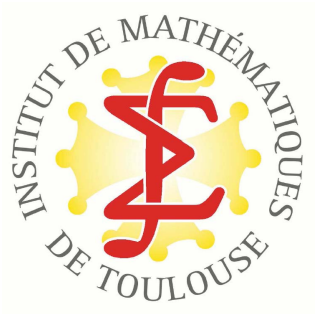


Apprentissage de données fonctionnelles

Formation continue
SII- Nov 2018

Béatrice Laurent-Philippe Besse

INSA Toulouse, Institut de Mathématiques de Toulouse

Introduction

Dans le contexte de **l'apprentissage supervisé**, on dispose d'un **échantillon d'apprentissage** composée d'observations de type **entrées/sorties** :

$$d_1^n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

avec $\boldsymbol{x}_i \in \mathbb{R}^p$, $y_i \in \mathcal{Y}$ pour $i = 1 \ldots n$.

Objectifs : A partir de l'échantillon d'apprentissage, on veut

- **Estimer** le lien entre le vecteur des entrées $\boldsymbol{x}$ (variables explicatives) et la sortie y (variable à expliquer) :

$$y = f(x^1, x^2, \ldots, x^p)$$

- **Prédire** la sortie y associé à une nouvelle entrée $\boldsymbol{x}$,
- **Sélectionner** les variables explicatives importantes parmi $x^1, \ldots, x^p$.

sortie quantitative	sortie qualitative
$\mathcal{Y} \subset \mathbb{R}^p$	$\mathcal{Y}$ fini
$\downarrow$	$\downarrow$
régression	**classification**

- Dans ce cours, on considère l'apprentissage supervisé pour la régression réelle ($\mathcal{Y} \subset \mathbb{R}$).
- Les variables explicatives $X^1, \dots X^p$ sont supposées être quantitatives
- Nous verrons l'importance du principe de parcimonie : "Il s'agit de déterminer un modèle qui fournit une bonne représentation des données, avec le moins de paramètres possibles ".

Evaluation du risque

On suppose que d_1^n est l'observation d'un n-échantillon $D_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ d'une loi conjointe P sur $\mathcal{X} \times \mathcal{Y}$, inconnue, et que x est une observation de la variable X, (X, Y) étant un couple aléatoire de loi conjointe P indépendant de D_1^n.

L'échantillon d_1^n est appelé **échantillon d'apprentissage**.

Une **règle de prédiction / régression ou classification** est une fonction (mesurable) $f : \mathcal{X} \to \mathcal{Y}$ qui associe la sortie $f(x)$ à l'entrée $x \in \mathcal{X}$.

Qualité de prédiction

Une fonction (mesurable) $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ est une **fonction de perte** si $l(y, y) = 0$ et $l(y, y') > 0$ pour $y \neq y'$.

Si f est une règle de prédiction, x une entrée, y la sortie qui lui est réellement associée, alors $l(y, f(x))$ mesure une perte encourue lorsque l'on associe à x la sortie $f(x)$.

En régression réelle : pertes $\mathbb{L}^p$ $(p \geq 1)$ $l(y, y') = |y - y'|^p$.
Si $p = 1$ on parle de perte absolue, si $p = 2$ de perte quadratique.

Qualité de prédiction

Etant donnée une fonction de perte l, le **risque** - ou l'**erreur de généralisation** - d'une règle de prédiction f est défini par

$$R_P(f) = \mathbb{E}_{(X,Y)\sim P}[l(Y, f(X))].$$

Soit $\mathcal{F}$ l'ensemble des règles de prédiction possibles.

On dira que f^* est une règle optimale si

$$R_P(f^*) = \inf_{f\in\mathcal{F}} R_P(f).$$

Peut-on construire des règles optimales ?

On appelle **fonction de régression** la fonction $f^* : \mathcal{X} \to \mathcal{Y}$ définie par
$$f^*(x) = \mathbb{E}[Y|X = x].$$

Cas de la régression réelle

$$\mathcal{Y} = \mathbb{R}, \quad l(y, y') = (y - y')^2$$

Theorem

La fonction de régression $f^* : x \mapsto \mathbb{E}[Y|X = x]$ *vérifie*
$$R_P(f^*) = \inf_{f \in \mathcal{F}} R_P(f).$$

Minimisation du risque empirique

- Parmi une collection de règles de prédictions possibles, on cherche à minimiser

