STATISTIQUE AVANCÉE : MÉTHODES NON-PARAMÉTRIQUES

Ecole Centrale de Paris

Arnak S. DALALYAN

Table des matières

1 Introduction **5**

2 Modèle de densité **7**

2.1 Estimation par histogrammes 7

2.2 Estimateur à noyau 14

2.3 Exercices 21

3 Modèle de régression **23**

3.1 Définitions 23

3.2 Régressogrammes 23

3.3 Moyenne Locale 23

3.4 Estimateur à Noyau 23

3.5 Estimateur par Polynômes Locaux 23

3.6 Lissage Linéaire et Validation Croisée 23

3.7 Estimation de la Variance 23

3.8 Exemple 23

3.9 Exercices 23

1
Introduction

La particularité de la statistique non-paramétrique est que le paramètre inconnu qu'on cherche à detecter, à estimer ou à classifier n'est pas supposé d'appartenir à une famille indicée par un petit nombre de paramètres réels. En général, dans la théorie non-paramétrique on suppose que le nombre de paramètres qui décrivent la loi des observations est une fonction croissant du nombre d'observations, ou encore que le nombre de paramètres est infini.

Pour donner un exemple concret, considérons le modèle linéaire multiple. C'est un modèle très populaire dans le milieu des praticiens et profondément étudié par les théoriciens. Il s'agit d'une expérience qui résulte en l'observation des couples (X_i, Y_i), $i = 1, \dots, n$ où, en général, X_i est un vecteur p-dimensionnel et Y_i est une valeur réelle. On suppose que pour un vecteur $\beta \in \mathbb{R}^p$ et un réel α, la dépendance de Y_i en X_i est expliquée par la fonction affine

$$\alpha + \beta^T x$$

à une erreur aléatoire près, c'est-à-dire

$$Y_i = \alpha + \beta^T X_i + \xi_i, \qquad i = 1, \dots, n \tag{1.1}$$

où ξ_i est tel que $\mathbf{E}[\xi_i | X_i] = 0$. Si l'on suppose de plus que les erreurs ξ_i sont indépendantes les unes des autres et suivent la loi Gaussienne centrée de variance σ^2, alors la loi des observations (X_i, Y_i) est entièrement caractérisée par les paramètres α, β et σ. C'est pourquoi, dans la littérature statistique, ce modèle est souvent considéré comme l'exemple type d'un modèle paramétrique. Cependant, cette considération doit être nuancée.

En réalité, il est conseillé d'appliquer les méthodes classiques de statistique au modèle (1.1) seulement dans le cas où la dimension p de β est significativement plus petite que n, la taille de l'échantillon. Si n et p sont comparables, ou encore si $p > n$, les méthodes classiques deviennent inefficaces. Il faut alors chercher de nouvelles approches pour effectuer une inférence statistique. C'est l'objectif poursuivi par la statistique non-paramétrique.

Le but de ce cours est de présenter les principes les plus basiques de la statistique non-paramétrique en insistant sur leurs avantages et leurs limites. Pour éviter des dévelopements

très techniques, nous nous concentrons uniquement sur l'étude de deux modèles : l'estimation de densité et l'estimation de la fonction de régression. Par ailleurs, nous présenterons uniquement la facette de la statistique non-paramétrique concernant le lissage et ne parlerons pas du tout d'une autre facette, historiquement plus ancienne, qui est l'inférence basée sur les rangs.

Le modèle de densité est un modèle simple qui permet de tester les différentes innovations statistiques sans rentrer dans des calculs très fastidieux. Cependant, certaines méthodes – comme, par exemple, l'estimation par projection – sont plus faciles à présenter dans le modèle de régression. C'est la raison pour laquelle on se focalise sur ces deux modèles.

La démarche générale pour effectuer une inférence statistique dans des problèmes non-paramétriques peut être décomposée en trois étapes suivantes.

1. Trouver une famille $\{\bar{f}_h : h > 0\}$ de fonctions simples qui approchent bien la fonction inconnue f, c'est-à-dire $\mathrm{dist}(\bar{f}_h, f) \downarrow 0$ lorsque $h \downarrow 0$. On dit alors que $\mathrm{dist}(\bar{f}_h, f)$ est l'erreur d'approximation.
2. Au lieu d'effectuer une inférence statistique sur f, faire comme si le vrai paramètre était $\bar{f}_h$ et appliquer une méthode de statistique paramétrique classique :
   – méthode du maximum de vraisemblance, méthode des moments ou méthode de contraste minimale pour l'estimation,
   – test de Neyman-Pearson, test du rapport de vraisemblance ou test de Wald pour les tests d'hypothèses.

   On obtient ainsi une procédure statistique $\hat{d}_h$ (estimateur ou test). On appelle alors erreur statistique, note par $r(\hat{d}_h)$, le risque de la procédure $\hat{d}_h$ calculé en utilisant $\bar{f}_h$ comme vraie valeur du paramètre f.
3. Choisir le paramètre h de façon optimale. D'une part, dans la plupart des cas, l'erreur statistique $r(\hat{d}_h)$ est une fonction décroissante de h. D'autre part, le risque associé à la parocédure $\hat{d}_h$ dans le problème d'origine où f est le paramètre inconnu se calcule comme une fonction $F\big(\mathrm{dist}(f, f_h); r(\hat{d}_h)\big)$ qui est décroissante par rapport à chacun des deux arguments. Comme les fonctions $h \mapsto \mathrm{dist}(f, f_h)$ et $h \mapsto r(\hat{d}_h)$ ont des sens de variation opposés, la minimisation du risque total $F\big(\mathrm{dist}(f, f_h); r(\hat{d}_h)\big)$ en fonction de h se fait par un compromis entre l'erreur d'approximation $\mathrm{dist}(f, f_h)$ et l'erreur statistique $r(\hat{d}_h)$.

Pour terminer cette introduction, nous allons reformuler la définition de statistique non-paramétrique. La statistique non-paramétrique étudie des problèmes statistiques dans lesquels la paramétrisation n'est pas considérée comme figée, mais il y a une liberté de choix entre plusieurs paramétrisation et le but est de trouver celle qui conduit vers les procédures les plus performantes.

2
Modèle de densité

Tout au long de ce chapitre, on suppose que les observations $X_1, \ldots, X_n$ sont des variables indépendantes de même loi (iid) de densité f. Pour simplifier, on suppose que les X_i sont à valeurs réelles et que f est la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Par conséquent,

$$\text{Prob}(X_i \in [a,b]) = \int_a^b f(x)\,dx, \qquad \forall\, a,b \in \mathbb{R}.$$

De plus, on supposera que f est deux fois continûment différentiable.

