Estimation non paramétrique : densité et régression

Executive Master Statistique et Big Data, Février-Avril 2024

Rapport à envoyer par mail (format pdf ou html produit à partir de Rmarkdown) avant le 2 Mai 2024 à l'adresse `sarah.ouadah@agroparistech.fr`

Contexte

On dispose de données $(x_i, y_i)_{1\leq i\leq 5000}$ où les x_i et les y_i sont les réalisations de variables aléatoires réelles $(X_i, Y_i)_{1\leq i\leq 5000}$ admettant la représentation

$$Y_i = r(X_i) + \xi_i, \quad i = 1, \dots 5000,$$

où

- les ξ_i sont indépendantes et identiquement distribuées, $\mathbb{E}[\xi_1] = 0$ et $\mathbb{E}[\xi_1^2] = \sigma^2$.
- les X_i sont indépendantes et identiquement distribuées de densité g, et indépendantes des ξ_i.

`Data.text` : la première colonne correspond aux x_i et la seconde colonne correspond aux y_i.

Ce projet porte sur l'étude de la densité g et la fonction de régression r.

Indications

Les sections et questions suivantes sont là pour vous guider, à vous de raconter votre analyse dans l'ordre et la façon que vous choisirez.

La question 2.3 sur le choix de la fenêtre de lissage des estimateurs joue un rôle central dans ce projet. Une bibliographie sur le sujet et une restitution claire sont attendues.

La rédaction est considérée comme un élément clé.

Par ailleurs, une discussion sur les avantages, inconvénients des méthodes utilisées ainsi que les difficultés rencontrées sera appréciée. Il serait également intéressant d'apporter une réflexion sur une possible application de l'estimation non paramétrique dans le cadre de vos activités professionnelles (décrire un contexte, des données que vous pourriez rencontrer).

Il n'y a pas de nombre de pages requis, en revanche le rapport ne doit pas dépasser 15 pages graphiques compris.

1 Estimation de la densité g

2.1. Vérifier via un histogramme et un QQ-plot ou si possible un test d'ajustement à une loi si la distribution g correspond à une loi de probabilité "connue".

2.2. Construire un estimateur à noyau $\widehat{g}_{n,h}(x)$ de $g(x)$ pour une fenêtre de lissage $h > 0$ et un noyau K donnés et donner son expression mathématique. Utiliser la fonction et le package R de votre choix et décrire en détail les arguments utilisés en entrée de la fonction.

2.3. Représenter graphiquement $x \mapsto \widehat{g}_{n,h}(x)$ pour différentes valeurs de h. Considérer plusieurs cas de choix de h et décrire leur principe. Discuter l'importance du choix de h.

2 Estimation de la fonction de régression r

3.1. Décrire la distribution des Y_i. Décrire la relation entre les Y_i et X_i.

3.2. Est-il plausible de penser que la fonction r est linéaire ?

3.3. Construire un estimateur non paramétrique $\widehat{r}_{n,h}(x)$ de $r(x)$ pour une fenêtre de lissage $h > 0$ bien choisie. Présenter son expression mathématique et utiliser la fonction et le package R de votre choix. Décrire en détail les arguments utilisés en entrée de la fonction.

3.4. Représenter graphiquement $x \mapsto \widehat{r}_{n,h}(x)$.

3.5. Écrire une fonction R qui construit l'estimateur de Nadaraya-Watson. Comparer le résultat produit à celui obtenu via le package utilisé précédemment.

3.6. Comparer et discuter l'ajustement des estimations paramétriques (via un simple modèle linéaire) et non paramétrique de r. (Facultatif : Explorer des modèles paramétriques non linéaires.)

3 Validation croisée

On coupe l'échantillon en deux, selon $i \in \mathcal{J}_- = \{1, \ldots, 2500\}$ et $i \in \mathcal{J}_+ = \{2501, \ldots, 5000\}$. On note $\widehat{r}^{(-)}_{n,h}(x)$ l'estimateur construit à l'aide de $(X_i, Y_i)_{i \in \mathcal{J}_-}$ et on pose

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

4.1. Proposer un estimateur du risque quadratique de $\widehat{r}_{n,h}$ à partir des $\widetilde{\xi}_i$: présenter son expression et préciser le nom de la méthode qui se base dessus pour sélectionner h.

4.2. Quel est l'intérêt d'avoir découpé le jeu de données selon $\mathcal{J}_+$ et $\mathcal{J}_-$?

4.3. Implémenter la *leave-one-out* validation croisée.

Indications pour Question 3.5 et Question 4.3.

```
# ########## Nadaraya Watson ##########
# NW <- function(x, X, Y, h, K = dnorm) {
```

```
#   Kx <- ?
#   W <- ?
#   ?(W %*% Y)
# }
#
# ########## Validation croisée ##########
# cv <- function(X, Y, h, K = dnorm) {
#   ?
# }
#
# cv.grid <- function(X, Y,  h.grid = ?, K = dnorm, plot.cv = ?)
# {
#   obj <- sapply(h.grid, function(h) cv(X = X, Y = Y, h = h, K = K))
#   h <- h.grid[which.min(obj)]
#   if (plot.cv) { plot(h.grid, obj)}
#  ?
# }
```

4 Discussion

À vous de jouer.