$$\mathbb{E}_{(X,Y)\sim P}\left[(Y - f(X))^2\right].$$

- Première idée : minimiser le risque empirique

$$\frac{1}{n}\sum_{i=1}^{n}\left[(y_i - f(\boldsymbol{x}_i))^2\right].$$

Minimisation du risque empirique

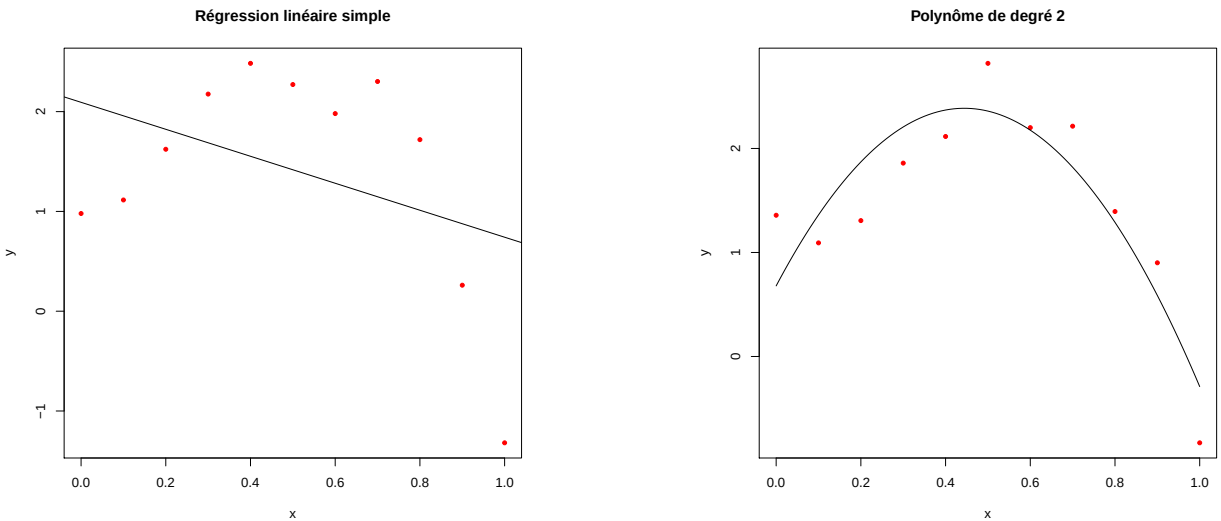


FIGURE: Régression polynomiale : modèle ajusté, à gauche : $y = \beta_0 + \beta_1 x + \epsilon$, $R^2 = 0.03$, à droite : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, $R^2 = 0.73$.

Minimisation du risque empirique

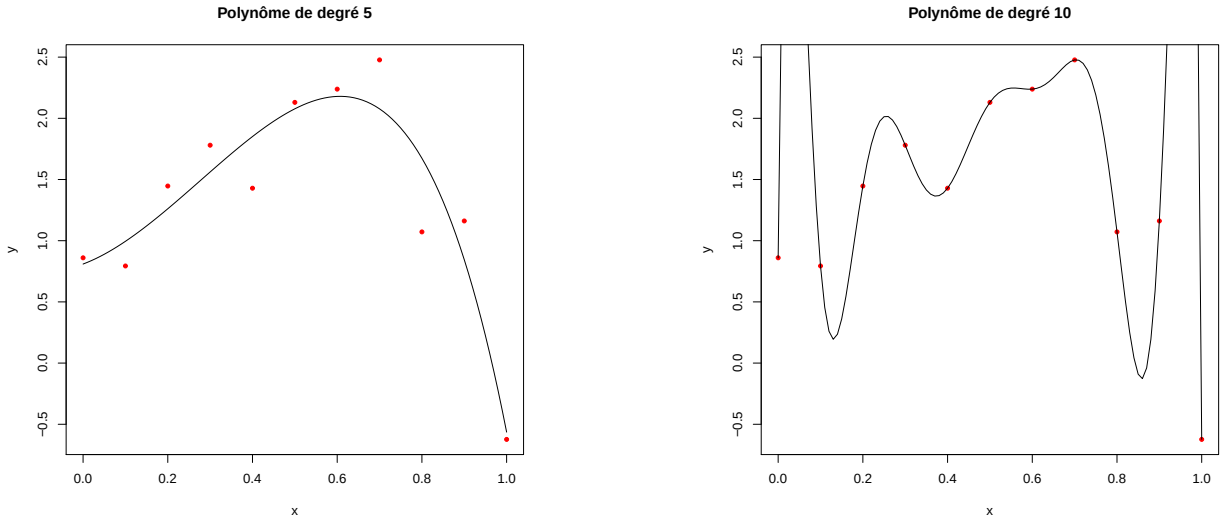


FIGURE: Régression polynomiale : modèle ajusté, à gauche : $y = \beta_0 + \beta_1 x + \ldots + \beta_5 x^5 + \epsilon$, $R^2 = 0.874$, à droite : $y = \beta_0 + \beta_1 x + \ldots + \beta_{10} x^{10} + \epsilon$, $R^2 = 1$.

Le risque empirique est égal à 0 pour le polynôme de degré $n-1$ (qui a n coefficients) et passe par tous les points de l'échantillon d'apprentissage.

Sélection de modèle

- Le meilleur modèle est celui qui réalise un bon compromis entre le terme de biais et le terme de variance.
- Minimiser le risque empirique n'est pas un bon critère pour comparer des modèles de complexité différentes.
- Solutions :
  - Minimiser le risque empirique pénalisé afin d'éviter le sur-ajustement
  - Utiliser la validation croisée.

Critères pénalisés

The Mallow's C_P criterion is

$$\text{Crit}_{C_P}(f_k) = \sum_{i=1}^{n} (y_i - f_k(\boldsymbol{x}_i))^2 + 2k\sigma^2,$$

and the BIC criterion penalizes more the dimension of the model with an additional logarithmic term.

$$\text{Crit}_{BIC}(f_k) = \sum_{i=1}^{n} (y_i - f_k(\boldsymbol{x}_i))^2 + \log(n)k\sigma^2.$$

The aim is to select the model (among all possible subsets) that minimizes one of those criterion. On the example of the polynomial models, we obtain the results summarized in the next Figure.

