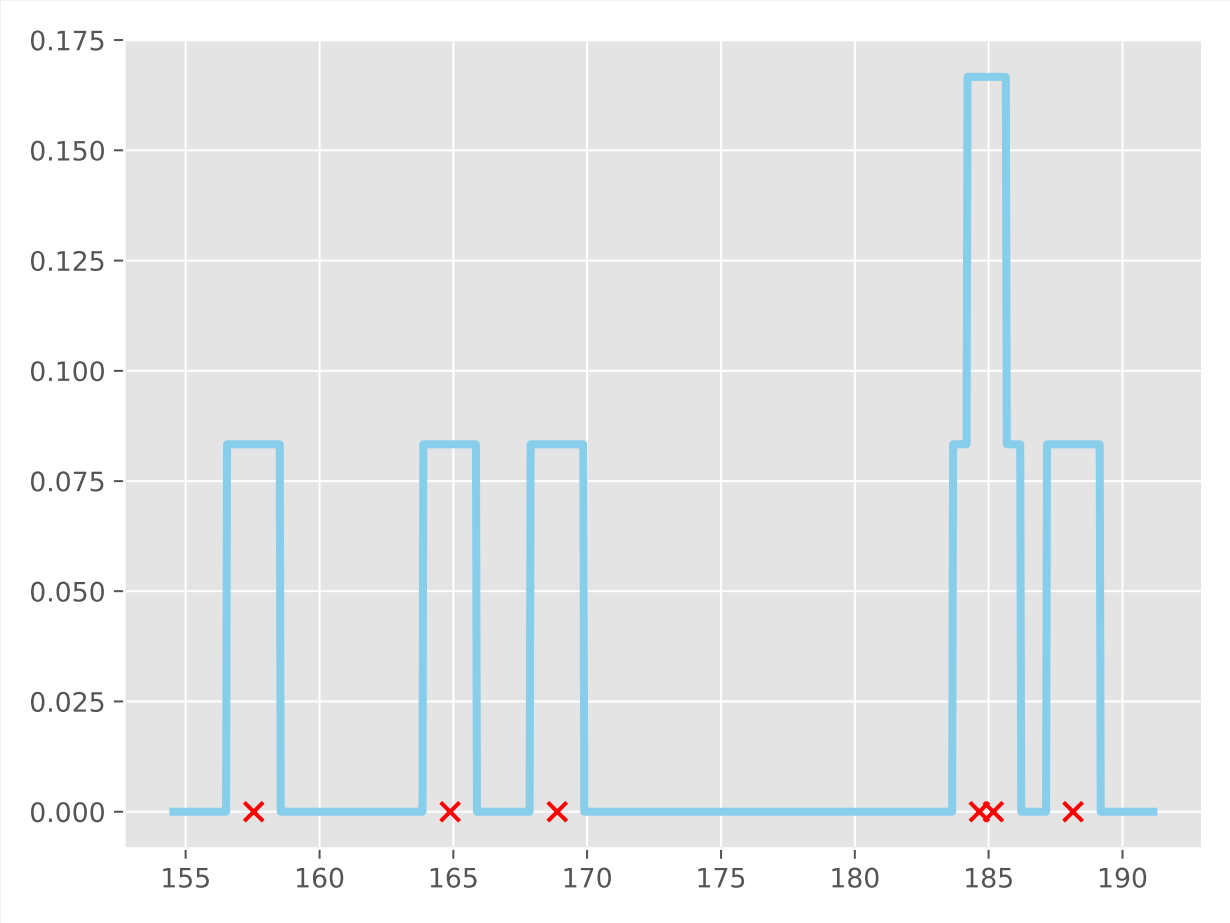
Une première idée pour résoudre un des limites de cette méthodes (le choix du support et de sa partition), et celle de supposer que chaque x_i dans notre échantillon est le centre d'une classe de l'histogramme de longueur ν.

