Modèles Additifs Parcimonieux

THÈSE

présentée pour l'obtention du grade de

Docteur de l'Université de Technologie de Compiègne – UTC
(spécialité Technologies de l'Information et des Systèmes)

par

Marta Avalos Fernández

Composition du jury

Rapporteurs :	Christian Jutten	Professeur, Université Joseph Fourier, Grenoble
	Florence d'Alché-Buc	Professeur, Université d'Evry, Val d'Essonne
Examinateurs :	Stéphane Canu	Professeur, INSA, Rouen
	Georges Oppenheim	Professeur, Université de Paris Sud, Orsay
	Christophe Ambroise	Maître de conférences, UTC, Compiègne
	Yves Grandvalet	Chargé de recherche C.N.R.S., UTC, Compiègne

HEUristique et DIAgnostic des SYstèmes Complexes — UMR CNRS 6599

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Science does not aim at simplicity ; it aims at parsimony.
K. Popper

Dans les sciences expérimentales, on utilise des modèles mathématiques pour représenter les phénomènes observés. Une simplicité excessive du modèle nuit à sa capacité à rendre compte de la réalité. Une trop grande complexité dans la modélisation rend difficile la compréhension du phénomène. Un modèle parcimonieux tend à une économie des moyens, afin d'obtenir un bon compromis entre simplicité et fidélité.

Notre quête de parcimonie est en partie motivée par la recherche de l'interprétabilité. Dans le cadre de l'apprentissage supervisé, la généralisation est la mesure habituelle de la performance. Cependant, dans certaines applications, un modèle de type "boîte noire" ne sera pas accepté par l'utilisateur final : seule une méthode explicitant la prédiction sera utilisable. L'interprétabilité du prédicteur est également nécessaire dans les études exploratoires, où l'objectif de l'apprentissage consiste à inspirer de nouvelles idées sur les relations entre les variables et à améliorer, ainsi, la compréhension du domaine.

La modélisation additive suppose que la somme des effets non linéaires des variables d'entrée explique la variable de sortie. L'effet de chaque variable est estimé par un fonction monovariée non paramétrique. La modélisation non paramétrique assure de la flexibilité au modèle dont la structure simple permet de représenter l'effet de chaque variable, ce qui permet l'interprétation des solutions.

L'estimation des modèles additifs nécessite définir au préalable le degré de flexibilité des fonctions monovariées, ce qui nécessite le réglage d'autant de paramètres de contrôle qu'il y a de variables d'entrée. Ce pré–requis n'est pas réaliste même pour une dimension modérée des entrées. L'application des modèles additifs se limite donc à des problèmes avec peu de variables, généralement sélectionnées par une étude préalable.

Sélection de modèle pour les modèles additifs

Le présent mémoire est consacré au développement d'un algorithme d'estimation des modèles additifs. Le réglage des hyper–paramètres contrôlant la complexité est, pour une bonne part, intégré dans la procédure d'estimation, ce qui simplifie considérablement l'utilisation des modèles additifs pour les problèmes de dimension supérieure ou égale à trois.

De plus, un terme de pénalisation adapté favorise les solutions parcimonieuses éliminant une partie de l'ensemble des variables d'entrée, tout en permettant une modélisation flexible de la dépendance sur les variables sélectionnées.

Plan du document

Le document est structuré en quatre parties. Le premier chapitre situe la régression par modèles additifs dans le cadre de la régression non paramétrique multidimensionnelle. Nous développons plus particulièrement la question de l'estimation des fonctions monovariées quand leur complexité est fixée, et nous justifions nos choix parmi les techniques existantes.

Le deuxième chapitre traite du problème du contrôle de la complexité pour les modèles additifs. La difficulté du problème quand le nombre de variables est modéré ou élevé est mise en évidence.

Nous introduisons notre approche dans le troisième chapitre, précédée des méthodes de pénalisation pour les modèles linéaires qui l'ont motivée. L'algorithme qui nous permet de calculer effectivement les solutions y est détaillé. La dernière partie de ce chapitre est consacrée à la question de la sélection de modèle.

Le quatrième chapitre traite quant à lui de la mise en œuvre de la méthode. La première partie définit les bases d'un benchmark pour les modèles additifs, ce qui nous permet, dans la deuxième partie, d'évaluer expérimentalement la performance des méthodes développées. Finalement, nous montrons un exemple d'application des modèles additifs parcimonieux sur deux jeux de données réelles.

Contributions au domaine

Nos contributions au domaine portent tout d'abord sur l'introduction d'une méthode permettant de réduire le modèle quand la complexité n'est pas adaptée. Il est ainsi possible d'identifier les variables à éliminer, les variables à effets linéaires et les variables à effets non linéaires.

Notre stratégie se base sur une décomposition des espaces de fonctions splines, comprenant, d'une part, les fonctions linéaires et, d'autre part, les fonctions strictement non linéaires.

La sélection d'un modèle de complexité adaptée est une étape clé pour les modèles d'apprentissage statistique. Cette étape est particulièrement difficile à mettre en œuvre pour les modèles additifs, car la complexité du modèle se définit au travers de celles des estimateurs monovariées. Ainsi, cette complexité est indexée par un vecteur de la dimension du nombre de variables d'entrée. Dans notre approche, le réglage de la complexité du modèle ne nécessite que deux paramètres, ce qui simplifie la mise en œuvre dès que le nombre de variables est supérieur à deux.

Nous proposons également un estimateur du nombre de degrés de liberté (ou nombre effectif de paramètres), mesure de la complexité du modèle. Celui–ci raffine, en ce qui concerne la contribution linéaire des composantes, les estimateurs existants. Quant à la contribution non linéaire, la décomposition des matrices permet son calcul aisé. Cette mesure de la complexité permet l'adaptation de critères analytiques pour

la sélection des deux paramètres de réglage de la complexité des modèles additifs parcimonieux.

Les performances des méthodes développées sont montrées expérimentalement. Un benchmark, spécifiquement conçu pour les modèles additifs, permet d'évaluer les méthodes là où le problème du contrôle de la complexité est particulièrement délicat.

Enfin, nous appliquons les méthodes proposées dans ce mémoire à la prédiction individualisée du risque cardio–vasculaire chez des patients présentant une hypertension artérielle, application qui s'inscrit dans le cadre du projet INDANA (Individual Data Analysis of Antihypertensive Intervention Trials).

Liste des abréviations et notations

Abréviations

v.a.i.i.d.	variables aléatoires indépendantes identiquement distribuées
i.i.d.	(variables aléatoires) indépendantes identiquement distribuées
ddl	nombre de degrés de liberté ou nombre effectif de paramètres
RSS	(*Residual Sum of Squares*), somme des carrés résiduels
MSE	(*Mean Squared Error*), espérance de l'erreur quadratique
MASE	(*Mean Average Squared Error*), espérance de l'erreur quadratique moyenne
ASE	(*Average Squared Error*), erreur quadratique moyenne
ISE	(*Integrated Squared Error*), intégrale de l'erreur quadratique
MISE	(*Mean Integrated Squared Error*), intégrale de l'erreur quadratique moyenne
APE	(*Average Predictive Error*), espérance de l'erreur quadratique de prédiction
PE	(*Predictive Error*), erreur de prédiction
EMT	erreur moyenne de test
AIC	(*Akaike Information Criteria*), critère d'information d'Akaike
BIC	(*Bayesian Information Criteria*), critère d'information bayésien
GCV	(*Generalized Cross Validation*), validation croisée généralisée
CV	(*Cross Validation*), validation croisée
OLS	(*Ordinary Least Squares*), moindres carrés ordinaires
RR	(*Ridge Regression*), pénalisation quadratique
AdR	(*Adaptive Ridge Regression*), pénalisation multiple adaptative

Scalaires–vecteurs–matrices

x	point de $\mathbb{R}$
$\mathbf{x} = (x_1, \ldots, x_d)^t$	point de $\mathbb{R}^d$
$\mathbf{e}_1$	$(1, 0, \ldots, 0)^t$
$\mathbf{1} = \mathbf{1}_n$	$(1, \ldots, 1)^t$
$\mathbf{I}_p$	matrice identité de dimension p

Probabilités

$\mathbb{E}_X[\cdot]$	espérance sur la v. a. X
$\mathrm{Var}_X[\cdot]$	variance sur la v. a. X
$\mathrm{Cov}_{(X,Y)}[\cdot,\cdot]$	covariance sur les v. a. X et Y
se	écart–type
h_X	densité de la variable aléatoire X

Echantillons

X	variable aléatoire
x	scalaire, réalisation de X
p	dimension des entrées, indexée normalement par j
$\boldsymbol{X} = (X_1, \ldots, X_p)$	vecteur de variables aléatoires
$\boldsymbol{x} = (x_1, \ldots, x_p)^t$	vecteur, réalisation de $\boldsymbol{X}$
Y	variable aléatoire
y	scalaire réalisation de Y
n	taille de l'échantillon, indexée normalement par i
$\{(X_{i1}, Y_i)\}_{i=1}^n$	échantillon i.i.d. de (X, Y) de taille n
$\{(x_{i1}, y_i)\}_{i=1}^n$	scalaires, réalisations de $\{(X_{i1}, Y_i)\}_{i=1}^n$
$\mathbf{x}_1 = (x_{11}, \ldots, x_{n1})^t$	vecteur $n \times 1$ des réalisations de $\{(x_{i1}, y_i)\}_{i=1}^n$
$\mathbf{y} = (y_1, \ldots, y_n)^t$	vecteur $n \times 1$ des réalisations de Y
$\{(X_{i1}, \ldots, X_{ip}, Y_i)\}_{i=1}^n$	échantillon i.i.d. de $(\boldsymbol{X}, Y)$ de taille n
$\{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^n$	scalaires, réalisations de $\{(X_{i1}, \ldots, X_{ip}, Y_i)\}_{i=1}^n$
$\mathbf{X} = \{x_{ij}\}_{i=1,\ldots,n;j=1,\ldots,p}$	matrice $n \times p$ des données
$\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^t$	vecteur $n \times 1$ des réalisations de $\{X_{ij}\}_i$, j–ème colonne de $\mathbf{X}$
$\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^t$	vecteur $p \times 1$ correspondant à la i–ème observation
$(\mathbf{x}^i)^t$	vecteur $1 \times p$, i–ème ligne de $\mathbf{X}$

Fonctions

$\mathbb{I}_T$	fonction indicatrice sur la partie T
$f^{(k)}$	dérivée d'ordre k de la fonction f
$(f)_+$	fonction partie positive de f (vaut f si $f > 0$, 0 sinon)
$x_{(1)}$	$\min\{x_1, \ldots, x_n\}$
$x_{(n)}$	$\max\{x_1, \ldots, x_n\}$
$x_{(1),j}$	$\min_i(\{x_{ij}\})$,
$x_{(n),j}$	$\max_i(\{x_{ij}\})$,
$\mathrm{D}f(\mathbf{x})$	différentielle de f en $\mathbf{x}$
$\langle \mathbf{x}, \mathbf{x}' \rangle$	produit scalaire des vecteurs $\mathbf{x}$ et $\mathbf{x}'$

Fonction de régression

$f(x) = \mathbb{E}[Y\|X = x]$	fonction de régression (cas unidimensionnel)
$\widehat{f}(x)$	estimation de f évaluée en x
$\widehat{\mathbf{f}}$	$\widehat{f}(\mathbf{x}_1) = \widehat{f}((x_{11}, \ldots, x_{n1})^t)$ (cas unidimensionnel)
$f(\boldsymbol{x}) = \mathbb{E}[Y\|\boldsymbol{X} = \boldsymbol{x}]$	fonction de régression (cas multidimensionnel)
$\widehat{f}(\boldsymbol{x})$	estimation de f évaluée en $\boldsymbol{x}$
$\widehat{\mathbf{f}}_j$	$\widehat{f}_j(\mathbf{x}_j) = \widehat{f}_j((x_{1j}, \ldots, x_{nj})^t)$
$\widehat{\mathbf{f}}$	$\widehat{\alpha} + \widehat{\mathbf{f}}_1 + \ldots + \widehat{\mathbf{f}}_p$ (cas multidimensionnel)

Paramètres

λ, μ	paramètres unidimensionnels qui règlent la complexité
$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$	paramètres multidimensionnels qui règlent la complexité
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$	

Ensembles

$[a, b]$	intervalle fermé, a, $b \in \mathbb{R}$
$]a, b[$	intervalle ouvert
$\mathcal{C}^2[a, b]$	$\{f\|f$ deux fois dérivable avec continuité sur $[a, b]\}$
$[\mathbf{x}_1, \ldots, \mathbf{x}_p]$	espace vectoriel généré par les vecteurs $\mathbf{x}_j \in \mathbb{R}^n$, $n > p$
$E_1 \oplus E_2$	somme directe des des sous-espaces vectoriels E_1, E_2
$\mathbb{M}_1(\mathbf{S})$	espace généré par les vecteurs propres de $\mathbf{S}$ de valeur propre 1

Chapitre 1

Modèles additifs

1.1 Introduction

L'approche classique d'estimation d'une fonction de régression consiste à supposer que la structure de la fonction est connue, dépendante de certains paramètres, et inclue dans un espace de fonctions de dimension finie. C'est l'approche paramétrique, dans laquelle les données sont utilisées pour estimer les valeurs inconnues de ces paramètres.

Le modèle de régression linéaire en est le paradigme. Dans le contexte paramétrique, les estimateurs dépendent généralement de peu de paramètres, ainsi ces modèles sont appropriés même pour des petits échantillons. Ils sont facilement interprétables, par exemple, dans le cas linéaire, les valeurs des coefficients indiquent l'influence de la variable explicative sur la variable réponse, et leur signe décrit la nature de cette influence. Cependant, un estimateur linéaire conduira à une erreur élevée, quelle que soit la taille de l'échantillon, si la vraie fonction qui a générée les données n'est pas linéaire et ne peut pas être approchée convenablement par des fonctions linéaires.

L'approche non paramétrique, elle, ne suppose pas de structure pré–déterminée de la fonction de régression. La relation fonctionnelle entre les variables explicatives et la variable réponse est ajustée à partir des données. Cette flexibilité permet de capter des traits inusuels ou inattendus, en revanche, la complexité du problème d'estimation est beaucoup plus importante.

1.2 Régression non paramétrique unidimensionnelle

1.2.1 Méthodes de lissage

Il existe plusieurs méthodes pour obtenir un estimateur non paramétrique de la fonction f :

$$Y = f(X) + \varepsilon, \tag{1.1}$$

où (X, Y) vecteur aléatoire, ε variable aléatoire indépendante de X telle que $\mathbb{E}(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2$. Par exemple, la dépendance entre les variables Y et X peut être estimée par des techniques de lissage. L'ajustement par ce type de techniques donne comme résultat des estimations de tendance moins variables que Y elle–même [Hastie et Tibshirani, 1990].

Soit $\{(X_{i1}, Y_i)\}_{i=1}^n$ un échantillon i.i.d. de (X, Y) de taille n, et $\{(x_{i1}, y_i)\}_{i=1}^n$ des réalisations de l'échantillon. Notons $\mathbf{y} = (y_1, \ldots, y_n)^t$ et $\mathbf{x}_1 = (x_{11}, \ldots, x_{n1})^t$. Les méthodes de lissage estiment $f(x_{i1})$ par un moyennage pondéré des $\{y_i\}_{i=1}^n$ sur un voisinage de x_{i1} : $\widehat{f}_i = \sum_{k=1}^n w_{ik} y_k$, où $\widehat{f}_i$ indique l'estimation de f en x_{i1}. Les pondérations $w_{ik} = w(x_{i1}, x_{k1})$ sont élevées quand $|x_{i1} - x_{k1}|$ est faible et s'approchent de zéro quand $|x_{i1} - x_{k1}|$ devient élevé.

Les applications des méthodes de lissage sont nombreuses. Elles sont utilisées comme outil d'analyse exploratoire car ces méthodes permettent d'inspecter graphiquement la forme des relations. Cela facilite la construction d'un modèle, qui peut être paramétrique, quand aucune information sur la relation entre les variables explicatives et la variable réponse n'est fournie. Elles peuvent également être utilisées comme mesure de la qualité de l'ajustement et comme hypothèse dans les tests d'hypothèses des modèles paramétriques [Simonoff, 1996].

Une classe particulière des méthodes de lissage sont les méthodes de lissage linéaires. Quand $\{w_{ik}\}$ dépend des $\{x_{i1}\}_{i=1}^n$ mais elle ne dépend pas des $\{y_i\}_{i=1}^n$, la méthode est dite linéaire. Il existe alors une matrice $\mathbf{S} = \{S_{ik}\}_{i,k=1}^n$, indépendante de $\mathbf{y}$, telle que l'estimateur s'écrit : $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$. Cette matrice est la matrice de lissage. Les méthodes de lissage linéaires classiques sont les fonctions noyaux et des fonctions splines.

1.2.2 Estimateurs à noyau

Les méthodes de lissage par noyaux sont intuitives et simples du point de vue mathématique. Ces techniques utilisent un ensemble de pondérations locales, définies par le noyau, pour construire l'estimateur en chaque valeur. Le noyau est, en général, une fonction K continue, bornée, non négative, symétrique telle que :

$$\int K(t)dt = 1 \qquad \int tK(t)dt = 0 \qquad \int t^2 K(t)dt < \infty. \tag{1.2}$$

Le poids assigné au point x' pour l'estimation au point x est défini par :

$$K_\lambda(x, x') = \frac{1}{\lambda} K\left(\frac{|x - x'|}{\lambda}\right), \tag{1.3}$$

où $\lambda > 0$ est le paramètre de lissage, largeur de la fenêtre, largeur de bande, hyper–paramètre, ou qui règle la complexité.

Des exemples classiques de noyaux sont :

$$\begin{aligned} K(t) &= \tfrac{1}{\sqrt{2\pi}} exp(-t^2/2), \\ K(t) &= \tfrac{3}{4}(1-t^2)\mathbb{I}_{|t|\leq 1}, \\ K(t) &= \tfrac{3}{8}(3-5t^2)\mathbb{I}_{|t|\leq 1}, \end{aligned} \tag{1.4}$$

les noyau gaussien, le noyau d'Epanechnikov, et le noyaux de variance minimum, respectivement.

Des estimateurs à noyau sont l'estimateur de Nadaraya–Watson, l'estimateur de Priestley–Chao et l'estimateur de Gasser–Müller (ces deux derniers, pour des x fixées). Les polynômes locaux généralisent ces estimateurs à noyaux.

1.2.2.1 Estimateur de Nadaraya–Watson

L'idée est de faire une partition de l'ensemble des valeurs de X et faire alors un moyennage pondéré des valeurs de Y dans chaque sous–intervalle [Wand et Jones, 1995]. Les sous–intervalles sont construits comme des voisinages centrés en chaque point x. La pondération des moyennes est donnée par une fonction noyau. L'estimateur de Nadaraya–Watson est :

$$\widehat{f}_{NW}(x) = \frac{\sum_{i=1}^{n} y_i K_\lambda(x, x_{i1})}{\sum_{i=1}^{n} K_\lambda(x, x_{i1})}. \tag{1.5}$$

Matriciellement, à l'aide de la matrice de lissage, l'estimation de la courbe aux points x_{i1}, $i = 1, \ldots, n$ s'écrit :

$$\begin{pmatrix} \widehat{f}(x_{11}) \\ \vdots \\ \widehat{f}(x_{n1}) \end{pmatrix} = \begin{pmatrix} \frac{K_\lambda(x_{11}, x_{11})}{\sum_{k=1}^{n} K_\lambda(x_{11}, x_{k1})} & \cdots & \frac{K_\lambda(x_{11}, x_{n1})}{\sum_{k=1}^{n} K_\lambda(x_{11}, x_{k1})} \\ \vdots & & \vdots \\ \frac{K_\lambda(x_{n1}, x_{11})}{\sum_{k=1}^{n} K_\lambda(x_{n1}, x_{k1})} & \cdots & \frac{K_\lambda(x_{n1}, x_{n1})}{\sum_{k=1}^{n} K_\lambda(x_{n1}, x_{k1})} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{1.6}$$

Cette matrice de lissage a une valeur propre égale à 1 de vecteur propre $\mathbf{1}_n = (1, \ldots, 1)^t$. En effet, $\mathbf{S}\mathbf{1}_n = \mathbf{1}_n$, puisque la somme de chaque ligne de la matrice est 1. Toutes les autres valeurs propres sont plus petites que 1, strictement positives[1]. Par conséquent, l'estimateur de Nadaraya–Watson préserve les fonctions constantes et "amortit" le tracé des points observés, dans les autres cas (voir section 1.2.3.2, page 21, pour une étude plus détaillée de la matrice de lissage en fonction de ses valeurs et vecteurs propres).

[1]Dans certains cas, les valeurs propres peuvent être négatives.

1.2.2.2 Polynômes locaux

Les polynômes locaux généralisent les estimateurs de type noyau. Si $f(x)$ est q fois dérivable dans un voisinage de x, alors l'expansion de Taylor peut être appliquée :

$$\begin{aligned} f(x') \approx & \quad f(x) + f^{(1)}(x)\,(x'-x) + \ldots + \tfrac{f^{(q)}(x)}{q!}\,(x'-x)^q \\ = & \quad \beta_0 + \beta_1\,(x'-x) + \ldots + \beta_q\,(x'-x)^q\,, \end{aligned} \tag{1.7}$$

où $\beta_k = f^{(k)}(x)/k!$, et $f^{(k)}$ indique la dérivée d'ordre k. Nous pouvons donc considérer le problème de régression polynomiale locale dans un voisinage de x. La fonction de régression est estimée en chaque point en ajustant localement un polynôme de degré q par moindres carrés pondérés. La pondération au point i, $i = 1, \ldots, n$ est choisie en fonction de l'amplitude de la fonction noyau centrée en ce point. L'estimateur de la régression en chaque point x est le polynôme local qui minimise [Fan et Gijbels, 2000] :

$$\sum_{i=1}^{n} (y_i - \beta_0 - \ldots - \beta_q (x_{i1} - x)^q)^2 K_\lambda(x_{i1}, x). \tag{1.8}$$

Matriciellement, le problème à résoudre est le suivant :

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{N}(x)\boldsymbol{\beta})^t \mathbf{W}(x) (\mathbf{y} - \mathbf{N}(x)\boldsymbol{\beta}), \tag{1.9}$$

où $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)^t$, $\mathbf{W}(x) = \mathrm{diag}[K_\lambda(x_{11}, x), \ldots, K_\lambda(x_{n1}, x)]$, et

$$\mathbf{N}(x) = \begin{pmatrix} 1 & (x_{11} - x) & \ldots & (x_{11} - x)^q \\ \vdots & & & \vdots \\ 1 & (x_{n1} - x) & \ldots & (x_{n1} - x)^q \end{pmatrix}. \tag{1.10}$$

Le vecteur $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \ldots, \widehat{\beta}_q\right)^t$ qui minimise (1.8) et (1.9) est :

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{f}(x), \ldots, \widehat{f^{(q)}}(x)/q!\right)^t = \left(\mathbf{N}(x)^t \mathbf{W}(x) \mathbf{N}(x)\right)^{-1} \mathbf{N}(x)^t \mathbf{W}(x) \mathbf{y}. \tag{1.11}$$

L'expression explicite de l'estimateur de $f(x)$ est donc :

$$\begin{cases} \widehat{f}(x) = \mathbf{e}_1^t \widehat{\boldsymbol{\beta}} = \mathbf{s}(x)^t \mathbf{y}, \quad \mathbf{e}_1 = (1, 0, \ldots, 0)^t \\ \\ \mathbf{s}(x)^t = \mathbf{e}_1^t \left(\mathbf{N}(x)^t \mathbf{W}(x) \mathbf{N}(x)\right)^{-1} \mathbf{N}(x)^t \mathbf{W}(x). \end{cases} \tag{1.12}$$

La forme générale de la matrice de lissage, pour q quelconque, s'écrit de façon explicite :

$$\mathbf{S} \;=\; \begin{pmatrix} \mathbf{s}(x_{11})^t \\ \vdots \\ \mathbf{s}(x_{n1})^t \end{pmatrix}. \tag{1.13}$$

Cette matrice dépend du paramètre de lissage λ par l'intermédiaire de $\mathbf{W}(x)$.
L'estimateur de Nadaraya–Watson correspond au polynôme local de degré 0.

1.2.2.3 Interprétation du paramètre largeur de bande

Pour l'estimation de la courbe en un point x, le poids assigné à un point y_i quand $q = 0$ (1.6) est :

$$\frac{K_\lambda(x_{i1}, x)}{\sum_{k=1}^{n} K_\lambda(x_{k1}, x)}. \tag{1.14}$$

Les observations proches de x ont plus d'influence sur l'estimation de la régression au point x que celles qui en sont éloignées. L'influence relative est contrôlée par le paramètre largeur de bande [Wand et Jones, 1995]. Si λ est petit ($\lambda \to 0$), l'ajustement local est fortement dépendant des observations proches de x, et donne comme résultat une courbe très fluctuante qui tend à l'interpolation des données. Si λ est grand ($\lambda \to \infty$), les poids donnés aux observations proches et éloignées, tendent à être égaux, et donne comme résultat une courbe qui approche la régression linéaire globale.

Quand $q > 0$, si λ est grand, le résultat est une courbe qui approche la régression polynomiale de degré q globale [Fan et Gijbels, 2000]. Si $\lambda \to 0$, les voisinages déterminés par λ sont de moins en moins denses, jusqu'à la limite où le voisinage est constitué par un seul point. Puisque pour définir un polynôme de degré q il faut un minimum de $q + 1$ points, l'estimateur n'est pas défini pour les x tels que leur λ–voisinage est constitué par moins de $q + 1$ points.

La figure (1.1) montre des estimations par régression linéaire locale, pour trois valeurs différentes du paramètre de lissage. Pour une valeur très faible de λ, la fonction d'estimation est proche d'une fonction régulière qui interpole les points. Pour une valeur élevée de λ, la fonction d'estimation s'approche de l'estimation moindres carrés ordinaires.

1.2.3 Splines

1.2.3.1 Splines de régression

Les splines de régression représentent un compromis entre la régression polynomiale globale, et les méthodes de lissage précédentes locales. L'idée consiste à construire des polynômes par morceaux se raccordant de façon lisse [Hastie et Tibshirani, 1990]. Les points de raccord entre les morceaux de polynômes sont les nœuds.

Pour représenter des splines, pour un ensemble fixé de nœuds $\{\xi_k\}_{k=1,\ldots,K}$, il faut déterminer une base de fonctions. Par exemple, la base de polynômes tronqués de degré q, pour l'ensemble de nœuds $\{\xi_k\}_{k=1,\ldots,K}$, évaluée en x est :

$$\{N_j(x)\}_{j=1}^{K+q+1} = \{1, x, x^2, x^3, \ldots, x^q, (x-\xi_1)_+^q, \ldots, (x-\xi_K)_+^q\}, \tag{1.15}$$

où $(.)_+$ indique la fonction partie positive. La représentation de $f(x)$ dans cette base est donnée par $f(x) = \sum_{j=1}^{K+q+1} \beta_j N_j(x)$. Les coefficients β_j sont estimés en minimisant

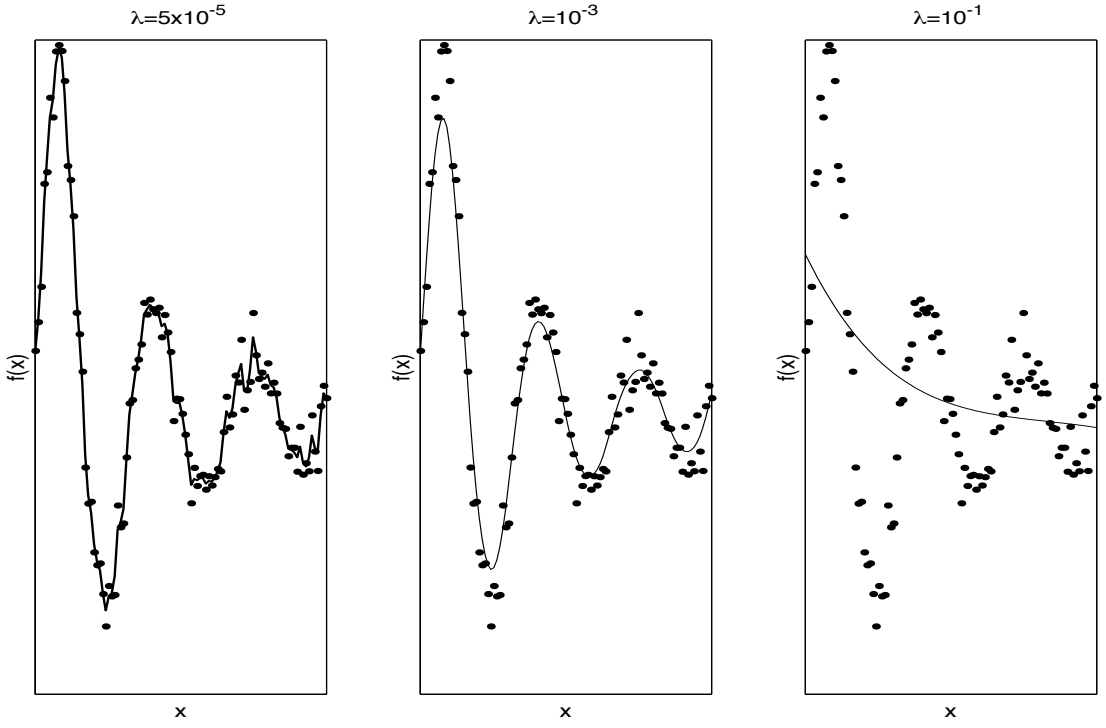

FIG. 1.1 – Estimation par des polynômes locaux de degré 1, à noyau gaussien, pour trois valeurs du paramètre de lissage : $\lambda = 5 \times 10^{-5}, 10^{-3}, 10^{-1}$. Les données sont générées par $y = \frac{\sin(6\pi x)}{(x+1)^2} + \varepsilon$, Les valeurs de x sont régulièrement espacées sur l'intervalle $[0, 1]$ ($n = 100$), et ε est une variable normale centrée d'écart–type $\sigma = 0.05$.

l'erreur quadratique :

$$\left\| \begin{pmatrix} 1 & \dots & x_{11}^q & (x_{11}-\xi_1)_+^q & \dots & (x_{11}-\xi_K)_+^q \\ \vdots & & & & & \vdots \\ 1 & \dots & x_{n1}^q & (x_{n1}-\xi_1)_+^q & \dots & (x_{n1}-\xi_K)_+^q \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{K+q+1} \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right\|^2 . \tag{1.16}$$

Une fonction spline classique est la fonction spline cubique : les polynômes par morceaux sont de degré 3 et ils sont contraints à avoir des dérivées de second ordre continues sur les nœuds. Si une autre restriction est imposée, linéarité au delà du domaine défini par les nœuds (dérivées d'ordre supérieur à 1 nulles), ces fonctions sont appelées splines cubiques naturelles. La condition naturelle de linéarité sur les bords implique l'expression suivante de la base naturelle des polynômes tronqués pour des splines cubiques [Hastie *et al.*, 2001] :

$$\begin{aligned} \{N_j(x)\}_{j=1}^K &= \{1, x, d_1(x) - d_{K-1}(x), \dots, d_{K-2}(x) - d_{K-1}(x)\}, \\ \text{où } d_k(x) &= \frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{\xi_k - \xi_K}. \end{aligned} \tag{1.17}$$

Quand la distribution des nœuds est loin d'être uniforme, des relations proches de la dépendance linéaire apparaissent entre les éléments de la base. Ceci se traduit par des systèmes linéaires mal conditionnés. D'autre part, comme les supports de ces fonctions ne sont pas bornés, l'évaluation sur des nouveaux points demande l'évaluation de presque toutes les fonctions de la base. Il est donc souhaitable l'utilisation d'une base où les supports des éléments de la base sont le plus petit possible [de Boor, 2001, Gyorfi *et al.*, 2002].

La base B–splines, en termes de quotients de différences, est plus adéquate pour effectuer des calculs. Les fonctions de cette base sont obtenues par des combinaisons linéaires des fonctions de la base de polynômes tronqués. Ce sont des fonctions à support local, ce qui implique que les matrices correspondantes sont des matrices bande. Cette base est constituée par $K+2$ fonctions, les B–splines ne sont pas des splines naturelles, elles ont des restrictions différentes sur les bords. Notons $\xi = \{\xi_j\}_{j=-3,\ldots,K+3}$ l'ensemble des nœuds, alors la fonction B–spline $B_{j,l,\xi}$ de degré l est définie de façon récursive [de Boor, 2001, Gu, 2002, Gyorfi *et al.*, 2002] :

$$B_{j,0,\xi} = \begin{cases} 1 & \text{si } \xi_j \leq x < \xi_{j+1}, \\ 0 & \text{autrement,} \end{cases} \tag{1.18}$$

où $j = -3, \ldots, K+2$, et

$$B_{j,l+1,\xi}(x) = \frac{x - \xi_j}{\xi_{j+l+1} - \xi_j} B_{j,l,\xi}(x) + \frac{-x + \xi_{j+l+2}}{\xi_{j+l+2} - \xi_{j+1}} B_{j+1,l,\xi}(x), \tag{1.19}$$

où $j = -3, \ldots, K+l-1$, $l = 0, \ldots, 2$. Dans cette formulation, $\xi_{j+l+1} - \xi_j = 0$ (ou $\xi_{j+l+2} - \xi_{j+1} = 0$) implique $B_{j,l,\xi}(x) = 0$ (ou $B_{j+1,l,\xi}(x) = 0$), par la convention $0/0 = 0$.

Le nombre et la position des nœuds sont les paramètres qui déterminent la complexité. En générale, les positions sont considérées fixées, par exemple les nœuds sont placés uniformément ou aux percentiles de la variable explicative [Hastie et Tibshirani, 1990].

1.2.3.2 Splines de lissage

L'estimateur par splines cubiques émerge également d'un problème d'optimisation, la minimisation de la somme des carrés résiduels pénalisés [Wahba, 1990, Hastie et Tibshirani, 1990] :

$$\min_f \sum_{i=1}^{n} (y_i - f(x_{i1}))^2 + \lambda \int \left[f^{(2)}(t)\right]^2 dt, \tag{1.20}$$

où $f \in \mathcal{C}^2[x_{(1)1}, x_{(n)1}]$, $x_{(1)1} = \min\{x_{11}, \ldots, x_{n1}\}$, $x_{(n)1} = \max\{x_{11}, \ldots, x_{n1}\}$, et $\mathcal{C}^2[x_{(1)1}, x_{(n)1}] = \{f | f$ deux fois dérivable avec continuité sur $[x_{(1)1}, x_{(n)1}]\}$.

Le premier terme mesure la fidélité aux données, alors que le deuxième terme pénalise les grandes fluctuations de la fonction. Le paramètre de lissage, ou hyper–paramètre ou paramètre de la complexité λ contrôle le compromis entre les deux termes. Il existe une solution unique à ce problème : la fonction qui minimise

l'équation (1.20) est une spline cubique avec des nœuds aux valeurs de $\{x_{i1}\}_{i=1}^n$, appelée spline cubique de lissage.

Soient $\mathbf{N}$ la matrice de la base évaluée en $\{x_{i1}\}_{i=1}^n$, $N_j^{(2)}(x)$ la dérivée seconde du j–ème élement de la base évaluée en x, et $\mathbf{\Omega}$ la matrice correspondant à la pénalisation de la dérivée seconde :

$$\Omega_{ij} = \int N_i^{(2)}(x) N_j^{(2)}(x) dx. \tag{1.21}$$

Pour un point x, la valeur de l'estimation est donnée par :

$$\widehat{f}(x) = \sum_j N_j(x)\widehat{\beta}_j. \tag{1.22}$$

Le vecteur de coefficients de $\widehat{f}$ sur la base fixée, $\widehat{\boldsymbol{\beta}}$, est la solution de :

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{N}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{N}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^t \mathbf{\Omega} \boldsymbol{\beta}. \tag{1.23}$$

L'expression explicite de $\widehat{\boldsymbol{\beta}}$ est :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{N}^t\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^t\mathbf{y}, \tag{1.24}$$

et donc,

$$\widehat{\mathbf{f}} = \mathbf{N}\widehat{\boldsymbol{\beta}} = \mathbf{N}(\mathbf{N}^t\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^t\mathbf{y}. \tag{1.25}$$

La matrice de lissage est donnée par $\mathbf{S} = \mathbf{N}(\mathbf{N}^t\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^t$, matrice symétrique, définie positive. Elle a 2 valeurs propres égales à 1, correspondantes aux fonctions propres contante et linéaire, et $n-2$ valeurs propres comprises strictement entre 0 et 1, correspondant aux fonctions propres d'ordre supérieur.

Considérons les valeurs propres de $\mathbf{S}$, en ordre décroissant : $\nu_1 \geq \ldots \geq \nu_n$, et les fonctions propres correspondant $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Nous avons la relation : $\mathbf{S}\mathbf{v}_i = \nu_i \mathbf{v}_i$, $i = 1, \ldots, n$. La représentation du vecteur réponse $\mathbf{y}$ en la base constituée par les fonctions propres est donnée par : $\mathbf{y} = \sum_{i=1}^n \gamma_i \mathbf{v}_i$, où $\{\gamma_i\}_i$ sont les coefficients associés à cette base. Le vecteur des estimations peut donc s'écrire en termes des valeurs et fonctions propres comme suit : $\widehat{\mathbf{y}} = \sum_{i=1}^n \gamma_i \mathbf{S}\mathbf{v}_i = \sum_{i=1}^n \gamma_i \nu_i \mathbf{v}_i$. La contribution des premières fonctions propres est très importante, car les valeurs propres correspondantes sont élevées, en revanche la contribution des fonctions d'ordre supérieur est plus faible, car les valeurs propres correspondantes sont proches de zéro [Ruppert *et al.*, 2003].

L'estimation par des splines cubiques préserve donc les fonctions constantes et linéaires : $\mathbf{S}\mathbf{1}_n = \mathbf{1}_n$, où $\mathbf{1}_n = (1, \ldots, 1)^t$, $\mathbf{S}\mathbf{x}_1 = \mathbf{x}_1$, où $\mathbf{x}_1 = (x_{11}, \ldots, x_{n1})^t$ et "amortit" (ou rétrécit) le tracé des points observés, dans les autres cas : $\|\mathbf{S}\mathbf{v}\| < \|\mathbf{v}\|$, pour une norme quelconque et pour $\mathbf{v}$ quelconque, n'appartenant pas au sous–espace généré par $\mathbf{1}_n$ et $\mathbf{x}_1$.

Les figures (1.2) et (1.3) montrent les valeurs propres et les fonctions propres des splines cubiques de lissage, respectivement. Les deux premières fonctions propres, correspondantes aux deux valeurs propres égales à 1, génèrent l'espace des fonctions linéaires et constantes. Les fonctions propres correspondantes aux plus petites valeurs propres deviennent plus oscillantes.

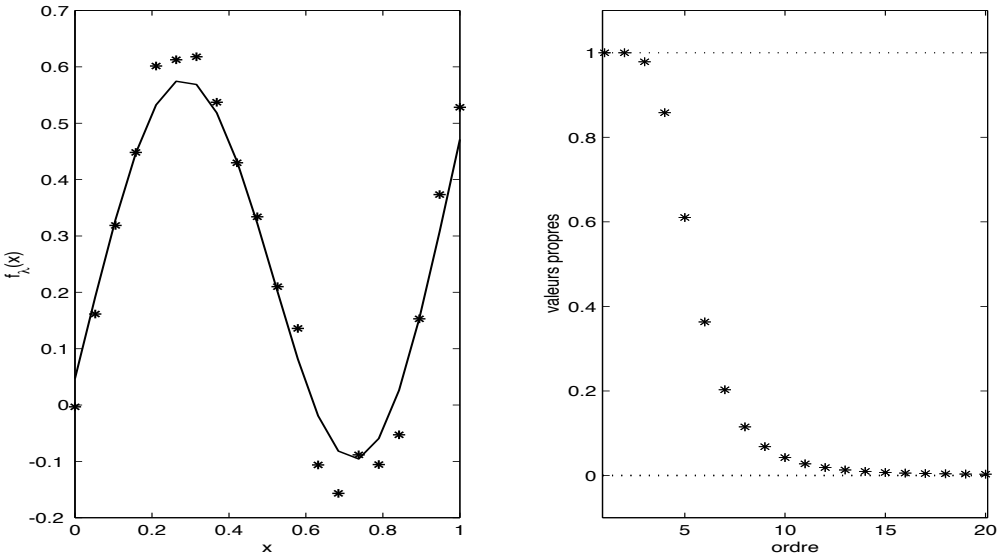

FIG. 1.2 – Estimation par des splines cubiques de lissage (à gauche). La valeur du paramètre de lissage est fixée à $\lambda = 10^{-3}$ Les données sont générées par $y = \frac{x}{2} + \frac{1}{2}\sin(2\pi x) + \varepsilon$, Les valeurs de x sont régulièrement espacées sur l'intervalle $[0, 1]$ $(n = 20)$, et ε est une variable normale centrée d'écart–type $\sigma = 0.04$. Valeurs propres pour les splines cubiques de lissage (à droite) correspondant aux fonctions propres de la figure (1.3).

1.2.3.3 Splines pénalisées

Quand le nombre d'observations est élevé, les splines cubiques de lissage présentent des difficultés numériques, conséquence du nombre élevé d'éléments dans la base de fonctions. Les splines pénalisées ou P–splines, comme les splines de lissage, sont le résultat d'un problème de minimisation des moindres carrés pénalisés, mais on n'utilise plus l'ensemble maximal de nœuds $\{x_{i1}\}$ [Eilers et Marx, 1996]. Ces méthodes reposent sur l'idée qu'on peut utiliser une base de splines de moins de n éléments avec une perte négligeable d'information.

La position et le nombre de nœuds doivent être alors déterminés, comme pour les splines de régression. Généralement, les points sont uniformément placés sur $[x_{(1)1}, x_{(n)1}]$ [Eilers et Marx, 1996], sur les percentiles [Ruppert, 2002], ou aléatoirement sur l'ensemble $\{x_{i1}\}$ [Gu et Kim, 2002, Kim et Gu, 2004]. Quant au nombre de nœuds, il a été observé que celui–ci peut être beaucoup plus petit que le nombre d'observations (de l'ordre de $kn^{2/(4r+1)}$, où $k \approx 10$ et $r \in [1, 2]$ dépend de la régularité de la vraie fonction[2]) sans affecter l'estimation [Gu et Kim, 2002, Kim et Gu, 2004].

Une autre technique qui vise à éviter les problèmes numériques est celle des

[2]Par exemple, pour les splines cubiques, $r = 1$ si la vraie fonction f vérifie uniquement $\int [f^{(2)}(t)]^2 dt < \infty$, $r = 1.5$ si $\int [f^{(3)}(t)]^2 dt < \infty$, et $r = 2$ si $\int [f^{(4)}(t)]^2 dt < \infty$.

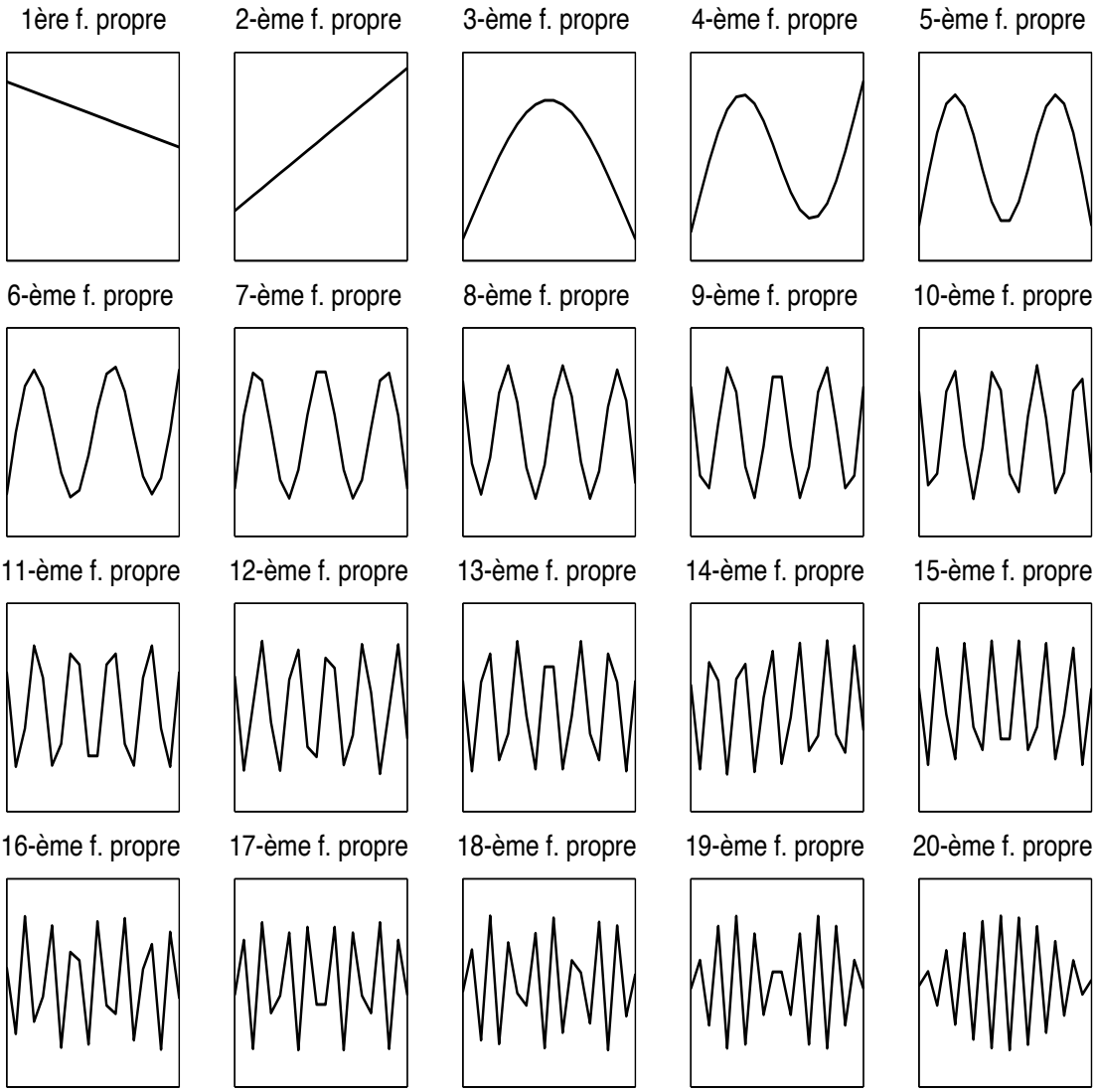

FIG. 1.3 – Fonctions propres correspondantes aux valeurs propres ordonnées (de façon décroissante) de la matrice de lissage des splines cubiques de lissage avec $n = 20$.

pseudo–splines, basée sur l'élimination des fonctions propres correspondantes aux valeurs propres de la matrice de lissage proches de zéro [Hastie, 1996]. Dans l'exemple de la figure (1.2), cela consiste à annuler exactement les dernières valeurs propres.

1.2.3.4 Différentes approches des splines de lissage

L'équation (1.20) témoigne du fait qu'il existe deux objectifs opposés en estimation fonctionnelle : maximiser l'ajustement aux données et minimiser les fluctuations de la courbe. La minimisation de la somme des carrés résiduels, en ajoutant un terme de pénalisation sur les grandes fluctuations, aborde directement le compromis

nécessaire. Cette approche est un cas particulier de la log–vraisemblance pénalisée [Green et Silverman, 1994] :

$$l_P(f) = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_{i1}))^2 - \frac{1}{2}\lambda\int\left[f^{(2)}(t)\right]^2 dt, \tag{1.26}$$

où σ^2 est la variance du bruit d'observation (supposé gaussien). Si le paramètre de lissage re–paramétré par $\mu = \lambda\sigma^2$, il est immédiat que maximiser de l_P est équivalent à minimiser le problème pénalisé (1.20).

Aussi, l'équation (1.20) est le lagrangien du problème d'optimisation sous contraintes :

$$\min_{f\in\mathcal{C}^2}\sum_{i=1}^{n}(y_i - f(x_{i1}))^2 \quad \text{sous contrainte} \quad \int\left[f^{(2)}(t)\right]^2 dt \leq \tau. \tag{1.27}$$

Il existe une relation monotone bijective entre le terme du lagrangien λ et "la limite de rugosité" (*the bound of roughness*) τ.

Il existe également une caractérisation bayésienne des splines cubiques de lissage [Green et Silverman, 1994, Hastie et Tibshirani, 2000]. Si une distribution a priori gaussienne est considérée, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^-\tau^2)$, la distribution a posteriori résultante est $\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbf{S}(\lambda)\mathbf{y}, \mathbf{S}(\lambda)\sigma^2)$, où $\lambda = \sigma^2/\tau^2$, et $\mathbf{K}$ matrice de pénalisation des splines cubiques naturelles, telle que $\int[f^{(2)}(x)]^2dx = \mathbf{f}^t\mathbf{K}\mathbf{f}$. La matrice $\mathbf{K}^-$ est une inverse généralisée de $\mathbf{K}$, telle que les valeurs propres de $\mathbf{K}$ égales à 0 (correspondantes aux vecteurs propres des fonctions constantes et linéaires) deviennent des valeurs propres de $\mathbf{K}^-$ égales à $+\infty$. Par conséquent, la distribution a priori attribue une variance "infinie" aux fonctions linéaires et constantes, ce qui veut dire qu'aucune contrainte n'est appliquée à ces fonctions.

1.2.3.5 Interprétation du paramètre de lissage

Le paramètre de lissage λ contrôle le compromis entre l'ajustement aux données et le lissage de la courbe, soit encore le compromis entre les effets du biais et les effets de la variance [Linde, 2000]. Quand $\lambda \to \infty$, la solution est une régression linéaire. Quand $\lambda \to 0$, la solution est une courbe très accidentée (mais régulière : $f \in \mathcal{C}^2$), interpolant les observations.

1.2.4 Relation entre les méthodes

De nombreuses méthodes d'estimation non paramétriques sont disponibles : des polynômes locaux (incluant le degré 0), des splines, des ondelettes, des séries de Fourier, des plus proches voisins, ... Les deux premières, traitées ici, possèdent de bonnes propriétés d'approximation de la vraie fonction, les bords inclus. Ces méthodes sont également faciles à comprendre : les polynômes locaux généralisent la régression linéaire, par l'incorporation de non linéarités locales (ce qui permet l'utilisation explicite du cadre théorique des moindres carrés pour la dérivation des propriétés des

estimateurs), l'estimateur par splines, lui, émerge de l'optimisation de la vraisemblance pénalisée (ce qui permet l'immersion dans un cadre théorique très général) [Simonoff, 1996].

Malgré leurs différences quant à l'origine, il existe une relation entre l'estimateur obtenu par les splines de lissage et un estimateur obtenu par un noyau quelconque : les splines de lissage sont, fondamentalement, une pondération locale par noyaux, avec une largeur de bande variable [Silverman, 1984, Green et Silverman, 1994, Simonoff, 1996].

Les splines de lissage sont une méthode de lissage linéaire en Y, l'estimation en x_0 s'écrit donc : $\widehat{f}(x_0) = 1/n \sum_{i=1}^{n} y_i S(x_0, x_{i1})$, où $S(x_0, x_{i1})$ est le poids en $\widehat{f}(x_0)$ associé au point x_{i1}. L'ensemble des pondérations $\{S(x_0, x_{i1})\}_i$ est dit noyau équivalent (*equivalent kernel*) en x_0. Sous certaines conditions (n grand, x n'est pas trop proche du bord du domaine défini par les observations, et λ n'est ni trop grand, ni trop petit), les approximations suivantes sont obtenues :

$$S(x_0, x) \approx \frac{K_{\mu(x)}(x_0, x)}{h_X(x)}, \tag{1.28}$$

où h_X est la densité locale de X et μ le paramètre (variable) largeur de bande, qui vérifie $\mu(x) = \lambda^{1/4} n^{-1/4} h_X(x)^{-1/4}$. La *fonction noyau* K est définie par :

$$K(t) = \frac{1}{2} \exp\Big(- \frac{|t|}{\sqrt{2}} \Big) \sin\Big(\frac{|t|}{\sqrt{2}} + \frac{\pi}{4} \Big). \tag{1.29}$$

Cette approximation de $S(x_0, x)$ met en évidence le fait que les splines de lissage définissent, approximativement, une convolution. Les observations qui se trouvent dans le voisinage, contribuent plus à l'estimation, et la vitesse avec laquelle l'influence des données se dissipe est contrôlée par $\mu(x)$. Aussi, cette approximation suggère une correspondance entre l'estimateur obtenu par les splines de lissage et un estimateur obtenu par un polynôme local de degré 2 ou 3.

Observons que la forme de K indique que l'influence des points sur l'estimation de la courbe se dissipe exponentiellement. Une modification du paramètre de lissage λ affecte le paramètre local largeur de bande, dans la même façon multiplicative partout. Souvent, la valeur optimale de λ varie énormément entre différents problèmes, en particulier si l'échelle de la variable explicative est différent. Le fait que λ soit proportionnelle à la puissance 4 de la largeur de bande locale pourrait expliquer cette grande variation.

Observons également que la largeur de bande locale dépend de la densité h_X. Cette situation est intermédiaire entre le paramètre de lissage global et le lissage basé sur un moyennage d'un nombre fixé de points du voisinage (largeur de bande locale proportionnelle à $1/h_X$). Ce comportement intermédiaire permet une bonne adaptabilité des splines de lissage aux effets dus à la variabilité de la densité de la variable d'entrée et aux changements rapides de la courbure de la fonction de régression.

1.2.4.1 Comparaison expérimentale des méthodes

Différentes méthodes de lissage, parmi lesquelles se trouvent les splines cubiques de lissage, les splines de régression et les noyaux, ont été comparées par [Breiman et Peters, 1992]. Ils concluent qu'aucune des méthodes ne domine dans tous les aspects étudiés. Les splines de lissage sont les plus précises (au sens des erreurs quadratiques) mais elles ont une tendance à sous–lisser, les splines de régression obtiennent la meilleure performance globale (pour toute valeur de n, pour toutes les distributions). Dans leurs expériences, la méthode basée sur les noyaux semble avoir des problèmes de biais et variance sur les bords.

1.3 Modèles additifs

1.3.1 Régression non paramétrique multidimensionnelle

La généralisation multidimensionnelle du problème (1.1) est la suivante :

$$Y = f(X_1, \ldots, X_p) + \varepsilon, \tag{1.30}$$

où $(\boldsymbol{X}, Y) = (X_1, \ldots, X_p, Y)$ vecteur aléatoire, ε variable aléatoire indépendante de $\boldsymbol{X}$ telle que $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

L'ajustement de Y à une surface p–dimensionnelle par lissage peut se faire en généralisant les noyaux par [Härdle et Muller, 2000] :

$$\mathcal{K}_{\boldsymbol{\Lambda}}(\mathbf{x}, \mathbf{x}') = \frac{1}{\det(\boldsymbol{\Lambda})} \mathcal{K}\left(\boldsymbol{\Lambda}^{-1}(\mathbf{x} - \mathbf{x}')\right), \tag{1.31}$$

où $\mathbf{x} = (x_1, \ldots, x_p)^t$, $\mathbf{x}' = (x'_1, \ldots, x'_p)^t$, et $\boldsymbol{\Lambda}$ est une matrice symmétrique, définie positive. De nombreuses possibilités existent pour définir la fonction $\mathcal{K}$. Par exemple, elle peut être définie par le produit de p noyaux unidimensionnels, $K : \mathcal{K}(\mathbf{t}) = K(t_1) \cdot \ldots \cdot K(t_p)$, ou par un seul noyau unidimensionnel : $\mathcal{K}(\mathbf{t}) = K(\|\mathbf{t}\|)$, où le choix de la norme détermine la forme des voisinages. Une autre possibilité consiste à généraliser directement les fonctions noyaux unidimensionnelles (par exemple, le noyau gaussien (1.4) p–dimensionnel).

La généralisation de l'estimateur de Nadaraya–Watson (1.5) est alors :

$$\widehat{\mathbf{f}}_{NW}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i \mathcal{K}_{\boldsymbol{\Lambda}}(\mathbf{x}, \mathbf{x}^i)}{\sum_{i=1}^n \mathcal{K}_{\boldsymbol{\Lambda}}(\mathbf{x}, \mathbf{x}^i)}, \tag{1.32}$$

où $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^t$.

La généralisation du problème de minimisation (1.8), dans le cas particulier de la régression locale linéaire, est :

$$\sum_{i=1}^n \left(y_i - \beta_0 - (\mathbf{x} - \mathbf{x}^i)^t \boldsymbol{\beta}_1\right)^2 \mathcal{K}_{\boldsymbol{\Lambda}}(\mathbf{x}, \mathbf{x}^i), \tag{1.33}$$

où $\boldsymbol{\beta}_1$ est un vecteur de dimension $p \times 1$.

Pour les splines cubiques, une possibilité est de généraliser la pénalisation de la dérivée seconde (1.20) à une pénalisation "plate" [Gu, 2000] :

$$\int\int \left\{ \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right\} dx_1 dx_2, \tag{1.34}$$

en dimension 2, ou

$$\int \cdots \int \left\{ \sum_{j=1}^{p} \left(\frac{\partial^2 f}{\partial x_j^2} \right)^2 + 2 \sum_{j>k} \left(\frac{\partial^2 f}{\partial x_j \partial x_k} \right)^2 \right\} \prod_{j=1}^{p} dx_j, \tag{1.35}$$

en général. Une autre possibilité est le produit tensoriel de splines, définis sur des espaces de Hilbert à noyau auto–reproduisant, RKHS (*Reproducing kernel Hilbert Spaces*)[3].

1.3.1.1 Les problèmes de la dimension élevée

Dans le cas multidimensionnel, la régression non paramétrique présente plusieurs problèmes. Premièrement, la représentation graphique n'est pas possible pour plus de 2 variables explicatives, et l'interprétation devient difficile.

Deuxièmement, l'approche des méthodes locales échoue en dimension élevée. C'est le problème dit du "fléau de la dimensionnalité" (*curse of dimensionality*) [Bellman, 1961], qui se manifeste de façons diverses. Par exemple, supposons que les observations des variables explicatives soient uniformément distribuées dans un cube unitaire p–dimensionnel ($p = 2$, $p = 10$). Pour recouvrir un pourcentage des données $r = 10\%$, il faut un sous–cube de côté $r^{1/p}$. La longueur du côté est 0.32, pour $p = 2$, et 0.79, pour $p = 10$. Pour p élevé, ces voisinages ne sont plus "locaux" (la longueur du côté est très proche de l'unité, et donc le sous–cube est très proche du cube global). Par conséquent, quand la dimension augmente, soit il faut prendre des voisinages plus grands, ce qui implique des moyennages globaux et donc, des grands biais, soit il faut réduire le pourcentage des données, r, ce qui implique des moyennages sur peu d'observations et donc, des grandes variances de l'ajustement [Hastie *et al.*, 2001].

Troisièmement, en dimension élevée, la plupart des ensembles de données se trouvent sur des variétés de dimension moins importantes. Si ces variétés sont des hyper–plans, on rencontre le problème de colinéarité des variables explicatives. Si ces variétés sont régulières, on rencontre le problème plus général de "concurvité". (voir section 1.3.4.4, page 34).

Finalement, si la surface à estimer est m fois continûment différentiable sur un domaine p–dimensionnel borné, alors le taux asymptotique optimal pour l'estimation de la surface de régression est de l'ordre de $n^{-2m/(2m+p)}$. Pour obtenir un taux du même ordre que dans le cas unidimensionnel, il faut supposer que la fonction est $m \times p$ continûment différentiable : quand la dimension croît, les surfaces pouvant être estimées sont de plus en plus régulières.

[3]Soit $\mathcal{H}$ un espace de Hilbert de fonctions sur un domaine $\mathcal{X}$. Si $\forall x \in \mathcal{X}$, l'opérateur d'évaluation $[x]f = f(x)$ est continu en $\mathcal{H}$, alors $\mathcal{H}$ est dit un espace RKHS.

