Apprentissage non paramétrique en régression

Résumé

*Différentes méthodes d'estimation non paramétriques en régression sont présentées. Tout d'abord les plus classiques : estimation par des polynômes, estimation sur des bases de splines, estimateurs à noyau et par projection sur des bases orthonormées (Fourier, ondelettes). Des contrôles du risque quadratique et des calculs de vitesses de convergences sont effectués. Nous présentons également les modèles additifs généralisés ainsi que les arbres de régression CART et la méthode KRLS (*kernel regression least square*).*

Retour au plan du cours

1 Introduction

On se place dans le cadre d'un modèle de régression :

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \dots n.$$

Nous supposerons que les variables $\boldsymbol{X}_i$ appartiennent à $\mathbb{R}^d$, les Y_i sont réelles.

- Soit les $\boldsymbol{X}_i$ sont déterministes, et nous supposerons les variables ε_i sont i.i.d., centrées, de variance σ^2.
- Soit les $\boldsymbol{X}_i$ sont aléatoires et nous supposerons les variables ε_i indépendantes des $\boldsymbol{X}_i$, i.i.d., centrées, de variance σ^2.

En l'absence de toute hypothèse sur la fonction de régression f, nous sommes dans un cadre non paramétrique. Nous allons proposer plusieurs types de méthodes d'apprentissage pour la fonction f : l' estimation par des splines, les estimateurs à noyaux et les estimateurs par projection sur des bases orthonormées, notamment des bases d'ondelettes. Nous verrons également une méthode qui permet de contourner le fléau de la dimension dans le cas des modèles additifs, enfin nous introduirons les arbres CART.

2 Estimation par des polynômes par morceaux.

Dans ce chapitre, on suppose que les X_i appartiennent à un compact de $\mathbb{R}$, que l'on peut supposer égal à $[0, 1]$.

2.1 Estimation par des constantes par morceaux

On peut estimer la fonction f par une fonction constante par morceaux sur une partition de $[0, 1]$. (Ces estimateurs sont les analogues en régression des estimateurs par histogramme en densité, on les appelle régressogrammes).
On découpe $[0, 1]$ en D intervalles de même taille :

$$I_{k,D} = \mathbf{1}_{]k/D,(k+1)/D]}, \ k = 0, \dots, D-1.$$

Il est naturel d'estimer la fonction f sur l'intervalle $I_{k,D}$ par la moyenne des valeurs de Y_i qui sont telles que $X_i \in I_{k,D}$, soit pour tout $x \in I_{k,D}$, on pose

$$\hat{f}_D(x) = \frac{\sum_{i, X_i \in I_{k,D}} Y_i}{\sharp\{i, X_i \in I_{k,D}\}}$$

si $\sharp\{i, X_i \in I_{k,D}\} \neq 0$ et

$$\hat{f}_D(x) = 0 \text{ si } \sharp\{i, X_i \in I_{k,D}\} = 0.$$

On peut aussi écrire $\hat{f}_D(x)$ sous la forme

$$\hat{f}_D(x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{X_i \in I_{k,D}}}{\sum_{i=1}^n \mathbf{1}_{X_i \in I_{k,D}}}.$$

On suppose dans la suite que $D < n$, si pour tout i, $X_i = i/n$, ceci entraîne que pour tout k, $\sharp\{i, X_i \in I_{k,D}\} \neq 0$.

Cet estimateur correspond à l'estimateur des moindres carrés de f sur le modèle paramétrique des fonctions constantes par morceaux sur les intervalles $I_{k,D}$:

$$\mathcal{S}_D = \{f(x) = \sum_{k=1}^{D} a_k \mathbf{1}_{x \in I_{k,D}}\}.$$

En effet, si on cherche à minimiser

$$h(a_1,\dots,a_D)=\sum_{i=1}^{n}\left(Y_i-\sum_{k=1}^{D}a_k\mathbf{1}_{X_i\in I_{k,D}}\right)^2=\sum_{k=1}^{D}\sum_{i,X_i\in I_{k,D}}(Y_i-a_k)^2, \tag{1}$$

la minimisation est obtenue pour

$$\hat{a}_l=\frac{\sum_{i,X_i\in I_{l,D}}Y_i}{\sharp\{i,X_i\in I_{l,D}\}},\quad \forall l.$$

2.2 Polynômes par morceaux

L'estimation par des polynômes par morceaux de degré m sur la partition définie par les intervalles $I_{k,D}, 1\leq k\leq D$ correspond à la minimisation du critère :

$$\sum_{i=1}^{n}\left(Y_i-\sum_{k=1}^{D}(a_{k,0}+a_{k,1}X_i+\ldots+a_{k,m}X_i^m)\mathbf{1}_{X_i\in I_{k,D}}\right)^2$$
$$=\sum_{k=1}^{D}\sum_{i,X_i\in I_{k,D}}(Y_i-a_{k,0}-a_{k,1}X_i-\ldots-a_{k,m}X_i^m)^2.$$

Sur tout intervalle $I_{k,D}$, on ajuste un polynôme de degré m, par la méthode des moindres carrés en minimisant le critère :

$$\sum_{i,X_i\in I_{k,D}}(Y_i-a_{k,0}-a_{k,1}X_i-\ldots-a_{k,m}X_i^m)^2.$$

Il s'agit simplement d'un modèle linéaire en les paramètres $(a_{k,0},\dots,a_{k,m})$, il y a donc une solution explicite. Le problème du choix des paramètres D et m se pose.

2.3 Ajustement des paramètres

Revenons au cas de l'estimation par des constantes par morceaux, et considérons le problème du choix du paramètre D. On peut alors distinguer deux cas extrêmes :

- Si D est de l'ordre de n, on a un seul point X_i par intervalle $I_{k,D}$ et on estime f par Y_i sur chaque intervalle $I_{k,D}$. On a une fonction très irrégulière, qui reproduit simplement les observations. On fait alors du sur-ajustement.

- Si $D=1$, on estime f sur $[0,1]$ par la moyenne de toutes les observations Y_i. Si f est très loin d'être une fonction constante, l'estimateur sera mal ajusté.

Il faut donc trouver un bon compromis entre ces deux situations extrêmes pour le choix de D.

2.4 Performances de l'estimateur.

Nous allons majorer le risque quadratique de l'estimateur, pour un choix convenable de D, dans le cas où la fonction de régression f est Lipschitzienne : on suppose que f est dans la classe de fonctions

$$\mathcal{S}_{1,R}=\{f\in\mathbb{L}^2([0,1]),\forall x,y\in[0,1],|f(x)-f(y)|\leq R|x-y|\}.$$

THÉORÈME 1. — *Dans le modèle*

$$Y_i=f(\frac{i}{n})+\varepsilon_i,\ \ i=1,\dots,n,$$

l'estimateur

$$\hat{f}_D(x)=\frac{\sum_{i=1}^{n}Y_i\mathbf{1}_{X_i\in I_{k,D}}}{\sum_{i=1}^{n}\mathbf{1}_{X_i\in I_{k,D}}},$$

avec

$$D=D(n)=[(nR^2)^{1/3}]$$

vérifie

$$\sup_{f\in\mathcal{S}_{1,R}}\mathbb{E}_f[\|\hat{f}_D-f\|_2^2]\leq C(\sigma)R^{\frac{2}{3}}n^{-\frac{2}{3}}.$$

Bien entendu, ce résultat est purement théorique, car en pratique, on ne sait pas si la fonction f appartient à la classe $\mathcal{S}_{1,R}$. Nous verrons à la Section 10 des méthodes pratiques de choix de D par validation croisée.

