# 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 ?**

# 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 $\nu$ est donné par :

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

# 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 $\nu$ 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 $\nu$, 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

# 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 toute 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 existent également et sont implémentés dans des libraires classiques en Python) :

# ESTIMATEUR À NOYAU

Suivant notre définition, a priori toute 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 existent également et sont implémentés dans des libraires classiques en Python) :

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

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

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

# ESTIMATEUR À NOYAU

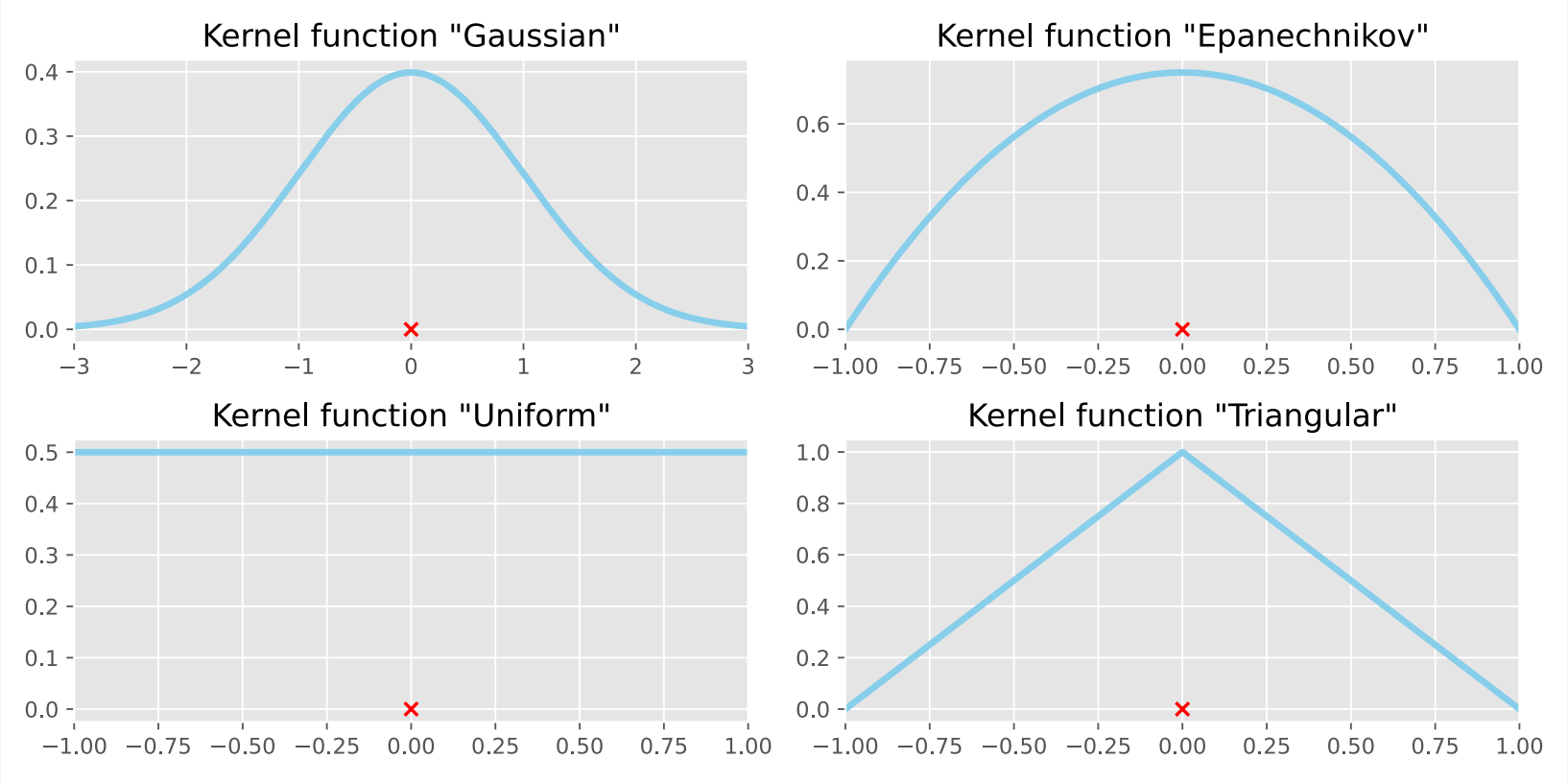
Suivant notre définition, a priori tout fonction $K$ non négative, paire et d'intégrale 1 peut être choisie comme noyau pour estimer une densité $f$ à partir d'un échantillon $\mathcal{D}_N$, mais voici quelques noyau couramment utilisés en pratique (d'autres existes également et sont implémentés dans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

# 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 $\nu$

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