1.3.1.2 Réduction de la dimension

Une solution aux problèmes de la dimension élevée consiste à supposer que la fonction de régression possède une structure déterminée. Ces techniques non paramétriques restent des outils flexibles. Le prix à payer est la possible spécification erronée du modèle.

Les techniques basées sur des principes de "réduction de la dimension" sont les modèles additifs, qui supposent que la fonction de régression est une somme de fonctions monovariées en chacune des variables, les modèles de projections révélatrices (*projection pursuit*), proches des réseaux de neurones de type perceptron multicouche, et les arbres.

Projections révélatrices

L'algorithme des projections révélatrices (*projection pursuit*) construit un modèle de régression additif de la forme [Friedman et Stuetzle, 1981, Klinke et Grassmann, 2000] :

$$Y = \sum_{k=1}^{K} f_k(\boldsymbol{\alpha}_k^T \boldsymbol{X}) + \varepsilon, \tag{1.36}$$

où ε est tel que $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$, et indépendant des variables d'entrée.

La matrice des données est projetée sur K directions $\boldsymbol{\alpha}_k$. La surface de régression est construite par l'estimation des régressions unidimensionnelles f_k appliquées aux projections. Les directions $\boldsymbol{\alpha}_k$ et le nombre de termes K sont choisis par des méthodes de sélection de modèle telles que la validation croisée généralisée.

L'avantage de cette technique est qu'elle permet le traitement facile des données peu denses. Le modèle est également peu contraint. Néanmoins, pour $K > 1$, ce modèle présente des difficultés d'interprétation : il est difficile d'évaluer les contributions de chaque variable. Pour $K = 1$, le modèle est connu sous le nom de modèle à indice simple (*single–index model*).

Les projections révélatrices sont souvent comparées aux perceptrons multicouches. Ces deux méthodes extraient des combinaisons linéaires des entrées, et modélisent, ensuite, la variable de sortie comme une fonction non linéaire de celles–ci. Cependant, les fonctions f_k des projections révélatrices sont différentes et non paramétriques, alors que les réseaux de neurones utilisent une fonction (d'activation) plus simple, normalement la fonction softmax (ou logistique). Dans le cas des projections révélatrices, le nombre de "couches" est fixé à deux et le nombre de fonctions K est également prédéfini, ce qui n'est pas le cas pour les réseaux de neurones.

Arbres

Les arbres divisent l'espace des variables explicatives en un ensemble d'hyper–cubes. Un modèle simple (par exemple, une constante) est alors ajusté sur chaque hyper–cube :

$$f(x) = \sum_{k=1}^{K} \alpha_k \mathbb{I}_{\{x \in R_k\}}, \tag{1.37}$$

où K le nombre de régions de la partition, R_k les régions disjointes, α_k la constante qui modélise la réponse dans la région R_k. L'algorithme décide simultanément la partition et les valeurs des paramètres α_k.

Les arbres ont l'avantage de la simplicité conceptuelle et la capacité d'interprétation. Leurs limitations sont l'instabilité et l'absence de continuité de la surface de régression.

1.3.2 Modèles additifs

Les modèles additifs supposent que la fonction de régression peut s'écrire comme une somme de fonctions des variables explicatives [Stone, 1985, Hastie et Tibshirani, 1986] :

$$Y = \alpha_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{1.38}$$

où ε est indépendant de $\boldsymbol{X} = (X_1, \ldots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$; α_0 est une constante, f_j, $j = 1, \ldots, p$ sont des fonctions unidimensionnelles telles que $\mathbb{E}_{X_j}[f_j] = 0$ (pour qu'il y ait unicité). Cette condition d'identifiabilité implique que $\mathbb{E}_{\boldsymbol{X}}[Y] = \alpha_0$ [Hastie et Tibshirani, 1990].

Généralisation du modèle linéaire

Les modèles additifs peuvent être introduits comme une généralisation du modèle de régression linéaire multiple. Celui-ci est l'outil de base pour modéliser la relation entre la variable réponse continue et les variables explicatives :

$$Y = \alpha_0 + X_1\alpha_1 + \ldots + X_p\alpha_p + \varepsilon, \tag{1.39}$$

où ε est indépendant de $\boldsymbol{X}$, $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

La supposition de dépendance linéaire de $\mathbb{E}_{\boldsymbol{X}}[Y]$ en chacune des variables explicatives est une hypothèse forte. Quand cette hypothèse n'est pas vérifiée, une façon d'étendre le modèle linéaire est le modèle additif. La forme non–paramétrique des f_j accorde plus de flexibilité au modèle, alors que la structure additive préserve la possibilité de représenter l'effet de chaque variable. Le modèle ajusté peut être représenté par p fonctions unidimensionnelles décrivant les rôles des variables explicatives dans la modélisation de la réponse, ce qui facilite l'interprétation. Cependant, la simplicité du modèle linéaire est perdue. Un nouveau problème apparaît : la sélection des paramètres de lissage, représentant la complexité de chaque composante du modèle.

1.3.3 Propriétés du modèle

Interprétabilité

L'effet conjoint des variables explicatives sur la variable réponse est exprimé comme une somme des effets individuels. Ces effets individuels montrent comment l'espérance de la réponse varie quand une des composantes varie alors que les autres sont fixées à des valeurs quelconque. Ainsi, les fonctions individuelles peuvent être représentées séparément afin de visualiser l'effet de chaque variable explicative, rendant intelligible le résultat. La possibilité de représenter les effets des variables directement (sans transformations) fournit au même temps des indications sur l'importance de chacune des variables.

Fléau de la dimensionnalité

En restreignant la nature des dépendances, les problèmes liés à la dimension élevée sont atténués : la réponse est modélisée comme la somme de fonctions unidimensionnelles des variables explicatives, au lieu d'être modélisée par des fonctions multidimensionnelles. Par conséquent, le nombre d'observations requis croit linéairement avec p (et non pas exponentiellement).

Considérons l'estimation de la fonction de régression (1.30). Le taux asymptotique optimal pour l'estimation de f est $n^{-[m/(2m+p)]}$, où m est un indice de la régularité de la fonction (f est $m-1$ fois continûment différentiable et ses m–èmes dérivés directionnelles existent)[Stone, 1982]. En revanche, si f est additive, le taux optimal atteint le taux de convergence unidimensionnel $n^{-[m/(2m+1)]}$ [Stone, 1986]. En ce sens–là, les modèles additifs sont considérés comme des techniques de réduction de la dimension.

Modèle incorrect

Le modèle est mal spécifié les variables explicatives interagissent. C'est à dire, que l'effet des variations d'une variable explicative sur la réponse dépend des valeurs adoptées par les autres variables explicatives.

Supposons le modèle de régression multiple général (1.30), où la fonction f est une fonction lisse. Supposant que les observations $\{x_{ij}\}$ sont contenues dans une région où la courbure de la fonction f est petite dans les espaces produit, alors l'additivité (et la linéarité) peut être justifiée par une expansion de Taylor de premier ordre : $f(\mathbf{x}) \approx f(\mathbf{x}') + \mathrm{D}f(\mathbf{x})(\mathbf{x}-\mathbf{x}')$, où $\mathbf{x}'$ est dans la région définie par les observations et $\mathrm{D}f$ indique la différentielle de f [Ruppert *et al.*, 2003]. Si la courbure de f est élevée dans la région définie par les produits cartesiens des observations, l'expansion de Taylor nécessite, au moins, des termes quadratiques et des termes croisés en 2 variables. Quand seulement les premiers sont nécessaires, le modèle est toujours additif, bien qu'il incorpore des termes "non linéaires".

Adaptabilité

L'intérêt des modèles additifs est leur capacité à modéliser la relation entre les variables d'une façon intuitive, mais aussi la possibilité d'adapter le modèle à des situations plus simples ou plus complexes. Quand des composantes ne demandent pas une modélisation non paramétrique, elles peuvent être réduites à des composantes

linéaires (voir modèles semi–paramétriques, section 1.4.4, page 49). Egalement, quand des interactions existent entre certaines variables, des termes quadratiques (ou d'ordre supérieur) peuvent être intégrés dans le modèle (voir modèles d'interaction, section 1.4.4, page 49).

1.3.4 Estimation

1.3.4.1 Complexité de l'estimation

La complexité de l'estimation des modèles additifs dépasse largement celle des modèles linéaires [Breiman, 1993]. L'espace des modèles linéaires pour p variables explicatives est p–dimensionnel (les moindres carrés choisissent un de ces espaces). L'espace des modèles additifs excède largement cette dimension. Souvent, la dimensionnalité n'est pas bien définie. Les données sont alors utilisées pour choisir entre de nombreuses alternatives.

Si l'objectif principal est la prédiction, les problèmes sont plus simples. Si on veut comprendre les relations entre variables, les problèmes sont parfois rendus difficiles par les problèmes de concurvité. Dans ce cas, il est possible d'obtenir un ensemble de modèles très différents, au niveau de la relation entre les variables explicatives, mais présentant la même précision en prédiction.

1.3.4.2 Équations normales

Afin de justifier les techniques de lissage pour les modèles additifs, le problème peut être formulé dans un espace de Hilbert [Hastie et Tibshirani, 1990].

Soit $[\mathcal{H}_{\boldsymbol{X}Y}, \langle ., . \rangle]$ l'espace de Hilbert des variables aléatoires, $g(\boldsymbol{X}, Y)$, fonctions de $\boldsymbol{X} = (X_1, \ldots, X_p)$ et de Y, centrées ($\mathbb{E}_{(\boldsymbol{X},Y)}(g(\boldsymbol{X}, Y)) = 0$), de variance finie ($\mathbb{E}_{(\boldsymbol{X},Y)}(g(\boldsymbol{X}, Y)^2) < \infty$), et produit scalaire défini par $\langle g_1(\boldsymbol{X}, Y), g_2(\boldsymbol{X}, Y) \rangle = \mathbb{E}_{(\boldsymbol{X},Y)}(g_1(\boldsymbol{X}, Y) \cdot g_2(\boldsymbol{X}, Y))$.

Soient $[\mathcal{H}_{X_j}, \langle ., . \rangle]$, $j = 1, \ldots, p$, des sous–espaces tels que $\mathcal{H}_{X_j}$ contient seulement des fonctions de X_j, centrées de carré intégrable. La relation entre ces sous–espaces est donnée par : $[\mathcal{H}_{X_j}, \langle ., . \rangle] \subset [\mathcal{H}_{X_1} \oplus \ldots \oplus \mathcal{H}_{X_p}, \langle ., . \rangle] \subset [\mathcal{H}_{\boldsymbol{X}}, \langle ., . \rangle] \subset [\mathcal{H}_{\boldsymbol{X}Y}, \langle ., . \rangle]$ (où $\oplus$ indique la somme directe).

Les fonctions individuelles $f_j(X_j)$ sont déterminées de façon unique par la condition $\mathbb{E}_{(\boldsymbol{X},Y)}(g(\boldsymbol{X}, Y)) = 0$. Cette condition implique que l'unique fonction constante qui appartient aux espaces définis est la fonction $\mathbf{0}$. Également, la fonction Y (l'identité appliquée à la variable Y) est supposée centrée.

La meilleure estimation possible de f est celle qui minimise le critère donné par l'erreur quadratique de prédiction sous la contrainte d'additivité :

$$\begin{cases} \min\limits_{f \in \mathcal{S}} \mathbb{E}_{\boldsymbol{X}Y}[Y - f(\boldsymbol{X})]^2 = \min\limits_{f \in \mathcal{S}} \langle Y - f(\boldsymbol{X}), Y - f(\boldsymbol{X}) \rangle_{\mathcal{H}_{\boldsymbol{X}Y}} \\ \\ \text{sous contrainte} \quad f(\boldsymbol{X}) = \sum_{j=1}^{p} f_j(X_j), \end{cases} \tag{1.40}$$

où $\mathcal{S}$ est la classe de fonctions de lissage (splines, noyaux,...).

Ce problème de minimisation sous contraintes revient à trouver l'élément du sous–espace $\mathcal{H}_{X_1} \oplus \ldots \oplus \mathcal{H}_{X_p}$ le plus proche du point $Y \in \mathcal{H}_{\boldsymbol{X}Y}$, ou de façon équivalente, le point $f(\boldsymbol{X}) \in \mathcal{H}_{\boldsymbol{X}}$ de la forme $\sum_{j=1}^{p} f_j(X_j)$ le plus proche du point $Y \in \mathcal{H}_{\boldsymbol{X}Y}$.

Sans contraintes, la solution au problème (1.40) est $f(\boldsymbol{X}) = E(Y|\boldsymbol{X})$. Avec la contrainte, on cherche la solution additive la plus proche.

Puisque $[\mathcal{H}_{X_1} \oplus \ldots \oplus \mathcal{H}_{X_p}, \langle ., . \rangle] \subset [\mathcal{H}_{\boldsymbol{X}}, \langle ., . \rangle]$ est fermé (sous des hypothèses techniques), par le théorème de la projection[4], il existe une solution unique de l'approximation optimale de $f(\boldsymbol{X})$ dans l'espace additif : $\sum_{j=1}^{p} f_j(X_j)$.

L'opérateur espérance conditionnelle, $P_j = \mathbb{E}(.|X_j) : \mathcal{H}_{\boldsymbol{X}Y} \longrightarrow \mathcal{H}_{X_j}$ est une projection orthogonale sur l'espace X_j. L'élément $f(\boldsymbol{X})$ minimisant (1.40) peut être caractérisé par le résidu, $Y - f(\boldsymbol{X})$, orthogonal à l'espace $\mathcal{H}_{X_1} \oplus \ldots \oplus \mathcal{H}_{X_p}$. Aussi, puisque cet espace est généré par les $\mathcal{H}_{X_j}$, le résidu est orthogonal aux $\mathcal{H}_{X_j}$ et donc, $P_j\,(Y - f(\boldsymbol{X})) = 0$, $\forall j$. Par l'équivalence :

$$\mathbb{E}[Y - f(\boldsymbol{X})|X_j] = 0 \Leftrightarrow f_j(X_j) = \mathbb{E}[Y - \textstyle\sum_{k \neq j} f_j(X_j)|X_j], \tag{1.41}$$

$j = 1, \ldots, p$, les fonctions $f(X_j)$ sont caractérisées. Cela implique la représentation matricielle du problème (1.40) :

$$\begin{pmatrix} I & P_1 & \ldots & P_1 \\ P_2 & I & \ldots & P_2 \\ \vdots & \vdots & \ddots & \vdots \\ P_p & \ldots & P_p & I \end{pmatrix} \quad \begin{pmatrix} f_1(X_1) \\ f_2(X_2) \\ \vdots \\ f_p(X_p) \end{pmatrix} \quad = \quad \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_p Y \end{pmatrix}. \tag{1.42}$$

De manière équivalente, les matrices des estimations de (1.42) sont considérées pour des observations. Soit $\{(X_{i1}, \ldots, X_{ip}, Y_i)\}_{i=1}^{n}$ un échantillon i.i.d. de $(\boldsymbol{X}, Y)$, et $\{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^{n}$ des réalisations. Soit $\mathbf{S}_j$ (ou $\mathbf{S}_j(\lambda_j)$) la matrice de lissage $n \times n$, pour λ_j fixé, alors $\widehat{\mathbf{f}}_j = \mathbf{S}_j \mathbf{y}$, où $\widehat{\mathbf{f}}_j = (\widehat{f}_{1j}, \ldots, \widehat{f}_{nj})^t$ et $\widehat{f}_{ij} = \widehat{f}(x_{ij})$.

En remplaçant P_j par $\mathbf{S}_j$ le problème est transformé en :

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_1 & \ldots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \ldots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \ldots & \mathbf{S}_p & \mathbf{I} \end{pmatrix} \quad \begin{pmatrix} \widehat{\mathbf{f}}_1 \\ \widehat{\mathbf{f}}_2 \\ \vdots \\ \widehat{\mathbf{f}}_p \end{pmatrix} \quad = \quad \begin{pmatrix} \mathbf{S}_1 \mathbf{y} \\ \mathbf{S}_2 \mathbf{y} \\ \vdots \\ \mathbf{S}_p \mathbf{y} \end{pmatrix}. \tag{1.43}$$

Ce système d'équations $np \times np$ des estimations est le système des équations normales. De façon plus compacte, on notera :

$$\mathbf{A}\widehat{\mathbf{f}} = \mathbf{B}\mathbf{y}. \tag{1.44}$$

[4]**Théorème de la Projection**
Soit $[\mathcal{H}, \langle ., . \rangle]$ un espace de Hilbert, et $\mathcal{K} \subset \mathcal{H}$ sous–espace convexe non vide et fermé.
$\forall x \in \mathcal{H}$, $\exists!\ y \in \mathcal{K}$ projection optimale ou orthogonale de x sur $\mathcal{K}$, i.e. $\langle x-y, x-y \rangle = \inf_{z \in \mathcal{K}} \langle x-z, x-z \rangle$.
On notera $y = P_{\mathcal{K}}(x)$. La projection orthogonale y est caractérisée par être l'unique élément de $\mathcal{K}$ tel que $\forall z \in \mathcal{K}\ Re\langle x - y, z - y \rangle \leq 0$.
L'opérateur $P_{\mathcal{K}}(.)$ est continu et linéaire.

1.3.4.3 Existence et unicité des solutions

Sous certaines conditions, les équations normales (1.43)–(1.44) sont consistantes (il existe au moins une solution) [Schimek et Turlach, 2000] :
Si $\forall j,\ j = 1, \ldots, p$, la matrice de lissage linéaire, $\mathbf{S}_j$, est symétrique et ses valeurs propres sont comprises dans l'intervalle $[0, 1]$, alors les équations normales sont consistantes pour tous les $\mathbf{y}$.

Cependant, pour que cette solution soit unique d'autres conditions sont nécessaires [Hastie et Tibshirani, 1990, Schimek et Turlach, 2000]. Pour le cas $p = 2$, la solution est unique si $\|\mathbf{S}_1\mathbf{S}_2\| < 1$, où $\|.\|$ est une norme matricielle quelconque (les conditions de symétrie et que les valeurs propres soient comprises à l'intervalle $]-1, 1]$ sont suffisantes pour la consistance, dans la cas $p = 2$). Pour $p > 2$, la condition d'unicité devient [Opsomer, 2000, Schimek et Turlach, 2000] :

$$\max_{\delta \in [2,p]} \left|\left| \sum_{j=1}^{\delta-1} \mathbf{S}_\delta \mathbf{S}_j \right|\right| < 1. \tag{1.45}$$

1.3.4.4 Concurvité

La concurvité est une cause de dégénérescence des équations normales, dont le résultat est la non unicité des solutions.

Dans le cas de la régression linéaire (1.39) la colinéarité désigne une relation exacte ou proche de la dépendance linéaire entre deux ou plusieurs variables explicatives (singularité exacte ou "mauvais conditionnement", respectivement) [Wetherill, 1986, Sen et M., 1990] : pour au moins un $j \in \{1, \ldots, p\}$, il existe un ensemble de scalaires $\{\alpha_k\}_{k \neq j}$, tels que $\mathbf{x}_j \approx \sum_{k \neq j} \alpha_k \mathbf{x}_k$. En présence de colinéarité, les équations normales,

$$\mathbf{X}^t\mathbf{X}\widehat{\boldsymbol{\alpha}} = \mathbf{X}^t\mathbf{y}, \tag{1.46}$$

sont dégénérées, ou la matrice $\mathbf{X}^t\mathbf{X}$ est proche de la singularité.

Dans le cas de la régression non paramétrique, la colinéarité se généralise à la concurvité [Buja *et al.*, 1989, Hastie et Tibshirani, 1990]. La concurvité désigne une relation régulière et additive proche de dépendance entre deux ou plusieurs variables explicatives : pour au moins un $j \in \{1, \ldots, p\}$, il existe un ensemble de fonctions $\{g_k\}_{k=1,\ldots,p}$, tels que $g_j(X_j) \approx \sum_{k \neq j} g_k(X_k)$. En termes d'équations normales, la concurvité est définie comme l'existence d'une solution non nulle à $\mathbf{Ag} = \mathbf{0}$ (1.44). Si un tel $\mathbf{g}$ existe, et que $\widehat{\mathbf{f}}$ est solution pour $\mathbf{A}\widehat{\mathbf{f}} = \mathbf{By}$, alors $\widehat{\mathbf{f}} + \gamma\mathbf{g}$ est aussi solution pour toute valeur de γ.

La concurvité, dans le cas consistant, peut être formulée en fonction des valeurs propres des matrices de lissage [Hastie et Tibshirani, 1990, Schimek, 2000]. Soit $\mathbb{M}_1(\mathbf{S}_j)$ l'espace généré par les vecteurs propres de $\mathbf{S}_j$ de valeur propre 1, $j = 1, \ldots, p$. Il existe de la concurvité si, et seulement si, les espaces $\mathbb{M}_1(\mathbf{S}_j)$ sont linéairement dépendants. Pour $p = 2$, cela est équivalent à En pratique, cela signifie qu'il existe un vecteur réponse, $\mathbf{y}$ qui peut être parfaitement expliqué soit par $\mathbf{x}_1 = (x_{11}, \ldots, x_{n1})^t$, soit par $\mathbf{x}_2 = (x_{12}, \ldots, x_{n2})^t$.

Dans la pratique, la singularité exacte (concurvité exacte) du système matriciel est improbable, cependant des systèmes proches de la singularité sont assez communs. La concurvité "approchée" peut être caractérisée par un mauvais conditionnement de $\mathbf{A}$ qui traduit des relations presque déterministes entre variables explicatives.

Conséquences de la concurvité

Dans le cas linéaire, si les colonnes de $\mathbf{X}$ sont proches de la dépendance linéaire, alors $\mathbf{X}^t\mathbf{X}$ est proche de la singularité, et l'estimateur des moindres carrés de $\boldsymbol{\alpha}$ est instable. L'erreur quadratique moyenne de l'estimateur des moindres carrés, $\mathrm{MSE}(\hat{\boldsymbol{\alpha}}) = \mathbb{E}[(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})] = \sigma^2\mathrm{tr}((\mathbf{X}^t\mathbf{X})^{-1})$, sera donc élevée. Par conséquent, en présence de colinéarité, l'erreur sera élevée et les estimations seront imprécises dans les directions des vecteurs correspondant aux valeurs propres petites [Wetherill, 1986].

Dans le cas additif, les conséquences sont similaires. Des problèmes d'instabilité numérique apparaissent en présence de concurvité [Buja *et al.*, 1989, Hastie et Tibshirani, 1990, Dominici *et al.*, 2002], la contribution des variables étant statistiquement instable, l'interprétation des composantes devient imprécise ou déformée [Gu, 1992b, Gu, 2002, Donnell *et al.*, 1994], et les estimations sont biaisées et leur variance sous–estimée [Ramsay *et al.*, 2003a, Ramsay *et al.*, 2003b].

Diagnostic de la concurvité

Des mesures basées sur les corrélations entre les variables explicatives (comme le facteur d'inflation de la variance) ou les valeurs propres de la matrice $\mathbf{X}^t\mathbf{X}$ standardisée (comme le conditionnement ou la plus petite valeur propre) sont souvent utilisées pour le diagnostic de la colinéarité [Wetherill, 1986, Sen et M., 1990].

La notion de concurvité (approchée) est plus complexe, car les relations considérées entre deux ou plusieurs variables sont non paramétriques. Les rares mesures de la concurvité proposées sont soit limitées à la relation entre les approximations linéaires des fonctions, soit d'application imprécise.

Plusieurs mesures, appelées globalement diagnostic des cosinus, sont proposées par [Gu, 1992b, Gu, 2002]. Des valeurs proches de 0 de $\cos(\widehat{\mathbf{f}}_j, \widehat{\mathbf{f}}_k)$ indiquent des relations proches de l'orthogonalité. Des valeurs proches de 0 du vecteur $\pi_j = \langle\widehat{\mathbf{f}}_j, \widehat{\mathbf{y}}\rangle/\|\widehat{\mathbf{y}}\|_2^2$, désignent des termes faisant partie du bruit. Inversement, des cosinus entre les $\widehat{\mathbf{f}}_j$ et les résidus proches de 0 indiquent des termes influents. Finalement, quand la norme de $\|\widehat{\mathbf{f}}_j\|_2$, comparée à celle des observations, est petite, alors le j–ème terme est considéré négligeable.

Ces mesures se réduisent à un diagnostic des dépendances entre les approximations linéaires des fonctions (le cosinus correspond aux produits scalaires des vecteurs normalisés) et de la pertinence des composantes.

L'application de la technique "les plus petites composantes principales additives" (*the smallest additive principal component*) à la détection de la concurvité est proposée par [Donnell *et al.*, 1994]. Cette technique est une généralisation des composantes principales (linéaires). La plus petite composante principale additive est une fonction additive des données $\sum_j \mathbf{g}_j(\mathbf{x}_j)$, de variance minimum, impliquant que les

données satisfont aussi fidèlement que possible la relation additive $\sum_j \mathbf{g}_j(\mathbf{x}_j) = \mathbf{0}$. De façon analogue aux composantes principales linéaires, les valeurs propres des plus petites composantes principales additives mesurent l'importance de la dégénérescence additive. Une valeur propre égale à 0 indique la dégénérescence exacte, une valeur propre petite révèle des problèmes d'instabilité. Si la plus petite valeur propre est 1, les espaces additifs sont orthogonaux.

Cette méthode présente quelques inconvénients. Tout d'abord, elle ne définit pas le seuil en dessous duquel une valeur propre est considérée petite. Aussi, l'interprétation en termes des valeurs propres n'est pas générale, en particulier elle n'est pas directement transférable aux régressions pénalisées, telles que les splines de lissage. Finalement, l'intégration de la sélection des paramètres de la complexité n'est pas traitée.

L'étude de la régression additive d'une des variables explicatives en fonction des autres est proposée par [Ramsay *et al.*, 2003a]. La corrélation entre la variable et la régression de celle–ci en fonction des autres variables est considérée indicatrice des problèmes de concurvité.

1.3.4.5 Méthodes de lissage

La généralisation du critère minimisé par les splines cubiques dans le cas unidimensionnel (1.20) au contexte de la régression additive admet l'expression explicite suivante [Hastie et Tibshirani, 1990, Wahba, 1990] :

$$\min_{\alpha_0, f_1, \ldots, f_p} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int \left[f_j^{(2)}(t) \right]^2 dt, \tag{1.47}$$

pour α_0 scalaire, $f_j \in \mathcal{C}^2$ tel que $\mathbb{E}[f_j(X_j)] = 0$ et pour des valeurs λ_j prédéfinies. Les fonctions $\widehat{f_j}$, $j = 1, \ldots, p$, qui minimisent ce critère sont effectivement des splines cubiques [Wahba, 1990].

Chaque fonction f_j est ainsi pénalisée par une constante λ_j, qui peut être différente sur chaque composante. Comme dans le cas unidimensionnel, si $\lambda_j = 0, \forall j$, la solution est un ensemble de fonctions d'interpolation, si chaque λ_j tend vers l'infini, alors les $\widehat{f}_j^{(2)} = 0$, $\forall j$, et donc les $\widehat{f_j}$ sont linéaires (le problème est, dans ce cas, celui des moindres carrés ordinaires).

La formulation du problème (1.47), en termes d'optimisation sous contraintes est la suivante :

$$\begin{aligned} &\min_{\alpha_0, f_1, \ldots, f_p} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 \\ &\text{sous contraintes} \quad \int \left[f_j^{(2)}(t) \right]^2 dt \leq \tau_j, \quad j = 1, \ldots, p, \end{aligned} \tag{1.48}$$

où τ_j dépend de λ_j.

Considérons une base de fonctions splines, par exemple la base naturelle B–spline, constituée de $n+2$ éléments. Soient $\{N_{kj}(x)\}_{k=1}^{n+2}$ les éléments de la base pour la j–ème

composante, évalués en x, $\mathbf{N}_j$ la matrice $n \times (n+2)$ de la base évaluée en $\{x_{ij}\}_i$, et $\mathbf{\Omega}_j$ la matrice $(n+2) \times (n+2)$ correspondant à la pénalisation de la dérivée seconde. La représentation de $f_j(x)$ dans cette base est donnée par $f_j(x) = \sum_{k=1}^{n+2} \beta_{kj} N_{kj}(x)$, où les coefficients $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{n+2,j})^t$, sont estimés par minimisation de l'expression matricielle du problème (1.47) :

$$\min_{\alpha_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p} \left\| \mathbf{y} - \alpha_0 - \sum_{j=1}^{p} \mathbf{N}_j \boldsymbol{\beta}_j \right\|_2^2 + \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j{}^t \mathbf{\Omega}_j \boldsymbol{\beta}_j. \tag{1.49}$$

Cette expression admet une solution analytique dont le calcul nécessite l'utilisation de procédures numériques adaptées.

L'expression de la généralisation des estimateurs à noyaux aux modèles additifs dépend de la procédure numérique choisie. Les méthodes itératives ne permettent pas d'obtenir d'expression analytique, tandis que celle–ci est possible par intégration marginale (section 1.3.5.3).

1.3.5 Procédures numériques

Les équations normales (1.44) sont un système linéaire de la forme $\mathbf{Af} = \mathbf{b}$, où $\mathbf{b} = \mathbf{By}$, $(\mathbf{A})_{ij} = a_{ij}$, $(\mathbf{b})_i = b_i$, $i, j = 1, \ldots, np$. En principe, le système

$$\begin{pmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_p \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{S}_1 & \ldots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \ldots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \ldots & \mathbf{S}_p & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_1\mathbf{y} \\ \mathbf{S}_2\mathbf{y} \\ \vdots \\ \mathbf{S}_p\mathbf{y} \end{pmatrix} \tag{1.50}$$

peut être résolu directement. Cependant, ceci implique la résolution d'un système de taille np, avec une matrice souvent mal conditionnée.

Plusieurs procédures numériques ont été proposées pour l'estimation des modèles additifs. Le *backfitting* [Buja *et al.*, 1989, Hastie et Tibshirani, 1990] est probablement la technique la plus utilisée. Elle semble bien fonctionner en pratique, cependant l'estimateur n'est pas explicitement défini, il est obtenu de manière itérative. Par conséquent, ses propriétés statistiques ne sont pas bien connues [Schimek, 2000]. Des méthodes apparues comme des alternatives au *backfitting* sont la projection itérative relaxée (*relaxed iterative projection*) [Schimek, 1996], technique également itérative adaptée aux cas proches de la singularité, l'intégration marginale (*marginal integration*) [Linton et Nielsen, 1995], pour les noyaux, et résolution directe pour les P–splines [Marx et Eilers, 1998, Ruppert, 2002]. Des procédures d'estimation ont été également proposées pour les modèles additifs ajustés par des ondelettes [Zhang et Wong, 2003, Sardy et Tseng, 2004].

La condition d'identifiabilité des modèles additifs implique que $\mathbb{E}_{\boldsymbol{X}}[Y] = \alpha_0$. L'estimation de la constante qui minimise l'erreur quadratique est la moyenne des observations : $\widehat{\alpha}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i$. Nous considérons ici que les observations de la variable réponse sont centrées, et donc $\widehat{\alpha}_0 = 0$.

1.3.5.1 Backfitting

La caractérisation des composantes additives de la solution (1.41) suggère un algorithme itératif pour l'estimation des fonctions unidimensionnelles $\mathbf{f}_j$. Pour une constante connue α_0 et pour des fonctions fixées $\mathbf{f}_k$, $k \neq j$, la fonction $\mathbf{f}_j$ peut être estimée à partir d'une régression unidimensionnelle sur les observations. Outre les méthodes de gradient, les deux procédure itératives standard pour la résolution des systèmes d'équations linéaires quand les systèmes sont non singuliers : Jacobi et Gauss–Seidel.

$$\begin{array}{rl} v_i^{[m]} = \left(b_i - \sum_{j=1, j\neq i}^{np} a_{ij} v_j^{[m-1]}\right) \big/ a_{ii} & \text{Jacobi} \\ v_i^{[m]} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} v_j^{[m]} - \sum_{j=i+1, j\neq i}^{np} a_{ij} v_j^{[m-1]}\right) \big/ a_{ii} & \text{Gauss–Seidel,} \end{array} \tag{1.51}$$

où $v_i^{[m]}$ est la solution itérative, $i = 1, \ldots, np$, m le nombre de l'itération.

La différence entre les deux algorithmes est que Jacobi calcule les estimations à l'itération $[m]$ en utilisant les estimations de l'itération $[m-1]$. Les estimations générées pendant l'itération $[m]$ ne seront utilisées que pour les estimations de l'itération $[m+1]$. La procédure Gauss–Seidel utilise toute l'information continûment actualisée.

Cette approche itérative appliquée à la régression non paramétrique multidimensionnelle est connue sous le nom de *backfitting*. L'idée est de déterminer les estimations des variables explicatives successivement, en profitant de la structure du problème d'estimation (1.43). Le principe du backfitting est d'ajuster les composantes $\mathbf{f}_j$ séparément aux résidus partiels, $\mathbf{r}_j = \mathbf{y} - \alpha_0 - \sum_{k\neq j} \mathbf{f}_k(\mathbf{x}_k)$. Cela conduit à une partition de la matrice $\mathbf{A}$ du système en p blocs de taille $n \times np$. Chaque bloc correspond à une variable explicative. La procédure de Gauss–Seidel est alors appliquée à ces blocs. Le résultat sont p vecteurs de taille n : $\mathbf{v}^{[M]}$, où M indique la dernière itération.

Description

L'algorithme backfitting comporte 3 étapes : 1. initialisation des fonctions, 2. estimation de chacune des fonctions à partir des résidus partiels, et 3. itération de l'étape 2. jusqu'à la convergence.

1. *Initialisation*

Si aucune information n'est connue sur les fonctions, les $\mathbf{f}_j$ sont normalement initialisées à la fonction $\mathbf{0}$ ou à la régression linéaire de $\mathbf{y}$ sur $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$: $\widehat{\mathbf{f}}_j = \mathbf{f}_j^{[0]}$.

2. *Estimation*

La fonction

$$f_j(x) = \mathbb{E}\left(Y - \alpha_0 - \textstyle\sum_{k\neq j} f_k(X_k) | X_j = x\right)$$

est estimée par

$$\widehat{\mathbf{f}}_j = \mathbf{S}_j \left(\mathbf{y} - \sum_{k \neq j} \widehat{\mathbf{f}}_k(\mathbf{x}_k) \right).$$

Les points $\{(x_{i1}, \ldots, x_{ij-1}, x_{ij+1}, \ldots, x_{ip}, y_i)\}_{i=1}^n$ sont utilisés pour le calcul de $(\mathbf{y} - \sum_{k \neq j} \widehat{\mathbf{f}}_k)$, et $\{x_{ij}\}_{i=1} n$ sont utilisés pour le calcul de $\mathbf{S}_j$. Ainsi, pour le calcul de chaque f_j, la dépendance de Y sur toutes les variables à exception de X_j est "éliminée".

Pour l'itération $[l]$ de la boucle, les principes de Jacobi ou de Gauss–Seidel peuvent être appliqués :

$$\begin{aligned} \widehat{\mathbf{f}}_j^{[l]} &= \mathbf{S}_j \left(\mathbf{y} - \textstyle\sum_{k=1}^{p} \widehat{\mathbf{f}}_k^{[l-1]}(\mathbf{x}_k) \right) && \text{Jacobi} \\ \widehat{\mathbf{f}}_j^{[l]} &= \mathbf{S}_j \left(\mathbf{y} - \textstyle\sum_{k=1}^{j-1} \widehat{\mathbf{f}}_k^{[l]}(\mathbf{x}_k) - \sum_{k=j+1}^{p} \widehat{\mathbf{f}}_k^{[l-1]}(\mathbf{x}_k) \right) && \text{Gauss–Seidel} \end{aligned} \tag{1.52}$$

Par application de Jacobi, l'estimation est réalisée à partir de la valeur des $\mathbf{f}_k$ de l'itération précédant. Cette approche est appliquée par [Opsomer et Ruppert, 1998]. Par l'application d'une méthode Gauss–Seidel, $\mathbf{f}_j^{[l]}$ est estimée en prenant en compte la dernière mise à jour des résidus partiels.

3. *Convergence*

L'étape 2 (estimation) est itérée jusqu'à la convergence. Les critères de convergence ne sont pas explicités dans la littérature. La description des critères d'arrêt se limitent à : "jusqu'à que les fonctions (globales ou individuelles) ne changent pas d'une itération à la suivante", ou "jusqu'à que la valeur de la somme des carrés résiduels,

$$\left\| \mathbf{y} - \sum_{j=1}^{p} \widehat{\mathbf{f}}_j(\mathbf{x}_j) \right\|_2^2, \tag{1.53}$$

soit plus petite qu'une certaine tolérance".

Le premier critère est utilisé par [Opsomer et Ruppert, 1998] :

$$\sum_{j=1}^{p} \frac{1}{n} \left\| \widehat{\mathbf{f}}_j^{[l]} - \widehat{\mathbf{f}}_j^{[l-1]} \right\|_2^2 < 10^{-3}. \tag{1.54}$$

Ce critère tient compte de la différence entre deux itérations de chaque fonction $\widehat{\mathbf{f}}_j$. SAS utilise les deux critères. La procédure "GAM" arrête le cycle quand la "valeur de la somme des carrés résiduels ne diminue pas" ou quand la différence relative est inférieure à un seuil :

$$\frac{\left\| \widehat{\mathbf{f}}^{[l]} - \widehat{\mathbf{f}}^{[l-1]} \right\|_2^2}{\left\| \widehat{\mathbf{f}}^{[l-1]} \right\|_2^2} < 10^{-8}. \tag{1.55}$$

Ce critère tient compte de la différence relative entre deux itérations de la fonction $\widehat{\mathbf{f}} = \sum_j \widehat{\mathbf{f}}_j$. La procédure "GAM" de S–plus utilise le premier critère. Une tolérance de 10^{-7}, avec un nombre maximal d'itérations égal à 30, est conseillée par

[Chambers et Hastie, 1993].

Propriétés

L'estimateur backfitting n'ayant pas d'expression analytique, des questions concernant la convergence et l'unicité des solutions, le comportement de l'algorithme, ainsi que les propriétés de l'estimateur ont été étudiées.

Convergence

La convergence et l'unicité ne sont garanties que dans certains cas. En particulier, des problèmes numériques peuvent être rencontrés en présence de concurvité.

Sous certaines conditions, les équations normales (1.43)–(1.44) sont consistantes (il existe au moins une solution) [Schimek et Turlach, 2000] :
Si $\forall j,\ j = 1, \ldots, p$, la matrice de lissage linéaire, $\mathbf{S}_j$, est symétrique et ses valeurs propres sont comprises dans l'intervalle $[0, 1]$, alors les équations normales sont consistantes pour tous les $\mathbf{y}$.

Cependant, pour que cette solution soit unique d'autres conditions sont nécessaires. Pour le cas $p = 2$, la solution des équations (1.50) est unique, et l'algorithme backfitting converge à cette solution, si $\|\mathbf{S}_1\mathbf{S}_2\| < 1$, où $\|.\|$ est une norme matricielle quelconque (les conditions de symétrie et que les valeurs propres soient comprises à l'intervalle $]-1, 1]$ sont suffisantes pour la consistance, dans la cas $p = 2$) [Buja *et al.*, 1989, Hastie et Tibshirani, 1990, Schimek et Turlach, 2000].

Pour $p > 2$, la condition devient [Opsomer, 2000, Schimek et Turlach, 2000] :

$$\max_{\delta \in [2,p]} \left|\left| \sum_{j=1}^{\delta-1} \mathbf{S}_\delta \mathbf{S}_j \right|\right| < 1. \tag{1.56}$$

Les splines et les projections satisfont ces conditions. En absence de concurvité, le backfitting converge vers la solution unique, indépendamment des valeurs initiales [Hastie et Tibshirani, 1990, Schimek, 2000].

Un exemple de convergence problématique en présence de concurvité est donné par [Dominici *et al.*, 2002]. Dans cette étude il a été observé que la convergence n'est pas assurée, et les estimations, ainsi que les écart–types, peuvent être biaisés quand il existe des problèmes de concurvité. Aussi, la convergence est lente quand le paramètre de lissage est petit.

D'autres études concernant la convergence de l'algorithme, avec des hypothèses moins sévères sont celles de [Härdle et Hall, 1993] pour des projections et [Ansley et Kohn, 1994] pour les splines.

L'estimateur backfitting

Les expressions de l'espérance de l'erreur quadratique, (ainsi que du biais et de la variance asymptotiques), ont été rapportées pour des polynômes locaux par [Opsomer et Ruppert, 1997]. Cependant, les hypothèses sur les matrices de lissage sont sévères (par exemple, l'indépendance des variables explicatives est exigée). Dans

une autre étude, la théorie des projections additives a été utilisée pour obtenir la convergence uniforme des polynômes locaux et des noyaux [Mammen *et al.*, 1999]. Les conditions sont plus faibles que dans l'étude précédant. En particulier, l'indépendance entre les variables explicatives n'est pas exigée.

La normalité asymptotique de l'estimateur backfitting pour des polynômes locaux a été déduite [Wand, 2000].

Il existe également une version bayésienne de l'algorithme backfitting qui bénéficie de l'approche bayésienne aux splines de lissage [Hastie et Tibshirani, 2000].

Problèmes numériques

Afin de garantir l'unicité et d'éviter la singularité de $\mathbf{A}$ (1.44), il est nécessaire de centrer les estimations et accumuler dans la constante les correspondantes déviations : $\mathbf{S}_j^* = (\mathbf{I} - \mathbf{1}\mathbf{1}^t/n)\mathbf{S}_j$. Cette transformation de la matrice de lissage assigne à la constante la valeur propre égale à 0.

L'efficacité de l'algorithme backfitting peut être améliorée si en plus du centrage, l'ensemble des projections sont calculés à part [Hastie et Tibshirani, 1990]. Dans plusieurs matrices de lissage, deux parties peuvent être différenciées : une projection et un "rétrécissement". Par exemple, les splines cubiques de lissage ont deux valeurs propres égales à 1 correspondants aux fonctions constantes et linéaires (projection), et les autres valeurs propres sont comprises strictement entre 0 et 1 (rétrécissement). En pratique, l'idée est de combiner les opérations correspondants aux projections de toutes les variables explicatives dans une seule opération, et utiliser seulement les parties de rétrécissement de chaque matrice de lissage dans la partie itérative de l'algorithme.

1.3.5.2 Projection itérative relaxée

Cette technique est aussi inspirée des techniques de résolution de systèmes d'équations linéaires quand la matrice $\mathbf{A}$ est presque singulière (1.44) [Schimek, 1996]. Ces techniques introduisent un terme de relaxation afin d'améliorer la vitesse de convergence. Les itérations (1.51) deviennent alors :

$$\begin{aligned} \mathbf{v}_c^{[m]} &= \mathbf{v}^{[m-1]} - \omega\left(\mathbf{A}\mathbf{v}_c^{[m-1]} - \mathbf{b}\right) \qquad \text{Jacobi} \\ \mathbf{v}_c^{[m]} &= (1-\omega)\mathbf{v}^{[m-1]} + \omega\mathbf{v}^{[m]} \qquad \text{Gauss–Seidel,} \end{aligned} \tag{1.57}$$

où $[m]$ indique l'itération, et c indique le terme corrigé.

La projection itérative relaxée introduit également un terme de relaxation dans la procédure itérative. Soit $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_{np})$, les colonnes de $\mathbf{A}$.

$$\begin{aligned} \mu^{[m]} &= \frac{\langle \omega \mathbf{u}^{[m-1]}, \mathbf{a}_m \rangle}{\langle \mathbf{a}_m, \mathbf{a}_m \rangle} \\ \mathbf{u}^{[m-1]} &= \mathbf{b} - \sum_{k=1}^{m-1} \mu^{[k]} \mathbf{a}_k, \end{aligned} \tag{1.58}$$

$\mathbf{a}_m$ est la colonne $1 + (m-1)$ module np pour $m > np$. Il est démontré que $\mathbf{f}_j = \sum_m \mu^{[m]}$, $j = 1, \ldots, p$, $m = j + (np)k$, $k = 1, 2, \ldots$.

La projection itérative relaxée semble se comporter mieux que le backfitting pour des matrices du système singulières ou proches de la singularité. Il faut, par ailleurs, choisir de manière adéquate le terme de relaxation ω. Pour des valeurs $0 < \omega < 2$ il est possible d'établir la convergence de la relaxation.

1.3.5.3 Intégration marginale

Cette méthode est fondée sur des moyennages marginaux plutôt que sur la solution itérative d'un système d'équations [Linton et Nielsen, 1995].

Comme défini en (1.38), le modèle additif suppose que la fonction de régression inconnue est une somme de fonctions des variables explicatives : $f(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \alpha_0 + \sum_{j=1}^{p} f_j(x_j)$. Afin de garantir l'unicité, il est aussi supposé $\mathbb{E}_{X_j}[f_j(X_j)] = 0$, et donc, $\mathbb{E}[Y] = \alpha_0$. Ces hypothèses impliquent :

$$\begin{aligned} &\mathbb{E}_{X_{\underline{j}}}\left[\alpha_0 + \sum_{k=1}^{j-1} f_k(X_k) + f_j(x_j) + \sum_{k=j+1}^{p} f_k(X_k)\right] = \\ &\int f(\boldsymbol{x}) \boldsymbol{h}_{\underline{j}}(\boldsymbol{x}_{\underline{j}}) \prod_{k \neq j} dx_k = \alpha_0 + f_j(x_j), \end{aligned} \tag{1.59}$$

où $\boldsymbol{h}_{\underline{j}}$ est la densité conjointe de $X_{\underline{j}} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$.

D'autre part, il est également possible d'obtenir f_j pour une fonction de pondérations quelconque w, satisfaisant $\mathbb{E}\left[w(X_{\underline{j}})\right] = 1$:

$$\mathbb{E}\left[w(X_{\underline{j}}) f(X_1, \ldots, X_{j-1}, x_j, X_{j+1}, \ldots, X_p)\right] = f_j(x_j) + C_w, \tag{1.60}$$

où C_w est une constante indépendante de x_j. De façon similaire, la substitution en (1.59) de la densité conjointe par une fonction de densité sur $\mathbb{R}^{p-1}$ quelconque, $\boldsymbol{q}$, permet d'obtenir

$$\int f(\boldsymbol{x}) \boldsymbol{q}(\boldsymbol{x}_{\underline{j}}) \prod_{k \neq j} dx_k = f_j(x_j) + C_q. \tag{1.61}$$

Alors, pour estimer la fonction f_j, une possibilité consiste à estimer directement l'intégrale en (1.59) ou en (1.61), en remplaçant les fonctions par une estimation non paramétrique [Linton et Nielsen, 1995]. Une autre possibilité consiste à estimer l'espérance en (1.59) ou en (1.60), par application de la loi des grands nombres [Linton et Härdle, 1996, Linton, 1997]. Un estimateur pondéré peut également être considéré à partir de (1.60) [Fan *et al.*, 1998].

Pour la deuxième approche, [Linton et Härdle, 1996] proposent l'estimateur

$$\widehat{f}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{f}}(X_{i,1},\ldots,X_{i,j-1},x_j,X_{i,j+1},\ldots,X_{i,p}) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{n}\mathcal{K}_{\boldsymbol{\lambda}_{\underline{j}}}(X_{i\underline{j}}-X_{k\underline{j}})K_{\lambda_j}(X_{kj}-x_j)y_k}{\sum_{l=1}^{n}\mathcal{K}_{\boldsymbol{\lambda}_{\underline{j}}}(X_{i\underline{j}}-X_{l\underline{j}})K_{\lambda_j}(X_{lj}-x_j)} = \tag{1.62}$$

$$\frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{n}\frac{\mathcal{K}_{\boldsymbol{\lambda}_{\underline{j}}}(X_{i\underline{j}}-X_{k\underline{j}})K_{\lambda_j}(X_{kj}-x_j)}{\sum_{l=1}^{n}\mathcal{K}_{\boldsymbol{\lambda}_{\underline{j}}}(X_{i\underline{j}}-X_{l\underline{j}})K_{\lambda_j}(X_{lj}-x_j)}\right)y_k,$$

où K_{λ_j} est un noyau à support fini, et $\mathcal{K}_{\boldsymbol{\lambda}_{\underline{j}}}$ un noyau multidimensionnel (1.31) à support fini.

L'estimateur de l'intégration marginale a ainsi des expressions explicites, ce qui permet, contrairement au backfitting, d'étudier ses propriétés, et de développer une théorie asymptotique de la distribution. Cependant, cette méthode peut montrer des difficultés d'application quand le nombre de variables est élevée. Des modifications et des extensions (comme par exemple l'inclusion d'une itération de backfitting), ont été proposées afin d'améliorer l'algorithme [Kim *et al.*, 1999, Mammen *et al.*, 1999, Linton et Nielsen, 2000].

1.3.5.4 Résolution directe pour les P–splines

Quand l'application des P–splines permet une réduction importante du rang des matrices de lissage, la résolution directe du système (1.50) est possible [Marx et Eilers, 1998]. Cependant, les calculs peuvent s'avérer lents et numériquement instables.

Des algorithmes ont été proposés afin d'assurer des simplifications et d'éviter ces problèmes. Ces algorithmes intègrent des diagonalisations [Ruppert, 2002], des factorisations QR [Gu et Wahba, 1991, Wood, 2000, Ruppert *et al.*, 2003], ainsi que des décompositions de Choleski et des décompositions en valeurs singulières [Wood, 2004]. Quelques–uns de ces algorithmes sont applicables à différentes bases de fonctions splines, d'autres sont spécifiques à certaines bases.

Algorithme de diagonalisation

Considérons tout d'abord le cas unidimensionnel. Soient $\mathbf{N}$ la matrice de la base choisie évaluée en $\{x_{i1}\}_{i=1}^{n}$ et $\boldsymbol{\Omega}$ une matrice de pénalisation quelconque, symétrique, semi–définie positive. Soit $\mathbf{B}$ une matrice carrée satisfaisant $\mathbf{B}^{-1}\mathbf{B}^{-t} = \mathbf{N}^t\mathbf{N}$ (par exemple, la décomposition de Cholesky de $\mathbf{N}^t\mathbf{N}$). Soit $\mathbf{U}$ une matrice orthogonale et $\mathbf{D}$ une matrice diagonale satisfaisant $\mathbf{U}\mathbf{D}\mathbf{U}^t = \mathbf{B}\boldsymbol{\Omega}\mathbf{B}^t$. Finalement, notons $\mathbf{Z} = \mathbf{N}\mathbf{B}^t\mathbf{U}$ et $\widehat{\boldsymbol{\gamma}} = \mathbf{U}^t\mathbf{B}^{-t}\widehat{\boldsymbol{\beta}} = (\mathbf{B}^t\mathbf{U})^{-1}\widehat{\boldsymbol{\beta}}$.

Alors, $\widehat{\boldsymbol{\gamma}}$ résout le système diagonal :

$$(\mathbf{I}+\lambda\mathbf{D})\widehat{\boldsymbol{\gamma}} = \mathbf{Z}^t\mathbf{y} = (\mathbf{U}^t\mathbf{B})\mathbf{N}^t\mathbf{y}. \tag{1.63}$$

D'autre part, $\mathbf{N}\widehat{\boldsymbol{\beta}} = \mathbf{Z}\widehat{\boldsymbol{\gamma}}$, ce qui implique que la matrice de lissage est $\mathbf{S}(\lambda) = \mathbf{Z}(\mathbf{I} + \lambda\mathbf{D})^{-1}\mathbf{Z}^t$.

Dans le cas additif, des matrices carrées, $\mathbf{B}_j$, $j = 1 \ldots, p$, satisfont

$$\mathbf{B}_j^{-1}\mathbf{B}_j^{-t} = (\mathbf{N}_1^t, \ldots, \mathbf{N}_p^t) \begin{pmatrix} \mathbf{N}_1 \\ \vdots \\ \mathbf{N}_p \end{pmatrix} + \sum_{k \neq j} \lambda_k \boldsymbol{\Omega}_k. \tag{1.64}$$

Alors, $\widehat{\boldsymbol{\beta}}_j$ est obtenu par résolution de

$$(\mathbf{B}_j^{-1}\mathbf{B}_j^{-t} + \lambda_j \boldsymbol{\Omega}_j)\widehat{\boldsymbol{\beta}}_j = (\mathbf{N}_1^t, \ldots, \mathbf{N}_p^t) \begin{pmatrix} \mathbf{y} \\ \vdots \\ \mathbf{y} \end{pmatrix}. \tag{1.65}$$

1.3.5.5 Comparaison des algorithmes

Les estimateurs intégration marginale et backfitting ont été comparés d'un point de vue théorique par [Nielsen et Linton, 1998]. L'estimateur intégration marginale est plus facile à interpréter que l'estimateur backfitting, car le premier est simplement obtenu par des pondérations, alors que le deuxième est la solution itérative des équations non linéaires. Pour le premier, les propriétés statistiques sont trivialement obtenues, cependant, en général, il n'est pas efficace. Des améliorations de l'algorithme passent par l'application d'une itération de backfitting [Linton, 1997].

Les deux procédures peuvent être considérées comme minimisant un critère basé sur l'intégrale de l'erreur quadratique :

$$\begin{aligned} &\min_{g_1 \in G_1, g_2 \in G_2, c \in \mathbb{R}} I(g_1, g_2, c) = \\ &\min_{g_1 \in G_1, g_2 \in G_2, c \in \mathbb{R}} \int\int \left[f(x_1, x_2) - g_1(x_1) - g_2(x_2) - c\right]^2 dh_{X_1, X_2}(x_1, x_2), \end{aligned} \tag{1.66}$$

où $h_{X_1,X_2}(x_1, x_2)$ probabilité conjointe, f fonction de régression additive, et G_1 et G_2 des espaces de fonctions monovariées centrées, tels que $\mathbb{E}_{X_1}[g_1(X_1)] = 0$, $\mathbb{E}_{X_2}[g_2(X_2)] = 0$. Soit $\widehat{f}$ un estimateur initial, le critère empirique est

$$\widehat{I}(g_1, g_2, c) = \int\int [\widehat{f}(x_1, x_2) - g_1(x_1) - g_2(x_2) - c]^2 dh(x_1, x_2). \tag{1.67}$$

Le backfitting optimise le critère avec des pondérations issues d'une densité empirique conjointe, ce qui correspond à minimiser l'intégrale de l'erreur quadratique moyenne. L'intégration marginale optimise le cas où la pondération est réalisée par rapport à la densité produit. Elle perd en efficacité quand les variables explicatives ne sont pas indépendantes. Un autre inconvénient de l'intégration marginale est sa sensibilité au fléau de la dimensionnalité. Par conséquent, des hypothèses fortes sont nécessaires à l'obtention d'un taux de convergence optimal en dimension élevée. Aussi, des bons

résultats ne sont pas assurés pour des petits échantillons.

Comparaison par simulations

Des travaux ont comparé les algorithmes. Les procédures backfitting, projection itérative relaxée et résolution directe pour les P–splines sont comparées par [Schimek, 2000]. Les conclusions obtenues sont les suivantes : dans une situation standard, l'utilisation du backfitting est conseillée ; s'il y a des raisons pour penser que le degré des splines devrait être supérieur à 3, alors la vraisemblance pénalisée est recommandée. Finalement, en présence de concurvité exacte, la projection itérative relaxée est la méthode qui montre le meilleur comportement. Dans les situations étudiées, le backfitting se comporte mieux que prévu, probablement d'après l'auteur, dû à une décomposition QR incorporée dans l'algorithme.

Le backfitting et l'intégration marginale ont été comparés par [Sperlich *et al.*, 1999]. Les auteurs concluent que le comportement des méthodes est très similaire. En particulier, quand la taille de l'échantillon est petite les deux méthodes n'obtiennent pas de bons résultats.

Le backfitting est plus performant que l'intégration marginale aux points des bords, en présence de corrélation entre les variables explicatives, et avec des données peu denses. L'estimateur de la fonction de régression $\widehat{\mathbf{f}}$ est, en général, meilleur (au sens de l'erreur quadratique empirique moyenne) que l'estimateur obtenu par l'intégration marginale. Cela peut être expliqué par le fait que l'estimateur du backfitting cherche, dans l'espace des modèles additifs, le meilleur ajustement de la variable réponse aux variables explicatives.

L'intégration marginale est plus performante quant à l'estimation des influences marginales de chaque composante, $\widehat{\mathbf{f}}_j$, spécialement en dimension $p > 2$. Cela peut être expliqué par le fait que l'estimateur de l'intégration marginale estime la fonction additive en intégrant sur les directions qui n'ont pas d'intérêt, donc cette méthode mesure l'influence marginale de chaque variable explicative.

1.4 Modèles additifs généralisés

Les modèles additifs généralisés sont une extension des modèles linéaires généralisés, permettant d'identifier et de décrire des effets non linéaires.

1.4.1 Modèles linéaires généralisés

La classe des modèles linéaires généralisés regroupe les modèles qui visent à exprimer l'espérance d'une variable de sortie en fonction d'une combinaison linéaire des variables d'entrée [Fahrmeir et Tutz, 2001].

Soient $(\boldsymbol{X}, Y) = (X_1, \ldots, X_p, Y)$ un vecteur de variables aléatoires et $(\boldsymbol{x}, y)$ une

Distribution	θ	$b(\theta)$	ϕ	$\mathbb{E}(y)$	$\text{var}(y)$
Normale $N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2	$\mu = \theta$	σ^2
Bernouilli $B(1, \pi)$	$\text{logit}(\pi)$	$\log(1 + e^{-\theta})$	1	$\pi = \dfrac{e^\theta}{(1 + e^\theta)}$	$\pi(1 - \pi)$
Poisson $P(\lambda)$	$\log(\lambda)$	e^θ	1	$\lambda = e^\theta$	λ
Gamma $G(\mu, \eta)$	$-1/\mu$	$-\log(-\theta)$	η^{-1}	$\mu = -1/\theta$	μ^2/η
Gaussienne inverse $IG(\mu, \sigma^2)$	$1/\mu^2$	$-(-2\theta)^{1/2}$	σ^2	$\mu = (-2\theta)^{-1/2}$	$\mu^3\sigma^2$

TAB. 1.1 – Paramètres des distributions de la famille exponentielle.

réalisation. Les modèles linéaires généralisés sont définis comme suit :

$$\begin{cases} h_Y(y; \theta; \phi) = exp\left(\dfrac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \text{densité de } Y \\ \mu = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = b^{(1)}(\theta) \\ g(\mu) = \nu = \alpha_0 + X_1\alpha_1 + \ldots, +X_p\alpha_p, \end{cases} \tag{1.68}$$

où h_Y est la fonction de densité de Y, issue de la famille exponentielle. Le paramètre θ est appelé paramètre naturel. Le paramètre de dispersion ϕ est un paramètre de nuisance intervenant, par exemple, lorsque la variance de la loi gaussienne est inconnue, mais égal à 1 pour les lois à un paramètre. La fonction g est le lien, et ν, le prédicteur linéaire. La fonction de lien, monotone et différentiable, exprime une relation fonctionnelle entre la composante aléatoire et le prédicteur linéaire. L'expression $b^{(m)}$ indique la dérivée d'ordre m de la fonction b.

La classe des modèles linéaires généralisés est caractérisée par trois composantes. La composante aléatoire identifie la distribution de probabilités de la variable à expliquer (parmi les distributions de la famille exponentielle : gaussienne, gaussienne inverse, Gamma, Poisson, binomiale,...). La composante déterministe du modèle est le prédicteur linéaire. La troisième composante exprime une relation fonctionnelle entre la composante aléatoire et le prédicteur linéaire, au moyen de la fonction de lien. Le tableau (1.1) montre les paramètres pour les distributions les plus usuelles de la famille exponentielle.

Dans les modèles additifs généralisés, le lien linéaire est remplacé par une fonction de lien additive : $g(\mu) = \alpha_0 + \sum_{j=1}^{p} f_j(X_j)$. Etudions ces modèles dans le cas concret du modèle logistique, pour lequel la distribution de probabilité de Y est binomiale (i.e. la distribution d'une variable binaire quelconque) et le lien est la fonction logit.