2.1 Estimation par histogrammes

La façon la plus simple d'estimer la densité f à partir des données est l'estimation par histogramme. Afin d'éviter des complications d'ordre technique, nous supposerons dans ce paragraphe que f est à support compact. De plus, sans perte de généralité, nous pouvons supposer que le support de f est inclus dans l'intervalle $[0,1]$.

2.1.1 Définition et propriétés de base

Pour commencer, on choisit une partition uniforme $C_1, \ldots, C_m$ de l'intervalle $[0,1[$:

$$C_j = \Big[\frac{j-1}{m}, \frac{j}{m}\Big[, \quad j = 1, \ldots, m.$$

Comme f est supposée être continue, pour m suffisament grand, elle est bien approchée par des fonctions en escalier, constantes par morceaux sur les intervalles $\{C_j\}$. Pour que nos notations reste en accord avec l'approche générale décrite dans l'introduction, on pose $h = 1/m$ et on approche f par la fonction

$$\bar{f}_h(x) = \sum_{j=1}^{m} \frac{p_j}{h}\, \mathbb{1}_{C_j}(x),$$

où $p_j = \int_{C_j} f(x)\,dx$. On ramène ainsi le problème d'estimation de f au problème d'estimation d'un paramètre m-dimensionnel $\boldsymbol{p} = (p_1, \ldots, p_m)$. Ceci peut se faire en utilisant, par exemple la méthode généralisée des moments. En effet, il est évident que

$$p_j = \int_{C_j} f(x)\,dx = \mathbf{E}_f[\mathbb{1}_{C_j}(X_1)], \qquad \forall\, j = 1, \ldots, m.$$

Par conséquent, il est naturel d'estimer le vecteur $\boldsymbol{p}$ par

$$\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m), \qquad \hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i).$$

Observons au passage que chaque $\hat{p}_j$ représente la proportion des observations X_i se trouvant dans l'intervalle C_j. Si la taille de l'échantillon est grande, il est légitime de s'attendre à ce que cette proportion, dite empirique, converge vers la proportion théorique correspondant à la probabilité qu'une observation tirée au hasard selon la densité f appartient à l'intervalle C_j.

Par substitution, nous définissons l'estimateur de f par histogramme à m classes comme suit :

$$\hat{f}_h(x) = \frac{1}{h}\sum_{j=1}^{m} \hat{p}_j\, \mathbb{1}_{C_j}(x).$$

Dans la terminologie statistique, on dit que chaque C_j est une classe et la longueur des classes h est une fenêtre.

Exercice 2.1. *Vérifier que l'estimateur par histogramme $\hat{f}_h$ est une densité de probabilité.*

Remarque 2.1. *Dans les applications, très souvent on utilise le terme histogramme pour la fonction $h\hat{f}_h(x)$, ce qui correspond à la proportion d'observations par intervalle C_j.*

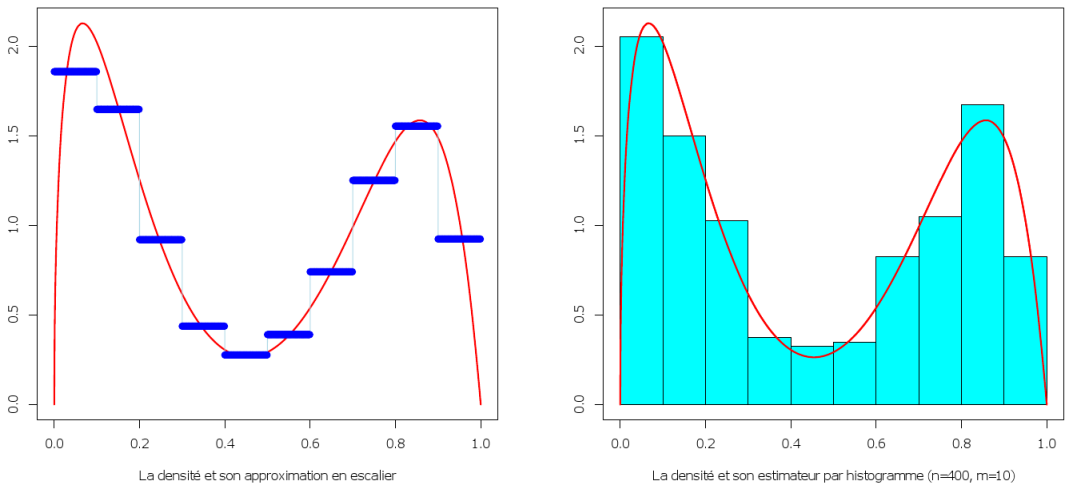


FIGURE 2.1 – A gauche : une densité de probabilité et son approximation constante par morceaux sur les intervalle $[j/10, (j+1)/10[$ pour $j = 1, \ldots, 10$. A droite : La même densité que celle du graphe de gauche et une estimation par histogramme basée sur un échantillon de taille 400.

2.1.2 Exemple : répartition des galaxies

A titre d'exemple, considérons un jeu de données astronomiques étudiées dans le livre de Wasserman et disponibles sur sa page WEB :

`http://www.stat.cmu.edu/~larry/all-of-nonpar/data.html`.

Ce qu'on veut montrer sur cet exemple, avant toute autre chose, est que le choix de la fenêtre h a un impact très important sur la qualité d'estimation de la densité f par l'histogramme $\hat{f}_h$.

Le jeu de données astronomiques précité contient 1253 valeurs numériques ; chaque valeur correspond au décalage vers le rouge (Redshift) d'un objet astronomique (galaxie, quasar, ...). Cette valeur peut être interprétée comme la distance entre la terre et l'objet en ques-

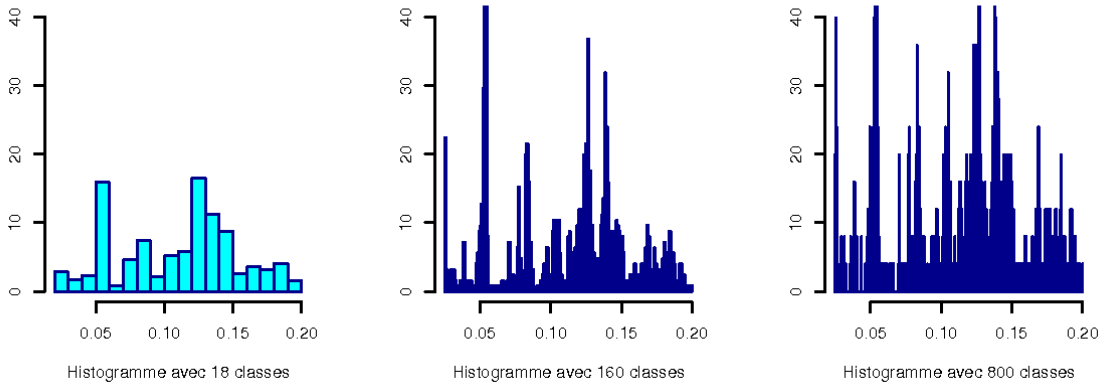


FIGURE 2.2 – Les histogrammes des données d'astronomie correspondant aux nombres de classes $m = 18$, $m = 160$ et $m = 850$.