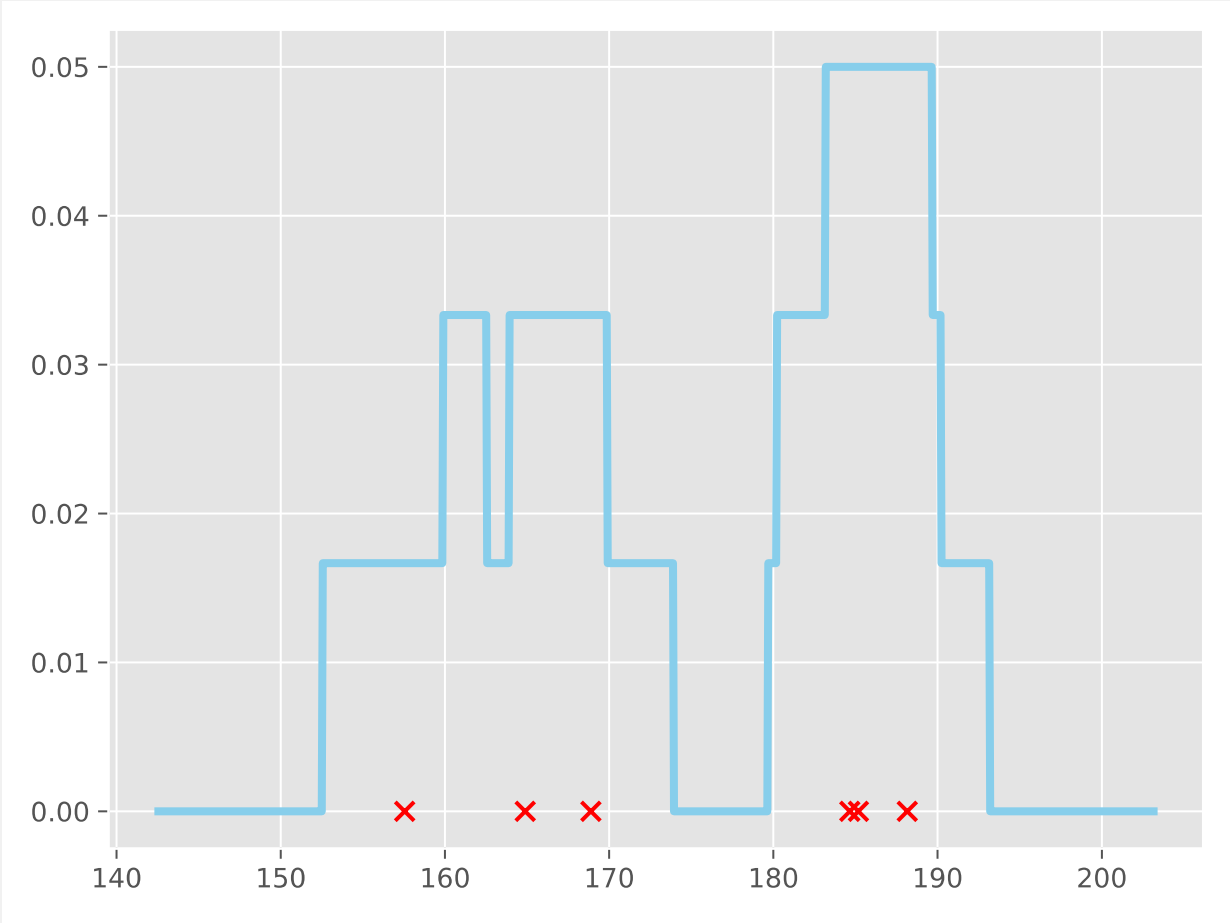
Cela nous amène à l'expression suivante :

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

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.

ESTIMATEUR PAR HISTOGRAMME : CONCLUSION





ESTIMATEUR PAR HISTOGRAMME : CONCLUSION

Pouvons-nous rendre cette estimation plus « lisse » ?

