App. statistique : méthodes non linéaires pour la régression
Modèles additifs

C. HELBERT

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs, prédicteurs)
- Y est la variable à expliquer quantitative.

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs, prédicteurs)
- Y est la variable à expliquer quantitative.

Modèles présentés dans la suite du cours :

- alternatives aux méthodes de régression (linéaires)
- hypothèses sous-jacentes à ces modèles permettent d'aborder la grande dimension $p >> n$

Plan

- Un inconvénient de la régression est de devoir spécifier à l'avance la base de projection : souvent linéaire en les prédicteurs, ou avec des interactions, éventuellement des termes quadratiques ...
- Quand p est grand, on se contente souvent de quantifier l'éffet linéaire de chaque variable explicative. Le problème de sélection (screening) étant déjà conséquent.

Dans le cas additif, l'hypothèse est la suivante :

$$E(Y|X) = \alpha + f_1(X_1) + ... + f_p(X_p)$$

Avantages :

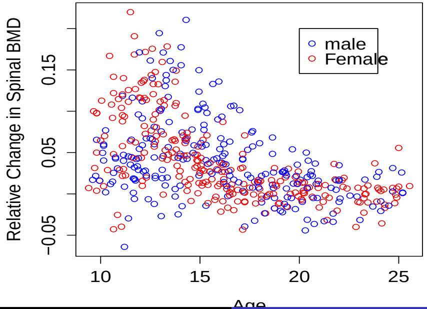
- les fonctions $f_1, ..., f_p$ sont **non spécifiées**.
- estimation non paramétrique (ex : smoothing splines, méthodes à noyaux ...)
- algorithme efficace d'estimation simultanée des p fonctions (même si $p > n$)

Plan

Exemple

Bone Mineral Density : 259 filles, 226 garçons, variables : âge et spnbmd (vitesse de densification osseuse).

Objectif : trouver les fonctions régulières (lisses) f_{male} et f_{female} telles que $BMD_{mle} = f_{male}(age) + \epsilon_{male}$ et $BMD_{female} = f_{female}(age) + \epsilon_{female}$.



Le problème est le suivant : trouver la fonction $f \in \mathcal{C}^2$ (deux dérivées continues) qui minimise le critère suivant :

$$RSS(f,\lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f''(t)^2 dt$$

où λ est un paramètre de lissage.

- le 1er terme : ajustement aux observations
- 2ème terme : pénalisation de la courbure de la fonction
- λ : compromis entre les deux termes, $\lambda \in [0,\infty[$
  - $\lambda = 0$: f interpole les données
  - $\lambda = \infty$: droite aux moindres carrés (dérivée seconde nulle)

Le problème admet une unique fonction optimale : la spline cubique naturelle aux noeuds $x_1, ...x_n$, i.e. une spline cubique (combinaison linéaire de fonction de degré 3 par morceaux et $\mathcal{C}^2$) linéaire avant x_1 et après x_n, i.e.

$$f(X) = \sum_{j=1}^{n} N_j(X)\theta_j$$

où les $N_1(X), ..., N_n(X)$ constituent la base de fonction des splines cubiques aux noeuds $x_1, ...x_n$.

Smoothing Splines

Les fonctions $N_1(X), ..., N_n(X)$ sont les fonctions suivantes :

- $N_1(X) = 1$
- $N_2(X) = X$
- $N_{k+2}(X) = d_k(X) - d_{n-1}(X)$ où $d_k(X) = \frac{(X-x_k)_+^3 - (X-x_n)_+^3}{x_n - x_k}$

Pour trouver les coefficients, on maximise le critère suivant :

$$RSS(f, \lambda) = (\mathbf{y} - \mathbf{N}\theta)^T(\mathbf{y} - \mathbf{N}\theta) + \lambda\theta^T\Omega_n\theta$$

où $\{\mathbf{N}\}_{ij} = N_j(x_i)$ et $\{\Omega_n\}_{jk} = \int N_j''(t)N_k''(t)dt$.

On obtient donc

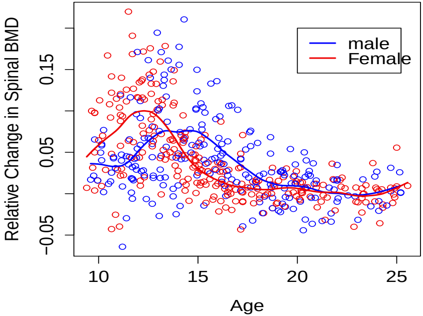
$$\hat{f}(x) = \sum_{j=1}^{n} N_j(x)\hat{\theta}_j$$

où $\hat{\theta} = (\mathbf{N}^T\mathbf{N} + \lambda\Omega_n)^{-1}\mathbf{N}^T\mathbf{y}$.

