I.2. 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 !).

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

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

ESTIMATEUR À NOYAU

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

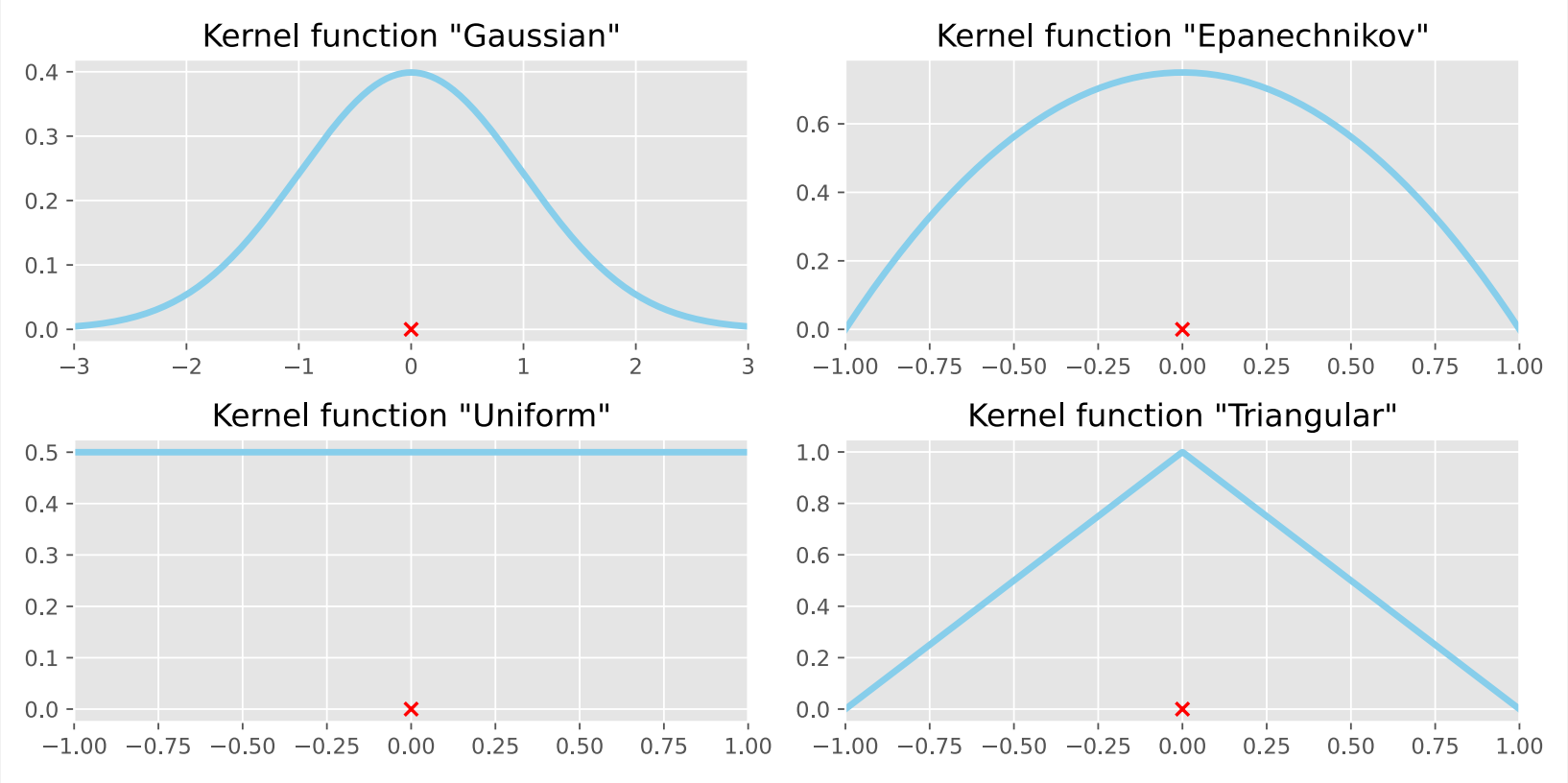
Suivant notre définition, a priori tout fonction K non négative, paire et d'intégrale 1 peut être choisie comme noyau pour estimer une densité f à partir d'un échantillon $\mathcal{D}_N$, mais voici quelques noyau couramment utilisés en pratique (d'autres existes également et sont implémentés dans des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

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 $\hat{f}_\nu^{K(x)}$, $\mathbb{E}\left[\hat{f}_\nu^K(x)\right]$ 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

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

Une première solution consiste à définir une méthode automatique pour estimer la fenêtre optimale, de la même manière que pour le cas des histogrammes, avec la validation croisée. Pour cela, il faut définir un estimateur sans biais, $\hat{J}(\nu)$, de la quantité : $J(\nu) = MISE_f(\nu) - ||f||_2^2$

On propose l'estimateur suivant :

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

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

ESTIMATEUR À NOYAU : FENÊTRE OPTIMALE

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

ESTIMATEUR À NOYAU : FENÊTRE OPTIMALE

- Le noyau **gaussien** : $K(z) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \rightarrow R(K) = 1/(2\sqrt{\pi})$
- Le noyau d' **Epanechnikov** : $K(z) \coloneqq \frac{3}{4}(1 - z^2)\mathbb{I}_{[-1,1]}(z) \rightarrow R(K) = 3/5$
- Le noyau **triangulaire** : $K(z) \coloneqq (1 - |z|)\mathbb{I}_{[-1,1]}(z) \rightarrow R(K) = 2/3$
- Le noyau **uniforme** : $K(z) \coloneqq \frac{1}{2}\mathbb{I}_{[-1,1]}(z) \rightarrow R(K) = 1/2$

ESTIMATEUR À NOYAU : FENÊTRE OPTIMALE

Cependant, cette méthode est en pratique difficile à réaliser et peut être très lente en fonction de la complexité du noyau.

Il y a donc d'autres méthodes, plus directes, pour choisir la fenêtre optimale.

Lorsque f esté suffisamment lisse, on peut faire l'hypothèse que les données suivent une loi normale (pouvez vous voir le paradoxe ici ?), ce qui nous permet de dériver la *règle de Silverman*.

Sous une hypothèse de Gaussianité, la valeur de ν minimisant le $MISE$ est donnée par :

$$\nu_N^{\text{opt}} := \hat{\sigma} \left(\frac{3}{4} N \right)^{-1/5}$$

ATTENTION: en pratique, nous remplaçons souvent la déviation standard empirique, $\hat{\sigma}$, par :

$$A := \min \left(\hat{\sigma}, \frac{IQR}{1.34} \right)$$

→ Retour au TP !