Estimateur par noyaux

4. ESTIMATEUR DE LA DENSITÉ À NOYAU

ESTIMATEUR À NOYAU

L'objectif est de pouvoir fournir une estimation de la densité plus lisse par rapport à celle obtenue par la méthodes des histogrammes. Un avantage est aussi celui de pouvoir intégrer dans notre estimation des propriétés qu'on peut supposer pour la densité d'origine, telle que la continuité, ou dérivabilité.

Qu'est-ce que un noyau ?

Ici un noyau peut être n'importe quelle fonction K qui satisfait les conditions suivantes :

1. $$K(x) \geq 0 \quad \forall x$$
2. $$\int_{\mathbb{R}} K(x)dx = 1$$

ESTIMATEUR À NOYAU

Une fois que notre choix de la fonction K a été faite, soit $\mathcal{D}_N := \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ un échantillon aléatoire composé de N observations de densité « réelle » f. L'estimateur de f à noyau K de taille ν est donné par :

$$\hat{f}_\nu^K(x) := \frac{1}{N\nu} \sum_{i=1}^{N} K\left(\frac{x - x_i}{\nu}\right)$$

Le plus souvent la fonction K est une fonction lisse et symétrique, et ν, comme dans le cas des histogrammes, contrôle l'ampleur du lissage. En pratique, K « lisse » chaque donnée x_i en des petites bosses (dont la forme est définie par la fonction K), puis additionne toutes ces petites bosses pour obtenir l'estimation finale de la densité.

ESTIMATEUR À NOYAU

A noter, l'estimateur vu précédemment, où les petits histogrammes étaient centrés sur chaque donné, est un premier exemple d'estimateur à noyau (même si pas lisse), où la fonction K choisie est $K(z) \coloneqq \mathbb{I}\left(|z| \leq \frac{1}{2}\right)$ (dans ce cas, on a donc un noyau uniforme !).

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

ESTIMATEUR À NOYAU

Propriétés :

- Si K satisfait les propriétés vu précédemment, alors $\hat{f}_{\mathcal{V}}^{K}$ est bien une densité de probabilité

EXERCICE. Démontrer que $\hat{f}_{\mathcal{V}}^{K}$ est une densité

Propriétés :

- Si K satisfait les propriétés vu précédemment, alors $\hat{f}_\nu^K$ est bien une densité de probabilité

SOLUTION.

$$\int \hat{f}_\nu^K(x)dx = \frac{1}{N\nu}\sum_{i=1}^{N}\int K\left(\frac{x-x_i}{\nu}\right)dx$$

$$= \frac{1}{N\nu}\sum_{i=1}^{N}\int K(u)\,\nu du$$

$$= \frac{1}{N\nu}\sum_{i=1}^{N}\nu = 1$$

ESTIMATEUR À NOYAU

Propriétés :

- Si K satisfait les propriétés vu précédemment, alors $\hat{f}_v^K$ est bien une densité de probabilité

- L'estimateur $\hat{f}_v^K$ est continu si K l'est. Il est même p-fois continument différentiable si K l'est.

