Projet en statistique non-paramétrique

Mamaou Lamine DIAMBAN

12 Décembre 2019

On considère le modèle de régression,

$$Y_i = g(\frac{i}{n}) + \epsilon_i, \quad 1 \leq i \leq n$$

.

On suppose ici $\epsilon_1, ..., \epsilon_n$ des variables aléatoires centrées et de variance σ^2 et dépendantes.
Elles vérifient la relation,

$$\epsilon_n = \eta_n \sqrt{\sigma^2(1-\alpha) + \alpha\epsilon_{n-1}^2}, \quad 0 \leq i \leq 1$$

,

avec $(\eta_n)_{n\geq 1}$ est une suite iid centrée de loi normal $\mathcal{N}(0,1)$ et η_n est indépendante de $\epsilon_1, ..., \epsilon_{n-1}$.
On définit, $\widehat{g}$ l'estimateur de g, par :

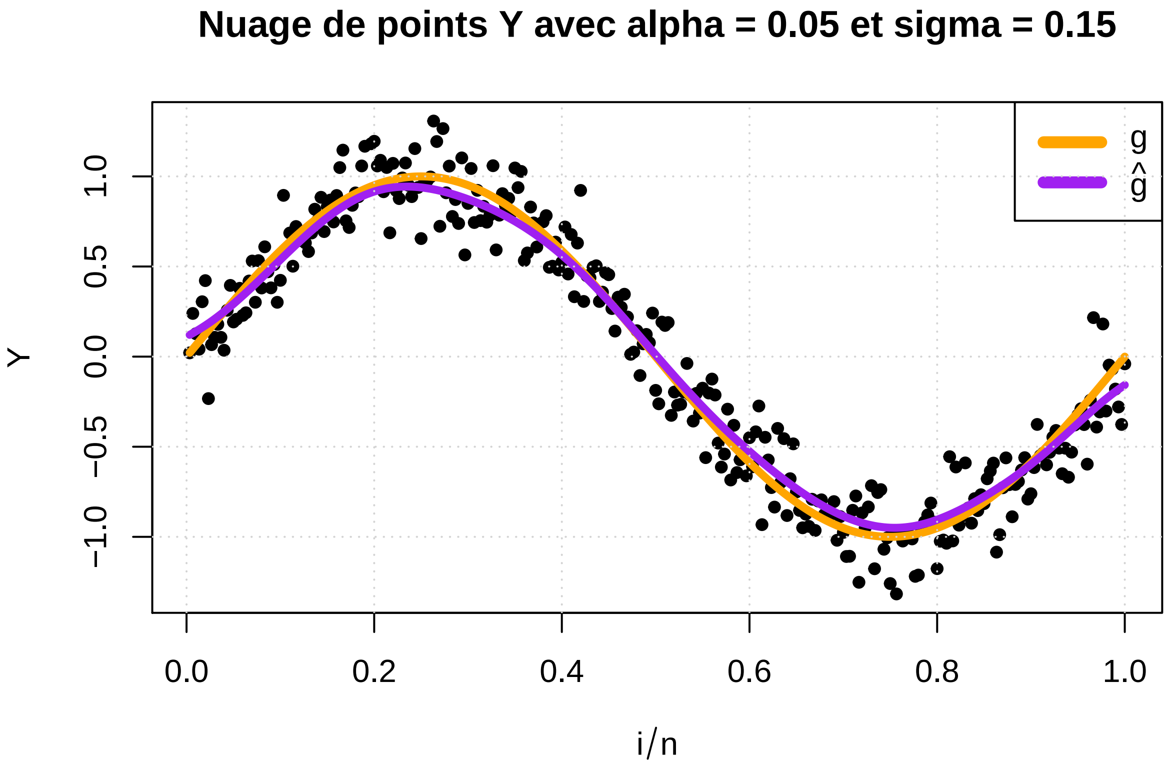
$$\widehat{g}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K(\frac{x - X_i}{h})$$

h est la fenêtre et K est un noyau pair et à support compact. L'objectif de ce projet est d'étudier empiriquement un bon choix de la fenêtre h. On prendra par la suite que :