Nous avons tracé dans la figure ci-dessus les histogrammes des données astronomiques basées sur 18, 160 et 850 classes. On constate que ces trois graphiques ont présentent des différences très importantes. Plus précisément, le graphe qui correspond à $m = 18$ est bien plus régulier que les deux autres. Dans la terminologie statistique, on dit que l'histogramme de gauche est trop lissé (en anglais *oversmoothing*) alors que l'histogramme de droite n'est pas lissé sufisamment (*undersmoothing*). Un problème crucial du point de vu des applications est donc de trouver la fenêtre h qui correspond au lissage optimal.

L'une des méthodes les plus utilisées fournissant une fenêtre proche de l'optimale est la méthode de validation croisée. La définition précise de cette méthode sera donnée plus tard dans ce chapitre. Notons simplement qu'elle consiste à définir une fonction $\hat{J}$ de h (ou, de façon équivalente, de m) qui est une estimation du risque de l'estimateur $\hat{f}_h$. Naturellement, la valeur de h est choisie en minimisant ce risque estimé. Lorsqu'on effectue une validation croisée sur les données astronomiques, on obtient la courbe ci-dessous pour la fonction $m \mapsto \hat{J}(m)$ et le minimum de ce

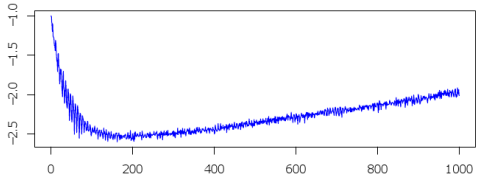


FIGURE 2.3 – La courbe de la fonction $\hat{J}(m)$. Le minimum est atteint au point $m = 163$.

2.1.3 Risque de l'estimateur par histogramme

Comme on a constaté sur l'exemple des données astronomique, la qualité de l'estimateur par histogramme dépend fortement de la fenêtre h. Afin de quantifier cette dépendance, nous introduisons le risque quadratique de $\hat{f}_h$ au point $x \in [0,1]$ comme étant la moyenne de l'erreur quadratique :

$$\mathsf{MSE}_f(x,h) = \mathbf{E}_f[\big(\hat{f}_h(x) - f(x)\big)^2],$$

où l'abbréviation MSE correspond à *Mean Squared Errror*.

Afin d'évaluer le MSE, on utilise d'abord la décomposition biais-variance :

$$\mathsf{MSE}_f(x,h) = \underbrace{\big(\mathbf{E}_f[\hat{f}_h(x)] - f(x)\big)^2}_{\text{carré du biais}} + \underbrace{\mathrm{Var}_f[\hat{f}_h(x)]}_{\text{variance}}.$$

Soit j l'indice de la classe contenant x ; $x \in C_j$. Remarquons que

$$\hat{f}_h(x) = \frac{\hat{p}_j}{h} = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i) = \frac{Z_j}{nh}, \qquad Z_j \sim \mathcal{B}(n, p_j), \tag{2.1}$$

car Z_j est la somme de n variables indépendantes de loi de Bernoulli de paramètre

$$\mathbf{P}_f(\mathbb{1}_{C_j}(X_1) = 1) = \mathbf{P}_f(X_1 \in C_j) = \int_{C_j} f(x)\,dx = p_j.$$

Il en découle immédiatement que pour tout $x \in C_j$,

$$\mathbf{E}_f[\hat{f}_h(x)] = \frac{p_j}{h}, \qquad \mathrm{Var}_f[\hat{f}_h(x)] = \frac{np_j(1-p_j)}{n^2h^2} = \frac{p_j(1-p_j)}{nh^2}\,. \tag{2.2}$$

Une première conséquence de ces formules est que le risque MSE est supérieur au carré du biais $(h^{-1}p_j - f(x))^2$. Par conséquent, si la fenêtre h est choisie indépendamment de la taille de l'échantillon n, l'estimateur par histogramme ne convergera pas vers la vraie densité lorsque $n \to \infty$ excepté la situation peu fréquente où f est constante sur l'intervalle C_j. Afin d'élargir la classe des densités pour lesquelles $\hat{f}_h$ est convergent, nous devons choisir h comme une fonction de n ; $h = h_n$ doit tendre vers 0 lorsque n tend vers $+\infty$. A partir de maintenant, on suppose que cette condition est satisfaite.

Rappelons que le but de ce paragraphe est d'évaluer le risque de l'estimateur $\hat{f}_h$. Afin d'avoir une évaluation globale valable pour tout point $x \in [0,1]$, on considère le risque quadratique intégré :

$$\mathsf{MISE}_f(h) = \int_{[0,1]} \mathsf{MSE}_f(x,h)\,dx = \mathbf{E}_f\Big[\int_0^1 \big(\hat{f}_h(x) - f(x)\big)^2 dx\Big]$$

(pour obtenir la dernière égalité nous avons utilisé le théorème de Foubini).

D'une part, en vertu de la propriété $\sum_j p_j = \int_0^1 f(x)\,dx = 1$, on a

$$\int_0^1 \mathrm{Var}_f[\hat{f}_h(x)]\,dx = \sum_{j=1}^{m}\int_{C_j} \mathrm{Var}_f[\hat{f}_h(x)]\,dx = \sum_{j=1}^{m}\frac{p_j(1-p_j)}{nh} = \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^{m} p_j^2.$$

D'autre part,

$$\begin{aligned}\int_0^1 \{\mathbf{E}_f[\hat{f}_h(x)] - f(x)\}^2\,dx &= \sum_{j=1}^m \int_{C_j} \Big(\frac{p_j}{h} - f(x)\Big)^2\,dx \\ &= \sum_{j=1}^m \frac{p_j^2}{h} - 2\frac{p_j}{h}\int_{C_j} f(x)\,dx + \int_0^1 f^2(x)\,dx \\ &= \int_0^1 f^2(x)\,dx - \frac{1}{h}\sum_{j=1}^m p_j^2.\end{aligned}$$

Nous avons donc démontré le résultat suivant :

Lemme 2.1. *Si $X_1, \ldots, X_n$ sont indépendantes de même loi de densité f supportée par $[0,1]$ et $\hat{f}_h$ est l'estimateur par histogramme avec $m = 1/h$ classes, alors*

$$\mathrm{MISE}_f(h) = \mathbf{E}_f\Big[\int_0^1 \big(\hat{f}_h(x) - f(x)\big)^2 dx\Big] = \int_0^1 f^2(x)\,dx + \frac{1}{nh} - \frac{n+1}{nh}\sum_{j=1}^m p_j^2.$$