Démonstration. —

- **Calcul de l'espérance**
  Pour tout $x \in I_{k,D}$,

$$\mathbb{E}_f\left(\hat{f}_D(x)\right) = \frac{\sum_{i,X_i \in I_{k,D}} f(X_i)}{\sharp\{i, X_i \in I_{k,D}\}}.$$

$$\mathbb{E}_f\left(\hat{f}_D(x)\right) - f(x) = \frac{\sum_{i,X_i \in I_{k,D}} (f(X_i) - f(x))}{\sharp\{i, X_i \in I_{k,D}\}}.$$

  Si on fait l'hypothèse que $f \in \mathcal{S}_{1,R}$ alors pour x et X_i dans le même intervalle $I_{k,D}$, $|x - X_i| \leq 1/D$, ce qui implique $|f(x) - f(X_i)| \leq RD^{-1}$. Ainsi

$$|\text{Biais}(\hat{f}_D(x))| = |\mathbb{E}_f\left(\hat{f}_D(x)\right) - f(x)| \leq \frac{R}{D}.$$

- **Calcul de la variance**

$$\begin{aligned} \text{Var}(\hat{f}_D(x)) &= \mathbb{E}_f[(\hat{f}_D(x) - \mathbb{E}_f(\hat{f}_D(x)))^2] \\ &= \frac{\sigma^2}{\sharp\{i, X_i \in I_{k,D}\}}. \end{aligned}$$

On utilise comme critère pour mesurer les performances de notre estimateur le risque $\mathbb{L}^2([0,1], dx)$ c'est-à-dire

$$L(\hat{f}_D, f) = \mathbb{E}_f[\int_0^1 (\hat{f}_D(x) - f(x))^2 dx].$$

On a aussi

$$L(\hat{f}_D, f) = \int_0^1 \mathbb{E}_f[(\hat{f}_D(x) - f(x))^2] dx.$$

Or,

$$\begin{aligned} \mathbb{E}_f[(\hat{f}_D(x) - f(x))^2] &= \mathbb{E}_f[\left(\hat{f}_D(x) - \mathbb{E}_f(\hat{f}_D(x)) + \mathbb{E}_f(\hat{f}_D(x)) - f(x)\right)^2] \\ &= \mathbb{E}_f[(\hat{f}_D(x) - \mathbb{E}_f(\hat{f}_D(x)))^2] + [\mathbb{E}_f(\hat{f}_D(x)) - f(x)]^2 \\ &= \text{Var}(\hat{f}_D(x)) + \text{Biais}^2(\hat{f}_D(x)) \\ &\leq \frac{\sigma^2}{\sharp\{i, X_i \in I_{k,D}\}} + R^2 D^{-2}. \end{aligned}$$

Puisque $X_i = i/n$, on remarque aisément que $\sharp\{i, X_i \in I_{k,D}\} \geq [n/D] \geq n/(2D)$ si on suppose $D \leq n/2$. Ceci implique :

$$L(\hat{f}_D, f) \leq \frac{2\sigma^2 D}{n} + R^2 D^{-2}.$$

Il reste à choisir D pour optimiser ce risque quadratique. En posant

$$D = [(nR^2)^{1/3}],$$

et on obtient

$$L(\hat{f}_D, f) \leq C(\sigma) R^{\frac{2}{3}} n^{-\frac{2}{3}}.$$

∎

3 Estimation sur des bases de splines

Nous supposons ici que les X_i appartiennent $\mathbb{R}$. Les estimateurs de la section précédente ne sont pas continus, pour obtenir des estimateurs qui sont des polynômes par morceaux et qui ont des propriétés de régularité, on utilise les bases de splines.

3.1 Splines linéaires et cubiques

$$f(x) = \beta_0 + \beta_1 x + \beta_2 (x-a)_+ + \beta_3 (x-b)_+ + \beta_4 (x-c)_+ + \ldots +$$

où $0 < a < b < c \ldots$ sont les points qui déterminent les intervalles de la partition (appelés les nœuds).

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x \text{ si } x \leq a \\ &= \beta_0 + \beta_1 x + \beta_2 (x-a)_+ \text{ si } a \leq x \leq b \\ &= \beta_0 + \beta_1 x + \beta_2 (x-a)_+ + \beta_3 (x-b)_+ \text{ si } b \leq x \leq c \end{aligned}$$

La fonction f est continue, si on veut imposer plus de régularité (par exemple f de classe C^2), on utilise des splines cubiques.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-a)_+^3 + \beta_5 (x-b)_+^3 + \beta_4 (x-c)_+^3 + \ldots +$$

La fonction $(x-a)^3$ s'annule ainsi que ses dérivées d'ordre 1 et 2 en a donc f est de classe C^2.
Pour éviter les problèmes de bords, on impose souvent des contraintes supplémentaires aux splines cubiques, notamment la linéarité de la fonction sur les deux intervalles correspondant aux extrémités.
On se place sur $[0,1]$. $\xi_0 = 0 < \xi_1 < \ldots < \xi_K < 1$.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3.$$

On impose $f''(0) = f^{(3)}(0) = 0$, $f''(\xi_K) = f^{(3)}(\xi_K) = 0$. On en déduit :

$$\beta_2 = \beta_3 = 0, \ \sum_{k=1}^{K} \theta_k (\xi_K - \xi_k) = 0, \ \sum_{k=1}^{K} \theta_k = 0.$$

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3] \\ &= \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k (\xi_K - \xi_k) [\frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{(\xi_K - \xi_k)}] \end{aligned}$$

On pose $\gamma_k = \theta_k(\xi_K - \xi_k)$ et $d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{(\xi_K - \xi_k)}$. $\sum_{k=1}^{K-1} \gamma_k = 0$.

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \gamma_k (d_k(x) - d_{K-1}(x)).$$

On obtient la base de splines naturels :

$$N_1(x) = 1, N_2(x) = x, \ \forall 1 \leq k \leq K-2, N_{k+2}(x) = d_k(x) - d_{K-1}(x).$$

On doit choisir la position et le nombre de nœuds.

3.2 Méthodes de régularisation

On se place dans un modèle de régression : $Y_i = f(X_i) + \epsilon_i$, $1 \leq i \leq n$. On minimise parmi les fonctions f splines naturels de nœuds en les X_i ($f(x) = \sum_{k=1}^{n} \theta_k N_k(x)$) le critère pénalisé :

$$C(f,\lambda) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f''(t))^2 dt,$$

où $\lambda > 0$. En notant $\Omega_{l,k} = \int_0^1 N_k''(x) N_l''(x) dx$ et $N_{i,j} = N_j(X_i)$, le critère à minimiser est

$$C(\theta,\lambda) = \|Y - N\theta\|^2 + \lambda \theta^* \Omega \theta.$$

La solution est :

$$\hat{\theta} = (N^* N + \lambda \Omega)^{-1} N^* Y$$

et

$$\hat{f}(x) = \sum_{k=1}^{n} \hat{\theta}_k N_k(x). \tag{2}$$

THÉORÈME 2. — *On note*

$$\mathcal{F} = \{f, \ C^2([0,1]), \int_0^1 f''^2(t) dt < +\infty\}.$$

On se donne $n \geq 2$, $0 < X_1 < \ldots < X_n < 1$ *et* $(y_1, \ldots, y_n) \in \mathbb{R}^n$. *Pour* $f \in \mathcal{F}$, *et* $\lambda > 0$, *on note*

$$C(f,\lambda) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f''(t))^2 dt.$$

Pour tout $\lambda > 0$, *il existe un unique minimiseur dans* $\mathcal{F}$ *de* $C(f,\lambda)$, *qui est la fonction définie en (2).*

4 Estimateurs à noyau

On considère le modèle

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{3}$$

où les $\boldsymbol{X}_i$ appartiennent à $\mathbb{R}^d$, les ε_i sont i.i.d. centrées de variance σ^2, les $\boldsymbol{X}_i$ et les ε_i sont indépendantes.

