Feuille d'exercices

Exercice 1 (Propriétés statistique de Parzen–Rosenblatt)**.**
On s'intéresse ici aux propriétés statistiques de l'estimateur de Parzen–Rosenblatt.

a) Rappelez l'expression de cet estimateur.

b) Montrez que si f est au moins $\mathscr{C}^2$ alors pour tout $x \in \mathbb{R}$

$$\text{Biais}\{\hat{f}_h(x)\} = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2), \qquad h \to 0,$$

où $\mu_2(K) = \int u^2 K(u)\mathrm{d}u$.

c) Montrez que pour $K \in L^2$, on a pour tout $x \in \mathbb{R}$

$$\text{Var}\{\hat{f}_h(x)\} = \frac{1}{nh} f(x) \int K(u)^2 \mathrm{d}u + o\left(\frac{1}{nh}\right), \qquad nh \to \infty.$$

d) En déduire une expression approchée pour la MISE.

e) Trouvez la fenêtre optimale minimisant cette MISE approchée.

Exercice 2 (La règle empirique de Silverman)**.**
Dans cet exercice nous allons essayer de mieux comprendre ce qu'il se cache derrière la règle empirique de Silverman et voir quelques modifications de cette dernière.

a) Soit $f \sim N(\mu, \sigma^2)$. Montrez que

$$\int f''(x)^2 \mathrm{d}x = \frac{3}{8\sigma^5\sqrt{\pi}}.$$

Astuce : on utilisera le fait que pour $Y \sim N(\mu, \sigma^2)$, $\mathbb{E}\{(X-\mu)^{2k}\} = (2k)!\sigma^{2k}/(2^k k!)$.

b) En déduire que le choix par défaut de la fenêtre selon Silverman est donné par

$$h_{\text{Silverman}} = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5}, \qquad \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

c) Expliquez l'intuition derrière la formule suivante

$$h_* = \left(\frac{4}{3n}\right)^{1/5} \min\left(\hat{\sigma}, \frac{X_{[3n/4]} - X_{[n/4]}}{1.349}\right),$$

où $X_{[np]}$ représente la $[np]$–ième statistique d'ordre, i.e., la $[np]$–ième plus petite valeur de l'échantillon $X_1, \ldots, X_n$.

Exercice 3 (Validation croisée « Leave one out »)**.**
On suppose que la vraie densité vérifie $f \in L^2$ et on pose $h > 0$.

a) Rappelez l'expression de ce type de validation croisée pour l'estimateur de Parzen–Rosenblatt.

b) Montrez que

$$\mathbb{E}\{CV(h)\} = MISE(h) - \int f(x)^2 \mathrm{d}x.$$

c) Qu'en déduisez vous ?

Exercice 4 (Old faithful geyser)**.**
Dans cet exercice nous allons mettre tout ce que nous avons vu sur l'estimation non paramétrique d'une densité de probabilité en s'appuyant sur le jeu de données `old faithfull geyser`. Ce jeu de données collecte (entre autre) le temps d'attente entre deux éruptions du geyser Old Faithful situté dans le parc de Yellowstone.

a) Importez le jeu de données et renseignez vous sur ce dernier via les commandes `R`

```
data(faithful)
?faithful
```

```
data(faithful)
```

b) Lisez la documentation de la fonction `density`.

c) Exécutez les commandes suivantes, dites ce qu'elles font et commentez les résultats

```
par(mfrow = c(1, 3), mar = c(4, 5, 0.5, 0))
for (bandwidth in c(0.5, 10, 4)){
    plot(density(faithful$waiting, kernel = "gaussian", bw = bandwidth),
         main = "")
    rug(faithful$waiting)
}
```

Exercice 5 (Mélange de gaussiennes)**.**
Soit la fonction

$$f(x) = 0.3\varphi(x) + \frac{0.7}{0.3}\varphi\left(\frac{x-1}{0.3}\right), \qquad x \in \mathbb{R},$$

où $\varphi(\cdot)$ correspond à la densité d'une $N(0,1)$.

a) Montrez que f est une densité de probabilité.

b) Ecrivez une fonction `R` qui génère un n–échantillon (iid) selon cette loi.

c) Simulez un n–échantillon (n choisi par vos soins) et obtenez une estimation de la densité. Vosu choisirez une fenêtre « optimale à l'oeil ».

d) Sur un même graphqiue, comparer cette estimation à la densité théorique.

Exercice 6 (Nadaraya–Watson)**.**
Dans cet exercice, nous allons retrouver la forme de l'estimateur de Nadaraya–Watson pour la régression non paramétrique.

a) Soit K_1 et K_2 deux noyaux sur $\mathbb{R}$ montrez que le noyau $(x, y) \mapsto K_1(x)K_2(y)$ est un noyau sur $\mathbb{R}^2$.

b) Considérons l'estimateur de la densité bivariée $f(x, y)$ suivant

$$\hat{f}_{h_1,h_2}(x, y) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K_{h_1}\left(\frac{X_i - x}{h_1}\right) K_{h_2}\left(\frac{Y_i - y}{h_2}\right).$$

Montrez que

$$\int y\hat{f}_{h_1,h_2}(x, y)\mathrm{d}y = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}\left(\frac{X_i - x}{h_1}\right) Y_i.$$

c) En déduire l'expression de l'estimateur de Nadaraya–Watson pour la régression non paramétrique.

d) Lisez la documentation de la fonction `ksmooth` et analysez le code suivant

```
data(faithful)
attach(faithful)
plot(eruptions, waiting)
fit <- ksmooth(eruptions, waiting, kernel = "normal")
lines(fit, col = "seagreen3", lwd = 2)
```

e) Jouez un peu avec l'argument `bandwidth` pour faire le lien avec le cours.

f) Ecrivez un bout de code `R` permettant de choisir une fenêtre adaptée par *leave-one-out*.