Le résultat du Lemme 2.1 est non-asymptotique : il est valable pour tout h et pour tout n. Ce qui nous intéresse maintenant c'est le comportement du risque MISE lorsque $h = h_n$ décroit vers zéro quand n tend vers $+\infty$. On vérifie aisément que

$$\begin{aligned}\int_{C_j} f(x)^2\,dx - h^{-1}p_j^2 &= \int_{C_j}\Big(f(x) - \frac{1}{h}\int_{C_j} f(u)\,du\Big)^2 dx \\ &= \frac{1}{h^2}\int_{C_j}\Big(\int_{C_j}\{f(x) - f(u)\}\,du\Big)^2 dx.\end{aligned}$$

Comme f est supposée deux fois continûment différentiable, on a $f(u) - f(x) = (u - x)f'(a_j) + O(h^2)$ pour tout $u, x \in C_j$, où a_j désigne l'extrémité gauche de l'intervalle C_j. Par conséquent,

$$\int_{C_j} f(x)^2\,dx - h^{-1}p_j^2 = \frac{f'(a_j)^2}{h^2}\int_{C_j}\Big(\int_{C_j}(x-u)\,du\Big)^2 dx + O(h^4).$$

En utilisant le changement de variable $(x,u) = (a_j + yh, a_j + zh)$, on obtient

$$\int_{C_j}\Big(\int_{C_j}(x-u)\,du\Big)^2 dx = h^5\int_0^1\Big(\int_0^1 (y-z)\,dz\Big)^2 dy = \frac{h^5}{12}.$$

Nous avons donc démontré que

$$\int_{C_j} f(x)^2\,dx - h^{-1}p_j^2 = \frac{h^3}{12}f'(a_j)^2 + O(h^4) = \frac{h^2}{12}\int_{C_j} f'(x)^2\,dx + O(h^4).$$

En conséquence,

$$\begin{aligned}\mathrm{MISE}_f(h) &= \sum_{j=1}^m\Big(\int_{C_j} f(x)^2\,dx - h^{-1}p_j^2\Big) + \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^m p_j^2 \\ &= \frac{h^2}{12}\int_0^1 f'(x)^2\,dx + O(h^3) + \frac{1}{nh} + O(1/n),\end{aligned}$$

où nous avons utilisé la relation $mO(h^4) = O(h^3)$. Ces calculs implique donc le résultat suivant :

Théorème 2.1. *Supposons que la densité de l'échantillon $X_1, \dots, X_n$ est deux fois continûment différentiable et s'annule en dehors de l'intervalle $[0,1]$. Si la fenêtre h de l'estimateur par histogramme $\hat{f}_h$ est telle que $h_n \to 0$ lorsque $n \to \infty$, alors*

$$\mathsf{MISE}_f(h) = \underbrace{\frac{h^2}{12}\int_0^1 f'(x)^2\,dx + \frac{1}{nh}}_{\textit{terme principal du risque}} + \underbrace{O(h^3) + O\Big(\frac{1}{n}\Big)}_{\textit{terme résiduel}}$$

lorsque $n \to \infty$.

Supposons un instant qu'on cannaît la quantité $\int_0^1 f'(x)^2\,dx$. Dans ce cas, on peut calculer le terme principal du risque $\mathsf{MISE}_f(h)$. Cela nous permet de trouver la valeur idéale de la fenêtre qui minimise le terme principal du risque. En effet, on voit aisément que le minimum de la fonction

$$h \mapsto \frac{h^2}{12}\int_0^1 f'(x)^2\,dx + \frac{1}{nh}$$

est atteint au point

$$h_{opt} = \Big(\frac{n}{6}\int_0^1 f'(x)^2\,dx\Big)^{-1/3}.$$

Cette fenêtre optimale est en général inaccessible au statisticien, car la densité f (ainsi que sa dérivée) est inconnue. Cependant, elle a le mérite de nous indiquer que la fenêtre optimale doit être de l'ordre de $n^{-1/3}$ lorsque n est grand. De plus, en injectant cette valeur de h dans l'expression de MISE obtenue dans le théorème précédent, on obtient

$$\mathsf{MISE}_f(h_{opt}) = (3/4)^{2/3}\Big(\int_0^1 f'(x)^2\,dx\Big)^{1/3} n^{-2/3} + O(n^{-1}).$$

Ce résultat nous indique les limites de l'estimateur par histogramme : pour les densités deux fois différentiables, la meilleure vitesse de convergence qu'on puisse espérer atteindre avec un estimateur par histogramme est de $n^{-2/3}$. Cette une vitesse honorable, mais elle est nettement moins bonne que la vitesse de convergence $1/n$ qui apparaît typiquement dans des problèmes paramétriques. Ceci n'est pas très surprenant, car l'estimation de densité est un problème non-paramétrique et, par conséquent, est plus difficile à résoudre qu'un problème paramétrique.

En revanche, on verra par la suite que, sous les mêmes hypothèses que celles considérées dans ce paragraphe, on peut construire un autre estimateur de la densité f qui converge à une meilleure vitesse $n^{-4/5}$. L'estimateur qui atteint cette vitesse s'appelle estimateur à noyau et on peut démontrer que cette vitesse ne peut pas être améliorée sans imposer de nouvelles condition sur f.

2.1.4 Choix de la fenêtre par validation croisée

Comme on l'a déjà fait remarquer, la fenêtre idéale h_{opt} définie dans le paragraphe précédent est inutilisable en pratique car elle fait intervenir la densité inconnue f à travers

l'intégrale du carré de sa dérivée. Afin de désigner une méthode de choix de h indépendant de f, nous commençons par estimer le risque [1] de l'estimateur $\widehat{f}_h$ en utilisant uniquement les observations $X_1, \ldots, X_n$. Soit $\widehat{J}(h, X_1, \ldots, X_n)$ un estimateur de $\mathsf{MISE}_f(h) - \|f\|_2^2$. Pour que la méthode de sélection de h conduise vers des résultats raisonnable, on demande de l'estimateur $\widehat{J}(h, X_1, \ldots, X_n)$ être sans biais [2], c'est-à-dire

$$\mathbf{E}_f[\widehat{J}(h, X_1, \ldots, X_n)] = \mathsf{MISE}_f(h) - \|f\|_2^2.$$

Une fois que nous avons à notre disposition cet estimateur $\widehat{J}$, on détermine la valeur de h en minimisant $\widehat{J}(h, X_1, \ldots, X_n)$ par rapport à $h \in]0, \infty[$. La valeur de h où ce minimum est atteint est sélectionnée comme fenêtre pour l'estimateur par histogramme. Voyons maintenant comment cette méthode peut être effectivement mise en oeuvre.