4.1 Définition des estimateurs a noyau.

DÉFINITION 3. — *On appelle noyau une fonction $K : \mathbb{R}^d \to \mathbb{R}$ telle que $\int K^2 < +\infty$ et $\int K = 1$.*

DÉFINITION 4. — *On se donne un réel $h > 0$ (appelé fenêtre) et un noyau K. On appelle estimateur à noyau de f dans le modèle (3) associé au noyau K et à la fenêtre h la fonction $\hat{f}_h$ définie par :*

$$\hat{f}_h(\boldsymbol{x}) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right)}.$$

Dans le cas où les $\boldsymbol{X}_i$ sont de loi uniforme sur $[0,1]^d$, on trouve aussi la définition suivante :

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^n Y_i K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right). \tag{4}$$

Si par exemple $d = 1$ et $K(u) = (1/2)\mathbf{1}_{|u|\leq 1}$, $\hat{f}_h(\boldsymbol{x})$ est la moyenne des Y_i tels que $|\boldsymbol{X}_i - \boldsymbol{x}| \leq h$. Il s'agit d'un estimateur constant par morceaux.
Cas extrêmes :
Supposons $d = 1$ et les X_i équirépartis sur $[0,1]$.
-Si $h = 1/n$, l'estimateur est très irrégulier et reproduit simplement les observations.
-Si $h \geq 1$, pour tout x, $\hat{f}_h(x) = \sum_{i=1}^n Y_i/n$.

Il faut donc, ici encore chercher une valeur de h qui réalise un bon compromis entre le terme de biais et le terme de variance.

Remarque : on utilise plus généralement des noyaux réguliers, ce qui permet d'obtenir des estimateurs réguliers.
Exemples de noyaux en dimension 1 **:**
-Le noyau triangulaire $K(x) = (1 - |x|)\mathbf{1}_{|x|\leq 1}$.
-Le noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
-Le noyau parabolique $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{|x|\leq 1}$.

4.2 Propriétés des estimateurs à noyau.

Pour simplifier les calculs, on se place dans un modèle où les X_i sont aléatoires, de loi uniforme sur $[0,1]$, et on considère l'estimateur défini en (4).

THÉORÈME 5. — *On suppose que $f \in \Sigma(\beta, R)$ définie par*

$$\Sigma(\beta, R) = \left\{ f \in \mathcal{C}^l([0,1]), \forall x, y \in [0,1], |f^{(l)}(x) - f^{(l)}(y)| \leq R|x-y|^\alpha \right\},$$

où $\beta = l + \alpha$ avec l entier et $\alpha \in]0,1]$.
On fait les hypothèses suivantes sur K :
H1 $\int u^j K(u)du = 0$ *pour* $j = 1, \ldots, l$.
H2 $\int |u|^\beta |K(u)|du < +\infty$.
En choisissant h de sorte que $h \approx (nR^2)^{-1/(1+2\beta)}$, on obtient, $\forall f \in \Sigma(\beta, R)$,

$$\mathbb{E}_f\left(\int_0^1 (\hat{f}_h(x) - f(x))^2\right) \leq C(\beta, \sigma, \|s\|_\infty) R^{\frac{2}{1+2\beta}} n^{-\frac{2\beta}{1+2\beta}}.$$

Démonstration. —

Calcul du biais : en notant $K_h = (1/h)K(./h)$,

$$\mathbb{E}_f(\hat{f}_h(x)) = \int_0^1 f(y)K_h(x-y)dy = f \star K_h(x).$$

On a alors, puisque $\int K = 1$,

$$\mathbb{E}_f(\hat{f}_h(x)) - f(x) = \int (f(x - uh) - f(x))K(u)du.$$

On utilise un développement de Taylor :

$$f(x-uh) = f(x) - f'(x)uh + f''(x)\frac{(uh)^2}{2} + \ldots + f^{(l)}(x-\tau uh)\frac{(-uh)^l}{l!}$$

avec $0 \leq \tau \leq 1$. En utilisant l'hypothèse **H1**,

$$\begin{aligned}\mathbb{E}_f(\hat{f}_h(x)) - f(x) &= \int f^{(l)}(x-\tau uh)\frac{(-uh)^l}{l!}K(u)du \\ &= \int (f^{(l)}(x-\tau uh) - f^{(l)}(x))\frac{(-uh)^l}{l!}K(u)du.\end{aligned}$$

Puisque $f \in \Sigma(\beta, R)$, et en utilisant l'hypothèse **H2**, on obtient

$$|\mathbb{E}_f(\hat{f}_h(x)) - f(x)| \leq R\tau^\alpha h^\beta \frac{1}{l!}\int |u|^\beta |K(u)|du.$$

Calcul de la variance :

$$\begin{aligned}Var(\hat{f}_h(x)) &= \frac{1}{n^2}\sum_{i=1}^n Var(Y_i K_h(x-X_i)). \\ &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_s[Y_i^2 K_h^2(x-X_i)] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[f^2(X_i)K_h^2(x-X_i) + \varepsilon_i^2 K_h^2(x-X_i)].\end{aligned}$$

De plus,

$$\begin{aligned}\mathbb{E}[f^2(X_i)K_h^2(x-X_i)] &= \int f^2(y)\frac{1}{h^2}K^2(\frac{x-y}{h})dy \\ &= \int f^2(x-uh)\frac{1}{h}K^2(u)du \\ &\leq \|f\|_\infty^2 \frac{1}{h}\int K^2.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\varepsilon_i^2 K_h^2(x-X_i)] &= \sigma^2 \int \frac{1}{h^2}K^2(\frac{x-y}{h})dy \\ &= \frac{\sigma^2}{h}\int K^2.\end{aligned}$$

Il en résulte que

$$Var(\hat{f}_h(x)) \leq C(\|f\|_\infty, \sigma)\frac{1}{nh}.$$

Puisque

$$\mathbb{E}_f\left(\int_0^1 (\hat{f}_h(x) - f(x))^2 dx\right) = \int_0^1 \left(Biais^2(\hat{f}_h(x)) + Var(\hat{f}_h(x))\right)dx,$$

on obtient

$$\mathbb{E}_f\left(\int_0^1 (\hat{f}_h(x) - f(x))^2 dx\right) \leq C(\beta, \sigma, \|f\|_\infty)\left(R^2 h^{2\beta} + \frac{1}{nh}\right).$$

En choisissant h de sorte que

$$R^2 h^{2\beta} \approx \frac{1}{nh},$$

c'est-à-dire $h \approx (nR^2)^{-1/(1+2\beta)}$, on obtient le résultat souhaité. ∎

