Université Paris-Dauphine – Année 2018-2019
Executive Master : Régression non-paramétrique

Attention ! *Il n'est pas nécessaire de traiter toutes les questions pour obtenir une bonne évaluation. Le devoir est conçu pour un travail personnel de 2 - 3 heures. Certaines questions sont exploratoires et n'admettent pas nécessairement de solution optimale unique.*

Modalités : A envoyer par mail (convertir au format pdf) avant le 8 janvier 2019 à l'adresse `celine.duval@parisdescartes.fr`

Situation

On dispose d'un jeu de données $(X_i, Y_i)_{1 \leq i \leq 10^4}$ où les X_i et les Y_i sont idéalisées comme des (réalisations de) variables aléatoires réelles admettant la représentation

$$Y_i = r(X_i) + \xi_i, \quad i = 1, \dots 10^4,$$

où les ξ_i sont indépendantes et identiquement distribuées, admettant une densité μ vérifiant $\mathbb{E}[\xi_1] = 0$ et $\mathbb{E}[\xi_1^2] = \sigma^2 > 0$. Les X_i sont indépendantes et identiquement distribuées de densité $g : [0,1] \to \mathbb{R}$, et indépendantes des ξ_i. La fonction $r : [0,1] \to \mathbb{R}$ vérifie $|r(x)| \leq 6$ pour tout $x \in [0,1]$. Les objectifs sont :

1. Reconstruire $x \mapsto g(x)$ graphiquement et étudier si g est la densité uniforme ou non.
2. Reconstruire $x \mapsto r(x)$ graphiquement.
3. Explorer les propriétés de $x \mapsto \mu(x)$ et estimer σ^2.

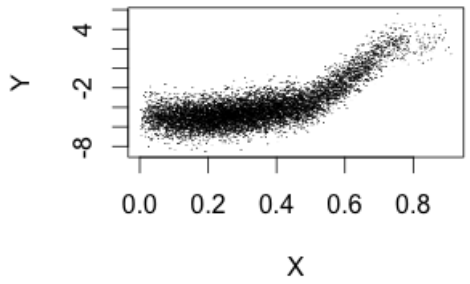

FIGURE 1 – Jeu de données `DataReg` représentant le vecteur $Z^{10^4} = (X_i, Y_i)_{1 \leq i \leq 10^4}$ (X_i en abscisse et Y_i en ordonnée).

Les valeurs du vecteur Z^{10^4} se trouvent dans le fichier `DataReg.csv`. La première colonne correspond aux X_i et la seconde colonne correspond aux Y_i.

1 Exploration des propriétés de $g(x)$

1. Construire un estimateur non-paramétrique $\widehat{g}_{n,h}(x)$ de $g(x)$ pour une fenêtre de lissage $h > 0$ donnée et représenter graphiquement $x \mapsto \widehat{g}_{n,h}(x)$ pour différentes valeurs de h que vous choisirez. On discutera la raison pour laquelle ce choix est important et ce qui se produit si h est mal choisi.

2. Représenter graphiquement $x \mapsto \widehat{g}_{n,\widehat{h}_n}(x)$, où $\widehat{h}_n$ est la fenêtre donnée par validation croisée ou par une autre méthode.
3. Implémenter un QQ-plot pour vérifier empiriquement l'hypothèse $g(x) = 1$ pour tout $x \in [0, 1]$. L'hypothèse selon laquelle g est uniforme semble-t-elle raisonnable ?

2 Reconstruction de $r(x)$

1. Est-il plausible de penser que la fonction r est linéaire ? Si on voulait implémenter un modèle linéaire par morceaux sur ces données, en quoi l'estimation non paramétrique de r serait une première étape nécessaire ?
2. Construire un estimateur non-paramétrique $\widehat{r}_{n,h}(x)$ de $r(x)$ pour une fenêtre de lissage $h > 0$ donnée et représenter graphiquement $x \mapsto \widehat{r}_{n,h}(x)$ pour différentes valeurs de h.
3. Discuter du choix automatique de la fenêtre. On distinguera deux cas :
   (a) Un cas où le h est le même pour tout x.
   (b) le cas ou le choix du h dépend localement de la fonction à estimer.

3 Étude de la loi des ξ_i

On considère deux estimateurs

$$U_n = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 \quad \text{et} \quad V_n = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (\tilde{Y}_{i+1} - \tilde{Y}_i)^2,$$

où $(\tilde{Y}_i)_i$ est obtenu en ordonnant l'échantillon $(Y_i)_i$ selon la permutation qui ordonne par ordre croissant les $(X_i)_i$ (Commande R pour générer $(\tilde{Y}_i)_i$: `tilde_Y=Y[order(X)]`).

1. Implémenter ces estimateurs sur le jeu de données.
2. Justifier (heuristiquement, à l'aide d'un argument mathématique ou bien empiriquement) ce qu'estiment ces deux quantités U_n et V_n.
   Indication : les données ont été générées avec $\sigma^2 = 1$.
3. On cherche à estimer $x \mapsto \mu(x)$. Pour cela, on coupe l'échantillon en deux, selon que $i \in \mathcal{J}_- = \{1, \dots, 5 \times 10^4\}$ ou que $i \in \mathcal{J}_+ = \{5 \times 10^4 + 1, 10^4\}$. On note $\widehat{r}^{(-)}_{n,h}(x)$ (pour un choix de h établi à la question précédente) l'estimateur construit à l'aide de $(X_i, Y_i)_{1 \le i \le 5 \times 10^4}$ et on pose

$$\widetilde{Y}_i = Y_i - \widehat{r}^{(-)}_{n,h}(X_i), \quad i \in \mathcal{J}_+.$$

   Quelle est la distribution approximative de $\widetilde{Y}_i$?
4. En déduire un estimateur de $x \mapsto \mu(x)$ et l'implémenter graphiquement.
5. *(Facultatif.)* La densité $x \mapsto \mu(x)$ peut-elle être gaussienne ? Proposer un protocole numérique pour le vérifier empiriquement et l'implémenter.