1.4.2 Modèle logistique

Le modèle de régression logistique est un outil standard en discrimination lorsque la compréhension de l'effet de chaque variable d'entrée sur la variable de sortie est un aspect crucial. Ce modèle permet de calculer la probabilité de survenue de l'événement auquel on s'intéresse, quand la valeur des variables explicatives est connue.

Considérons le vecteur aléatoire $(\boldsymbol{X}, Y) = (X_1, \ldots, X_p, Y)$, où Y est une variable binaire (codée 0–1), et la réalisation $\boldsymbol{x} = (x_1, \ldots, x_p)^t$. Le modèle de régression logistique s'écrit

$$\text{logit}[P(Y = 1|\boldsymbol{X} = \boldsymbol{x})] = \log \frac{P(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{1 - P(Y = 1|\boldsymbol{X} = \boldsymbol{x})} = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_p x_p. \tag{1.69}$$

Pour un problème de décision où l'objectif est de minimiser le taux d'erreur de classement (coût $\{0, 1\}$), la frontière de décision est alors définie par l'hyper–plan $\{\mathbf{x} \in \mathbb{R}^p | \alpha_0 + \sum_{j=1}^p \alpha_j x_j = 0\}$, et la relation inverse donne la probabilité *a posteriori*,

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\alpha_0 + \sum_{j=1}^p \alpha_j x_j)}{1 + \exp(\alpha_0 + \sum_{j=1}^p \alpha_j x_j)}, \tag{1.70}$$

qu'on reconnaît comme la fonction softmax des réseaux de neurones.

1.4.2.1 Modèle logistique additif

La simplicité du modèle logistique en fait, avec les arbres de décision, une des méthodes de discrimination les plus interprétables. Cependant, l'hypothèse de dépendance linéaire est souvent trop restrictive : dans le "monde réel" les effets sont généralement non linéaires. Le modèle logistique additif est une généralisation permettant d'identifier et de décrire les effets non linéaires. Il remplace chaque composante linéaire par une fonction plus générale :

$$\log \frac{P(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{1 - P(Y = 1|\boldsymbol{X} = \boldsymbol{x})} = \alpha_0 + f_1(x_1) + \ldots + f_p(x_p), \tag{1.71}$$

où les f_j sont des fonctions lisses. On retrouve le modèle additif avec des erreurs gaussiennes (1.38), avec la fonction logit comme variable réponse.

1.4.3 Estimation

Soient $\{(X_{i1}, \ldots, X_{ip}, Y_i)\}_{i=1}^n$ échantillon i.i.d. de $(\boldsymbol{X}, Y)$ de taille n et $\{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^n$ des réalisations. Notons $\mathbf{X}$ la matrice des observations des variables d'entrée et $\mathbf{y}$ le vecteur des sorties observées. Nous introduisons le problème d'estimation pour le modèle additif par celui du modèle linéaire, plus simple.

1.4.3.1 Modèle linéaire

L'estimation des paramètres $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$ est calculée en maximisant la log–vraisemblance du modèle linéaire généralisé. La log–vraisemblance du modèle

logistique s'écrit :

$$l(\boldsymbol{\alpha}) = \sum_{i=1}^{n} y_i \log P_i + (1 - y_i) \log(1 - P_i), \tag{1.72}$$

où $P_i = P(Y_i = 1 | X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip})$. La méthode d'optimisation standard pour résoudre le problème de maximisation de la log–vraisemblance est le *Fisher scoring*, basé sur un algorithme Newton–Raphson [Hastie et Tibshirani, 1990]. La connaissance des dérivées de premier et deuxième ordre sont alors nécessaires :

$$\frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = [\mathbf{1}\ \mathbf{X}]^t(\mathbf{y} - \mathbf{P}), \quad \frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^t} = -[\mathbf{1}\ \mathbf{X}]^t \mathbf{W} [\mathbf{1}\ \mathbf{X}], \tag{1.73}$$

où $\mathbf{P} = (P_1, \ldots, P_n)^t$, $\mathbf{W} = \text{diag}\,[P_1(1 - P_1), \ldots, P_n(1 - P_n)]$, et $[\mathbf{1}\ \mathbf{X}]$ indique la matrice des données précédée d'une colonne de uns, afin d'incorporer l'estimation de α_0. Le problème d'optimisation étant convexe en $\boldsymbol{\alpha}$, la maximisation de (1.72) consiste à résoudre un système de $p+1$ équations $\dfrac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \mathbf{0}$. Ces équations sont non linéaires en α et elles sont résolues itérativement jusqu'à obtention d'un point fixe. Cette mise à jour permet également s'écrire sous une forme légèrement différente :

$$\boldsymbol{\alpha}^{[k+1]} = \boldsymbol{\alpha}^{[k]} - \left(\frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^t} \bigg|_{\boldsymbol{\alpha}^{[k]}} \right)^{-1} \frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \bigg|_{\boldsymbol{\alpha}^{[k]}} = ([\mathbf{1}\ \mathbf{X}]^t \mathbf{W} [\mathbf{1}\ \mathbf{X}])^{-1} [\mathbf{1}\ \mathbf{X}]^t \mathbf{W} \mathbf{z}, \tag{1.74}$$

où $\mathbf{z} = [\mathbf{1}\ \mathbf{X}]\boldsymbol{\alpha}^{[k]} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{P})$, ($\mathbf{P}$, et donc $\mathbf{W}$ et $\mathbf{z}$, dépendent de $\boldsymbol{\alpha}$), et k indique l'itération en cours.

Sous cette forme, l'algorithme est appelé moindres carrés pondérés itératifs (IRLS), car à chaque itération, le problème résolu est équivalent à un problème de moindres carrés pondérés :

$$\boldsymbol{\alpha}^{[k+1]} = \arg\min_{\boldsymbol{\alpha}} \left|\left| \mathbf{W}^{1/2} \big(\mathbf{z} - [\mathbf{1}\ \mathbf{X}]\boldsymbol{\alpha} \big) \right|\right|_2^2. \tag{1.75}$$

Cette analogie explique que $\mathbf{z}$ soit souvent dénommée "réponse de travail".

1.4.3.2 Modèle additif

Les premières étapes de résolution sont identiques pour les modèles additifs. L'estimation des paramètres α_j est généralisée à celle des fonctions f_j. Considérons le cas où les f_j sont des splines cubiques de lissage, définies maintenant d'une façon plus générale, comme la solution au problème de régularisation suivant : parmi les fonctions deux fois continûment dérivables, retenons celles minimisant la fonction coût (ici, la log–vraisemblance) [Wahba, 1990] :

$$\begin{aligned} &\min_{\alpha_0 \in \mathbb{R}, f_j \in \mathcal{C}^2} -l(\alpha_0, f_j) + \sum_{j=1}^{p} \lambda_j \int \left[f_j^{(2)}(t) \right]^2 dt = \\ &\min_{\alpha_0 \in \mathbb{R}, f_j \in \mathcal{C}^2} -\sum_{i=1}^{n} y_i \log P_i + (1 - y_i) \log(1 - P_i) + \sum_{j=1}^{p} \lambda_j \int \left[f_j^{(2)}(t) \right]^2 dt, \end{aligned} \tag{1.76}$$

où

$$P_i = \frac{\exp[\alpha_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip})]}{1 + \exp[\alpha_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip})]}. \tag{1.77}$$

De façon analogue au problème de type gaussien, le premier terme de (1.76) mesure l'ajustement aux données, et le deuxième terme pénalise les solutions de courbure forte. Les paramètres de lissage λ_j déterminent le compromis entre les deux objectifs. Les fonctions obtenues $\widehat{f_j}$ sont des splines cubiques en x_j, avec des nœuds sur les x_{ij}. Les contraintes $\sum_i \widehat{f_j}(\mathrm{x}_{ij}) = 0$, $j = 1, \ldots, p$ assurent l'unicité de la solution.

L'algorithme IRLS permet de résoudre (1.76), et le problème à minimiser pour estimer les fonctions f_j devient un problème quadratique pondéré :

$$\min_{\alpha_0 \in \mathbb{R}, f_j \in \mathcal{C}^2} \left|\left| \mathbf{W}^{1/2}\big(\mathbf{z} - \alpha_0 - \sum_{j=1}^{p} f_j\big) \right|\right|_2^2 + \sum_{j=1}^{p} \lambda_j \int \left[f_j^{(2)}(t)\right]^2 dt, \tag{1.78}$$

où la réponse de travail $\mathbf{z}$ est maintenant définie comme suit : $\mathbf{z} = \widehat{\mathbf{f}}^{[k]}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{P})$, où $\widehat{\mathbf{f}}^{[k]}(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \widehat{\alpha}_0^{[k]} + \widehat{\mathbf{f}}_1^{[k]}(\mathbf{x}_1) + \ldots + \widehat{\mathbf{f}}_p^{[k]}(\mathbf{x}_p)$.

Les procédures numériques analysées pour le cas gaussien (section 1.3.5, page 37), sont maintenant applicables à la résolution du problème (1.78).

1.4.4 D'autres extensions du modèle additif

Par rapport à la partie déterministe du modèle

Modèle semi–paramétrique

Un modèle semiparamétrique incluant les modèles (1.38) et (1.39) est donné par [Hastie et Tibshirani, 1990, Ruppert *et al.*, 2003] :

$$Y = \alpha_0 + \sum_{j=1}^{q} X_j \alpha_j + \sum_{j=q+1}^{p} f_j(X_j) + \varepsilon, \tag{1.79}$$

où ε est indépendant de $\mathbf{X}$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$; α_0, α_j, $j = 1, \ldots, q$ sont des constantes, et f_j, $j = q+1, \ldots, p$ sont des fonctions unidimensionnelles telles que $\mathbb{E}_{X_j}[f_j] = 0$.

Des méthodes d'estimation efficaces ont été proposées pour les modèles semi–paramétriques [Carroll *et al.*, 1997, Huang, 1999]. Aussi, des tests ont été déduits pour comparer les hypothèses H_0 : la composante est linéaire vs H_1 : la composante est lisse (voir section 2.4.4, page 73).

Modèle d'interaction

Le modèle (1.38) implique l'hypothèse de non interaction entre les variables explicatives. Des méthodes ont été proposées pour tester la pertinence de cette hypothèse [Eubank *et al.*, 1995, Chen *et al.*, 1995a, Sperlich *et al.*, 2002, Ruppert *et al.*, 2003,

Härdle *et al.*, 2004a]. Aussi, une généralisations des modèles additifs permet la prise en compte des interactions [Wahba, 1990, Hastie et Tibshirani, 1990] :

$$Y = \alpha_0 + \sum_{j=1}^{p} f_j(X_j) + \sum_{j<k} f_{j,k}(X_j, X_k) + \ldots + f_{1,\ldots,p}(X_1, \ldots, X_p) + \varepsilon, \quad (1.80)$$

où ε est indépendant de $\mathbf{X}$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$; α_0 est une constante, et les autres composantes additives sont des fonctions unidimensionnelles d'espérance (par rapport à chacun de leurs arguments) nulle. Cette décomposition peut être vue comme une version fonctionnelle de l'analyse de la variance [Chen, 1993, Gu, 2002]. Dans la pratique, l'interprétation et l'estimation des modèles incluant des interactions d'ordre élevé est largement plus difficile que des modèles incluant les effets principaux (termes additifs) et juste des interactions d'ordre inférieur.

Le modèle d'interaction peut être ajusté par des produits des bases de fonctions splines, MARS (*multivariate adaptive regression splines*) [Friedman, 1991].

Par rapport à la partie aléatoire du modèle

Les modèles additifs ont été généralisés aux situations où les erreurs sont corrélées ou hétéroscédastiques, par exemple dans le cadre des séries chronologiques [Kohn *et al.*, 2000, Fan, 2003], ou dans le cadre des mesures répétées [Gu, 2002] ou longitudinales [Martinussen et Scheike, 1999, Ruppert *et al.*, 2003].

D'autres extensions incluent les modèles additifs mixtes [Lin et Zhang, 1999, Fahrmeir et Tutz, 2001, Ruppert *et al.*, 2003] et les réponses multiples [Yee et Wild, 1996].

Par rapport au phénomène à modéliser

Les modèles de survie sont utilisés lorsque la variable de sortie Y est binaire (codée 0–1), et qu'on s'intéresse à la date de survenue de l'événement $Y = 1$ [Fahrmeir et Tutz, 2001]. Trois cas peuvent se produire. 1) Si Y_i est devenu 1 au cours de l'étude, le temps de participation ou de survie du i–ème sujet est le délai entre son entrée dans l'étude et la date de survenue de l'événement. 2) Si Y_i est resté 0 au cours de l'étude, le temps de survie du i–ème sujet est le délai entre l'entrée dans l'étude et la date fixée pour la fin de celle–ci. 3) Si le i–ème sujet est perdu de vue avant la date fixée pour la fin de l'étude, alors que Y_i était encore 0, son temps de survie est le délai entre l'entrée dans l'étude et la date de ses dernières nouvelles. On dit que les informations concernant ce sujet sont censurées.

La forme des données est la suivante : $\{(x_{i1}, \ldots, x_{ip}, t_i, \delta_i)\}_{i=1}^{n}$, où t_i indique le temps de survie, et la variable binaire δ_i indique si l'information concernant le i–ème sujet est complète ou censurée. Une quantité d'intérêt dans ce contexte est le risque de survenue de l'événement à l'instant t, noté $h(t)$, qui est lié à la probabilité de survie au–delà du temps t : $s(t) = \exp\big(-\int_0^t h(u)du\big)$.

Un des modèles de survie des plus employés est le modèle de Cox, aussi connu sous le nom de modèle des risques proportionnels (*proportional hazards model*), car l'association entre les facteurs de risque potentiels et la survenue de l'événement est

supposée constante au cours du temps :

$$h(t|X_1 = x_1, \ldots, X_p = x_p) = h_0(t) \exp\Big(\sum_{j=1}^{p} x_j \beta_j\Big). \tag{1.81}$$

La forme de la fonction h_0 (la valeur de base) n'est pas précisée, on ne peut donc pas évaluer le risque propre à un sujet, mais seulement le risque supplémentaire apporté par l'exposition à tel ou tel facteur de risque. Sous certaines conditions (période de suivi égal, événement rare, absence de censure), les paramètres du modèle logistique coïncident avec ceux du modèle de Cox.

L'extension au cas additif est immédiate [Hastie et Tibshirani, 1995] :

$$h(t|X_1 = x_1, \ldots, X_p = x_p) = h_0(t) \exp\Big(\sum_{j=1}^{p} f_j(x_j)\Big), \tag{1.82}$$

où f_j, $j = 1, \ldots, p$, sont des fonctions unidimensionnelles telles que $\mathbb{E}_{X_j}[f_j] = 0$. L'estimation du modèle de Cox repose sur la maximisation d'une log–vraisemblance partielle, de façon similaire aux modèles linéaires et additifs généralisés.

1.5 En bref

Dans ce premier chapitre nous avons situé la régression par modèles additifs dans le cadre de la régression non paramétrique multidimensionnelle. Premièrement, nous avons étudié l'estimation des fonctions monovariées quand leur complexité est fixée. Deuxièmement, nous avons traité l'estimation des modèles additifs. Dans cette section nous justifions nos choix parmi les techniques existantes.

Nous avons accordé plus d'attention aux méthodes splines qu'aux méthodes à noyaux. Les splines de lissage et les P–splines montrent des bonnes propriétés d'adaptabilité, mais surtout, le fait que ces méthodes émergent également d'un problème de minimisation d'une fonction de coût pénalisée, nous semble particulièrement attractif. Ceci permettra la généralisation aux modèles additifs des méthodes de sélection de variables pour les modèles linéaires basées sur une pénalisation (chapitre 3).

Nous avons considéré la modélisation de phénomènes réguliers ($\mathcal{C}^2$). et nous avons choisi des splines cubiques, qui aboutissent à des estimations ayant l'allure recherchée. D'autres splines très employées sont les splines linéaires, qui constituent une bonne approximation des phénomènes admettant des changements brusques de direction. L'adaptation des méthodes développées est dans ce cas directe.

Quant aux bases de fonctions splines, nous avons considéré initialement la base des polynômes par morceaux, plus facile à comprendre. Cependant, son instabilité numérique nous a conduit à l'utilisation des B–splines.

En ce qui concerne les procédures numériques pour les modèles additifs, nous avons prêté une attention spéciale au backfitting. Cette méthode est efficace en pratique.

La projection itérative relaxée, méthode également itérative, est rarement utilisée. Elle semble améliorer les performances du backfitting seulement dans des situations très spécifiques.

L'intégration marginale est une technique plutôt adressée aux méthodes noyaux et sa performance est perturbée en dimension élevée ou même modérée. Ces deux facteurs ont déterminé notre désintérêt.

Quant à la résolution directe pour les P–splines, les premiers travaux ne proposent pas de méthodes numériques avantageuses. Initialement, elles ne sont pas adressées à des échantillons de taille élevée. Cette méthode devient intéressante lorsque des algorithmes intégrant des décompositions (QR, en valeurs singulières, ou Choleski) sont proposés, notamment par [Wood, 2000, Ruppert *et al.*, 2003, Wood, 2004]. Son invention récente est la principale cause de son omission. Nous avons évalué le comportement de notre méthode quand des P–splines associées au backfitting sont appliquées à l'estimation, mais il resterait à évaluer le comportement quand des P–splines associées à la résolution directe sont utilisées. Cette approche offre, en effet, de nouvelles perspectives.

En ce qui concerne l'application de l'algorithme backfitting, nous avons utilisé l'approche qui prend en compte la dernière mise à jour des résidus partiels et nous avons adopté le critère d'arrêt suivant :

$$\max_{j=1,\ldots,p} \frac{\left\| \widehat{\mathbf{f}}_j^{[l]}(\mathbf{x}_j) - \widehat{\mathbf{f}}_j^{[l-1]}(\mathbf{x}_j) \right\|_2^2}{1 + \left\| \widehat{\mathbf{f}}_j^{[l]}(\mathbf{x}_j) \right\|_2^2} < 10^{-5}, \tag{1.83}$$

où $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^t$, avec un nombre maximal d'itérations égal à 30. L'ajout de 1 dans le dénominateur nous permet d'éviter des problèmes numériques. Si $\left\| \widehat{\mathbf{f}}_j^{[l]}(\mathbf{x}_j) \right\|_2^2$ est faible, le quotient est équivalent à la différence absolue, si cette valeur est importante, le quotient est équivalent à la différence relative.

Chapitre 2

Complexité

2.1 Introduction

La sélection de modèle consiste à déterminer la structure du modèle la plus adaptée aux données. Dans le cadre des modèles additifs avec des observations i.i.d., elle comporte deux sous–problèmes : la sélection de variables et la sélection de la complexité. Le premier consiste à sélectionner le groupe de variables d'entrée les plus prédictives de la variable de sortie [Miller, 1990, Bi *et al.*, 2003]. Le deuxième aborde la question "quelle est la bonne proportion de lissage ?" [Härdle, 1990, Hastie et Tibshirani, 1990]. Dans ce chapitre, nous passons revue du problème de la sélection de la complexité.

Le paramètre de lissage introduit un ordre entre les modèles, lesquels s'étendent du plus simple (une droite, pour les splines cubiques, un polynôme global, pour les polynômes locaux), au plus complexe (l'interpolation des données). La somme des carrés résiduels sur les données d'apprentissage ne proportionne donc pas un bon critère pour estimer les paramètres de lissage : ce critère sélectionnerait dans tous les cas l'interpolation, dont les résidus sont nuls, tandis que ce modèle n'est pas approprié pour la prédiction de nouvelles données.

Des critères plus adaptés à ce problème se basent sur l'erreur en prédiction (ou erreur de généralisation) [Hastie *et al.*, 2001]. Le développement de ces critères permet de mettre en évidence que deux termes sont essentiels : le biais et la variance. Le terme du biais correspond à la différence au carré entre la vraie fonction de régression et l'estimation. Ce terme décroît quand la complexité du modèle augmente. Le terme de la variance correspond à la variance de l'estimation, qui augmente quand la complexité du modèle augmente. Les relations opposées se produisent quand la complexité diminue. La sélection de la complexité du modèle est ainsi un problème difficile, qui demande à trouver le bon compromis entre le biais et la variance.

Cette difficulté explique, en partie, que dans certaines applications, les techniques non paramétriques sont difficilement acceptées par l'utilisateur final [Sperlich, 2003]. Des limites concernent la sélection et l'interprétabilité des paramètres de lissage, ainsi que l'automatisation des techniques.

Bien que des aspects fondamentaux tels que la flexibilité (capacité du modèle de fournir des ajustements précis dans un vaste éventail de situations), et le traitement en

dimension élevée (par des techniques basées sur une réduction de la dimension) soient respectés par des modèles non paramétriques, des aspects tels que l'automatisation ou l'interprétabilité (capacité de rendre compréhensible la structure sous–jacente) constituent, en effet, des problèmes ouverts [Sperlich, 2003].

Ces problèmes s'aggravent quand la dimension augmente.

2.2 Nombre de degrés de liberté

Dans les statistiques paramétriques, la notion de degrés de liberté joue un rôle important admettant plusieurs lectures [Ye, 1998]. La complexité d'un modèle ajusté par moindres carrés est mesurée par les degrés de liberté, lesquels correspondent au nombre de paramètres (supposant que la matrice des données est non singulière). En particulier, dans le cas linéaire, la complexité est directement liée à la dimension de l'espace engendré et, donc, au nombre de variables d'entrée, p.

De façon plus générale, on peut obtenir le nombre de degrés de liberté comme la trace de la matrice chapeau, qui est la matrice $\mathbf{H}$, indépendante des $\mathbf{y}$, telle que $\widehat{\mathbf{f}} = \mathbf{H}\mathbf{y}$. Dans le cas linéaire, supposant que la matrice $\mathbf{X}$ est non singulière :

$$\text{ddl} = \text{tr}\,(\mathbf{H}) = \text{tr}\left(\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right) = \text{tr}\,(\mathbf{I}_p) = \text{rang}(\mathbf{X}) = p, \tag{2.1}$$

ou, afin d'incorporer l'estimation de la constante :

$$\begin{aligned} \text{ddl} = \quad & \text{tr}\,(\mathbf{H}) = \text{tr}\left([\mathbf{1}\ \mathbf{X}]([\mathbf{1}\ \mathbf{X}]^t[\mathbf{1}\ \mathbf{X}])^-[\mathbf{1}\ \mathbf{X}]^t\right) = \text{rang}(\mathbf{X}) + 1 = \\ & \text{tr}\left(\mathbf{H}^{\text{C}} + \mathbf{H}^{\text{LI}}\right) = \text{tr}\left(\frac{1}{n}\mathbf{1}\mathbf{1}^t\right) + \text{tr}\left(\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right) = p + 1, \end{aligned} \tag{2.2}$$

où $[\mathbf{1}\ \mathbf{X}]$ indique la matrice des données précédée d'une colonne de uns, $\mathbf{H}^{\text{C}}$ est la matrice $n \times n$ telle que tous les éléments sont égaux à $1/n$, elle agit sur la partie constante, et $\mathbf{H}^{\text{LI}}$ est la matrice chapeau du problème précédant, elle agit sur la partie linéaire.

Le nombre de degrés de liberté correspond ainsi à la somme de la sensibilité de chaque valeur ajustée par rapport à la valeur observée correspondante. Ce nombre représente également le coût de la procédure d'estimation, et donc il peut être utilisé pour obtenir des estimateurs non–biaisés de la variance de l'erreur. Aussi, cette quantité permet la comparaison de différents modèles.

L'interprétation initiale du nombre de degrés de liberté comme nombre de paramètres n'est plus satisfaisante quand le critère d'ajustement est modifié. Par exemple, elle ne tient pas compte des contraintes sur les paramètres quand ces derniers sont pénalisés et elle n'est pas directement transférable au contexte non paramétrique. En revanche, la notion de complexité d'un modèle en termes de la trace de la matrice chapeau est facilement généralisable aux méthodes de lissage linéaires.

2.2.1 Régression non paramétrique unidimensionnelle

Dans le contexte non paramétrique, le paramètre de lissage contrôle la complexité du modèle, mais son interprétation dépend de la méthode utilisée (par exemple,

le paramètre qui contrôle la pénalisation pour les splines, ou la largeur de bande pour les polynômes locaux) ainsi que de la formulation du problème (par exemple, la formulation des splines en termes de vraisemblance pénalisée ou en termes de problème d'optimisation sous contraintes). La généralisation du nombre de degrés de liberté à la régression non paramétrique permet la comparaison de différents modèles en termes de complexité.

Les lissages linéaires s'écrivent sous la forme $\widehat{f}(x) = \sum_i^n w(x, x_{i1})y_i$, ils sont donc linéaires vis à vis du vecteur d'observations (voir section 1.2.1, page 15). Par analogie au nombre de paramètres du modèle linéaire, le nombre effectif de paramètres ou nombre de degrés de liberté est défini comme la trace de la matrice de lissage[1] [Hastie et Tibshirani, 1990], et s'exprime simplement comme :

$$\text{ddl} = \text{tr}(\mathbf{S}_\lambda) = \sum_i w(x_{i1}, x_{i1}). \tag{2.3}$$

Cette somme correspond exactement au nombre de paramètres pour les modèles ajustés par moindres carrés. Chacun des éléments $w(x_{i1}, x_{i1})$ mesure la contribution de y_i dans le calcul de $\widehat{f}(x_{i1})$. Pour un apprentissage par coeur $w(x_{i1}, x_{i1}) = 1$, et le nombre effectif de paramètres est égal à la taille de l'échantillon. La notion de nombre effectif de paramètres généralise ainsi la mesure de complexité à l'ensemble des méthodes de lissage linéaires. Elle est moins générale que la dimension de Vapnik–Chervonenkis [Vapnik, 1995], en revanche, elle est facilement calculable.

La définition (2.3) implique que les ddl sont également la somme des valeurs propres de $\mathbf{S}_\lambda$. Ainsi, par exemple, les splines cubiques, dont la matrice de lissage a 2 valeurs propres égales à 1, correspondantes aux fonctions constantes et linéaires, et $n-2$ valeurs propres dans l'intervalle $[0, 1[$, correspondantes aux fonctions d'ordre supérieur, vérifient $2 \leq \text{ddl} \leq n$. La valeur minimale 2 est obtenue dans le cas le plus simple, quand le problème réduit à la régression linéaire ($\text{ddl}(\lambda = \infty) = 2$). La valeur maximale n est obtenue dans le cas le plus complexe, l'interpolation ($\text{ddl}(\lambda = 0) = n$). La relation entre λ et ddl est ici décroissante.

Afin de ne pas tenir compte de la valeur propre correspondant aux fonctions constantes, la définition suivante est parfois utilisée :

$$\text{ddl} = \text{tr}(\mathbf{S}^*_\lambda) = \text{tr}(\mathbf{S}_\lambda) - 1, \tag{2.4}$$

où $\mathbf{S}^*_\lambda = (\mathbf{I} - \mathbf{1}\mathbf{1}^t/n)\mathbf{S}_\lambda$, est la matrice de lissage centrée.

D'autres définitions sont celle des degrés de liberté de l'erreur :

$$\text{ddl}^{\text{err}} = n - \text{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}^t_\lambda), \tag{2.5}$$

ce qui dans le cas linéaire correspond à $n - p$, car l'espérance de la somme des carrés résiduels, RSS (*Residual Sum of Squares*) admet la factorisation suivante :

$$\mathbb{E}[\text{RSS}] = \mathbb{E}\left[\sum_{i=1}^n \big(y_i - \widehat{f}_\lambda(x_{i1})\big)^2\right] = \Big[n - \text{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}^t_\lambda)\Big]\sigma^2 + \mathbf{b}^t_\lambda\mathbf{b}_\lambda, \tag{2.6}$$

[1]Notons la matrice de lissage $\mathbf{S}$ ou, afin de souligner sa dépendance vis à vis du paramètre de lissage, $\mathbf{S}_\lambda$ (ou encore $\mathbf{S}(\lambda)$). Notons également l'estimation de f par $\widehat{f}$, $\widehat{f}_\lambda$ ou $\widehat{f}(\lambda)$.

où $\mathbf{b}_\lambda = \mathbf{f} - \mathbb{E}[\mathbf{S}_\lambda \mathbf{y}] = \mathbf{f} - \mathbf{S}_\lambda \mathbf{f}$ est le biais.

Les degrés de liberté de la variance sont définis :

$$\text{ddl}^{\text{var}} = \text{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda^t), \tag{2.7}$$

puisque dans le cas linéaire $\sum_{i=1}^n \text{Var}\left[\widehat{f}_\lambda(x_{i1})\right] = p\sigma^2$ et, pour les méthodes de lissage linéaires, $\sum_{i=1}^n \text{Var}\left[\widehat{f}_\lambda(x_{i1})\right] = \text{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda^t)\sigma^2$.

Si $\mathbf{S}_\lambda$ est une projection symétrique, telle que les splines de régression, alors $\text{ddl}^{\text{var}} = \text{ddl} = n - \text{ddl}^{\text{err}}$. Les splines de lissage eux vérifient : $\text{ddl}^{\text{var}} \leq \text{ddl} \leq n - \text{ddl}^{\text{err}}$.

2.2.1.1 Estimation de la variance de l'erreur, des écart–types et intervalles de confiance

Un estimateur non biaisé de la variance de l'erreur est le suivant [Hastie et Tibshirani, 1990] :

$$\widehat{\sigma}^2 = \frac{\text{RSS}(\lambda)}{\text{ddl}^{\text{err}}(\lambda)} = \frac{(\mathbf{y} - \widehat{\mathbf{f}}_\lambda)^t (\mathbf{y} - \widehat{\mathbf{f}}_\lambda)}{n - \text{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda \mathbf{S}_\lambda^t)}. \tag{2.8}$$

La matrice de covariances des estimations $\widehat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$ s'écrit $\text{Cov}(\widehat{\mathbf{f}}_\lambda) = \mathbf{S}_\lambda \mathbf{S}_\lambda^t \sigma^2$ [Hastie et Tibshirani, 1990]. Disposant d'un estimateur de la variance de l'erreur (2.8), nous pouvons utiliser la matrice de covariances pour obtenir les écart–types ponctuels : $\text{se}_i = \sigma\sqrt{(\mathbf{S}_\lambda \mathbf{S}_\lambda^t)_{ii}}$, $i = 1, \ldots, n$. Supposant que les erreurs sont Gaussiennes et le biais négligeable, elle peut également être utilisée pour obtenir des intervalles de confiance ponctuels : $\widehat{f}_\lambda(x_{i1}) \pm z_{\alpha/2}\text{se}_i$, où $z_{\alpha/2}$ est le $\alpha/2$–ème percentile de la distribution normale.

D'autres moyens de construire des intervalles de confiance ont été proposés, par exemple, des intervalles de confiance bayésiens pour les méthodes splines [Wahba, 1990, Gu, 2002] : $\widehat{f}_\lambda(x_{i1}) \pm z_{\alpha/2}\sigma\sqrt{(\mathbf{S}_\lambda)_{ii}}$, et des intervalles de confiance bootstrap pour les splines [Wahba, 1990], et pour des méthodes à noyaux [Mammen, 2000].

2.2.2 Modèles additifs

Chacune des définitions de degrés de liberté présentées dans la section précédant admet une définition analogue dans le cadre des modèles additifs [Hastie et Tibshirani, 1990] :

$$\begin{aligned} \text{ddl} &= \text{tr}(\mathbf{R}_{\boldsymbol{\lambda}}) \\ \text{ddl}^{\text{err}} &= n - \text{tr}(2\mathbf{R}_{\boldsymbol{\lambda}} - \mathbf{R}_{\boldsymbol{\lambda}} \mathbf{R}_{\boldsymbol{\lambda}}^t) \\ \text{ddl}^{\text{var}} &= \text{tr}(\mathbf{R}_{\boldsymbol{\lambda}} \mathbf{R}_{\boldsymbol{\lambda}}^t), \end{aligned} \tag{2.9}$$

où $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^t$ et $\mathbf{R}_{\boldsymbol{\lambda}}$ est la matrice qui génère le vecteur des prédictions : $\widehat{\mathbf{f}}_{\boldsymbol{\lambda}} = \mathbf{R}_{\boldsymbol{\lambda}} \mathbf{y}$ (par exemple, quand la procédure backfitting est appliquée, cette matrice est obtenue à la dernière itération de l'algorithme).

Pour le modèle additif, les contributions individuelles sont aussi intéressantes, les degrés de liberté associés à la j–ème composante sont :

$$\begin{array}{l} \mathrm{ddl}_j = \mathrm{tr}(\mathbf{R}_j) \\ \mathrm{ddl}_j^{\mathrm{err}} = \mathrm{tr}(2\mathbf{R}_{\boldsymbol{\lambda}} - \mathbf{R}_{\boldsymbol{\lambda}}\mathbf{R}_{\boldsymbol{\lambda}}^t) - \mathrm{tr}(2\mathbf{R}_{(j)} - \mathbf{R}_{(j)}\mathbf{R}_{(j)}^t) \\ \mathrm{ddl}_j^{\mathrm{var}} = \mathrm{tr}(\mathbf{R}_j\mathbf{R}_j^t), \end{array} \tag{2.10}$$

où $\mathbf{R}_j$ est la matrice de convergence telle que $\widehat{\mathbf{f}}_j = \mathbf{R}_j\mathbf{y}$, et $\mathbf{R}_{(j)}$ est la matrice de convergence pour le modèle additif sans la composante j–ème.

Le calcul des matrices $\mathbf{R}_{\boldsymbol{\lambda}}$, $\mathbf{R}_j$, ou $\mathbf{R}_{(j)}$ peut s'avérer difficile. La somme des traces des matrices de lissage individuelles $\mathbf{S}_j(\lambda_j)$ ne correspond pas exactement à la trace de la matrice $\mathbf{R}_{\boldsymbol{\lambda}}$ mais elle en est une bonne approximation :

$$\begin{array}{rcl} \widehat{\mathbf{f}}_{\boldsymbol{\lambda}} & = & \mathbf{R}_{\boldsymbol{\lambda}}\mathbf{y} = \widehat{\alpha}_0 + \widehat{\mathbf{f}}_1(\lambda_1) + \ldots + \widehat{\mathbf{f}}_p(\lambda_p) = \widehat{\alpha}_0 + (\mathbf{R}_1 + \ldots, +\mathbf{R}_p)\,\mathbf{y} \\ & \approx & \widehat{\alpha}_0 + (\mathbf{S}_1 + \ldots, +\mathbf{S}_p)\,\mathbf{y}, \end{array} \tag{2.11}$$

exceptant les cas où les variables d'entrée sont très corrélées et les cas où les valeurs des paramètres de lissage sont très petites [Buja *et al.*, 1989, Hastie et Tibshirani, 1990]. Les approximations suivantes sont donc adoptées :

$$\begin{array}{ll} \mathrm{ddl}_j \approx \mathrm{tr}(\mathbf{S}_j) - 1 & \mathrm{ddl} \approx \sum_{j=1}^p \mathrm{ddl}_j \\ \mathrm{ddl}_j^{\mathrm{err}} \approx \mathrm{tr}(2\mathbf{S}_j - \mathbf{S}_j\mathbf{S}_j^t) & \mathrm{ddl}^{\mathrm{err}} \approx n - 1 - \sum_{j=1}^p (\mathrm{tr}(\mathbf{S}_j) - 1) \\ \mathrm{ddl}_j^{\mathrm{var}} = \mathrm{tr}(\mathbf{S}_j\mathbf{S}_j^t) & \mathrm{ddl}^{\mathrm{var}} \approx \sum_{j=1}^p \mathrm{ddl}_j^{\mathrm{var}}. \end{array} \tag{2.12}$$

L'approximation $\mathrm{ddl} \approx 1 + \sum_{j=1}^p \mathrm{ddl}_j$ est également utilisée quand l'estimation de la constante est prise en compte.

2.2.2.1 Estimation de la variance de l'erreur, des écart–types et intervalles de confiance

Un estimateur non biaisé de la variance de l'erreur est donnée par [Hastie et Tibshirani, 1990] : $\widehat{\sigma}^2 = \dfrac{\mathrm{RSS}(\lambda)}{\mathrm{ddl}^{\mathrm{err}}(\lambda)}$, où $\mathrm{RSS} = \|\widehat{\alpha}_0 + \widehat{\mathbf{f}}_1(\lambda_1) + \ldots + \widehat{\mathbf{f}}_p(\lambda_p) - \mathbf{y}\|^2$, et $\mathrm{ddl}^{\mathrm{err}}$ est calculé suivant l'approximation précédant.

Les intervalles de confiance ponctuels sont également basés sur des approximations [Hastie et Tibshirani, 1990] : $\mathrm{Cov}(\widehat{\mathbf{f}}_j) = \mathbf{R}_j\mathbf{R}_j^t\sigma^2 \approx \mathbf{S}_j\mathbf{S}_j^t\sigma^2$ ou encore $\approx \mathbf{S}_j\sigma^2$.

Des intervalles de confiance bootstrap, avec une base théorique plus fondée, ont été proposés [Härdle *et al.*, 2004a].

2.2.3 Modèles additifs généralisés

Pour les modèles additifs généralisés, les expressions correspondantes aux définitions (2.9–2.10) se basent sur l'approximation du vrai prédicteur additif, $\nu = g\left(\mathbb{E}[Y|X_1 = x_1, \ldots, X_p = x_p]\right) = \alpha_0 + x_1\alpha_1 + \ldots, +x_p$ (où g est la fonction lien, voir définition des modèles linéaires et additifs généralisés (1.4.1), page 45), par son

estimation à la dernière itération de l'algorithme IRLS, $\widehat{\nu}$ [Hastie et Tibshirani, 1990, Chambers et Hastie, 1993] :

$$\widehat{\boldsymbol{\nu}} = \mathbf{R}_{\boldsymbol{\lambda}} \left(\widehat{\boldsymbol{\nu}} + \left[\frac{-\partial^2 l}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^t} \right]^{-1} \frac{-\partial l}{\partial \boldsymbol{\nu}} \right) = \mathbf{R}_{\boldsymbol{\lambda}} \mathbf{z}, \tag{2.13}$$

où l est la log–vraisemblance, $\mathbf{R}_{\boldsymbol{\lambda}}$ est la matrice pondérée qui génère le vecteur des prédictions, et $\mathbf{z}$ est la réponse de travail, asymptotiquement normale (voir (1.75)–(1.78), page 48).

L'extension de la notion de ddl est donnée simplement par la trace de $\mathbf{R}_{\boldsymbol{\lambda}}$ et, comme dans le cas gaussien, des approximations sont appliquées :

$$\text{ddl} = \text{tr}(\mathbf{R}_{\boldsymbol{\lambda}}) \approx \sum_{j=1}^{p} [\text{tr}(\mathbf{S}_j) - 1], \tag{2.14}$$

où $\mathbf{S}_j$ est la j–ème matrice de lissage pondérée, obtenue dans la dernière itération de l'algorithme IRLS. L'approximation $\text{tr}(\mathbf{R}_{\boldsymbol{\lambda}}) \approx 1 + \sum_{j=1}^{p} [\text{tr}(\mathbf{S}_j) - 1]$ est également utilisée quand on veut tenir compte de l'estimation de la constante.

La définition de ddl^{err} dans le cas gaussien est justifiée par la factorisation de l'espérance de la somme des résidus carrés (RSS). La mesure qui généralise la RSS aux modèles additifs généralisés est la déviance, D. Son expression asymptotique est

$$D \approx (\mathbf{y} - \widehat{\boldsymbol{\mu}})^t \left[\frac{-\partial^2 l}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^t} \right]^{-1} (\mathbf{y} - \widehat{\boldsymbol{\mu}}) \approx (\mathbf{z} - \widehat{\boldsymbol{\nu}})^t \frac{-\partial^2 l}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^t} (\mathbf{z} - \widehat{\boldsymbol{\nu}}), \tag{2.15}$$

où $\widehat{\boldsymbol{\mu}}$ est l'estimation de l'espérance conditionnelle de Y, $\dfrac{-\partial^2 l}{\partial \boldsymbol{\nu}}$ est, en pratique, la matrice des pondérations de l'algorithme IRLS, $\mathbf{W}$. Alors les degrés de liberté de l'erreur sont ici :

$$\text{ddl}^{\text{err}} = n - \text{tr}\left(2\mathbf{R}_{\boldsymbol{\lambda}} - \mathbf{R}_{\boldsymbol{\lambda}}^t \mathbf{W} \mathbf{R}_{\boldsymbol{\lambda}} \mathbf{W}^{-1}\right). \tag{2.16}$$

Les contributions individuelles sont également approchées par $\text{tr}(\mathbf{S}_j) - 1$ et, les degrés de liberté de l'erreur globaux par $n - 1 - \sum_{j=1}^{p} [\text{tr}(\mathbf{S}_j) - 1]$.

Notons que l'expression $\widehat{\boldsymbol{\nu}} = \mathbf{R}_{\boldsymbol{\lambda}} \mathbf{z}$ ne correspond pas ici à un lissage linéaire : la matrice $\mathbf{R}_{\boldsymbol{\lambda}}$ dépend des $\mathbf{y}$ par le biais de la matrice de pondérations. Considérons donc sa version asymptotique, $\mathbf{R}_0$. Alors,

$$\text{Cov}(\widehat{\boldsymbol{\nu}}) \approx \phi \mathbf{R}_0 \text{Cov}(\mathbf{z}) \mathbf{R}_0^t \approx \phi \mathbf{R}_{\boldsymbol{\lambda}} \mathbf{W}^{-1} \mathbf{R}_{\boldsymbol{\lambda}}^t, \tag{2.17}$$

où ϕ est le paramètre de dispersion. Celui–ci est connu dans certaines distributions (binomiale ou Poisson, par exemple) et inconnu dans des autres, pour lesquelles il doit être estimé. De façon similaire,

$$\text{Cov}(\widehat{\mathbf{f}_j}) \approx \phi \mathbf{R}_j \mathbf{W}^{-1} \mathbf{R}_j^t \quad \left(\text{ ou } \approx \phi \mathbf{R}_j \mathbf{W}^{-1} \right), \tag{2.18}$$

où $\mathbf{R}_j$ est la matrice qui génère les $\widehat{\mathbf{f}_j}$ à partir des $\mathbf{z}$. Le nombre de degrés de liberté de la variance est défini par :

$$\text{ddl}^{\text{var}} \approx \sum_j \text{ddl}_j^{\text{var}} = \text{tr}\left(\mathbf{R}_j \mathbf{W}^{-1} \mathbf{R}_j^t\right). \tag{2.19}$$

2.2.3.1 Estimation des écart–types et intervalles de confiance

L'expression (2.18) permet la construction des écart–types ponctuels. Le terme $\mathrm{Cov}(\mathbf{z})$ est approché par l'inverse de la matrice de pondérations à la dernière itération de l'algorithme IRLS, les matrices $\mathbf{R}_j$ sont approchées par $\mathbf{S}_j$, qui apportent seulement l'information marginale [Chambers et Hastie, 1993].

Ces approximations évitent des calculs difficiles. Cependant, en présence de concurvité, des problèmes de sous–estimation de la variance des estimations ont été rapportés [Dominici *et al.*, 2002, Ramsay *et al.*, 2003a, Ramsay *et al.*, 2003b].

Des intervalles de confiance bootstrap ont également été proposés pour les modèles additifs généralisés [Härdle *et al.*, 2004a]. Cette méthode semble mieux se comporter en présence de concurvité [Figueiras *et al.*, 2003].

2.3 Formalisation des objectifs

Le paramètre de lissage optimal (ou, de façon équivalente, le nombre effectif de paramètres optimal) est celui qui minimise la distance entre l'estimation $\widehat{f}_\lambda$ et la vraie fonction de régression f. Nous considérons ici différentes mesures de cette distance [Hastie et Tibshirani, 1990, Hastie *et al.*, 2001].

Soient $\{(X_{i1}, Y_i)\}_{i=1}^n$ échantillon i.i.d. des variables aléatoires parentes (X, Y) et $\{(x_{i1}, y_i)\}_{i=1}^n$ réalisations de $\{(X_{i1}, Y_i)\}_{i=1}^n$.

Une mesure de la distance entre f et son estimation est l'espérance de l'erreur quadratique, MSE (*Mean Squared Error*) :

$$\mathrm{MSE}(x,\lambda) = \mathbb{E}_{\{(X_{i1},Y_i)\}_{i=1}^n}\left[\left(\widehat{f}(x,\lambda) - f(x)\right)^2\right] = \mathrm{Biais}^2\left[\widehat{f}(x,\lambda)\right] + \mathrm{Var}\left[\widehat{f}(x,\lambda)\right]. \tag{2.20}$$

L'espérance est prise par rapport à toutes les possibles échantillons de (X, Y) (notons que $\widehat{f}(x,\lambda)$ est fonction des variables aléatoires $\{(X_{i1}, Y_i)\}_{i=1}^n$). La version conditionnelle est considérée quand l'intérêt porte sur le comportement des estimateurs pour la réalisation disponible, plutôt que pour tous les réalisations possibles de la densité de X :

$$\mathrm{MSE}(x,\lambda) = \mathbb{E}_{\{Y_i\}_{i=1}^n}\left[\left(\widehat{f}(x,\lambda) - f(x)\right)^2 \middle| X_{11} = x_{11}, \ldots, X_{n1} = x_{n1}\right], \tag{2.21}$$

(par abus de notation nous noterons la version conditionnelle de l'espérance de l'erreur quadratique encore par MSE).

La moyenne de cette mesure par rapport aux observations, $\sum_i \mathrm{MSE}(x_{i1},\lambda)/n$, est connue sous le nom d'espérance de l'erreur quadratique moyenne, MASE (*Mean Average Squared Error*). D'autres mesures ne comprennent pas d'espérance, comme l'erreur quadratique moyenne, ASE (*Average Squared Error*) :

$$\mathrm{ASE}(\lambda) = \frac{1}{n}\sum_{i=1}^n \left(\widehat{f}(X_{i1},\lambda) - f(X_{i1})\right)^2, \tag{2.22}$$

qui est donc une variable aléatoire.

Les mesures introduites jusqu'ici sont ponctuelles, d'autres mesures globales de la distance entre les deux courbes incorporent l'intégrale par rapport à la densité de X, comme l'intégrale de l'erreur quadratique, ISE (*Integrated Squared Error*) :

$$\mathrm{ISE}(\lambda) = \int \left(\widehat{f}(x,\lambda) - f(x)\right)^2 h_X(x)dx = \left|\left|\widehat{f}_\lambda - f\right|\right|^2_{L_2(h_X)}, \tag{2.23}$$

où h_X est la densité de X, et $\|\cdot\|^2_{L_2(h_X)}$ est la norme de l'espace de Hilbert des fonctions de carré intégrable. Une autre mesure globale est l'espérance (conditionnelle ou pas) de l'intégrale de l'erreur quadratique moyenne, MISE (*Mean Integrated Squared Error*) :

$$\mathrm{MISE}(x_1,\ldots,x_n,\lambda) = \mathbb{E}_{\{Y_i\}_{i=1}^n}\left[\mathrm{ISE}(\lambda)\middle| X_1 = x_1 \ldots, X_p = x_p\right]. \tag{2.24}$$

Les mesures basées sur une espérance (conditionnelle ou pas) admettent, comme MSE (2.20), une factorisation en termes du biais et de la variance. En particulier, pour les lissages linéaires, dans le cas de MASE, les termes du biais et de la variance ont les expressions suivantes :

$$\mathrm{MASE}(\lambda) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[\widehat{f}_\lambda(x_{i1})] + \frac{1}{n}\sum_{i=1}^n b^2_\lambda(x_{i1}) = \frac{\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda^t)}{n}\sigma^2 + \frac{\mathbf{b}_\lambda\mathbf{b}_\lambda^t}{n}, \tag{2.25}$$

(voir (2.6) et justification de (2.7), page 55).

Une mesure qui diffère de MASE seulement par une fonction constante de σ^2 est l'espérance de l'erreur quadratique de prédiction, APE (*Average Predictive Error*) :

$$\mathrm{APE}(\lambda) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\{(X_{i1},Y_i)\}_{i=1}^n}\left[\left(\widehat{f}(x_{i1},\lambda) - Y_i^*\right)^2\right] = \mathrm{MASE}(\lambda) + \sigma^2, \tag{2.26}$$

où Y_i^* est une nouvelle observation en x_{i1}, i.e., $Y_i^* = f(x_{i1}) + \varepsilon_i^*$, ε_i^* iid des ε_i, $i = 1,\ldots,n$.

Une mesure plus générale est l'erreur de prédiction, PE (*Predictive Error*) :

$$\mathrm{PE}(\lambda) = \mathbb{E}_{(X,Y)}\left[\left(\widehat{f}(X,\lambda) - Y\right)^2\right]. \tag{2.27}$$

Ces mesures reposent sur la fonction de coût quadratique. D'autres quantités basées sur des fonctions de coût différentes (L_1 et L_∞, notamment) ont été étudiées, mais leur analyse est plus complexe. Une quantité différente est l'information de Kullback–Leibler qui mesure de la perte occasionnée par l'approximation d'une fonction de densité g par une autre, h [Sakamoto *et al.*, 1986] :

$$I(g,h) = \int g(x)\log\frac{g(x)}{h(x)} = \mathbb{E}\log[g(X)/h(X)]. \tag{2.28}$$

Dans le contexte de la régression, la vraisemblance du vraie modèle et celle du modèle considéré remplacent les fonctions de densité g et h, respectivement.

2.4 Critères de sélection de la complexité

Les mesures ci–dessus dépendent de la fonction inconnue f. Elles ne peuvent, donc, pas être calculées directement. Différentes méthodes ont été proposées pour les estimer. Nous abordons tout d'abord ces méthodes dans le cas unidimensionnel gaussien et, ensuite, nous considérons l'extension au cas additif et additif généralisé.

Dans le cas unidimensionnel, la résolution de (1.5), page 17, pour les noyaux, de (1.9), page 18, pour les polynômes locaux, ou de (1.20), page 21, pour les splines cubiques de lissage, nécessite pré–définir la valeur d'un seul paramètre de lissage λ, tandis que dans le cas multidimensionnel, la résolution de (1.62), page 43, pour les noyaux, ou de (1.47), page 36, pour les splines cubiques, nécessite de définir au préalable p paramètres de lissage $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$. Bien que l'estimation de chaque composante additive soit un problème de nature unidimensionnelle, le choix du degré de lissage reste un problème multidimensionnel. La généralisation des méthodes automatiques de sélection du paramètre de lissage unidimensionnelles devient alors difficile et présente des problèmes divers.

Une première solution au problème de la sélection de la complexité consiste à diviser l'échantillon en deux sous–ensembles : en utiliser une partie, l'ensemble d'apprentissage, pour l'estimation des fonctions, et la partie restante, l'ensemble de test, pour la sélection des paramètres de la complexité. Néanmoins, cette solution n'est pas réalisable quand le nombre d'observations n'est pas suffisamment élevé.

Des alternatives ont été proposées. Fondamentalement, quatre classes de méthodes abordent le problème de la sélection de la complexité : les méthodes d'évaluation sur une grille (ou p–cube), les méthodes de resubstitution, les tests d'hypothèses, et les méthodes bayésiennes.

Les méthodes d'évaluation sur une grille englobent les méthodes consistant à évaluer un critère sur une collection de points et choisir, ensuite, la valeur minimisant le critère. L'estimation de l'hyper–paramètre optimal est donc le point de la collection pour lequel le critère prend la valeur minimale. Des approches très différentes sont inclues dans cette classe, telles que des méthodes de rééchantillonnage ou des critères bayésiens.

Les méthodes de resubstitution sont propres aux méthodes non paramétriques. Elles sont développées pour les méthodes à noyau et se basent sur des résultats asymptotiques. Les tests d'hypothèses sont des extensions du cas paramétrique. Les résultats sont souvent seulement approximatifs. Finalement, les méthodes bayésiennes attribuent une probabilité *a priori* aux paramètres de la complexité. Dans l'approche bayésienne empirique, les données interviennent dans la détermination de l'*a priori*.

2.4.1 Méthodes d'évaluation sur une grille de type rééchantillonage

Les techniques de rééchantillonage sont basées sur le principe de divisions multiples de l'échantillon en un ensemble d'apprentissage et un ensemble de validation. Ces techniques sont précises, car la totalité de l'ensemble d'apprentissage est utilisé pour déterminer $\widehat{f}$ et $\widehat{\lambda}$. Elles ont l'avantage de ne faire d'hypothèses ni sur la fonction

de régression, ni sur la forme du bruit. En particulier, ces techniques ne requièrent pas l'estimation du nombre effectif de paramètres ni de la variance de l'erreur. En revanche, chacune des divisions réclame un nouvel apprentissage, ce qui implique une considérable quantité d'opérations, elles sont donc très coûteuses en temps de calcul. Cela limite leur application au cas unidimensionnel ou multidimensionnel avec p peu élevée (2 ou 3).

2.4.1.1 Validation croisée

Cas unidimensionnel

Dans la validation croisée, CV (*cross validation*), l'ensemble de validation est constitué d'un ensemble de points de taille V tiré de l'ensemble des données. L'ensemble de points restants est utilisé pour l'apprentissage. L'ensemble de mesures $\mathcal{L}$ est découpé en K ensembles de même taille V : $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K$. Pour chaque k, $k = 1, \ldots, K$, l'estimateur $\widehat{f}_\lambda^{-\mathcal{L}_k}$ est construit à partir de l'ensemble d'apprentissage $(\mathcal{L} - \mathcal{L}_k)$, avec λ fixé. Il est ensuite évalué sur ensemble de validation $\mathcal{L}_k$:

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{V} \sum_{x_{i1}, y_i \in \mathcal{L}_k} \left(\widehat{f}_\lambda^{-\mathcal{L}_k}(x_{i1}) - y_i \right)^2. \tag{2.29}$$

Cette version de la CV est dite à K sous-ensembles (*K–fold CV*). Une version plus simple est la validation croisée *leave–one–out*, où les n ensembles de validation sont constitués d'un seul élément :

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{f}_\lambda^{-i}(x_{i1}) - y_i \right)^2, \tag{2.30}$$

où, pour chaque i, l'estimation $\widehat{f}_\lambda^{-i}$ est construite à partir de l'ensemble d'apprentissage $\{x_{k1}\}_{k \neq i}$, et évaluée, ensuite, en x_{i1}. Pour les méthodes de lissage linéaires, cette valeur a une expression en termes de la matrice de lissage : $\widehat{f}_\lambda^{-i}(x_{i1}) = \sum_{k \neq i} \frac{(\mathbf{S}_\lambda)_{ik}}{1 - (\mathbf{S}_\lambda)_{ii}} y_k$ [Hastie et Tibshirani, 1990]. La CV *leave–one–out*, en fonction de la matrice de lissage, a donc l'expression suivante :

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\widehat{f}_\lambda(x_{i1}) - y_i}{1 - (\mathbf{S}_\lambda)_{ii}} \right)^2. \tag{2.31}$$

Le nombre de calculs demandé par cet expression est inférieur à ceux demandés par d'autres versions de la CV et comparable au nombre de calculs demandé par les méthodes d'évaluation sur une grille de type analytique.

La validation croisée est un estimateur d'APE qui vérifie, pour les lissages linéaires, $\mathbb{E}\left(\mathrm{CV}(\lambda)\right) \approx \mathrm{APE}(\lambda) + \frac{2}{n} \sum_{i=1}^{n} (\mathbf{S}_\lambda)_{ii} (\mathbf{b}_\lambda)_i^2$, où $\mathbf{b}_\lambda = \mathbf{f} - \mathbf{S}_\lambda \mathbf{f}$ est le biais [Hastie et Tibshirani, 1990].

Cette technique présente des bonnes propriétés théoriques, telles que la convergence en probabilité d'ASE (et ISE) [Simonoff, 1996, Hart, 1997] :

$$\lim_{n\to\infty} \frac{\mathrm{ASE}(\lambda_{\mathrm{CV}})}{\mathrm{ASE}(\lambda_{\mathrm{opt}})} = 1. \tag{2.32}$$

Cependant, elle présente des problèmes de variance élevée (qui se traduit par une variabilité importante des estimations $\widehat{\lambda}_{\mathrm{CV}}$), ainsi qu'une tendance à sous–lisser (elle sélectionne des complexités trop élevées) [Simonoff, 1996, Herrmann, 2000]. Des modifications de la CV ont été proposées afin de corriger le problème de la variance élevée, au détriment du biais. Il a été observé qu'un facteur important de cette variabilité est la corrélation négative entre $\widehat{\lambda}_{\mathrm{CV}}$ et la vraie valeur [Hart, 1997].

Modèles additifs

Le critère (2.31) peut être adapté aux modèles additifs [Schimek et Turlach, 2000] :

$$\mathrm{CV}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{\mathbf{f}}_{\boldsymbol{\lambda}}(\mathbf{x}_i) - y_i}{1-(\mathbf{R}_{\boldsymbol{\lambda}})_{ii}}\right)^2, \tag{2.33}$$

où $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$, $\widehat{\mathbf{f}}_{\boldsymbol{\lambda}}(\mathbf{x}_i) = \widehat{f}_0 + \sum_{j=1}^{p} \widehat{f}_{\lambda_1}(x_{ij})$. Cependant, la validation croisée pour les modèles additifs apparaît rarement dans la littérature.

2.4.1.2 Bootstrap

Le *bootstrap* [Efron et Tibshirani, 1993] utilise le principe de resubstitution (le remplacement les paramètres inconnus par des estimations, voir section 2.4.3) de manière intensive. Cette méthode consiste à remplacer la loi F dont sont issues les données par la densité empirique $\widehat{F}$ pour calculer l'estimateur d'une quantité donnée. Ainsi, un échantillon s_l de taille l, tiré indépendamment selon F est remplacé par un échantillon de taille l, tiré indépendamment selon $\widehat{F}$. Ce dernier échantillon est obtenu en faisant l tirages équiprobables avec remise sur l'échantillon s_l. En répétant la procédure B fois sur s_l, on obtient ainsi B échantillons différents de taille l : $s_l^1, \ldots, s_l^b, \ldots, s_l^B$. Pour chaque b, on calcule l'estimateur $\widehat{\mathbf{f}}_{\lambda}^{b}$ sur l'ensemble d'apprentissage s_l^b.

Cas unidimensionnel

Le bootstrap appliqué à la sélection du paramètre de la complexité, dans le contexte non paramétrique, se réduit aux méthodes à noyau. Les estimations bootstrap se construisent à partir d'estimations par resubstitution de MISE (voir section 2.4.3, page 70) [Mammen, 2000].

L'avantage du bootstrap sur les méthodes de resubstitution découle du fait que le terme du biais de MISE est estimé avec une précision élevée. Cette consistance

d'ordre élevé n'est pas atteinte par les méthodes de resubstitution, qui reposent sur une approximation asymptotique de MISE. L'implémentation consistante du bootstrap demande, en revanche, des conditions de régularité sévères sur la fonction de régression f.

Le bootstrap améliore, en général, l'instabilité de la validation croisée [Efron et Tibshirani, 1995]. Toutefois, le nombre d'opérations et le temps de calcul, pour la sélection de l'hyper–paramètre, sont considérablement supérieurs.

Modèles additifs

Le bootstrap pour la régression non paramétrique proposé par Mammen [Mammen, 2000] est théoriquement généralisable à la régression additive. Néanmoins, cette application n'a pas été concrétisée.

2.4.2 Méthodes d'évaluation sur une grille de type analytique

En dimension p élevée (ou même modérée) les techniques de rééchantillonnage deviennent irréalisables : le nombres de points à évaluer sur le p–cube augmente de façon exponentielle avec la dimension, et pour chacun de ces points, de nombreuses estimations de la fonction de régression sont nécessaires. Des critères analytiques sont alors proposés.