5 Estimation ponctuelle par des polynômes locaux

Dans la section 2, nous nous étions donné une partition à priori, elle ne dépendait pas des observations. L'estimation de la fonction de régression en un point x était construite à partir des observations pour lesquelles X_i était dans le même intervalle de la partition que x, ce qui conduit à des estimateurs irréguliers. Une idée naturelle est d'estimer la fonction de régression en un point x à partir des observations pour lesquelles X_i est "proche" de x. Plus généralement, on introduit une fonction de poids $(w_i(x))$ construite à partir

d'un noyau : $w_i(x) = K((X_i - x)/h)$ qui va attribuer un poids plus important aux observations pour lesquelles X_i est "proche" de x, et on minimise (en a) la somme des carrés pondérée :

$$\sum_{i=1}^{n} w_i(x)(Y_i - a)^2.$$

La solution est donnée par

$$a = \hat{f}_n(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}, \tag{5}$$

ce qui correspond à l'estimateur à noyau de la fonction de régression ! On peut généraliser la formule ci-dessus en remplaçant la constante a par un polynôme de degré p : on se donne un point x en lequel on souhaite estimer la fonction de régression. Pour u dans un voisinage de x, on considère le polynôme

$$P_x(u, a) = a_0 + a_1(u - x) + \ldots + \frac{a_p}{p!}(u - x)^p.$$

On cherche à estimer la fonction de régression au voisinage de x par le polynôme $P_x(u, a)$ où le vecteur $a = (a_0, \ldots a_p)$ est obtenu par minimisation de la somme des carrés pondérée :

$$\sum_{i=1}^{n} w_i(x)(Y_i - a_0 - a_1(X_i - x) - \ldots - \frac{a_p}{p!}(X_i - x)^p)^2.$$

La solution obtenue est le vecteur $\hat{a}(x) = (\hat{a}_0(x), \ldots \hat{a}_p(x))$, l'estimateur local de la fonction de régression f est

$$\hat{f}_n(u) = \hat{a}_0(x) + \hat{a}_1(x)(u - x) + \ldots + \frac{\hat{a}_p(x)}{p!}(u - x)^p.$$

Au point x, où l'on souhaite réaliser l'estimation, on obtient :

$$\hat{f}_n(x) = \hat{a}_0(x).$$

Attention, cet estimateur ne correspond pas à celui que l'on obtient en (5), qui correspond à $p = 0$ (c'est l'estimateur à noyau). Si $p = 1$, on parle de régression linéaire locale. On peut expliciter la valeur de $\hat{a}_0(x)$ à partir d'un critère des moindres carrés pondérés : soit X_x la matrice

$$X_x = \begin{pmatrix} 1 & X_1 - x & \ldots & \frac{(X_1-x)^p}{p!} \\ 1 & X_2 - x & \ldots & \frac{(X_2-x)^p}{p!} \\ . & . & . & . \\ . & . & . & . \\ 1 & X_n - x & \ldots & \frac{(X_n-x)^p}{p!} \end{pmatrix}.$$

Soit W_x la matrice diagonale de i-ème élément sur la diagonale $w_i(x)$. On a alors :

$$\sum_{i=1}^{n} w_i(x)(Y_i - a_0 - a_1(X_i - x) - \ldots - \frac{a_p}{p!}(X_i - x)^p)^2 = (Y - X_x a)^* W_x (Y - X_x a).$$

Minimiser l'expression ci-dessus conduit à l'estimateur des moindres carrés pondérés :

$$\hat{a}(x) = (X_x^* W_x X_x)^{-1} X_x^* W_x Y,$$

et l'estimateur par polynômes locaux au point x correspond à $\hat{f}_n(x) = \hat{a}_0(x)$, c'est-à-dire au produit scalaire du vecteur Y avec la première ligne de la matrice $(X_x^* W_x X_x)^{-1} X_x^* W_x$. On obtient le théorème suivant :

THÉORÈME 6. — *L'estimateur par polynômes locaux au point x est*

$$\hat{f}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i$$

où $l(x)^* = (l_1(x), \ldots, l_n(x))$,

$$l(x)^* = e_1^* (X_x^* W_x X_x)^{-1} X_x^* W_x,$$

avec $e_1^* = (1, 0, \ldots, 0)$.

$$\mathbb{E}(\hat{f}_n(x)) = \sum_{i=1}^{n} l_i(x) f(X_i)$$

$$Var(\hat{f}_n(x)) = \sigma^2 \sum_{i=1}^{n} l_i^2(x).$$

6 Estimateurs par projection

On se place dans le modèle

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n. \tag{6}$$

Soit $(\phi_j, j \geq 1)$ une base orthonormée de $\mathbb{L}^2([0,1])$. On se donne $D \geq 1$ et on pose

$$S_D = \text{Vect}\{\phi_1, \dots, \phi_D\}.$$

On note f_D la projection orthogonale de f sur S_D dans $\mathbb{L}^2([0,1])$:

$$f_D = \sum_{j=1}^{D} \langle f, \phi_j \rangle \phi_j,$$

où

$$\theta_j = \langle f, \phi_j \rangle = \int_0^1 f(x)\phi_j(x)dx.$$

Il est naturel d'estimer θ_j par

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i).$$

En effet, si les X_i sont déterministes,

$$\mathbb{E}(\hat{\theta}_j) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)\phi_j(X_i),$$

et si $f\phi_j$ est régulière et les X_i équirépartis sur $[0,1]$, ceci est proche de θ_j. Si les X_i sont aléatoires, de loi uniforme sur $[0,1]$, on a

$$\mathbb{E}(\hat{\theta}_j) = \theta_j.$$

On introduit alors l'estimateur

$$\hat{f}_D(x) = \sum_{j=1}^{D} \hat{\theta}_j \phi_j(x),$$

appelé *estimateur par projection.*

Exemple de la base Fourier On note $(\phi_j, j \geq 1)$ la base trigonométrique de $\mathbb{L}^2([0,1])$:

$$\begin{aligned}
\phi_1(x) &= \mathbf{1}_{[0,1]}, \\
\phi_{2k}(x) &= \sqrt{2}\cos(2\pi kx) \quad \forall k \geq 1 \\
\phi_{2k+1}(x) &= \sqrt{2}\sin(2\pi kx) \quad \forall k \geq 1.
\end{aligned}$$

On obtient pour tout $D \geq 1$, l'estimateur

$$\hat{f}_D(x) = \frac{1}{n} \sum_{j=1}^{D} \sum_{i=1}^{n} Y_i \phi_j(X_i)\phi_j(x).$$

Nous allons énoncer les performances de l'estimateur, lorsque la fonction de régression f appartient à une classe de fonctions périodiques, régulières.

DÉFINITION 7. — *Soit $L > 0$ et $\beta = l + \alpha$ avec $l \in \mathbb{N}$ et $\alpha \in]0,1]$. On définit la classe $\Sigma^{per}(\beta, R)$ par*

$$\Sigma^{per}(\beta, R) = \Big\{ f \in \mathcal{C}^l([0,1]), \forall j = 0, \dots, l, \quad f^{(j)}(0) = f^{(j)}(1),$$
$$\forall x, y \in [0,1], |f^{(l)}(x) - f^{(l)}(y)| \leq R|x-y|^{\alpha} \Big\}.$$

THÉORÈME 8. — *Dans le modèle*

$$Y_i = f(\frac{i}{n}) + \varepsilon_i, \quad i = 1, \dots, n,$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$, l'estimateur $\hat{f}_D$ défini pour tout $x \in [0,1]$ par :

$$\hat{f}_D(x) = \frac{1}{n} \sum_{j=1}^{D} \sum_{i=1}^{n} Y_i \phi_j(X_i)\phi_j(x)$$

avec $D = [(nR^2)^{1/(1+2\beta)}]$, vérifie pour tout $\beta > 1$, $R > 0$,

$$\sup_{f \in \Sigma^{per}(\beta,R)} \mathbb{E}_f \left(\|\hat{f}_D - f\|_2^2 \right) \leq C(\beta, \sigma) R^{\frac{-2}{1+2\beta}} n^{\frac{2\beta}{1+2\beta}}.$$

Nous introduisons, dans le chapitre suivant, la définition des bases d'ondelettes, qui sont utilisées en particulier si la fonction à estimer est très irrégulière.