Pour toute densité f et pour tout histogramme $\hat{f}_h$, soit

$$J_f(h) = \mathsf{MISE}_f(h) - \|f\|_2^2 = \frac{1}{nh} - \frac{n+1}{nh} \sum_{j=1}^{m} p_j^2, \tag{2.3}$$

en vertu du Lemme 2.1. Rappelons que p_j représente la proportion théorique des observations qui se situent dans la classe C_j, pour tout $j = 1, \ldots, m$. Il découle de cette rélation que pour désigner un estimateur sans biais de $J_f(h)$, il suffit de désigner un estimateur sans biais de p_j^2, pour tout $j = 1, \ldots, m$. Une approche naïve consisterait à estimer p_j^2 par $\hat{p}_j^2$, où

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i)$$

est la proportion empirique des observations se situant dans C_j. Comme $n\hat{p}_j$ suit la loi binômiale de paramètres (n, p_j), on a $\mathrm{Var}_f[\hat{p}_j] = p_j(1 - p_j)/n$ et, par conséquent,

$$\mathbf{E}_f[\hat{p}_j^2] = \mathrm{Var}_f[\hat{p}_j] + (\mathbf{E}_f[\hat{p}_j])^2 = p_j^2\Big(1 - \frac{1}{n}\Big) + \frac{p_j}{n}. \tag{2.4}$$

Cette égalité nous montre d'une part que l'idée naïve d'estimer p_j^2 par $\hat{p}_j^2$ ne conduit pas vers un estimateur sans biais. Mais, d'autre part, ce petit calcul que nous venons d'effectuer prépare le terrain pour déterminer l'estimateur utilisé par la méthode de validation croisée. En effet, comme $\hat{p}_j$ est un estimateur sans biais de p_j, il résulte de (2.4) que $\hat{p}_j^2 - \hat{p}_j/n$ est un estimateur sans biais de $p_j^2(1 - 1/n)$. Par conséquent, pour tout $j = 1, \ldots, m$,

$$\tilde{p}_j^2 = \frac{\hat{p}_j^2 - \hat{p}_j/n}{1 - 1/n} = \frac{n}{n-1}\, \hat{p}_j^2 - \frac{1}{n-1}\, \hat{p}_j$$

est un estimateur sans biais de p_j^2. En injectant cet estimateur dans le membre droit de l'égalité (2.3) et en utilisant le fait que $\sum_{j=1}^{m} \hat{p}_j = 1$, nous obtenons le résultat suivant.

1. En pratique, il est préférable d'estimer non pas le risque $\mathsf{MISE}_f(h)$ de $\widehat{f}_h$, mais la différence entre le risque de $\widehat{f}_h$ et celui de l'estimateur trivial $\widehat{f}_{\mathrm{triv}} \equiv 0$.

2. Idéalement, il faudrait également pouvoir contrôler la variance de $\hat{J}(h, X_1, \ldots, X_n)$, mais cet aspect ne sera pas évoqué dans le cadre de ce cours.

Proposition 2.1. *Si f est une densité de carré intégrable et si $\hat{f}_h$ est l'histogramme à $m = 1/h$ classes basé sur l'échantillon $X_1, \ldots, X_n$ ayant f pour densité de probabilité, alors*

$$\widehat{J}(h, X_1, \ldots, X_n) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h} \sum_{j=1}^{m} \hat{p}_j^2$$

est un estimateur sans biais de $\mathsf{MISE}_f(h) - \|f\|_2^2$.

Nous pouvons à présent énoncer la méthode de validation croisée. Nous allons le faire dans le cadre général, sans supposer que les observations sont incluses dans $[0,1]$. Dans ce cas, on peut poser $a = \min_i X_i$ et $b = \max_i X_i$ et pour tout $m \in \mathbb{N}$ choisir la fenêtre $h = (b-a)/m$. On définit alors les classes $C_j = [a + (j-1)h; a + jh[$ pour $j = 1, \ldots, m-1$ et $C_m = [b-h; b]$.

Algorithm de validation croisée pour choisir la fenêtre d'un histogramme.

Entrée : $X_1, \ldots, X_n$;
Sortie : $\widehat{h}_{CV}$;
Définir $a \leftarrow \min_i X_i$;
$b \leftarrow \max_i X_i$;
Initialiser
$m \leftarrow 1$;
$m_{CV} \leftarrow 1$;
$J_{CV} \leftarrow -1$;
Tant que $(m < n)$ **effectuer** :
Poser $J \leftarrow \dfrac{2m}{n-1} - \dfrac{(n+1)m}{n-1} \displaystyle\sum_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i) \right)^2$;
Si $(J < J_{CV})$ **alors**
$m_{CV} \leftarrow m$;
$J_{CV} \leftarrow J$;
FinSi
$m \leftarrow m + 1$;
Fin Tant que
$\hat{h}_{CV} \leftarrow (b-a)/m_{CV}$;

Une fois la fenêtre $\hat{h}_{CV}$ déterminée, nous pouvons calculer et tracer la courbe de l'histogramme ayant comme fenêtre $\hat{h}_{CV}$.

2.2 Estimateur à noyau

L'estimation de la densité par histogrammes est une méthode naturelle très répendue car elle est facilement implémentable. Cependant, l'estimateur de densité fournit par un histogramme ne peut pas être adapté à la situation assez courant où nous disposons d'une

information à priori sur la régularité de la densité à estimer. Plus précisément, si l'on sait par avance que la densité de l'échantillon observé est, par exemple, deux fois continûment différentiable, on aurait naturellement envie d'estimer cette densité par une fonction qui, elle aussi, est deux fois continûment différentiable. Or, les histogrammes sont des fonctions qui ne sont même pas continues. Il est naturel alors de vouloir "lisser" les histogrammes. On s'attend alors à ce que le résultat du lissage améliore non seulement l'aspect visuel de l'estimateur, mais produise de plus un estimateur plus proche de la vraie densité que l'estimateur par histogramme.