$$g(x) = \sin(2\pi x)$$

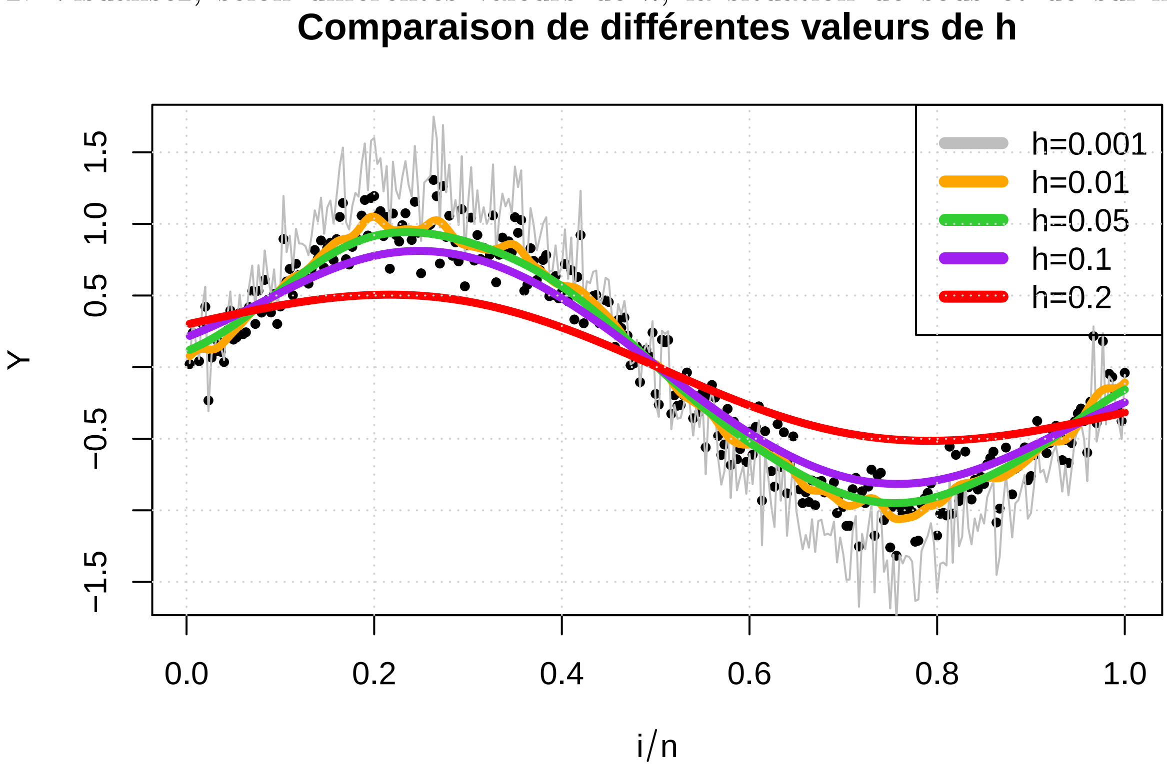
1. Représenter sur un même graphique le nuage des points $(\frac{i}{n}, Y_i)_{1\leq i\leq n}$, la fonction g et l'estimateur $\hat{g}$ pour un choix K, de α et de σ^2 que vous préciserz. Afin de trouver le modèle de regression Y, nous avons d'abord fixé les paramètres des erreurs(ϵ_i) :
$\sigma^2 = 0.0225$
$\alpha = 0.05$.

Puis pour calculer l'estimateur $\hat{g}$, on a pris un noyeau gaussien pour $K, K(u) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-u^2}{2})$ et une fenêtre $h = 0.03$.



Pour un choix des paramètres cités ci-dessous, on peut voir que Y a une forme sinisoïdale et la fonction g et son estimateur $\hat{g}$ sont très proches.
Et dans notre cas, lorsque Y augmente, g est sous-estimée par $\hat{g}$. Et inversement, lorsque Y diminue, la fonction g est sur-estimée par $\hat{g}$.