7 Bases d'ondelettes et estimation par seuillage

Dans ce chapitre, on s'intéresse à l'estimation de fonctions spatialement inhomogènes, c'est-à-dire qui peuvent être très régulières dans certaines zones puis très irrégulières (présentant des pics) dans certaines parties de l'espace. Les bases d'ondelettes sont des bases orthonormées, qui sont bien adaptées pour l'estimation de fonctions de ce type. Nous supposerons ici que les X_i appartiennent à $[0,1]$, mais nous traiterons également en TP des exemples en dimension 2 dans le cadre du traitement d'images.

7.1 Bases d'ondelettes

Base de Haar

La base de Haar est la base d'ondelettes la plus simple. L'ondelette père (ou fonction d'échelle) est définie par

$$\begin{aligned}\phi(x) &= 1 \text{ si } x \in [0,1[,\\ &= 0 \text{ sinon.}\end{aligned}$$

L'ondelette mère (ou fonction d'ondelette) est définie par

$$\begin{aligned}\psi(x) &= -1 \text{ si } x \in [0,1/2],\\ &= 1 \text{ si } x \in]1/2,1].\end{aligned}$$

Pour tout $j \in \mathbb{N}$, $k \in \mathbb{N}$, on pose

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k).$$

THÉORÈME 9. — *Les fonctions* $(\phi, \psi_{j,k}, j \in \mathbb{N}, k \in \{0,\dots,2^j-1\})$ *forment une base orthonormée de* $\mathbb{L}^2([0,1])$.

Il résulte de ce théorème que l'on peut développer une fonction de $\mathbb{L}^2([0,1])$ dans cette base :

$$f(x) = \alpha\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x).$$

$\alpha = \int_0^1 f(x)\phi(x)dx$ est appelé "coefficient d'échelle" et les $\beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)dx$ sont appelés "détails". On appelle approximation de f au niveau de résolution J la fonction

$$f_J = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x).$$

Cette expression comporte 2^J coefficients. Comme l'espace engendré par les fonctions $(\phi, \psi_{j,k}, 0 \le j \le J-1, 0 \le k \le 2^j-1)$ est l'espace des fonctions constantes par morceaux sur les intervalles de longueur $1/2^J$, c'est-à-dire l'espace engendré par les fonctions $(\phi_{J,k}, 0 \le k \le 2^J-1)$, on a aussi

$$f_J = \sum_{k=0}^{2^J-1}\alpha_{J,k}\phi_{J,k}(x),$$

où $\alpha_{J,k} = \int_0^1 f(x)\phi_{J,k}(x)dx$.

La base de Haar est simple à définir, les fonctions sont à support compact, néanmoins cette base fournit des approximations qui ne sont pas régulières. Il existe d'autres bases d'ondelettes à la fois à support compact et régulières, par exemple les ondelettes de Daubechies (voir Daubechies (1992) : *Ten Lectures on wavelets*).

7.2 Estimation d'une fonction de régression avec des ondelettes

Les ondelettes sont bien adaptées pour l'analyse des signaux recueillis sur une grille régulière, dyadique. On les utilise en traitement du signal et de l'image. On considère le modèle

$$Y_k = f(\frac{k}{N}) + \epsilon_k,\ k = 1,\dots,N = 2^J,$$

On considère les $N = 2^J$ premières fonctions d'une base d'ondelettes sur $[0,1]$: $(\phi, \psi_{j,k}, 0 \leq j \leq J-1, 0 \leq k \leq 2^j - 1)$. On note W la matrice $N * N$

$$W = \frac{1}{\sqrt{N}} \begin{pmatrix} \phi(1/N) & \psi_{0,0}(1/N) & \dots & \psi_{J-1,2^J-1}(1/N) \\ . & . & . & . \\ \phi(i/N) & \psi_{0,0}(i/N) & \dots & \psi_{J-1,2^J-1}(i/N) \\ . & . & . & . \\ \phi(N/N) & \psi_{0,0}(N/N) & \dots & \psi_{J-1,2^J-1}(N/N) \end{pmatrix}$$

Dans le cas de la base de Haar, W est une matrice orthogonale (la base est orthonormée pour le produit scalaire discret). On note W^* la transposée de W et

$$\hat{\theta} = W^* Y,$$

la tranformée en ondelettes du vecteur Y.
Il s'agit de l'estimateur des moindres carrés de θ dans le modèle $Y = W\theta + \varepsilon$ si W est orthogonale.

$$\begin{aligned} \hat{\theta}_{j,k} = \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N}) Y_l &= \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N}) f(\frac{l}{N}) + \tilde{\epsilon}_l \\ &\approx \sqrt{N}\beta_{j,k} + \tilde{\epsilon}_l \end{aligned}$$

où

$$\begin{aligned} \tilde{\epsilon}_l &= \frac{1}{\sqrt{N}} \sum_{l=1}^{N} \psi_{j,k}(\frac{l}{N}) \epsilon_l \\ &\sim \mathcal{N}(0, \frac{\sigma^2}{N} \sum_{l=1}^{N} \psi_{j,k}^2(\frac{l}{N})). \end{aligned}$$

Dans le cas de la base de Haar, $\frac{\sigma^2}{N} \sum_{l=1}^{N} \psi_{j,k}^2(\frac{l}{N}) = \sigma^2$. On peut reconstruire le signal à partir de sa transformée en ondelettes par la transformation inverse :

$$Y = (W^*)^{-1} \hat{\theta}.$$

$Y = W\hat{\theta}$dans le cas de la base de Haar.

Débruitage par approximation linéaire :

On approxime la fonction de régression f par projection orthogonale de f sur V_{J_0} :

$$f_{J_0} = \alpha\phi + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k},$$

ce qui correspond à regarder seulement les 2^{J_0} premiers coefficients d'ondelettes. Pour estimer f_{J_0}, dans $\hat{\theta}$, on ne garde que les 2^{J_0} premiers coefficients, les autres sont annulés, cela forme le vecteur noté $\hat{\theta}_{J_0}$, puis on reconstruit le signal débruité :

$$\hat{Y}_{J_0} = (W^*)^{-1} \hat{\theta}_{J_0}.$$

La fonction de régression f est alors estimée par

$$\hat{f}_{J_0}(x) = \frac{1}{\sqrt{N}} (\phi(x), \psi_{0,0}(x), \dots, \psi_{J-1,2^J-1}(x)) \hat{\theta}_{J_0}.$$

$$\hat{f}_{J_0}(x) = \hat{\alpha}\phi(x) + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}\psi_{j,k}(x)$$

où $\hat{\theta}_{J_0} = \sqrt{N}(\hat{\alpha}, \hat{\beta}_{j,k}, j = 0, \dots J_0 - 1, k = 0, \dots, 2^j - 1, 0, \dots, 0)$. Il faut choisir le paramètre J_0 de manière optimale.