2.4.2.1 Validation croisée généralisée

Cas unidimensionnelle

La validation croisée généralisée, GCV (*generalized cross validation*) est un exemple d'approximation de la validation croisée [Craven et Wahba, 1979] :

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{f}_\lambda(x_{i1}) - y_i}{1 - \mathrm{ddl}(\lambda)/n}\right)^2 = \frac{1}{n}\frac{\|\mathbf{S}_\lambda \mathbf{y} - \mathbf{y}\|_2^2}{(1 - \mathrm{ddl}(\lambda)/n)^2} = \frac{n\mathrm{RSS}(\lambda)}{(n - \mathrm{ddl}(\lambda))^2}, \quad (2.34)$$

où l'expression de $\mathrm{ddl}(\lambda)$ est celle de (2.3). En effet, l'élément $(\mathbf{S}_\lambda)_{ii}$ de l'équation (2.31) est remplacé par la moyenne des éléments $\sum_i (\mathbf{S}_\lambda)_{ii}/n$. Cette méthode conserve un des avantages de la CV sur d'autres méthodes : elle ne requière pas d'estimation de la variance de l'erreur.

La GCV, initialement développée pour la sélection du paramètre de lissage des splines cubiques pour la fonction coût quadratique, a été adaptée a des nombreuses méthodes telles que les machines à vecteurs de support [Lin *et al.*, 2000] ou les méthodes de pénalisation pour les modèles linéaires [Tibshirani, 1996].

La GCV aspire à atteindre des propriétés d'invariabilité que la CV ne possède pas [Wahba, 1990]. Considérons les problèmes de régression :

$$\begin{aligned} &y_i = f_\lambda(x_{i1}) + \varepsilon_i \quad \text{où } \mathbb{E}(\varepsilon_i) = 0 \text{ et } Var(\varepsilon_i) = \sigma^2 \\ &\widetilde{y}_i = \widetilde{f}_\lambda(x_{i1}) + \widetilde{\varepsilon}_i \quad \text{où } \widetilde{y} = \Gamma y,\ \widetilde{f}_\lambda(x_{i1}) = \Gamma f(x_{i1}) \text{ et } \widetilde{\varepsilon} = \Gamma\varepsilon, \end{aligned} \quad (2.35)$$

où Γ est une matrice orthogonale $n \times n$. Les deux problèmes d'estimation de f sont le même problème, puisque $\mathbb{E}(\widetilde{\varepsilon}_i) = 0$, $Var(\widetilde{\varepsilon}_i) = \sigma^2$. Cependant, en général, les hyper–paramètres estimés par validation croisée sont différents. La validation croisée généralisée, elle reste invariante par application d'une rotation.

Il existe une justification bayésienne de la validation croisée généralisée [Golub *et al.*, 1979], ainsi qu'une justification asymptotique. Des bons résultats ne sont pas assurés pour des échantillons de petite taille : il existe une probabilité, faible mais non nulle, que GCV sélectionne incorrectement un hyper–paramètre très petit, conduisant à des problèmes de sous–lissage [Wahba et Wang, 1995]. Afin de corriger les éventuels problèmes de sous–lissage, la modification suivante a été proposée [Kim et Gu, 2004] :

$$\mathrm{GCV}(\lambda) = \frac{1}{n} \frac{\|\mathbf{S}_\lambda \mathbf{y} - \mathbf{y}\|_2^2}{(1 - \gamma \mathrm{ddl}(\lambda)/n)^2}, \tag{2.36}$$

où $\gamma \geq 1$ est une constante pré–définie. Quand γ augmente, on obtient des estimations plus lisses. Expérimentalement, des bons résultats ont été obtenus pour $\gamma \in [1.2, 1.4]$.

Le résultat suivant a été démontré. Il existe une suite d'hyper–paramètres $\{\lambda_n\}_{n\in\mathbb{N}}$ minimisant l'espérance de la GCV et telle que [Craven et Wahba, 1979, Golub *et al.*, 1979] :

$$\lim_{n\to\infty} \frac{\mathrm{MASE}(\lambda_n)}{\inf_{\lambda\geq 0} \mathrm{MASE}(\lambda)} = 1. \tag{2.37}$$

Un résultat plus fort est le suivant : il existe une suite d'hyper–paramètres $\{\lambda_n\}_{n\in\mathbb{N}}$ minimisant la GCV, qui converge en probabilité [Li, 1986] :

$$\lim_{n\to\infty} \frac{\mathrm{ASE}(\widehat{\lambda}_n)}{\inf_{\lambda\geq 0} \mathrm{ASE}(\lambda)} = 1. \tag{2.38}$$

Comme dans le cas de la CV, une corrélation négative entre $\widehat{\lambda}_{\mathrm{GCV}}$ et le vrai paramètre de lissage a été observée. Cette particularité a été étudiée par [Gu, 1998]. L'auteur conclue que les résultats dépendent de la formulation du problème (par exemple, si la complexité des splines est indexée en termes du paramètre de pénalisation, λ, ou du seuil de la contrainte, τ, section 1.2.3.4).

La GCV est robuste dans des situations où les erreurs ne sont pas normales [Xiang et Wahba, 1996] ou hétéroscedastiques [Andrews, 1991]. Cependant, GCV montre une mauvaise performance quand les erreurs sont corrélées [Wahba, 1990].

Pour des problèmes de grande taille (tel que la restauration d'images) l'utilisation de la GCV devient difficile. La version randomisées de la validation croisée généralisée, RGCV (*randomized generalized cross–validation*), réduit les calculs de la trace, en utilisant que $\mathbb{E}(\epsilon^t \mathbf{S} \epsilon)/\sigma_\epsilon^2 = \mathrm{tr}(\mathbf{S})$, où ϵ est un vecteur de dimension n avec une distribution $N(\mathbf{0}, \sigma_\epsilon \mathbf{I}_n)$ [Girard, 1991, Wahba et Luo, 1997].

Modèles additifs

L'adaptation de la GCV aux modèles additifs est donnée par [Gu, 2002] :

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{\mathbf{f}}_{\boldsymbol{\lambda}}(\mathbf{x}_i) - y_i}{1 - \mathrm{ddl}(\boldsymbol{\lambda})/n}\right)^2 = \frac{1}{n}\frac{||\mathbf{R}_{\boldsymbol{\lambda}}\mathbf{y} - \mathbf{y}||_2^2}{(1 - \mathrm{ddl}(\boldsymbol{\lambda})/n)^2} = \frac{n\mathrm{RSS}(\boldsymbol{\lambda})}{(n - \mathrm{ddl}(\boldsymbol{\lambda}))^2}, \tag{2.39}$$

où l'expression de $\mathrm{ddl}(\boldsymbol{\lambda})$ est ici celle de (2.9).

Modèles additifs généralisés

L'adaptation de la GCV aux modèles additifs généralisés est donnée par [Gu, 1992a, Gu, 2002] :

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{1}{n}\frac{\left|\left|\mathbf{W}^{1/2}\left(\mathbf{R}_{\boldsymbol{\lambda}}\mathbf{z} - \mathbf{z}\right)\right|\right|_2^2}{(1 - \mathrm{ddl}(\boldsymbol{\lambda})/n)^2}. \tag{2.40}$$

où l'expression du nombre effectif de paramètres est celle de (2.14), $\mathbf{z}$ est la réponse de travail et $\mathbf{W}$ la matrice des pondérations de la dernière itération de l'algorithme IRLS.

Pour des réponses binaires, une version de la GCV basée sur la log–vraisemblance pénalisée (1.76), au lieu du coût quadratique, a été également proposée, GACV (*generalized approximated Cross Validation*) [Xiang et Wahba, 1996, Gu et Xiang, 2001, Gu, 2002] :

$$\mathrm{GACV}(\boldsymbol{\lambda}) = -\frac{1}{n}l(\alpha_0, \widehat{\mathbf{f}}_{\boldsymbol{\lambda}}) + \frac{1}{n}\sum_{i=1}^{n} y_i(y_i - P_i)\frac{\mathrm{tr}(\mathbf{R}\mathbf{W}^{-1})}{n - \mathrm{tr}(\mathbf{R})}, \tag{2.41}$$

L'application de ce dernier critère au modèle de Poisson ne semble pas obtenir de bons résultats [Yuan et Wahba, 2001].

Des versions aléatoires de GCV et GACV existent également.

2.4.2.2 Maximum de vraisemblance généralisée

Cas unidimensionnel

Basé sur l'interprétation bayésienne des splines de lissage, le critère maximum de vraisemblance généralisée, GML (*generalized maximum likelihood*) est donné par [Wahba, 1990] :

$$\mathrm{GML}(\lambda) = \frac{\mathbf{y}^t(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}}{n\left[\det{}^+(\mathbf{I} - \mathbf{S}_\lambda)\right]^{\frac{1}{n-m}}}, \tag{2.42}$$

où $\det^+(\mathbf{I} - \mathbf{S}_\lambda)$ indique le produit des $n - m$ valeurs propres différentes de 0 de $(\mathbf{I} - \mathbf{S}_\lambda)$, et $m = 2$.

Les comportements de la GML et la GCV ont été comparés par rapport à ASE [Wahba, 1985]. Les conclusions des simulations sont les suivantes. Quand la vraie fonction de régression f est lisse et n élevé, GML a tendance à sous-lisser, comparativement à GCV et $\widehat{\lambda}_{\mathrm{GML}}$ converge à la valeur optimale plus lentement

que $\widehat{\lambda}_{\mathrm{GCV}}$. Quand f est lisse et n petit, si le bruit n'est pas important, la GCV se comporte mieux. Dans ce dernier cas, si le bruit est élevé, les deux méthodes sont équivalentes. Quand f n'est pas lisse (régulière mais avec une courbure élevée), les deux méthodes sont équivalentes.

Modèles additifs

L'adaptation de la GML aux modèles additifs est donnée par [Gu et Wahba, 1991, Gu, 2000] :

$$\mathrm{GML}(\boldsymbol{\lambda}) = \frac{\mathbf{y}^t(\mathbf{I} - \mathbf{R}_{\boldsymbol{\lambda}})\mathbf{y}}{n\left[\det{}^{+}(\mathbf{I} - \mathbf{R}_{\boldsymbol{\lambda}})\right]^{\frac{1}{n-m}}}, \tag{2.43}$$

où m est la dimension du noyau de $(\mathbf{I} - \mathbf{R}_{\boldsymbol{\lambda}})$.

2.4.2.3 Critère d'information d'Akaike, critère d'information bayésien et C_p de Mallow

Cas unidimensionnel

Les expressions de l'espérance de RSS (2.6) (en appliquant le facteur $1/n$), et d'APE (en appliquant la factorisation (2.25)), diffèrent de $2\mathrm{tr}(\mathbf{S}_\lambda)\sigma^2/n$. La statistique C_p de Mallow utilise ce terme correcteur pour estimer APE à partir de RSS [Hastie et Tibshirani, 1990] :

$$C_p(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_\lambda(x_{i1}) - y_i\right)^2 + \frac{2\mathrm{ddl}(\lambda)\sigma^2}{n} = \frac{\mathrm{RSS}(\lambda)}{n} + \frac{2\mathrm{ddl}(\lambda)\sigma^2}{n}. \tag{2.44}$$

La variance de l'erreur est estimée par $\mathrm{RSS}(\widetilde{\lambda})/\mathrm{ddl}^{\mathrm{err}}(\widetilde{\lambda})$ (2.8), où $\widetilde{\lambda}$ comporte une complexité élevée (afin d'obtenir une estimation peu biaisée). Par exemple, pour les splines, on peut même prendre $\widetilde{\lambda} = 0$, à condition que le calcul soit numériquement stable [Ruppert *et al.*, 2003].

Les mêmes résultats asymptotiques obtenus pour la GCV sont obtenus pour C_p [Li, 1986]. Ceci n'est pas étonnant, car une approximation de la GCV est $\mathrm{RSS}(\lambda)/n + 2\mathrm{ddl}(\lambda)\mathrm{RSS}(\lambda)/n$, qui est très proche de C_p.

Des versions stochastiques de C_p ont été également proposées [Girard, 1991].

Le critère d'information d'Akaike, AIC (*Akaike information criterion*) est la version log–vraisemblance de C_p. Les deux statistiques sont équivalentes pour des erreurs gaussiennes iid :

$$\mathrm{AIC}(\lambda) = -2\widehat{l}(\lambda) + 2\mathrm{ddl}(\lambda), \tag{2.45}$$

où $\widehat{l}$ est la log–vraisemblance maximisée. Le critère d'information bayésien, BIC (*Bayesian information criterion*) est défini par :

$$\mathrm{BIC}(\lambda) = -2\widehat{l}(\lambda) + \log(n)\mathrm{ddl}(\lambda). \tag{2.46}$$

Le critère AIC est un estimateur de l'information de Kullback–Leibler, tandis que BIC est basé sur des arguments bayésiens. Minimiser AIC et BIC est équivalent à

maximiser la fonction de vraisemblance par rapport aux observations y_i, pénalisée par une mesure de la complexité [Herrmann, 2000, Smith *et al.*, 2000]. La méthode BIC pénalise plus fortement le nombre de paramètres effectifs, alors elle choisi des modèles plus simples. BIC est un critère consistent : quand le nombre d'observations tend vers l'infini, cette méthode sélectionne le modèle correct, ce qui n'est pas le cas pour AIC.

Afin d'éviter la variabilité élevée et la tendance à sous–lisser d'AIC (et GCV), une correction a été proposée [Hurvich *et al.*, 1998] :

$$\mathrm{AICc}(\lambda) = -2\widehat{l}(\lambda) + \frac{2n\mathrm{ddl}(\lambda)}{n - \mathrm{ddl}(\lambda) - 1}. \tag{2.47}$$

Modèles additifs

L'adaptation de C_p aux modèles additifs est donnée par :

$$C_p(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mathbf{f}}_{\boldsymbol{\lambda}}(\mathbf{x}_i) - y_i\right)^2 + \frac{2\mathrm{ddl}(\boldsymbol{\lambda})\sigma^2}{n} = \frac{\mathrm{RSS}(\boldsymbol{\lambda})}{n} + \frac{2\mathrm{ddl}(\boldsymbol{\lambda})\sigma^2}{n}, \tag{2.48}$$

où l'expression de $\mathrm{ddl}(\boldsymbol{\lambda})$ est ici celle de (2.9). (Voir l'estimateur de la variance de l'erreur dans la section 2.2.2.1, page 57).

Modèles additifs généralisés

L'adaptation d'AIC aux modèles additifs généralisés est donnée par [Hastie et Tibshirani, 1990] :

$$\mathrm{AIC}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} D(y_i, \widehat{\mu}_i) + \frac{2\mathrm{ddl}(\boldsymbol{\lambda})\phi}{n} \approx \frac{1}{n}\left|\left|\mathbf{W}^{1/2}\left(\mathbf{R}_{\boldsymbol{\lambda}}\mathbf{z} - \mathbf{z}\right)\right|\right|_2^2 + \frac{2\mathrm{ddl}(\boldsymbol{\lambda})\phi}{n}, \tag{2.49}$$

où les expressions de la déviance et du nombre effectif de paramètres sont, respectivement, celles de (2.15) et (2.14), $\mathbf{z}$ est la réponse de travail et $\mathbf{W}$ la matrice des pondérations de la dernière itération de l'algorithme IRLS.

Ce critère, qui est également connu comme UBR (*unbiased risk estimate*), est spécialement utilisé pour le modèle logistique et le modèle de Poisson, pour lesquels $\phi = 1$ [Gu, 1992a, Gu, 2002].

Des formulations équivalentes peuvent être obtenues pour BIC.

2.4.2.4 Algorithmes

Différentes approches ont abordé le problème de l'implémentation des méthodes d'évaluation sur une grille de type analytique. Ces approches ont été, le plus souvent, centrées sur la GCV. Bien que quelques–unes utilisent des caractéristiques propres à la GCV, la plupart sont transposables aux autres critères.

L'approche la plus simple consiste à considérer de façon indépendante les procédures de minimisation du critère (par rapport à $\boldsymbol{\lambda}$) et d'estimation de la fonction de régression. On définie alors une grille (ou p–cube), c'est à dire, une partition de m

points de chaque espace : $\{\lambda_{k_j j}\}_{k_j=1,\ldots,m}$, $j = 1, \ldots, p$, (et donc m^p points sur l'espace p–dimensionnel : $\{(\lambda_{k_1 1}, \ldots, \lambda_{k_p p})\}_{k_j=1,\ldots,m}$), et on estime la fonction de régression, avec $\boldsymbol{\lambda}$ fixé à un des points de la grille. On répète la même procédure pour chacun des points, et ensuite on minimise le critère, évalué sur chaque point de la grille.

Bien que la convergence soit assurée pour chacune des étapes, cette stratégie est très lourde à niveau des calculs. D'autres approches se basent sur l'alternance des étapes, et/ou sur des méthodes newtoniennes, et/ou la réduction du problème d'optimisation d'un critère p–dimensionnel à un problème itératif d'optimisation d'un critère uni–dimensionnel.

BRUTO

Cet algorithme combine la sélection des paramètres de lissage avec le *back-fitting* [Hastie et Tibshirani, 1990]. A chaque étape du backfitting, où la j–ème composante est ajustée supposant les autres fixes, le critère GCV est minimisé par rapport à λ_j, supposant les $\{\lambda_k\}_{k\neq j}$ fixes. L'approximation aux dégrées de liberté ddl $\approx 1 + \sum_{j=1}^{p}[\mathrm{tr}(\mathbf{S}_j) - 1]$ est appliquée dans l'expression de GCV. Quand la procédure de sélection est stabilisée, les paramètres de la complexité sont fixés et le backfitting est itéré jusqu'à convergence. La convergence de BRUTO n'est pourtant pas assurée.

Simplification du critère par des transformations

L'application d'un algorithme de diagonalisation aux splines pénalisées (section 1.3.5.4) permet de simplifier les calculs nécessaires à l'estimation, ainsi que le calcul du critère GCV [Ruppert, 2002]. Les procédures d'estimation (par résolution directe d'un système d'équations) et d'optimisation de la GCV sont ici indépendants. La GCV est minimisée, initialement, par rapport à un seul paramètre de lissage, commun à toutes les variables $\lambda = \lambda_1 = \ldots = \lambda_p$. Ensuite, le critère est minimisé par rapport à λ_j, supposant les $\{\lambda_k\}_{k\neq j}$ fixés.

Méthode de Newton modifiée

Une alternative à l'exploration de tout l'espace sont les méthodes d'optimisation newtoniennes. Une première méthode est proposée par [Gu et Wahba, 1991] pour des splines généraux. Après une re–paramétrisation des splines de lissage :

$$\lambda_j = \rho/\theta_j, \tag{2.50}$$

les dérivées de premier et deuxième ordre de la GCV sont déduites. L'algorithme proposé applique des factorisations QR ainsi qu'un algorithme de (tri–)diagonalisation. Ensuite, de façon itérative, les calculs de ρ et des θ_j sont alternés. Le premier paramètre (la complexité "globale"), pour lequel les θ_j sont considérés fixes, est obtenu par minimisation du critère unidimensionnel GCV. Ensuite, considérant le paramètre ρ fixe, les dérivés suivies de l'ensemble des paramètres θ_j sont actualisées (de façon similaire aux réponses de travail de l'algorithme IRLS).

Cette méthode est raffinée par [Wood, 2000]. L'inclusion de contraintes linéaires dans le problème d'optimisation permet de tenir compte, entre d'autres, des B–splines ou des splines naturelles, qui incluent des contraintes sur les bords. Des expressions plus explicites des estimations sont disponibles pour les splines pénalisées et les splines de lissage. Des simplifications sont possibles par l'application d'algorithmes QR, de (tri–)diagonalisation, ainsi que des décompositions de Choleski. Cet algorithme a été adapté aux modèles additifs généralisés, en appliquant ici des décompositions QR et des décompositions en valeurs singulières [Wood, 2004].

Une méthode newtonienne, inspirée des précédentes, a été également proposée pour les polynômes locaux [Kauermann et Opsomer, 2004].

Modèles additifs généralisés

Deux approches sont proposées pour la résolution de (2.40). La première consiste à fixer $\boldsymbol{\lambda}$, évaluer GCV seulement à la dernière itération et minimiser ensuite GCV par rapport à $\boldsymbol{\lambda}$. La deuxième consiste à alterner les itérations IRLS et la minimisation du critère : après une itération IRLS, on sélectionne $\boldsymbol{\lambda}$ (dans le voisinage du $\boldsymbol{\lambda}$ précédant), on actualise ensuite les éléments dépendant de $\boldsymbol{\lambda}$, et on applique à nouveau un itération IRLS. On répète la procédure jusqu'à que les valeurs de $\boldsymbol{\lambda}$ soient stabilisées et l'algorithme IRLS converge. Le deuxième algorithme semble mieux se comporter à la pratique, cependant, sa convergence n'est pas assurée [Gu, 1992a, Gu, 2002].

Deux stratégies de calcul sont également considérées pour AIC (2.49). La plus utilisée est celle qui alterne les itérations IRLS et la minimisation du critère. La convergence n'est pourtant pas assurée.

2.4.3 Méthodes de resubstitution

La resubstitution (*plug in*) consiste à remplacer les paramètres inconnus par des estimations de ces paramètres, dans une équation quelconque. Pour la régression non paramétrique, cette méthode est présentée comme une alternative aux techniques d'évaluation sur une grille, qui demandent l'ajustement de plusieurs modèles. L'application de la resubstitution est réduite aux méthodes à noyaux, où des expressions asymptotiques relativement simples des mesures présentées à la section 2.3 sont déduites facilement.

Cas unidimensionnel

Dans ce contexte, la resubstitution consiste à minimiser, par rapport à la largeur de bande, des approximations asymptotiques de la distance entre f et $\widehat{f}_\lambda$, et remplacer, ensuite, les quantités inconnues dans les expressions résultantes par des estimations. La difficulté de cette méthode est que les quantités inconnues dépendent, indirectement, du paramètre largeur de bande.

Par exemple, pour l'estimateur de Nadaraya–Watson, sous des conditions de régularité de la fonction de régression et de la densité g_X, supposant $\lambda \to 0$ et $n\lambda \to \infty$, l'expression asymptotique de MSE, (prenant en compte jusqu'aux termes

quadratiques de l'expansion de Taylor) est déduite [Härdle *et al.*, 2004b] :

$$\mathrm{MSE}(x,\lambda) \approx \left(\frac{1}{n\lambda}\frac{\sigma^2}{g_X(x)}\|K\|_2^2\right) + \left(\frac{\lambda^4}{4}\left[f^{(2)}(x) + 2\frac{f^{(1)}g_X^{(1)}(x)}{g_X(x)}\right]^2\left[\int t^2K(t)dt\right]^2\right), \tag{2.51}$$

pour x tel que la densité $g_X(x) > 0$, pour un noyau K tel que $\int |K(t)|dt < \infty$, $\lim_{|t|\to\infty} tK(t) = 0$ et pour $\mathbb{E}Y^2 < \infty$. L'erreur quadratique moyenne peut s'écrire $\frac{1}{n\lambda}C_1 + \lambda^4 C_2$, où les termes C_1 et C_2, correspondent au terme de la variance asymptotique et du biais asymptotique, respectivement. Ils sont constants en n et λ, mais ils dépendent, respectivement, de σ^2 et de $f^{(2)}$, dont les estimateurs eux dépendent de n et λ.

Pour les polynômes locaux, (en particulier les polynômes locaux de degré 1, dont les formules sont plus simples) la formule diffère de la précédant seulement dans le terme du biais [Härdle *et al.*, 2004b] :

$$\mathrm{MSE}(x,\lambda) \approx \left(\frac{1}{n\lambda}\frac{\sigma^2}{g_X(x)}\|K\|_2^2\right) + \left(\frac{\lambda^4}{4}\left[f^{(2)}(x)\right]^2\left[\int t^2K(t)dt\right]^2\right). \tag{2.52}$$

La valeur de λ minimisant cette fonction est la suivante :

$$\lambda_{\mathrm{opt}} = \left[\frac{\sigma^2\|K\|_2^2}{n\left[\int t^2K(t)dt\right]^2\left[f^{(2)}(x)\right]^2 g_X(x)}\right]^{1/5}. \tag{2.53}$$

Des nombreuses techniques de resubstitution ont été proposées [Wand et Jones, 1995]. Elles divergent par rapport à la distance considérée (généralement MISE, MASE), par rapport à la flexibilité des hypothèses (aboutissant à des expressions asymptotiques plus ou moins simples), et par rapport à la stratégie adoptée pour estimer et calculer les quantités inconnues. Le *plug in* direct [Ruppert *et al.*, 1995] découle de l'approximation et estimation du paramètre de lissage optimal (selon MASE), λ_{opt}, suivantes :

$$\lambda_{\mathrm{opt}} \simeq \left[\frac{\sigma^2\|K\|_2^2(x_{(n)} - x_{(1)})}{n\left[\int t^2K(t)dt\right]^2\theta}\right]^{1/5} \qquad \widehat{\lambda} = C(K)\left[\frac{\hat{\sigma}^2(x_{(n)} - x_{(1)})}{n\widehat{\theta}}\right]^{1/5}, \tag{2.54}$$

où $C(K)$ est une constante qui dépend du noyau K, $x_{(1)} = \min_i(\{x_{i1}\})$, $x_{(n)} = \max_i(\{x_{i1}\})$, et $\theta = \mathbb{E}[f^{(2)}(x)]^2$. Les estimations $\widehat{\sigma}^2$ et $\widehat{\theta}$ sont obtenues séparément (puisque $\widehat{\theta}$ dépend de σ^2) par des régressions polynomiales locales, où des estimateurs initiaux sont trouvés à l'aide du critère C_p de Mallow.

Modèles additifs

Le *plug in* direct a été généralisé aux modèles additifs ajustés par backfitting [Opsomer et Ruppert, 1997, Opsomer et Ruppert, 1998]. Cette méthode est présentée comme une alternative aux méthodes basées sur l'évaluation d'un p–cube, cependant elle demande parfois des hypothèses très restrictives.

Considérons la régression linéaire locale additive pour deux variables explicatives. Sous certaines conditions de régularité et de compacité des densités et du noyau, des condition de régularité de la fonction de régression, supposant $\lambda_j \to 0$, $n\lambda_j/\log(n) \to \infty$, $j = 1, 2$, et supposant que l'éventuelle dépendance entre X_1 et X_2 n'est pas sévère : $\sup_{x_1,x_2}\left|\frac{g_{X_1,X_2}(x_1,x_2)}{g_{X_1}(x_1)g_{X_2}(x_2)} - 1\right| < 1$, alors :

$$\begin{aligned}\text{MASE} \approx \; & \frac{1}{4}\left[\int t^2 K(t)dt\right]^2 (\lambda_1^4\theta_{11} + \lambda_1^2\lambda_2^2\theta_{12} + \lambda_2^4\theta_{22}) + \\ & \sigma^2\|K\|_2^2\left(\frac{x_{(n),1} - x_{(1),1}}{n\lambda_1} + \frac{x_{(n),2} - x_{(1),2}}{n\lambda_2}\right),\end{aligned} \tag{2.55}$$

où $x_{(1),j} = \min_i(\{x_{ij}\})$, $x_{(n),j} = \max_i(\{x_{ij}\})$,

$$\begin{aligned}\theta_{11} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \left(\mathbf{t}_i^t \mathrm{D}^2\mathbf{f}_1 - \mathbf{v}_i^t\mathbb{E}(f_1^{(2)}(x_{i1})|X_2)\right)^2 \\ \theta_{22} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \left(\mathbf{v}_i^t \mathrm{D}^2\mathbf{f}_2 - \mathbf{t}_i^t\mathbb{E}(f_2^{(2)}(x_{i2})|X_1)\right)^2 \\ \theta_{12} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \left(\mathbf{t}_i^t \mathrm{D}^2\mathbf{f}_1 - \mathbf{v}_i^t\mathbb{E}(f_1^{(2)}(x_{i2})|X_2)\right)\left(\mathbf{v}_i^t \mathrm{D}^2\mathbf{f}_2 - \mathbf{t}_i^t\mathbb{E}(f_2^{(2)}(x_{i2})|X_1)\right),\end{aligned} \tag{2.56}$$

$\mathbf{t}_i$ et $\mathbf{v}_i$ sont, respectivement, la i-ème ligne et colonne de l'approximation asymptotique de $(\mathbf{I} - \mathbf{S}_1^*\mathbf{S}_2^*)^{-1}$, et $\mathrm{D}^2\mathbf{f}_j = (f_j^{(2)}(x_{1j}), \ldots, f_j^{(2)}(x_{nj}))^t$.

Sous l'hypothèse d'indépendance de X_1 et X_2, des simplifications de θ_{11}, θ_{22}, et l'annulation de θ_{12} permettent déduire des expressions analytiques des paramètres de lissage optimaux :

$$(\lambda_{1\text{opt}}, \lambda_{2\text{opt}}) = \left(\left[\frac{\sigma^2\|K\|_2^2(x_{(n),1} - x_{(1),1})}{n\left[\int t^2K(t)dt\right]^2\theta_{11}}\right]^{1/5}, \left[\frac{\sigma^2\|K\|_2^2(x_{(n),2} - x_{(1),2})}{n\left[\int t^2K(t)dt\right]^2\theta_{22}}\right]^{1/5}\right) \tag{2.57}$$

Comme dans le cas unidimensionnel, le calcul de $(\widehat{\lambda}_1, \widehat{\lambda}_2)$ nécessite des estimations $\widehat{\sigma}^2$ et $\widehat{\theta}_{jj}$.

Etant donné que $\widehat{\theta}_{jj}$ dépend de l'estimation de la variance et que les deux estimations demandent une sélection du paramètre de lissage, la stratégie suivante est adoptée. Des estimations de σ^2 et de $f^{(4)}$ sont obtenues par des régressions polynomiales locales, où des estimateurs initiaux sont trouvés à l'aide du critère C_p de Mallow. Ces valeurs sont remplacées dans les expressions asymptotiques des θ_{jj} permettant la déduction des paramètres optimaux. Ces paramètres sont remplacés par leurs estimations, $\widehat{\theta}_{jj}$. Ensuite, une nouvelle estimation des paramètres de lissage est calculée et appliquée à une actualisation de σ^2. Finalement $(\widehat{\lambda}_1, \widehat{\lambda}_2)$ est calculé à partir de $\widehat{\sigma}^2$ et $\widehat{\theta}_{jj}$.

Bien que les résultats théoriques ne demandent pas l'hypothèse d'indépendance dans le cas bi–dimensionnel, cette hypothèse est nécessaire pour obtenir des expressions analytiques des estimations $(\widehat{\lambda}_1, \widehat{\lambda}_2)$. Les calculs deviennent compliqués dans le cas contraire. Cependant, pour l'extension des résultats au cas p–dimensionnel, $p > 2$,

cette hypothèse est insuffisante pour démontrer théoriquement que la méthode minimise la valeur MASE et assurer la convergence.

A partir des simulations, les auteurs concluent que la méthode fonctionne bien pour $p = 2, 3$, et qu'elle est relativement insensible aux violations de l'hypothèse d'indépendance pour $p = 2$. A partir d'un exemple, les auteurs concluent que des résultats cohérents sont obtenus pour $p = 5$. Cependant, quand la dimension augmente, le nombre de calculs est très important (quoique inférieur à celui des méthodes basées sur l'évaluation d'un p–cube de type analytique) [Kauermann et Opsomer, 2004]. Une autre limite de la méthode de resubstitution est que l'extension aux modèles additifs généralisés n'est pas facile.

2.4.4 Tests d'hypothèses

Une approche différente dans la sélection de la complexité consiste à choisir, à l'aide d'un test statistique, entre deux alternatives. Pour les modèles additifs, cette approche évite l'optimisation d'un critère multidimensionnel. Cependant, des tests séquentiels, comparant à chaque étape deux complexités différentes sont rarement employés, car la multitude de tests à effectuer, qui ne sont pas généralement indépendants, est problématique. A la pratique, ces méthodes sont appliquées pour comparer deux valeurs pré–définies, au lieu de rechercher la valeur optimale. On teste si l'estimation donnée décrit la relation entre les variables explicatives et la variable expliquée de façon satisfaisante ou un ajustement plus flexible est nécessaire. En particulier on peut tester si une composante est linéaire ou lisse, ou aussi, comme dans le cas paramétrique, pour décider si une composante est pertinente (voir section 3.2.2, page 90).

Cas unidimensionnel

Par analogie à la régression linéaire, un test F approximatif peut être déduit [Hastie et Tibshirani, 1990]. Supposons que nous voulons choisir entre deux paramètres de lissage :

$$\begin{array}{l} H_0 : \lambda = \lambda^0 \\ H_1 : \lambda = \lambda^1, \end{array} \tag{2.58}$$

où la courbe $\widehat{\mathbf{f}}(\lambda^1) = \mathbf{S}(\lambda^1)\mathbf{y}$ est plus complexe que $\widehat{\mathbf{f}}(\lambda^0) = \mathbf{S}(\lambda^0)\mathbf{y}$. Supposons que $\widehat{\mathbf{f}}(\lambda^1)$ est non biaisée, et que $\widehat{\mathbf{f}}(\lambda^0)$ est non biaisée sous l'hypothèse nulle. Soiet $\mathrm{ddl}_k^{\mathrm{err}}$, les degrés de liberté de l'erreur (2.5) et RSS_k la sommes des erreurs quadratiques, $k = 0, 1$, alors

$$\frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\mathrm{ddl}_0^{\mathrm{err}} - \mathrm{ddl}_1^{\mathrm{err}})}{\mathrm{RSS}_1/\mathrm{ddl}_1^{\mathrm{err}}} \sim F_{\mathrm{ddl}_0^{\mathrm{err}} - \mathrm{ddl}_1^{\mathrm{err}}, \mathrm{ddl}_1^{\mathrm{err}}}. \tag{2.59}$$

Modèles additifs

Une adaptation du test F aux modèles additifs est la suivante [Bowman et Azzalini, 1997].

$$\begin{array}{l} H_0 : \mathrm{ddl}_j = \mathrm{ddl}_j^0 \\ H_1 : \mathrm{ddl}_j = \mathrm{ddl}_j^1, \end{array} \tag{2.60}$$

où ddl_j indique les degrés de liberté pour la composante j (2.4), $j = 1, \ldots, p$, supposant les autres composantes fixées, ddl_j^1 et ddl_j^0 sont deux valeurs pré–définies. Un test F approximatif est :

$$\frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\mathrm{ddl}_j^1 - \mathrm{ddl}_j^0)}{\mathrm{RSS}_1/(n - \mathrm{ddl}_j^1)} \sim F_{\mathrm{ddl}_j^1 - \mathrm{ddl}_j^0, n - \mathrm{ddl}_j^1}. \tag{2.61}$$

Un test spécifique pour les splines de lissage est le suivant [Cantoni et Hastie, 2002] :

$$\begin{array}{l} H_0 : \mathrm{ddl}_j = \mathrm{ddl}_j^0 \\ H_1 : \mathrm{ddl}_j = \mathrm{ddl}_j^1 > \mathrm{ddl}_j^0, \end{array} \tag{2.62}$$

supposant les composantes $l \neq j$ fixées. Quand $\mathrm{ddl}_j^0 = 1$, il s'agit de tester la linéarité de la j–ème composante.

La statistique de test est :

$$F_{\mathrm{AM}} = \frac{\mathbf{y}^t(\mathbf{R}(\boldsymbol{\lambda}^1) - \mathbf{R}(\boldsymbol{\lambda}^0))\mathbf{y}}{\mathbf{y}^t(\mathbf{I} - \mathbf{R}(\boldsymbol{\lambda}^1))\mathbf{y}}, \tag{2.63}$$

où $\mathbf{R}(\boldsymbol{\lambda}^k)$, $k = 0, 1$ est la matrice de convergence de l'algorithme *backfitting* (voir section 2.2.2, page 56), pour les valeurs des paramètres de lissage λ_j^0 et λ_j^1, respectivement.

Le degré de signification (p–*value*) est calculée par :

$$\begin{aligned} P(F_{\mathrm{AM}} > f_{\mathrm{AM,obs}}) &= P\left(\mathbf{y}^t[\mathbf{R}(\boldsymbol{\lambda}^1) - \mathbf{R}(\boldsymbol{\lambda}^0) - f_{AM,\mathrm{obs}}(\mathbf{I} - \mathbf{R}(\boldsymbol{\lambda}^1))]\mathbf{y} > 0\right) \\ &= P\left(R_{\mathrm{AM}} > 0\right), \end{aligned} \tag{2.64}$$

où $f_{\mathrm{AM,obs}}$ est la valeur de F_{AM} évaluée sur les observations.

Sous H_0, R_{AM} suit une distribution χ^2_δ, où δ dépend de $\mathbf{R}(\boldsymbol{\lambda}^k)$, $k = 0, 1$, $\mathbf{S}_l(\lambda_l)$, $l \neq j$, et $\mathbf{S}_j(\lambda_j^0)$. Le calcul de cette valeur est lourd quand la dimension p ou la taille de l'échantillon n sont élevées. Dans ces cas il est préférable l'utilisation d'un test F approximatif.

Des tests d'hypothèses ont été proposés spécifiquement pour tester la linéarité d'une des composantes. Une statistique de type F, basée sur la différence des sommes des erreurs quadratiques sous l'hypothèse nulle (ajustement linéaire) et sous l'hypothèse alternative (ajustement par noyaux) est proposée par [Azzalini et Bowman, 1993]. Pour éviter le problème de l'estimation du paramètre de lissage, les auteurs analysent le graphique du degré de signification en fonction du paramètre de lissage. L'indépendance des variables explicatives est supposée, et la partie non paramétrique du modèle semi–paramétrique est constituée d'une

seule composante. Ces hypothèses sont partagées par [Shively *et al.*, 1994], pour un ajustement avec des splines. L'estimation du paramètre de lissage est également évité, ainsi on compare $H_0 : \boldsymbol{\lambda} = \mathbf{0}$, $H_1 : \boldsymbol{\lambda} = \lambda^1$, avec $\boldsymbol{\lambda}^1$ tel que la puissance du test demandée soit atteinte, pour un niveau de signification fixé.

Modèles additifs généralisés

On s'intéresse à la comparaison de deux modèles emboîtés qui divergent seulement à la j–ème composante : $\widehat{\boldsymbol{\nu}}^0$ (le plus simple) et $\widehat{\boldsymbol{\nu}}^1$ (le plus complexe) [Hastie et Tibshirani, 1990]. Sous l'hypothèse nulle, le modèle le plus simple est correcte, et

$$\mathbb{E}D(\widehat{\boldsymbol{\nu}}^0, \widehat{\boldsymbol{\nu}}^1)/\phi \approx \mathrm{ddl}^{\mathrm{err}}(\widehat{\boldsymbol{\nu}}^0) - \mathrm{ddl}^{\mathrm{err}}(\widehat{\boldsymbol{\nu}}^1). \tag{2.65}$$

Pour les modèles linéaires généralisés, si le paramètre de dispersion est connu, la distribution asymptotique de $D(\widehat{\boldsymbol{\nu}}^0, \widehat{\boldsymbol{\nu}}^1)$ est une distribution χ^2. Pour les modèles additifs généralisés, la distribution asymptotique de la déviance est une $\chi^2_{\mathrm{ddl}^{\mathrm{err}}(\widehat{\boldsymbol{\nu}}^0) - \mathrm{ddl}^{\mathrm{err}}(\widehat{\boldsymbol{\nu}}^1)}$ seulement approximativement. Cette approximation peut être améliorée par une correction du premier et deuxième moment. Quand le paramètre de dispersion est inconnu, un test F approximatif est plus adéquate.

Une modification du test χ^2 pour tester la linéarité d'une des composantes du modèle semi–paramétrique généralisé est proposée par [Härdle *et al.*, 2004a]. Une correction du biais de l'estimateur non paramétrique et une correction de la statistique de test sont appliquées, au moyen d'une procédure bootstrap. La statistique de test est asymptotiquement normale, mais la convergence est lente. Les auteurs proposent donc d'utiliser aussi le bootstrap pour calculer les valeurs critiques.

2.4.5 Méthodes bayésiennes

En utilisant la formulation bayésienne des splines de régression (voir section 1.2.3.4, page 24), des techniques bayésiennes d'estimation des modèles additifs, qui intègrent la sélection des paramètres de la complexité ont été proposées [Wong et Kohn, 1996]. Des méthodes de Monte Carlo par chaînes de Markov sont appliquées à l'estimation. Des probabilités *a priori* sont considérées pour les hyper–paramètres : $P(\sigma^2) \propto 1/\sigma^2$, et $P(\tau_j) \propto \exp[-10^{-10}/\tau_j^2]$, $j = 1, \ldots, p$, où σ^2 est la variance de l'erreur et τ_j est tel que $\lambda_j = \sigma^2/\tau_j^2$. Des distributions plus complexes, telle que la gamma inverse sont également considérées pour ces hyper–paramètres. Les auteurs soulignent la robustesse de cette méthode par rapport à des observations aberrantes, en comparaison à d'autres approches.

2.5 En bref

Nous avons traité ici le problème du contrôle de la complexité pour les modèles additifs. Les difficultés de la mise en œuvre des différentes méthodes ont été étudiées. En effet, l'application des méthodes de réechantillonnage est limitée par la quantité de calculs nécessaire. Les bases théoriques des méthodes de resubstitution ne

garantissent pas leur application pour $p > 2$, même sous des hypothèses fortes, telle que l'indépendance des variables explicatives. Les méthodes basées sur des tests rendent difficile l'automatisation des procédures, et la multitude de tests effectués (si le nombre de variables est modéré ou important) ne sont généralement pas indépendants. Les approches bayésiennes sont peu développées. Seules les méthodes analytiques, pour lesquelles des algorithmes efficaces ont été proposés, aboutissent à des résultats satisfaisants.

Chapitre 3

Modèles additifs parcimonieux

3.1 Introduction

Nous abordons dans ce chapitre le problème général de la sélection de modèle pour les modèles additifs, avec une spéciale attention à la sélection de variables.

Dans le chapitre précédant nous avons analysé les différents problèmes que les méthodes de sélection de modèle comportent. Seules les méthodes analytiques aboutissent à des résultats satisfaisants quand il s'agit d'estimer la complexité des composantes additives. Cependant, quand la sélection de la complexité doit aboutir à la suppression de variables, le problème devient impraticable même pour un nombre modéré d'entrées.

Une approche différente consiste à explorer seulement une partie préalablement déterminée de la grille de valeurs des paramètres de la complexité. Il existe la possibilité de "laisser choisir aux données" la partie de la grille à inspecter. Cela peut être réalisé par l'application de méthodes de régularisation et l'introduction d'une information *a priori* qui nous dirige vers la partie à explorer.

Dans ce chapitre, nous passons revue des méthodes de régularisation pour les modèles linéaires, dans une optique de sélection de variables. Deuxièmement, nous étudions les méthodes de sélection de variables proposées pour les modèles additifs. Nous présentons ensuite notre approche, qui est motivée par les méthodes de régularisation dans le cadre linéaire. Celle–ci accomplit une sélection parcimonieuse des variables, qui a pour objectif d'éviter la sur–estimation ainsi que la sous–estimation.

3.2 Sélection de variables : état de l'art

La sélection de variables consiste à sélectionner le groupe de variables d'entrée les plus prédictives de la variable de sortie [Miller, 1990, Bi *et al.*, 2003]. Les objectifs sont multiples : améliorer la précision en prédiction (par exemple, par la réduction de l'instabilité), faciliter la compréhension du processus sous–jacent qui a généré les données (par exemple, par l'obtention de modèles simples, facilitant leur interprétation), et réduire le temps d'obtention des solutions [Guyon et Elisseeff, 2003].

Le problème de la sélection de variables peut être considéré dans un cadre plus général et être abordé d'une perspective de régularisation.

Régulariser un problème "mal posé", c'est de le remplacer par un autre, "bien posé", de sorte que l'erreur commise soit compensée par le gain de stabilité [Tikhonov et Arsenin, 1977]. Un problème est dit bien posé si les conditions suivantes sont vérifiées : 1) la solution existe, 2) elle est unique, 3) elle dépend continûment des données. Dans le cas contraire, le problème est dit mal posé (*ill–posed problem*). Ce sont des problèmes pour lesquels l'information spécifiée n'est pas suffisante et des hypothèses supplémentaires sont nécessaires. Ainsi, pour résoudre l'instabilité, une information *a priori* est introduite.

La forme générale d'une classe de problèmes de régularisation est la suivante :

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L\Big(y_i, f(x_{i1}, \ldots, x_{ip})\Big) + \mu J(f), \tag{3.1}$$

où L est une fonction de coût, J est un opérateur de pénalisation, $\mathcal{F}$ est un espace de fonctions, et $\mu \geq 0$ est le paramètre de régularisation, qui contrôle le compromis entre l'ajustement de la fonction aux données, évalué par le coût, et la pénalisation de la fonction.

3.2.1 Modèles linéaires

L'estimation des moindres carrés, OLS (*ordinary least squares*), obtenue par minimisation de l'erreur quadratique, est souvent peu satisfaisante. Ses limites concernent la précision en prédiction : bien que l'estimation OLS ait un biais faible, sa variance est souvent élevée. Aussi, en pratique, un nombre important de variables d'entrée est considéré, afin d'atténuer le biais résultant de la modélisation. Un tel nombre de variables, parfois non pertinentes ou apportant une information redondante, rend difficile la compréhension du phénomène.

Les méthodes de régularisation, appliquées à la régression linéaire, ont pour objectif de réduire la variance des estimateurs, (améliorant ainsi leur précision) et/ou obtenir des modèles plus simples, avec peu de coefficients non–nuls (facilitant l'interprétation). En général, ces techniques sacrifient un peu de biais afin d'obtenir une réduction de la variance. Un problème commun à ces méthodes est l'estimation du paramètre de régularisation, aussi connu dans ce contexte sous le nom de paramètre de complexité ou hyper–paramètre.

Passons la revue des principales méthodes de régularisation pour les modèles linéaires. Par simplicité, nous considérons au cour de la section (3.2.1) que la variable réponse est centrée, ce qui réduit l'estimation de la constante à 0 (considérant que les variables d'entrée sont centrées et réduites).

3.2.1.1 Principales méthodes de régularisation

Sélection de sous–ensembles

La sélection de sous–ensembles (*subset selection*), consiste à déterminer un sous–ensemble des variables explicatives, qui sont estimées par moindres carrés ordinaires,

et éliminer les autres [Miller, 1990]. De nombreuses procédures automatiques ont été proposées pour sélectionner le sous–ensemble de variables (procédure pas à pas ascendante, descendante, mixte [Miller, 1990], ou d'autres procédures plus minutieuses telles que *stagewise* et *lars* (*least angle regression*) [Efron *et al.*, 2004]), ainsi que de nombreux critères de sélection (basées sur des tests, sur des statistiques telles que AIC ou BIC [Miller, 1990], sur les coefficients de corrélations entre les variables d'entrée et les résidus, ...). Cette méthode comporte des problèmes combinatoires : 2^p modèles sont disponibles quand p variables d'entrée sont considérées.

Pénalisation quadratique

La pénalisation quadratique, RR (*ridge regression*) [Hoerl et Kennard, 1970] rétrécit les coefficients de la régression, par une pénalisation de la norme l_2 du vecteur de coefficients :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \mu \|\boldsymbol{\alpha}\|_2^2, \tag{3.2}$$

où μ est une valeur prédéfinie. La solution explicite du problème (3.2) est donnée par :

$$\boldsymbol{\alpha}^{\mathrm{RR}} = \left(\mathbf{X}^t\mathbf{X} + \mu\mathbf{I}_p\right)^{-1}\mathbf{X}^t\mathbf{y}. \tag{3.3}$$

Dans le cas de la sélection de sous–ensembles, la variance des prédictions est réduite par l'élimination d'un certain nombre de variables. Pour la pénalisation quadratique, la réduction de la variance peut être expliquée au moyen d'une décomposition en valeurs singulières. Supposons $\mathbf{X} = \mathbf{UDV}$, où $\mathbf{U}$ est une matrice de dimension $n \times p$, telle que $\mathbf{U}^t\mathbf{U} = \mathbf{I}$, $\mathbf{V}$ est une matrice de dimension $p \times p$, telle que $\mathbf{V}^t\mathbf{V} = \mathbf{I}$, et $\mathbf{D}$ est une matrice diagonale $p \times p$ constituée des valeurs singulières $\{d_j\}_{j=1,\ldots,p}$. L'estimateur pénalisation quadratique (3.3) admet l'expression :

$$\boldsymbol{\alpha}^{\mathrm{RR}} = \mathbf{V}^t\left(\mathbf{D}^2 + \mu\mathbf{I}_p\right)^{-1}\mathbf{DU}^t\mathbf{y} = \mathbf{V}^t \begin{pmatrix} \frac{d_1}{d_1^2+\mu} & & \\ & \ddots & \\ & & \frac{d_p}{d_p^2+\mu} \end{pmatrix} \mathbf{U}^t\mathbf{y}, \tag{3.4}$$

et sa variance, $\mathrm{Var}(\boldsymbol{\alpha}^{\mathrm{RR}}) = \sigma^2\mathbf{V}^t\mathrm{diag}\left[\left\{\frac{d_j^2}{(d_j^2+\mu)^2}\right\}_j\right]\mathbf{V}$. L'addition de μ réduit la contribution des plus petites valeurs singulières, celles qui dominent la variance.

Un résultat sur l'efficacité de la pénalisation quadratique est le suivant. Il existe une valeur strictement positive du paramètre de réguralisation ($\mu > 0$), telle que l'erreur de l'estimateur RR est plus petite que celle de l'estimateur OLS ($\mu = 0$), au sens de MASE [Gruber, 1998]. Ce résultat est démontré à partir de la dérivée de MASE par rapport à μ, qui est facilement déduite en termes du biais et la variance et décomposition en valeurs singulières.

La sélection de sous–ensembles donne lieu à des modèles simples, mais elle est instable[1], car la procédure est discrète : la variable est retenue ou éliminée [Breiman, 1996, Tibshirani, 1996]. La pénalisation quadratique, elle, est stable [Breiman, 1996, Gruber, 1998], mais elle ne donne pas lieu à des coefficients nuls. En effet, la solution explicite en termes de la décomposition en valeurs singulières (3.4) permet d'observer que, bien que le rétrécissement soit appliqué à tous les coefficients, celui–ci est proportionnel : les coefficients plus petits sont moins rétrécis que les coefficients plus importants, les empêchant d'atteindre le zéro. La pénalisation quadratique fonctionne mieux quand toutes les variables d'entrée sont également pertinentes.

Cela a motivé la recherche d'autres méthodes, combinant les avantages de la sélection de sous–ensembles et de la pénalisation quadratique.

Lasso

Le lasso (*least absolute shrinkage and selection operator*) [Tibshirani, 1996], comme la pénalisation quadratique, rétrécit les coefficients de la régression, mais par pénalisation de la norme l_1 du vecteur de coefficients :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \mu \|\boldsymbol{\alpha}\|_1. \tag{3.5}$$

Une formulation plus classique du lasso est donnée en termes du problème d'optimisation sous contraintes :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 \qquad \text{sous contrainte} \qquad \sum_{j=1}^{p} |\alpha_j| \leq \tau, \tag{3.6}$$

où τ est une valeur prédéfinie. Une particularité de la pénalisation l_1 est que certains coefficients sont rétrécis, alors que les autres sont annulés exactement, effectuant ainsi l'estimation des coefficients et la sélection de variables de façon simultanée. Aussi, la forme lisse de la pénalisation conduit à des modèles moins variables que ceux obtenus par la sélection de sous–ensembles [Tibshirani, 1996].

La comparaison expérimentale des trois méthodes de pénalisation, a montré que la sélection de sous–ensembles est la mieux adaptée, suivie du lasso, lorsque le nombre de variables ayant un effet important est petit. Dans la situation inverse, quand le nombre de variables ayant un effet faible est élevé, la pénalisation quadratique est la méthode la plus appropriée, suivie du lasso. Enfin, dans les situations intermédiaires, le lasso est le plus performant, suivie de la pénalisation quadratique [Tibshirani, 1996].

Bridge regression

[1]Selon [Breiman, 1996], une méthode de pénalisation est instable si pour deux échantillons peu différents, l'algorithme d'estimation de la fonction de régression donne des résultats très variables

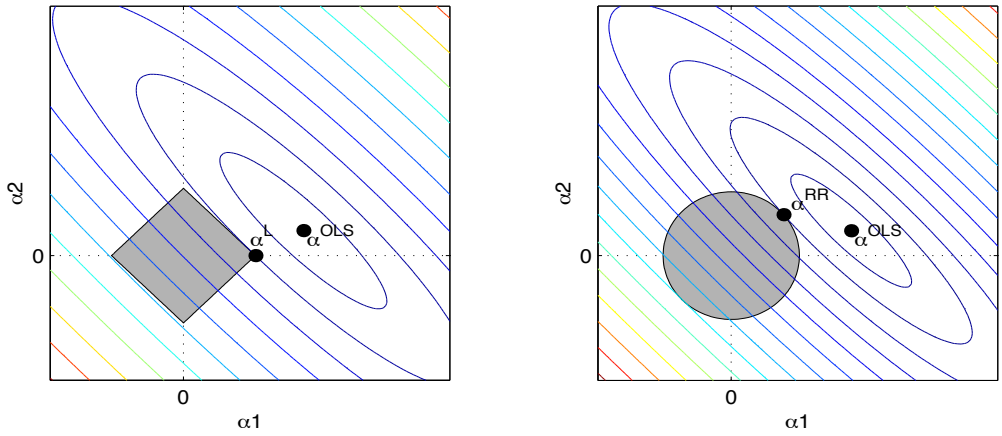

FIG. 3.1 – Solution du lasso (à gauche), notée $\boldsymbol{\alpha}^{\mathrm{L}}$, et de la pénalisation quadratique (à droite) notée $\boldsymbol{\alpha}^{\mathrm{RR}}$, pour $\tau = 1$ et $p = 2$. Les aires grises sont les régions définies par les contraintes $\|\boldsymbol{\alpha}\|_q^q \leq 1$, où $q = 1$, pour le lasso et et $q = 2$, pour la pénalisation quadratique. Les ellipses sont les contours de l'erreur quadratique en fonction de $\boldsymbol{\alpha}$, autour de la solution OLS.

Ces trois méthodes sont englobées par une méthode plus générale, la *bridge regression* [Frank et Friedman, 1993] :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \mu \sum_{j=1}^{p} |\alpha_j|^q, \tag{3.7}$$

où $q \geq 0$ et $\mu \geq 0$ sont des hyper–paramètres. De façon équivalente,

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 \qquad \text{sous contrainte} \qquad \sum_{j=1}^{p} |\alpha_j|^q \leq \tau. \tag{3.8}$$

En effet, la sélection de sous–ensembles correspond à $q = 0$, la fonction de pénalisation rendant compte du nombre de coefficients non nuls. La pénalisation quadratique et le lasso correspondent à $q = 2$ et $q = 1$, respectivement.

Convexité et stabilité

Les régions $\|\boldsymbol{\alpha}\|_q^q \leq \tau$ telles que $q \geq 1$ sont strictement convexes, et inversement, pour $q < 1$, les régions définies par la contrainte sont concaves. L'optimisation de ces derniers problèmes devient alors plus difficile. Aussi, pour des fonctions non convexes, les solutions $\widehat{\boldsymbol{\alpha}}_q(\tau)$ ne sont pas continues en τ et subséquemment des problèmes d'instabilité surgissent [Knight, 2004]. En particulier, la sélection de sous–ensembles ($q = 0$), comme signalé précédemment, est une méthode instable.

La figure (3.1) montre les solutions du lasso et de la pénalisation quadratique dans le cas bi–dimensionnel pour $\tau = 1$. La fonction cible (représentée par les

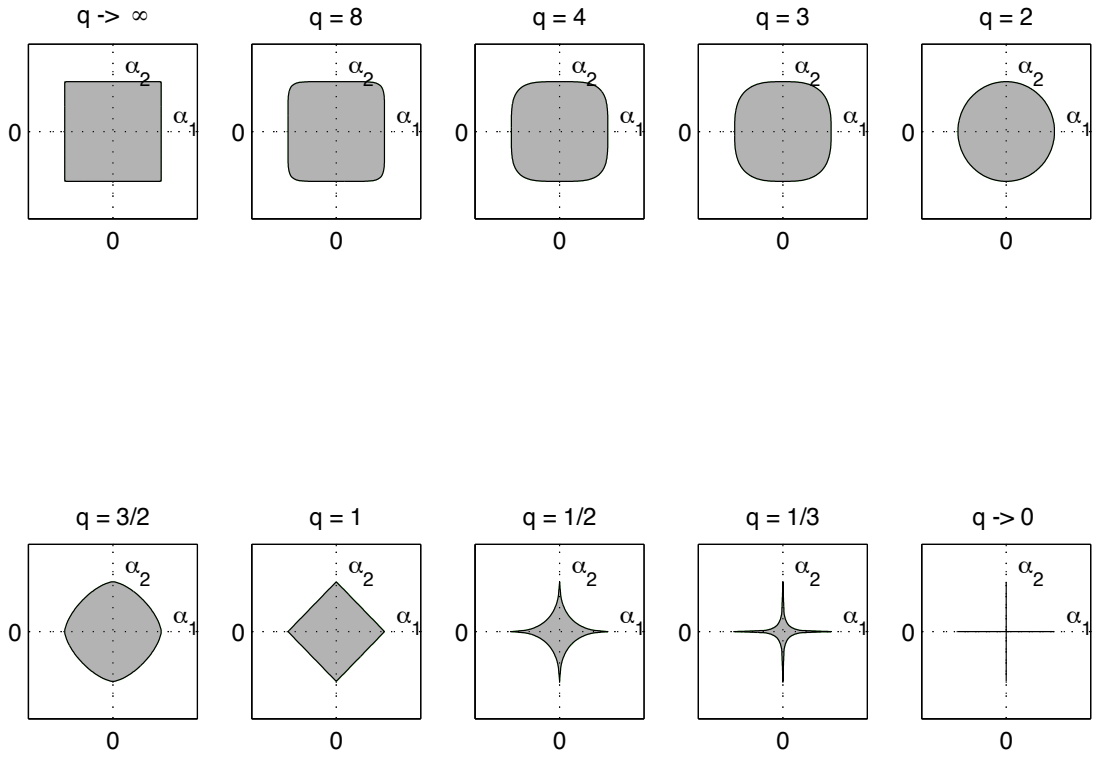

FIG. 3.2 – Régions définies par les contraintes $\|\boldsymbol{\alpha}\|_q^q \leq 1$, pour des valeurs différentes de q et pour $p = 2$.

ellipses) et la contrainte (représentée par l'aire grise en forme de losange, pour le lasso, et de cercle, pour la pénalisation quadratique) étant strictement convexes, il existe une solution unique à ces deux problèmes qui se trouve sur la frontière de la région définie par la contrainte.

Coefficients nuls

La pénalisation quadratique correspond à l'unique valeur de q telle que le contour $\|\boldsymbol{\alpha}\|_q^q = \tau$, pour $\tau > 0$ quelconque, n'as pas d'"angles" (i.e. des points saillants et qui sont, donc, plus faciles à atteindre). Pour $q \neq 2$, les contours ont des "angles", mais ils ne sont situés sur les axes que pour $q < 2$. D'autre part, quand q décroît, il devient plus facile d'atteindre un des points sur les axes et les possibilités d'obtenir des coefficients nuls augmentent. Cela donne des intuitions sur le fait que pour $q \leq 1$ des coefficients peuvent être annulés exactement.

Dans la figure (3.1) la solution du lasso se trouve sur un des axes, annulant exactement le paramètre de la variable à importance faible, tandis que la pénalisation quadratique applique un rétrécissement proportionnel qui ne rend nul aucun des paramètres. La figure (3.2) montre les régions définies par les contraintes $\|\boldsymbol{\alpha}\|_q^q \leq 1$, dans le cas bi–dimensionnel, pour un éventail de valeurs de q, comprenant la sélection

de sous–ensembles, le lasso, la pénalisation quadratique, et des valeurs intermédiaires.