2. Visualisez, selon différentes valeurs de h, la situation de sous et de sur-lissage.



Nous avons fait varier la fenêtre h entre 10−3 et 0.2.
Et il en résulte que pour des valeurs de $h \in [0.001, ..., 0, 1]$, la fonction g est sur-lissée.
Tant disque pour des valeurs de $h \geq 0.1$, la fonction g est sous-lissée.
On peut donc conclure que la vraie valeur du paramètre de lissage h se situe dans la décimale 2.

3. Ecrire un programme qui calcule la valeur optimale du paramètre de lissage en fonction du ASE `ASE (Average square error)` est définit par,

$$ASE(h) = \frac{1}{h}\sum_{i=1}^{n}(\hat{r}(x_i) - r(x_i))^2$$

Soit $\hat{h}_0$ cette valeur optimale du ASE(h), c'est-à-dire,

$$\hat{h}_0 = argmin_{h>0} ASE(h)$$

h.optimale	0.026
ASE.optimale	0.0015

4. Même question, en remplacant ASE(h) pour le critère de validation croisé CV(h)

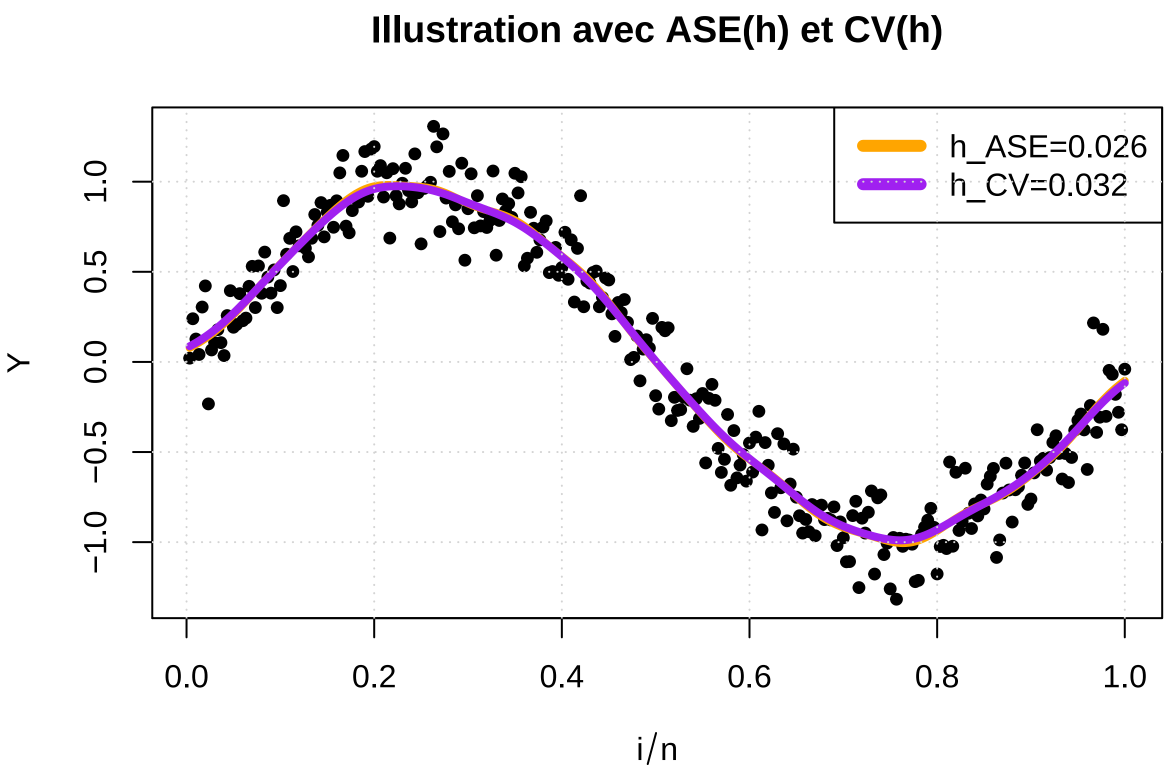
`CV(h)` est définit comme suit,

$$CV(h) = \frac{1}{h}\sum_{i=1}^{n}(\frac{\hat{r}(x_i) - Y_i}{1 - L_{i,i}})^2$$

avec $L_{i,i} = \frac{K(0)}{nh}$. On pose,

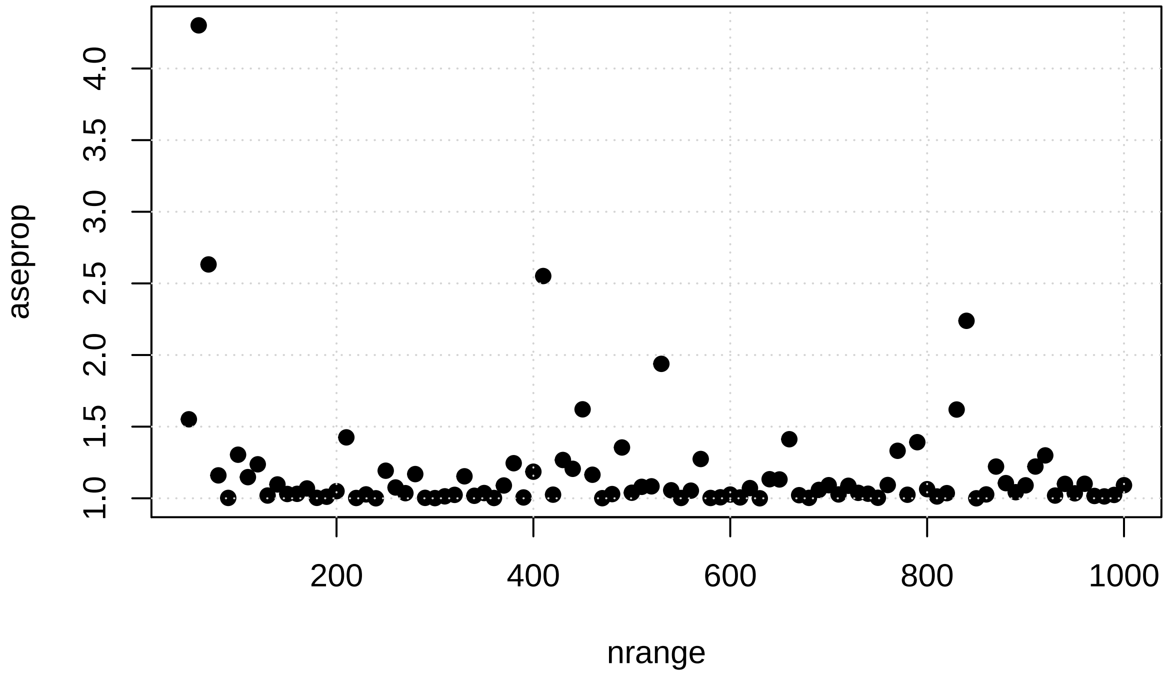
$$\hat{h} = argmin_{h>0}CV(h)$$

h.optimale	0.032
CV.optimale	0.0246



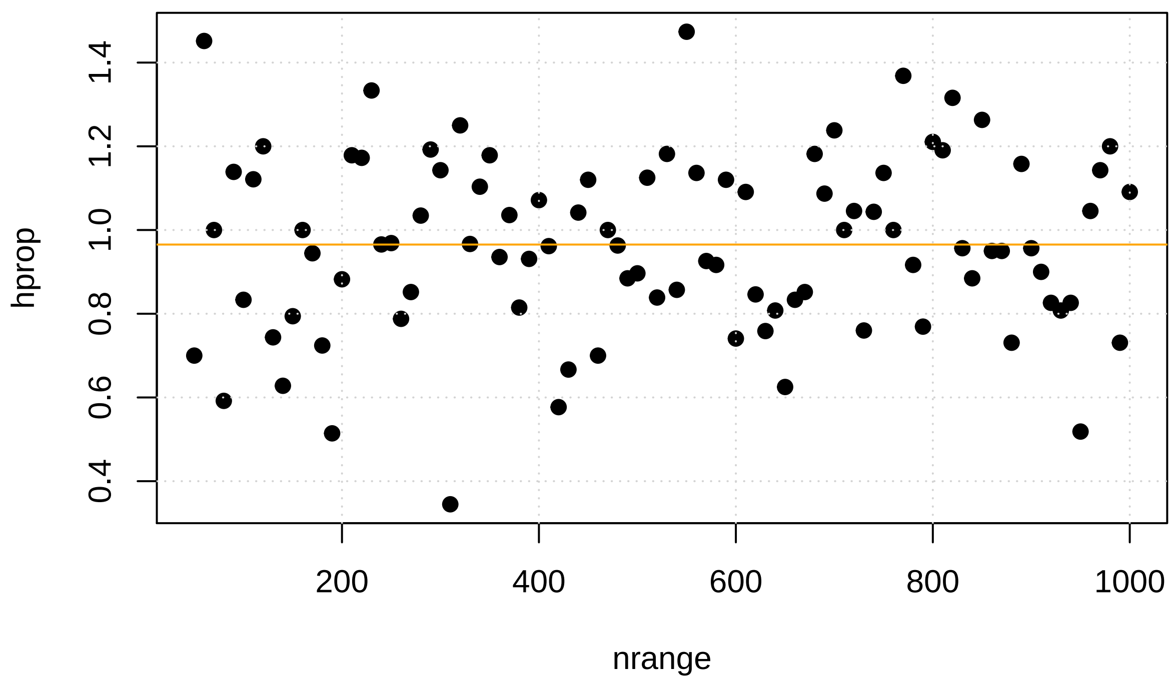
Bien que le paramètre de lissage soit supérieur avec la crosse validation($h = 0.032$), il apparaît cependant que toutes les deux fournissent un lissage très satisfaisant.