Retour sur l'exemple

Bone Mineral Density : 259 filles, 226 garçons, variables : âge et spnbmd (vitesse de densification osseuse).

Objectif : trouver les fonctions régulières (lisses) f_{male} et f_{female} .



Plan

Dans le cas additif, l'hypothèse est la suivante :

$$Y = \alpha + f_1(X_1) + ... + f_p(X_p) + \epsilon$$

On se place dans le contexte des splines de lissage pour chaque fonction f_j. Le problème est alors de minimiser la quantité suivante :

$$PRSS(\alpha, f_1, ..., f_p) = \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt$$

On peut montrer que le "minimiseur" est un modèle de splines cubiques adaptatives, i.e. chaque fonction f_j est une spline cubique en la variable X_j aux noeuds $x_{1j}, ..., x_{nj}$.

Attention : le problème n'est pas identifiable, la constante α n'est pas identifiable. Par convention, on ajoute p contraintes d'identificabilité $\sum_{i=1}^{n} f_j(x_{ij}) = 0$.

L'algorithme "Backfitting" :

1 Initialisation : $\hat{\alpha} = \frac{1}{n} \sum_1^n y_i, \hat{f}_j \equiv 0, \forall i, j$
2 Cycle : $j = 1, 2, ..., p, ..., 1, 2, ...p, ...$

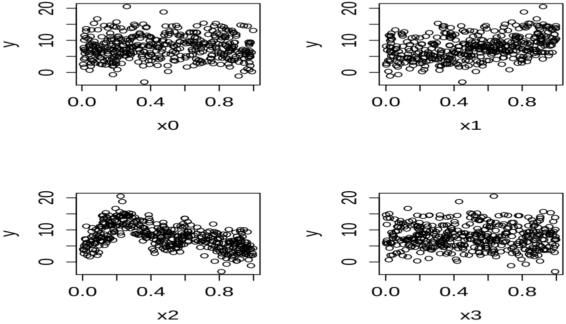
- $\hat{f}_j \leftarrow \mathcal{S}_j[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^n]$
- $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n} \sum_{i=1}^n \hat{f}_j(x_{ij})$

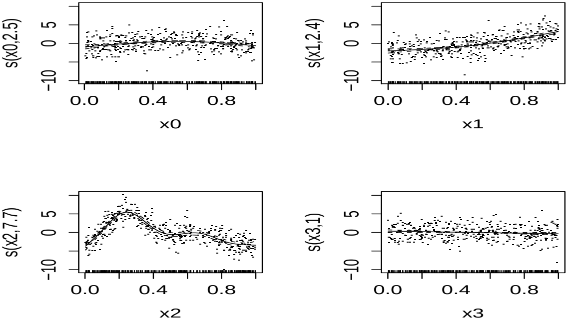
jusqu'à ce que les fonctions changent moins qu'un certain seuil.

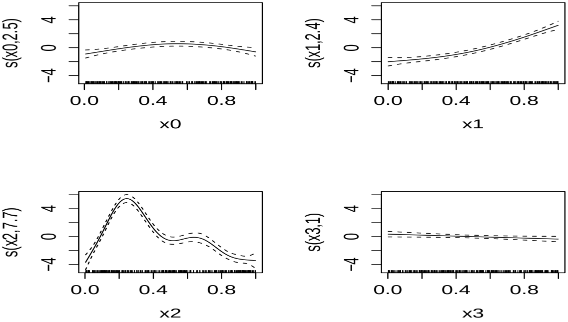
Modèle additif : conslusion

Ces modèles ajustent une fonction dans chaque direction. Si p est vraiment grand, on perd du temps (et des degrés de libertés) pour estimer des fonctions non significatives. Il y a donc des techniques (COSSO "COmponent Selection and Smoothing Operator" ou SpAM "Sparse Additive Models" ou ...) qui continuent à être développées pour faire l'estimation non linéaire et la sélection conjointement.

Plan


y
20
10
0
0.0
0.4
0.8
x0
x1
x2
x3


s(x0,2.5)
−10
0
5
0.0
0.4
0.8
x0
s(x1,2.4)
−10
0
5
0.0
0.4
0.8
x1
s(x2,7.7)
−10
0
5
0.0
0.4
0.8
x2
s(x3,1)
−10
0
5
0.0
0.4
0.8
x3


s(x0,2.5)
4
0
-4
0.0
0.4
0.8
x0
s(x1,2.4)
x1
s(x2,7.7)
x2
s(x3,1)
x3