En effet, pour $q \leq 1$, il a été démontré, par des résultats asymptotiques, la capacité d'annuler exactement les coefficients dont la vraie valeur est 0 [Knight et Fu, 2000]. Des résultats non asymptotiques permettent également aux auteurs de conclure que les pénalisations du type $q > 1$ ont des avantages sur les pénalisations du type $q \leq 1$ seulement dans les cas où tous les vrais paramètres ont des valeurs faibles (relativement à n). En revanche, en présence de paramètres avec un effet important, le rétrécissement appliqué sera proportionnel à la taille des paramètres (comme déjà précisé pour la pénalisation quadratique), et donc un petit rétrécissement sera attribué aux petits paramètres. Contrairement, quand $q \leq 1$, les paramètres à valeurs faibles seront estimés nuls, même en présence de paramètres de grande taille.

Biais

Pour $q > 1$, le rétrécissement appliqué à un coefficient est proportionnel à la taille du vrai paramètre, alors pour des paramètres ayant un effet élevé, le biais de leur estimation sera trop important [Knight et Fu, 2000]. Pour le lasso ($q = 1$), le biais des estimations est plus "contrôlable" dans le sens où, pour τ fixé, le biais est borné par une constante qui dépend des données, mais qui est indépendante des vraies valeurs des coefficients α_j [Knight, 2004]. Pour $q < 1$, les paramètres non nuls peuvent être estimés sans biais (asymptotique) [Knight et Fu, 2000].

Interprétation bayésienne

D'un point de vu bayésien, $|\alpha_j|^q$ peut être interprété comme la log–densité *a priori* de α_j, les paramètres de régularisation admettent également une représentation bayésienne [Hastie *et al.*, 2001]. Pour la pénalisation quadratique, la distribution a priori est une normale centrée :

$$h(\alpha_j) = \frac{1}{\sqrt{2\pi}\tau} exp\left(-\alpha_j^2/2\tau^2\right), \tag{3.9}$$

où τ^2 est inversement proportionnel à μ. Pour $q \leq 1$, cette densité n'est pas uniforme sur les directions, elle concentre plus de masse sur les directions des axes. La densité *a priori* correspondant à $q = 1$ est une distribution de Laplace :

$$h(\alpha_j) = \frac{1}{2\tau} exp\left(-|\alpha_j|/\tau\right), \tag{3.10}$$

où τ est inversement proportionnel à μ [Tibshirani, 1996, Hastie *et al.*, 2001].

"Garrot non négatif"

La méthode "garrot non négatif" (*non–negative garrote*) [Breiman, 1995] est une modification de la pénalisation quadratique. La pénalisation appliquée ici inclue l'information apportée par l'estimation OLS :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \mu \sum_{j=1}^{p} \frac{\alpha_j^2}{(\alpha_j^{\mathrm{OLS}})^2}. \tag{3.11}$$

Les coefficients avec une estimation OLS petite, sont plus sévèrement pénalisés. Géométriquement, cela donne lieu à des régions elliptiques à la place des cercles.

Pénalisation multiple adaptative

La pénalisation multiple adaptative, AdR (*adaptive ridge regression*), est une modification de la pénalisation quadratique qui attribue à chaque coefficient une pénalisation en accord avec son importance [Grandvalet, 1998, Grandvalet et Canu, 1998] :

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2 \\ &\text{sous contraintes} \qquad \frac{1}{p}\sum_{j=1}^{p} \frac{1}{\mu_j} = \frac{1}{\mu}, \quad \mu_j > 0, \end{aligned} \tag{3.12}$$

où μ est prédéfini et les μ_j sont réglés automatiquement, à partir des données. Elle est basée sur le principe du garrot non négatif, mais ici, les μ_j sont choisis par la résolution du problème quadratique pénalisé :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2. \tag{3.13}$$

Cependant, une contrainte supplémentaire devient nécessaire, puisqu'une minimisation directe rétrécirait tous les paramètres à 0 :

$$\frac{1}{p}\sum_{j=1}^{p} \frac{1}{\mu_j} = \frac{1}{\mu}, \quad \mu_j > 0. \tag{3.14}$$

L'origine de cette contrainte est liée à l'interprétation bayésienne des moindres carrés pénalisés. Les paramètres suivent des lois *a priori* gaussiennes centrées :

$$h(\alpha_j) = \frac{1}{\sqrt{2\pi}\tau_j} exp\left(-\alpha_j^2/2\tau_j^2\right), \tag{3.15}$$

où les variances τ_j^2 sont inversement proportionnelles à μ_j [Grandvalet, 1998, Grandvalet et Canu, 1998]. La contrainte (3.14) relie donc les variances individuelles en imposant que la variance moyenne soit constante et inversement proportionnelle à μ. Ainsi, pour la pénalisation multiple adaptative, chaque coefficient a sa propre distribution *a priori*, ce qui est plus approprié quand les variables explicatives n'ont pas la même importance.

Pour la pénalisation multiple adaptative, comme pour la méthode garrot non négatif, une pénalisation quadratique pondérée conduit à des régions elliptiques. Quand l'ellipse est très étroite sur la direction d'un des paramètres, il est possible d'annuler ce paramètre.

La pénalisation multiple adaptative a été comparée à la sélection de sous–ensembles et la pénalisation quadratique, en termes d'erreur de prédiction [Boukari et Grandvalet, 1998]. Les auteurs concluent que la sélection de sous–ensembles est la mieux adaptée lorsque le nombre d'entrées significatives est très petit devant le nombre total de variables explicatives, et que ces variables sont peu corrélées. La pénalisation quadratique est appropriée quand la majorité des entrées sont significatives ou qu'elles sont très corrélées. Enfin, la pénalisation multiple adaptative donne les meilleurs résultats dans les cas intermédiaires. Toutefois, ses résultats sont proches de ceux de la meilleure méthode. Les auteurs attribuent la robustesse de cette méthode, par rapport à la sélection de sous–ensembles et à la pénalisation quadratique, au fait qu'elle fait à la fois de la sélection et du rétrécissement.

Pénalisation non concave

Avec l'objectif de réunir les différents avantages des méthodes précédentes, et, au même temps corriger le biais, [Fan et Li, 2001] proposent une pénalisation non concave, SCAD (*smoothly clipped absolute deviation penalty*) :

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \sum_{j=1}^{p} J_\mu(|\alpha_j|), \tag{3.16}$$

où J_μ est une fonction de pénalisation telle que

$$J_\mu^{(1)}(\theta) = \mathbb{I}_{\theta \leq \mu} + \frac{(a\mu - \theta)_+}{(a-1)\mu} \mathbb{I}_{\theta > \mu}, \tag{3.17}$$

pour $\theta > 0$, où $a > 2$ et μ sont des valeurs prédéfinies. Cette méthode comporte donc deux paramètres de la complexité.

Cette pénalisation permet d'annuler certains coefficients, et elle a des propriétés de régularité conduisant à des solutions stables. Aussi, cette méthode réduit le biais des estimations. Cependant, son implémentation est plus complexe, dû à la forme moins simple de l'estimateur, qui rend également la procédure moins intuitive.

3.2.1.2 Le lasso en détail

Parmi les différentes méthodes englobées par la *bridge regression*, le lasso est l'unique méthode strictement convexe, stable, raisonnablement biaisée qui sélectionne de variables. Ces bonnes propriétés ont motivé dans les dernières années des nombreuses études sur cette méthode.

Résultats théoriques

La consistance de l'estimateur lasso découle de sa distribution asymptotique, déduite par [Knight et Fu, 2000]. Il a également été démontré qu'il existe une valeur strictement positive du paramètre de réguralisation ($\mu > 0$), telle que l'erreur de

l'estimateur lasso est plus petite que celle de l'estimateur OLS ($\mu = 0$), au sens d'APE [Huang, 2003, Rosset et Zhu, 2003]. Dans le cas du lasso la démonstration de ce résultat n'est pas directe, contrairement à la pénalisation quadratique, qui admet une expression analytique du biais et la variance de l'estimateur.

Stabilité

Selon [Knight, 2004], la stabilité du lasso (et en général, des méthodes telles que $q \geq 1$) découle de l'association de la convexité stricte de l'erreur pénalisée comme fonction de $\boldsymbol{\alpha}$, pour μ fixé,

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 + \mu\|\boldsymbol{\alpha}\|_1, \tag{3.18}$$

et de la continuité en μ des solutions du lasso, notées $\boldsymbol{\alpha}^{\mathrm{L}}$:

$$\boldsymbol{\alpha}^{\mathrm{L}}(\mu) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 + \mu\|\boldsymbol{\alpha}\|_1. \tag{3.19}$$

La comparaison expérimentale de la stabilité de la sélection de sous–ensembles, de la pénalisation quadratique et du lasso a été effectuée par [Tibshirani, 1996]. Les résultats des simulations montrent la grande variabilité des coefficients estimés par la sélection de sous–ensembles, par rapport au lasso et la pénalisation quadratique.

Coefficients nuls

La géométrie du lasso explique sa capacité à annuler exactement des coefficients [Tibshirani, 1996]. Des résultats théoriques justifient également cette propriété [Knight et Fu, 2000].

Expérimentalement, [Tibshirani, 1996] met en évidence que le lasso sélectionne le modèle correct peu souvent, en revanche le modèle sélectionné contient le modèle correct dans la plupart des situations. Quant à la sélection de sous–ensembles, le modèle correct est sélectionné plus souvent, cependant, cette méthode élimine fréquemment des variables pertinentes.

Dans les applications pratiques du lasso, on peut trouver des exemples où la méthode élimine peu de variables [Steyerberg *et al.*, 2000] ainsi que des exemples où des modèles très parcimonieux sont obtenus [Li *et al.*, 2004].

Lasso VS pénalisation multiple adaptative

La pénalisation multiple adaptative et le lasso sont équivalents, dans le sens où ils génèrent les mêmes estimations [Grandvalet, 1998, Grandvalet et Canu, 1998]. La pénalisation multiple adaptative admet plusieurs formulations. En termes du lagrangien :

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j\right)^2 + \sum_{j=1}^{p}\mu_j\alpha_j^2 + \nu\left(\sum_{j=1}^{p}\frac{1}{\mu_j} - \frac{p}{\mu}\right), \tag{3.20}$$

où ν est le paramètre du Lagrangien. De façon équivalente,

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 \qquad \text{sous contrainte} \qquad \frac{1}{p} \left(\sum_{j=1}^{p} |\alpha_j| \right)^2 \leq \tau^2, \tag{3.21}$$

ou, encore,

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 + \frac{\mu}{p} \left(\sum_{j=1}^{p} |\alpha_j| \right)^2. \tag{3.22}$$

A partir de ces formulations, il a été établie l'équivalence entre les deux méthodes, pour un coût différentiable quelconque [Grandvalet, 1998] (voir section A.2, page 141).

Bien que le lasso (3.6) et la pénalisation multiple adaptative (3.12) soient équivalents, leur complexité n'est pas indexée de la même façon. Le paramètre de régularisation du lasso, τ, varie dans l'intervalle $[0, \|\boldsymbol{\alpha}^{\text{OLS}}\|_1]$, tandis que celui de la pénalisation multiple adaptative, μ, varie varie dans l'intervalle $[0, \infty)$. Le premier dépend de l'estimateur OLS, très sensible aux problèmes de conditionnement de la matrice des données. L'estimation de ce paramètre est donc plus difficile, et la procédure globale peut subir une perte de stabilité [Grandvalet et Canu, 1998].

Algorithmes

La résolution numérique du problème (3.6) n'est pas triviale. Il s'agit d'un problème d'optimisation convexe, ce qui implique l'existence d'une solution unique, qui se trouve sur la frontière de la région définie par la contrainte : $\|\boldsymbol{\alpha}\|_1 = \tau$. La fonction cible $\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2$ est quadratique, cependant, les contraintes d'inégalité sont non–linéaires et non–différentiables, rendant le problème difficile.

Un premier algorithme est proposé par [Tibshirani, 1996]. La reformalisation des contraintes : $\boldsymbol{\delta}_k^t \boldsymbol{\alpha} \leq \tau$, $k = 1, 2, \ldots, 2^p$, où $\boldsymbol{\delta}_k$ sont les 2^p vecteurs de dimension p dont les éléments sont $+1$ ou -1, permet l'application d'une méthodes avec activation de contraintes pour des contraintes d'inégalité linéaires (voir section A.1, page 137).

Le problème, exprimé en termes de pénalisation multiple adaptative (3.12), peut être résolu par un algorithme du type point fixe [Grandvalet, 1998, Grandvalet et Canu, 1998]. A chaque étape, on estime $\boldsymbol{\alpha}$ avec les paramètres des lois a priori μ_j fixés :

$$\boldsymbol{\alpha} = \left(\mathbf{X}^t\mathbf{X} + \mathbf{M} \right)^{-1} \mathbf{X}^t\mathbf{y}, \tag{3.23}$$

où $\mathbf{M} = \text{diag}(\mu_j)$. Les μ_j sont à leur tour estimés à $\boldsymbol{\alpha}$ fixé :

$$\mu_j = \mu \frac{\sum_{j=1}^{p} |\alpha_j|}{p|\alpha_j|}. \tag{3.24}$$

L'algorithme converge vers un minimum local, cependant, les conditions générales de convergence globale ne sont pas claires. Si des critères existent, ils dépendent des conditions initiales.

Les solutions du problème (3.7) peuvent être recherchées par la résolution du système [Fu, 1998] :

$$\frac{\partial(\mathbf{X}\boldsymbol{\alpha}-\mathbf{y})^t(\mathbf{X}\boldsymbol{\alpha}-\mathbf{y})}{\partial\alpha_j}+\mu q|\alpha_j|^{q-1}\text{sign}(\alpha_j)=0, \qquad j=1,\ldots,p, \tag{3.25}$$

au moyen d'une méthode de Newton–Raphson. Pour $1<q<2$, une modification est introduite afin d'assurer la convergence, car pour $q\leq 2$, $\mu q|\alpha_j|^{q-1}\text{sign}(\alpha_j)$ n'est pas différentiable en $\alpha_j=0$. Pour $q=1$, il est également utilisé que $\lim_{q\to 1^+}\widehat{\boldsymbol{\alpha}}(\mu,q)=\boldsymbol{\alpha}^{\text{L}}$.

Un algorithme efficace a été proposé par [Osborne *et al.*, 2000b, Osborne *et al.*, 2000a]. L'idée consiste à générer une direction de descente, $\mathbf{h}$, pour le $\boldsymbol{\alpha}$ courant, et résoudre une approximation "linéaire" :

$$\min_{\mathbf{h}} f(\boldsymbol{\alpha}+\mathbf{h}) \quad \text{sous contrainte} \quad [\text{sign}(\boldsymbol{\alpha})_\sigma]^t(\boldsymbol{\alpha}_\sigma+\mathbf{h}_\sigma)\leq\tau, \tag{3.26}$$

où $\sigma=\{j|\alpha_j\neq 0\}$, à la place du problème initial :

$$\min_{\mathbf{h}} f(\boldsymbol{\alpha}+\mathbf{h}) \quad \text{sous contrainte} \quad [\text{sign}(\boldsymbol{\alpha}+\mathbf{h})]^t(\boldsymbol{\alpha}+\mathbf{h})\leq\tau. \tag{3.27}$$

La fonction cible étant quadratique et les contraintes actives linéarisées localement, la résolution par programmation quadratique est possible (voir section A.1, page 137).

Finalement, l'algorithme lars, basé sur une version plus minutieuse de la sélection ascendante pas à pas, peut être modifié afin d'obtenir l'estimateur lasso [Efron *et al.*, 2004]. Cet algorithme est plus directe que le précédant, et il semble plus efficace sur des applications pratiques [Segal *et al.*, 2003].

Variance de l'estimateur

L'estimateur lasso est une fonction non–linéaire en $\mathbf{y}$, ce qui rend difficile l'obtention d'une estimation précise de sa variance.

Une première proposition est basée sur une reformalisation des solutions du lasso ($|\alpha_j^{\text{L}}|=(\alpha_j^{\text{L}})^2/|\alpha_j^{\text{L}}|$), qui permet une formulation de type pénalisation quadratique [Tibshirani, 1996] :

$$\boldsymbol{\alpha}^{\text{L}}=\left(\mathbf{X}^t\mathbf{X}+\mu\mathbf{A}^-\right)^{-1}\mathbf{X}^t\mathbf{y} \tag{3.28}$$

où $\mathbf{A}=\text{diag}(|\alpha_j^{\text{L}}|)$, et $\mathbf{A}^-$ indique la pseudo–inverse.

La matrice de covariance peut alors être approchée par :

$$\text{Var}(\boldsymbol{\alpha}^{\text{L}})=\left(\mathbf{X}^t\mathbf{X}+\mu\mathbf{A}^-\right)^{-1}\mathbf{X}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}+\mu\mathbf{A}^-\right)^{-1}\widehat{\sigma}^2, \tag{3.29}$$

où $\widehat{\sigma}^2$ est une estimation de la variance de l'erreur. Cette approximation implique que la variance des coefficients estimés nuls est 0. Une deuxième approximation, rendant des variances positives pour toutes les estimations est proposée par [Osborne *et al.*, 2000b] :

$$\text{Var}(\boldsymbol{\alpha}^{\text{L}})=\left(\mathbf{X}^t\mathbf{X}+\frac{\mathbf{X}^t\mathbf{r}\mathbf{r}^t\mathbf{X}}{\|\boldsymbol{\alpha}^{\text{L}}\|_1\|\mathbf{X}^t\mathbf{r}\|_\infty}\right)^{-1}\mathbf{X}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}+\frac{\mathbf{X}^t\mathbf{r}\mathbf{r}^t\mathbf{X}}{\|\boldsymbol{\alpha}^{\text{L}}\|_1\|\mathbf{X}^t\mathbf{r}\|_\infty}\right)^{-1}\widehat{\sigma}^2, \tag{3.30}$$

où $\mathbf{r} = \mathbf{X}\boldsymbol{\alpha}^{\mathrm{L}} - \mathbf{y}$. Cette estimation suppose que les estimateurs sont approximativement des transformations linéaires, ce qui n'est pas le cas. Une approche alternative consiste à appliquer une méthode bootstrap [Knight et Fu, 2000].

Nombre effectif de paramètres

La régression pénalisation quadratique s'écrit

$$\widehat{\mathbf{f}} = \mathbf{H}_{\mu}\mathbf{y} = \mathbf{X}\boldsymbol{\alpha}^{\mathrm{RR}} = \mathbf{X}(\mathbf{X}^t\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y}, \tag{3.31}$$

ce qui permet de calculer simplement le nombre degrés de liberté comme la trace de la matrice chapeau, $\mathrm{ddl} = \mathrm{tr}\,[\mathbf{H}_{\mu}]$.

La reformalisation des solutions du lasso (3.28) permet l'estimation de ddl suivante [Tibshirani, 1996] :

$$\mathrm{ddl}(\mu) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \mu\mathbf{A}^{-}\right)^{-1}\mathbf{X}^t\right]. \tag{3.32}$$

Pour des problèmes basés sur un coût différent, tel que la log–vraisemblance (par exemple, le modèle logistique) ou log–vraisemblance partielle (par exemple, le modèle de Cox), l'estimation du nombre de degrés de liberté incorpore la matrice de pondérations, $\mathbf{W}$, obtenue à la dernière itération de l'algorithme IRLS :

$$\mathrm{ddl}(\mu) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mu\mathbf{A}^{-}\right)^{-1}\mathbf{X}^t\mathbf{W}\right]. \tag{3.33}$$

Dans ces estimations, la pénalisation sur les variables jugées non pertinentes n'est pas prise en compte. Une modification est proposée [Fu, 1998][2] :

$$\mathrm{ddl}(\mu) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \mu\mathbf{A}^{-}\right)^{-1}\mathbf{X}^t\right] - p_0, \tag{3.34}$$

où p_0 est le nombre de coefficients estimés nuls.

Paramètre de régularisation

Les résultats obtenus par [Huang, 2003, Rosset et Zhu, 2003] témoignent de l'importance de bien choisir le paramètre de régularisation. En effet, l'estimateur moindres carrés peut toujours être amélioré, en termes de l'erreur quadratique de prédiction, en lui appliquant une proportion adéquate de rétrécissement de type lasso.

Il a été démontré que la trajectoire des solutions optimales du lasso (3.19), comme fonction de μ, est linéaire par morceaux : il existe $\infty > \mu_0 > \mu_1 > \ldots > \mu_m = 0$ tels que $\forall\mu$, $\mu_k \geq \mu \geq \mu_{k+1}$, il est vérifié

$$\boldsymbol{\alpha}^{\mathrm{L}}(\mu) = \boldsymbol{\alpha}^{\mathrm{L}}(\mu_k) - (\mu - \mu_k)\gamma_k, \tag{3.35}$$

où γ_k est la direction de la k–ième itération de l'algorithme lars pour le lasso [Efron *et al.*, 2004, Rosset et Zhu, 2004]. Ce résultat est généralisable aux fonctions

[2]Par uniformité des notations, une ré–paramétrisation de la définition de [Fu, 1998], qui intègre les constantes dans le paramètre de régularisation, a été appliquée.

de coût convexes, deux fois différentiables presque partout, tel que le coût de Huber ou des critères basés sur la marge (*hinge loss*).

Une conséquence de ce résultat est que le lasso peut être résolu de façon efficace pour toutes les valeurs de $\mu \in [0, \infty)$, en utilisant un algorithme incrémental. Ensuite l'hyper–paramètre optimal peut être estimé par une méthode de sélection de la complexité.

Les méthodes CV et GCV :

$$\mathrm{GCV}(\mu) = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})}{n \left(1 - \mathrm{ddl}(\mu)/n\right)^2}, \tag{3.36}$$

dans le cas où les entrées sont aléatoires, ainsi qu'un estimateur analytique non–biaisé du risque, dans le cas où les entrées sont fixes, sont proposées par [Tibshirani, 1996] pour estimer l'hyper–paramètre. Les résultats des expériences montrent que la GCV est la méthode la plus performante.

Des techniques de rééchantillonnage (*bootstrap* et CV) ont été également testées expérimentalement par [Boukari et Grandvalet, 1998]. Bien que le *bootstrap* (.632) soit, en général, le meilleur critère, les techniques de validation croisée ont des performances du même ordre alors qu'elles demandent moins de calculs.

Finalement, une statistique de type C_p a été proposée par [Efron *et al.*, 2004]. Des réflexions sur sa pertinence ont été apportées par [Ishwaran, 2004, Loubes et Massart, 2004, Stine, 2004, Weisberg, 2004].

Extensions du lasso

Des extensions aux modèles linéaires généralisés [Tibshirani, 1996, Klinger, 2001] et au modèle de Cox [Tibshirani, 1997] ont été proposées pour le lasso, ainsi que pour la *bridge regression* [Fu, 2003].

Le cas à réponses multiples a été étudié par [Turlach *et al.*, 2001]. L'objectif ici est la recherche d'un sous–ensemble de variables explicatives commun à toutes les variables réponse. L'implémentation du lasso a été également adaptée au cas $n < p$ [Osborne *et al.*, 2000b, Osborne *et al.*, 2000a, Efron *et al.*, 2004], une situation commune dans des applications telles que l'analyse de biopuces en génomique [Segal *et al.*, 2003, Ghosh *et al.*, 2003].

La pénalisation l_1 de fonctions de coût différentes au traditionnel coût quadratique, notamment des fonctions de coût robustes, tel que le coût de Huber ou l_1, a été étudiée par [Bakin, 1999, Roth, 2001, Rosset et Zhu, 2004].

3.2.2 Modèles additifs

La sélection de variables pour les modèles additifs se réduit, jusqu'à présent, à la sélection de sous–ensembles. La sélection de sous–ensembles appliquée aux modèles additifs exploite le fait que la régression additive généralise la régression linéaire. Dans le contexte linéaire, comme précisé précédemment, (section 3.2.1.1, page 78), la sélection de sous–ensembles donne lieu à des modèles simples et interprétables, mais des problèmes combinatoires et d'instabilité sont rencontrés. L'application de

ces techniques aux modèles additifs comporte de nouveaux problèmes : ne seulement il faut choisir les composantes à inclure dans le modèle, mais aussi leur proportion de lissage. Par conséquent, ces méthodes sont réduites aux cas avec peu de variables en entrée.

Concernant les procédures de sélection, celles du type pas à pas descendantes (par exemple, [Brumback *et al.*, 1999]) sont moins exposées aux problèmes de sous–estimation. Toutefois, elles semblent moins adéquates pour les modèles non–paramétriques, car elles supposent l'estimation d'un modèle complet, et donc la sélection de la complexité, en tant que problème p–dimensionnel. En revanche, les procédures de type pas à pas ascendantes (par exemple, [Chambers et Hastie, 1993]), ont à traiter avec un modèle complet, seulement dans le cas où $p-1$ variables ont été considérées pertinentes.

Des différentes stratégies, plus ou moins minutieuses, peuvent être considérées pour les procédures de type pas à pas ascendantes. Par exemple, la complexité d'une composante sélectionnée peut être estimée une seule fois, au moment de son inclusion dans le modèle, ce qui implique un problème de sélection de la complexité uni–dimensionnel à chaque étape d'inclusion. Une autre possibilité consiste à estimer la complexité de chacune des composantes sélectionnées et celle de la nouvelle composante conjointement, ce qui implique un problème multi–dimensionnel à chaque étape d'inclusion (si $q < p$ variables ont été sélectionnées, à la prochaine étape d'inclusion, le problème de sélection de la complexité sera dimension $q+1$). Les détails concernant les différentes stratégies sont, néanmoins, rarement spécifiés dans la littérature.

Quant aux critères de sélection, les méthodes de sélection de la complexité basées sur l'évaluation des points d'une grille (section 2.4.1, page 61 et section 2.4.2, page 64), ainsi que celles basées sur des tests (section 2.4.4, page 73), peuvent intégrer la sélection de variables, considérant le cas $\mathrm{ddl}_j = 0$. Les critères GCV ou AIC [Hastie et Tibshirani, 1990] ou les tests proposés par [Hastie et Tibshirani, 1990, Cantoni et Hastie, 2002, Härdle *et al.*, 2004a] sont utilisés de cette façon.

Il est également possible d'utiliser ces critères pour la sélection de variables et la sélection de la complexité séparément, en deux étapes différentes. Ainsi, pour la sélection de variables, des critères tels que AIC [Chambers et Hastie, 1993, Brumback *et al.*, 1999], GCV, BIC [Brumback *et al.*, 1999], ou CV (en tant qu'estimateur d'ISE) [Vieu, 1994, Härdle *et al.*, 2004b], sont utilisés comme critères d'arrêt. Un critère différent, basé sur la comparaison des valeurs $\|\mathbf{f}_j\|_\infty$, est proposé par [Härdle et Korostelev, 1996].

Aussi, des tests ont été proposés spécifiquement pour la sélection de variables. Par exemple, des tests de type χ^2 pour les splines pénalisés [Wood, 2000], des tests basés sur l'association du critère GCV et une méthode bootstrap pour les splines [Chen, 1993], ainsi que des tests basés sur l'idée que, une valeur élevée de (l'estimation de) $\mathbb{E}[f_j^2(X_j)]$, traduit une influence importante de la variable X_j, pour l'intégration marginale [Chen *et al.*, 1996].

Des approches bayésiennes à la sélection de variables ont été également développées [Smith et Kohn, 1996, Shively *et al.*, 1999].

3.3 Modèles additifs parcimonieux

La résolution du problème (1.47) nécessite de définir au préalable p paramètres de lissage λ_j, qui vont régler la complexité des fonctions f_j. Ce pré–requis n'est pas réaliste quand le nombre de variables est important.

Dans le cas des splines cubiques de lissage, le problème a été abordé par le biais de la pénalisation adaptative [Grandvalet, 1998, Grandvalet et Canu, 1998], qui incorpore l'estimation des λ_j dans la procédure d'estimation des paramètres :

$$\min_{\substack{\alpha_0, f_1, \ldots, f_p \\ \lambda_1, \ldots, \lambda_p}} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int \left[f_j^{(2)}(t) \right]^2 dt, \tag{3.37}$$

$$\text{sous contraintes} \quad \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{1}{\lambda}, \quad \lambda_j > 0, \tag{3.38}$$

pour α_0 scalaire et $f_j \in \mathcal{C}^2$ tel que $\mathbb{E}[f_j(X_j)] = 0$. Seul λ doit être défini avant la procédure d'estimation. Matriciellement, pour la base des B–splines naturelles,

$$\min_{\substack{\alpha_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p \\ \lambda_1, \ldots, \lambda_p}} \left(\mathbf{y} - \alpha_0 - \sum_{j=1}^{p} \mathbf{N}_j \boldsymbol{\beta}_j \right)^t \left(\mathbf{y} - \alpha_0 - \sum_{j=1}^{p} \mathbf{N}_j \boldsymbol{\beta}_j \right) + \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j, \tag{3.39}$$

$$\text{sous contraintes} \quad \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{1}{\lambda}, \quad \lambda_j > 0. \tag{3.40}$$

Le problème (3.39)–(3.40), qui peut être motivé par une approche bayésienne hiérarchique, est également fortement lié à la pénalisation l_1. En effet, le terme de pénalisation $\sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j}$ généralise la pénalisation du lasso $\sum_{j=1}^{p} |\alpha_j| = \sum_{j=1}^{p} \sqrt{\alpha_j^2}$. Ce terme résume la non–linéarité de f_j, il peut être donc interprété comme un indice de la pertinence. Les solutions de ce problème ont tendance à être parcimonieuses, en ce sens que, pour certaines variables, $\widehat{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widehat{\boldsymbol{\beta}}_j = 0$. Cependant ce critère ne sélectionne pas de variables, car la composante linéaire de $\widehat{f_j}$ appartient au noyau de $\boldsymbol{\Omega}_j$. Ainsi, même si $\widehat{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widehat{\boldsymbol{\beta}}_j = 0$, la j–ième variable n'est pas éliminée mais linéarisée.

3.3.1 Principe de décomposition

Les modèles additifs ajustés par des splines cubiques de lissage s'inscrivent dans le cadre théorique des espaces hilbertiens des fonctions L_2 (voir section 1.3.4.2, page 32). Cela permet d'appliquer des résultats généraux.

Soit $\mathcal{H}_j$ l'espace de Hilbert des fonctions mesurables centrées de variance finie, et de produit scalaire défini par $\langle f, g \rangle = \mathbb{E}_{X_j}(f(X_j) \cdot g(X_j))$. Chaque sous–espace $\mathcal{H}_j$ admet une décomposition $\mathcal{H}_j^L \oplus \mathcal{H}_j^{NL}$, où L indique le sous–espace des composantes linéaires et NL indique le sous–espace des composantes non linéaires. Cette

décomposition apparaît clairement sur la base des polynômes par morceaux pour les splines cubiques (1.15)–(1.17), page 19.

L'idée consiste donc à considérer les parties linéaire et non linéaire séparément. En ajoutant en (3.39)–(3.40) un terme de pénalisation agissant sur la composante linéaire, il est possible de supprimer l'influence de certaines variables sur le modèle. Le problème d'optimisation s'écrit :

$$\min_{\substack{\alpha_0,\alpha_1,\ldots,\alpha_p,\widetilde{\boldsymbol{\beta}}_1,\ldots,\widetilde{\boldsymbol{\beta}}_p \\ \mu_1,\ldots,\mu_p,\lambda_1,\ldots,\lambda_p}} \left\| \mathbf{y} - \alpha_0 - \sum_{j=1}^{p} \mathbf{x}_j \alpha_j - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j \right\|_2^2 + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \sum_{j=1}^{p} \lambda_j \widetilde{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j, \tag{3.41}$$

sous contraintes

$$\sum_{j=1}^{p} \frac{1}{\mu_j} = \frac{p}{\mu}, \quad \mu_j > 0, \quad \sum_{j=1}^{p} \frac{1}{\lambda_j} = \frac{p}{\lambda}, \quad \lambda_j > 0, \quad \mathbf{x}_j^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0, \quad \mathbf{1}^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0, \tag{3.42}$$

où μ et λ sont les paramètres qui règlent la complexité du modèle, et $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^t$, $\widetilde{\boldsymbol{\beta}}_j$ sont les coefficients de la partie linéaire et non linéaire, respectivement. Les contraintes $\mathbf{x}_j^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0$ et $\mathbf{1}^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0$ assurent l'orthogonalité des sous–espaces linéaires et non–linéaires.

Quand, après convergence, $1/\mu_j = 0$ et $1/\lambda_j = 0$, la j–ème variable est éliminée. Si $1/\mu_j > 0$ et $1/\lambda_j = 0$, la j–ème variable est linéairisée. Lorsque $1/\mu_j = 0$ et $1/\lambda_j > 0$, la j–ème variable est estimée strictement non linéaire.

La parcimonie du problème (3.41)–(3.42) découle de l'équivalence entre pénalisation multiple adaptative et pénalisation l_1 [Grandvalet et Canu, 1998] (voir section 3.2.1.2).

Dans le cas des modèles additifs généralisés, le problème d'optimisation s'écrit :

$$\min_{\substack{\boldsymbol{\alpha},\widetilde{\boldsymbol{\beta}}_1,\ldots,\widetilde{\boldsymbol{\beta}}_p \\ \mu_1,\ldots,\mu_p;\lambda_1,\ldots,\lambda_p}} - l(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_p) + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \sum_{j=1}^{p} \lambda_j \widetilde{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j, \tag{3.43}$$

sous les contraintes (3.42), où l indique la log–vraisemblance. Dans le cas particulier du modèle logistique,

$$\min_{\substack{\boldsymbol{\alpha},\widetilde{\boldsymbol{\beta}}_1,\ldots,\widetilde{\boldsymbol{\beta}}_p \\ \mu_1,\ldots,\mu_p,\lambda_1,\ldots,\lambda_p}} - \sum_{i=1}^{n} y_i \log P_i + (1 - y_i) \log(1 - P_i) + \sum_{j=1}^{p} \mu_j \alpha_j^2 + \sum_{j=1}^{p} \lambda_j \widetilde{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j, \tag{3.44}$$

sous les contraintes (3.42), pour

$$P_i = \frac{\exp[\alpha_0 + \sum_{j=1}^{p} x_{ij} \alpha_j - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\boldsymbol{\beta}_j}]}{1 + \exp[\alpha_0 + \sum_{j=1}^{p} x_{ij} \alpha_j - \sum_{j=1}^{p} \mathbf{N}_j \widetilde{\boldsymbol{\beta}_j}]}. \tag{3.45}$$

Algorithme du point fixe :

1. Fixer μ et λ et initialiser $\widetilde{\boldsymbol{\beta}}_j$, μ_j, λ_j, $j = 1, \ldots, p$.
2. Décomposition en valeurs singulières :
   (a) Décomposition en valeurs propres : $\boldsymbol{\Omega}_j = \mathbf{P}_j \mathbf{D}_j \mathbf{P}_j^t$.
   (b) Remplacement des valeurs propres correspondant aux fonctions linéaires et constantes par une valeur positive (1 par défaut).
   (c) Notation : $\mathbf{Q}_j = \mathbf{N}_j \mathbf{P}_j \mathbf{D}_j^{-1/2}$.
   (d) Décomposition en valeurs singulières : $\mathbf{Q}_j = \mathbf{U}_j \mathbf{Z}_j \mathbf{V}_j^t$.
   (e) Matrices chapeau :
   $\mathbf{H}_j = [\mathbf{1}\ \mathbf{x}_j]([\mathbf{1}\ \mathbf{x}_j]^t[\mathbf{1}\ \mathbf{x}_j]^t)^{-1}[\mathbf{1}\ \mathbf{x}_j]^t, \qquad \mathbf{H} = [\mathbf{1}\ \mathbf{X}]\left([\mathbf{1}\ \mathbf{X}]^t[\mathbf{1}\ \mathbf{X}]\right)^{-1}[\mathbf{1}\ \mathbf{X}]^t,$
   (f) Matrices de lissage et de rétrécissement :
   $\mathbf{S}_j = \mathbf{U}_j \mathbf{Z}_j \left(\mathbf{Z}_j^t \mathbf{Z}_j + \lambda_j \mathbf{I}\right)^{-1} \mathbf{Z}_j^t \mathbf{U}_j^t, \qquad \widetilde{\mathbf{S}}_j = \mathbf{S}_j - \mathbf{H}_j.$
3. Composantes linéaires :
   (a) Estimation des coefficients :
   $\boldsymbol{\alpha} = \left([\mathbf{1}\ \mathbf{X}]^t[\mathbf{1}\ \mathbf{X}] + \mathbf{M}\right)^{-1}[\mathbf{1}\ \mathbf{X}]^t\mathbf{y}$, où $\mathbf{M} = \mathrm{diag}[0, \mu_1, \ldots, \mu_p]$.
   (b) Ré–estimation des termes de pénalisation : $\mu_j = \mu \dfrac{\sum_{k=1}^{p} |\alpha_k|}{p|\alpha_j|}$.
   (c) Itérer 3.(a) et 3.(b) jusqu'à convergence.
4. Composantes non linéaires :
   (a) Estimation des coefficients par backfitting :
      i. Résidus partiels : $\widetilde{\mathbf{r}}_j = \mathbf{y} - \mathbf{H}\mathbf{y} - \sum_{k \neq j} \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j$.
      ii. Coefficients : $\mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = \widetilde{\mathbf{S}}_j \widetilde{\mathbf{r}}_j$.
   (b) Ré–estimation des termes de pénalisation : $\lambda_j = \lambda \dfrac{\sum_{j=1}^{p} \sqrt{\widetilde{\boldsymbol{\beta}}_j \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j}}{p\sqrt{\widetilde{\boldsymbol{\beta}}_j \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j^t}}$.
   (c) Itérer 4.(a) et 4.(b) jusqu'à convergence.

FIG. 3.3 – Algorithme de résolution du problème de minimisation quadratique sous contraintes.

3.3.2 Estimation

Nous présentons tout d'abord la procédure d'estimation pour le problème de type gaussien (3.41)–(3.42). Celui–ci consiste à englober l'algorithme de backfitting dans un algorithme de point fixe.
Ensuite, nous présentons l'algorithme d'estimation pour le modèle additif logistique (3.44)–(3.42). Celui–ci englobe l'algorithme pour le problème gaussien dans un algorithme IRLS.

3.3.2.1 Modèles additifs

Application du principe de décomposition

La matrice de lissage des splines cubiques, $\mathbf{S}_j = \mathbf{N}_j(\mathbf{N}_j^t\mathbf{N}_j + \lambda_j\boldsymbol{\Omega}_j)^{-1}\mathbf{N}_j^t$, a deux valeurs propres égales à 1, correspondant aux fonctions propres constante et linéaire, et $n-2$ valeurs propres dans l'intervalle $]0, 1[$, correspondant aux fonctions d'ordre supérieur (voir section 1.2.3.2, page 21). Aussi, $\mathbf{S}_j$ est symétrique, ce qui permet la décomposition suivante $\mathbf{S}_j = \mathbf{H}_j + \widetilde{\mathbf{S}}_j$ [Hastie et Tibshirani, 1990], où $\mathbf{H}_j$ est la

matrice de projection sur l'espace des fonctions propres constante et linéaires (la matrice chapeau correspondant à la régression moindres carrés sur $[\mathbf{1}, \mathbf{x}_j]$), et $\widetilde{\mathbf{S}}_j$ est la matrice de rétrécissement, correspondant à l'espace des fonctions propres d'ordre supérieur.

L'intégration de cette décomposition de la matrice dans l'algorithme backfitting, permet de différencier deux étapes : 1. estimation de la partie projection, $\mathbf{g} = \mathbf{H}(\mathbf{y} - \sum \widetilde{\mathbf{f}}_j)$, où $\mathbf{H}$ est la matrice chapeau correspondant à la régression par moindres carrés sur $[\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p]$, et 2. estimation de la partie rétrécissement $\widetilde{\mathbf{f}}_j = \widetilde{\mathbf{S}}_j(\mathbf{y} - \mathbf{g} - \sum_{k \neq j} \widetilde{\mathbf{f}}_k)$. L'estimation de la fonction globale est donnée par $\widehat{\mathbf{f}} = \mathbf{g} + \sum \widetilde{\mathbf{f}}_j$ (voir section 1.3.5.1, page 38).

D'autre part, on peut observer que les composantes linéaires et non linéaires se trouvent effectivement dans des espaces orthogonaux : $\mathbf{H}\widetilde{\mathbf{f}}_j = \mathbf{0}$ and $\widetilde{\mathbf{S}}_j\mathbf{g} = \mathbf{0}$ [Avalos *et al.*, 2003]. Les étapes 3. et 4. sont ainsi complètement indépendantes, ce qui implique, en particulier, que lorsque plusieurs valeurs de μ (m_μ valeurs) et λ (m_λ valeurs) sont évaluées, la taille de la grille n'est pas quadratique ($m_\mu \times m_\lambda$) mais linéaire ($m_\mu + m_\lambda$) .

Décomposition en valeurs singulières

Considérons la décomposition en valeurs propres de $\mathbf{\Omega}_j$: $\mathbf{\Omega}_j = \mathbf{P}_j\mathbf{D}_j\mathbf{P}_j^t$, où $\mathbf{D}_j$ est une matrice diagonale constituée des valeurs propres de la matrice de pénalisation, et $\mathbf{P}_j$ est une matrice orthonormale ($\mathbf{P}_j^t\mathbf{P}_j = \mathbf{P}_j\mathbf{P}_j^t = \mathbf{I}_{n+2}$) dont les colonnes sont les vecteur propres correspondants.

Définissons la matrice $\mathbf{Q}_j = \mathbf{N}_j\mathbf{P}_j\mathbf{D}_j^{-1/2}$, de dimension $n \times n+2$. Soit $\mathbf{Q}_j = \mathbf{U}_j\mathbf{Z}_j\mathbf{V}_j^t$ sa décomposition en valeurs singulières, où $\mathbf{Z}_j$ est une matrice diagonale de la même dimension que $\mathbf{Q}_j$ et avec des éléments non négatifs sur la diagonale, ζ_{ij}, en ordre décroissant ; $\mathbf{U}_j$ et $\mathbf{V}_j$ sont des matrices orthonormales ($\mathbf{U}_j^t\mathbf{U}_j = \mathbf{U}_j\mathbf{U}_j^t = \mathbf{I}_n$, $\mathbf{V}_j^t\mathbf{V}_j = \mathbf{V}_j\mathbf{V}_j^t = \mathbf{I}_{n+2}$).

Alors la matrice de lissage pour la j–ème composante additive s'écrit [Avalos *et al.*, 2004c] :

$$\mathbf{S}_j = \mathbf{U}_j\mathbf{Z}_j(\mathbf{Z}_j^t\mathbf{Z}_j + \lambda_j\mathbf{I})^{-1}\mathbf{Z}_j^t\mathbf{U}_j^t. \tag{3.46}$$

Cette décomposition suppose que $\mathbf{\Omega}_j$ est de rang plein, ce qui n'est pas le cas, puisque la dérivée seconde des fonctions linéaires et constante est la fonction nulle. Cependant, si on ajoute une pénalisation sur la partie linéaire, alors $\mathbf{\Omega}_j$ devient de rang plein. Ceci n'a pas d'effet sur l'estimation des coefficients car les parties linéaires et non linéaires sont traitées de façon indépendante.

On veut donc $\mathbf{\Omega}'_j$ de rang plein, tel que $\mathbf{S}'_j = \mathbf{N}_j(\mathbf{N}_j^t\mathbf{N}_j + \lambda_j\mathbf{\Omega}'_j)^{-1}\mathbf{N}_j^t$ a les mêmes valeurs propres et vecteurs propres que $\mathbf{S}_j$ sauf sur l'espace $[\mathbf{1}, \mathbf{x}_j]$, correspondant aux fonctions propres constante et linéaire. Prenant en compte que $\mathbf{x}_j$ est centrée réduite,

la matrice de pénalisation recherchée est, dans ce cas,

$$\begin{aligned}\boldsymbol{\Omega}'_j &= \boldsymbol{\Omega}_j + \mathbf{N}_j^t \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_2} \frac{\mathbf{x}_j^t}{\|\mathbf{x}_j\|_2} \mathbf{N}_j + \mathbf{N}_j^t \frac{\mathbf{1}}{\|\mathbf{1}\|_2} \frac{\mathbf{1}^t}{\|\mathbf{1}\|_2} \mathbf{N}_j \\ &= \boldsymbol{\Omega}_j + \frac{1}{n} \mathbf{N}_j^t \mathbf{x}_j \mathbf{x}_j^t \mathbf{N}_j + \frac{1}{n} \mathbf{N}_j^t \mathbf{1}\mathbf{1}^t \mathbf{N}_j.\end{aligned} \tag{3.47}$$

Composantes linéaires

L'estimation de composantes linéaires comporte l'estimation des fonctions linéaires et constante. Afin d'incorporer l'estimation de cette dernière, la matrice des données précédée d'une colonne de uns, $[\mathbf{1}\ \mathbf{X}]$, est utilisée (étape 3.(a)). Cependant, α_0 n'étant pas pénalisé, le premier élément de la matrice diagonale $\mathbf{M}$ est 0 (étape 3.(b)).

L'algorithme de point fixe (figure 3.3) résout le problème d'estimation des paramètres de pénalisation des composantes linéaires (étape 3.(b)) [Grandvalet et Canu, 1998].

Le critère d'arrêt (étape 3.(c)) porte sur la convergence des coefficients $\boldsymbol{\alpha}$, le maximum des variations absolues et relatives entre deux itérations,

$$\max_{j=1,\ldots,p} \left(\frac{|\alpha_j^{[k]} - \alpha_j^{[k-1]}|}{1 + |\alpha_j^{[k]}|} \right), \tag{3.48}$$

où k indique l'itération, devant être inférieur à 10^{-6}.

Composantes non linéaires

L'algorithme backfitting est utilisé pour ajuster les composantes non linéaires (étape 4.(a)). Le calcul explicite des coefficients $\widetilde{\boldsymbol{\beta}}_j$ (étape 4.(a)ii.) n'est pas nécessaire pour l'estimation de $\widetilde{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j$ (étape 4.(b)). Ce dernier peut être calculé directement comme $\widetilde{\boldsymbol{\gamma}}_j^t \widetilde{\boldsymbol{\gamma}}_j = \widetilde{\mathbf{r}}_j^t \mathbf{U}_j \mathbf{Z}_j (\mathbf{Z}_j^t \mathbf{Z}_j + \lambda_j \mathbf{I})^{-2} \mathbf{Z}_j^t \mathbf{U}_j^t \widetilde{\mathbf{r}}_j$, où $\widetilde{\boldsymbol{\gamma}}_j = \boldsymbol{\Omega}_j^{1/2} \widetilde{\boldsymbol{\beta}}_j$ et $\widetilde{\mathbf{r}}_j$ sont les résidus partiels non linéaires.

L'algorithme de point fixe (figure 3.3) résout, également, le problème d'estimation des paramètres de pénalisation des composantes non linéaires (étape 4(b)) [Grandvalet et Canu, 1998].

La décomposition en valeurs singulières (étapes 2(a)–2(d)) permet de simplifier les calculs, en évitant, en particulier, l'inversion de matrices. Aussi, la plupart des calculs demandés par la décomposition en valeurs singulières ne dépendent pas de λ. Les méthodes de sélection de la complexité basées sur l'évaluation sur un grille de valeurs μ et λ y trouvent aussi avantage.

Le critère d'arrêt (étape 4.(c)) porte sur la convergence des coefficients $\widetilde{\boldsymbol{\gamma}}_j$ (ou de façon équivalente sur des coefficients $\widetilde{\boldsymbol{\beta}}_j$), le maximum des normes quadratiques des

Algorithme IRLS :

1. Fixer μ et λ et initialiser α_0, α_j, $\widetilde{\boldsymbol{\beta}}_j$, μ_j, λ_j,
   $\mathbf{f}_j = \mathbf{x}_j\alpha_j + \widetilde{\mathbf{f}}_j$, $\widetilde{\mathbf{f}}_j = \mathbf{N}_j\widetilde{\boldsymbol{\beta}}_j$, $f_{ij} = \mathrm{x}_{ij}\alpha_j + \widetilde{f}_j(\mathrm{x}_{ij})$, $j = 1, \ldots, p$, $i = 1, \ldots, n$.
2. Calculer :
   (a) $\mathbf{P} = (P_1, \ldots, P_n)^t$, où $P_i = \dfrac{\exp(\alpha_0 + \sum_{j=1}^p f_{ij})}{1 + \exp(\alpha_0 + \sum_{j=1}^p f_{ij})}$.
   (b) $\mathbf{W} = \mathrm{diag}\,[P_1(1-P_1), \ldots, P_n(1-P_n)]$.
   (c) $\mathbf{z} = \alpha_0 + \sum_{j=1}^p \mathbf{f}_j + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{P})$.
3. Optimiser
   $$\min_{\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_p} \left|\left| \mathbf{W}^{1/2}\Big(\mathbf{z} - [\mathbf{1}\ \mathbf{X}]\boldsymbol{\alpha} - \sum_{j=1}^p \mathbf{N}_j\widetilde{\boldsymbol{\beta}}_j\Big) \right|\right|_2^2 + \sum_{j=1}^p \mu_j\alpha_j^2 + \sum_{j=1}^p \lambda_j \left(\widetilde{\boldsymbol{\beta}}_j\right)^t \boldsymbol{\Omega}_j \left(\widetilde{\boldsymbol{\beta}}_j\right)$$
   sous contraintes $\sum_{j=1}^p \dfrac{1}{\mu_j} = \dfrac{p}{\mu}$, $\mu_j > 0$, $\sum_{j=1}^p \dfrac{1}{\lambda_j} = \dfrac{p}{\lambda}$, $\lambda_j > 0$, $\mathbf{x}_j^t\mathbf{N}_j\widetilde{\boldsymbol{\beta}}_j = 0$, $\mathbf{1}^t\mathbf{N}_j\widetilde{\boldsymbol{\beta}}_j = 0$.
4. Itérer 2. et 3. jusqu'à convergence.

FIG. 3.4 – Algorithme d'estimation du modèle logistique additif parcimonieux.

variations absolues et relatives entre deux itérations,

$$\max_{j=1,\ldots,p} \left(\frac{\|\widetilde{\boldsymbol{\gamma}}_j^{[k]} - \widetilde{\boldsymbol{\gamma}}_j^{[k-1]}\|_2^2}{1 + \|\widetilde{\boldsymbol{\gamma}}_j^{[k]}\|_2^2} \right), \tag{3.49}$$

devant être inférieur à 10^{-6}.

3.3.2.2 Modèles additifs généralisés : logistique

La procédure IRLS (figure 3.4) résout le problème de minimisation quadratique pondéré (3.44) sous contraintes (3.42) [Avalos *et al.*, 2004a, Avalos *et al.*, 2004b]. L'étape 3, détaillée sur la figure (3.5), est résolue par la version pondérée de l'algorithme décrit pour les modèles additifs dans la section précédente. Cependant, le problème de minimisation quadratique de l'algorithme itératif IRLS présente de nouvelles difficultés : les estimations des coefficients linéaires et non linéaires ne sont plus indépendantes car elles interagissent par le biais de la matrice de pondérations $\mathbf{W}$. D'une part, la quantité de calculs s'en trouve augmentée. D'autre part, la décomposition en valeurs singulières est moins intéressante car elle doit être recalculée à chaque itération de l'algorithme IRLS.

Le critère d'arrêt (étape 4. de la figure 3.4) porte sur la convergence des coefficients $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}_j$, le maximum des variations absolues et relatives entre deux itérations (pour chacun des coefficients), devant être inférieur à 10^{-6}.

3.3.2.3 Amélioration de l'algorithme

Un algorithme efficace pour trouver les solutions du lasso a été proposé par [Osborne *et al.*, 2000b] et, plus récemment par [Efron *et al.*, 2004] (voir section 3.2.1.2). Ceux–ci pourraient être appliqués à l'estimation des composantes linéaires (étapes 3.(a)–3.(c) dans la figure (3.3), et 2.(a)–2.(c) dans la figure (3.5)). Il serait

Algorithme du point fixe :

1. Matrices de lissage :

   $\mathbf{H}_j = [\mathbf{1}\ \mathbf{x}_j]([\mathbf{1}\ \mathbf{x}_j]^t\mathbf{W}[\mathbf{1}\ \mathbf{x}_j]^t)^{-1}[\mathbf{1}\ \mathbf{x}_j]^t\mathbf{W}$, $\quad \mathbf{H} = [\mathbf{1}\ \mathbf{X}]\left([\mathbf{1}\ \mathbf{X}]^t\mathbf{W}[\mathbf{1}\ \mathbf{X}]\right)^{-1}[\mathbf{1}\ \mathbf{X}]^t\mathbf{W}$,
   $\mathbf{S}_j = \mathbf{N}_j(\mathbf{N}_j^t\mathbf{W}\mathbf{N}_j^t + \lambda_j\mathbf{\Omega}_j)^{-1}\mathbf{N}_j^t\mathbf{W}$, $\quad \widetilde{\mathbf{S}}_j = \mathbf{S}_j - \mathbf{H}_j$.

2. Composantes linéaires :
   (a) Estimation des coefficients :
   $\boldsymbol{\alpha} = \left([\mathbf{1}\ \mathbf{X}]^t\mathbf{W}[\mathbf{1}\ \mathbf{X}] + \mathbf{M}\right)^{-1}[\mathbf{1}\ \mathbf{X}]^t\mathbf{W}\mathbf{z}$, où $\mathbf{M} = \mathrm{diag}[0, \mu_1, \ldots, \mu_p]$.
   (b) Ré–estimation des termes de pénalisation : $\mu_j = \mu\dfrac{\sum_{k=1}^p |\alpha_k|}{p|\alpha_j|}$, $j = 1, \ldots, p$.
   (c) Itérer 2.(a) et 2.(b) jusqu'à convergence.
3. Composantes non linéaires :
   (a) Estimation des coefficients par backfitting :
      i. Résidus partiels : $\widetilde{\mathbf{r}}_j = \mathbf{y} - \mathbf{H}\mathbf{y} - \sum_{k\neq j}\mathbf{N}_j\widetilde{\boldsymbol{\beta}}_j$.
      ii. Coefficients : $\mathbf{N}_j\widetilde{\boldsymbol{\beta}}_j = \widetilde{\mathbf{S}}_j\widetilde{\mathbf{r}}_j$.
   (b) Ré–estimation des termes de pénalisation : $\lambda_j = \lambda\dfrac{\sum_{j=1}^p\sqrt{\widetilde{\boldsymbol{\beta}}_j\mathbf{\Omega}_j\widetilde{\boldsymbol{\beta}}_j}}{p\sqrt{\widetilde{\boldsymbol{\beta}}_j\mathbf{\Omega}_j\widetilde{\boldsymbol{\beta}}_j^t}}$.
   (c) Itérer 3.(a) et 3.(b) jusqu'à convergence.
4. Itérer 2. et 3. jusqu'à convergence.

FIG. 3.5 – Algorithme de résolution du problème de minimisation quadratique pondéré sous contraintes.

plus intéressant l'adaptation de ces algorithmes à l'estimation des composantes non linéaires (étapes 4.(a)–4.(c) dans la figure (3.3), et 3.(a)–3.(c) dans la figure (3.5)), car c'est cette étape de l'algorithme qui demande la plupart des calculs.

Les systèmes à résoudre dans la partie non linéaire peuvent également être simplifiés par l'utilisation des P–splines à la place des splines de lissage (voir section 1.2.3.2, page 21). Dans ce cas, si le nombre de nœuds n'est pas trop élevé, les systèmes peuvent être résolus directement (voir section 1.3.5.4, page 43).

Aussi, des décompositions QR et de Choleski combinées avec les décompositions en valeurs singulières permettrait des actualisations plus rapides [Wood, 2000, Kim et Gu, 2004].

3.3.3 D'autres méthodes de régularisation pour les modèles additifs

Des méthodes de régularisation basées sur la pénalisation l_1 ont été proposées préalablement pour les modèles additifs. L'objectif principal de ces approches est de choisir un nombre réduit de fonctions de la base de représentation, susceptibles de rendre compte de la partie la plus significative des données. Cependant, ces stratégies n'encouragent pas nécessairement la sélection de variables.

Plasm

Une généralisation du lasso a été proposée par [Bakin, 1999] : *plasm* (*probing least*

absolute squares modelling). La sélection est ici réalisée par groupes de coefficients, plutôt que de façon individuelle, et la matrice des données est remplacée la matrice d'une base de fonctions évaluées sur les points observés.

Pour le cas additif (centré) cela s'écrit :

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{A}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{A}\boldsymbol{\beta}) \qquad \text{sous contrainte} \qquad \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\beta}_j} \leq \tau, \tag{3.50}$$

où τ est prédéfini, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{d_j j})^t$ est un vecteur de dimension d_j, $j = 1, \ldots, p$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_p^t)^t$ est un vecteur de dimension $\sum_{j=1}^{p} d_j$, $\mathbf{A}$ est une matrice par blocs $\mathbf{A}_j$ de dimension $n \times d_j$, dont les éléments sont déterminés par l'évaluation d'une base B_j sur les x_{ij}. Ainsi, $\mathbf{f}_1(\mathbf{x}_1) + \ldots + \mathbf{f}_p(\mathbf{x}_p) = \mathbf{A}\boldsymbol{\beta} = \mathbf{A}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{A}_p\boldsymbol{\beta}_p$ et $\mathbf{f}_j(\mathbf{x}_j) = \mathbf{A}_j\boldsymbol{\beta}_j = \sum_{k=1}^{d_j} B_{jk}(x_{ij})\beta_{jk}$.

L'idée consiste donc à choisir le modèle en termes de groupes de coefficients $\{\beta_{jk}, k = 1, \ldots, d_j\}_{j=1}^{p}$. Quand chaque groupe de coefficients est constitué par un seul élément, $d_j = 1$, $j = 1, \ldots, p$, on retrouve le lasso. Si un seul groupe de coefficients est considéré, $p = 1$, on retrouve la pénalisation quadratique.

Un algorithme du type programmation quadratique successive (voir section A.1, page 137) est utilisé pour résoudre le problème d'optimisation, en termes du Lagrangien.

Une modification de plasm permet l'estimation du problème plus général, qui englobe les splines cubiques de lissage :

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{A}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{A}\boldsymbol{\beta}) \qquad \text{sous contrainte} \qquad \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\beta}_j} \leq \tau, \tag{3.51}$$

où $\boldsymbol{\Omega}_j$ sont des matrices de pénalisation symétriques, (semi–)définies positives.

On retrouve alors, sous la formulation de problème d'optimisation sous contraintes, le même problème de pénalisation au sens de la norme l_1 appliquée, par [Grandvalet et Canu, 1998], à la régression additive ajustée des splines cubiques de lissage. Cependant, comme précisé précédemment, les variables sélectionnées ne sont pas éliminées, elles sont linéarisées.

Likelihood basis pursuit

En exploitant le fait que les splines cubiques de lissage constituent un espace de Hilbert à noyau auto–reproduisant (RKHS), dotés de la norme $\|f\|_2^2 = \int [f^{(2)}(t)]^2 dt$, [Zhang *et al.*, 2003] utilisent représentation des splines en termes de bases de noyaux. Avec cette formulation, il est possible de pénaliser les composantes linéaires et non linéaires séparément, au sens de la norme l_1, avec le coût basé sur la log–vraisemblance. Le nom de cette méthode s'inspire des méthodes de pénalisation l_1 pour les ondelettes [Chen *et al.*, 1995b].

Après ré–paramétrisation des contraintes, afin de les linéariser, le problème est résolu par programmation non linéaire.

Les fonctions f_j sont donc estimées par un sous–ensemble réduit d'éléments de la base de noyaux, cependant la sélection de variables n'est pas forcément encouragée.

Cosso

Une approche similaire, pour des splines de lissage plus générales, qui constituent également un RKHS, mais dotés cette fois–ci de la norme $\|f\|^2 = [\int f(t)dt]^2 + [\int f^{(1)}(t)dt]^2 + \int [f^{(2)}(t)]^2 dt$, est connue sous le nom de cosso (*component selection and smoothing operator*) [Lin et Zhang, 2003]. Dans ce cas, un seul terme pénalise les composantes linéaires et non linéaires de façon simultanée.