5. Illustrer le comportement asymptotique lorsque n tend vers l'infini de $\frac{ASE(\hat{h})}{ASE(\hat{h}_0)}$.



Lorsque $n \to \infty$, le rapport des erreurs du paramètre de lissage est presque constante et est proche de 1. Cela est d'autant plus marquant lorsque $n > 600$, toutes les valeurs sont comprises entre $1 et 1.5$. Tant disque lorsque $n < 600$, on peut voir qu'il existe des valeurs aberrantes pouvant aller jusqu'à ≈ 3.

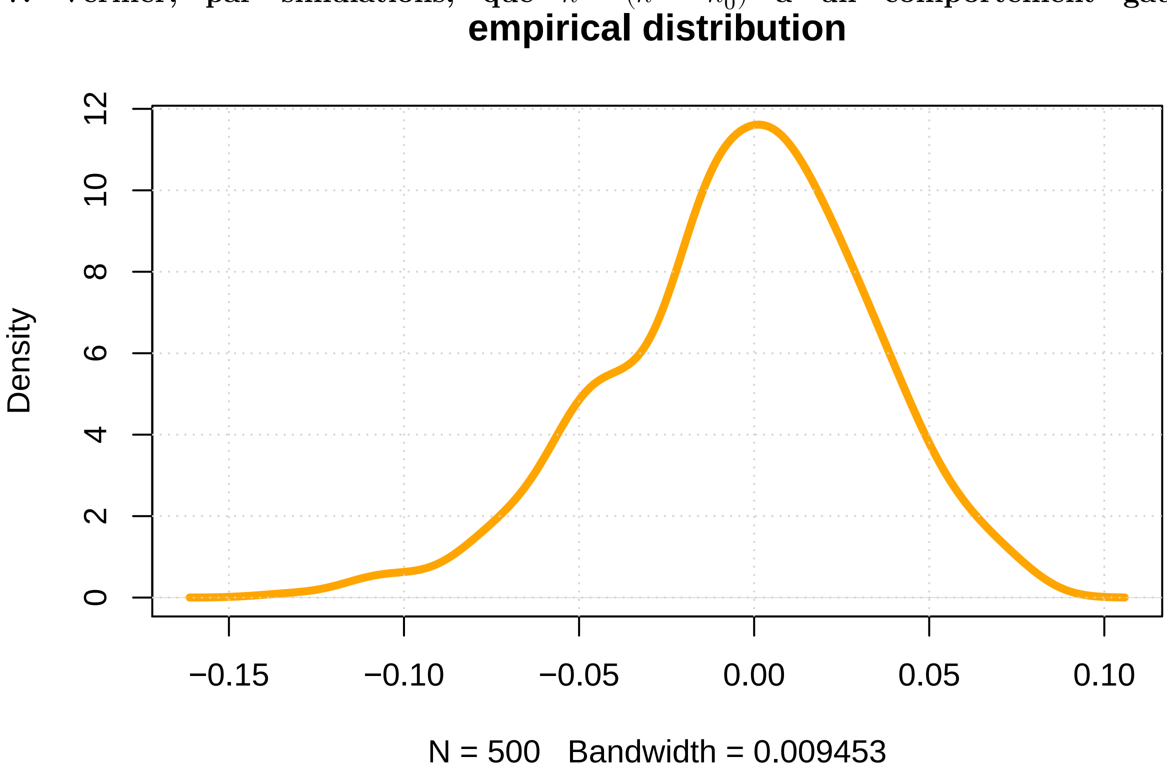
6. Illustrer le comportement asymptotique lorsque n tend vers l'infini de $\frac{\hat{h}}{\hat{h}_0}$.



