Fiche 8 - TIAD - M1 IM
Estimation à noyaux d'une densité

Tran Viet Chi, `chi.tran@math.univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Programmation sous R)
Dans un premier temps, on va générer un échantillon de densité connue (une $\mathcal{N}(0,1)$) que l'on utilisera pour reconstruire la densité et la comparer au résultat théorique. Dans un second temps, on considèrera l'estimation de densité sur un jeu de vraies données.

Partie A

1. Générer un échantillon `x` de n=1000 v.a. i.i.d. $X_1, \dots X_n$ de loi $\mathcal{N}(0,1)$. Représenter l'histogramme des données et superposer la densité estimée par la fonction `density` de R. Dans la suite de l'exercice, on va reprogrammer cet estimateur "à la main".

2. On considèrera dans la suite 5 noyaux : la densité de la loi $\mathcal{N}(0,1)$ et les 4 autres noyaux suivants

$$\texttt{Tri}(x) = (1-|x|)\mathbf{1}_{|x|\leq 1}, \qquad \texttt{Rect}(x) = \frac{\mathbf{1}_{[-1,1]}(x)}{2}, \qquad \texttt{EP}(x) = \frac{3}{4}(1-x^2)\mathbf{1}_{|x|\leq 1}, \qquad \texttt{sinc}(x) = \frac{\sin(x)}{x}.$$

Représenter graphiquement ces noyaux.

3. On définit une grille 500 pas sur laquelle sera calculée la densité. On considère l'estimation sur l'intervalle $[a,b]$ où $a = \min(X_i) - E$, $b = \max(X_i) + E$ où $E = \max(X_i) - \min(X_i)$ est l'étendue des valeurs de l'échantillon.

Coder une fonction `KernelEst` qui :
- prend en arguments : 1) `x` l'échantillon dont il faut reconstruire la densité f, 2) `h` la fenêtre, 3) une chaîne de caractère, `Tri`, `Rect`, `EP`, `Gaus`, `sinc` qui indiquera quel noyau utiliser pour l'estimation de la densité, 4) le vecteur `abs` des abscisses des 500 points où l'estimateur de la densité sera calculé.
- calcule l'estimateur à noyaux de la densité $\widehat{f}$, dont on rappelle la définition :

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big).$$

- renvoie les valeurs $\widehat{f}(x)$ pour les points x de la grille définie ci-dessus, ainsi que les valeurs de ces points x.

4. Utiliser la fonction précédente pour estimer la densité de `x` en utilisant les noyaux `EP` et `sinc` avec `h=0.6`. Commenter. Comment améliorer l'estimation avec le noyau `sinc` ?

5. Avec `h=0.8`, comparer l'estimation de la densité de `x` en utilisant les noyaux `EP`, `Rect`, `Tri` et le noyau gaussien.

6. Calculer un estimateur Monte-Carlo du MSE en x, basé sur N simulations indépendantes de `x` :

$$\frac{1}{N}\sum_{j=1}^{N} |\widehat{f}^{(j)}(x) - f(x)|^2.$$

Pour chacun des noyaux, sortir des statistiques résumant la distribution des MSE(x).

7. En déduire le MISE pour chacun des noyaux.

8. Tracer l'évolution du MISE pour $h = n^{1/5}$ en fonction de n. Pour cela, on simulera un échantillon de 1000 v.a. i.i.d. $\mathcal{N}(0,1)$ et on en considèrera les sous échantillons $X_1, \dots X_n$ pour n variant de 100 à 1000

par pas de 10.

9. On souhaite maintenant minimiser le MISE en h, pour un choix de noyau fixé. Rappeler quelle est la fonction $J(h)$ de h qu'il suffit de minimiser pour trouver $\arg\min_h MISE(h)$? Donner un estimateur sans biais $CV(h)$ de $J(h)$.
Programmer une fonction qui renvoie la valeur de $CV(h)$ et la valeur de h qui minimise cette fonction lorsque l'on fait varier h sur une grille de N pas espacés de $\frac{(b-a)}{10\ N}$ où a et b sont définis à la question 3) et où $N = \frac{n}{2}$ si $n \leq 100$ et $\frac{n}{4}$ si $n > 100$.

10. Quelle est la fenêtre optimale au sens du MISE pour estimer la densité de `x` ?

11. Tracer la fenêtre optimale en fonction de la taille n de l'échantillon. Pour cela, on simulera un échantillon de 1000 v.a. i.i.d. $\mathcal{N}(0,1)$ et on déterminera la fenêtre optimale au sens du MISE pour les sous échantillons $X_1, \ldots X_n$ pour n variant de 100 à 1000 par pas de 10.

Partie B

On cherche maintennat à étudier les données correspondant au mouvement de 82 galaxies.

1. Charger les données en tapant les commandes

```
library(MASS)
help(galaxies)
```

2. Donner quelques statistiques descriptives.

3. Obtenir les estimations de la densité en utilisant les fonctions pré-programmées ou programmées de la partie A. Quelle est la fenêtre optimale ?