Débruitage par approximation non linéaire via le seuillage :

La méthode de seuillage consiste à minimiser en $\theta \in \mathbb{R}^N$ le critère pénalisé avec une pénalité de type l_1 :

$$C(\theta) = \|Y - W\theta\|^2 + 2\lambda\|\theta\|_1,$$

avec $\|\theta\|_1 = \sum_{i=1}^{N} |\theta_i|$. Nous supposerons ici que la matrice W est orthogonale, ce qui permet de trouver une solution explicite.

$$\begin{aligned} C(\theta) &= \|Y\|^2 + \|W\theta\|^2 - 2\langle Y, W\theta\rangle + 2\lambda\|\theta\|_1, \\ &= \|Y\|^2 + \|\theta\|^2 - 2\theta^* W^* Y + 2\lambda\|\theta\|_1. \end{aligned}$$

Minimiser $C(\theta)$ équivaut à minimiser en θ

$$\begin{aligned} C'(\theta) &= \|\theta\|^2 - 2\theta^* W^* Y + 2\lambda\|\theta\|_1 \\ &= -2\sum_{i=1}^N \theta_i\hat{\theta}_i + 2\lambda\sum_{i=1}^N |\theta_i| + \sum_{i=1}^N \theta_i^2. \end{aligned}$$

Ceci est minimal si pour tout i, θ_i est du même signe que $\hat{\theta}_i$. On a donc $\theta_i\hat{\theta}_i = |\theta_i||\hat{\theta}_i|$.

$$\begin{aligned} C'(\theta) &= -2\sum_{i=1}^N |\theta_i||\hat{\theta}_i| + 2\lambda\sum_{i=1}^N |\theta_i| + \sum_{i=1}^N \theta_i^2 \\ &= \sum_{i=1}^N \Big(|\theta_i| - (|\hat{\theta}_i| - \lambda)\Big)^2 - \sum_{i=1}^N (|\hat{\theta}_i| - \lambda)^2. \end{aligned}$$

Minimiser ce critère en θ équivaut à minimiser

$$\sum_{i=1}^N \Big(|\theta_i| - (|\hat{\theta}_i| - \lambda)\Big)^2.$$

La solution est donc :

$$\begin{aligned} |\tilde{\theta}_i| &= |\hat{\theta}_i| - \lambda \text{ si } |\hat{\theta}_i| \geq \lambda \\ &= 0 \text{ si } |\hat{\theta}_i| \leq \lambda \end{aligned}$$

$$\tilde{\theta}_i = \text{signe}(\hat{\theta}_i)(|\hat{\theta}_i| - \lambda)\mathbf{1}_{|\hat{\theta}_i|\geq\lambda}.$$

Il s'agit du seuillage dit "doux" (soft thresholding), on applique une fonction continue à $\hat{\theta}_i$. Le seuillage dur ("soft thresholding") consiste à poser

$$\tilde{\theta}_i = \hat{\theta}_i\mathbf{1}_{|\hat{\theta}_i|\geq\lambda}.$$

on reconstruit le signal débruité :

$$\tilde{Y} = W\tilde{\theta}.$$

La fonction de régression f est estimée par

$$\hat{f}_N(x) = \frac{1}{\sqrt{N}}(\phi(x), \psi_{0,0}(x), \ldots, \psi_{J-1,2^J-1}(x))\tilde{\theta}.$$

En notant $\tilde{\theta} = \sqrt{N}(\tilde{\alpha}, \tilde{\beta}_{j,k}, j = 0, \ldots, J-1, k = 0, \ldots 2^{j-1})$, on obtient

$$\hat{f}_N(x) = \tilde{\alpha}\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \tilde{\beta}_{j,k}\psi_{j,k}(x).$$

En pratique, il faut choisir le seuil λ, on prend généralement $\lambda = \sigma\sqrt{2\log(N)}$.

$$\hat{\theta} = W^* Y = \frac{1}{\sqrt{N}}\sum_{l=1}^N \psi_{j,k}(\frac{l}{N})f(\frac{l}{N}) + \tilde{\epsilon}_l$$

avec

$$\tilde{\epsilon} = W^*\epsilon \sim \mathcal{N}_N(0, \sigma^2 I_N).$$

On peut montrer que

$$E\left(\sup_{1\leq i\leq N} |\tilde{\epsilon}_i|\right) \approx \sigma\sqrt{2\log(N)}.$$

Les coefficients qui sont inférieurs à $\sigma\sqrt{2\log(N)}$ sont considérés comme du bruit et sont annulés. Ces méthodes de seuillages fournissent des estimateurs permettant d'estimer des signaux très irréguliers (notamment des fonctions avec des pics).

8 Modèles additifs généralisés

Les méthodes d'estimation présentées précédemment vont se heurter au fléau de la dimension. Sous certaines hypothèses de structure sur la fonction de régression, on peut contourner ce problème. Nous allons nous intéresser ici à des fonctions de régression additives. Nous nous plaçons dans le modèle

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i,$$

où les ε_i sont i.i.d. centrées de variance σ^2, et les $\boldsymbol{X}_i \in \mathbb{R}^d$. Nous supposons que la fonction de régression f est additive, c'est-à-dire que

$$f(\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,d}) = \alpha + f_1(\boldsymbol{X}_{i,1}) + \ldots + f_d(\boldsymbol{X}_{i,d}).$$

Pour assurer l'unicité d'une telle écriture, on impose que

$$\int_{\mathbb{R}} f_j(x_j)dx_j = 0, \quad \forall j = 1, \ldots, d.$$

Nous allons décrire dans ce chapitre une méthode d'estimation des composantes de ce modèle additif, il s'agit des modèles GAM (Generalized Additive Models). Nous supposerons que chacune des fonctions unidimensionnelles est estimée à l'aide de Splines comme dans la section 3.2. On introduit alors le critère pénalisé :

$$\begin{aligned} \text{Crit}(\alpha, f_1, f_2, \ldots, f_p) &= \sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{j=1}^{d} f_j(X_{i,j}) \right)^2 \\ &+ \sum_{j=1}^{d} \lambda_j \int (f_j'')^2(x_j)dx_j, \end{aligned}$$

où les $\lambda_j \geq 0$ sont des paramètres de régularisation. On peut montrer que la solution de la minimisation de ce critère est un modèle de additif de splines cubiques, chaque fonction $\hat{f}_j$ étant un spline cubique de la variable x_j, dont les nœuds correspondent aux valeurs différentes des $X_{i,j}, i = 1, \ldots n$. Pour garantir l'unicité du minimiseur, on impose les contraintes

$$\forall j = 1, \ldots, d, \quad \sum_{i=1}^{n} f_j(X_{i,j}) = 0.$$

Sous ces conditions, on obtient $\hat{\alpha} = \sum_{i=1}^{n} Y_i/n$, et si la matrice des variables d'entrées $X_{i,j}$ n'est pas singulière, on peut montrer que le critère est strictement convexe, et admet donc un unique minimiseur. L'algorithme suivant, appelé algorithme de backfitting, converge vers la solution :

Algorithme de backfitting pour les modèles GAM :

1. Initialisation : $\hat{\alpha} = \sum_{i=1}^{n} Y_i/n$, $\hat{f}_j = 0 \; \forall j$.
2. Pour $l = 1$ à Niter
   Pour $j = 1$ à d
   - $\hat{f}_j$ minimise

$$\sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(X_{i,k}) - f_j(X_{i,j}) \right)^2 + \lambda_j \int (f_j'')^2(x_j)dx_j,$$

   - $\hat{f}_j := \hat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_j(X_{i,j})$.

   Arrêt lorsque toutes les fonctions $\hat{f}_j$ sont "stabilisées".