2.2.1 Définition et propriétés élémentaires

Soit $x \in \mathbb{R}$ et $h > 0$. Si l'on suppose que x est le centre d'une classe de l'histogramme et que h est la longueur des classes, l'estimateur de $f(x)$ par histogramme peut s'écrire comme

$$\widehat{f}_h^H(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}(|X_i - x| \leq h/2) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}\Big(\frac{|X_i - x|}{h} \leq \frac{1}{2}\Big).$$

Une façon de généraliser les histogramme consiste à utiliser la formule ci-dessus pour tout $x \in \mathbb{R}$ et pas seulement pour les centres des classes. Cette généralisation est certe utile, car elle conduit vers un estimateur qui est constant par morceaux comme les histogrammes, mais a l'avantage d'avoir des plateaux de longueurs variables. Cependant, cela ne nous conduit pas vers un estimateur continu. On remarque aisément que la discontinuité de l'estimateur défini ci-dessus est une conséquence de la discontinuité de la fonction indicatrice. Par conséquent, en remplaçant $\mathbb{1}(|z| \leq \frac{1}{2})$ par une fonction K quelconque, on obtient l'estimateur

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

qui est continu et même ℓ-fois continûment différentiable du moment où la fonction K l'est. On arrive ainsi à la définition suivante.

Définition 2.1. *Soit $K : \mathbb{R} \to \mathbb{R}$ une fonction quelconque et soit h un réel positif. On appelle estimateur à noyau la fonction*

$$\widehat{f}_h^K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big). \tag{2.5}$$

On dit alors que K est le noyau de cet estimateur et h est la fenêtre.

Selon cette définition, toute fonction K peut servir comme noyau d'estimation d'une densité f. Les noyaux les plus couramment utilisés en pratique sont

– le noyau rectangulaire :

$$K(u) = \frac{1}{2}\mathbb{1}_{[-1,1]}(u),$$

– le noyau triangulaire :

$$K(u) = (1 - |u|)\mathbb{1}_{[-1,1]}(u),$$

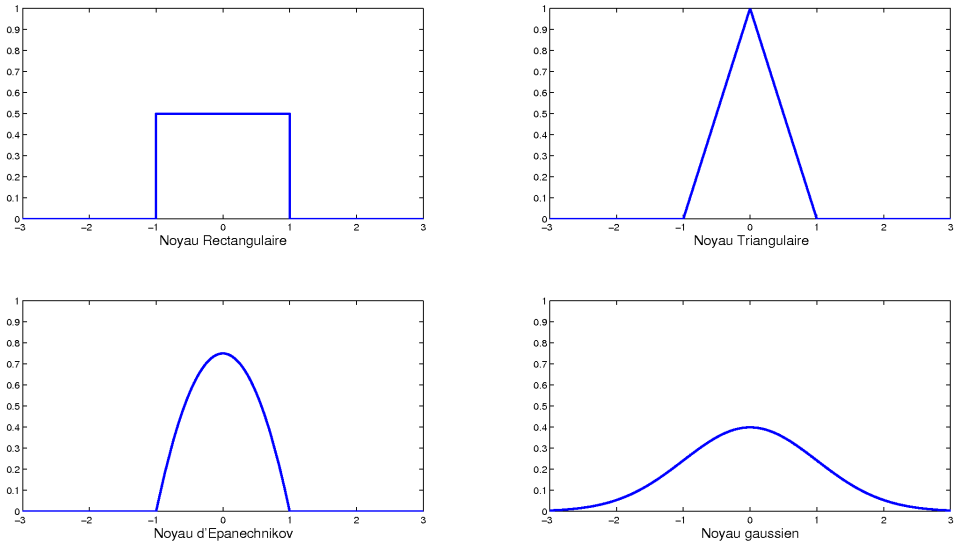
– le noyau d'Epanechnikov :

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{[-1,1]}(u),$$

– le noyau gaussien :

$$K(u) = \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}.$$

Les courbes de ces noyaux sont présentées ci-dessous :



Lemme 2.2. *Si K est positive et $\int_{\mathbb{R}} K(u)\,du = 1$, alors $\widehat{f}_h^K(\cdot)$ est une densité de probabilité. De plus, $\widehat{f}_h^K$ est continue si K est continue.*

Démonstration. L'estimateur à noyau est positive et continue car la somme des fonctions positives et continues est elle-même une fonction positive et continue. Il faut donc vérifier que l'intégrale de $\widehat{f}_h^K(\cdot)$ vaut un. En effet,

$$\begin{aligned}
\int_{\mathbb{R}} \widehat{f}_h^K(x)\,dx &= \int_{\mathbb{R}} \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)\,dx \\
&= \frac{1}{nh}\sum_{i=1}^n \int_{\mathbb{R}} K\Big(\frac{X_i - x}{h}\Big)\,dx \quad (u = (X_i - x)/h) \\
&= \frac{1}{nh}\sum_{i=1}^n \int_{\mathbb{R}} K(u)\,h du = 1.
\end{aligned}$$

□

On voit donc que, tout comme l'estimateur par histogramme, l'estimateur à noyau est une densité de probabilité. Il a de plus l'avantage d'être continu à condition que K le soit, ce

qui n'était pas le cas pour les histogrammes. Par conséquent, lorsqu'on estime une densité continue, il est naturel de s'attendre que l'estimateur à noyau soit meilleur que l'estimateur par histogramme. Le but de la suite de ce chapitre est de donner des résultats quantitatives caractérisant le gain obtenu par l'utilisation de $\widehat{f}_h^K$ par rapport à $\widehat{f}_h^H$.

2.2.2 Etude du biais et de la variance

Lorsqu'on définit un estimateur à noyau, on a non-seulement le choix de la fenêtre $h > 0$ mais aussi celui du noyau K. Il y a un certain nombre de conditions qui sont considérées comme usuelles pour les noyaux et qui permettent d'analyser le risque de l'estimateur à noyau qui en résulte.

HYPOTHÈSE K : On suppose que K vérifie les 4 conditions suivantes :

1. $\int_{\mathbb{R}} K(u)\,du = 1$,
2. K est une fonction paire ou, plus généralement, $\int_{\mathbb{R}} u\,K(u)\,du = 0$,
3. $\int_{\mathbb{R}} u^2|K(u)|\,du < \infty$,
4. $\int_{\mathbb{R}} K(u)^2\,du < \infty$.

Proposition 2.2. *Si les trois premières conditions de l'hypothèse K sont remplies et f est une densité bornée dont la dérivée seconde est bornée, alors*

$$\left|Biais\big(\widehat{f}_h^K(x)\big)\right| \leq C_1 h^2,$$

où $C_1 = \frac{1}{2}\sup_{z\in\mathbb{R}}|f''(z)|\int_{\mathbb{R}} u^2|K(u)|\,du$.
Si, de plus, la condition 4 de l'hypothèse K est satisfaite, alors

$$Var\big(\widehat{f}_h^K(x)\big) \leq \frac{C_2}{nh}$$

avec $C_2 = \sup_{z\in\mathbb{R}} f(z)\int_{\mathbb{R}} K(u)^2\,du$.