ESTIMATEUR À NOYAU

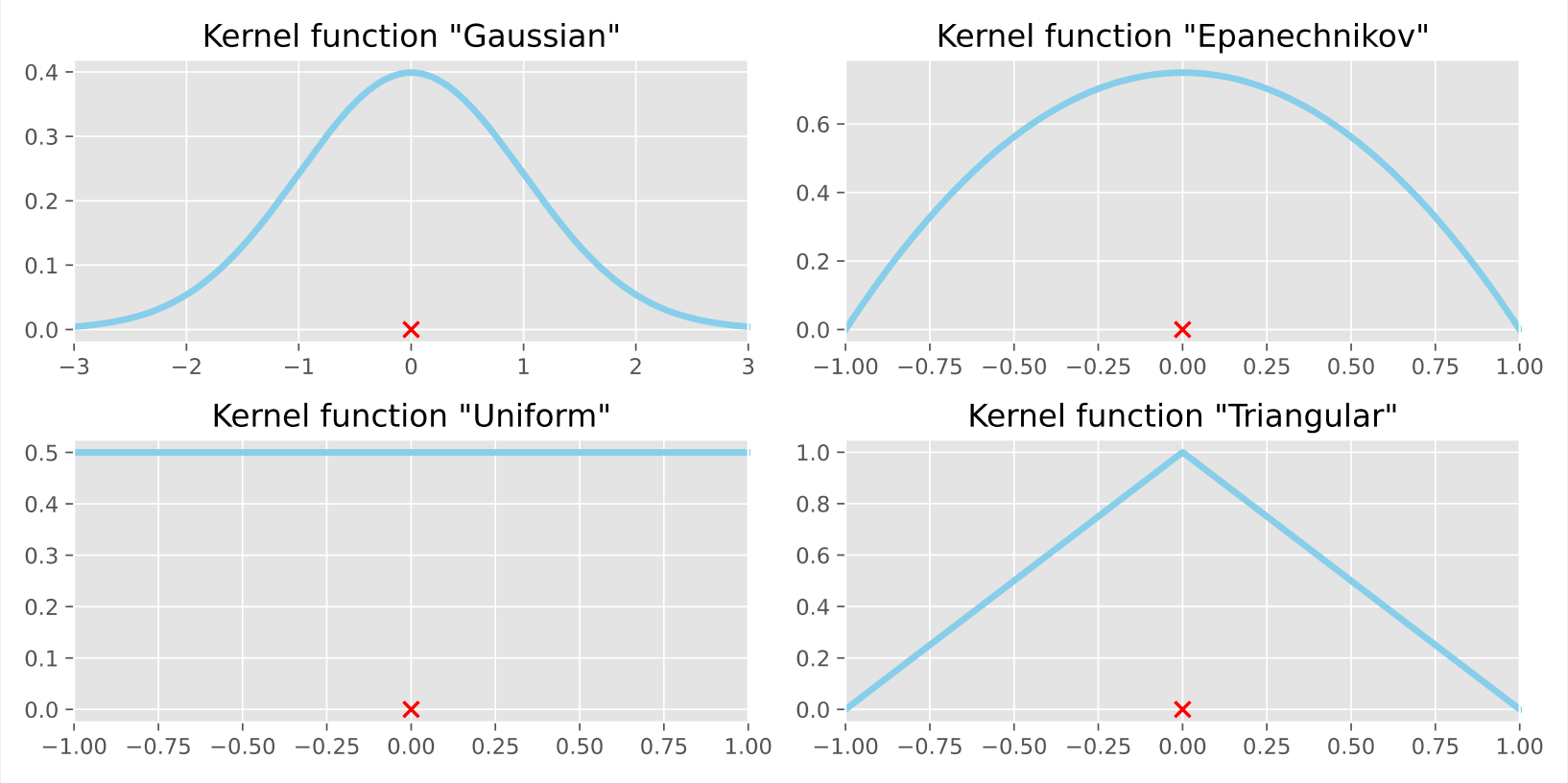
Suivant notre définition, a priori tout fonction K non négative, paire et d'intégrale 1 peut être choisie comme noyau pour estimer une densité f à partir d'un échantillon $\mathcal{D}_N$, mais voici quelques noyau couramment utilisés en pratique (d'autres existes également et sont implémentés dans des libraires classiques en Python) :

- Le noyau gaussien : $K(z) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
- Le noyau d' Epanechnikov : $K(z) \coloneqq \frac{3}{4}(1 - z^2)\mathbb{I}_{[-1,1]}(z)$
- Le noyau triangulaire : $K(z) \coloneqq (1 - |z|)\mathbb{I}_{[-1,1]}(z)$
- Le noyau uniforme : $K(z) \coloneqq \frac{1}{2}\mathbb{I}_{[-1,1]}(z)$

ESTIMATEUR À NOYAU

$$K(z) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$K(z) \coloneqq \frac{3}{4}(1 - z^2)\mathbb{I}_{[-1,1]}(z)$$



$$K(z) \coloneqq \frac{1}{2}\mathbb{I}_{[-1,1]}(z)$$

$$K(z) \coloneqq (1 - |z|)\mathbb{I}_{[-1,1]}(z)$$

ESTIMATEUR À NOYAU

Quels paramètre avons-nous du fixer ?