La stratégie adoptée pour résoudre le problème, sous la forme d'optimisation sous contraintes, consiste à itérer deux étapes. La première considère les paramètres de la complexité fixes, et estime les coefficients par résolution des splines. La deuxième considère les coefficients fixes, et une formulation de type non–negative garrote permet d'estimer les paramètres de la complexité. Les auteurs signalent que cette procédure est néanmoins lente à converger.

Comme dans le cas précédant, les fonctions f_j sont ajustées par peu de termes mais ceci n'encourage pas la sélection de variables.

3.4 Sélection des paramètres de la complexité

La sélection d'un modèle de complexité adaptée est une étape clé pour les modèles d'apprentissage statistique. Nous avons pu constater au cour du chapitre 2 que cette étape est difficile à mettre en œuvre pour les modèles additifs. En effet, comme ces modèles sont composés d'autant de fonctions que de variables, l'espace de recherche de la complexité est de dimension p. Notre modèle ne présente quant à lui que deux paramètres de réglage de la complexité, ce qui facilite considérablement sa mise en œuvre dès que le nombre de variables est supérieur à deux. Cette simplification est la conséquence des contraintes sur μ_j et λ_j (3.42), qui permettent à chacune d'explorer un espace de dimension $p-1$ dans la procédure d'estimation des paramètres. Il ne reste plus qu'à fixer les valeurs de μ et λ par estimation de l'erreur de généralisation.

3.4.1 Estimation du nombre effectif de paramètres

Dans les section (3.3.1)–(3.3.2), nous avons constaté que les composantes linéaires et non linéaires sont estimées de façon indépendante. L'estimation du modèle additif pénalisé admet alors l'expression suivante, linéaire vis à vis des observations et additive par rapport aux composantes constante, linéaires et non linéaires :

$$\widehat{\mathbf{y}} = \left(\boldsymbol{H}^{\mathrm{C}} + \boldsymbol{H}_{\mu}^{\mathrm{LI}} + \boldsymbol{S}_{\lambda}^{\mathrm{NL}}\right)\mathbf{y}. \tag{3.52}$$

La matrice "chapeau" correspondante à la composante constante, $\boldsymbol{H}^{\mathrm{C}}$, est la matrice $n \times n$ telle que tous les éléments sont égaux à $1/n$:

$$\boldsymbol{H}^{\mathrm{C}} = \tfrac{1}{n}\mathbf{1}_{n\times n}. \tag{3.53}$$

La complexité résultante des paramètres non pénalisés est mesurée comme le nombre de paramètres. La participation de la constante α_0 aux degrés de liberté est, donc, simplement 1.

La déduction des matrices "chapeau" correspondantes aux composantes linéaires,

$$\boldsymbol{H}_{\mu}^{\mathrm{LI}} = \bar{\mathbf{X}} \left(\bar{\mathbf{X}}^t \bar{\mathbf{X}} + \bar{\mathbf{M}} \right)^{-1} \bar{\mathbf{X}}^t, \tag{3.54}$$

et des matrices de "rétrécissement" correspondantes aux composantes non linéaires,

$$\boldsymbol{S}_{\lambda}^{\mathrm{NL}} \approx \sum_{j=1}^{p} \widetilde{\mathbf{S}}_j, \tag{3.55}$$

dans les sections suivantes permettra de définir le nombre effectif de paramètres associé aux composantes linéaires, $\mathrm{ddl}^{\mathrm{LI}}(\mu)$, et aux composantes non linéaires, $\mathrm{ddl}^{\mathrm{NL}}(\lambda)$, respectivement.

Le nombre effectif de paramètres total est alors obtenu comme l'addition du nombre effectif de paramètres associé à la constante, aux composantes linéaires et aux composantes non linéaires :

$$\mathrm{ddl}(\mu, \lambda) = \mathrm{ddl}^{\mathrm{LI}}(\mu) + \mathrm{ddl}^{\mathrm{NL}}(\lambda) + 1. \tag{3.56}$$

Cette addition est justifiée par l'orthogonalité des composantes non linéaires, linéaires et constante, qui est assurée par les contraintes d'orthogonalité : $\mathbf{x}_j^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0$ et $\mathbf{1}^t \mathbf{N}_j \widetilde{\boldsymbol{\beta}}_j = 0$ (3.42).

3.4.1.1 Nombre effectif de paramètres associé aux composantes linéaires

Considérons le problème linéaire sans la constante (qui est traitée séparément). Des estimations du nombre de degrés de liberté pour l'estimateur lasso ont été proposées (voir section 3.2.1.2, page 85). Une première estimation basée sur une reformalisation des solutions du lasso, est proposée par [Tibshirani, 1996]. Une modification de l'estimation précédant permet de prendre en compte la pénalisation sur les variables jugées non pertinentes [Fu, 1998].

Cependant, pour cette dernière estimation, quand $\mathbf{X}$ n'est pas orthogonale, le nombre de degrés de liberté lié aux coefficients nuls ne coïncide pas avec le nombre de coefficients annulés. Nous considérons une modification de cette définition, ne prenant en compte que les colonnes de $\mathbf{X}$ et $\mathbf{A}$ pour lesquelles les coefficients $\widehat{\alpha}_j$ sont non nuls ($\bar{\mathbf{X}}$ et $\bar{\mathbf{A}}$, respectivement) :

$$\mathrm{ddl}^{\mathrm{LI}}(\mu) = \mathrm{tr} \left[\bar{\mathbf{X}} \left(\bar{\mathbf{X}}^t \bar{\mathbf{X}} + \mu \bar{\mathbf{A}}^{-} \right)^{-1} \bar{\mathbf{X}}^t \right]. \tag{3.57}$$

Cette définition induit une prédiction plus conservatrice. En effet, on peut montrer que l'expression (3.57) est plus grande que celle de Fu (3.34) (voir section A.3, page 143). Elle est également plus avantageuse numériquement, car la dimension des matrices en (3.34) est plus élevée.

En abordant le problème sous la forme de la pénalisation adaptative, nous proposons l'estimation suivante :

$$\mathrm{ddl}^{\mathrm{LI}}(\mu) = \mathrm{tr}\left[\boldsymbol{H}_{\mu}^{\mathrm{LI}}\right] = \mathrm{tr}\left[\bar{\mathbf{X}}\left(\bar{\mathbf{X}}^t\bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1}\bar{\mathbf{X}}^t\right], \tag{3.58}$$

où $\bar{\mathbf{M}}$ comprend les colonnes de $\mathbf{M} = \mathrm{diag}(\mu_j)$ pour lesquelles les coefficients $\widehat{\alpha}_j$ sont non nuls, et $\boldsymbol{H}_{\mu}^{\mathrm{LI}}$ est la matrice chapeau introduite à (3.52).

Ces estimations souffrent quelques imprécisions. D'une part, les définitions (3.34) et (3.57) supposent la linéarité du modèle, ce qui implique que la matrice $\mathbf{A}$ de (3.34) ou la matrice $\bar{\mathbf{A}}$ de (3.57) sont supposées ne pas être affectées par les observations y_i. Cette hypothèse simplificatrice n'étant pas respectée, ces définitions ne proposent qu'une borne inférieure du nombre effectif de paramètres. Notre définition (3.58) étant la plus conservatrice, elle est la plus proche de la réalité.

D'autre part, le coût de l'estimation des coefficients α_j, dans les définitions (3.34) et (3.57) et de l'estimation des termes de pénalisation μ_j, dans la définition (3.58), n'est pas pris en compte : le nombre de degrés de liberté est calculé comme si les estimations étaient connues a priori, ce qui introduit du biais dans l'estimation du nombre effectif de paramètres associé aux composantes linéaires. Ce problème est souvent rencontré dans le contexte de la sélection de modèles : le modèle sélectionné est supposé être connu a priori et donc l'effet de la sélection est négligé [Tibshirani et Knight, 1997, Ye, 1998].

Pour les problèmes de type non gaussien, basés sur la maximisation de la log–vraisemblance, l'estimation du nombre de degrés de liberté incorpore la matrice de pondérations, $\mathbf{W}$, obtenue à la dernière itération de l'algorithme IRLS :

$$\mathrm{ddl}^{\mathrm{LI}}(\mu) = \mathrm{tr}\left[\boldsymbol{H}_{\mu}^{\mathrm{LI}}\right] = \mathrm{tr}\left[\bar{\mathbf{X}}\left(\bar{\mathbf{X}}^t\mathbf{W}\bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1}\bar{\mathbf{X}}^t\mathbf{W}\right]. \tag{3.59}$$

Il faut signaler que l'hypothèse de linéarité vis à vis du vecteur d'observations est ici transgressée par la matrice des termes de pénalisation individuels, $\bar{\mathbf{M}}$, ainsi que par la matrice des pondérations, $\mathbf{W}$.

3.4.1.2 Nombre effectif de paramètres associé aux composantes non linéaires

Par analogie au modèle linéaire, le nombre de degrés de liberté est défini, pour les modèles additifs comme la trace de la matrice $\mathbf{R}$, indépendante des observations, qui génère la prédiction, $\widehat{\mathbf{f}} = \mathbf{R}\mathbf{y}$, où $\widehat{\mathbf{f}} = \widehat{\mathbf{f}}_1 + \ldots + \widehat{\mathbf{f}}_p$ (voir sections 2.2.2, page 56 et 2.2.3, page 57). Cette matrice correspond à la dernière itération de l'algorithme backfitting et le calcul de sa trace peut s'avérer difficile. L'approximation de la trace de la matrice $\mathbf{R}$ par la somme des traces des matrices de lissage individuelles $\mathbf{S}_j$ (2.12)–(2.14) est donc adoptée.

Afin de ne pas tenir compte des valeurs propres correspondantes aux fonctions constante et linéaires, l'estimation est effectuée sur les matrices de rétrécissement. Le nombre effectif de paramètres associé aux composantes non linéaires est ainsi estimé

par :

$$\mathrm{ddl}^{\mathrm{NL}}(\lambda) = \mathrm{tr}\left[\boldsymbol{S}_{\lambda}^{\mathrm{NL}}\right] \approx \mathrm{tr}\left[\sum_{j=1}^{p} \widetilde{\mathbf{S}}_j\right] = \sum_{j=1}^{p} \mathrm{tr}[\widetilde{\mathbf{S}}_j] = \sum_{j=1}^{p} \mathrm{tr}\left[\mathbf{S}_j - \mathbf{H}_j\right], \tag{3.60}$$

où $\widetilde{\mathbf{S}}_j$ sont les matrices de rétrécissement et $\mathbf{H}_j$ sont les matrices chapeau (figures (3.3) et (3.5)), dont la trace est simplement 2 : la somme des valeurs propres égales à 1 correspondantes aux fonctions linéaire et constante. La matrice $\boldsymbol{S}_{\lambda}^{\mathrm{NL}}$ est la matrice de rétrécissement correspondante aux composantes non linéaires introduite à (3.52).

Pour les problèmes de type non gaussien, l'estimation de $\mathrm{ddl}^{\mathrm{NL}}(\lambda)$ est basée sur l'approximation par les matrices de rétrécissement obtenues à la dernière itération de l'algorithme IRLS.

La décomposition en valeurs singulières de la matrice de lissage (3.46) nous permet de simplifier le calcul de $\mathrm{ddl}^{\mathrm{NL}}(\lambda)$:

$$\begin{aligned} \mathrm{ddl}^{\mathrm{NL}}(\lambda) &= \sum_{j=1}^{p} \left[\mathrm{tr}(\mathbf{S}_j) - 2\right] = \sum_{j=1}^{p} \left[\mathrm{tr}\left(\mathbf{Z}_j^t \mathbf{Z}_j (\mathbf{Z}_j^t \mathbf{Z}_j + \lambda_j \mathbf{I})^{-1}\right) - 2\right] \\ &= \sum_{j=1}^{p} \left[\sum_{i=1}^{n} \frac{\zeta_{ij}^2}{\lambda_j + \zeta_{ij}^2} - 2\right], \end{aligned} \tag{3.61}$$

où ζ_{ij} sont les éléments de la matrice diagonale $\mathbf{Z}_j$.

Comme dans le cas linéaire, le coût de l'estimation des termes de pénalisation individuels λ_j n'est pas intégré dans l'estimation du nombre effectif de paramètres associé aux composantes non linéaires, introduisant ainsi un biais [Tibshirani et Knight, 1997, Ye, 1998].

3.4.1.3 Estimation des écart–types et des intervalles de confiance

De façon analogue aux modèles additifs, des intervalles de confiance ponctuelles basées sur des approximations sont déduits pour les modèles additifs parcimonieux. En utilisant que les composantes linéaires et non linéaires sont orthogonales et supposant μ et λ fixés, une estimation de la covariance des estimations est la suivante :

$$\mathrm{Cov}(\widehat{\mathbf{f}}_j) \approx \left(\mathbf{x}_j \left[\left(\bar{\mathbf{X}}^t \bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1} \bar{\mathbf{X}}^t \bar{\mathbf{X}} \left(\bar{\mathbf{X}}^t \bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1}\right]_{jj} \mathbf{x}_j^t + \widetilde{\mathbf{S}}_j \widetilde{\mathbf{S}}_j^t\right) \sigma^2, \tag{3.62}$$

$j = 1, \ldots, p$, pour $1/\mu_j > 0$ et $1/\lambda_j > 0$. Un estimateur non biaisé de la variance de l'erreur est également donné par :

$$\widehat{\sigma}^2 = \frac{\left\|\mathbf{y} - \left(\boldsymbol{H}^{\mathrm{C}} + \boldsymbol{H}_{\mu}^{\mathrm{LI}} + \boldsymbol{S}_{\lambda}^{\mathrm{NL}}\right)\mathbf{y}\right\|_2^2}{n - \mathrm{ddl}(\mu, \lambda)}, \tag{3.63}$$

où $\mathrm{ddl}(\mu, \lambda)$ est défini à (3.56) et $\boldsymbol{H}^{\mathrm{C}}$, $\boldsymbol{H}_{\mu}^{\mathrm{LI}}$ et $\boldsymbol{S}_{\lambda}^{\mathrm{NL}}$ sont définies à (3.53), (3.54) et (3.55), respectivement.

Les écart–types ponctuelles s'écrivent alors :

$$\mathrm{se}_{ij} = \widehat{\sigma}\sqrt{\left(\mathbf{x}_j \left[\left(\bar{\mathbf{X}}^t\bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1}\bar{\mathbf{X}}^t\bar{\mathbf{X}}\left(\bar{\mathbf{X}}^t\bar{\mathbf{X}} + \bar{\mathbf{M}}\right)^{-1}\right]_{jj}\mathbf{x}_j^t + \widetilde{\mathbf{S}}_j\widetilde{\mathbf{S}}_j^t\right)_{ii}} \tag{3.64}$$

$i = 1, \ldots, n$. Supposant que les erreurs sont Gaussiennes et le biais négligeable, les écart–types peuvent être utilisés pour obtenir des intervalles de confiance ponctuelles : $\widehat{f}_j(x_{ij}) \pm z_{\alpha/2}\mathrm{se}_{ij}$, où $z_{\alpha/2}$ est le $\alpha/2$–ème centile de la distribution normale.

Néanmoins, dans le chapitre 2 (sections 2.2.2.1 et 2.2.3.1) nous avons noté que ce type d'approximations peut entraîner des sous–estimations de la variance des $\widehat{\mathbf{f}}_j$. La déduction d'intervalles de confiance de type bootstrap est donc souhaitable.

3.4.2 Adaptation des méthodes de sélection

Parmi les méthodes de sélection de la complexité introduites à la section 2.4, les méthodes d'évaluation sur une grille de type réechantillonage sont applicables directement à notre problème. Cependant, ces méthodes requièrent, en général, un nombre considérable de calculs. En revanche, l'expression de la validation croisée *leave–one–out* en fonction de la matrice de lissage (2.31) demande une quantité de calculs comparable à celle des méthodes analytiques. Nous pouvons approcher la validation croisée au moyen des matrices "chapeau" (3.52), introduites dans la section précédente :

$$\mathrm{CV}(\mu, \lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{\alpha}_0 + \sum_{j=1}^{p} x_{ij}\widehat{\alpha}_j + \sum_{j=1}^{p}\widetilde{f}_j(x_{ij}) - y_i}{1 - \left(\boldsymbol{H}^{\mathrm{C}} + \boldsymbol{H}_{\mu}^{\mathrm{LI}} + \boldsymbol{S}_{\lambda}^{\mathrm{NL}}\right)_{ii}}\right)^2. \tag{3.65}$$

La définition du nombre de degrés de liberté (3.56), nous permet également l'adaptation des critères analytiques (section 2.4.2) à la sélection des deux paramètres de réglage de la complexité des modèles additifs parcimonieux. La validation croisée généralisée, le critère d'information d'Akaike, ainsi que sa version corrigée, et le critère d'information bayésien, s'écrivent :

$$\mathrm{GCV}(\mu, \lambda) = \frac{\left|\left|\widehat{\mathbf{f}}_{\mu,\lambda} - \mathbf{y}\right|\right|_2^2}{n(1 - \mathrm{ddl}(\mu, \lambda)/n)^2}, \tag{3.66}$$

$$\mathrm{AIC}(\mu, \lambda) = \frac{1}{n}\left|\left|\widehat{\mathbf{f}}_{\mu,\lambda} - \mathbf{y}\right|\right|_2^2 + \frac{2\mathrm{ddl}(\mu, \lambda)\sigma^2}{n}, \tag{3.67}$$

$$\mathrm{AICc}(\mu, \lambda) = \frac{1}{n}\left|\left|\widehat{\mathbf{f}}_{\mu,\lambda} - \mathbf{y}\right|\right|_2^2 + \frac{2n\mathrm{ddl}(\mu, \lambda)}{n - \mathrm{ddl}(\mu, \lambda) - 1}\sigma^2, \tag{3.68}$$

$$\mathrm{BIC}(\mu, \lambda) = \frac{1}{n}\left|\left|\widehat{\mathbf{f}}_{\mu,\lambda} - \mathbf{y}\right|\right|_2^2 + \frac{\log(n)\mathrm{ddl}(\mu, \lambda)\sigma^2}{n}, \tag{3.69}$$

où $\widehat{\mathbf{f}}_{\mu,\lambda} = \widehat{\alpha}_0 + \sum_{j=1}^{p}\mathbf{x}_j\widehat{\alpha}_j + \sum_{j=1}^{p}\widetilde{\mathbf{f}}_j(\mathbf{x}_j) = \left(\boldsymbol{H}^{\mathrm{C}} + \boldsymbol{H}_{\mu}^{\mathrm{LI}} + \boldsymbol{S}_{\lambda}^{\mathrm{NL}}\right)\mathbf{y}$.

Ces méthodes sont évaluées sur une grille de valeurs des deux paramètres de la complexité. Pour des réponses gaussiennes, cette grille n'est pas quadratique mais linéaire, car les composantes linéaires et non linéaires sont orthogonales.

Pour les critères AIC, AICc et BIC, la variance de l'erreur est estimée par (3.63), pour des valeurs de μ et λ comportant une complexité élevée (afin d'obtenir une estimation peu biaisée). Une possibilité est de fixer μ et λ aux plus petites valeurs considérées dans la grille.

Pour les modèles additifs généralisés parcimonieux, les méthodes précédant peuvent s'écrire en termes des réponses de travail de l'algorithme IRLS :

$$\mathrm{GCV}(\mu,\lambda) = \frac{\left|\left|\mathbf{W}^{1/2}\left(\widehat{\mathbf{f}}_{\mu,\lambda}-\mathbf{z}\right)\right|\right|_2^2}{n(1-\mathrm{ddl}(\mu,\lambda)/n)^2}, \tag{3.70}$$

$$\begin{aligned}\mathrm{AIC}(\mu,\lambda) &= -2l(\mu,\lambda)+2\mathrm{ddl}(\mu,\lambda)\\ &\approx \frac{1}{n}\left|\left|\mathbf{W}^{1/2}\left(\widehat{\mathbf{f}}_{\mu,\lambda}-\mathbf{z}\right)\right|\right|_2^2+\frac{2\mathrm{ddl}(\mu,\lambda)\phi}{n},\end{aligned} \tag{3.71}$$

$$\begin{aligned}\mathrm{AICc}(\mu,\lambda) &= -2l(\mu,\lambda)+\frac{2n\mathrm{ddl}(\mu,\lambda)}{n-\mathrm{ddl}(\mu,\lambda)-1}\\ &\approx \frac{1}{n}\left|\left|\mathbf{W}^{1/2}\left(\widehat{\mathbf{f}}_{\mu,\lambda}-\mathbf{z}\right)\right|\right|_2^2+\frac{2n\mathrm{ddl}(\mu,\lambda)}{n-\mathrm{ddl}(\mu,\lambda)-1}\phi,\end{aligned} \tag{3.72}$$

$$\begin{aligned}\mathrm{BIC}(\mu,\lambda) &= -2l(\mu,\lambda)+\log(n)\mathrm{ddl}(\mu,\lambda)\\ &\approx \frac{1}{n}\left|\left|\mathbf{W}^{1/2}\left(\widehat{\mathbf{f}}_{\mu,\lambda}-\mathbf{z}\right)\right|\right|_2^2+\frac{\log(n)\mathrm{ddl}(\mu,\lambda)\phi}{n},\end{aligned} \tag{3.73}$$

où $\widehat{\mathbf{f}}_{\mu,\lambda}=\widehat{\alpha}_0+\sum_{j=1}^p \mathbf{x}_j\widehat{\alpha}_j+\sum_{j=1}^p \widetilde{\mathbf{f}}_j(\mathbf{x}_j)$ est l'estimation obtenue à la dernière itération de l'algorithme IRLS, et ϕ est le paramètre de dispersion.

Nous considérons les problèmes d'estimation fonctionnelle et de sélection de la complexité séparément, afin d'assurer la convergence. Les fonctions sont donc ajustées pour chacune des valeurs des paramètres de la complexité considérées dans la grille. Ensuite, les valeurs (μ,λ) minimisant les critères précédant, sont sélectionnées. L'application de méthodes d'optimisation newtoniennes pour la GCV [Gu et Wahba, 1991, Wood, 2000, Wood, 2004] semblent pouvoir s'adapter à notre problème. Elles pourraient présenter une alternative à l'exploration de tout l'espace : $[0,\infty[+[0,\infty[$, dans le cas Gaussien, $[0,\infty[\times[0,\infty[$, dans le cas non Gaussien.

3.5 En bref

Deux faits ont motivé notre généralisation du lasso aux modèles additifs. D'une part, dans le cadre linéaire, plusieurs travaux ont étudié les propriétés de stabilité et la capacité de sélectionner de variables des méthodes de pénalisation, telles que le lasso. Leurs résultats nous permettent de mieux comprendre les raisons pour lesquelles,

dans certaines situations, elles sont plus performantes que des méthodes classiques telles que la sélection pas à pas.

D'autre part, dans le cadre des modèles additifs, l'application des techniques de sélection pas à pas comporte de nouveaux problèmes : non seulement il faut choisir les composantes à inclure dans le modèle, mais aussi leur proportion de lissage. Par conséquent, ces méthodes sont réduites aux cas avec peu de variables en entrée.

Pour généraliser le lasso aux modèles additifs, nous observons que 1) le lasso peut être exprimé en termes de pénalisation multiple adaptative, les deux méthodes sont équivalentes. 2) La pénalisation multiple adaptative est une modification de la pénalisation quadratique (*ridge regression*) qui attribue à chaque coefficient une pénalisation en accord avec son importance et qui aboutisse à une sélection de variables. 3) Les splines cubiques de lissage (ou des P–splines) s'écrivent comme un coût pénalisée par un terme quadratique, elles sont une généralisation de la pénalisation quadratique.

La généralisation du lasso passe donc par l'adaptation de la pénalisation multiple adaptative aux splines. Cependant, quand une composante est très pénalisée cela implique l'une des possibilités suivantes : soit elle est non pertinente soit elle est linéaire, mais on ne peut pas discerner entre ces deux possibilités.

Afin d'identifier les variables à éliminer, les variables à effets linéaires et les variables à effets non linéaires, nous considérons les sous–espaces vectoriels linéaires et non linéaires séparément et nous les pénalisons indépendamment. Ainsi, quand la composante non linéaire d'une fonction est très pénalisée, cela implique qu'elle n'est pas non linéaire (donc, linéaire ou nulle) Si sa composante linéaire est également très pénalisée, la variable est éliminée, dans le cas contraire, elle reste linéairement dans le modèle.

Les sous–espaces linéaires et non linéaires sont orthogonaux, ce qui permet le calcul effectif des solutions. L'espace généré par les composantes linéaires est facile à traiter, on retrouve simplement le cas linéaire, l'espace généré par les composantes non linéaires est simplement l'espace total moins l'espace linéaire.

Les algorithmes proposés semblent bien se comporter en pratique. Nous avons rencontré des problèmes de stabilité numérique seulement dans les cas où les paramètres μ ou λ (particulièrment ce dernier) sont très petits. Néanmoins, ils peut être considérablement simplifiés et améliorés afin d'accélerer les calculs.

La complexité du modèle est ainsi contrôlée par seulement deux paramètres, l'un contrôle la complexité des parties linéaires, l'autre celle des parties non linéaires. Les paramètres qui contrôlent les complexités individuelles de chaque composante sont réglés automatiquement. Afin de choisir ces deux hyper–paramètres, des critères de sélection de modèle ont été adaptés et des approximations du nombre effectif de paramètres ont été définies.

Chapitre 4

Expériences

4.1 Introduction

Ce chapitre est consacré à la mise en œuvre des modèles additifs parcimonieux. La première partie est dédiée à définir les bases d'un benchmark pour les modèles additifs, ce qui nous permet, dans la deuxième partie, d'évaluer expérimentalement la performance des méthodes développées.

La comparaison avec d'autres méthodes, notamment la sélection de variables pas à pas, nous permettent de déduire les conditions d'application de chaque algorithme. Lors des simulations, nous testons également les différentes techniques de sélection de modèle présentées dans le chapitre précédent.

Finalement, nous montrons un exemple d'application des modèles additifs parcimonieux sur deux jeux de données réelles.

4.2 Benchmark

Notre objectif est de définir les bases d'un *benchmark* (banc d'essais ou protocole de référence) pour les modèles additifs. Un benchmark est défini par un générateur de données paramétré, un plan d'expérience sur les paramètres, des simulations, et une analyse, à la suite, des résultats.

Pour définir les bases d'un benchmark, il faut étudier, d'abord, quelles sont les situations qui rendent difficile l'estimation, et ensuite, quels paramètres nous permettent de contrôler ces situations. Pour cela, nous prenons comme repère le benchmark proposé par [Breiman, 1996] pour la régression linéaire, et nous étudions les différents plans d'expérience proposés pour la régression additive.

4.2.1 Modèles linéaires

[Breiman, 1996] établit un benchmark afin de comparer différentes méthodes de régularisation dans le cas linéaire (parmi lesquelles se trouvent la sélection de sous–ensembles, la pénalisation quadratique et "garrot non négatif").

Le problème du contrôle de la complexité est particulièrement délicat quand la taille de l'échantillon est du même ordre que le nombre de degrés de liberté du modèle qui a généré les données. Dans ce cas, il n'y a pas suffisamment d'information pour estimer de manière précise les paramètres du modèle. Le rapport n/p est alors celui d'une situation extrême, où les méthodes peuvent se comporter mal.

Le nombre de variables significatives est également un facteur déterminant pour le choix d'une méthode de pénalisation (section 3.2.1.1). En effet, il est bien connu que les situations où seulement quelques coefficients sont non nuls favorisent des méthodes telles que la sélection de sous–ensembles, tandis que les situations où la majorité des entrées sont significatives favorisent des méthodes telles que la pénalisation quadratique. La corrélation des variables explicatives influe également sur l'estimation. Elle est une source d'instabilité à laquelle la sélection de sous–ensembles est spécialement sensible, alors que la pénalisation quadratique est conçue pour résister à l'instabilité.

Dans l'étude de [Breiman, 1996], le comportement de chaque méthode est évalué par la perte en prédiction PL (*predictive loss*), qui correspond à la différence entre l'erreur en prédiction ((2.27) ou (2.26), page 60) commise par la méthode, pour le paramètre de complexité sélectionné, moins l'erreur en prédiction de la même méthode, pour le paramètre de complexité optimal (celui qui minimise l'erreur en prédiction). La perte en prédiction mesure donc la perte occasionnée par la sélection de modèle.

4.2.2 Modèles additifs

Si les modèles additifs sont à l'origine d'une littérature abondante, où les conjectures et les résultats théoriques sont généralement évalués par des simulations, peu nombreuses sont les études qui justifient leur choix du plan d'expériences. Aussi, l'absence d'uniformité entre les différents scénarios rend les généralisations difficiles. Les critères appliqués pour mesurer le comportement des méthodes étudiées sont également très variés.

Certains paramètres de contrôle sont, néanmoins, souvent utilisés. Prenant comme base les idées du benchmark de [Breiman, 1996] pour le cas linéaire, et les points d'intérêt dans les simulations pour le cas additif, nous étudions premièrement quelles sont les situations qui rendent difficile l'estimation des modèles additifs, et ensuite, quels paramètres nous permettent de contrôler ces situations [Avalos *et al.*, 2003].

4.2.2.1 Situations rendant l'estimation difficile

Concernant les variables d'entrée

L'estimation du modèle est plus difficile quand la distribution empirique des variables d'entrée est clairsemée. En effet, quand le nombre d'observations est faible, ou quand la densité est loin d'être uniforme (combinant des régions denses avec les régions peu denses), l'information est localement pauvre. L'influence du nombre d'observations est étudiée par [Gu et Wahba, 1991, Breiman, 1993, Linton et Härdle, 1996, Opsomer et Ruppert, 1998, Kauermann et Opsomer, 2004]. L'effet de la variation de la densité des variables d'entrée est également analysée par [Breiman, 1993, Sperlich *et al.*, 1999].

D'autre part, la corrélation ou, plus généralement, la concurvité entre les variables d'entrée est source d'instabilité numérique (section 1.3.4.4). L'influence de la corrélation entre les variables d'entrée est analysée par [Breiman, 1993, Linton et Härdle, 1996, Opsomer et Ruppert, 1998, Sperlich *et al.*, 1999, Schimek, 2000, Kauermann et Opsomer, 2004], et celle de la concurvité est étudiée par [Bakin, 1999].

Concernant les fonctions sous–jacentes

La complexité de chaque composante peut être, en partie, contrôlée au moyen des fonctions sous–jacentes : plus la structure des vraies fonctions est "complexe", plus elle est difficile à estimer, et plus le nombre d'observations nécessaire à l'estimation est élevé.

La notion de complexité d'une fonction n'a pas vraiment de sens. En revanche, certaines fonctions sont plus faciles à estimer que d'autres (ce qui est conditionné par la méthode d'estimation).

D'un point de vue bayésien, l'estimation consiste à mettre à jour notre connaissance a priori sur les fonctions. Si notre "connaissance" a priori est bonne, l'estimation consiste à faire une petite mise à jour. C'est donc un problème simple. Si notre "connaissance" a priori ne reflète pas bien la réalité, le rôle de l'estimation est plus important et sa mise en œuvre est donc plus délicate.

Nous estimons ici des fonctions par des splines de lissage. La connaissance a priori encodée par le terme de régularisation, qui stipule que les fonctions plus lisses (telles que l'intégrale de la dérivée seconde au carré est petite) sont a priori plus plausibles que les fonctions irrégulières.

La difficulté du processus d'estimation sera donc affectée par la régularité des fonctions.

L'influence de la régularité des fonctions sous–jacentes est étudiée par [Breiman, 1993, Linton et Härdle, 1996, Schimek, 2000], et la diversité des fonctions sous–jacentes est analysée par [Gu et Wahba, 1991, Breiman, 1993, Opsomer et Ruppert, 1998, Sperlich *et al.*, 1999, Bakin, 1999, Brumback *et al.*, 1999, Kauermann et Opsomer, 2004].

Concernant la variable de sortie

La variable de sortie est générée par une fonction perturbée par un bruit additif. Dans la procédure d'estimation, le but est de discerner le phénomène sous–jacent du bruit. Par conséquent, plus le bruit est élevé, plus l'estimation est difficile. Afin d'éviter la sensibilité à l'échelle des données, l'effet du bruit peut être contrôlé par le biais du coefficient de détermination, R^2.

L'effet du bruit est étudié par [Bakin, 1999, Schimek, 2000].

Concernant la complexité globale

Quand la taille de l'échantillon est élevée par rapport au nombre de degrés de liberté du modèle qui a généré les données, le problème posé est trop facile, toute

méthode "correcte" trouvera une bonne solution. A l'inverse, quand le rapport est trop faible, le problème est insolvable. Il faut que le rapport permette de discerner les méthodes qui se comportent bien dans les situations difficiles.

Plusieurs facteurs participent au nombre de degrés de liberté (ou complexité globale) du modèle, par exemple, comme signalé précédemment, la complexité individuelle de chaque composante. Un autre facteur est, comme dans le cas linéaire, le nombre de variables d'entrée [Gu et Wahba, 1991, Breiman, 1993, Opsomer et Ruppert, 1998, Schimek, 2000].

Un autre aspect important est la reconnaissance du processus qui a réellement généré la réponse. Il est connu que la performance de certaines méthodes dépend du nombre d'entrées significatives sur le nombre total de variables explicatives. L'effet de la pertinence des entrées est analysé par [Gu et Wahba, 1991, Breiman, 1993, Bakin, 1999, Brumback *et al.*, 1999].

4.2.2.2 Paramètres de contrôle

Corrélation, concurvité et dispersion

La corrélation des variables explicatives peut être contrôlée dans le cas gaussien au moyen d'une matrice de corrélations. La dispersion plus ou moins importante des données peut être contrôlée par la variance : $\mathbf{X} \sim \mathcal{N}(0, \sigma^2_{\mathbf{X}}\mathbf{\Gamma})$, où $\Gamma_{ij} = \rho^{|i-j|}$. La matrice $\mathbf{\Gamma}$ est celle de corrélations, et $\sigma^2_{\mathbf{X}}\mathbf{\Gamma}$ correspond à la matrice de variances–covariances. Les paramètres ρ et $\sigma^2_{\mathbf{X}}$ contrôlent la corrélation et la dispersion, respectivement.

Un possible moyen d'introduire de la concurvité de façon contrôlée est la suivante : $X_j = h(X_k) + \varepsilon_k$, où $\varepsilon \sim \mathcal{N}(0, \sigma^2_k)$, $j \neq k$, j, $k \in \{1, \ldots, p\}$, et h une fonction régulière quelconque. Le paramètre ε_k contrôle le degré de concurvité.

Les fonctions sous–jacentes

La courbure des vraies fonctions peut être contrôlée par les fonctions sinus et cosinus. L'inclusion de fonctions linéaires, ainsi que de la combinaison de deux types de fonctions (linéaire et trigonométrique), permettent de prendre en compte un éventail assez large de fonctions, par rapport à leur complexité :

$$\begin{aligned} f_j(\mathbf{x}_j) &= \sin 2\pi k_j \mathbf{x}_j \\ f_j(\mathbf{x}_j) &= \cos 2\pi k_j \mathbf{x}_j \\ f_j(\mathbf{x}_j) &= a_j \mathbf{x}_j \\ f_j(\mathbf{x}_j) &= a_j \mathbf{x}_j + \sin 2\pi k_j \mathbf{x}_j \\ f_j(\mathbf{x}_j) &= a_j \mathbf{x}_j + \cos 2\pi k_j \mathbf{x}_j. \end{aligned} \tag{4.1}$$

Les paramètres k_j nous permettent de jouer sur la courbure de chaque fonction, les paramètres a_j nous permettent de jouer sur l'influence de la variable.

Nombre d'observations, nombre de variables explicatives, nombre de variables pertinentes et rapport nombre d'observations–nombre de degrés de liberté

Le problème du contrôle de la complexité est crucial quand la taille de l'échantillon est du même ordre que le nombre de degrés de liberté nécessaires pour modéliser les données. Dans le cas linéaire, cette mesure est facile à contrôler, car le nombre de degrés de liberté est simplement la dimension des entrées. Dans le cas non linéaire, il est plus simple d'approcher ce paramètre de contrôle par le nombre d'observations et le nombre de variables explicatives, ainsi que par le nombre de variables pertinentes.

Niveau de bruit

Nous considérons le cas gaussien : $Y = \sum_{j=1}^{p} f_j + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Le paramètre σ^2 contrôle le bruit. Cependant, il convient d'utiliser le coefficient de détermination, R^2, qui dépend de la variance de l'erreur, car c'est une mesure insensible à l'échelle des données.

4.2.2.3 Critères de comparaison

Dans le cas linéaire, [Breiman, 1996] utilise la perte en prédiction pour mesurer la perte occasionnée par la sélection de modèle. L'application de ce critère dans le cadre non paramétrique est compliquée. Par exemple, pour la sélection de sous–ensembles l'estimation de la perte en prédiction implique trouver les paramètres de la complexité optimaux pour le sous–ensemble optimal, ce qui est impraticable même pour p peu élevé.

Nous considérons donc l'erreur en prédiction (2.26) commise par chaque méthode. Aussi, la capacité d'éliminer les variables non pertinentes ou redondantes et de sélectionner les variables pertinentes est analysée. Finalement, le temps de calcul de chaque méthode est pris en compte.

4.3 Données contrôlées

4.3.1 Méthodes en comparaison

Dans les simulations suivantes nous comparons les modèles additifs parcimonieux, la sélection de variables pas à pas ascendante pour les modèles additifs et la pénalisation quadratique généralisée aux modèles additifs. La généralisation de la pénalisation quadratique est le modèle additif avec un seul paramètre de la complexité commun à toutes les composantes : $\lambda = \lambda_j$, $j = 1, \ldots, p$.

Les variables significatives et les paramètres de lissage de la sélection pas à pas sont sélectionnés par le critère GCV (section 3.2.2), évalué sur une grille de 8 valeurs à échelle logarithmique (cas linéaire, $\lambda_j = \infty$, inclus), pour chaque composante additive.

Le paramètre de lissage de la pénalisation quadratique est sélectionné par le critère GCV, évalué sur une grille de 8 valeurs. Le modèle linéaire, estimé par moindres carrés ordinaires ($\lambda_j = \infty$, $j = 1, \ldots, p$), est également inclus.

Dans le cas des modèles additifs parcimonieux, les fonctions AIC (3.67), AICc (3.68), BIC (3.69), GCV (3.66) et CV (3.65), sont évaluées sur une grille 8×8 de valeurs de (μ, λ) à échelle logarithmique. Les performances obtenues sont comparées

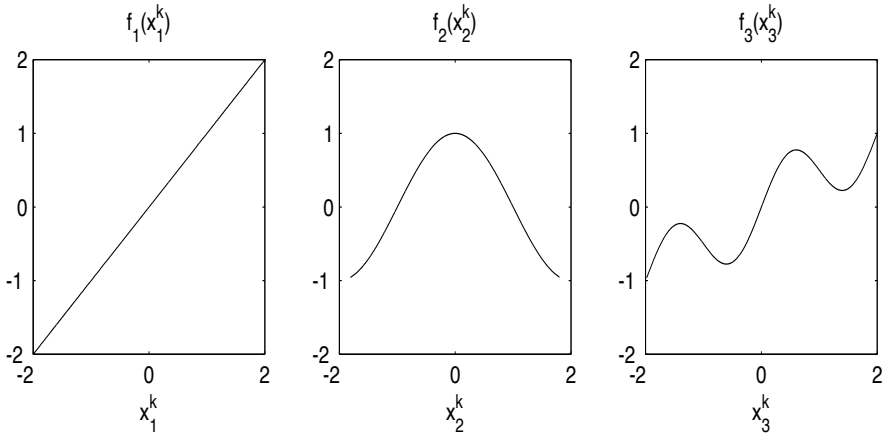

FIG. 4.1 – Fonctions sous–jacentes pour chaque groupe k, $k = 1, \ldots, 6$.

à la performance optimale (celle obtenue par une méthode de sélection choisissant le modèle d'erreur en prédiction minimale), également calculée sur la grille 8×8 de (μ, λ).

Dans tous les cas, les intervalles de recherche des paramètres de la complexité sont définis par les mêmes valeurs extrêmes, $[5 \times 10^{-3}, 5 \times 10^{2}]$. Pour la pénalisation quadratique et la sélection pas à pas, la valeur $\lambda_j = \infty$ est ensuite ajoutée, afin de tenir compte du cas linéaire.

Les performances obtenues par le modèle constant, estimé par la moyenne des réponses, sont rapportées pour référence.

4.3.2 Protocole expérimental

Des données ont été générées aléatoirement avec des solutions pré-spécifiées comme suit. Il y a au total $p = 18$ variables explicatives issues d'une distribution normale standard, et 1 variable réponse. Les variables explicatives sont partitionnées en 6 groupes de 3 variables : $\mathbf{X}^k = (X_1^k, X_2^k, X_3^k)$, $k = 1, \ldots, 6$. Les variables appartenant à des groupes différents sont indépendantes, les variables appartenant au même groupes sont corrélées : $\mathbf{X}^k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$, $\Lambda_{ij} = \rho^{|i-j|}$, où ρ est le paramètre qui contrôle la corrélation.

La partition en groupes réduits de variables corrélées nous permet, d'une part, de contrôler le "niveau" de corrélation d'une façon simple et d'autre part, d'identifier facilement quelles sont les variables apportant de l'information rédondante.

Les fonctions sous–jacentes dans chaque groupe sont (figure 4.1) : $f_1(x_1^k) = x_1^k$, $f_2(x_2^k) = \cos(\frac{\pi}{2} x_2^k)$, $f_3(x_3^k) = \frac{1}{2} x_3^k + \frac{1}{2} \sin(\pi x_3^k)$, $k = 1, \ldots, 6$. La réponse est générée par

$$y = \sum_{k=1}^{6} \delta^k [f_1(x_1^k) + f_2(x_2^k) + f_3(x_3^k)] + \varepsilon, \tag{4.2}$$

où $\delta^k \in \{0, 1\}$ contrôle la pertinence du k–ème groupe et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Le bruit est contrôlé au moyen de R^2, qui dépend de σ^2.

En fonction des paramètres de contrôle, nous considérons les situations suivantes :

Cas	Corrélation	N° de variables pertinentes	Bruit	N° d'observations
1	faible ($\rho = 0.1$)	faible ($d = 6$)	faible ($R^2 = 0.95$)	faible ($n = 50$)
2	faible ($\rho = 0.1$)	faible ($d = 6$)	faible ($R^2 = 0.95$)	modéré ($n = 200$)
3	faible ($\rho = 0.1$)	faible ($d = 6$)	modéré ($R^2 = 0.75$)	faible ($n = 50$)
4	faible ($\rho = 0.1$)	faible ($d = 6$)	modéré ($R^2 = 0.75$)	modéré ($n = 200$)
5	faible ($\rho = 0.1$)	élevé ($d = 15$)	faible ($R^2 = 0.95$)	faible ($n = 50$)
6	faible ($\rho = 0.1$)	élevé ($d = 15$)	faible ($R^2 = 0.95$)	modéré ($n = 200$)
7	faible ($\rho = 0.1$)	élevé ($d = 15$)	modéré ($R^2 = 0.75$)	faible ($n = 50$)
8	faible ($\rho = 0.1$)	élevé ($d = 15$)	modéré ($R^2 = 0.75$)	modéré ($n = 200$)
9	forte ($\rho = 0.9$)	faible ($d = 6$)	faible ($R^2 = 0.95$)	faible ($n = 50$)
10	forte ($\rho = 0.9$)	faible ($d = 6$)	faible ($R^2 = 0.95$)	modéré ($n = 200$)
11	forte ($\rho = 0.9$)	faible ($d = 6$)	modéré ($R^2 = 0.75$)	faible ($n = 50$)
12	forte ($\rho = 0.9$)	faible ($d = 6$)	modéré ($R^2 = 0.75$)	modéré ($n = 200$)
13	forte ($\rho = 0.9$)	élevé ($d = 15$)	faible ($R^2 = 0.95$)	faible ($n = 50$)
14	forte ($\rho = 0.9$)	élevé ($d = 15$)	faible ($R^2 = 0.95$)	modéré ($n = 200$)
15	forte ($\rho = 0.9$)	élevé ($d = 15$)	modéré ($R^2 = 0.75$)	faible ($n = 50$)
16	forte ($\rho = 0.9$)	élevé ($d = 15$)	modéré ($R^2 = 0.75$)	modéré ($n = 200$)

TAB. 4.1 – Résumé des situations analysées, en fonction des paramètres de contrôle.

- Corrélation (intra–groupe) faible ($\rho = 0.1$) ou forte ($\rho = 0.9$),
- Nombre faible ($\delta^1 = \delta^2 = 1$, $\delta^3 = \delta^4 = \delta^5 = \delta^6 = 0$) ou nombre élevé ($\delta^1 = \delta^2 = \delta^3 = \delta^4 = \delta^5 = 1$, $\delta^6 = 0$) de variables pertinentes ($d = 6$ sur 18 et $d = 15$ sur 18, respectivement).
- Bruit faible ($R^2 = 0.95$) ou modéré ($R^2 = 0.75$),
- Taille des échantillons petite ($n = 50$) ou modérée ($n = 200$).

Le tableau (4.1) montre les différents cas, en fonction des paramètres de contrôle. Pour chacune des 16 situations, 50 expériences ont été effectuées, les résultats sont donnés en termes de la moyenne (écart–type). Nous comparons l'erreur en prédiction commise par chaque méthode, en l'estimant sur un ensemble de test de taille 10000. Le nombre de variables éliminées, le nombre de variables non pertinentes éliminées et les degrés de liberté sont également rapportés.

4.3.3 Résultats

Comparaison des méthodes de pénalisation par rapport à l'erreur en prédiction

Les estimations de l'erreur en prédiction pour la fonction constante (pour référence), pour la pénalisation quadratique, la sélection pas à pas et le modèle additif

Cas	Constante	Pénalisation quadratique	Pas à pas	Modèle additif parcimonieux
1	3.593 (1.043)	1.028 (0.249)	**0.402** (0.806)	0.630 (0.219)
2	3.305 (0.436)	0.192 (0.019)	**0.141** (0.018)	0.180 (0.030)
3	3.632 (1.032)	1.911 (0.295)	1.700 (1.008)	**1.523** (0.268)†
4	3.790 (0.577)	0.886 (0.072)	**0.772** (0.077)	0.815 (0.064)
5	12.723 (4.221)	2.628 (0.449)	4.065 (6.541)	**2.392** (0.490)†
6	12.688 (2.172)	**0.463** (0.048)	0.475 (0.097)	0.642 (0.135)
7	13.248 (3.499)	4.983 (0.853)	7.139 (5.891)	**4.525** (0.594)†
8	14.321 (1.927)	**2.212** (0.167)	2.401 (0.407)	2.346 (0.207)†
9	4.436 (1.148)	0.971 (0.178)	**0.438** (0.209)	0.484 (0.112)
10	4.268 (0.627)	0.247 (0.020)	**0.202** (0.025)	0.213 (0.023)
11	5.481 (1.379)	2.412 (0.452)	2.020 (0.509)	**1.682** (0.251)†
12	4.956 (0.588)	1.318 (0.072)	1.245 (0.097)	**1.191** (0.070)†
13	14.718 (4.054)	2.346 (0.537)	2.376 (1.319)	**1.838** (0.283)†
14	15.330 (2.294)	0.614 (0.060)	**0.594** (0.120)	0.701 (0.080)
15	17.217 (4.356)	6.301 (0.976)	6.502 (2.766)	**4.660** (0.545)†
16	17.637 (2.342)	3.274 (0.168)	3.473 (0.447)	**3.211** (0.172)†

TAB. 4.2 – Erreur moyenne de test pour la pénalisation quadratique, la sélection pas à pas et pour le modèle additif parcimonieux, ainsi que pour le modèle constant. La sélection de modèle est effectuée par GCV. Les valeurs correspondent à la médiane (écart–type) sur 50 simulations. Pour chacune des situations, la plus petite valeur de l'erreur est marquée en gras. Le symbol † indique que la valeur est plus petite que celle de la sélection pas à pas.

parcimonieux sont montrées sur le tableau (4.2). Pour les trois méthodes, la sélection de modèle est effectuée par GCV. Les valeurs sont des médianes (plus robuste que la moyenne) et des écart–types[1]. Pour chacune des 16 situations, la plus petite valeur de l'erreur parmi les trois méthodes, est marquée en gras. Pour chacune des 16 situations, le symbol † indique que l'erreur en prédiction du modèle additif parcimonieux est plus petite que celle de la sélection pas à pas.

Lorsque le nombre de variables pertinentes est faible devant le nombre total de variables explicatives, la pénalisation quadratique est la méthode qui obtient les plus mauvais résultats. En général, la sélection pas à pas est performante, néanmoins, en présence de corrélation (cas 11, 12), de bruit (cas 3, 11, 12) ou quand la taille de l'échantillon est petite (cas 3, 11), son comportement est perturbé. Finalement, de façon complémentaire à la sélection pas à pas, le modèle additif parcimonieux obtient

[1]Les écart–types, se, sont calculés par $\mathrm{se}(x_{11},\ldots,x_{n1}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{\mathbf{x}}_1)^2$, où $\bar{\mathbf{x}}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$.

les meilleurs résultats quand la corrélation, le bruit ou le faible nombre d'observations rendent difficile l'estimation du modèle.

Inversement, lorsque le nombre de variables pertinentes est important devant le nombre total de variables explicatives, et que la corrélation est faible, la pénalisation quadratique et le modèle additif parcimonieux partagent les meilleurs résultats. En présence d'une corrélation élevée, le modèle additif parcimonieux est le plus performant. Le cas 14, pour lequel le bruit est faible, est une exception pour laquelle la sélection pas à pas obtient les meilleurs résultats.

Comparaison des méthodes de pénalisation par rapport à la stabilité

Exception faite des cas 2 et 6, l'erreur commise par la sélection pas à pas est plus variable que l'erreur commise par les autres méthodes (l'estimation de la sélection pas à pas est donc plus variable). Dans certains cas, notamment quand le nombre de variables pertinentes est élevé et le nombre d'observations faible, cette variabilité est très importante. Les boîtes à moustaches pour la pénalisation quadratique, la sélection pas à pas, le modèle additif parcimonieux (sélectionné par GCV) et le modèle additif parcimonieux (minimisant l'erreur de test) sont représentées dans la figure (4.2). Pour la sélection pas à pas des valeurs éloignées du centre sont souvent observées.

Comparaison des critères de sélection de modèle par rapport à l'erreur en prédiction

Les estimations de l'erreur en prédiction pour les méthodes de sélection du modèle additif parcimonieux sont montrées sur le tableau (4.3). Les performances obtenues par les méthodes AICc, BIC, GCV et CV (*leave–one–out*) sont comparées à la performance optimale (celle obtenue par une méthode minimisant l'erreur moyenne de test, EMT). Pour chacune des 16 situations, la plus petite valeur de l'erreur moyenne de test (parmi les méthodes de sélection) est marquée en gras. Le symbol † indique que la valeur est plus petite que celle de la sélection pas à pas.

Parmi les méthodes de sélection des paramètres de la complexité pour le modèle additif parcimonieux, la GCV obtient les résultats les plus proches des résultats optimaux. La perte occasionnée par la sélection de modèle est, pour la plupart des cas, relativement faible. La méthode CV est, en général, très proche de la GCV. Les résultats obtenus par les méthodes AICc et BIC sont proches entre eux, aussi. Nous ne rapportons pas les résultats obtenus par AIC. Quand la taille de l'échantillon est modérée, ceux–ci coïncident avec ceux obtenus par AICc, en revanche, la version corrigée améliore légèrement la performance d'AIC quand la taille de l'échantillon est petite.

Comparaison des méthodes de pénalisation par rapport à la sélection/élimination de variables

Le tableau (4.4) présente le nombre (moyen) de variables éliminées, le nombre (moyen) de variables non pertinentes éliminées et le nombre de degrés de liberté pour

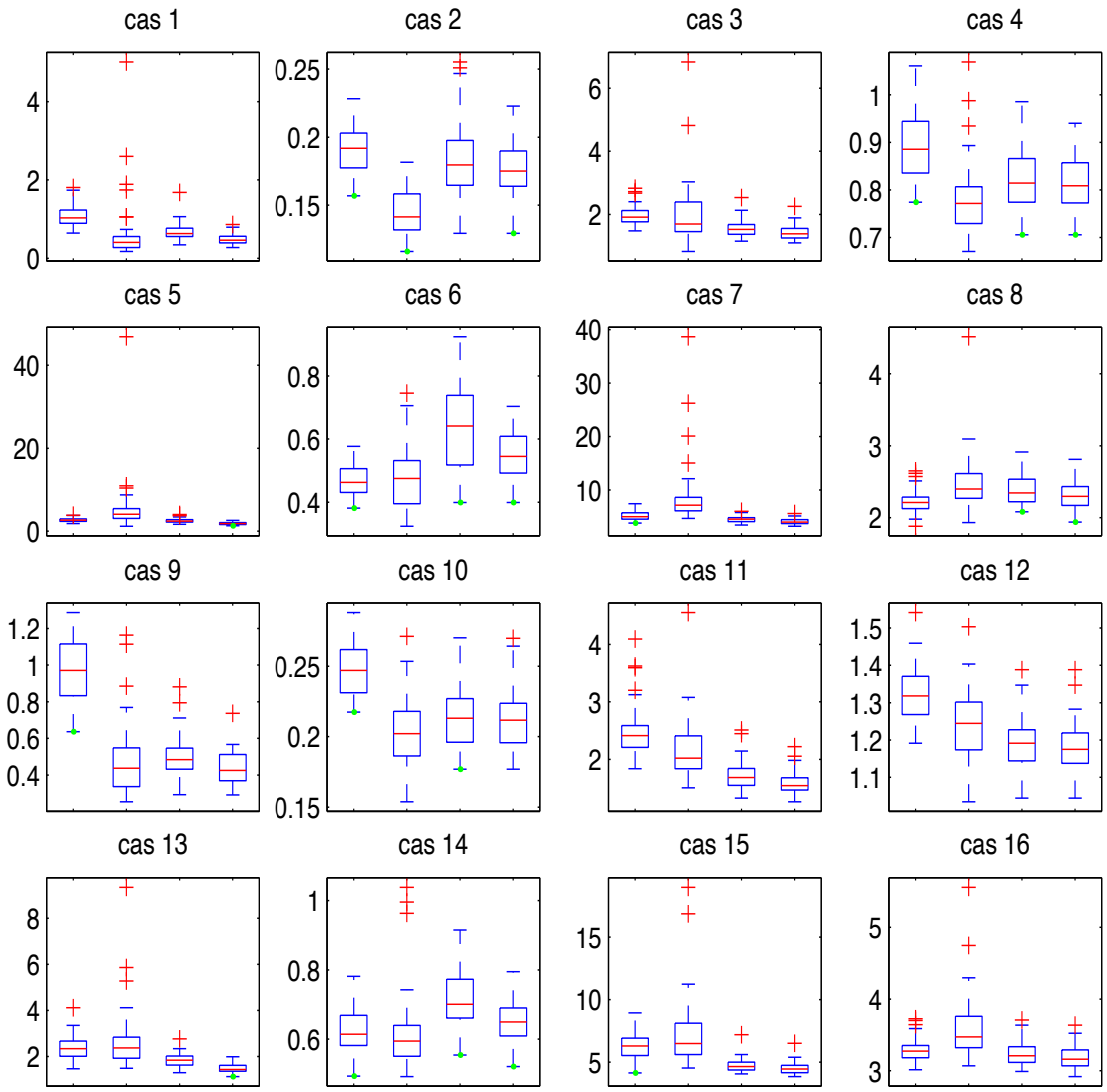

FIG. 4.2 – Boîtes à moustaches pour (de gauche à droite) la pénalisation quadratique, la sélection pas à pas, le modèle additif parcimonieux (sélectionné par GCV) et le modèle additif parcimonieux (minimisant l'erreur de test), pour chacun des 16 cas.

la sélection pas à pas et le modèle additif parcimonieux. Pour ce dernier, les méthodes GCV, AICc et EMT sont considérées. Encore une fois, les résultats obtenus par CV sont proches de ceux obtenus par GCV et les résultats obtenus par AIC et BIC sont proches de ceux obtenus par AICc.

Dans les cas 1–4 et 9–12, il y a 12 variables réellement non pertinentes $(p - d)$ et, dans les cas 5–8 et 13–16, il y a 3 variables non pertinentes. Cependant, la constitution du modèle optimal n'est pas évidente. Par exemple, en présence de corrélation, il est convenable, en général, d'éliminer l'information redondante, mais, lorsque le bruit est élevé, des variables corrélées peuvent apporter de l'information complémentaire.

Cas	EMT	GCV	CV	AICc	BIC
1	0.458 (0.137)	**0.630** (0.219)	0.703 (0.198)	0.668 (0.214)	0.646 (0.228)
2	0.175 (0.020)	**0.180** (0.030)	0.202 (0.044)	0.204 (0.023)	0.204 (0.023)
3	1.388 (0.227)†	**1.523** (0.268)†	**1.523** (0.268)†	1.608 (0.314)†	1.608 (0.304)†
4	0.809 (0.058)	**0.815** (0.064)	**0.815** (0.065)	1.084 (0.082)	1.084 (0.084)
5	1.824 (0.328)†	2.392 (0.490)†	2.394 (0.496)†	**2.304** (0.460)†	**2.304** (0.409)†
6	0.545 (0.073)	0.642 (0.135)	0.738 (0.175)	**0.568** (0.065)	**0.568** (0.065)
7	4.010 (0.510)†	4.525 (0.594)†	**4.391** (0.589)†	4.400 (0.754)†	4.487 (0.810)†
8	2.298 (0.172)†	**2.346** (0.207)†	2.404 (0.218)	2.836 (0.240)	2.836 (0.240)
9	0.426 (0.088)†	**0.484** (0.112)	0.510 (0.119)	0.493 (0.136)	0.493 (0.135)
10	0.212 (0.023)	**0.213** (0.023)	**0.213** (0.023)	0.260 (0.031)	0.260 (0.031)
11	1.541 (0.200)†	1.682 (0.251)†	**1.673** (0.254)†	1.819 (0.370)†	1.819 (0.404)†
12	1.175 (0.065)†	**1.191** (0.070)†	**1.191** (0.071)†	1.562 (0.134)	1.566 (0.135)
13	1.436 (0.196)†	1.838 (0.283)†	1.887 (0.286)†	**1.817** (0.308)†	1.854 (0.319)†
14	0.650 (0.069)	**0.701** (0.080)	0.712 (0.076)	0.731 (0.081)	0.731 (0.081)
15	4.459 (0.463)†	**4.660** (0.545)†	**4.660** (0.524)†	5.069 (0.912)†	5.122 (0.914)†
16	3.162 (0.163)†	**3.211** (0.172)†	3.253 (0.168)†	3.941 (0.296)	3.941 (0.298)

TAB. 4.3 – Erreur moyenne de test des modèles additifs parcimonieux, pour les différentes méthodes de sélection GCV, CV, AICc, et BIC, ainsi que pour le modèle optimal, EMT. Les valeurs correspondent à la médiane (écart–type) sur 50 simulations. Pour chacune des 16 situations, la valeur qui s'approche le plus à l'erreur minimale (EMT) est marquée en gras. Le symbol † indique que la valeur est plus petite que celle de la sélection pas à pas.

La sélection pas à pas élimine un nombre important de variables dans tous les cas où le nombre de variables pertinentes est élevé et le nombre d'observations faible (cas 5, 7, 13 et 15). Dans ces cas, les variables non pertinentes sont identifiées correctement, en revanche un nombre élevé de variables pertinentes sont éliminées, tant en présence qu'en absence de corrélation ou de bruit. Dans les cas 11 et 16, le nombre de variables éliminées est supérieur au nombre de variables réellement non pertinentes, ce qui peut être justifié par la corrélation élevée qui caractérise ces cas.