Démonstration. Commençons par calculer le biais :

$$\begin{aligned}
\mathbf{E}_f[\widehat{f}_h^K(x)] &= \frac{1}{nh}\sum_{i=1}^n \mathbf{E}_f\Big[K\Big(\frac{X_i - x}{h}\Big)\Big] \\
&= \frac{1}{nh}\sum_{i=1}^n \int_{\mathbb{R}} K\Big(\frac{y-x}{h}\Big) f(y)\,dy \\
&= \frac{1}{h}\int_{\mathbb{R}} K\Big(\frac{y-x}{h}\Big) f(y)\,dy \qquad (y = x + uh, dy = hdu) \\
&= \int_{\mathbb{R}} K(u)\,f(x+uh)\,du.
\end{aligned}$$

En effectuant un développement limité à l'ordre 2, il vient

$$\begin{aligned}
\mathbf{E}_f[\widehat{f}_h^K(x)] &= \int_{\mathbb{R}} K(u)\,f(x+uh)\,du \\
&= \int_{\mathbb{R}} K(u)\,\big[f(x) + (uh)f'(x) + \frac{(uh)^2}{2}f''(\xi_u)\big]\,du \quad (\xi_u \in [x, x+uh]) \\
&= f(x)\underbrace{\int_{\mathbb{R}} K(u)\,du}_{=1} + hf'(x)\underbrace{\int_{\mathbb{R}} uK(u)\,du}_{=0} + \frac{h^2}{2}\int_{\mathbb{R}} u^2K(u)f''(\xi_u)\,du.
\end{aligned}$$

Il en résulte que

$$\begin{aligned}\big|\mathrm{Biais}\big(\widehat{f}_h^K(x)\big)\big| &= \big|\mathbf{E}_f[\widehat{f}_h^K(x)] - f(x)\big| \\ &\le \frac{h^2}{2}\Big|\int_{\mathbb{R}} u^2 K(u)\, f''(\xi_u)\, du\Big| \\ &\le \frac{h^2}{2}\int_{\mathbb{R}} u^2 |K(u)|\, |f''(\xi_u)|\, du \\ &\le h^2 \underbrace{\frac{\max_x |f''(x)|}{2}\int_{\mathbb{R}} u^2 |K(u)|\, du}\, C_1\end{aligned}$$

d'où la première assertion de la proposition.

Pour prouver la seconde assertion, on utilise le faite que les variables aléatoires $Y_i = K\big((X_i - x)/h\big)$, $i = 1, \ldots, n$ sont i.i.d. et que la variance de la somme de variables indépendantes coïncide avec la somme des variances :

$$\begin{aligned}\mathrm{Var}_f[\widehat{f}_h^K(x)] &= \frac{1}{(nh)^2}\,\mathrm{Var}_f\Big[\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)\Big] \\ &= \frac{1}{(nh)^2}\sum_{i=1}^n \mathrm{Var}_f\Big[K\Big(\frac{X_i - x}{h}\Big)\Big] \\ &= \frac{1}{(nh)^2} \times n \times \mathrm{Var}_f\Big[K\Big(\frac{X_1 - x}{h}\Big)\Big] \\ &\le \frac{1}{nh^2}\,\mathbf{E}_f\Big[K\Big(\frac{X_1 - x}{h}\Big)^2\Big] \\ &= \frac{1}{nh^2}\int_{\mathbb{R}} K\Big(\frac{y - x}{h}\Big)^2 f(y)\, dy \qquad (y = x + uh, dy = h du) \\ &= \frac{1}{nh}\int_{\mathbb{R}} K(u)^2 f(x + uh)\, du \\ &\le \frac{1}{nh}\underbrace{\sum_z f(z)\int_{\mathbb{R}} K(u)^2\, du}_{C_2}.\end{aligned}$$

C'est exactement ce qu'il fallait démontrer. □

2.2.3 Quelques remarques

Les évaluations du biais et de la variance que l'on vient de démontrer ont un certain nombre de conséquences résumées ci-dessous.

Vitesse de convergence : On déduit de la Proposition 2.2 que le risque MSE de $\hat{f}_h^K(x)$ admet la majoration suivante :

$$\mathsf{MSE}\big(\hat{f}_h^K(x)\big) \le C_1^2 h^4 + \frac{C_2}{nh}.$$

On vérifie aisément que la valeur de la fenêtre h qui minimise le majorant du MSE est $h_{opt} = (C_2/4C_1^2)^{1/5} n^{-1/5}$. En injectant cette valeur dans l'expression du MSE on obtient :

$$\mathsf{MSE}\big(\hat{f}_{h_{opt}}^K(x)\big) \le \mathrm{Const} \cdot n^{-4/5}.$$

Cela montre que la vitesse de convergence de l'estimateur à noyau est de $n^{-4/5}$. Elle est donc meilleure que la vitesse $n^{-2/3}$ obtenue pour les histogrammes. Par conséquent, les estimateurs à noyau sont préférables aux histogrammes lorsqu'il s'agit d'estimer une densité deux fois continûment différentiable.

Optimalité de la vitesse : On peut démontrer qu'il est impossible d'estimer f à une vitesse meilleure que $n^{-4/5}$ sans imposer des hypothèses supplémentaires (de régularité ou de structure) sur la densité inconnue f.

Sur-lissage et sous-lissage : Lorsque la fenêtre h est très petit, le biais de l'estimateur à noyau est très petit face à sa variance et c'est cette dernière qui détermine la vitesse de convergence du risque quadratique. Dans ce type de situation, l'estimateur est très volatile et on parle de sous-lissage (under-smoothing, en anglais). En revanche, lorsque h grandit, la variance devient petite et c'est le biais qui devient dominant. L'estimateur est alors très peu variable et est de moins à moins influencé par les données. On parle alors d'un effet de sur-lissage (over-smoothing en anglais). En pratique, il est primordial de trouver la bonne dose de lissage qui permet d'éviter le sous-lissage et le sur-lissage.

Décriptage de la vitesse de convergence : On peut se demander d'où viennent le 4 et lle 5 dans la vitesse de convergence $n^{-4/5}$ présentée ci-dessus. En fait, si l'on estime une densité non pas univariée, mais d-variée [3], et l'on suppose que f est k fois continûment différentiable, alors la vitesse de convergence optimale est de $n^{-2k}/(2k+d)$. Dans le cas où $d = 1$ et $k = 2$, on retrouve la vitesse $n^{-4/5}$.

Comparaison avec le cadre paramétrique : Dans la théorie statistique paramétrique classique, la vitesse de convergence usuelle pour le risque quadratique est de n^{-1}, où n est le nombre d'observations. On voit que la vitesse $n^{-4/5}$ obtenue pour l'estimateur à noyau est meilleure que $n^{-2/3}$ obtenu pour l'estimateur par histogramme mais reste quand-même inférieure à la vitesse paramétrique. Ceci est tout à fait naturelle et traduit la complexité de l'estimation non-paramétrique comparée à l'estimation paramétrique. On peut remarquer également que lorsque la régularité de la densité tend vers l'infinie ($\beta \to \infty$), la vitesse de convergence se rapproche de plus en plus de la "vitesse paramétrique".

