Méthodes non-paramétriques

Méthodes à noyau

Wiam Chaoui Sophie Manuel Stéphane Sadio

11/03/2021

Contents

Méthodes non-paramétriques

Estimation de densité par les estimateurs à noyau

Notre but est d'estimer la densité f. Pour cela, on s'appuiera sur un échantillon iid $X = (X_1, ..., X_n)$ où chacune des variables X_i admet la densité f (par rapport à la mesure de Lebesgue).

Pour estimer une densité on peut utiliser une méthode à noyau. Les méthodes à noyau sont des méthodes non-paramétriques qui permettent de proposer une estimation de la densité plus lisse que celle obtenue par un histogramme.

Comment construit-on un estimateur à noyau ?

L'idée pour la construction de cet estimateur est d'utiliser l'approximation suivante , valable lorsque h est petit :

$$f(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

Pour estimer la densité f on peut passer par un estimateur $\hat{F}_n$ de la fonction de répartion F. $\hat{F}_n$ est la fonction de répartition empirique ($\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h}\mathbb{1}_{X_i \in]x-h,x+h[}$).

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h}\mathbb{1}_{X_i \in]x-h;x+h]}$$

Notons $\hat{f}(x)$ l'estimateur à noyau de la densité f, alors celui-ci s'écrit :

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

où h est la fenêtre (ou paramètre de lissage), n le nombre d'observations, et K le noyau. Cette formule n'est valable que si h est petit et positif.

ici $K(u) = \frac{1}{2}\mathbb{1}_{u \in]-1;1]}$, il s'agit du noyau de Rosenblatt, mais il existe d'autres noyaux.

Explication ce qu'est un noyau ?

Définition : Un noyau (*kernel* en anglais) est une application $K : \mathbb{R} \to \mathbb{R}$ intégrable et centrée telle que :

$$\int_{\mathbb{R}} K(u)du = 1 \qquad \text{et} \qquad \int_{\mathbb{R}} uK(u) = 0$$

si le noyau est en plus positif alors il correspond à une fonction de densité.

Exemples de noyau :

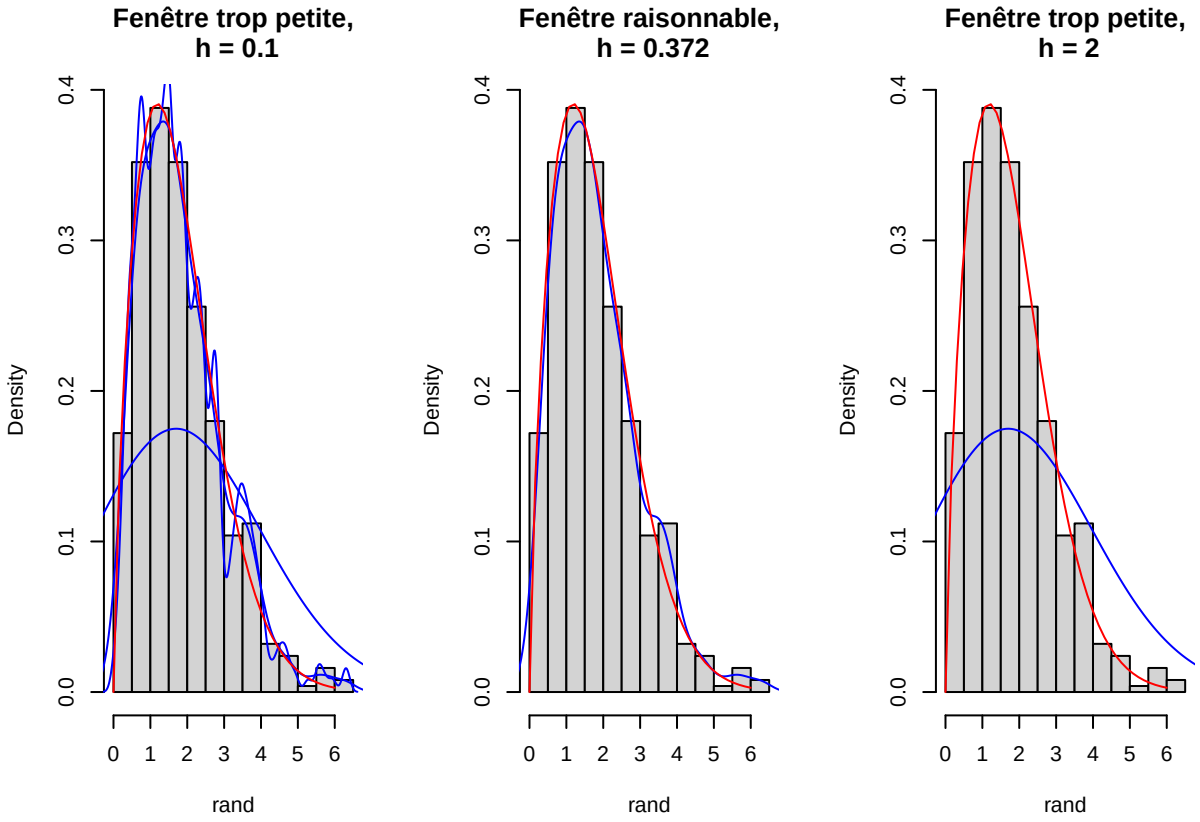
- Noyau de Rosenblatt, ou rectangulaire : $K(u) = \frac{1}{2}\mathbb{1}_{u \in]-1;1]}$
- Noyau Gaussien : $K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2})$
- Noyau d'Epanechnikov : $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{[-1,1]}(u)$
- Noyau triangulaire : $K(u) = (1- \mid u \mid)\mathbb{1}_{[-1,1]}(u)$
- Noyau Biweight : $K(u) = \frac{15}{16}(1-u^2)^2\mathbb{1}_{[-1,1]}(u)$