Critères pénalisés

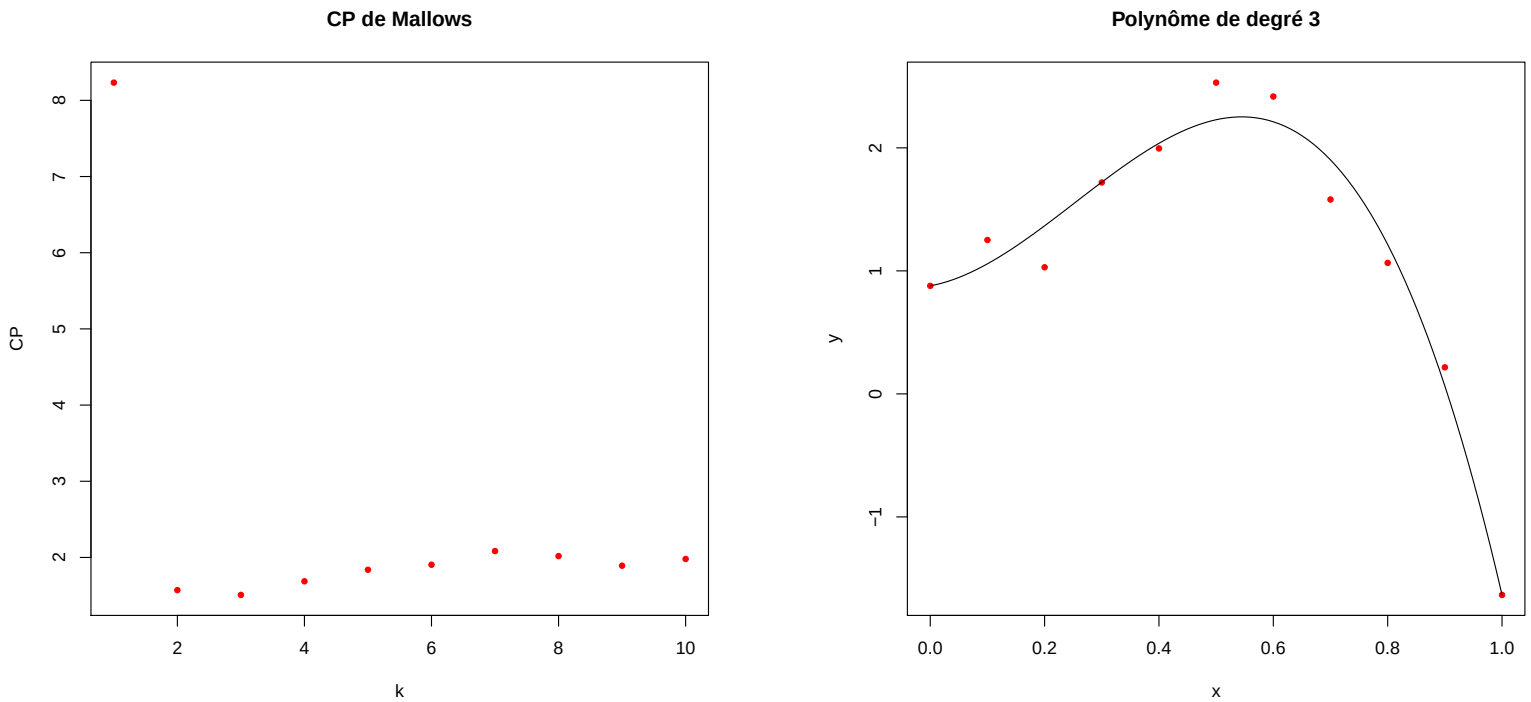


FIGURE: Mallows' C_P in function of the degree of the polynomial. Selected model : polynomial with degree 3.

Validation croisée

- Supposons que l'on dispose d'une procédure d'estimation qui dépend d'un paramètre λ. Comment choisir ce paramètre de manière à sélectionner le meilleur estimateur ?
- La plupart des logiciels utilisent la **validation croisée :**
- On découpe les données d'apprentissage en K sous-ensembles. Pour l de 1 à K :
  - On calcule l'estimateur $\hat{f}_\lambda$ associé au paramètre λ à partir de tous les sous-ensembles, sauf le l-ième (qui sera l'échantillon "test").
  - On note $\hat{\boldsymbol{f}}_\lambda^{(-l)}$ l'estimateur obtenu.
  - On note $\tau(i)$ le sous-ensemble contenant la ième observation. On calcule le critère :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{f}}_\lambda^{(-\tau(i))}(\boldsymbol{x}_i))^2.$$

  - On choisit la valeur de λ qui minimise $CV(\lambda)$.
- Le principe est donc de tester les performances d'un l'estimateur sur des données qui n'ont pas servi à le construire.

Plan de l'exposé

- Estimation non paramétrique en régression réelle.
  - Polynômes par morceaux
  - Splines
  - Estimateurs à noyaux
  - Estimateurs par projection
- Modèles additifs

Modèle de régression

- On se place dans le cadre d'un modèle de régression : les $(X_i, Y_i)_{1 \leq i \leq n}$ sont i.i.d. et obéissent au modèle

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

- Les variables X_i appartiennent à $\mathbb{R}^d$, les Y_i sont réelles.
  - On suppose que

$$\mathbb{E}(\varepsilon_i / X_i) = 0,$$
$$Var(\varepsilon_i / X_i) = \sigma^2.$$

- La fonction de régression $f^* : x \mapsto \mathbb{E}[Y|X = x] = f(x)$.
- En l'absence de toute hypothèse sur la fonction de régression f, nous sommes dans un cadre non paramétrique.

