INSA – Génie Mathématique Année 2018-2019

GM5 – Régression Non Linéaire avec Application sous R – Feuille TP 5

Estimation non-paramétrique de la fonction de régression

L'objectif de ce TP est double : étudier par simulations le comportement de l'estimateur de Nadaraya-Watson à l'aide de la fonction `ksmooth` du logiciel R, d'une part, et d'expérimenter cet outil sur un jeu de données réelles d'autre part.

Partie 1 Etude par simulations

1. Ecrire un module R qui permette de simuler 1000 réalisations $(x_i, y_i)_{1\leq i\leq 1000}$ du couple (X, Y) de variables aléatoires, sachant que $X \sim \mathcal{N}(0, 1)$ et

$$Y \;=\; f(X) \;+\; \varepsilon$$

où $f(x) = (x^2 - 1)$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2 = 0.25)$.

Dans les questions qui suivent, on appréciera la qualité de l'estimation de la fonction f à l'aide de graphiques et de l'erreur quadratique moyenne définie par

$$\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \widehat{f_n}(x_i)\Big)^2$$

et qui est à comparer à σ^2.

2. Pour $n = 1000$, et à l'aide de la fonction `ksmooth`, dont vous aurez préalablement consulté l'aide en ligne, étudier l'influence sur l'estimateur, d'une fenêtre de la forme $s_X n^{-\alpha}$ où $\alpha \in]0, 1[$ et s_X est l'estimation de l'écart-type de X calculée à partir des données (x_i). On essaiera différentes valeurs de α bien choisies, afin d'illustrer le compromis biais-variance.
3. Etudiez, en fonction de la taille n de l'échantillon, la qualité de l'estimation de la fonction f par l'estimateur de Nadaraya-Watson $\widehat{f_n}$ construit à partir des données $(x_i, y_i)_{1\leq i\leq n}$. On regardera par exemple, les valeurs $n = 100, 200, 500$ et 1000. On utilisera le noyau gaussien et la fenêtre optimale. Pour une taille d"échantillon donnée, on pourra représenter sur un même graphique la vraie fonction et son estimation, calculées sur une grille $t_1, t_2, \cdots, t_q$ de $[-3, 3]$. On pourra regrouper les graphiques par 4.
4. Pour $n = 1000$ et le bon réglage de fenêtre, construire à l'aide d'une méthode Monte-Carlo, une bande de confiance pour l'estimation de f. On utilisera la méthode des pourcentiles simples pour déterminer en toute valeur de t de votre intervalle d'étude, l'intervalle de confiance pour $f(t)$.

5. Commenter les différents résultats obtenus.
6. Reprendre les questions 1 à 5. avec le modèle

$$Y \ = \ f(X) \ + \ \varepsilon$$

où $X \sim \mathcal{U}_{[0,1]}$, $\varepsilon \sim \mathcal{N}(0,1)$ et

$$f(x) = 0.2 * x^{11} * (10 * (1-x))^6 \ + \ 10 * (10 * x)^3 * (1-x)^{10}$$

Sur cet exemple, on calculera l'estimateur en les points d'une grille $t_1, t_2, \cdots, t_q$ de l'intervalle $[0,1]$.

Deux utilisations possibles de `ksmooth`. Si `X` et `Y` contiennent respectivement les valeurs simulées $(x_i)_{1leqi\leq n}$ et $(y_i)_{1leqi\leq n}$, en exécutant

```
res <- ksmooth(X,Y)
```

alors `res$x` et `res$y` contiennent respectivement les valeurs (z_j) et $(\widehat{f}_n(z_j))$ où les (z_j) sont des points équidistants, ordonnés par ordre croissant, tels que $z_1 = \min(x_i)$ et $z_q = \max(x_i)$. En revanche, si on exécute les instructions

```
t <- seq(-3,3,0.05)
res <- ksmooth(X,Y,x.point = t)
```

alors `res$x` et `res$y` contiennent respectivement les valeurs `t` $= (t_j)_{1\leq j\leq q}$ et les valeurs $\left(\widehat{f}_n(t_i]\right)_{1\leq j\leq q}$ où q est la dimension du vecteur `t`, ie. q `<- length(t)`.

Ainsi, pour calculer les résidus $\widehat{\varepsilon}_i = y_i - \widehat{f}_n(x_i))$, il faut procéder astucieusement. Une manière de faire consiste à exécuter :

```
res <- ksmooth(X,Y,x.point = X)
```

Alors `res$x` et `res$y` contiennent respectivement les valeurs de X ordonnées par ordre croissant, que l'on notera $(x_{(i)})_{1\leq i\leq n}$ et les valeurs $\left(\widehat{f}_n(x_{(i)}]\right)_{1\leq j\leq n}$. On obtient alors les résidus en calculant :

```
ind = order(X)
eps_chap <- Y[ind] - res$y
```

Mais attention, ces résidus sont "classés" en fonctions des valeurs de `X`. Pour construire la suite $\left(\widehat{\varepsilon}_i = y_i - \widehat{f}_n(x_i))\right)_{1\leq i\leq n}$, il faut utiliser une boucle.

```
ind = order(X)
for (i in 1:n) {
ind_i = ind[i]
eps_chap[ind_i] <- Y[ind_i] - res$y[i]
}
```

Vous pouvez alors apprécier la qualité du modèle estimé en étudiant les résidus, vous pouvez vérifier la normalité des résidus estimés, comparer la moyenne quadratique des résidus à σ^2, etc ...

Partie 2 Expérimentation sur un jeu de données réelles

Reprendre le jeu de données "Barber.txt" du TP4, disponible à l'adresse
`http://lmi2.insa-rouen.fr/~bportier/Data/barber.txt`
et modéliser le taux de croissance (h^{-1}) de la bactérie *Escherichia coli* à différentes températures (°C) à l'aide d'un modèle de régression non linéaire en la température.

Estimer la fonction de régression à l'aide de l'estimateur de Nadaraya-Watson, en utilisant les données de léchantillon d'apprentissage. Il faudra régler "à la main" la valeur de la fenêtre. On pourra commencer par laisser la fonction `ksmooth` fixer la valeur de la fenêtre, puis essayer la fenêtre $h_n = sd(T)\, n^{-0.2}$ où $sd(T)$ désigne l'écart-type des températures observées. A vous, ensuite, de trouver la bonne fenêtre.

On comparera les performances de ce modèle avec celles obtenues avec le modèle non linéaire paramétrique du TP4, et on concluera sur l'intérêt ou non de cette approche.

A l'aide d'une méthode boodstrap basée sur les résidus, construire une bande de confiance à 95% pour la fonction estimée. On procèdera comme pour la bande de confiance du TP4. Comparer alors les bandes de confiance obtenues dans les deux TP.