Les modèles additifs parcimonieux sélectionnent plus fréquemment les variables pertinentes et éliminent moins fréquemment des variables non pertinentes que la sélection pas à pas. En général, le nombre de variables éliminées est en accord avec le nombre de variables non pertinentes éliminées. Les plus grands écarts sont observés dans les cas 7, 13 et 15 où le nombre de variables pertinentes est élevé et le nombre d'observations faible. Dans les deux derniers cas, ceci peut être justifié par la corrélation entre les variables explicatives.

Parmi les techniques de sélection pour les modèles additifs parcimonieux, AICc

		Variables éliminées				Non pertinentes éliminées				Degrés de liberté			
Cas	$p-d$	Pas	GCV	AICc	EMT	Pas	GCV	AICc	EMT	Pas	GCV	AICc	EMT
1	12	10.3	3.9	2.8	1.4	10.2	3.8	2.8	1.4	24.2	13.2	19.0	66.0
2	12	9.3	1.9	0.0	0.6	9.3	1.9	0.0	0.6	32.9	42.8	108.3	55.8
3	12	10.3	4.6†	2.3†	4.5†	9.2	4.4†	2.2†	4.3†	21.6	11.3	20.5	19.5
4	12	9.2	1.8	0.0	1.7	9.2	1.8	0.0	1.7	30.2	29.7	115.0	31.1
5	3	8.6	2.3†	1.2†	0.3†	2.4	1.1†	0.6†	0.2†	26.7	12.6	20.2	99.9
6	3	2.4	0.2	0.0	0.0	2.4	0.2	0.0	0.0	58.5	55.2	117.4	92.2
7	3	9.8	3.1†	1.3†	2.4†	2.3	1.0†	0.4†	0.8†	24.4	12.5	23.4	16.4
8	3	2.5	0.4†	0.0	0.1†	2.4	0.4†	0.0	0.1†	48.2	39.0	119.9	51.1
9	12	11.5	6.0	5.5	4.6†	10.7	5.9	5.4	4.6†	21.7	12.4	15.8	27.2
10	12	9.4	5.4	0.0	3.9	9.4	5.4	0.0	3.9	35.7	21.9	98.0	30.1
11	12	12.4	6.4†	3.3†	7.5†	10.0	5.7†	2.9†	7.0†	19.3	10.5	20.0	9.4
12	12	9.7	4.7†	0.0	4.1†	9.4	4.7†	0.0	4.1†	31.6	18.4	106.6	20.9
13	3	9.3	4.1†	2.2†	1.4†	2.6	1.8†	1.0†	0.7†	25.5	13.0	20.1	62.3
14	3	2.7	1.0	0.0	0.3	2.6	1.0	0.0	0.3	57.5	36.5	105.0	60.7
15	3	11.1	4.6†	2.1†	4.3†	2.4	1.4†	0.6†	1.4†	21.9	11.0	21.2	13.1
16	3	5.9	1.2†	0.0	0.7†	2.4	1.0†	0.0	0.6†	40.6	29.3	108.7	34.5

TAB. 4.4 – Nombre de variables réellement non pertinentes, noté $p-d$, nombre total de variables éliminées (en moyenne) et nombre de variables non pertinentes éliminées (en moyenne) par la sélection pas à pas (notée simplement "Pas") et pour le modèle additif parcimonieux. Pour ce dernier, les méthodes GCV et AICc, ainsi que le modèle optimal (EMT) sont considérés. Le symbol † rappelle quand la méthode de sélection pour le modèle additif parcimonieux est plus performante que la sélection pas à pas, en termes d'erreur en prédiction.

sur–estime et GCV sous–estime, en général, la complexité du modèle, par rapport à la complexité choisie par EMT.

Dans les cas où les modèles additifs parcimonieux sont plus performants que la sélection pas à pas (en termes de l'erreur en prédiction), le nombre de degrés de liberté des premiers est souvent inférieur à celui de la sélection pas à pas. On peut déduire que, même si peu de variables non pertinentes ou redondantes sont éliminées, ces variables restent très pénalisées dans le modèle.

Les figures (4.3) et (4.4) montrent les boîtes à moustaches sur 50 simulations pour les termes de pénalisation individuels linéaires et non linéaires du modèle additif parcimonieux sélectionné par GCV, pour les 18 variables d'entrée et les 16 situations.

La figure (4.3) correspond aux cas où le nombre de variables pertinentes est faible ($j = 1, \ldots, 6$). La figure (4.4) correspond aux cas où le nombre de variables pertinentes est élevé ($j = 1, \ldots, 15$). En moyenne, les valeurs de $1/\mu_j$ et $1/\lambda_j$ correspondant aux

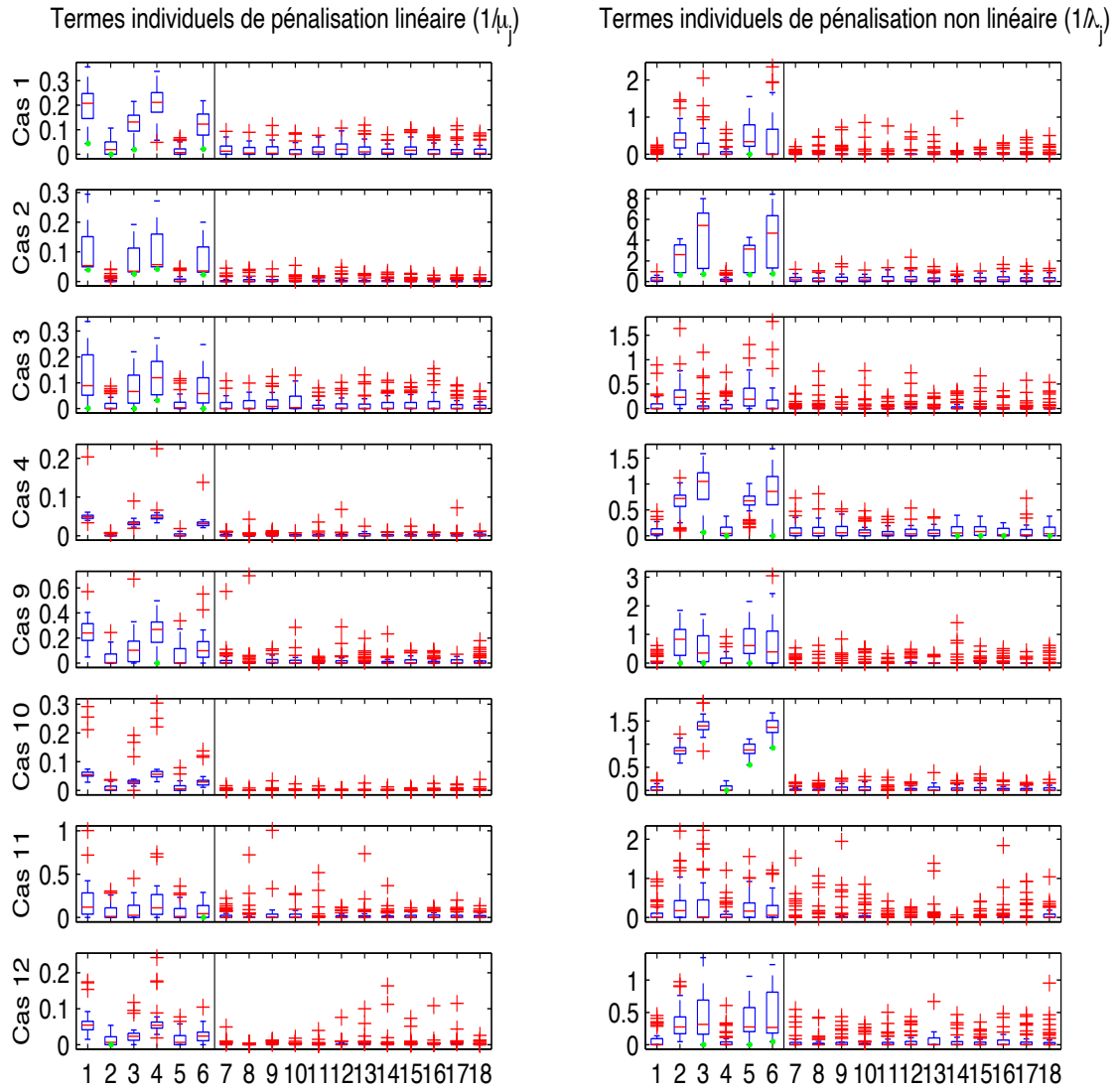

FIG. 4.3 – Boîtes à moustaches pour les termes de pénalisation individuels linéaires $(1/\mu_j)$ et non linéaires $(1/\lambda_j)$, du modèle additif parcimonieux sélectionné par GCV, pour les 18 variables d'entrée et les 8 cas correspondant à 6 variables pertinentes. La ligne verticale dans chaque graphique indique la séparation entre variables pertinentes et non pertinentes.

variables non pertinentes, ($j = 7, \ldots, 18$, pour la figure (4.3), et $j = 16, \ldots, 18$, pour la figure (4.4)), sont proches de zéro, mais rarement exactement nulles. La variabilité de $1/\mu_j$ et $1/\lambda_j$ augmente quand le nombre d'observations est faible (cas impairs) ou le bruit élevé (cas 3, 4, 11 et 12, pour la figure (4.3), et cas 7, 8, 15 et 16, pour la figure (4.4)). Les valeurs $1/\lambda_j$ des variables dont la fonction sous–jacente est strictement linéaire ($j = 1, 4$, pour la figure (4.3), et $j = 1, 4, 7, 10, 13$, pour la figure (4.4)), sont également très pénalisées. Les composantes linéaires des variables dont la fonction

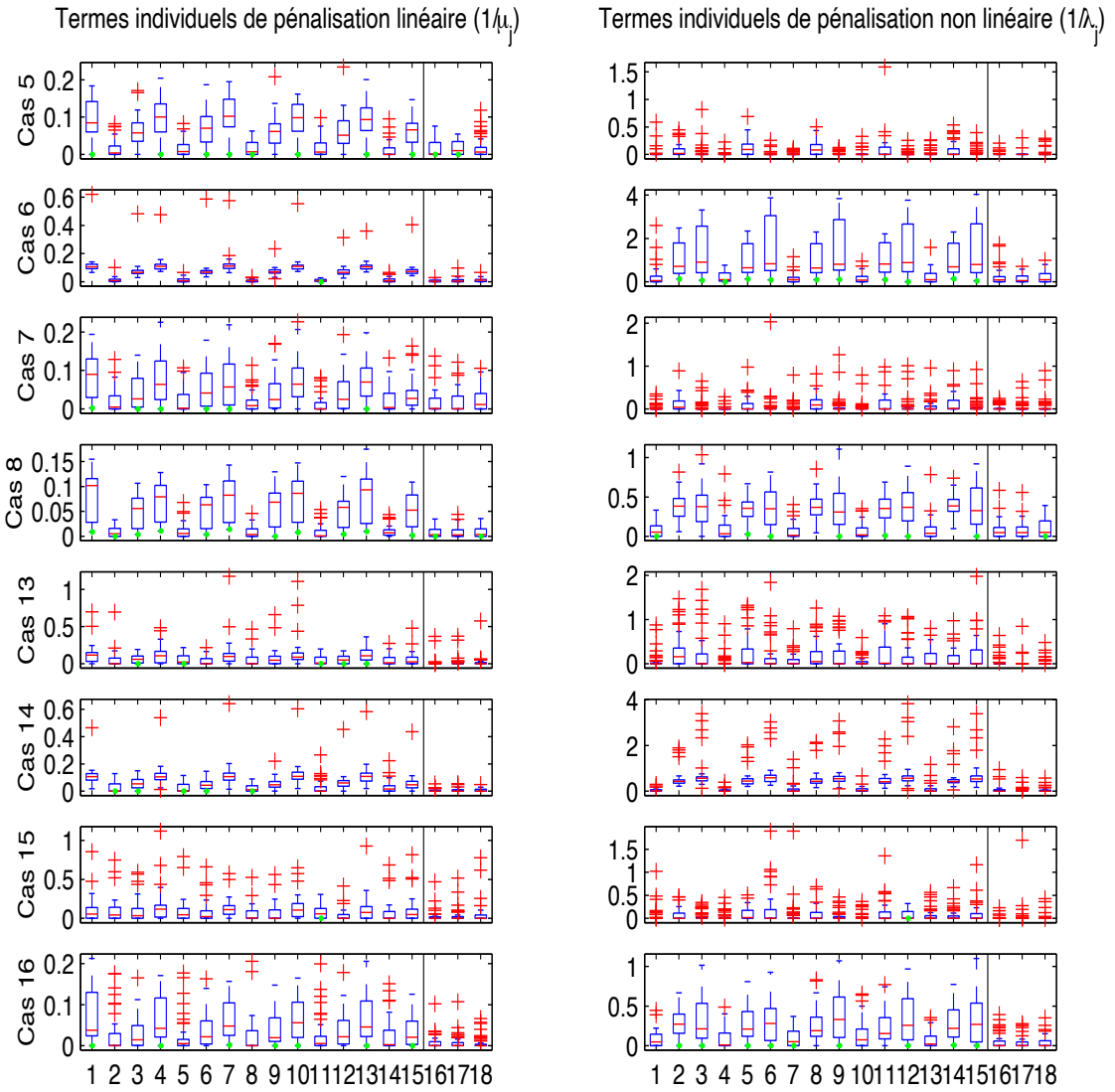

FIG. 4.4 – Boîtes à moustaches pour les termes de pénalisation individuels linéaires $(1/\mu_j)$ et non linéaires $(1/\lambda_j)$, du modèle additif parcimonieux sélectionné par GCV, pour les 18 variables d'entrée et les 8 cas correspondant à 15 variables pertinentes. La ligne verticale dans chaque graphique indique la séparation entre variables pertinentes et non pertinentes.

sous–jacente est un cosinus ($j = 2, 5$, pour la figure (4.3), et $j = 2, 5, 8, 11, 14$, pour la figure (4.4)) sont très pénalisées, mais leurs composantes non linéaires ont un effet important.

En présence de corrélation (cas 9–12 et 13–16), les variables apportant de l'information redondante ne semblent pas être plus pénalisées. En revanche, si on compare une situation où la corrélation est élevée avec son équivalent à corrélation faible (même nombre de variables pertinentes, même niveau de bruit, même taille

Variables pertinentes	Observations	Sélection pas à pas	Modèle additif parcimonieux
$d = 6$	$n = 50$	25.6 (9.4)	33.3 (8.2)
$d = 15$	$n = 50$	30.1 (11.6)	35.1 (8.9)
$d = 6$	$n = 200$	588.8 (132.4)	506.5 (85.6)
$d = 15$	$n = 200$	951.5 (145.5)	556.4 (90.7)

TAB. 4.5 – Temps de calcul en secondes pour la sélection pas à pas et pour le modèle additif parcimonieux sélectionné par GCV. Les valeurs sont des moyennes (écart–type) sur les 50 simulations. Les situations qui diffèrent par rapport au bruit et à la corrélation ont été confondues.

de l'échantillon), on peut observer qu'une pénalisation plus sévère est appliquée à l'ensemble des variables inter–corrélées.

Comparaison des méthodes de pénalisation par rapport au temps de calcul

Finalement, le tableau (4.5) montre le temps de calcul en secondes pour la sélection pas à pas et le modèle additif parcimonieux sélectionné par GCV (les résultats sont similaires pour les autres méthodes). Les valeurs correspondent aux moyennes et écart–types sur les 50 simulations. Les moyennages sont également faits sur les situations qui diffèrent par rapport à la corrélation et au bruit, lesquelles n'introduisent pas de variation du temps de calcul. Seulement les situations qui diffèrent par rapport au nombre de variables pertinentes, d, et au nombre d'observations, n, sont examinées. Les chiffres ne sont qu'indicatifs, car les deux algorithmes peuvent être améliorés de façon importante. Toutefois, on peut conclure que, lorsque le nombre d'observations est assez élevé, le temps de calcul de la sélection pas à pas dépend très fortement du nombre de variables pertinentes, tandis que le temps de calcul du modèle additif parcimonieux n'augmente que légèrement quand le nombre de variables pertinentes augmente.

4.3.4 Conclusions

Les résultats de nos expériences concordent et prolongent, en général, ceux déjà obtenus pour la régression linéaire. Les résultats obtenus ici, dans le contexte non linéaire, sont néanmoins moins catégoriques. D'une part, la situation en soi est plus complexe, d'autre part, plusieurs facteurs agissent sur la difficulté du problème.

Comparaison des méthodes de pénalisation par rapport à l'erreur en prédiction

Rappelons que la comparaison expérimentale dans le contexte linéaire des méthodes de pénalisation montre qu'il n'y a pas de méthode optimale pour toute situation et que la corrélation des variables explicatives ainsi que le nombre d'entrées

réellement pertinentes sont deux facteurs déterminants pour le choix d'une méthode de pénalisation [Breiman, 1996, Tibshirani, 1996, Boukari et Grandvalet, 1998] (section 3.2.1.1). La sélection pas à pas est la mieux adaptée, suivie du lasso, lorsque le nombre d'entrées significatives est très petit devant le nombre total de variables explicatives, et que ces variables sont peu corrélées. La pénalisation quadratique est appropriée quand la majorité des entrées sont significatives ou qu'elles sont très corrélées. Enfin, le lasso obtient les meilleurs résultats dans les cas intermédiaires et, dans les autres cas, ses résultats sont proches de ceux de la meilleure méthode.

Les résultats de nos expériences dans le domaine de la régression additive permettent de conclure qu'il n'y a pas de méthode optimale pour toute situation, mais, contrairement au cas linéaire, une des méthodes (la pénalisation quadratique) n'est pas adaptée au problème. La corrélation des variables explicatives, le nombre d'entrées réellement pertinentes, le bruit et la taille de l'échantillon sont des facteurs importants pour le choix d'une méthode de pénalisation :

- La sélection de variables pas à pas est la mieux adaptée lorsque le nombre de variables pertinentes est faible devant le nombre total de variables explicatives, et que l'occurrence des autres facteurs de difficulté n'est pas simultanée : soit on est seulement en présence de corrélation élevée, soit on est seulement en présence de bruit, soit on est seulement en présence d'un nombre faible d'observations.
- La pénalisation quadratique est appropriée uniquement quand la majorité des entrées sont significatives, peu corrélées et que la taille de l'échantillon est modérée.
- Le modèle additif parcimonieux obtient les meilleurs résultats lorsque 1) le nombre de variables pertinentes est faible devant le nombre total de variables explicatives, et qu'au moins deux facteurs de difficulté (corrélation, bruit et/ou nombre d'observations faible) sont présents ; 2) la majorité des entrées sont significatives, exceptant quelques cas où l'occurrence des facteurs de difficulté n'est pas simultanée.

Les résultats obtenus pour la généralisation de la pénalisation quadratique à la régression non paramétrique additive diffèrent de ceux obtenus pour la pénalisation quadratique dans le contexte linéaire. Des trois, cette méthode est la moins performante. L'hypothèse d'une complexité commune à toutes les composantes est trop restrictive dans le cas non linéaire, où les fonctions sous–jacentes peuvent présenter des courbures très différentes.

Les résultats obtenus pour la généralisation de la sélection pas à pas et du lasso à la régression non paramétrique additive concordent, en général, avec ceux déjà obtenus pour la régression linéaire. Ils sont néanmoins moins nets.

Comparaison des méthodes de pénalisation par rapport à la stabilité

Nous avons traité précédemment (section 3.2.1) la stabilité du lasso et de la pénalisation quadratique, en opposition à l'instabilité de la sélection pas à pas. De façon similaire, nous constatons que dans le contexte non paramétrique additif la sélection pas à pas présente, le plus souvent, une variabilité supérieure aux autres méthodes.

Comparaison des critères de sélection de modèle par rapport à l'erreur en prédiction

Parmi les techniques de sélection de modèle pour les modèles additifs parcimonieux, la GCV est la plus performante (proche de la meilleure performance possible). La CV est proche de la GCV mais ses résultats sont légèrement moins bons. Les méthodes AIC, AICc et BIC sont proches entre elles et éloignées des autres. Une explication possible de la bonne performance de la GCV par rapport à AIC, AICc et BIC est que, contrairement à ces trois dernières, la GCV ne nécessite pas d'estimation de la variance de l'erreur. Dans le cas linéaire, le lasso sélectionné par GCV obtient des meilleurs résultats que le lasso sélectionné par CV à 5 blocs [Tibshirani, 1996]. Cependant, dans notre cas il s'agit de la version *leave–one–out* de la CV, on s'attend donc à que les résultats de deux méthodes soient similaires.

Comparaison des méthodes de pénalisation par rapport à la sélection/élimination de variables

La sélection pas à pas identifie correctement les variables non pertinentes, cependant dans certaines situations (notamment quand le nombre de variables pertinentes est élevé et le nombre d'observations faible), un nombre élevé de variables pertinentes sont éliminées, tant en présence qu'en absence de corrélation ou de bruit.

Parmi les techniques de sélection pour les modèles additifs parcimonieux, BIC choisi un modèle légèrement plus simple que AIC et AICc, cependant les trois méthodes sur–estiment, en général, la complexité du modèle, par rapport à la complexité choisie par EMT. Une explication possible est que ces méthodes demandent l'estimation de la variance de l'erreur, laquelle est estimée pour des valeurs de μ et λ comportant une complexité élevée [Ruppert *et al.*, 2003]. Concrètement, μ et λ sont les plus petites valeurs considérées dans la grille, ce qui pourrait provoquer des problèmes d'instabilité numérique.

La méthode GCV, elle, sous–estime, en général, la complexité du modèle. Ceci contraste avec les résultats obtenus par l'application de la GCV à d'autres modèles tels que les splines avec une seule variable d'entrée [Wahba et Wang, 1995] ou le modèle additif standard [Kim et Gu, 2004]. La définition du nombre de degrés de liberté, induisant une prédiction conservatrice (section 3.4.1), est plausiblement à l'origine de décalage du problème.

Les modèles additifs parcimonieux identifient correctement les variables pertinentes. En revanche, ils éliminent peu de variables non pertinentes et de variables redondantes. Ces variables restent très pénalisées dans le modèle. Ces résultats coïncident avec ceux déjà obtenus pour le cas linéaire. En effet, des travaux montrent que, généralement, le lasso élimine peu de variables [Tibshirani, 1996, Steyerberg *et al.*, 2000]. Dans le cas non linéaire, l'élimination d'une variable demande l'élimination de ses parties linéaire et non linéaire, ce qui peut rendre plus difficile l'annulation exacte de la variable. Une solution possible consiste à introduire un seuil à partir duquel les variables très pénalisées, mais non

nulles, seraient éliminées. Une idée proche est développé par [Perkins *et al.*, 2003], en proposant un critère d'optimisation intégrant les pénalisations l_0, l_1 et l_2.

Comparaison des méthodes de pénalisation par rapport au temps de calcul

Finalement, par rapport au temps de calcul, le modèle additif parcimonieux est plus avantageux que la sélection pas à pas lorsque le nombre d'observations et le nombre de variables pertinentes sont modérés ou élevés. En effet, le nombre de variables pertinentes a un effet plus drastique sur le temps de calcul de la sélection pas à pas que sur celui du modèle additif parcimonieux.

D'autres simulations

Nous n'avons pas exploré ici des situations différentes au niveau de la dispersion des entrées ou de la concurvité, afin de ne pas rendre les résultats confus.

La concurvité est prise en compte par [Avalos *et al.*, 2004c], pour des échantillons de petite taille. Cependant, la capacité d'éliminer des variables apportant de l'information redondante n'a pas été analysée par rapport à la nature des dépendances (linéaires et non linéaires).

4.4 Données réelles

Dans cette section nous évaluons le modèle logistique additif parcimonieux sur des jeux de données médicales réelles.

Le modèle logistique (section 1.4.2) est fréquemment utilisé en épidémiologie lorsque la variable réponse Y est binaire et que la fréquence de l'événement auquel on s'intéresse (décès, maladie,...) est mesurée par un risque [Bouyer *et al.*, 1995, Avalos *et al.*, 2004d]. C'est le cas s'il s'agit d'une enquête cas–témoins (exemple de la section 4.4.1) ou si on s'intéresse à la survenue de l'événement au cours d'une période fixée (exemple de la section 4.4.2).

Deux raisons principales ont conduit au choix de la fonction logistique : 1) Elle permet d'évaluer l'association entre l'événement et l'exposition aux facteurs de risque de façon cohérente avec l'odds ratio, OR, mesure usuelle de la relation entre la maladie et les facteurs de risque. 2) Cette fonction a une forme sigmoïde qui correspond à une forme de relation souvent observée entre une dose et la fréquence de l'événement.

4.4.1 Difformité vertébrale post–opératoire

Dans un premier temps, nous reprenons l'étude de cas *kyphosis* utilisée par [Hastie et Tibshirani, 1990, Tibshirani, 1996] pour illustrer les différences entre la régression logistique par modèle additif standard, et la régression logistique pénalisée. La variable réponse indique la présence ou l'absence de difformité vertébrale post–opératoire (*kyphosis*) chez des enfants. Il y a 83 exemples, dont 18 étiquetés *kyphosis*. Les variables d'entrée sont l'âge des enfants en mois (X_1), le nombre de vertèbres touchées par l'opération (X_2) et la position de la première vertèbre concernée (X_3).

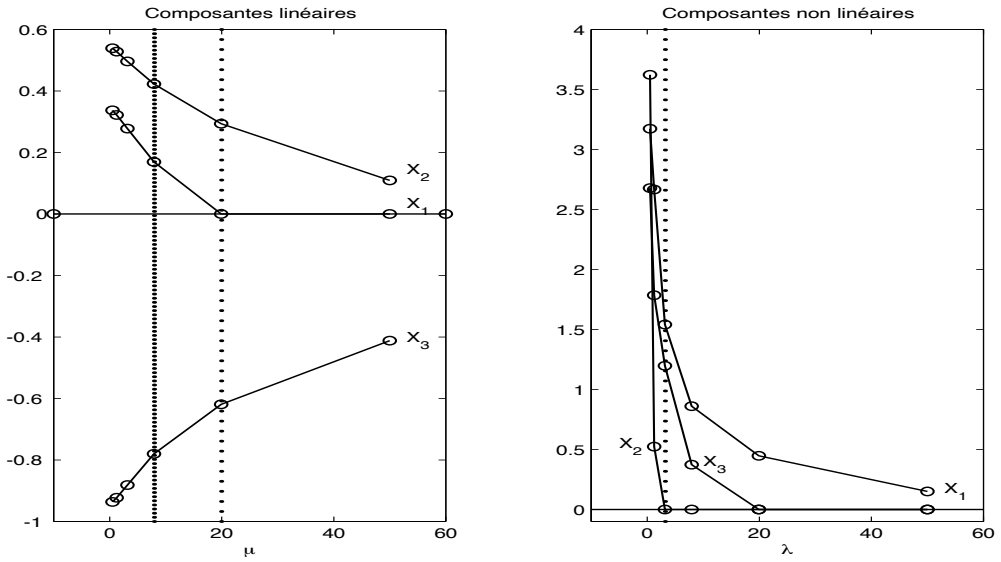

FIG. 4.5 – Coefficients des composantes linéaires, α_j, et norme des coefficients des composantes non linéaires, $(\widetilde{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j)^{1/2}$, en fonction des paramètres de la complexité correspondants. Le graphique de gauche correspond à $\lambda = 1.2$, et celui de droite à $\mu = 4.2$, mais l'allure des courbes est similaire pour tous les μ et λ. Les lignes verticales indiquent la complexité choissie par les critères.

Les données sont centrées réduites, afin de rendre les pénalisations comparables. Les critères de sélection sont évalués sur une grille 6×6 de valeurs de (μ, λ) à échelle logarithmique.

La partie gauche de la figure (4.5) montre les coefficients linéaires en fonction du paramètre de pénalisation linéaire. La norme des coefficients non linéaires en fonction du paramètre de pénalisation non linéaire est représenté à droite. Les lignes verticales pointillées indiquent la valeur sélectionnée par les critères AIC, AICc, GCV et BIC. Pour la partie linéaire, BIC (ligne pointillée peu dense) a choisi un modèle plus simple que les autres critères (ligne pointillée dense). Pour la partie non linéaire, les quatre méthodes ont effectué le même choix (ligne pointillée).

La figure (4.6) montre l'effet de chaque variable sur la fonction logit, estimé par 4 modèles : logistique additive avec 3 degrés de liberté pour chaque composante, M1 ; logistique linéaire pénalisée (lasso), M2 ; et logistique additive pénalisée, pour les paramètres de la complexité sélectionnés par les différents critères, M3 (AIC, AICc et GCV) et M4 (BIC).

Etant donnée la difficulté de sélectionner les p paramètres de lissage du modèle additif standard, celui–ci est souvent appliqué comme un outil d'analyse exploratoire. Ainsi, le modèle M1 [Hastie et Tibshirani, 1990] suggère des termes quadratiques, lesquels sont intégrés dans le modèle paramétrique M2 [Tibshirani, 1996]. La répartition automatique de la complexité est alors possible, aboutissant à un modèle linéaire en X_2 et X_3 et quadratique en X_1. La méthode que nous proposons permet de distribuer la complexité de chaque variable de façon automatique, sans l'intermédiaire

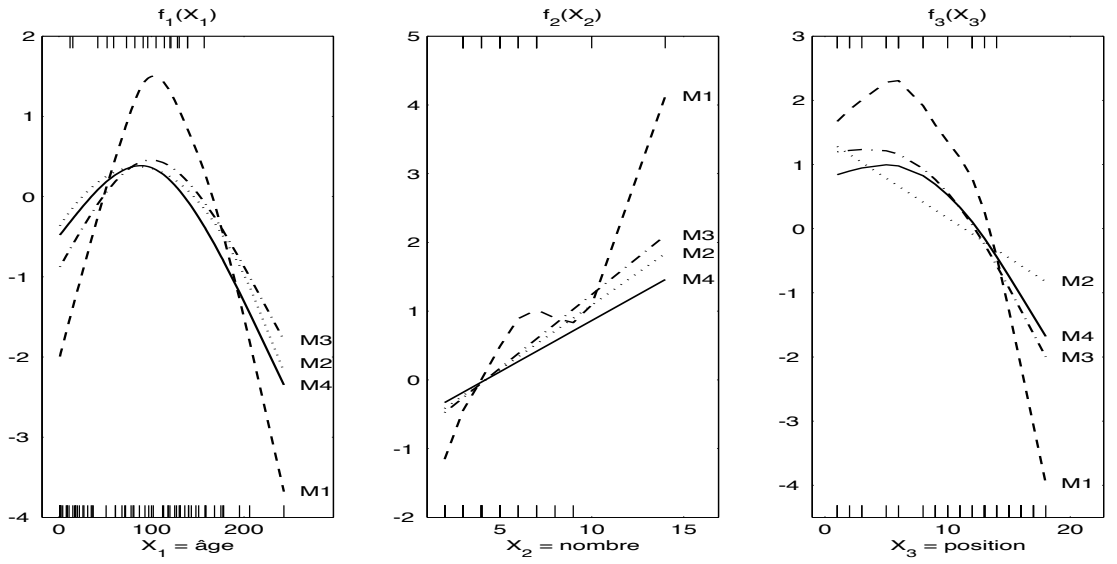

FIG. 4.6 – Composantes additives ajustées par : le modèle logistique additif (M1, ligne discontinue) ; le modèle logistique lasso (M2, ligne pointillée) ; le modèle logistique parcimonieux sélectionné par AIC, AICc et GCV (M3, ligne point-tirets) ; le modèle logistique parcimonieux sélectionné par BIC (M4, ligne continue). Les bâtons en haut et en bas des graphiques indiquent les observations de présence et absence de kyphosis, respectivement.

d'une approximation paramétrique.

Les courbes M3 et M4 sont similaires à la courbe M2 pour les variables X_1 et X_2, et plus complexes pour la troisième variable. En observant l'estimation obtenue par le modèle additif M1, on s'attend à ce que le modèle paramétrique conserve le terme quadratique. En fait, cet exemple montre que le lasso appliqué sur le modèle paramétrique peut produire des résultats contre–intuitifs. Ici, dans la base $\{1, x, x^2\}$ une composante en $\alpha(x+1)^2$ est jugée plus "complexe" qu'une composante en αx^2. En effet, la première se développant en $\alpha(x^2+2x+1)$, elle est pénalisée par 3α, alors que la première est pénalisée par α. Dans la base $\{1, x, (x+1)^2\}$ le phénomène inverse serait observé. Notre algorithme, pour lequel la pénalisation de la partie non linéaire n'est affectée que par la courbure de la fonction, est insensible à ce problème de représentation.

La représentation graphique permet de déduire que : 1) le risque de kyphosis (le odds ratio, plus précisément) augmente jusqu'à l'âge moyen (environ 85 mois), et ensuite il décroît ; 2) le risque augmente quand le nombre de vertèbres touchées augmente (l'augmentation du risque est de 4.5% par vertèbre, selon M3, et de 3%, selon M4) ; et 3) le risque, élevé, stagne jusqu'à une certaine position (vertèbre numéro 10), et décroît ensuite rapidement. D'autre part, nous remarquons que, pour la classe "absence de kyphosis", des valeurs élevées de la variable âge, éloignées du centre des observations, semblent avoir une forte influence sur l'estimation. Il s'agit de trois sujets âgées de 195, 206 et 243 mois, respectivement (16, 17 et 20 ans, respectivement).

	AIC/AICc	BIC	GCV	CV	CV_v
(μ, λ)	$(10, 10)$	$(100, 100)$	$(10, 0.01)$	$(0.01, 0.1)$	$(1, 10)$
Erreur	0.71 (0.14)	0.78 (0.10)	0.95 (0.14)	0.63 (0.14)	0.69 (0.14)
Sensibilité	66.7%	86.7%	40.0%	66.7%	60.0%
Spécificité	61.3%	28.2%	66.1%	69.4%	71.0%

TAB. 4.6 – Valeurs de (μ, λ) choisies par les techniques de sélection de modèle, ainsi que leur erreur moyenne (écart–type), sensibilité et spécificité sur l'ensemble de test.

Il conviendrait considérer la pertinence de ces observations.

4.4.2 Risque cardio–vasculaire

Le projet INDANA (Individual Data Analysis of Antihypertensive Intervention Trials) s'inscrit dans le cadre de la prédiction individualisée du risque cardio–vasculaire chez des patients présentant une hypertension artérielle, en vue d'aider la décision des médecins praticiens dans le domaine de la prévention cardio–vasculaire [Gueyffier *et al.*, 1995]. La base de données INDANA réunit les données individuelles de 10 essais thérapeutiques (contrôlés randomisés) conduits pour évaluer l'efficacité des traitements anti–hypertenseurs. Cette base de données a été mise en forme et est maintenue dans l'Unité de Pharmacologie Clinique de L'Université de Lyon 1 (chef de projet : F. Gueyffier).

Le modèle logistique additif a été préalablement utilisé sur un des essais de la base INDANA (Shep). La sensibilité (proportion des décédés bien classés) et la spécificité (proportion des non décédés bien classés) sur une validation croisée à 10 blocs stratifiée ont mesuré la performance de la méthode. Les valeurs de sensibilité et de spécificité obtenues par la régression additive sont de 66.36% et 66.37%, respectivement. Sur ces données, le modèle logistique additif a donné les meilleurs résultats par rapport aux autres méthodes testées, à savoir *balanced–bagging*, *C4.5*, *Fôret–Floue–T–norme*, *Fôret–Floue–stricte*, *Framingham*, *GloBoost*, *Pocock*[2].

La régression logistique additive parcimonieuse est appliquée ici au groupe de contrôle d'un des essais (Coope) [Avalos *et al.*, 2004a, Avalos *et al.*, 2004b]. Les données extraites sont constituées de 9 variables d'entrée : sexe, tabagisme, facteur de risque (antécédent d'angor, d'infarctus myocardiaque, d'accident cardio–vasculaire, ou hypertrophie ventriculaire), âge, pression systolique, pression diastolique, cholestérol, uricémie, et indice pondéral. Les trois premières variables sont binaires. Elles sont modélisées et pénalisées linéairement, ce qui correspond à assigner un coefficient par modalité. Les six dernières variables, continues, ont été centrées et réduites pour les estimations. La variable de sortie est le décès cardio–vasculaire. Il y a 413 exemples, dont 43 décès.

Deux tiers de la base Coope (274 exemples, dont 28 décès) constituent l'ensemble d'apprentissage, et un tiers (139 exemples, dont 15 décès) constitue l'ensemble de

[2]Ces résultats sont disponibles sur Internet à
`http ://www.grappa.univ-lille3.fr/~torre/Recherche/Indana`

test. Le critère de comparaison est la proportion d'exemples mal classés. Cette erreur est calculée en appliquant le coût $\{1, 10\}$, afin d'équilibrer les données. La sensibilité et spécificité de chaque méthode sont également calculées.

Nous testons les méthodes AIC, AICc, BIC, GCV, ainsi que validation croisée stratifiée sur 10 sous–ensembles. Pour cette dernière, nous rapportons les résultats obtenus selon deux critères : l'erreur de classement (avec les coûts $\{1, 10\}$), CV, d'une part, et la vraisemblance, CV_v, d'autre part.

Les méthodes analytiques et de rééchantillonnage sont évaluées sur une grille 5×5 de valeurs de (μ, λ), régulièrement espacées sur une échelle logarithmique. Afin d'éviter des problèmes numériques, qui peuvent apparaître pour les splines de lissage lorsque le nombre d'observations est élevé, des P–splines ont été utilisées (sections 1.2.3.2 et 1.3.5.4). Ainsi, 100 nœuds (au lieu de 274) ont été placés sur les centiles de chacune des variables (continues) d'entrée. Les résultats sont présentés dans le tableau (4.6).

Les deux versions de la validation croisée réalisent la meilleure performance. Celle basée sur l'erreur de classification, CV, est proche de la performance de test optimale obtenue, 0.62 (0.14), pour les valeurs $(0.01, 100)$ de (μ, λ). Néanmoins, cette méthode attribue une complexité excessive aux parties non linéaires. Une explication possible est que CV sélectionne un minimum local de l'erreur de classification. La CV_v, basée sur la vraisemblance, sur–estime la pénalisation sur les parties linéaires et sous–estime, légèrement, la pénalisation sur les parties non linéaires. Les valeurs de (μ, λ) sélectionnées par cette méthode sont, toutefois, très proches (échelle logarithmique) des valeurs $(10, 100)$, qui maximisent la log–vraisemblance sur l'ensemble de test. L'estimation en termes de probabilités ou de frontière de décision n'aboutit pas nécessairement à la même solution [Friedman, 1997].

Les méthodes analytiques basées sur la vraisemblance, AIC, AICc et BIC, induisent des erreurs élevées. Cependant, pour les deux premières, (lesquelles aboutissent à des résultats identiques), les erreurs ne sont pas éloignées de celle du maximum de vraisemblance sur l'ensemble de test et de CV_v. Elles sous–estiment la complexité des parties linéaires, contrairement à ce qu'on pourrait attendre (section 3.4.1.1), et sur–estiment légèrement la complexité des parties non linéaires. La GCV est la méthode analytique dont l'erreur est la plus grande. Les exemples "décès" sont particulièrement mal classés, ce qui conduit à une erreur de classement (avec les coûts $\{1, 10\}$) très élevée.

La figure (4.7) montre les coefficients linéaires en fonction du paramètre de pénalisation linéaire, à gauche, et la norme des coefficients non linéaires en fonction du paramètre de pénalisation non linéaire, à droite. Les courbes en noir correspondent aux trois variables binaires (ligne discontinue pour X_1, ligne continue pour X_2 et ligne point-tirets pour X_3). Les autres couleurs correspondent aux six variables continues (rouge pour X_4, bleu clair pour X_5, rose pour X_6, vert pour X_7, jaune pour X_8 et bleu pour X_9). Le graphique de gauche correspond à $\lambda = 100$, et celui de droite à $\mu = 0.01$. L'allure des courbes est similaire pour tous les μ, mais elle diffère légèrement pour les valeurs de λ. Ainsi, pour une valeur faible de λ, la variable X_2 a toujours un effet plus important que la variable X_3, tandis que pour une valeur forte, telle que dans la figure (4.7), leurs courbes se croissent. La variable X_9 a dans tous les cas un effet

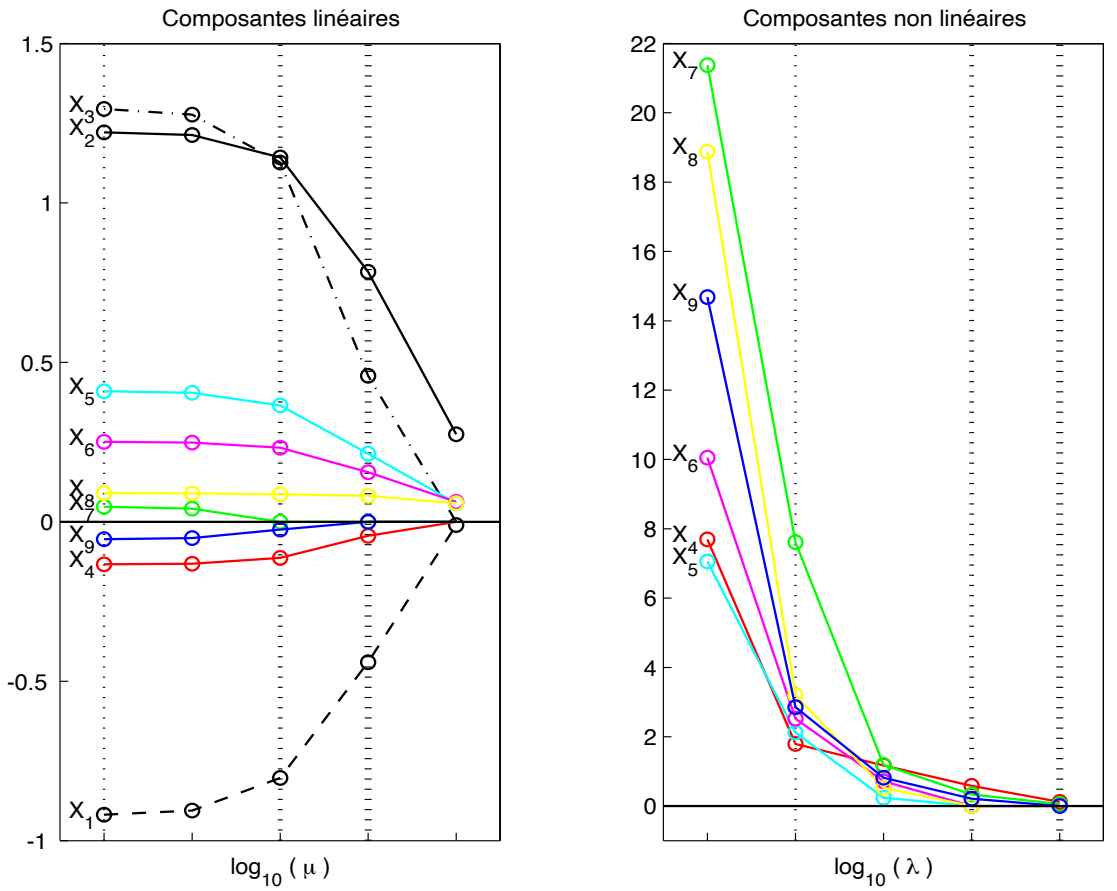

FIG. 4.7 – Coefficients des composantes linéaires, α_j, et norme des coefficients des composantes non linéaires, $(\widetilde{\boldsymbol{\beta}}_j^t \boldsymbol{\Omega}_j \widetilde{\boldsymbol{\beta}}_j)^{1/2}$, en fonction des paramètres de la complexité correspondants. Les lignes verticales indiquent des valeurs des paramètres de la complexité sélectionnées par les différents critères.

linéaire très faible, mais celui–ci est positif pour une valeur faible de λ et négatif, comme dans la figure (4.7), pour λ important.

Les lignes verticales pointillées indiquent différentes valeurs des paramètres de la complexité. Dans le cas linéaire, ces valeurs correspondent à la valeur sélectionnée par CV, coïncidant avec la valeur qui minimise l'erreur de classification, (ligne pointillée fine) ; la valeur sélectionnée par CV_v (ligne pointillée moyenne) ; et la valeur qui maximise la vraisemblance, (ligne pointillée épaisse). Dans le cas non linéaire, ces valeurs correspondent à la valeur sélectionnée par CV (ligne pointillée fine) ; la valeur sélectionnée par CV_v (ligne pointillée moyenne) ; et la valeur qui minimise l'erreur de classification et qui maximise la vraisemblance, (ligne pointillée épaisse).

La représentation graphique des résultats permet d'utiliser le modèle additif logistique en tant qu'outil d'analyse exploratoire. La figure (4.8) montre les estimations des fonctions additives pour la valeur de (μ, λ) minimisant l'erreur de classification, $(0.01, 100)$. La fonction logit (proche de la fonction risque de décès cardio–vasculaire)

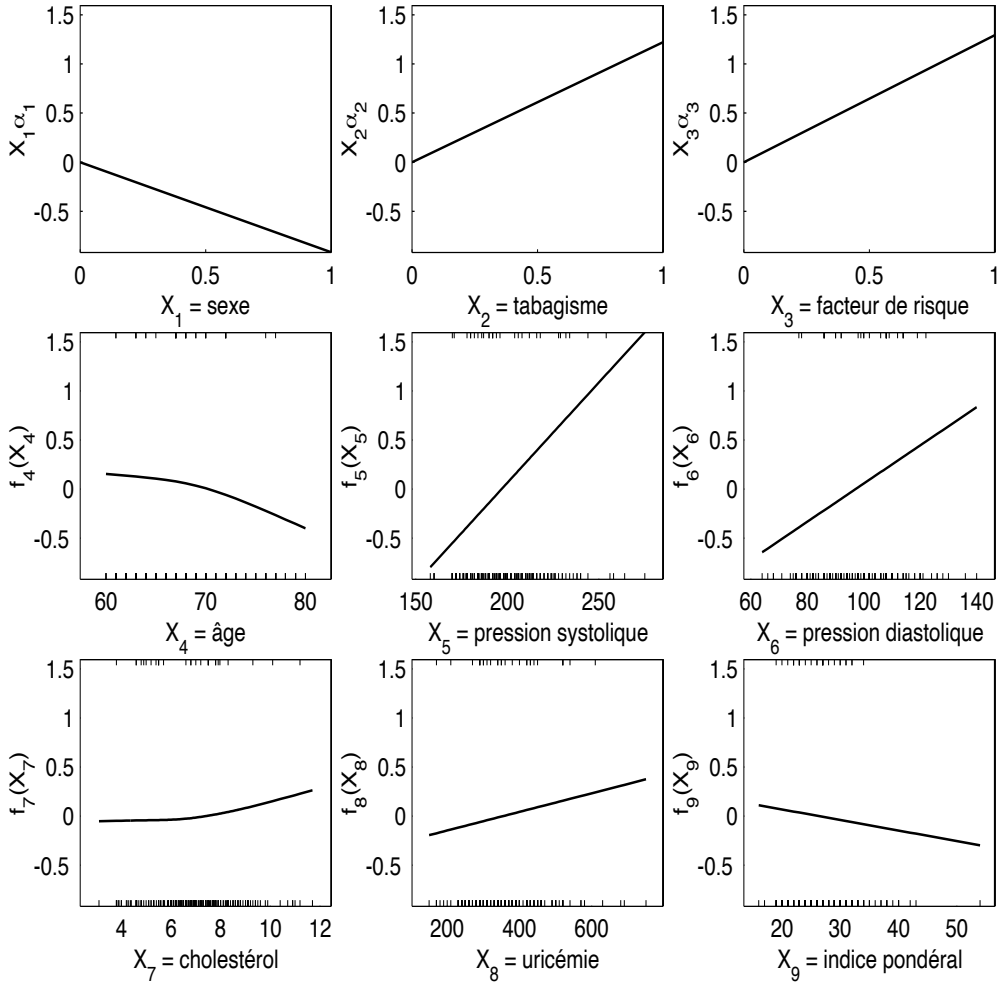

FIG. 4.8 – Composantes additives du modèle logistique additif parcimonieux évaluées sur l'ensemble d'apprentissage. Les valeurs de (μ, λ) sont celles qui minimisent l'erreur de classification. Les bâtons en haut et en bas des graphiques indiquent si les observations correspondent à un sujet décédé ou vivant, respectivement.

est obtenue comme la somme de ces fonctions : logit$=\widehat{\alpha}_0+\sum_{j=1}^{3} x_j\widehat{\alpha}_j+\sum_{j=4}^{9}\widehat{\mathbf{f}}_j(x_j)$. Les trois premières variables, binaires, sont modélisées et pénalisées linéairement. Parmi les six variables continues, l'effet des variables X_5, X_6, X_8 et X_9 est estimé linéaire. Le modèle additif parcimonieux réduit ainsi à un modèle plus simple, quand une complexité élevée n'est pas adaptée. Par ailleurs, la base de données a été préalablement traitée. En particulier, des variables considérées non significatives ont été éliminées par des procédures non automatiques. Ceci explique le fait qu'aucune variable ne soit éliminée.

Afin de comparer la contribution de chaque variable sur la réponse, les fonctions estimées sont représentées sur la même échelle. On peut observer que les variables binaires sont les facteurs les plus influents. Ces variables obtiennent, en effet, les

valeurs des coefficients linéaires les plus élevées (ou de façon équivalente, les termes de pénalisation linéaire les plus faibles), suivies de la pression systolique et de la pression diastolique. Le risque de décès cardio–vasculaire est ainsi supérieur pour les hommes (codé 0), pour les fumeurs (codé 1), et en cas d'antécédent d'angor, ou d'antécédent d'infarctus myocardiaque, ou d'antécédent d'accident cardio–vasculaire, ou d'hypertrophie ventriculaire (codé 1). L'augmentation du risque en fonction de la pression systolique et diastolique est de 41.0% et 25.1% par unité de pression, respectivement.

Bien que la variable âge n'ait pas d'influence importante sur le risque, celle–ci est négative : quand l'âge augmente le risque décroît, ce qui est étonnant. Il conviendrait d'analyser si la faible représentation des âges élevés dans la classe des sujets décédés est liée aux critères d'inclusion des individus dans l'étude. Une situation similaire est rencontrée pour l'indice pondéral. En effet, une valeur très élevée de cette variable dans la classe des sujets non décédés, correspondant à un sujet présentant une obésité sévère, a une forte contribution sur l'estimation de la fonction. La suppression de ce point a comme conséquence l'élimination de la variable indice pondéral. Le risque, faible, stagne jusqu'à une certaine valeur de la variable cholestérol (≈ 6.5), et croît ensuite mais lentement. Finalement, l'augmentation du risque en fonction de la variable uricémie est de 4.7% par unité.

4.5 En bref

Dans ce chapitre nous avons traité la mise en œuvre des méthodes développées. Le plan d'expériences utilisé dans nos simulations tient compte des situations dans lesquelles le contrôle de la complexité est particulièrement délicat pour les modèles additifs. Au même temps, nous avons paramétrisé ces situations visant à rendre les résultats concis et clairs.

Les résultats de nos expériences concordent et prolongent, en général, ceux déjà obtenus pour la régression linéaire. Sommairement, on peut dire que la sélection pas à pas est plus performante dans les situations moins complexes, tandis que les modèles additifs parcimonieux obtiennent les meilleurs résultats dans les situations plus complexes. Ces derniers sont également plus stables.

En ce qui concerne la sélection de modèle pour les modèles additifs parcimonieux, la validation croisée généralisée est la méthode la plus performante, proche de la meilleure performance possible.

Quant à la sélection et élimination des variables du modèle, la sélection pas à pas identifie correctement les variables non pertinentes, mais elle peut éliminer un nombre élevé de variables pertinentes. Les modèles additifs parcimonieux identifient correctement les variables pertinentes. En revanche, ils éliminent peu de variables non pertinentes et de variables redondantes. Ces variables restent très pénalisées dans le modèle.

Finalement, par rapport au temps de calcul, le modèle additif parcimonieux est plus avantageux que la sélection pas à pas lorsque le nombre d'observations et le nombre de variables pertinentes sont modérés ou élevés.

En ce qui concerne l'application à des données réelles, nous constatons l'utilité de ces modèles en tant qu'outil d'analyse exploratoire. En effet, la représentation graphique des résultats permet d'étudier les effets de chaque variable sur le risque. En outre, le modèle logistique additif parcimonieux permet la répartition automatique de la complexité parmi les composantes additives. Seuls les deux paramètres qui règlent la complexité globale doivent être déterminés par des critères de sélection de modèle. Néanmoins, l'adaptation de ces méthodes pour le modèle logistique est moins satisfaisante.

Conclusion

Les modèles additifs généralisent les modèles linéaires en proposant une solution flexible qui préserve la capacité à décrire graphiquement les dépendances. En conséquence, ces modèles sont appliqués dans de nombreux domaines tels que l'économie [Smith et Kohn, 1996, Beck et Jackman, 1998], l'ingénierie [Walker et Wright, 2002], ou l'épidémiologie [Bacchetti et Quale, 2002, Dominici *et al.*, 2002].

La plupart de ces applications se limitent à un nombre réduit de variables d'entrée (exceptionnellement plus de 5), sélectionnées par une étude préalable. Les difficultés actuelles de la mise en œuvre des modèles additifs expliquent leur application restreinte. Les méthodes analytiques de sélection de modèle, pour lesquelles des algorithmes efficaces ont été proposés, aboutissent à des résultats satisfaisants quand il s'agit d'estimer la complexité des composantes additives. Cependant, quand la sélection de la complexité doit aboutir à la suppression de variables, le problème devient impraticable même pour un nombre modéré d'entrées.

Dans ce mémoire, nous avons proposé une nouvelle méthode d'estimation fonctionnelle pour les modèles additifs basée sur une généralisation du lasso. Celle–ci est motivée par les bons résultats obtenus par cette méthode de pénalisation dans le cadre linéaire.

Notre stratégie se base sur une décomposition des espaces de fonctions splines, comprenant, d'une part, les fonctions linéaires et, d'autre part, les fonctions strictement non linéaires. Les sous–espaces linéaires et non linéaires sont orthogonaux, ce qui permet un calcul efficace des solutions.

La complexité du modèle est contrôlée par seulement deux paramètres, l'un contrôlant la complexité des parties linéaires, l'autre celle des parties non linéaires. Les paramètres qui contrôlent les complexités individuelles de chaque composante sont répartis automatiquement lors de l'estimation des fonctions monovariées. Il reste alors à choisir ces deux hyper–paramètres par des critères de sélection de modèle, tels que ceux que nous avons adaptés en approximant le nombre effectif de paramètres.

Nous avons évalué expérimentalement les performances des modèles additifs parcimonieux. Les résultats de nos expériences concordent et prolongent, en général, ceux déjà obtenus pour la régression linéaire. Sommairement, on peut dire que la sélection pas à pas est plus performante dans les situations moins complexes, où peu de variables ont une influence sur la sortie, tandis que les modèles additifs parcimonieux obtiennent les meilleurs résultats dans les situations plus complexes. Ces derniers sont également plus stables.

Dans notre comparaison des critères analytiques de sélection de modèle, la va-

lidation croisée généralisée s'est révélée être le critère le plus performant. Le coût occasionné par l'estimation des hyper–paramètres par validation croisée généralisée est faible. La performance du modèle sélectionné est proche de celle du meilleur modèle testé.

Quant à la sélection des variables du modèle, la sélection pas à pas identifie correctement les variables non pertinentes, mais elle peut éliminer un nombre élevé de variables pertinentes. Les modèles additifs parcimonieux identifient correctement les variables pertinentes. En revanche, ils éliminent peu de variables non pertinentes et de variables redondantes, lesquelles sont néanmoins très pénalisées.

Le modèle additif parcimonieux est également plus avantageux que la sélection pas à pas à niveau des calculs, lorsque le nombre d'observations et le nombre de variables pertinentes sont modérés ou élevés.

En ce qui concerne l'application à des données réelles, nous constatons l'utilité de ces modèles en tant qu'outil d'analyse exploratoire. En effet, la représentation graphique des résultats permet d'étudier les effets de chaque variable sur la sortie.

Discussion et perspectives

Les perspectives de recherches prolongeant le travail exposé sont variées. Elles concernent premièrement l'optimisation des algorithmes, deuxièmement l'applicabilité de l'outil modèles additifs à un plus grand nombre de problèmes, et troisièmement, l'extension à l'étude de différents types de phénomènes.

Premièrement, l'adaptation de nouveaux algorithmes devrait accélérer considérablement nos calculs. Tout d'abord, nous avons estimé les modèles additifs par backfitting. Cependant des nouveaux algorithmes basés sur la résolution directe d'un système de rang peu élevé (au moyen des P–splines) [Wood, 2000, Ruppert *et al.*, 2003, Wood, 2004] permettrait une résolution plus rapide. Ces algorithmes sont également associés à un algorithme de descente pour le calcul de la validation croisée généralisée. Vu que ce critère obtient les meilleurs résultats parmi les critères de sélection de modèle pour des réponses de type gaussien, il serait intéressant d'optimiser son calcul. En outre, le nouvel algorithme pour le lasso [Efron *et al.*, 2004] peut être directement appliqué à l'estimation des parties linéaires. Il serait pour autant plus intéressant son adaptation à l'estimation des parties non linéaires.

Deuxièmement, notre approche offre des nouvelles perspectives pour la modélisation parcimonieuse des effets (éventuellement) non linéaires de plusieurs variables continues sur une variable réponse. Nous ne prétendons pas que les modèles additifs parcimonieux soient la solution pour étendre l'application des modèles additifs aux problèmes de grande dimension (au sens de $n \times p$ grand). L'application de ces modèles quand le nombre d'observations est très élevé n'est pas approprié, car l'estimation non linéaire demande un nombre de calculs élevé. Toutefois, les méthodes d'estimation et de sélection de modèle que nous avons proposé permettent d'attaquer des problèmes où le nombre des variables explicatives est modéré.

Une étude plus approfondie sur les raisons de la mauvaise performance des critères de sélection de modèle analytiques pour les modèles additifs généralisés parcimo-

nieux serait souhaitable. Ces méthodes sont originairement proposées dans le cadre linéaire et, dans le cas de la validation croisée généralisée, pour les problèmes de type gaussien. Pour étendre ces méthodes au contexte non paramétrique gaussien, des approximations sont considérées. L'extension aux modèles additifs généralisés comporte des approximations non négligeables, ce qui pourrait expliquer cette mauvaise performance. En effet, le contexte non paramétrique avec des réponses de type non gaussien est inéluctablement non linéaire. Des versions de la validation croisée généralisée, approchant une version de la validation croisée plus adaptée aux réponses de type non gaussien ont été proposées [Xiang et Wahba, 1996, Gu et Xiang, 2001, Yuan et Wahba, 2001]. Cependant, leurs performances ne sont pas générales. Ainsi, des résultats correctes ont été obtenus pour les réponses de type binaire (modèle logistique additif), tandis que de mauvais résultats ont été obtenus pour les réponses de type Poisson (modèle de Poisson additif).

L'accélération des calculs, discuté précédemment, offrirait également la possibilité d'utiliser des techniques de rééchantillonage, qui supposent moins d'hypothèses et donc moins d'approximations.

Un autre aspect qui mériterait d'être étudié est l'introduction d'un seuil à partir duquel les variables très pénalisées, mais non nulles, seraient éliminées. Les résultats seraient plus catégoriques et donc plus faciles à interpréter, cependant la méthode pourrait perdre de la stabilité.

Troisièmement, malgré les limitations d'application actuelles, les modèles additifs sont utilisés dans de nombreux domaines, pour leur interprétabilité. En particulier, leur utilisation est largement répandue dans les études des effets de la pollution de l'air sur la santé, au moyen du modèle de Poisson, ainsi que dans les études de survie, au moyen du modèle de Cox. Notre généralisation du lasso au modèle additif logistique n'est pas spécifique de la distribution binomiale, elle peut donc être appliquée à ces modèles, mais l'évaluation pourrait montrer des difficultés similaires pour le contrôle de la complexité. Les versions parcimonieuses correspondantes mériteraient d'être explorées.