Estimation par des polynômes par morceaux

Estimation par des constantes par morceaux : régressogramme

- On suppose que les X_i appartiennent à $[0,1]$, on découpe $[0,1]$ en D intervalles de même taille :

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

- Sur l'intervalle $I_{k,D}$, f est estimée par la moyenne des valeurs de Y_i qui sont telles que $X_i \in I_{k,D}$:

$$\forall x \in I_{k,D}, \hat{f}_D(x) = \frac{\sum_{i, X_i \in I_{k,D}} Y_i}{\sharp\{i, X_i \in I_{k,D}\}} \text{ si } \sharp\{i, X_i \in I_{k,D}\} \neq 0,$$

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

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

Constantes par morceaux

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

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

- On minimise $h(a_1, \ldots, a_D) = \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{D} a_k \mathbb{1}_{X_i \in I_{k,D}} \right)^2$.
- La minimisation est obtenue pour

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

Sélection de modèles

- Choix de D dans le cas de l'estimation par des constantes par morceaux, deux cas extrêmes :
  1. 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 est en situation de surajustement.
  2. 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.

Performances de l'estimateur

Theorem

Soit $\mathcal{S}_{1,R} = \left\{ f \in \mathbb{L}^2([0,1]), \forall x, y \in [0,1], |f(x) - f(y)| \leq R|x-y| \right\}$.
Dans le modèle $Y_i = f(\frac{i}{n}) + \varepsilon_i, \quad i = 1, \ldots, n$, *l'estimateur*

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

avec

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

vérifie

$$\sup_{f \in \mathcal{S}_{1,R}} \mathbb{E}_f \left[\int_0^1 (\hat{f}_D - f)^2 \right] \leq C(\sigma) R^{\frac{2}{3}} n^{-\frac{2}{3}}$$

où $C(\sigma)$ *est une constante positive qui dépend de* σ.

Performances de l'estimateur

- Le théorème donne une majoration du risque quadratique de l'estimateur dans le cas où la fonction de régression f est Lipschitzienne.
- Résultat théorique, en pratique, on ne sait pas si la fonction f appartient à la classe $\mathcal{S}_{1,R}$.
- En pratique, on peut utiliser la validation croisée pour le choix de D.

Plan de l'exposé

- Estimation non paramétrique en régression réelle.
  - Polynômes par morceaux
  - Splines
  - Estimateurs à noyaux
  - Estimateurs par projection
- Modèles additifs

Estimation sur des bases de splines

- On suppose $X_i \in \mathbb{R}$.
- Les estimateurs construits à la section précédente sont discontinus.
- Nous allons introduire des estimateurs qui sont des polynômes par morceaux, et qui ont des propriétés de régularité.
- Pour cela, on utilise les bases de Splines.

Splines cubiques

Si on veut imposer une régularité de classe C^2 pour l'estimateur de la fonction de régression, on utilise des splines cubiques.

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

où $a < b < c < \dots$ sont les noeuds.

- 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.

Splines cubiques

- 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 $f''(0) = f^{(3)}(0) = 0$, $f''(\xi_K) = f^{(3)}(\xi_K) = 0$.
- Alors

$$f(x) = \sum_{k=1}^{K} \theta_k N_k(x),$$

avec

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

$$\forall 1 \leq k \leq K-1, d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{(\xi_K - \xi_k)}.$$

Splines cubiques - Méthodes de régularisation

- Comment choisir le nombre et la position des noeuds ? On prend tous les points X_i de la base d'apprentissage.
- On introduit un critère pénalisé : on cherche parmi les fonctions de la forme $f(x) = \sum_{k=1}^{K} \theta_k N_k(x)$, celle qui minimise le critère :

$$C(f, \lambda) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f''(t))^2 dt, \lambda > 0.$$

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 :

$$\hat{\theta} = (N^* N + \lambda \Omega)^{-1} N^* Y, \quad \hat{f}(x) = \sum_{k=1}^{n} \hat{\theta}_k N_k(x).$$