- ~~La valeur (ou coordonnée) m~~
- ~~Le nombre total d'intervalles (ou boîtes) b~~
- Le noyau K
- La longueur (ou volume) de chaque intervalle/boîte v

Comme nous l'avons fait pour l'estimation par histogrammes, nous allons voir empiriquement l'effet de ces choix à l'aide d'un exemple. Nous allons aussi observer à nouveau combien la taille de l'échantillon va jouer sur notre estimation.

- Télécharger le fichier TP2_Noyau_partie1.ipynb : ibalelli.github.io → Teaching → Modélisation statistique avancée
- Ouvrir un terminal, aller dans le dossier où vous avez enregistré le fichier → jupyter notebook

ESTIMATEUR À NOYAU

Comme dans le cas des histogrammes, nous pouvons faire les observations suivantes à propos du choix de ν :

- Si ν est trop petit, cela entraîne un sous-lissage : le tracé de la densité ressemblera à une combinaison de pics individuels (un pic pour chaque élément de l'échantillon).
- Si ν est trop grand, cela entraîne un sur-lissage : le tracé de la densité ressemblera à une distribution unimodale et cachera toutes les propriétés de la distribution, notamment si elle est multimodale.

Cela nous amène à nouveau à s'interroger sur la possibilité de déterminer le ν optimale, étant donné l'échantillon et un noyau K fixé.

ESTIMATEUR À NOYAU

Une première possibilité est de procéder de façon empirique, et tester plusieurs choix possibles de valeurs de ν sur un intervalle qui nous semble « raisonnable ». Cela revient à faire une *grid search*.

→ Essayer cela dans notre exemple pratique

ESTIMATEUR À NOYAU : RISQUE

Comme pour l'histogramme, nous pouvons aussi regarder le risque de notre estimateur à noyau, ce qui va nous donner des pistes pour répondre à notre question.

Rappel:

$$\mathrm{MSE}_f(x^*, \nu) = \left(\mathbb{E}\left[\hat{f}_\nu^K(x^*)\right] - f(x^*)\right)^2 + \mathrm{Var}\left[\hat{f}_\nu^K(x^*)\right]$$

ESTIMATEUR À NOYAU : RISQUE

Comme pour l'histogramme, nous pouvons aussi regarder le risque de notre estimateur à noyau, ce qui va nous donner des pistes pour répondre à notre question.

Il est possible de démontrer les faits suivants, sous des hypothèse raisonnables de régularité de K :

- La valeur absolue du biais de $\hat{f}_\nu^{K(x)}$, est majoré par $C_1\nu^2$, où C_1 est une constante qui dépends de f'' et K:

$$|\text{biais}_f(x^*, \nu)| := \left|\mathbb{E}\left[\hat{f}_\nu^K(x^*)\right] - f(x^*)\right| \leq C_1\nu^2$$

ESTIMATEUR À NOYAU : RISQUE

Comme pour l'histogramme, nous pouvons aussi regarder le risque de notre estimateur à noyau, ce qui va nous donner des pistes pour répondre à notre question.

Il est possible de démontrer les faits suivants, sous des hypothèse raisonnables de régularité de K :

- La valeur absolue du biais de $\hat{f}_\nu^{K(x)}$, est majoré par $C_1\nu^2$, où C_1 est une constante qui dépends de f'' et K
- La variance de $\hat{f}_\nu^{K(x)}$, $Var\left[\hat{f}_\nu^K(x)\right]$ est majoré par $\frac{C_2}{N\nu}$, où C_2 est une constante qui dépends de f et K

$$\mathrm{Var}\left[\hat{f}_\nu^K(x)\right] \leq \frac{C_2}{N\nu}$$

ESTIMATEUR À NOYAU : RISQUE

Comme pour l'histogramme, nous pouvons aussi regarder le risque de notre estimateur à noyau, ce qui va nous donner des pistes pour répondre à notre question.