Finalement, dans cette optique d'applicabilité, on pourrait envisager l'implémentation des modèles additifs parcimonieux en R, logiciel disposant d'un grand nombre de méthodes d'analyse de données, les modèles additifs inclus. En effet, il serait intéressant de mettre cette méthode complètement automatique à disponibilité des praticiens.

Annexe A

A.1 Quelques rappels sur l'optimisation sous contraintes

On considère le problème de minimisation d'une fonction $f : \Omega \to \mathbb{R}$, Ω ouvert de $\mathbb{R}^n$, en présence de contraintes données par les fonctions $c_E : \Omega \to \mathbb{R}^{m_E}$, et $c_I : \Omega \to \mathbb{R}^{m_I}$, ($m_E$, m_I sont des entiers positifs) [Bonnans *et al.*, 1997, Gill *et al.*, 1981]. La fonction c_E définit des contraintes d'égalité et c_I des contraintes d'inégalité. On cherche donc un point $\boldsymbol{\alpha}_* \in \Omega$ minimisant f sur l'*ensemble admissible* $\Omega^a = \{\boldsymbol{\alpha} \in \Omega : c_E(\boldsymbol{\alpha}) = \mathbf{0}, c_I(\boldsymbol{\alpha}) \leq \mathbf{0}\}$. Les inégalités vectorielles doivent se comprendre composante par composante. Donc $c_I(\boldsymbol{\alpha}) \leq \mathbf{0}$ signifie que toutes les composantes du vecteur $c_I(\boldsymbol{\alpha}) \in \mathbb{R}^{m_I}$ doivent être négatives.

Le problème s'écrit :

$$(\mathrm{P_{EI}}) \quad \begin{cases} \min f(\boldsymbol{\alpha}) \\ c_E(\boldsymbol{\alpha}) = \mathbf{0} \\ c_I(\boldsymbol{\alpha}) \leq \mathbf{0} \\ \boldsymbol{\alpha} \in \Omega. \end{cases} \tag{A.1}$$

On appelle *solution globale* du problème ($\mathrm{P_{EI}}$) un point $\boldsymbol{\alpha}_* \in \Omega$ minimisant f sur l'ensemble admissible Ω^a :

$$f(\boldsymbol{\alpha}_*) \leq f(\boldsymbol{\alpha}), \quad \forall \boldsymbol{\alpha} \in \Omega^a. \tag{A.2}$$

Une *solution locale* de ($\mathrm{P_{EI}}$) est un point $\boldsymbol{\alpha}_*$ admissible, minimisant f localement sur l'ensemble admissible Ω^a :
Il existe $\varepsilon > 0$ tel que

$$f(\boldsymbol{\alpha}_*) \leq f(\boldsymbol{\alpha}), \quad \forall \boldsymbol{\alpha} \in \Omega^a \cap B(\boldsymbol{\alpha}_*, \varepsilon), \tag{A.3}$$

où $B(\boldsymbol{\alpha}_*, \varepsilon)$ est la boule ouverte de centre $\boldsymbol{\alpha}_*$ et rayon ε.

On supposera que les fonctions f, c_E et c_I sont régulières, $\mathcal{C}^2(\Omega)$[1]. Soit $\boldsymbol{\alpha} \in \Omega$. Si $c_i(\boldsymbol{\alpha}) = 0$, on dit que la contrainte i est *active* ou *saturée* en $\boldsymbol{\alpha}$. On note

$$I^0(\boldsymbol{\alpha}) = \{i \in I : c_I(\boldsymbol{\alpha}) = \mathbf{0}\} \tag{A.4}$$

[1]La condition de régularité n'est pas vérifiée par $\|.\|_1$.

l'ensemble d'indices des contraintes d'inégalité *actives* en $\boldsymbol{\alpha} \in \Omega$.

On dira que les contraintes sont *qualifiées* en $\boldsymbol{\alpha}$ si l'une des conditions suivantes est vérifiée :

- Les contraintes d'indices $i \in E \cup I^0$ sont affines dans un voisinage de $\boldsymbol{\alpha}$.
- Les gradients des contraintes d'inégalité actives et des contraintes d'égalité, $\{\nabla c_i(\boldsymbol{\alpha}) : i \in E \cup I^0(\boldsymbol{\alpha})\}$, sont linéairement indépendants.
- Si $\sum_{i \in E \cup I^0(\boldsymbol{\alpha})} \alpha_i \nabla c_i(\boldsymbol{\alpha}) = 0$, avec $\alpha_i \geq 0$ pour $i \in E \cup I^0(\boldsymbol{\alpha})$, alors $\alpha_i = 0$ pour tout $i \in E \cup I^0(\boldsymbol{\alpha})$ (Qualification de Mangasirian–Fromovitz).

A.1.0.1 Conditions d'optimalité du premier ordre

Soit $\boldsymbol{\alpha}_*$ une solution locale de ($\mathrm{P_{EI}}$), alors si les contraintes sont qualifiées en $\boldsymbol{\alpha}_*$, il existe $\boldsymbol{\mu}_* \in \mathbb{R}^{m_E+m_I}$, $\boldsymbol{\mu}_* = ((\boldsymbol{\mu}_*)_E, (\boldsymbol{\mu}_*)_I)^t$, tel que l'on ait les conditions de KKT (Karush, Kuhn et Tucker) suivantes :

$$\text{(KKT)} \quad \begin{cases} (a) & \nabla f(\boldsymbol{\alpha}_*) + \nabla c_E(\boldsymbol{\alpha}_*)^t (\boldsymbol{\mu}_*)_E + \nabla c_I(\boldsymbol{\alpha}_*)^t (\boldsymbol{\mu}_*)_I = \mathbf{0} \\ (b) & c_E(\boldsymbol{\alpha}_*) = \mathbf{0} \\ (c) & c_I(\boldsymbol{\alpha}_*) \leq \mathbf{0} \\ (d) & (\boldsymbol{\mu}_*)_I \geq \mathbf{0} \\ (e) & (\boldsymbol{\mu}_*)_I^t c_I(\boldsymbol{\alpha}_*) = \mathbf{0}. \end{cases} \tag{A.5}$$

L'identité (a) est l'*équation d'optimalité* proprement dite. Cette équation peut encore s'écrire $\nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}_*, \boldsymbol{\mu}_*) = \mathbf{0}$, où $\mathcal{L}$ est le *Lagrangien* associé au problème ($\mathrm{P_{EI}}$) : $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}) = f(\boldsymbol{\alpha}) + (\boldsymbol{\mu})_E^t c_E(\boldsymbol{\alpha}) + (\boldsymbol{\mu})_I^t c_I(\boldsymbol{\alpha})$. Le vecteur $\boldsymbol{\mu}_*$ s'appelle le *multiplicateur de Lagrange*.

On reconnaît l'admissibilité de $\boldsymbol{\alpha}_*$ en (b) et (c).

Les conditions (d) et (e) sont propres aux contraintes d'inégalité. Par (d), on exprime que les multiplicateurs correspondant aux contraintes d'inégalité ont un signe, qui dépend de la forme sous laquelle on formule le problème ($\mathrm{P_{EI}}$) (problème de minimisation, contraintes d'inégalité négatives et signe "+" dans l'équation (a), et donc dans la définition du Lagrangien). L'identité (e) porte le nom de *conditions de complémentarité*. Comme $(\boldsymbol{\mu}_*)_I \geq \mathbf{0}$ et $c_I(\boldsymbol{\alpha}_*) \leq \mathbf{0}$, cela revient à écrire que $(\boldsymbol{\mu}_*)_i c_i(\boldsymbol{\alpha}_*) = 0$, $\forall i \in I$. Autrement dit, les multiplicateurs correspondant aux contraintes inactives sont nuls : $c_i(\boldsymbol{\alpha}_*) < 0 \Rightarrow (\boldsymbol{\mu}_*)_i = 0$. Cela vient du fait que (KKT) exprime la stationnarité de $\boldsymbol{\alpha}_*$ qui est une propriété locale : si $c_i(\boldsymbol{\alpha}_*) < 0$, la contrainte c_i ne doit pas intervenir dans (A.5) car une petite perturbation de cette contrainte ne modifie pas la stationnarité de $\boldsymbol{\alpha}_*$. Dans certains cas, on a l'équivalence $c_i(\boldsymbol{\alpha}_*) < 0 \Leftrightarrow (\boldsymbol{\mu}_*)_i = 0$. On dit alors que l'on a *complémentarité stricte*.

Un couple $(\boldsymbol{\alpha}_*, \boldsymbol{\mu}_*)$ vérifiant (KKT) est appelé solution *primale–duale* de ($\mathrm{P_{EI}}$) et $\boldsymbol{\alpha}_*$ est dit *stationnaire*.

Observons que, sous la deuxième condition de qualification, il y a au plus un multiplicateur $\boldsymbol{\mu}_*$ vérifiant (KKT) pour une solution primale $\boldsymbol{\alpha}_*$ donnée. La condition de Mangasarian–Fromovitz est plus faible que la deuxième condition. Il s'agit d'une sorte de "sous–surjectivité" de la jacobienne des contraintes actives d'inégalité et des

contraintes d'égalité, alors que la deuxième condition exprime la surjectivité[2] de cette même jacobienne.

A.1.0.2 Conditions d'optimalité du second ordre

Soit $S(\boldsymbol{\alpha}_*)$ la surface de $\mathbb{R}^n$ définie par les contraintes d'égalité et la saturation des contraintes d'inégalité actives en $\boldsymbol{\alpha}$. Soit $S^+(\boldsymbol{\alpha}_*)$ le sous–ensemble de $S(\boldsymbol{\alpha}_*)$ tel que le multiplicateur de Lagrange associé aux contraintes d'inégalité actives soit strictement positif. Et soit $T^+(\boldsymbol{\alpha}_*)$ le plan tangent en $\boldsymbol{\alpha}_*$ à la surface $S^+(\boldsymbol{\alpha}_*)$:

$$
\begin{aligned}
S(\boldsymbol{\alpha}_*) &= \{\boldsymbol{\alpha} \in \Omega : c_i(\boldsymbol{\alpha}) = \mathbf{0}, i \in E \cup I^0(\boldsymbol{\alpha})\} \\
S^+(\boldsymbol{\alpha}_*) &= \{\boldsymbol{\alpha} \in \Omega : c_i(\boldsymbol{\alpha}) = \mathbf{0}, i \in E \cup I^0(\boldsymbol{\alpha}), (\boldsymbol{\mu}_*)_{I^0(\boldsymbol{\alpha})} > \mathbf{0}\} \qquad \text{(A.6)} \\
T^+(\boldsymbol{\alpha}_*) &= \{\mathbf{v} \in \mathbb{R}^n : \nabla c_i(\boldsymbol{\alpha}_*)^t \mathbf{v} = 0, i \in E \cup I^0(\boldsymbol{\alpha}_*), (\boldsymbol{\mu}_*)_{I^0(\boldsymbol{\alpha}_*)} > \mathbf{0}\}.
\end{aligned}
$$

Conditions nécessaires

Soit $\boldsymbol{\alpha}_*$ une solution locale de ($\mathrm{P_{EI}}$), alors si les contraintes sont qualifiées en $\boldsymbol{\alpha}_*$, il existe $\boldsymbol{\mu}_* \in \mathbb{R}^{m_E+m_I}$, tel que les conditions (KKT) soient vérifiées et on a

$$
\mathbf{v}^t \nabla^2_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}_*, \boldsymbol{\mu}_*) \mathbf{v} \geq 0, \forall \mathbf{v} \in T^+(\boldsymbol{\alpha}_*). \qquad \text{(A.7)}
$$

Conditions suffisantes

Supposons qu'il existe un multiplicateur $\boldsymbol{\mu}_* \in \mathbb{R}^{m_E+m_I}$ tel que les conditions d'optimalité (KKT) soient vérifiées et que

$$
\mathbf{v}^t \nabla^2_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}_*, \boldsymbol{\mu}_*) \mathbf{v} > 0, \forall \mathbf{v} \in T^+(\boldsymbol{\alpha}_*) \backslash \{\mathbf{0}\}. \qquad \text{(A.8)}
$$

Alors $\boldsymbol{\alpha}_*$ est un minimum local strict du problème ($\mathrm{P_{EI}}$).

A.1.0.3 Méthodes de résolution

Problème avec contraintes linéaires

On s'intéresse tout d'abord à un problème avec contraintes d'égalité linéaires

$$
(\mathrm{P_{EL}}) \quad \left\{ \begin{array}{c} \min f(\boldsymbol{\alpha}) \\ \mathbf{A}\boldsymbol{\alpha} - \boldsymbol{\tau} = \mathbf{0} \\ \boldsymbol{\alpha} \in \Omega. \end{array} \right. \qquad \text{(A.9)}
$$

Les algorithmes les plus efficaces pour résoudre le problème (A.9) génèrent une série d'itérations admissibles. La méthode du gradient projeté et la méthode de quasi–Newton projeté sont souvent utilisées.

[2]Une application $f : U \to V$ est dite surjective si Im(f)=V, où $\mathrm{Im}(f) = \{v \in V | \exists u \in U \text{ tel que } f(u) = v\}$.

On s'intéresse maintenant à un problème avec contraintes d'inégalité linéaires

$$(\mathrm{P_{IL}}) \quad \begin{cases} \min f(\boldsymbol{\alpha}) \\ \mathbf{A}\boldsymbol{\alpha} - \boldsymbol{\tau} \leq \mathbf{0} \\ \boldsymbol{\alpha} \in \Omega. \end{cases} \tag{A.10}$$

Les méthodes avec activation de contraintes se basent sur le fait que seulement les contraintes actives en $\boldsymbol{\alpha}_*$ contribuent dans les conditions d'optimalité. Les conditions d'optimalité du premier ordre (KKT) s'écrivent :

$$\begin{array}{l} \nabla f(\boldsymbol{\alpha}_*) - \bar{\mathbf{A}}^t \boldsymbol{\mu}_* = \mathbf{0} \\ \boldsymbol{\mu}_* \geq \mathbf{0}, \end{array} \tag{A.11}$$

où $\bar{\mathbf{A}}$ indique la matrice dont les lignes sont les lignes de $\mathbf{A}$ correspondantes aux contraintes actives. Si l'ensemble de contraintes actives était connu *a priori*, le problème $(\mathrm{P_{IL}})$ serait équivalent au problème $(\mathrm{P_{EL}})$.

Un cas particulier de (A.10) est le problème des moindres carrés linéaires avec contraintes d'inégalité linéaires. On se trouve face à un problème de programmation quadratique ($f(\boldsymbol{\alpha})$ quadratique et contraintes linéaires) convexe (matrice Hessienne semi–définie positive), avec diverses simplifications dans les calculs.

Problème avec contraintes non–linéaires

Un cas spécial de problème avec contraintes non–linéaires est la programmation convexe. Un problème de programmation convexe est un problème d'optimisation tel que $f(\boldsymbol{\alpha})$ est convexe, les contraintes d'égalité sont linéaires et les contraintes d'inégalité sont concaves (matrice Hessienne semi–définie négative).

Une propriété fondamentale de la programmation convexe est que les minima locaux sont aussi globaux. Ceci, pour les cas où $f(\boldsymbol{\alpha})$ est strictement convexe, implique l'unicité des solutions.

Finalement, il existe un problème dual équivalant au problème primal convexe, donc une stratégie duale peut être appliquée, ainsi qu'une stratégie combinant les deux problèmes.

Comme le cas précédent des contraintes linéaires, un cas particulier de la programmation convexe est le problème des moindres carrés linéaires avec contraintes d'inégalité concaves. Aussi dans ce cas diverses simplifications dans les calculs sont possibles.

La *programmation quadratique successive* (PQS), est un ensemble de techniques fondées sur la méthode de Newton pour résoudre un problème d'optimisation non linéaire (fonction à minimiser et contraintes peuvent toutes être non linéaires). L'idée de base est de linéariser les conditions d'optimalité du problème et d'exprimer le système linéaire résultant sous une forme propice au calcul.

L'intérêt de la linéarisation est de fournir un algorithme à convergence locale rapide. La PQS transforme ainsi un problème d'optimisation non linéaire en une suite de problèmes quadratiques (critère à minimiser quadratique sous contraintes d'égalité et d'inégalité linéaires) plus simples à résoudre. Cette démarche est efficace car on

dispose de bons algorithmes pour résoudre les problèmes quadratiques : méthodes avec activation de contraintes et méthodes de points intérieurs.

Un concept important dans le cadre de la PQS est la *pénalisation exacte*. Celui-ci est central pour forcer la convergence des algorithmes quel que soit l'itéré initial. Aussi, des versions quasi–newtoniennes des algorithmes, dans lesquelles les matrices contenant les dérivées secondes sont remplacées par des matrices mises à jour par des formules adéquates, sont développées pour la PQS.

A.2 Equivalence entre le lasso et AdR

Une démonstration de l'équivalence du lasso et la pénalisation multiple adaptative a été apportée par [Grandvalet, 1998]. Nous détaillons à continuation cette démonstration.

Soit L une fonction de coût différentiable quelconque. Supposons, pour simplifier que les réponses sont centrées. La solution pénalisation multiple adaptative, indiquée ici $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)$, est la valeur minimisant le problème

$$\begin{cases} (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\mu}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\mu}} L(\boldsymbol{\alpha}) + \sum_{j=1}^{p} \mu_j \alpha_j^2, \\ \text{sous contrainte} \qquad \sum_{j=1}^{p} \dfrac{1}{\mu_j} = \dfrac{p}{\mu}, \qquad \mu_j > 0, \end{cases} \tag{A.12}$$

où $\mu \in]0, +\infty[$. Une paramétrisation qui permet d'éviter les solutions divergentes est la suivante

$$\gamma_j = \sqrt{\frac{\mu_j}{\mu}}\alpha_j \qquad \text{and} \qquad c_j = \sqrt{\frac{\mu}{\mu_j}} \qquad \text{pour } j = 1, \ldots, p \tag{A.13}$$

Le problème d'optimisation défini par la pénalisation multiple adaptative est donc

$$\begin{cases} (\widehat{\mathbf{c}}, \widehat{\boldsymbol{\gamma}}) = \arg\min_{\mathbf{c},\boldsymbol{\gamma}} L(\mathbf{c}, \boldsymbol{\gamma}) + \mu \sum_{j=1}^{p} \gamma_j^2, \\ \text{sous contrainte} \qquad \sum_{j=1}^{p} c_j^2 = p, \qquad c_j \geq 0. \end{cases} \tag{A.14}$$

Le Lagrangien associé $\mathcal{L}$ est

$$\mathcal{L}(\mathbf{c}, \boldsymbol{\gamma}) = L(\mathbf{c}, \boldsymbol{\gamma}) + \mu \sum_{j=1}^{p} \gamma_j^2 + \nu \left(\sum_{j=1}^{p} c_j^2 - p \right) - \boldsymbol{\xi}^t \mathbf{c}, \tag{A.15}$$

où ν et $\boldsymbol{\xi}$ sont les multiplicateurs de Lagrange correspondant, respectivement, aux contraintes d'égalité et les contraintes positives sur $\{c_j\}$. Les équations normales pour (A.15) sont alors

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \dfrac{\partial L(\mathbf{c}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} + 2\mu\boldsymbol{\gamma} \\ \dfrac{\partial \mathcal{L}}{\partial \mathbf{c}} = \dfrac{\partial L(\mathbf{c}, \boldsymbol{\gamma})}{\partial \mathbf{c}} + 2\nu\mathbf{c} - \boldsymbol{\xi}. \end{cases} \tag{A.16}$$

Tout d'abord, une relation entre les dérivées partielles de L par rapport à $\mathbf{c}$ et $\boldsymbol{\gamma}$ est précisée. Cette relation découle de la relation $\boldsymbol{\alpha} = \text{diag}(\mathbf{c})\boldsymbol{\gamma}$:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{\gamma}} = \text{diag}(\mathbf{c})\dfrac{\partial L}{\partial \boldsymbol{\alpha}} \\ \dfrac{L}{\partial \mathbf{c}} = \text{diag}(\boldsymbol{\gamma})\dfrac{\partial L}{\partial \boldsymbol{\alpha}}. \end{cases} \tag{A.17}$$

Pour ce système, nous constatons

$$\text{diag}(\boldsymbol{\gamma})\frac{\partial L(\mathbf{c},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \text{diag}(\mathbf{c})\frac{\partial L(\mathbf{c},\boldsymbol{\gamma})}{\partial \mathbf{c}}. \tag{A.18}$$

Cette dernière équation permet de déduire une relation entre $\widehat{c}_j$ et $\widehat{\gamma}_j$, indépendamment de L et des multiplicateurs de Lagrange :

$$\begin{cases} \text{diag}(\widehat{\boldsymbol{\gamma}})\dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \text{diag}(\widehat{\boldsymbol{\gamma}})\dfrac{\partial L(\mathbf{c},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\Big|_{(\widehat{\mathbf{c}},\widehat{\boldsymbol{\gamma}})} + 2\mu\text{diag}(\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\gamma}} \\ \text{diag}(\widehat{\mathbf{c}})\dfrac{\partial \mathcal{L}}{\partial \mathbf{c}} = \text{diag}(\widehat{\mathbf{c}})\dfrac{\partial L(\mathbf{c},\boldsymbol{\gamma})}{\partial \mathbf{c}}\Big|_{(\widehat{\mathbf{c}},\widehat{\boldsymbol{\gamma}})} + 2\nu\text{diag}(\widehat{\mathbf{c}})\widehat{\mathbf{c}} - \text{diag}(\widehat{\mathbf{c}})\boldsymbol{\xi}. \end{cases} \tag{A.19}$$

Etant donné que les multiplicateurs de Lagrange pour les contraintes inactives sont zéro, nous constatons $\text{diag}(\widehat{\mathbf{c}})\boldsymbol{\xi} = 0$. Puisque (A.18) est vérifié pour $(\widehat{\mathbf{c}},\widehat{\boldsymbol{\gamma}})$, et que l'optimalité de $(\widehat{\mathbf{c}},\widehat{\boldsymbol{\gamma}})$ implique $\dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \dfrac{\partial \mathcal{L}}{\partial \mathbf{c}} = \mathbf{0}$, alors, à partir de (A.19), il est déduit

$$\widehat{c}_j^2 = \frac{\mu}{\nu}\widehat{\gamma}_j^2, \quad \forall j. \tag{A.20}$$

La contrainte d'égalité (A.14) sur $\{c_j\}$ implique :

$$\widehat{c}_j = \frac{\sqrt{p}|\widehat{\gamma}_j|}{\sqrt{\sum_{k=1}^{p}\widehat{\gamma}_k^2}}, \quad \forall j. \tag{A.21}$$

Finalement, cette équation permet de donner les conditions d'optimalité en fonction des variables initiales $\widehat{\alpha}_j$. Puisque $|\widehat{\alpha}_j| = \widehat{c}_j|\widehat{\gamma}_j|$, il est vérifié

$$|\widehat{\alpha}_j| = \frac{\sqrt{p}\widehat{\gamma}_j^2}{\sqrt{\sum_{k=1}^{p}\widehat{\gamma}_k^2}} \Rightarrow \frac{|\widehat{\alpha}_j|}{\sum_{k=1}^{p}|\widehat{\alpha}_k|} = \frac{\widehat{\gamma}_j^2}{\sum_{k=1}^{p}\widehat{\gamma}_k^2} \Leftrightarrow \widehat{c}_j^2 = \frac{p|\widehat{\alpha}_j|}{\sum_{k=1}^{p}|\widehat{\alpha}_k|}. \tag{A.22}$$

Cette valeur de $\widehat{c}_j$ est maintenant remplacée dans la première du système (A.16) évalué en $(\widehat{\mathbf{c}},\widehat{\boldsymbol{\gamma}})$, en utilisant la première équation du système (A.17) :

$$\widehat{c}_j = \frac{\partial L}{\partial \alpha_j}\Big|_{\widehat{\alpha}_j} + 2\mu\widehat{\gamma}_j = 0, \quad \forall j. \tag{A.23}$$

Par conséquent, soit $\widehat{c}_j = \widehat{\gamma}_j = \widehat{\alpha}_j = 0$, soit $\left.\dfrac{\partial L}{\partial \alpha_j}\right|_{\widehat{\alpha}_j} + 2\mu\dfrac{\widehat{\gamma}_j}{\widehat{c}_j} = 0$. A partir de (A.22) et en utilisant $\boldsymbol{\alpha}$, $\widehat{\gamma}_j/\widehat{c}_j$ peut être ré–écrit de la façon suivante :

$$\begin{aligned} \frac{\widehat{\gamma}_j}{\widehat{c}_j} &= \widehat{\gamma}_j \widehat{c}_j \frac{1}{\widehat{c}_j^2} \\ &= \widehat{\alpha}_j \frac{\sum_{k=1}^{p} |\widehat{\alpha}_k|}{p|\widehat{\alpha}_j|} \\ &= \frac{1}{p}\text{sign}(\widehat{\alpha}_j) \sum_{k=1}^{p} |\widehat{\alpha}_k|. \end{aligned} \tag{A.24}$$

Les conditions d'optimalité sont ainsi

$$\begin{cases} \left.\dfrac{\partial L}{\partial \alpha_j}\right|_{\widehat{\alpha}_j} + 2\dfrac{\mu}{p}\text{sign}(\widehat{\alpha}_j) \displaystyle\sum_{k=1}^{p} |\widehat{\alpha}_k| = 0, & \forall j, \\ \text{ou} \quad \widehat{\alpha}_j = 0, \end{cases} \tag{A.25}$$

qui sont en fait les équations normales de

$$L(\boldsymbol{\alpha}) + \frac{\mu}{p}\left(\sum_{k=1}^{p} |\alpha_k| \right)^2, \tag{A.26}$$

ce qui conclue la démonstration.

A.3 Relation entre les définitions des ddl

Considérons la définition des degrés de liberté proposée par [Fu, 1998] :

$$\text{ddl}_1 = \text{tr}\left[\mathbf{X} \left(\mathbf{X}^t\mathbf{X} + \mu \mathbf{A}^- \right)^{-1} \mathbf{X}^t \right] - p_0, \tag{A.27}$$

où p_0 est le nombre de coefficients estimés nuls, ainsi que la modification de cette définition que nous proposons :

$$\text{ddl}_2 = \text{tr}\left[\mathbf{X}_\sigma \left(\mathbf{X}_\sigma^t\mathbf{X}_\sigma + \mu \mathbf{A}_\sigma^{-1} \right)^{-1} \mathbf{X}_\sigma^t \right], \tag{A.28}$$

où $\sigma = \{j | \alpha_j^{\text{L}} \neq 0\}$. L'objectif est de montrer que cette dernière définition induit une prédiction plus conservatrice, c'est à dire $\text{ddl}_1 \leq \text{ddl}_2$. Pour cela, nous utilisons le lemme de l'inversion de matrices par blocs :

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{E} & -\mathbf{EBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CD}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CEBD}^{-1} \end{pmatrix}, \tag{A.29}$$

où $\mathbf{E} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1}$.

Considérons, sans perte de généralité, que les colonnes de la matrice des données sont ordonnées de telle sorte que $\mathbf{X} = [\mathbf{X}_\sigma \mathbf{X}_{\bar{\sigma}}]$, avec $\bar{\sigma}$ le complémentaire de σ. Supposons que la matrice $\mathbf{X}^t\mathbf{X}$ (et donc la matrice $\mathbf{X}_\sigma^t\mathbf{X}_\sigma$) est de rang plein. Considérons les

matrices $\mathbf{X}^t\mathbf{X}$ et $\mathbf{A}^-\mathbf{M}$ en termes de la décomposition par blocs, où $\mathbf{A} = \mathrm{diag}(|\widehat{\alpha}_j^{\mathrm{L}}|)$, $\mathbf{A}^-$ la pseudo–inverse, et $\mathbf{M}$ matrice carré de rang plein quelconque :

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} \mathbf{X}_\sigma^t\mathbf{X}_\sigma & \mathbf{X}_\sigma^t\mathbf{X}_{\bar\sigma} \\ \mathbf{X}_{\bar\sigma}^t\mathbf{X}_\sigma & \mathbf{X}_{\bar\sigma}^t\mathbf{X}_{\bar\sigma} \end{pmatrix}, \qquad \mathbf{A}^-\mathbf{M} = \begin{pmatrix} \mathbf{A}_\sigma^{-1}\mathbf{M}_{\sigma\sigma} & \mathbf{A}_\sigma^{-1}\mathbf{M}_{\sigma\bar\sigma} \\ 0 & 0 \end{pmatrix}. \tag{A.30}$$

Les $p-p_0$ valeurs propres de $\mathbf{A}_\sigma^{-1}\mathbf{M}_{\sigma\sigma}$ sont valeurs propres de $\mathbf{A}^-\mathbf{M}$, qui a également p_0 valeurs propres nulles. Les degrés de liberté s'écrivent alors,

$$\mathrm{ddl}_1 = \mathrm{tr}\left[\left(\mathbf{I}_p + \mu\mathbf{A}^-(\mathbf{X}^t\mathbf{X})^{-1}\right)^{-1}\right] - p_0 = \mathrm{tr}\left[\left(\mathbf{I}_{p-p_0} + \mu\mathbf{A}_\sigma^{-1}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{\sigma\sigma}\right)^{-1}\right], \tag{A.31}$$

$$\mathrm{ddl}_2 = \mathrm{tr}\left[\left(\mathbf{I}_{p-p_0} + \mu\mathbf{A}_\sigma^{-1}\left(\mathbf{X}_\sigma^t\mathbf{X}_\sigma\right)^{-1}\right)^{-1}\right]. \tag{A.32}$$

Par application du lemme de l'inversion de matrices par blocs au bloc $\sigma\sigma$ de $(\mathbf{X}^t\mathbf{X})^{-1}$, on obtient :

$$\begin{aligned}
\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{\sigma\sigma} &= (\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1} + \\
&\quad (\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t\mathbf{X}_{\bar\sigma}(\mathbf{X}_{\bar\sigma}^t\mathbf{X}_{\bar\sigma} - \mathbf{X}_{\bar\sigma}^t\mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t\mathbf{X}_{\bar\sigma})^{-1}\mathbf{X}_{\bar\sigma}^t\mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1} \\
&= (\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1} + \\
&\quad (\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t\mathbf{X}_{\bar\sigma}(\mathbf{X}_{\bar\sigma}^t(\mathbf{I}_n - \mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t)\mathbf{X}_{\bar\sigma})^{-1}\mathbf{X}_{\bar\sigma}^t\mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1} \\
&= (\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1} + \\
&\quad \mathbf{Z}(\mathbf{U}[\mathbf{I}_n - \mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t]\mathbf{U}^t)^{-1}\mathbf{Z}^t,
\end{aligned} \tag{A.33}$$

où $\mathbf{Z} = (\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t\mathbf{X}_{\bar\sigma}$ et $\mathbf{U} = \mathbf{X}_{\bar\sigma}^t$.

La matrice $\mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t$ a $p-p_0$ valeurs propres égales à 1 et $n-(p-p_0)$ valeurs propres égales à 0. Les valeurs propres de $\mathbf{Z}(\mathbf{U}[\mathbf{I}_n - \mathbf{X}_\sigma(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}\mathbf{X}_\sigma^t]\mathbf{U}^t)^{-1}\mathbf{Z}^t$ sont donc positives, ce qui implique que l'ensemble des valeurs propres de $\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{\sigma\sigma}$ sont $\geq$ à celles de $(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}$. Comme $\mathbf{A}_\sigma^{-1}$ est diagonale ≥ 0, les valeurs propres de $\mathbf{A}_\sigma^{-1}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{\sigma\sigma}$ sont $\geq$ aux valeurs propres de $\mathbf{A}_\sigma^{-1}(\mathbf{X}_\sigma^t\mathbf{X}_\sigma)^{-1}$. Par conséquent, les valeurs propres de $\mathbf{I}_{p-p_0} + \mu\mathbf{A}_\sigma^{-1}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{\sigma\sigma}$ sont $\geq$ aux valeurs propres de $\mathbf{I}_{p-p_0} + \mu\mathbf{A}_\sigma^{-1}\left(\mathbf{X}_\sigma^t\mathbf{X}_\sigma\right)^{-1}$. Et, prenant les inverses, $\mathrm{ddl}_1 \leq \mathrm{ddl}_2$.

Bibliographie

[Andrews, 1991] Andrews, D. W. K. (1991). Asymptotic optimality of generalized cl, cross–validation, and generalized cross–validation in regression with heteroskedastic errors. *J. Econ.*, 47(2–3):359–377.

[Ansley et Kohn, 1994] Ansley, C. F. et Kohn, R. (1994). Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society Series A*, 57:316–329.

[Avalos *et al.*, 2003] Avalos, M., Grandvalet, Y. et Ambroise, C. (2003). Regularization methods for additive models. *In* Berthold, M. R., Lenz, H. J., Bradley, E., Kruse, R. et Borgelt, C., éditeurs : *5th International Symposium on Intelligent Data Analysis.*, pages 509–520. Springer. LNCS.

[Avalos *et al.*, 2004a] Avalos, M., Grandvalet, Y. et Ambroise, C. (2004a). Discrimination par modèles additifs parcimonieux. *In* Liquière, M. et Sebban, M., éditeurs : *Conférence d'Apprentissage CAp 2004*, pages 17–32.

[Avalos *et al.*, 2004b] Avalos, M., Grandvalet, Y. et Ambroise, C. (2004b). Discrimination par modèles additifs parcimonieux. *Revue d'Intelligence Artificielle. Numéro spécial sur l'apprentissage (meilleurs articles de la conférence CAp 2004). Accepté.*

[Avalos *et al.*, 2004c] Avalos, M., Grandvalet, Y. et Ambroise, C. (2004c). Généralisation du lasso aux modèles additifs. *In* Berlinet, A., éditeur : *XXXVIèmes Journées de Statistique.*

[Avalos *et al.*, 2004d] Avalos, M., Grandvalet, Y. et Ambroise, C. (2004d). Penalized additive logistic regression for cardiovascular risk prediction. *In* Auget, J. L., Balakrishnan, N., Mesbah, M. et Molenberghs, G., éditeurs : *International Conference on Statistics in Health Sciences*, pages 301–303.

[Azzalini et Bowman, 1993] Azzalini, A. et Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, B*, 55(2):549–557.

[Bacchetti et Quale, 2002] Bacchetti, P. et Quale, C. (2002). Generalized additive models with interval–censored data and time–varying covariates : application to human immunodeficiency virus infection in hemophiliacs. *Biometrics*, 58(2):443–447.

[Bakin, 1999] Bakin, S. (1999). *Adaptive Regression and Model Selection in Data Mining Problems.* Thèse de doctorat, School of Mathematical Sciences, The Australian National University, Canberra.

[Beck et Jackman, 1998] Beck, N. et Jackman, S. (1998). Beyond linearity by default : Generalized additive models. *American Journal of Political Science*, 42: 596–627.

[Bellman, 1961] Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press.

[Bi *et al.*, 2003] Bi, J., Bennett, K. P., Embrechts, M., Breneman, K. M. et Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 3:1229–1243.

[Bonnans *et al.*, 1997] Bonnans, J. F., Gilbert, J. C., Lemaréchal, C. et Sagastizábal, C. (1997). *Optimisation Numérique. Aspects Théoriques et Pratiques*, volume 27 de *Mathématiques et Applications*. Springer, Paris.

[Boukari et Grandvalet, 1998] Boukari, H. et Grandvalet, Y. (1998). Pénalisation multiple adaptative. *In 13èmes Journées Francophones sur l'Apprentissage, Arras*, pages 186–197. Hermès.

[Bouyer *et al.*, 1995] Bouyer, J., Hémon, D., Cordier, S., Derriennic, F., Stücker, I., Stengel, B. et Clavel, J. (1995). *Epidémiologie. Principes et Méthodes Quantitatives*. Les Editions INSERM, Paris.

[Bowman et Azzalini, 1997] Bowman, A. W. et Azzalini, E. (1997). *Applied Smoothing Techniques for Data Analysis*, volume 18 de *Oxford Statistical Science Series*. Oxford.

[Bratko, 1997] Bratko, I. (1997). Machine learning : between accuracy and interpretability. *In* Della Riccia, G. e. a., éditeur : *Learning, networks and statistics. ISSEK'96 workshop*, CISM Courses Lect.382, pages 163–177. Springer.

[Breiman, 1993] Breiman, L. (1993). Fitting additive models to regression data. diagnostics and alternative views. *Comput. Stat. Data Anal.*, 15(1):13–46.

[Breiman, 1995] Breiman, L. (1995). Better subset selection using the non–negative garrote. *Technometrics*, 3:373–384.

[Breiman, 1996] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383.

[Breiman et Peters, 1992] Breiman, L. et Peters, S. (1992). Comparing automatic smoothers (a public service enterprise). *Int. Stat. Rev.*, 60(3):271–290.

[Brumback *et al.*, 1999] Brumback, B. A., D., R. et P., W. M. (1999). Comment on "Variable selection and function estimation in additive nonparametric regression using a data–based prior" by Shively, T. S. and Khon, R. and Wood, S. *Journal of the American Statistical Association*, 94(447):794–797.

[Buja *et al.*, 1989] Buja, A., Hastie, T. J. et J., T. R. (1989). Linear smoothers and additive models. *Annals of Statistics*, 17:453–510.

[Cantoni et Hastie, 2002] Cantoni, E. et Hastie, T. J. (2002). Degrees of freedom tests for smoothing splines. *Biometrika*, 89:251–263.

[Carroll *et al.*, 1997] Carroll, R. J., Fan, J., Gijbels, I. et Wand, M. P. (1997). Generalized partially linear single–index models. *Journal of the American Statistical Association*, 92:477–489.

[Chambers et Hastie, 1993] Chambers, J. M. et Hastie, T. J. (1993). *Statistical Models in S*. Computer Science Series. Chapman & Hall, London.

[Chen *et al.*, 1996] Chen, R., Härdle, W., Linton, O. B. et Severance-Lossin, E. (1996). Nonparametric estimation of additive separable regression models. *In* Härdle, W. et Schimek, M. G., éditeurs : *Statistical Theory and Computational Aspects of Smoothing : Proceedings of the COMPSTAT'94 Satellite Meeting*, Contributions to Statistics, pages 247–265, Heidelberg. Physica–Verlag.

[Chen *et al.*, 1995a] Chen, R., Liu, J. S. et Tsay, R. S. (1995a). Additivity tests for nonlinear autoregressions. *Biometrika*, 82:369–383.

[Chen *et al.*, 1995b] Chen, S., Donoho, D. et Saunders, M. (1995b). Atomic decomposition by basis pursuit. Rapport technique 479, Department of Statistics, Stanford University.

[Chen, 1993] Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *J. R. Statist. Soc. B*, 55(2):473–491.

[Craven et Wahba, 1979] Craven, P. et Wahba, G. (1979). Smoothing noisy data with spline functions : estimating the correct degree of smoothing by the method of generalized cross–validation. *Numerische Mathematik*, 31:377–403.

[de Boor, 2001] de Boor, C. (2001). *A Practical Guide to Splines. Revised Edition*, volume 27 de *Applied Mathematical Sciences*. Springer, New York.

[Dominici *et al.*, 2002] Dominici, F., McDermott, A., Zeger, S. L. et Samet, J. M. (2002). On the use of generalized additive models in time–series studies of air pollution and health. *American Journal of Epidemiology*, 156(3):193–203.

[Donnell *et al.*, 1994] Donnell, D. J., Buja, A. et Stuetzle, W. (1994). Analysis of additive dependencies and concurvities using smallest additive principal components. *Annals of Statistics*, 22(4):1635–1673.

[Efron *et al.*, 2004] Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

[Efron et Tibshirani, 1993] Efron, B. et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, volume 57 de *Monographs on Statistics and Applied Probability*. Chapman & Hall.

[Efron et Tibshirani, 1995] Efron, B. et Tibshirani, R. J. (1995). Cross-validation and the bootstrap : Estimating the error rate of a prediction rule. Rapport technique 477, Stanford University, Stanford, CA.

[Eilers et Marx, 1996] Eilers, P. H. C. et Marx, B. D. (1996). Flexible smoothing with B–splines and penalties (with discussion). *Statistical Science*, 11:89–121.

[Eubank *et al.*, 1995] Eubank, R. L., Hart, J. D., Simpson, D. G. et Stefanski, L. A. (1995). Testing for additivity in nonparametric regression. *Annals of Statistics*, 23:1896–1920.

[Fahrmeir et Tutz, 2001] Fahrmeir, L. et Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition*. Springer Series in Statistics. Springer, New York.

[Fan, 2003] Fan, J. (2003). *Nonlinear times series*. Springer Series in Statistics. Springer, New York.

[Fan et Gijbels, 2000] Fan, J. et Gijbels, I. (2000). Local polynomial fitting. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 229–276. John Wiley & sons.

[Fan *et al.*, 1998] Fan, J., Härdle, W. et Mammen, E. (1998). Direct estimation of low–dimensional components in additive models. *Annals of Statistics*, 26(3):943–971.

[Fan et Li, 2001] Fan, J. et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[Figueiras *et al.*, 2003] Figueiras, J., Roca-Pardiñas, J. et Cadarso-Suárez, C. (2003). Avoiding the effect of concurvity in generalized additive models in time–series studies of air pollution. *In The ISI International Conference on Environmental Statistics and Health*, Santiago de Compostela.

[Frank et Friedman, 1993] Frank, I. E. et Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148.

[Friedman, 1991] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19:1–141.

[Friedman, 1997] Friedman, J. H. (1997). On bias, variance, 0/1 loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77.

[Friedman et Stuetzle, 1981] Friedman, J. H. et Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

[Fu, 1998] Fu, W. J. (1998). Penalized regression : the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

[Fu, 2003] Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, 59(1):126–132.

[Ghosh *et al.*, 2003] Ghosh, D., Barette, T. R., Rhodes, D. et Chinnaiyan, A. (2003). Statistical issues and methods for meta–analysis of microarray data : a case study in prostate cancer. *Funct. Integr. Genomics*, 3:180–188.

[Gill *et al.*, 1981] Gill, P., Murray, W. et Wright, M. H. (1981). *Practical Optimization*. Academic Press, New York.

[Girard, 1991] Girard, D. (1991). Asymptotic optimality of the fast randomized versions of GCV and CL in ridge regression and regularization. *Annals of Statistics*, 19(4):1950–1963.

[Golub *et al.*, 1979] Golub, G., Health, M. et Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–224.

[Grandvalet, 1998] Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. *In* Niklasson, L., Bodén, M. et Ziemske, T., éditeurs : *ICANN'98*, volume 1 de *Perspectives in Neural Computing*, pages 201–206. Springer.

[Grandvalet et Canu, 1998] Grandvalet, Y. et Canu, S. (1998). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. *In* Kearns, M., Solla, S. et Cohn, D., éditeurs : *Advances in Neural Information Processing Systems 11*, pages 445–451. MIT Press.

[Green et Silverman, 1994] Green, P. J. et Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, volume 58 de *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

[Gruber, 1998] Gruber, M. H. J. (1998). *Improving efficiency by shrinkage*, volume 156 de *Statistics : Textbooks and Monographs*. Marcel Dekker, Inc., New York.

[Gu, 1992a] Gu, C. (1992a). Cross–validating non–Gaussian data. *J. Comput. Graph. Stats.*, 1:169–179.

[Gu, 1992b] Gu, C. (1992b). Diagnostics for nonparametric regression models with additive terms. *Journal of the American Statistical Association*, 87(420):1051–1058.

[Gu, 1998] Gu, C. (1998). Model indexing and smoothing parameter selection in nonparametric function estimation. *Statistica Sinica*, 8(3):607–646.

[Gu, 2000] Gu, C. (2000). Multivariate spline regression. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 229–356. John Wiley & sons.

[Gu, 2002] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer, New York.

[Gu et Kim, 2002] Gu, C. et Kim, Y.-J. (2002). Penalized likelihood regression : general formulation and efficient approximation. *Canadian Journal of Statistics*, 30(4):619–628.

[Gu et Wahba, 1991] Gu, C. et Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the newton method. *SIAM J. Sci. Statist. Comput.*, 12:383–398.

[Gu et Xiang, 2001] Gu, C. et Xiang, D. (2001). Cross–validating non–Gaussian data : Generalized approximate cross–validation revisited. *Journal of Computational and Graphical Statistics*, 10:581–591.

[Gueyffier *et al.*, 1995] Gueyffier, F., Boutitie, F., Boissel, J. P., Coope, J., Cutler, J., Ekbom, T., Fagard, R., Friedman, L., Perry, H. M. et Pocock, S. (1995). INDANA : a meta–analysis on individual patient data in hypertension. protocol and preliminary results. *Therapie*, 50(4):353–362.

[Guyon et Elisseeff, 2003] Guyon, I. et Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 3:1157–1182.

[Gyorfi *et al.*, 2002] Gyorfi, L., Kohler, M., Krzyzak, A. et Walk, H. (2002). *A distribution–free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York.

[Härdle, 1990] Härdle, W. (1990). *Applied Nonparametric Regression*, volume 19 de *Economic Society Monographs*. Cambridge University Press, New York.

[Härdle et Hall, 1993] Härdle, W. et Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, 47:157–178.

[Härdle *et al.*, 2004a] Härdle, W., Huet, S., Mammen, E. et Sperlich, S. (2004a). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20:265–300.

[Härdle et Korostelev, 1996] Härdle, W. et Korostelev, A. (1996). Search for significant variables in nonparametric additive regression. *Biometrika*, 83(3):541–549.

[Härdle et Muller, 2000] Härdle, W. et Muller, M. (2000). Multivariate and semiparametric kernel regression. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 357–392. John Wiley & sons.

[Härdle *et al.*, 2004b] Härdle, W., Müller, M., Sperlich, S. et Werwatz, A. (2004b). *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, New York.

[Hart, 1997] Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*, volume 43 de *Springer Series in Statistics*. Springer-Verlag.

[Hastie, 1996] Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society B*, 58:379–396.

[Hastie et Tibshirani, 1986] Hastie, T. et Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, 1:297–318.

[Hastie et Tibshirani, 1995] Hastie, T. et Tibshirani, R. (1995). Generalized additive models for medical research. *Statistical Methods in Medical Research*, 4:187–196.

[Hastie et Tibshirani, 2000] Hastie, T. et Tibshirani, R. (2000). Bayesian backfitting. (with comments and a rejoinder). *Statistical Science*, 15(3):196–223.

[Hastie et Tibshirani, 1990] Hastie, T. J. et Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43 de *Monographs on Statistics and Applied Probability*. Chapman & Hall.

[Hastie *et al.*, 2001] Hastie, T. J., Tibshirani, R. J. et Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York.

[Herrmann, 2000] Herrmann, E. (2000). Variance estimation and bandwidth selection for kernel regression. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 71–107. John Wiley & sons.

[Hoerl et Kennard, 1970] Hoerl, A. et Kennard, R. (1970). Ridge regression : biased estimation for non-orthogonal problems. *Technometrics*, 8:27–51.

[Huang, 2003] Huang, F. (2003). Prediction error property of the lasso estimator and its generalization. *Australian & New Zealand Journal of Statistics*, 45(2):217–228.

[Huang, 1999] Huang, J. (1999). Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics*, 26:242–272.

[Hurvich *et al.*, 1998] Hurvich, C. M., Simonoff, J. S. et Tsai, C. L. (1998). Smoothing parameter selection in non parametric regression using an improved akaike information criteria. *Journal of the Royal Statistical Society, B*, 60(2):271–293.

[Ishwaran, 2004] Ishwaran, H. (2004). Comments on "least angle regression" by Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Annals of Statistics*, 32(2): 452–458.

[Kauermann et Opsomer, 2004] Kauermann, G. et Opsomer, J. D. (2004). Generalized cross–validation for bandwidth selection of backfitting estimators in generalized additive models. *Journal of Computational and Graphical Statistics*, 13:66–89.

[Kim *et al.*, 1999] Kim, W., Linton, O. B. et Hengartner, N. W. (1999). A computational efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, 8:278–297.

[Kim et Gu, 2004] Kim, Y.-J. et Gu, C. (2004). Smoothing spline Gaussian regression : more scalable computation via efficient approximation. *Journal of the Royal Statistical Society, B*, 66:337–356.

[Klinger, 2001] Klinger, A. (2001). Inference in high dimensional generalized linear models based on soft thresholding. *Journal of the Royal Statistical Society, B*, 63(2):377–392.

[Klinke et Grassmann, 2000] Klinke, S. et Grassmann, J. (2000). Projection pursuit regression. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 471–496. John Wiley & sons.

[Knight, 2004] Knight, K. (2004). Comments on "least angle regression" by Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Annals of Statistics*, 32(2):458–460.

[Knight et Fu, 2000] Knight, K. et Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378.

[Kohn *et al.*, 2000] Kohn, R., Schimek, M. G. et Smith, M. (2000). Spline and kernel regression for dependent data. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 135–158. John Wiley & sons.

[Li, 1986] Li, K. (1986). Asymptotic optimality of cl and generalized cross–validation in ridge regression with application to spline smoothing. *Annals of Statistics*, 14:1101–1112.

[Li *et al.*, 2004] Li, L., Huang, J., Sun, S., Jianzhao, S., Unverzagt, F. W., Gao, S., Hendrie, H. H., Hall, K. et Hui, S. L. (2004). Selecting pre–screening items for early intervention trials of dementia – a case study. *Statistics in Medicine*, 23:271–283.

[Lin et Zhang, 1999] Lin, X. et Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, B*, 61(2):381–400.

[Lin *et al.*, 2000] Lin, Y., Wahba, G., Zhang, H. et Yoonkyung, L. (2000). Statistical properties and adaptive tuning of support vector machines. Rapport technique 1022, University of Winconsin.

[Lin et Zhang, 2003] Lin, Y. et Zhang, H. H. (2003). Component selection and smoothing in smoothing spline analysis of variance models. Rapport technique 1072r, University of Winconsin – Madison and North Carolina State University.

[Linde, 2000] Linde, A. (2000). Variance estimation and smoothing–parameter selection for spline regression. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 19–41. John Wiley & sons.

[Linton, 1997] Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, 84:469–473.

[Linton et Härdle, 1996] Linton, O. B. et Härdle, W. (1996). Estimation for additive regression models with known links. *Biometrika*, 83:529–540.

[Linton et Nielsen, 1995] Linton, O. B. et Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–100.

[Linton et Nielsen, 2000] Linton, O. B. et Nielsen, J. P. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, 16:502–523.

[Loubes et Massart, 2004] Loubes, J. M. et Massart, P. (2004). Comments on "least angle regression" by Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Annals of Statistics*, 32(2):460–465.

[Mammen, 2000] Mammen, E. (2000). Resampling methods for nonparametric regression. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 425–450. John Wiley & sons.

[Mammen *et al.*, 1999] Mammen, E., Linton, O. et Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27:1443–1490.

[Martinussen et Scheike, 1999] Martinussen, T. et Scheike, T. H. (1999). A semiparametric additive regression model for longitudinal data. *Biometrika*, 86(3):691–702.

[Marx et Eilers, 1998] Marx, B. D. et Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28:193–209.

[Miller, 1990] Miller, A. J. (1990). *Subset selection in regression*, volume 40 de *Monographs on Statistics and Applied Probability*. Chapman & Hall.

[Nielsen et Linton, 1998] Nielsen, J. P. et Linton, O. B. (1998). An optimization interpretation of integration and back–fitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, Series B*, 60:217–222.

[Opsomer, 2000] Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of the American Statistical Association*, 93:605–619.

[Opsomer et Ruppert, 1997] Opsomer, J. D. et Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25:186–211.

[Opsomer et Ruppert, 1998] Opsomer, J. D. et Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *J. Multivariate Analysis*, 73:166–179.

[Osborne *et al.*, 2000a] Osborne, M. R., Presnell, B. et Turlach, B. A. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–404.

[Osborne *et al.*, 2000b] Osborne, M. R., Presnell, B. et Turlach, B. A. (2000b). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337.

[Perkins *et al.*, 2003] Perkins, S., Lacker, K. et Theiler, J. (2003). Grafting : Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 3:1333–1356.

[Popper, 1961] Popper, K. (1961). *The logic of scientific discovery*. Sciences Editions, New York.

[Ramsay *et al.*, 2003a] Ramsay, T. O., Burnett, R. T. et Krewski, D. (2003a). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, 14(1):18–23.

[Ramsay *et al.*, 2003b] Ramsay, T. O., Burnett, R. T. et Krewski, D. (2003b). Exploring bias in a generalized additive model for spatial air pollution data. *Environmental Health Perspectives*, 111(10):1283–1288.

[Rosset et Zhu, 2003] Rosset, S. et Zhu, J. (2003). Corrected proof of the result of "a prediction error property of the lasso estimator and its generalization" by Huang, f. Rapport technique, Stanford University, Stanford, CA.

[Rosset et Zhu, 2004] Rosset, S. et Zhu, J. (2004). Comments on "least angle regression" by Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Annals of Statistics*, 32(2):469–475.

[Roth, 2001] Roth, V. (2001). Sparse kernel regressors. *In* Dorfner, G., Bischof, H. et Hornik, K., éditeurs : *Artificial Neural Networks–ICANN 2001*, pages 339–346. Springer, LNCS 2130.

[Ruppert, 2002] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

[Ruppert *et al.*, 1995] Ruppert, D., Sheather, S. J. et Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

[Ruppert *et al.*, 2003] Ruppert, D., Wand, M. P. et Carroll, R. J. (2003). *Semiparametric regression*, volume 12 de *Cambridge Series on Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[Sakamoto *et al.*, 1986] Sakamoto, Y., Ishiguro, M. et Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Mathematics and Its Applications. Japanese Series. KTK Scientific Publishers, Reidel Publishing Company, Tokio.

[Sardy et Tseng, 2004] Sardy, S. et Tseng, P. (2004). Amlet and gamlet : Automatic nonlinear fitting of additive models and generalized additive models with wavelets. *Journal of Computational and Graphical Statistics (To appear in)*.

[Schimek, 1996] Schimek, M. G. (1996). An iterative projection algorithm and some simulation results. *In* Prat, A., éditeur : *Proceedings of the COMPSTAT'96 Satellite Meeting*, Contributions to Statistics, Heidelberg. Physica–Verlag.

[Schimek, 2000] Schimek, M. G. (2000). Gam spline algorithms : a direct comparison. *In* Bethlehem, J. et van der Heijden, P., éditeurs : *Proceedings of the COMPSTAT'00 Satellite Meeting*, Contributions to Statistics, Heidelberg. Physica–Verlag.

[Schimek et Turlach, 2000] Schimek, M. G. et Turlach, B. A. (2000). Additive and generalized additive models. *In* Schimek, M. G., éditeur : *Smoothing and Regression : Approaches, Computation and Application*, Wiley Series in Probability and Mathematical Statistics, pages 229–276. John Wiley & sons.

[Segal *et al.*, 2003] Segal, M. R., Dahlquist, K. D. et Conklin, B. R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980.

[Sen et M., 1990] Sen, A. et M., S. (1990). *Regression Analysis. Theory, Methods and Applications*. Springer Texts in Statistics. Springer–Verlag, New York.

[Shively *et al.*, 1999] Shively, T. S., Khon, R. et Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data–based prior. *Journal of the American Statistical Association*, 94(447):777–806.

[Shively *et al.*, 1994] Shively, T. S., Kohn, R. et Ansley, C. F. (1994). Testing of linearity in a semiparametric regression model. *Journal of Econometrics*, 64(1–2):77–96.

[Silverman, 1984] Silverman, B. (1984). Spline smoothing : the equivalent variable kernel method. *Annals of Statistics*, 12(3):898–916.

[Simonoff, 1996] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer, New York.

[Smith et Kohn, 1996] Smith, M. et Kohn, R. (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343.

[Smith *et al.*, 2000] Smith, M., Kohn, R. et Yau, P. (2000). Nonparametric bayesian bivariate surface estimation. *In* Schimek, M. G., éditeur : *Smoothing and*

Regression : Approaches, Computation and Application, Wiley Series in Probability and Mathematical Statistics, pages 545–580. John Wiley & sons.

[Sperlich, 2003] Sperlich, S. (2003). About sense and nonsense of non– and semiparamentric analysis in applied econometrics. *In The 2003 Semiparametrics Conference*, Berlin.

[Sperlich *et al.*, 1999] Sperlich, S., Linton, O. B. et Härdle, W. (1999). Integration and backfitting methods in additive models – finite samples properties and comparison. *Test*, 8(2):419–458.

[Sperlich *et al.*, 2002] Sperlich, S., Tjø stheim, D. et Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, 18(2):197–251.

[Steyerberg *et al.*, 2000] Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E. J. et Habbema, J. D. F. (2000). Pronostic modelling with logistic regression analysis : a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, 19:1059–1079.

[Stine, 2004] Stine, R. A. (2004). Comments on "least angle regression" by Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Annals of Statistics*, 32(2):475–481.

[Stone, 1982] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053.

[Stone, 1985] Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705.

[Stone, 1986] Stone, C. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics*, 14:590–606.

[Tibshirani, 1996] Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58(1):267–288.

[Tibshirani, 1997] Tibshirani, R. J. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

[Tibshirani et Knight, 1997] Tibshirani, R. J. et Knight, K. (1997). The covariance inflation criterion for adaptive model selection. Rapport technique, University of Toronto.

[Tikhonov et Arsenin, 1977] Tikhonov, A. N. et Arsenin, V. Y. (1977). *Solution of ill-posed problems*. W. H. Wilson, Washington, D. C.

[Turlach *et al.*, 2001] Turlach, B. A., Venables, W. N. et Wright, S. J. (2001). Simultaneous variable selection. Rapport technique, The University of Western Australia, Crawley WA 6009, Australia.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Series in Statistics. Springer, New York.

[Vieu, 1994] Vieu, P. (1994). Variable selection ? ? *Computational Statistics & Data Analysis*, 17.

[Wahba, 1985] Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 13:1378–1402.

[Wahba, 1990] Wahba, G. (1990). *Spline Models for Observational Data*. Numéro 59 de Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PA.

[Wahba et Luo, 1997] Wahba, G. et Luo, Z. (1997). Smoothing spline anova fits for very large, nearly regular data sets, with applications to historical global climate data. *Ann. Numer. Math.*, 4(1–4):579–597.

[Wahba et Wang, 1995] Wahba, G. et Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Stat. Probab. Lett.*, 25(2):105–111.

[Walker et Wright, 2002] Walker, E. et Wright, S. P. (2002). Comparing curves using additive models. *Journal of Quality Technology*, 34(1):118–129.

[Wand et Jones, 1995] Wand, J. R. et Jones, M. C. (1995). *Kernel Smoothing*, volume 60 de *Monographs on Statistics and Applied Probability*. Chapman Hall, New York.

[Wand, 2000] Wand, M. P. (2000). A central limit theorem for local polynomial backfitting estimators. *Journal of Multivariate Analysis*, 70:57–65.

[Weisberg, 2004] Weisberg, S. (2004). Comments on "least angle regression" by Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Annals of Statistics*, 32(2): 490–494.

[Wetherill, 1986] Wetherill, G. B. (1986). *Regression Analysis with Applications*, volume 27 de *Monographs on Statistics and Applied Probability*. Chapman & Hall.

[Wong et Kohn, 1996] Wong, C. M. et Kohn, R. (1996). A bayesian approach to additive semiparametric regression. *Journal of Econometrics*, 74(2):209–235.

[Wood, 2000] Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc. B*, 62(2):413–428.

[Wood, 2004] Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models (A paraître). *Journal of the American Statistical Association*.

[Xiang et Wahba, 1996] Xiang, D. et Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, 6(3):675–692.

[Ye, 1998] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131.

[Yee et Wild, 1996] Yee, T. W. et Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society, Series B*, 58:481–493.

[Yuan et Wahba, 2001] Yuan, M. et Wahba, G. (2001). Automatic smoothing for poisson regression. Rapport technique 1083, University of Winconsin.

[Zhang *et al.*, 2003] Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. et Klein, B. (2003). Variable selection and model building via likelihood basis pursuit. Rapport technique 1059r, University of Winconsin.

[Zhang et Wong, 2003] Zhang, S. et Wong, M. Y. (2003). Wavelet threshold estimation for additive regression models. *Annals of Statistics*, 31(1):152–173.