Splines cubiques - Méthodes de régularisation

Theorem

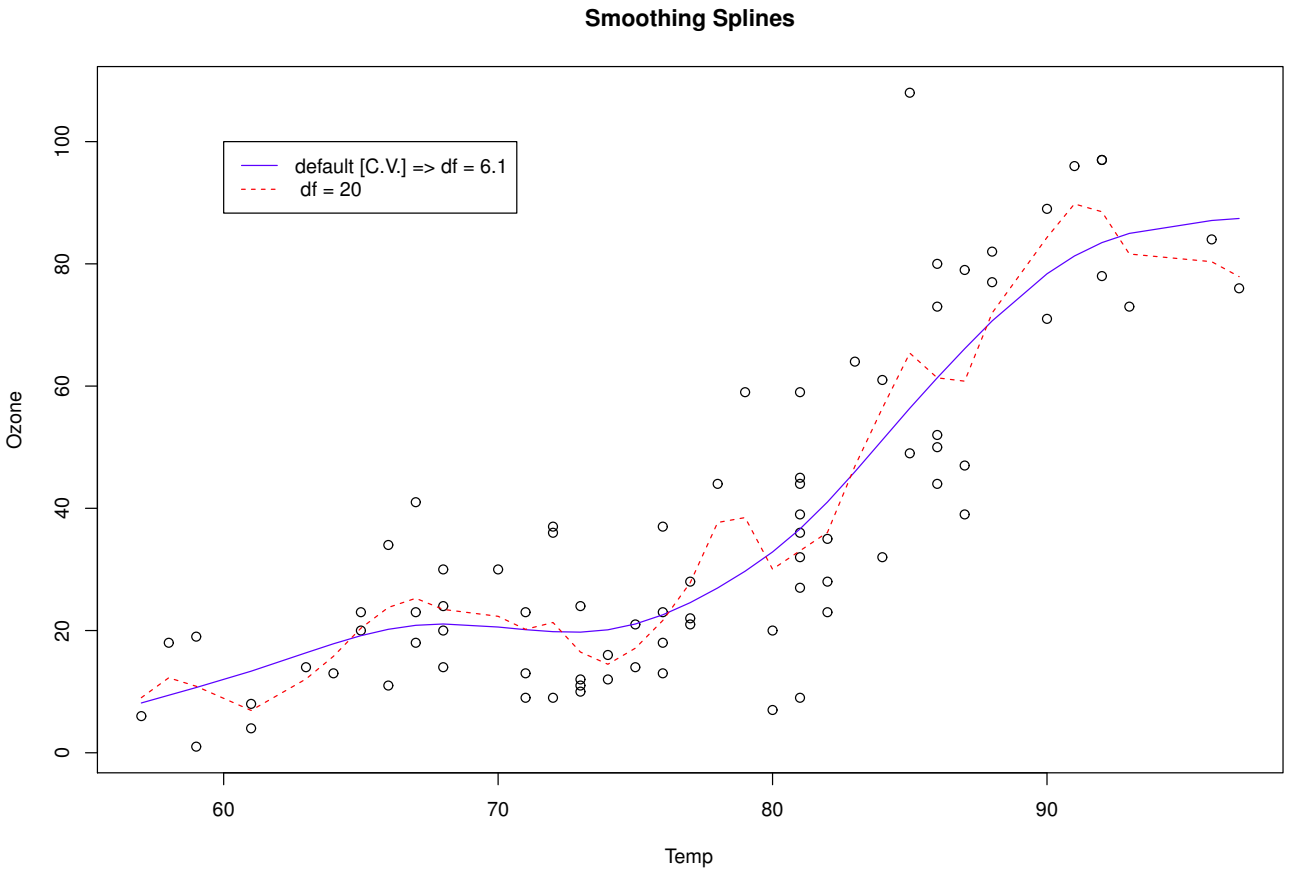
On note $\mathcal{F} = \left\{ f,\ C^2([0,1]), \int_0^1 f''^2(t)dt < +\infty \right\}$.
On se donne $n \geq 2$, $0 < X_1 < \ldots < X_n < 1$ *et* $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$. *Pour* $f \in \mathcal{F}$, *et* $\lambda > 0$, *on note*

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

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

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

Estimateur sur une base de Splines



Plan de l'exposé

- Estimation non paramétrique en régression réelle.
  - Polynômes par morceaux
  - Splines
  - Estimateurs à noyaux
  - Estimateurs par projection
- Modèles additifs

Estimateurs à noyau

On considère le modèle

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

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

Definition

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

Estimateurs à noyau

Definition

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

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

Dans le cas où les X_i sont de loi uniforme sur $[0,1]^d$, on trouve aussi la définition suivante :

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right). \tag{2}$$

Exemples de noyaux en dimension 1

- Le noyau fenêtre $K(x) = (1/2)\mathbb{1}_{|x|\leq 1}$
- Le noyau triangulaire $K(x) = (1 - |x|)\mathbb{1}_{|x|\leq 1}$
- Le noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
- Le noyau parabolique $K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{|x|\leq 1}$

Propriétés des estimateurs à noyau.

Theorem

On se place dans un modèle de régression où les X_i sont aléatoires, de loi uniforme sur $[0,1]$, et on considère l'estimateur défini par

