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.

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:

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

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

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 $\hat{J}(\nu) = MISE_f(\nu) - ||f||_2^2$:

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

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 $\hat{J}(\nu) = MISE_f(\nu) - ||f||_2^2$:

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

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

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 $\hat{J}(\nu) = MISE_f(\nu) - ||f||_2^2$:

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

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

ESTIMATEUR À NOYAU : FENÊTRE OPTIMALE

- Le noyau **gaussien** : $K(z) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ → $\boxed{R(K) = 1/(2\sqrt{\pi})}$
- 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** : $K(z) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \rightarrow \boxed{R(K) = 1/(\sqrt{2\pi})}$

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

- Le noyau **triangulaire** : $K(z) \coloneqq (1 - |z|)\mathbb{I}_{[-1,1]}(z) \rightarrow \boxed{R(K) = 2/3}$

- Le noyau **uniforme** : $K(z) \coloneqq \frac{1}{2}\mathbb{I}_{[-1,1]}(z) \rightarrow \boxed{R(K) = 1/2}$

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 directs, de choisir la fenêtre optimale. La règle plus simple pour cela est appelée la règle de Silverman, qui repose sur l'hypothèse que les données suivent une loi normale (pouvez vous voir le paradoxe ici ?).

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 directs, de choisir la fenêtre optimale. La règle plus simple pour cela est appelée la règle de Silverman, qui repose sur l'hypothèse que les données suivent une loi normale (pouvez vous voir le paradoxe ici ?).

Dans ce cas, la valeur de ν minimisant le $MISE$ est donnée par :

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

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 !