Les propriétés du noyau (continuité, différentiabilité...) se transmettent à l'estimateur $\hat{f}_n$.

Comment choisir les paramètres de la méthode ?

Dans la méthode d'estimation à noyau le choix du noyau n'est pas le plus important, le vrai enjeu de cette méthode est le choix de la fenêtre h (*bandwidth*). En effet, la fenêtre détermine l'influence des données dans l'estimation. Si h est petit, l'effet local est important donc on aura beaucoup de bruit. Si h est grand on aura une estimation plus douce, plus lisse.

Nous pouvons constater l'influence du paramètre h sur l'exemple suivant : Nous avons simulé 500 variables suivant une loi de Weibull de paramètres ($\alpha = 1.7$, $\lambda = 2$) représentées dans l'histogramme. La courbe en rouge est la vraie fonction de densité et la bleue est l'estimation avec la méthode des noyaux sur les variables simulées.



La fenêtre h du second graphique est calculé automatiquement par la fonction `density`de R.

Comment choisir la fenêtre optimale ?

Afin d'expliquer comment choisir une fenêtre, nous devons définir quelques objets. - Notons d la distance, nous préciserons laquelle si besoin. - Soit $\omega : \mathbb{R} \to \mathbb{R}^+$ telle que ω est convexe et $\omega(0) = 0$ alors ω est la fonction de perte. - On appelle risque de l'estimateur $\hat{f}_n$:

$$\hat{R}(\hat{f}_n, f) = \mathbb{E}[\omega(d(\hat{f}_n, f))]$$

Dans la suite on prendra le risque quadratique donc $\omega : x \to x^2$ et $d : (f, g) \to \sqrt{\int_{\mathbb{R}}(f(x) - g(x))^2 dx}$ la distance dans L^2.

On peut déterminer un h optimal par validation croisée, avec la fonction de risque. Puisque h dépend de la régularité de la fonction qui est inconnue, nous allons estimer cette fenêtre optimale par un estimateur $\hat{h}$. Pour cela, on cherche à minimiser en h le risque quadratique dans L^2.

$$R(\hat{f}_{n,h}, f) = \mathbb{E}[\| \hat{f}_{n,h} - f \|_2^2]$$

En pratique on ne peut pas calculer ce risque car il dépend de f qui est inconnu, donc on utilisera un estimateur de ce risque. On remarquera que minimiser $R(\hat{f}_{n,h}, f)$ en h équivaut à minimiser $R(\hat{f}_{n,h}, f) - \| f \|_2^2$ en h. On cherchera un estimateur (sans biais) de cette dernière. On suppose aussi que $R(\hat{f}_{n,h}, f) < \infty$ et que $f \in L^2$. Alors, d'après un théorème,

$$\hat{R}(h) = \| \hat{f}_{n,h} \|_2^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{1}{h} K\left(\frac{X_i - X_j}{h}\right)$$

est un estimateur sans biais de $R(\hat{f}_{n,h}, f) - \| f \|_2^2$.

On en déduit donc que :

$$\hat{h} = \arg\min_{h>0} \hat{R}(h)$$

Lorsque le minimum est atteint on obtient un $\hat{h}$ optimal.