Le même algorithme peut être utilisée avec d'autres méthodes d'ajustement que les splines : estimateurs par polynômes locaux, à noyaux, par projection .. Les modèles additifs généralisés sont une extension des modèles linéaires, les rendant plus flexibles, tout en restant facilement interprétables. Ces modèles sont très largement utilisés en modélisation statistique, néanmoins, en très grande dimension, il est difficile de les mettre en œuvre, et il sera utile de les combiner à un algorithme de sélection (pour réduire la dimension).

9 Kernel Regression Least Square

Un exemple élémentaire de *machine à noyau*.

- L'objectif est ici de présenter une méthode qui fournit des prédicteurs non linéaires.
- Le point commun avec les méthodes présentées précédemment est qu'il s'agit d'une méthode de régularisation basée sur un critère des moindres carrés pénalisés.
- On note $(\boldsymbol{X_i}, Y_i)_{1 \leq i \leq n}$ les observations, avec $\boldsymbol{X_i} \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$.
- On se donne un noyau k défini sur $\mathbb{R}^p$, symétrique, semi-défini positif :

$$k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x}); \quad \sum_{i,j=1}^{n} c_i c_j k(\boldsymbol{X_i}, \boldsymbol{X_j}) \geq 0.$$

- Exemples de noyaux sur $\mathbb{R}^p$:
  - Linéaire :
  $$k(\boldsymbol{X_i}, \boldsymbol{X_j}) = \boldsymbol{X_i}' \boldsymbol{X_j} = \langle \boldsymbol{X_i}, \boldsymbol{X_j} \rangle$$
  - Polynomial :
  $$k(\boldsymbol{X_i}, \boldsymbol{X_j}) = (\boldsymbol{X_i}' \boldsymbol{X_j} + 1)^d$$

– Gaussien :

$$k(\boldsymbol{X_i}, \boldsymbol{X_j}) = \exp\left(\frac{-\|\boldsymbol{X_i} - \boldsymbol{X_j}\|^2}{\sigma^2}\right).$$

- On cherche un prédicteur de la forme

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} c_j k(\boldsymbol{X_j}, \boldsymbol{x}),\ \boldsymbol{c} \in \mathbb{R}^n.$$

- On note $\boldsymbol{K}$ la matrice définie par $\boldsymbol{K}_{i,j} = k(\boldsymbol{X_i}, \boldsymbol{X_j})$.
- La méthode consiste à minimiser pour f de la forme ci-dessus le critère des moindres carrés pénalisés :

$$\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X_i}))^2 + \lambda\|f\|^2_{\boldsymbol{K}},$$

où

$$\|f\|^2_{\boldsymbol{K}} = \sum_{i,j=1}^{n} c_i c_j k(\boldsymbol{X_i}, \boldsymbol{X_j}).$$

- De manière équivalente, on minimise pour $c \in \mathbb{R}^n$ le critère

$$\|\boldsymbol{Y} - \boldsymbol{K}\boldsymbol{c}\|^2 + \lambda \boldsymbol{c'}\boldsymbol{K}\boldsymbol{c}.$$

- La solution est explicite :

$$\hat{\boldsymbol{c}} = (\boldsymbol{K} + \lambda I_n)^{-1} Y.$$

- On obtient le prédicteur

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \hat{c}_j k(\boldsymbol{X_j}, \boldsymbol{x}).$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{K}\hat{\boldsymbol{c}}.$$

- Avec le noyau correspondant au produit scalaire, on retrouve un estimateur linéaire :

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X'}, \hat{\boldsymbol{c}} = (\boldsymbol{X}\boldsymbol{X'} + \lambda I_n)^{-1} Y,$$

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \hat{c}_j \langle \boldsymbol{X_j}, \boldsymbol{x} \rangle.$$

- La méthode fournit des estimateurs non linéaires pour les noyaux polynomiaux ou gaussiens par exemple.
- Un intérêt important de la méthode précédente est la possibilité de généralisation à des prédicteurs $\boldsymbol{X_i}$ qui ne sont pas nécessairement dans $\mathbb{R}^p$ mais qui peuvent être de nature complexe (graphes, séquence d'ADN ..) dès lors que l'on sait définir un noyau $k(\boldsymbol{x}, \boldsymbol{y})$ symétrique et semi-défini positif agissant sur ces objets.
- Ceci fait appel à la théorie des **RKHS** *Reproducing Kernel Hilbert Spaces* ou *Espaces de Hilbert à noyau reproduisant.*

10 Arbres de régression CART

Les méthodes basées sur les arbres reposent sur une partition de l'espace des variables d'entrée, puis on ajuste un modèle simple (par exemple un modèle constant) sur chaque élément de la partition. On suppose que l'on a un échantillon de taille n : $(\boldsymbol{X}_i, Y_i)_{1 \leq i \leq n}$ avec $\boldsymbol{X}_i \in \mathbb{R}^d$ et $Y_i \in \mathbb{R}$. L'algorithme CART permet de définir, à partir de l'échantillon d'apprentissage, une partition automatique de l'espace des variables d'entrées $\boldsymbol{X}_i$. Supposons que l'espace où varient les $\boldsymbol{X}_i$ soit partitionné en M régions, notées $R_1, \ldots R_M$. On introduit la classe F des fonctions constantes par morceaux sur chacune des régions :

$$F = \{f, f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathbf{1}_{\boldsymbol{x} \in R_m}\}.$$

L'estimateur des moindres carrés de la fonction de régression f sur la classe F minimise le critère

$$\sum_{m=1}^{M}(Y_i - f(\boldsymbol{X}_i))^2,$$

parmi les fonctions $f \in F$. La solution est

$$\hat{f}(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{c}_m \mathbf{1}_{\boldsymbol{x} \in R_m},$$

où $\hat{c}_m$ est la moyenne des observations Y_i pour lesquelles $\boldsymbol{X}_i \in R_m$. Pour construire la partition, CART procède de la manière suivante : étant donné

une variable de séparation $X^{(j)}$ et un point de séparation s, on considère les demi-espaces

$$R_1(j,s) = \{\boldsymbol{X} = (X^{(1)}, \ldots, X^{(d)}) / X^{(j)} \leq s\} \text{ et } R_2(j,s) = \{\boldsymbol{X} / X^{(j)} > s\}.$$

La variable de séparation $X^{(j)}$ et un point de séparation s sont choisis de manière à résoudre

$$\min_{j,s}[\sum_{i,\boldsymbol{X}_i \in R_1(j,s)} (Y_i - \hat{c}_1)^2 + \sum_{i,\boldsymbol{X}_i \in R_2(j,s)} (Y_i - \hat{c}_2)^2].$$

Ayant déterminé j et s, on partitionne les données en les deux régions correspondantes, puis on recommence la procédure de séparation sur chacune des deux sous-régions, et ainsi de suite sur chacune des sous-régions obtenues. La taille de l'arbre est un paramètre à ajuster, qui va gouverner la complexité du modèle : un arbre de trop grande taille va conduire à un sur-ajustement (trop grande variance), au contraire un arbre de petite taille va mal s'ajuster à la fonction de régression (biais trop élevé). Il est donc nécessaire de choisir une taille "optimale" de manière adaptative à partir des observations. La stratégie adoptée consiste à construire un arbre de grande taille, puis à l'élaguer en introduisant un critère pénalisé. On dira que T est un sous-arbre de T_0 si T peut être obtenu en élaguant T_0, c'est-à-dire en réduisant le nombre de nœuds de T_0. On note $|T|$ le nombre de nœuds terminaux de l'arbre T et $R_m, m = 1, \ldots |T|$, la partition correspondant à ces nœuds terminaux. On note N_m le nombre d'observations pour lesquelles $\boldsymbol{X}_i \in R_m$. On a donc

$$\hat{c}_m = \frac{1}{N_m} \sum_{i,\boldsymbol{X}_i \in R_m} Y_i,$$

et on introduit le critère

$$C_\lambda(T) = \sum_{m=1}^{|T|} \sum_{i,\boldsymbol{X}_i \in R_m} (Y_i - \hat{c}_m)^2 + \lambda |T|.$$

Pour tout λ, on peut montrer qu'il existe un unique arbre minimal T_λ qui minimise le critère $C_\lambda(T)$. Pour trouver l'arbre T_λ, on supprime par étapes successives le nœud interne de l'arbre T qui réduit le moins le critère $\sum_m \sum_{i,\boldsymbol{X}_i \in R_m} (Y_i - \hat{c}_m)^2$. Ceci donne une succession de sous-arbres, dont on peut montrer qu'elle contient l'arbre T_λ.