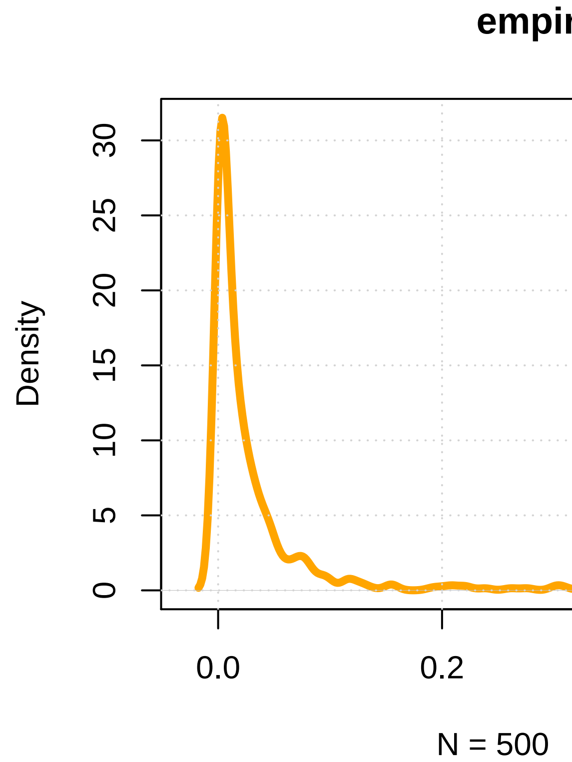
On a une asymétrie du rapport $\frac{\hat{h}}{\hat{h}_0}$ qui ne s'atténue pas, lorsque n tend vers l'infini avec une moyenne

$\approx 1.$

7. Vérifier, par simulations, que $n^{3/10}(\hat{h} - \hat{h}_0)$ a un comportement gaussien.



$n^{3/10}(\hat{h} - \hat{h}_0)$ suit bien une loi $\mathcal{N}(-0.0076, 0.037^2)$



8. Que peut être la loi asymptotique de $n(ASE(\hat{h}) - ASE(\hat{h}_0))$.

$n(ASE(\hat{h}) - ASE(\hat{h}_0))$ suit une loi de Poisson.

9. Conclure quant au critère $CV(h)$. D'une part, la vrai fonction $r(x)$ nous a été donnée de sorte qu'on puisse facilement calculer $ASE(h) = \frac{1}{h}\sum_{i=1}^{n}(\hat{r}(x_i) - r(x_i))^2$ et trouver la fenêtre optimale qui rend ASE minimum. Cependant, dans la pratique, il est impossible de connaître la vraie fonction qui produit les données $r(x)$.

Et d'autre part, même si $\hat{h}$ est sensiblement plus grande que $\hat{h}_0$, les résultats montrent que, $\frac{ASE(\hat{h})}{ASE(\hat{h}_0)}$ et $\frac{\hat{h}}{\hat{h}_0}$ sont respectivement proches de 1, ce qui nous permet de conclure que dans la pratique, nous pouvons utiliser la méthode $CV(h)$ à la place de ASE pour calculer la fenêtre optimale h.