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

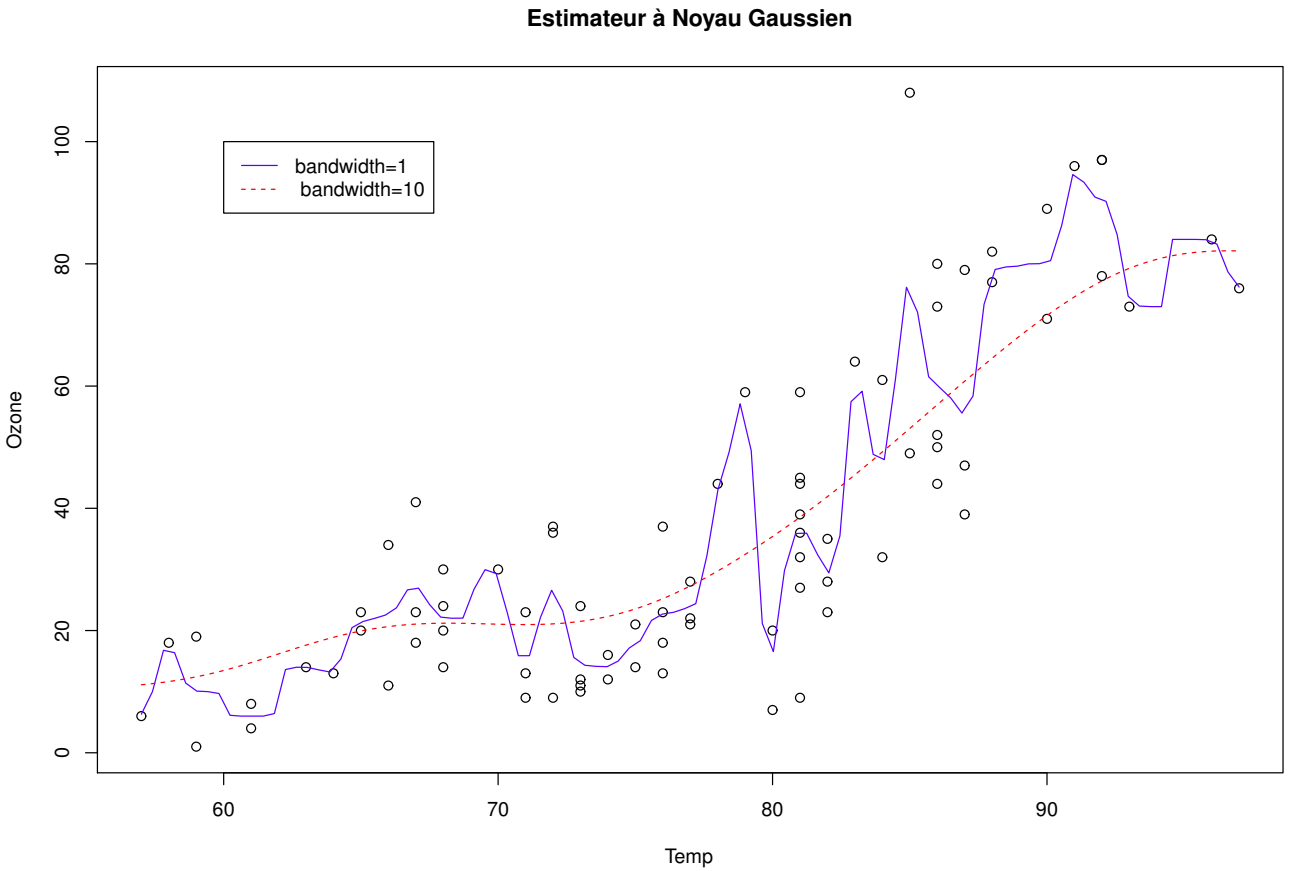
On suppose que $f \in \Sigma(\beta, R)$ définie par

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

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

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

Estimateur à noyau



Plan de l'exposé

- Estimation non paramétrique en régression réelle.
  - Polynômes par morceaux
  - Splines
  - Estimateurs à noyaux
  - Estimateurs par projection
- Modèles additifs

Estimateurs par projection

On se place dans le modèle de régression

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

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 $\mathbb{L}^2([0,1])$:

$$f_D = \sum_{j=1}^{D} \theta_j \phi_j, \quad \theta_j = \langle f, \phi_j \rangle = \int_0^1 f(x)\phi_j(x)dx.$$

On définit

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

Estimateurs par projection

- 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

$$\hat{\theta}_j = \theta_j.$$

- On introduit alors l'estimateur de f :

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

appelé *estimateur par projection*.

Exemple de la base Fourier

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

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

Performances de l'estimateur

On introduit une classe de fonctions périodiques, régulières.

Definition

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

$$\Sigma^{per}(\beta, R) = \Big\{ f \in \mathcal{C}^l([0,1]), \forall j = 0, \dots, l, \quad f^{(j)}(0) = f^{(j)}(1),$$

$$\forall x, y \in [0,1], |f^{(l)}(x) - f^{(l)}(y)| \leq R|x - y|^{\alpha} \Big\}.$$

Performances de l'estimateur

Theorem

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

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

Bases d'ondelettes et estimation par seuillage

- Les ondelettes sont bien adaptées pour l'estimation de fonctions présentant des irrégularités (pics) dans certaines parties de l'espace.
- Elles sont très utilisées en traitement du signal et de l'image (notamment pour le débruitage et la compression).
- Les bases d'ondelettes sont des bases orthonormées.
  Pour simplifier, nous nous plaçons sur $[0, 1]$.