Le paramètre de régularisation λ doit à son tour être calibré pour réaliser un bon compromis entre le biais et la variance de l'estimateur ainsi obtenu, ou de manière équivalente entre un bon ajustement aux données et une taille pas trop importante pour l'arbre. La méthode de validation croisée, décrite en annexe, peut être utilisée.

Annexe : Choix d'un paramètre de lissage par validation croisée

Dans le cas des estimateurs à noyaux, et pour les estimateurs par polynômes locaux, on doit choisir la fenêtre h ; pour les estimateurs constants par morceaux (ou polynômes par morceaux), ainsi que pour les estimateurs par projection, on doit choisir un paramètre D (nombre de morceaux de la partition ou dimension de l'espace de projection sur lequel on réalise l'estimation), pour les arbres CART, on doit choisir le paramètre λ de la procédure d'élaguage. Dans ce chapitre, nous allons décrire la méthode de validation croisée, qui est une méthode possible pour choisir ces paramètres, ce qui correspond à sélectionner un estimateur dans une collection d'estimateurs.

Notons λ le paramètre à choisir. Soit $\hat{f}_{n,\lambda}$ l'estimateur de la fonction de régression f associé à ce paramètre λ. On considère l'erreur quadratique moyenne :

$$R(\lambda) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{n,\lambda}(\boldsymbol{X}_i) - f(\boldsymbol{X}_i))^2\right).$$

Idéalement, on souhaiterait choisir λ de manière à minimiser $R(\lambda)$, mais cette quantité dépend de la fonction inconnue f.

Une première idée est d'estimer $R(\lambda)$ par l'**erreur d'apprentissage** :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2,$$

mais cette quantité sous-estime $R(\lambda)$ et conduit à un sur-ajustement. Ceci est dû au fait que l'on utilise les mêmes données pour construire l'estimateur $\hat{f}_{n,\lambda}$

(qui est construit pour bien s'ajuster à l'échantillon d'apprentissage) et pour estimer l'erreur commise par cet estimateur. Pour avoir une meilleure estimation du risque, on doit construire l'estimateur du risque avec des observations qui n'ont pas été utilisées pour construire l'estimateur $\hat{f}_{n,\lambda}$. Idéalement, si on avait assez d'observations, on pourrait les séparer en un échantillon d'apprentissage et un échantillon test. Ce n'est généralement pas le cas, et on souhaite utiliser l'ensemble des données d'apprentissage pour la construction de l'estimateur. On va alors avoir recours à la validation croisée. On partitionne l'échantillon d'apprentissage en V blocs, notés $B_1, \dots B_V$, de tailles à peu près identiques. Pour tout v de 1 à V, on note $\hat{f}_{n,\lambda}^{(-v)}$ l'estimateur obtenu en supprimant de l'échantillon d'apprentissage les données appartenant au bloc B_v.

DÉFINITION 10. — *On définit le score de validation croisée V-fold par :*

$$CV = \hat{R}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{f}_{n,\lambda}^{(-v(i))}(\boldsymbol{X}_i))^2,$$

où $\hat{f}_{n,\lambda}^{(-v(i))}$ est l'estimateur de f obtenu en enlevant les observations du bloc qui contient l'observation i.

Le principe de la validation croisée est de choisir une valeur $\hat{\lambda}$ de λ qui minimise la quantité $\hat{R}(\lambda)$. Un cas particulier correspond à la validation croisée leave-one-out, obtenue quand on considère n blocs, chacun réduits à une observation.

DÉFINITION 11. — *Le score de validation croisée leave-one-out est défini par :*

$$CV = \hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2,$$

où $\hat{f}_{n,\lambda}^{(-i)}$ est l'estimateur de f obtenu en enlevant l'observation $(\boldsymbol{X}_i, Y_i)$.

L'idée de la validation croisée leave-one-out vient du calcul suivant :

$$\begin{aligned}\mathbb{E}((Y_i - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2) &= \mathbb{E}((Y_i - f(\boldsymbol{X}_i) + f(\boldsymbol{X}_i) - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2)\\ &= \sigma^2 + \mathbb{E}((f(\boldsymbol{X}_i) - \hat{f}_{n,\lambda}^{(-i)}(\boldsymbol{X}_i))^2)\\ &\simeq \sigma^2 + \mathbb{E}((f(\boldsymbol{X}_i) - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2).\end{aligned}$$

On obtient donc $\mathbb{E}(\hat{R}(\lambda)) \simeq \sigma^2 + R(\lambda)$.

Le calcul de $\hat{R}(\lambda)$ peut s'avérer long, mais dans certains cas, il n'est pas nécessaire de recalculer n fois un estimateur de la fonction de régression. Pour la plupart des méthodes traitées dans ce chapitre, l'estimateur correspond à un algorithme de moyennes locales, c'est-à-dire est de la forme

$$\hat{f}_{n,\lambda}(\boldsymbol{x}) = \sum_{j=1}^{n} Y_j l_j(\boldsymbol{x}),$$

avec $\sum_{j=1}^{n} l_j(\boldsymbol{x}) = 1$, et on peut montrer que

$$\hat{f}_{n,\lambda}^{(-i)}(x) = \sum_{j=1}^{n} Y_j l_j^{(-i)}(\boldsymbol{x}),$$

avec

$$\begin{aligned} l_j^{(-i)}(\boldsymbol{x}) &= 0 \text{ si } j = i\\ &= \frac{l_j(\boldsymbol{x})}{\sum_{k\neq i} l_k(\boldsymbol{x})} \text{ si } j \neq i.\end{aligned}$$

THÉORÈME 12. — *Sous les hypothèses ci-dessus concernant l'estimateur, le score de validation croisée leave-one-out est égal à :*

$$CV = \hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i)}{1 - l_i(\boldsymbol{X}_i)}\right)^2.$$

On trouve également dans les logiciels une définition légèrement différente :

DÉFINITION 13. — *On appelle score de validation croisée généralisée la quantité :*

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i)}{1 - \nu/n}\right)^2,$$

où $\nu/n = \sum_{i=1}^{n} l_i(\boldsymbol{X}_i)/n$.

Dans cette définition, $l_i(\boldsymbol{X}_i)$ est remplacé par la moyenne des $l_i(\boldsymbol{X}_i)$. En pratique, les deux méthodes donnent généralement des résultats assez proches. En utilisant l'approximation $(1-x)^{-2} \approx 1+2x$ pour x proche de 0, on obtient :

$$GCV(\lambda) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2 + \frac{2\nu\hat{\sigma}^2}{n},$$

où $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,\lambda}(\boldsymbol{X}_i))^2$. Cela correspond au critère C_p de Mallows.