Exercice 1

Dans un premier temps, on va générer un échantillon $X_1, \ldots, X_n$ de densité connue (une loi $N(0,1)$) 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 1

1. Générer un échantillon x de $n = 1000$ v.a. i.i.d $X_1, \ldots, X_n$ de loi $N(0,1)$.
2. On considèrera dans la suite 5 noyaux : la densité de la loi $N(0,1)$ et les 4 autres noyaux suivants :

   $$\text{Tri}(x) = (1-|x|)\,\mathbb{1}(|x| \leq 1),\ \text{Rect}(x) = \frac{1}{2}\,\mathbb{1}(|x| \leq 1), \text{EP}(x) = \frac{3}{4}(1-x^2)\,\mathbb{1}(|x| \leq 1), \text{sinc}(x) = \frac{\sin(x)}{x}.$$

   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 *R 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 $\hat{f}$, dont on rappelle la définition

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

   — renvoie les valeurs $\hat{f}(x)$ 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 :

   $$\text{MSE}(x) = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{f}^{(j)}(x) - f(x)\right)^2$$

   Pour chacun des noyaux, donner 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. $N(0,1)$ et on en considèrera les sous échantillons $X_1, \ldots, X_n$ pour n variant de 100 à 1000 par pas de 10. 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 $\arg\min_h \text{MISE}(h)$? Donner un estimateur sans biais $\text{CV}(h)$ 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 $\frac{(b-a)}{10N}$ où a et b sont définis à la question 3) et où $N = n/2$ si $n \leq 100$ et $n/4$ si $n > 100$.

Partie 2
On cherche maintenant à étudier les données correspondant au mouvement de 82 galaxies.

1. Charger les données en tapant les commandes
   library(MASS)
   data(galaxies)
2. Donner quelques statistiques descriptives.
3. Obtenir les estimations de la densité en utilisant les fonctions programmées de la partie 1. Quelle est la fenêtre optimale ?

Exercice 2

1. Simuler un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille 1000 où les X_i proviennent d'une loi uniforme sur $[0, 1]$ et pour $i = 1, \ldots, n$
   $$Y_i = m(X_i) + \varepsilon_i$$
   avec ε_i une suite de variables aléatoires indépendantes de loi $N(0, \sigma^2 = 0.02)$ et
   $$m(X_i) = \sin^3(2\pi X_i^3).$$
2. Estimer la fonction de régression m par
   — un estimateur à noyau de Nadaraya-Watson,
   — par une régression polynomiale locale,
   — par une régression spline.
3. Déterminer pour l'estimateur de Nadaraya-Waston la taille de la fenêtre/le paramètre de lissage par la méthode de la validation croisée.
4. Déterminer un intervalle de confiance de la fonction de lien/de la fonction de régression m de niveau de confiance 95%.

Exercice 3
Dans cet exercice, nous appliquons les modèles de régression non paramétrique à des données acquises à hautes fréquences (10 Hz) et développons des modèles de régression fonctionnelle afin de proposer un outil de surveillance de la qualité des eaux (biomonitoring) et de mesures des effets du réchauffement global sur la planète.

La fréquence d'échantillonnage est fixée à une mesure toutes les 0.1 s. Pour avoir un nombre d'animaux représentatif, nous travaillons sur un groupe de 16 animaux. La première colonne correspond aux huîtres (numérotées de 1 à 16), la deuxième à l'amplitude d'ouverture en mm, et la troisième au inter-temps en ms. A noter que dans cette configuration, avec 16 huîtres chaque animal est interrogé toute les 1.6 secondes puisque la fréquence d'échantillonnage est de 10 Hz. Les mesures de l'activité valvaire commencent à partir de minuit tous les jours.

1. Représenter graphiquement pour l'huître numéro 6 la distance en mm entre les 2 électrodes sur 24 h.
2. Estimer la fonction de régression par l'estimateur de Nadaraya-Watson pour les données brutes de la même huître avec pour choix de taille de fenêtre fixé $h_n = 0.001$ et pour noyau le noyau gaussien.
3. Déterminer la taille de la fenêtre par la méthode de la validation croisée.
4. Effectuer une comparaison avec la méthode de la régression polynomiale locale et la régression spline.

Exercice 4

1. Déterminer l'estimateur de la fonction de lien par la méthode de Nadaraya-Watson, de la régression polynomiale locale et de la régression spline sur les données *faithful* et *Cars*.