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

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

Bases d'ondelettes

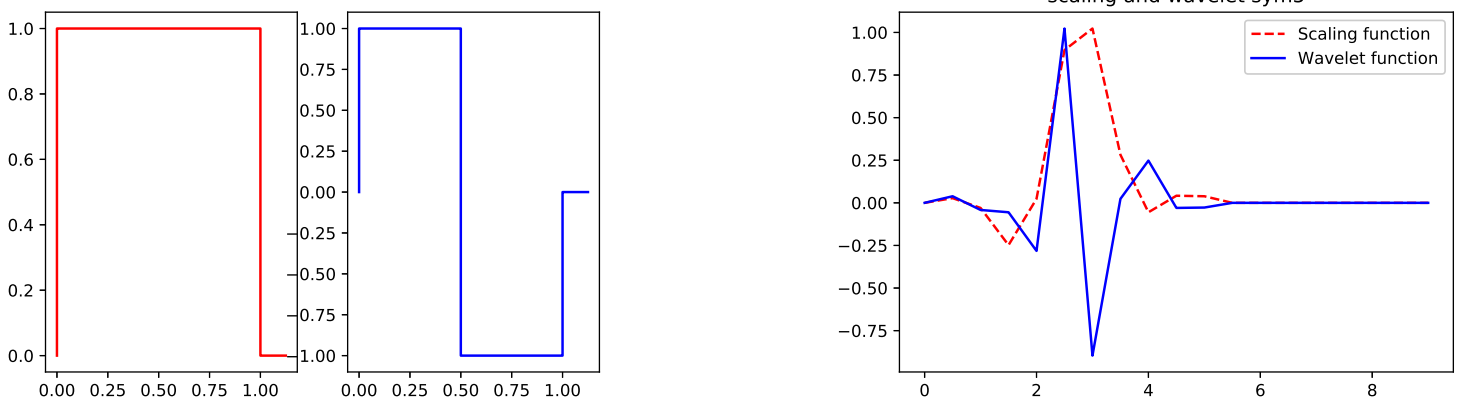


FIGURE: Ondelettes de Haar à gauche , à droite : Symlet 5

Base de Haar

Theorem

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

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"
- les $\beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)dx$ sont appelés "coefficient d'ondelette" ou "détails".

Bases d'ondelettes

$$
\begin{aligned}
f(x) &= \alpha\phi + \sum_{j\geq 0}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k} \\
&= \alpha\phi + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k} + \sum_{j\geq J_0}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}.
\end{aligned}
$$

On a la propriété suivante pour les ondelettes :

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

Bases d'ondelettes

Ainsi, pour tout $J_0 \geq 0$,

$$f(x) = \sum_{k=0}^{2^{J_0}-1} \alpha_{J_0,k}\phi_{J_0,k} + \sum_{j \geq J_0} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}.$$

$$f(x) = f_{J_0}(x) + \sum_{j \geq J_0} D_j(x).$$

- La fonction f_{J_0} est appelée "approximation", elle comporte 2^{J_0} coefficients ; les fonctions D_j sont appelées "détails".
- Les ondelettes sont des bases localisées : si les fonctions ϕ et ψ sont à support compact, plus j augmente, plus le support des fonctions $\psi_{j,k}$ est de petite taille.
- D'autres types de bases (Daubechies, symmlet ..) sont plus adaptées à la description de fonctions régulières.

Estimation d'une fonction de régression avec des ondelettes

- On observe un signal sur une grille dyadique, régulière :

$$Y_l = f(\frac{l}{N}) + \epsilon_l,\ l = 1, \ldots, N = 2^J,$$

- 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 f_J l'approximation de f au niveau J :

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

$$\hat{\alpha}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_{j,k}(X_i), \quad \hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{j,k}(X_i).$$

$$\hat{f}_J = \hat{\alpha}\phi + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}\psi_{j,k} = \sum_{k=0}^{2^J-1} \hat{\alpha}_{J,k}\phi_{J,k}.$$

Débruitage par approximation linéaire

- La fonction f_J comporte autant de paramètre à estimer que d'observations $\Longrightarrow$ surajustement !
- On se donne $J_0 \leq J$. On considère f_{J_0} l'approximation de f au niveau J_0 : elle comporte 2^{J_0} coefficients.

$$\begin{aligned}\hat{f}_{J_0}(x) &= \hat{\alpha}\phi(x) + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\hat{\beta}_{j,k}\psi_{j,k}(x) = \sum_{k=0}^{2^{J_0}-1}\hat{\alpha}_{J_0,k}\phi_{J_0,k}(x) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{2^{J_0}-1} Y_i\phi_{J_0,k}(X_i)\phi_{J_0,k}(x)\end{aligned}$$

- J_0 doit être optimisé pour un bon compromis biais/variance.

Débruitage par seuillage

- Nous allons considérer des approximations non linéaires.
- Au lieu de garder les 2^{J_0} premiers coefficients d'ondelette, l'idée est de garder uniquement les plus grands coefficients.
- Ceci permet de décrire des fonctions irrégulières avec seulement un petit nombre de coefficients non nuls (approximations parcimonieuses).
- On réduit ainsi la variance des estimateurs.
- C'est utile également pour la compression.