Exercice 2.2. *Soit $\beta > 0$, $L > 0$ et soit [4] $k = \lfloor \beta \rfloor$. On suppose que la densité f appartient à la classe de Hölder $\mathscr{H}(\beta, L)$ définie par :*

$$f \in \mathscr{H}(\beta, L) \quad \Longleftrightarrow \quad f \in C^k \quad et \quad |f^{(k)}(y) - f^{(k)}(x)| \leq L|x-y|^{\beta-k},\ \forall x, y.$$

1. *Montrer que si le noyau K vérifie les conditions K et $\int_{\mathbb{R}} u^j K(u) = 0$, $\forall j = 1, \ldots, k$, et $\int |u|^\beta |K(u)| du < \infty$ alors il existe des constantes C_1 et C_2 telles que*

$$MSE_f[\hat{f}_h^K(x)] \leq C_1 h^{2\beta} + \frac{C_2}{nh}.$$

2. *En déduire la valeur h_{opt} de la fenêtre h qui minimise le majorant du risque. Quelle est la vitesse de convergence du risque associé à cette fenêtre optimale ?*
3. *Montrer que si le noyau K vérifie les conditions ci-dessus et si $\beta > 2$, alors l'estimateur $\hat{f}_h^K$ n'est pas une densité de probabilité.*

3. c'est-à-dire $f : \mathbb{R}^d \to \mathbb{R}$
4. $\lfloor \beta \rfloor$ désigne le plancher de β, c'est-à-dire le plus grand nombre entier strictement plus petit que β

2.2.4 Validation croisée

Pour désigner une méthode automatique pour le choix de la fenêtre h, on utilise souvent la méthode de la validation croisée. Il s'agit de proposer dans un premier temps (pour un h fixé) un estimateur $\hat{J}(h)$ sans biais de la quantité $J(h) = \mathsf{MISE}(\hat{f}_h^K) - \|f\|_2^2$ et, dans un deuxième temps, de minimiser cet estimateur $\hat{J}(h)$ sur un ensemble fini de candidats pour h.

Proposition 2.3. *La statistique*

$$\hat{J}(h) = \|\hat{f}_h^K\|_2^2 - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K\Big(\frac{X_i - X_j}{h}\Big)$$

est un estimateur sans biais de $J(h)$.

Démonstration. D'une part, comme la densité jointe du couple (X_i, X_j) est $f(x)f(y)$, on a

$$\begin{aligned}
\mathbf{E}_f[\hat{J}(h)] &= \mathbf{E}_f[\|\hat{f}_h^K\|_2^2] - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathbf{E}_f\left[K\Big(\frac{X_i - X_j}{h}\Big)\right] \\
&= \mathbf{E}_f[\|\hat{f}_h^K\|_2^2] - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \int_{\mathbb{R}^2} K\Big(\frac{x-y}{h}\Big) f(x)f(y)\,dxdy \\
&= \mathbf{E}_f[\|\hat{f}_h^K\|_2^2] - \frac{2}{n(n-1)h} \times n(n-1) \times \int_{\mathbb{R}^2} K\Big(\frac{x-y}{h}\Big) f(x)f(y)\,dxdy \\
&= \mathbf{E}_f[\|\hat{f}_h^K\|_2^2] - \frac{2}{h} \int_{\mathbb{R}^2} K\Big(\frac{x-y}{h}\Big) f(x)f(y)\,dxdy.
\end{aligned}$$

D'autre part,

$$\begin{aligned}
J(h) &= \mathsf{MISE}(\hat{f}_h^K) - \|f\|_2^2 = \mathbf{E}_f[\|\hat{f}_h^K - f\|_2^2] - \|f\|_2^2 \\
&= \mathbf{E}_f\Big[\|\hat{f}_h^K\|_2^2 - 2\langle \hat{f}_h^K, f\rangle + \|f\|_2^2\Big] - \|f\|_2^2 \\
&= \mathbf{E}_f\big[\|\hat{f}_h^K\|_2^2\big] - 2\mathbf{E}_f\Big[\int_{\mathbb{R}} \hat{f}_h^K(x) f(x)\,dx\Big] \\
&= \mathbf{E}_f\big[\|\hat{f}_h^K\|_2^2\big] - 2\int_{\mathbb{R}} \mathbf{E}_f\big[\hat{f}_h^K(x)\big]\, f(x)\,dx.
\end{aligned}$$

Or, on a vu déjà (voir la démonstration de la Prop. 2.2) que $\mathbf{E}_f\big[\hat{f}_h^K(x)\big] = \frac{1}{h}\int_{\mathbb{R}} K\big(\frac{y-x}{h}\big)\, f(y)\,dy$. Par conséquent,

$$\begin{aligned}
J(h) &= \mathbf{E}_f\big[\|\hat{f}_h^K\|_2^2\big] - 2\int_{\mathbb{R}} \mathbf{E}_f\big[\hat{f}_h^K(x)\big]\, f(x)\,dx \\
&= \mathbf{E}_f\big[\|\hat{f}_h^K\|_2^2\big] - 2\int_{\mathbb{R}} \frac{1}{h}\int_{\mathbb{R}} K\Big(\frac{y-x}{h}\Big)\, f(y)\,dy\, f(x)\,dx \\
&= \mathbf{E}_f\big[\|\hat{f}_h^K\|_2^2\big] - \frac{2}{h}\int_{\mathbb{R}}\int_{\mathbb{R}} K\Big(\frac{y-x}{h}\Big)\, f(y)f(x)\,dy\,dx \\
&= \mathbf{E}_f[\widehat{J}(h)],
\end{aligned}$$

ce qui équivaut à dire que $\widehat{J}(h)$ est un estimateur sans biais de $J(h)$. □

En utilisant cet estimateur $\widehat{J}(h)$, on définit l'algorithm de validation croisée (cross validation, en anglais) de la même manière que pour les estimateurs par histogramme.

2.3 Exercices

3

Modèle de régression

3.1 Définitions

3.2 Régressogrammes

3.3 Moyenne Locale

3.4 Estimateur à Noyau

3.5 Estimateur par Polynômes Locaux

3.5.1 Définition et Propriétés de bases

3.5.2 Etude du Biais et de la Variance

3.5.3 Vitesse de convergence

3.6 Lissage Linéaire et Validation Croisée

3.7 Estimation de la Variance

3.8 Exemple

3.9 Exercices