Il est possible de démontrer les faits suivants, sous des hypothèse raisonnables de régularité de K :

- La valeur absolue du biais de $\hat{f}_v^{K(x)}$, $\mathbb{E}\left[\hat{f}_v^K(x)\right]$ est majoré par $C_1 v^2$, où C_1 est une constante qui dépends de f'' et K
- La variance de $\hat{f}_v^{K(x)}$, $Var\left[\hat{f}_v^K(x)\right]$ est majoré par $\frac{C_2}{Nv}$, où C_2 est une constante qui dépends de f et K

EXERCICE. Déduire la fenêtre optimale qui minimise le majorant du MSE.

ESTIMATEUR À NOYAU : RISQUE

Comme pour l'histogramme, nous pouvons aussi regarder le risque de notre estimateur à noyau, ce qui va nous donner des pistes pour répondre à notre question.

Il est possible de démontrer les faits suivants, sous des hypothèse raisonnables de régularité de K :

- La valeur absolue du biais de $\hat{f}_\nu^{K(x)}$, $\mathbb{E}\left[\hat{f}_\nu^K(x)\right]$ est majoré par $C_1\nu^2$, où C_1 est une constante qui dépends de f'' et K
- La variance de $\hat{f}_\nu^{K(x)}$, $Var\left[\hat{f}_\nu^K(x)\right]$ est majoré par $\frac{C_2}{N\nu}$, où C_2 est une constante qui dépends de f et K

SOLUTION.

- D'après la définition du MSE et les majorations données, on sait que :

$$\mathrm{MSE}_f(x^*, \nu) \leq C_1^2\nu^4 + \frac{C_2}{N\nu} := g(\nu)$$

$$\frac{dg(\nu)}{d\nu} = 4C_1^2\nu^3 - \frac{C_2}{N\nu^2} \Rightarrow \frac{dg(\nu)}{d\nu} = 0 \Leftrightarrow \nu = \left(\frac{C_2}{4C_1^2}\right)^{1/5} N^{-1/5}$$

ESTIMATEUR À NOYAU : FENÊTRE OPTIMALE

Nous pouvons à nouveau définir une méthode automatique pour estimer la fenêtre optimale, de la même manière que pour le cas des histogrammes. Dans ce cas, nous utiliserons l'estimateur sans biais de $\hat{J}(\nu) = MISE_f(\nu) - ||f||_2^2$:

$$\hat{J}_K(\nu, x_1, \ldots, x_N) := ||\hat{f}_\nu^K||_2^2 - \frac{2}{N(N-1)\nu} \sum_{i=1}^{N} \sum_{j \neq i} K\left(\frac{x_i - x_j}{\nu}\right)$$

ESTIMATEUR À NOYAU : FENÊTRE OPTIMALE

Nous pouvons à nouveau définir une méthode automatique pour estimer la fenêtre optimale, de la même manière que pour le cas des histogrammes. Dans ce cas, nous utiliserons l'estimateur asymptotiquement sans biais de $\hat{J}(\nu) = MISE_f(\nu) - ||f||_2^2$:

$$\hat{J}_K(\nu, x_1, \ldots, x_N) := ||\hat{f}_\nu^K||_2^2 - \frac{2}{N(N-1)\nu} \sum_{i=1}^{N} \sum_{j \neq i} K\left(\frac{x_i - x_j}{\nu}\right)$$

$$\frac{1}{N\nu^2} \|K\|_2^2 = \frac{1}{N\nu^2} \underbrace{\int_{\mathbb{R}} (K(x))^2 \, dx}_{R(K)}$$