Débruitage par seuillage

- Deux types de seuillage
  - Seuillage dur (Hard Thresholding)

$$\hat{\theta}_i^H(\lambda) = \hat{\theta}_i \mathbb{1}_{|\hat{\theta}_i| \geq \lambda}$$

  - Seuillage doux (Soft Thresholding)

$$\hat{\theta}_i^S(\lambda) = \text{signe}(\hat{\theta}_i)(|\hat{\theta}_i| - \lambda)\mathbb{1}_{|\hat{\theta}_i| \geq \lambda}.$$

Débruitage par seuillage doux

- La fonction de régression f est estimée par

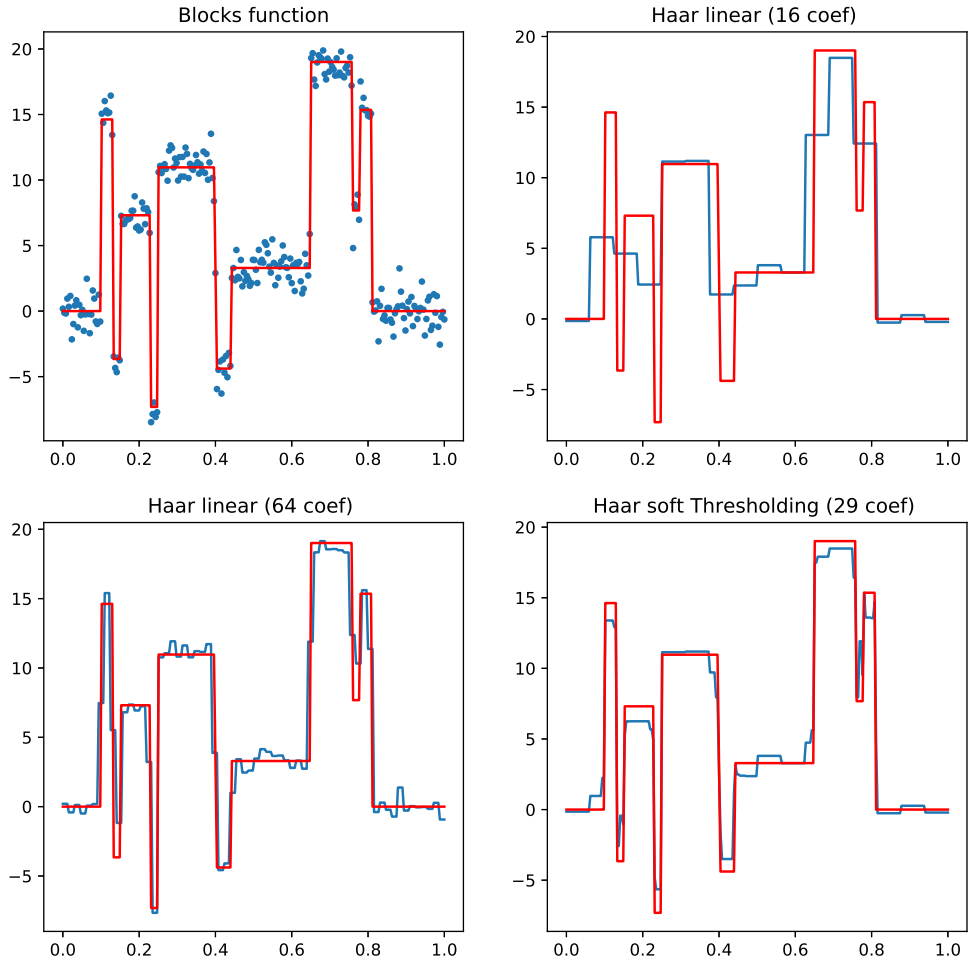
$$\hat{f}_\lambda^S(x) = \hat{\alpha}\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}^S(\lambda)\psi_{j,k}(x)$$

- Généralement, on ne seuille pas les premiers niveaux : on se donne $J_1 < J$

$$\hat{f}_\lambda^S(x) = \sum_{k=0}^{2^{J_1}-1} \hat{\alpha}_{J_1,k}\phi_{J_1,k}(x) + \sum_{j=J_1}^{J-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}^S(\lambda)\psi_{j,k}(x)$$

- En théorie, on doit choisir λ de l'ordre de $\sigma\sqrt{2\log(N)}$.
- Les coefficients qui sont inférieurs à $\sigma\sqrt{2\log(N)}$ sont considérés comme du bruit et sont annulés.

Débruitage par ondelettes



Plan de l'exposé

- Estimation non paramétrique en régression réelle.
  - Polynômes par morceaux
  - Splines
  - Estimateurs à noyaux
  - Estimateurs par projection
- Modèles additifs

Modèles additifs généralisés

- Les procédures d'estimations précédentes se heurtent au fléau de la dimension.
- Sous certaines hypothèses sur la structure de la fonction f, on peut contourner ce problème.
- C'est le cas par exemple si la dépendance de f en les différentes variables est additive.
- On considère le modèle

$$f(X_{i,1}, \ldots, X_{i,d}) = \alpha + f_1(X_{i,1}) + \ldots + f_d(X_{i,d})$$

  appelé modèle additif généralisé (GAM).
- Pour assurer l'unicité d'une telle écriture, on impose que

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

Modèles additifs généralisés

- Chacune des fonctions unidimensionnelles est estimée à l'aide de Splines cubiques, par exemple.
- 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 $\hat{f}_j$ est un spline cubique de la variable X_j, dont les noeuds correspondent aux valeurs différentes des $X_{i,j}, i = 1, \ldots n$.

Modèles additifs généralisés

- 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

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

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

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

Modèles additifs généralisés

- Le même algorithme peut être utilisé 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 oeuvre, et il sera utile de les combiner à un algorithme de sélection (pour réduire la dimension).

Références

- The elements of Statistical Learning by T. Hastie et al (2009).
- Introduction to nonparametric statistics (2009) by A. Tsybakov (2009)
- Introduction to High-Dimensional Statistics by C. Giraud (2015)